

Chapter 6

A Comparative Study of Learning Object Metadata, Learning Material Repositories, Metadata Annotation & an Automatic Metadata Annotation Tool

*Devshri Roy, Sudeshna Sarkar, Sujoy Ghose**

Abstract: One of the most important components of an e-learning system is the learning material. The popularity of e-learning has led to the development of many learning object repositories that store high quality learning materials specifically created for e-learning. High quality learning materials are expensive to create. So it is very important to ensure reuse of learning materials. Reuse of learning materials are made possible by semantically tagging them with standard metadata. In this chapter, we present a comparative study of available learning object metadata and learning object repositories. The learning material can be tagged either manually or automatically. Manual annotation is a time consuming and expensive process. We have explored the feasibility of tagging learning materials automatically with a set of IEEE LOM metadata specification. Here, we present a standard classification approach using probabilistic neural network to automatically identify the topic of the learning material. The classifier is tested and the result shows a fair degree of accuracy.

Keywords: Learning Object Metadata, Learning Object Repository, Metadata Annotation

1. 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 promising approach to facilitate and enhance learning through computer and communication technologies. One of the resources of E-learning is the online learning object repositories. Learning object repositories are essentially the storage of learning materials. Learning material repositories store good quality learning materials specifically created for e-learning. For efficient retrieval of learning materials according to the requirement of an e-learner, the learning materials are tagged with a set of metadata which describes educational artifacts such as topic of the document, type of the document etc. However, in order to facilitate the sharing and reuse of learning materials across different information repositories or learning management systems, it is recommended that the learning materials should be associated with some common metadata standard. Several metadata standards have emerged for the description of learning resources. The Dublin Core metadata initiative

* { droy, sudeshna, sujoy}@cse.iitkgp.ernet.in
Indian Institute of Technology, Kharagpur, India

(<http://dublincore.org/>) is an open forum engaged in the development of interoperable online metadata standards that supports a broad range of purposes and business models. Although Dublin Core attributes that contains metadata such as authors, title or granularity, are definitely useful for describing learning content, but Dublin Core does not contain attributes describing the pedagogical perspective of a document. In order to cope with the educational concerns, various metadata standards were defined such as IMS Metadata (<http://www.imsglobal.org/>), CanCore (<http://www.cancore.ca/>) and IEEE Learning Object Metadata (<http://ltsc.ieee.org/wg12/index.html>). The IMS Global Learning Consortium develops and promotes the adoption of open technical specifications for interoperable learning technology. The Advance Distributed Learning [5] (<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 SCORM-compliant courses are reusable, accessible, interoperable and durable. IEEE LOM aims to develop accredited technical standards, recommended practices, and guidelines for learning technology.

Most of the available online learning object repositories have been developed manually. The authors, contributors and developers of the open repositories have the responsibility of manually attributing meta information to the learning objects. In the Health Education Assets Library (<http://www.healcentral.org>) and iLumina (<http://www.ilumina-dlib.org/>), the contributors are required to follow strict guidelines and fill up many forms to carefully ensure that the learning objects associated to the repository are according to their requirements. In LearnAlberta Online Curriculum Repository (<http://www.learnalberta.ca/login.aspx>), the developer has to follow the specifications of resource development guideline such as learning object development guideline, metadata guidelines, instructional design guidelines etc.

Associating meta information to learning objects by humans is a labour intensive activity. Many contributors find the task of manual annotation and assigning of meta-tags uninteresting, and sometimes the tagging is not done satisfactorily. The development of a repository with manually annotated learning materials is expensive in terms of the time and effort required. Many people feel (Ochoa 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 *Bibliographic Control of Web Resources: A Library of Congress Action Plan* (<http://lcweb.loc.gov/catdir/bibcontrol/actionplan.pdf>) recognizes the need of automatic metadata annotation and targets the development of *automatic metadata generation tool*.

Automating part of the process of building repositories will facilitate the building of large learning object repositories with little effort. It will be useful if the contributors/ developers can be relieved from filling up many forms while submitting the learning material into the repository. This can be made possible by automating the tagging of the learning objects. Further, the process of collection of learning objects can be made easier by building a system that is able to select learning objects from existing materials present in various sources like the Internet, as well as learning objects present in the other repositories.

Automatic indexing is especially important if we wish to harness the large number of documents available in the Web. The Internet includes a huge storehouse of learning materials. In order to be able to make use of such materials in e-learning systems, one will have to search and filter the learning materials from this storehouse and use these for building learning object repository. These materials are often not tagged. The challenge is to make such materials usable by learning management systems by incorporating appropriate meta tags automatically.

We have worked to take forward the process of automatic metadata generation from learning materials. We have worked for automatically extracting a subset of IEEE LOM metadata from learning materials (Roy, Sarkar and Ghose 2008). The work presented in this chapter focuses on the semantic tagging of learning materials with a very important metadata *topic* of the learning material. The metadata is automatically extracted from the document. Generally, the syllabus of any subject is defined in terms of a set of topics and it is expected that learners' objective is to learn the topics of the syllabus of their grade level. When a learner gives a keyword in *keyword based* search engine for searching learning materials of their interest, he may get many irrelevant documents. The problem with *keyword based* search is that for the given keyword it selects all documents which contain the keyword and the same keyword may belong to many

different topics. If the topic of a learning material is identified and annotated with the learning content, it helps in retrieving learning materials according to the learner's interested topic.

In this chapter, we present

- The different available open metadata standards
- A survey of some of the available learning object repositories and their comparison in terms of metadata standard used, type of annotation, subject domain, types of searching facilities etc.
- Advantages of automatic annotation over manual annotation and previous work done on automatic annotation
- An automatic annotation tool developed by us and the machine learning approach used to identify the topic of the learning materials using an example based probabilistic neural network classifier.

2. THE METADATA STANDARDS

In recent years, several open metadata standards have been emerged. The term "meta" comes from a Greek word that denotes "alongside, with, after, next". Metadata can be thought of as *data about other data*. The metadata system is common in libraries. The library catalog contains a set of records with elements that describe a book or other library item such as the author, title, date of creation or publication, subject coverage, and the index number specifying the location of the item on the shelf. Metadata is the Internet-age term for information that librarians traditionally have put into catalogs, and it most commonly refers to the descriptive information about web resources. A metadata record consists of a set of attributes or elements, necessary to describe the resource. The advantages of tagging documents with metadata are as follows.

- It makes search, acquisition, and use of learning objects easier by the learner.
- It enables the retrieval module of a retrieval system to retrieve personalized learning objects for an individual learner. It helps the tutoring module of a tutoring system in the tutoring processes.
- It facilitates reusability of learning objects i.e. the learning objects can be reused in different instructional contexts.
- It facilitates interoperability of learning objects i.e. the sharing and the exchange of learning objects across any technology supported learning system.

The worldwide interest in metadata standards and practices has exploded with the growth of e-learning and digital libraries or learning object repositories.

The Dublin Core Metadata Initiative (<http://dublincore.org/>) is an open forum engaged in the development of interoperable online metadata standards that supports a broad range of purposes and business models. The Dublin Core standard includes two levels: Simple and Qualified. The simple Dublin Core contains fifteen elements. The elements are *Title, Subject, Description, Type, Source, Relation, Coverage, Creator, Publisher, Contributor, Rights, Date, Format, Identifier* and *Language*. The qualified Dublin Core includes three additional elements *Audience, Provenance* and *RightsHolder*, as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in a way that may be useful in resource discovery. The Dublin Core metadata contains metadata elements useful for general purpose applications but it does not contain attributes describing the pedagogical perspective of a document. In order to cope with educational concerns, various other metadata standards have been defined such as IMS Metadata, SCORM Metadata, CanCore and IEEE Learning Object Metadata.

The IEEE Learning Object Metadata (<http://ltsc.ieee.org/wg12/index.html>) aims to develop accredited technical standards, recommended practices, and guides for learning technology. This standard specifies learning object metadata. It specifies a conceptual data schema that defines the structure of a metadata instance for a learning object. For this standard, a learning object is defined as any entity digital or non-digital that may be used for learning, education or training. For this standard, a metadata instance for a learning object describes relevant characteristics of the learning object to which it applies. Such characteristics are grouped in *General, Life cycle, Meta-metadata, Educational, Technical, Rights,*

Relation, Annotation, and Classification categories. It is intended to reference by other standards that define the implementation descriptions of the data schema so that a metadata instance for a learning object can be used by a learning technology system to manage, locate, evaluate or exchange learning objects.

The IMS Global Learning Consortium (<http://www.imsglobal.org/>) develops and promotes the adoption of open technical specifications for interoperable learning technology. The IMS Content Packaging Information Model defines a standardized set of structure that can be used to exchange the learning content. These structures provide the basis for standardized data bindings that allow the software developers and the implementers to create instructional materials that are interoperable across authoring tools, learning management systems, and run time environments.

The Advance Distributed Learning Initiative (<http://www.adlnet.org>) aims to establish a distributed learning environment that facilitates the interoperability of e-learning tools and course contents on a global scale. The Sharable Content Object Reference Model is a collection of standards and specifications adapted from multiple sources to provide a comprehensive suite of e-learning capabilities that enable interoperability, accessibility and reusability of web-based learning content. The SCROM is often described as a bookshelf that houses specifications that are originated in other organizations like ARIADNE, AICC, IMS and IEEE. The SCROM has three parts, *Overview, Content Aggregation Model* and the *Run Time Environment*. The first part covers the overview about the model, vision and the future. The second part, the *Content Aggregation Model (CAM)* covers many specifications. The first specification in the CAM (from IEEE/ARIADNE/Dublin Core and IMS) is the "Learning Object Metadata". This is a dictionary of tags that are used to describe the learning content in a variety of ways. The second specification in the CAM is the XML "binding" for the metadata tags (from IMS). This defines how to code the tags in XML so that they are machine (and human) readable. The third specification in the CAM is the IMS Content Packaging Specification. This defines how to package a collection of learning objects, their metadata, and the information about how the content is to be delivered to the user. Packaging defines how learning contents of all types can be exchanged between different systems in a standardized way. The third part is the *Run Time Environment*. During the evolution of the SCORM suite of specifications, a standardized way is needed for sending the information back and forth between the learner (content) and the learning management system. An application program interface is defined that provides a standard way of communication with the learning management system, regardless of what tools are used to develop the content.

The CanCore Learning Resource Metadata Initiative (<http://www.cancore.ca/>) enhances the ability of educators, researchers and students in Canada and around the world to search and to locate materials from online collections of educational resources. CanCore is fully compatible with the IEEE Learning Object Metadata standard and the IMS Learning Resource Meta-data specification

A complete list of metadata elements of IEEE LOM, Dublin Core and CanCore metadata standards is given in Table 1.

Table 1 A complete list of metadata elements of IEEE LOM, CanCore LOM and Dublin Core Metadata Standards

IEEE LOM	CanCore LOM	Dublin Core Metadata
1. General	1. General	1. Contributor
1.1 Identifier	1.1 Identifier	2. Coverage
1.2 Catalog	1.1.1 Catalog	3. Creator
1.3 Entry	1.1.2 Entry	4. Date
1.4 Title	1.2 Title	5. Description
1.5 Language	1.3 Language	6. Format
1.6 Description	1.4 Description	7. Identifier
1.7 Keyword	1.5 Keyword	8. Language
1.8 Coverage	1.6 Coverage	9. Publisher
1.9 Structure	1.7 Structure	10. Relation
1.10 Aggregation Level	1.8 Aggregation Level	11. Rights
2. Life Cycle	2. Life Cycle	12. Source
2.1 Version	2.1 Version	13. Subject
2.2 Status	2.2 Status	14. Title
2.3 Contribute	2.3 Contribute	15. type
2.4 Role	2.3.1 Role	
2.5 Entity	2.3.2 Entity	
2.6 Date	2.3.3 Date	
3 Meta-Metadata	3 Meta-Metadata	
3.1 Identifier	3.1 Identifier	
3.2 Catalog	3.1.1 Catalog	
3.3 Entry	3.1.2 Entry	
3.4 Contribute	3.2 Contribute	
3.5 Role	3.2.1 Role	
3.6 Entity	3.2.2 Entity	
3.7 Date	3.2.3 Date	
3.8 Metadata Schema	3.3 Metadata Schema	
3.9 language	3.4 Language	
4. Technical	4. Technical	
4.1 Format	4.1 Format	
4.2 Size	4.2 Size	
4.3 Location	4.3 Location	
4.4 Requirement	4.4 Requirement	
4.5 OrComposite	4.4.1 OrComposite	
4.6 Type	4.4.1.1 Type	
4.7 Name	4.4.1.2 Name	
4.8 Min Version	4.4.1.3 Minimum Version	
4.9 Max Version	4.4.1.4 Maximum Version	
4.10 Installation Remarks	4.5 Installation Remarks	
4.11 Other Platform Requirements	4.6 Other Platform Requirements	
4.12 Duration	4.7 Duration	
5. Educational	5. Educational	
5.1 Learning Resource Type	5.1 Interactivity Type	
5.2 Interactivity Level	5.2 Learning Resource Type	
5.3 Semantic Density	5.3 Interactivity Level	
5.4 Intended End User Role	5.4 Semantic Density	
5.5 Context	5.5 Intended End User Role	
5.6 Typical Age Range	5.6 Context	
5.7 Difficulty	5.7 Typical Age Range	
5.8 Typical Learning Time	5.8 Difficulty	
	5.9 Typical Learning Time	

5.9 Description 5.10 Language 6. Rights 6.1 Cost 6.2 Copyright & other restrictions 6.3 Description 7. Relation 7.1 Kind 7.2 Resource 7.3 Identifier 7.4 Catalog 7.5 Entry 7.6 Description 8 Annotation 8.1 Entity 8.2 Date 8.3 Description 9. Classification 9.1 Purpose 9.2 Taxon Path 9.3 Source 9.4 Taxon 9.5 Id 9.6 Entry 9.7 Description 9.8 Keyword	5.10 Description 5.11 Language 6. Rights 6.1 Cost 6.2 Copyright & other restrictions 6.3 Description 7. Relation 7.1:Kind 7.2:Resource 7.2.1:Identifier 7.2.1.1:Catalog 7.2.1.2:Entry 7.2.2:Description 8 Annotation 8.1 Entity 8.2 Date 8.3 Description 9. Classification 9.1:Purpose 9.2:Taxon Path 9.2.1:Source 9.2.2:Taxon 9.2.2.1:Id 9.2.2.2:Entry 9.3:Description 9.4:Keyword	
---	--	--

3. LEARNING OBJECT METADATA BASED REPOSITORIES

Learning objects are the content components that are meant to be reusable in different contexts. These learning objects are associated with metadata, so that they can easily be searchable and manageable. As the international standardization in this area is making a fast progress, the number of learning object repositories are also growing rapidly. A LOM repository or learning object repository stores both learning objects (LOs') and their metadata.

A learning object repository allows users to search and retrieve learning materials from the repository. It supports simple and advanced search, as well as browsing through the materials. In simple search, it returns the search results against the given input keywords. A learner needs to search specific learning materials according to his requirements. The advanced search allows users to specify values for specific metadata elements to filter learning materials to meet the user's specific need. Browsing allows the users to descend in a tree of disciplines and sub-disciplines to access the learning objects available in the repository. Here we will discuss the features and characteristics of some of the existing learning object repositories.

ARIADNE (<http://www.ariadne-eu.org>), the European digital library project, was initiated in 1996 by European commission's telematics for education and training program. Since then, an infrastructure has been developed in Belgium and Switzerland for the production of reusable learning content, including its description, distributed storage, and discovery, as well as its exploitation in structured courses. The core of this infrastructure is a distributed library of digital, reusable educational components called the Knowledge Pool System (KPS). It is actively used in both academic and corporate contexts. The KPS content (*Duval and Hodgins 2004*) is oriented more toward technical science, strongly represented by computer science, economics, electronics, health science, transportation and life science. The KPS is a reference library. The KPS includes descriptions (metadata), as well as the documents themselves, making it easier to replicate documents across all nodes of the system, ensuring convenient access without excessive download times.

The ARIADNE includes a set of metadata from general, technical and educational categories. The ARIADNE includes the traditional metadata *title*, *author* and *publication date*, which are generally used in a library. It includes metadata describing the technical characteristics of the document i.e. uncompressed size of the document and the requirements with respect to the computing platform. Educational metadata includes *document type* (active or expository), *format* (questionnaire, simulation, hypertext and others), *usage remarks* (explaining how the documents can be used in a sound way in any learning environment), *didactical context* and *course level* (describing the kind of learners for whom the document is intended), *difficulty level*, *interactivity level*, *semantic density* and *pedagogical duration*. A user can search learning materials using a tool SILO (search and index learning objects). It provides the facility of simple search using keywords, advanced search and federated search. Advanced search can be done on the *document title*, *usage right*, *author's name* and the *main concept*. Federated search provides the facility of searching learning materials from other repositories namely MERLOT, EdNA, CGIAR along with the ARIADNE. The authoring tool of ARIADNE allows indexing pedagogical material and inserting it into the knowledge pool system.

The National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) (<http://www.smete.org>) is constructed to meet learner's and educator's need. The digital library offers direct access to the learning resources. It promotes learning through personal ownership and management of the learning process while connecting the learner with the content and communities of learners and educators. Users can create their profile and submit it to the repository. It recommends learning objects based on their profile and past user interaction with the repository. Contents and services provided through the digital library includes multimedia courseware, digital problem sets and exercises, educational software applications, related articles, journals and instructional technology services for the educators and the students, both commercial and non-commercial, organized and labeled for the purpose of education and instruction. To search and obtain more precise learning resources from this digital repository, user can give input to the different search fields apart from the keyword. The search fields are *learning resource type* (applet, case study, course, demonstration, educational games, images/diagrams/graphs, links, laboratory/experimental support, lecture/presentation, lesson plan, practical problems, exercise etc), *grade* (starting from primary education to higher education), *title*, *author/creator*, *collection* and the *publication year*. The search results shows the URL of the learning object along with meta information title, author, publisher, subject, description, grade, format, rating. It allows the users to browse through the repository over subject headings.

Multimedia Educational Resources for Learning and Online Teaching (<http://www.merlot.org>) is an open repository designed primarily for faculty and students. Links to online learning materials are collected here along with annotations such as peer reviews and assignments. The learning materials are peer reviewed by the reviewers. The primary purpose of these reviews is to allow the faculty from any institution of higher education to decide that the online teaching-learning materials that they are examining will work in their course(s). Peer Reviews are performed by evaluation standards that divide the review into three dimensions: *Quality of Content*, *Potential Effectiveness as a Teaching Tool*, and *Ease of Use*. Each of these dimensions is evaluated separately. In addition to the written findings (review) by the reviewers, there is a rating for each of the learning materials with three dimensions (1-5 stars, 5 being the highest). A review must average three stars (or textual equivalent) to be posted to the MERLOT site. It provides the facility of simple search, advanced search and browsing by discipline. Advanced search can be done on fields *subject*, *sub-category (topic)*, *material types*, *title*, *content URL*, *description*, *primary audience*, *Technical format*, *learning management system*, *language* and *author*.

The Health Education Assets Library (<http://www.healcentral.org>) is a digital library that provides freely accessible learning resources that meet the needs of today's health sciences educators and learners. The HEAL provides the facility of search, either with a simple keyword or with an advanced search. Advanced search can be done on fields like *learning resource type*, *title*, *description*, *contributors*, *medical subject heading*, and *primary audience*. User can browse the learning materials by Medical Subject Heading (MeSH). User can also view the detailed cataloging information (metadata) about the resource. The metadata schema is based on the international IEEE LOM standard and includes extensions specific to the health sciences. The health science specific elements in the HEAL metadata schema are *Specimen Type*

(cell, tissue, organ, organ system), *Radiograph* (radiology technology used to generate the multimedia item), *Magnification* (magnification of microscopic image), *Disease Process* (indicate disease process displayed, discussed, or implied in the item), and *Clinical History* (clinical history of patient). In addition to the health science specific extensions, there is a set of elements which satisfy the functional requirement of the HEAL system. These are *Inappropriate for minors*, *annotated* (indicates that the multimedia is labeled or not), *Context URL* (The context in which the item can be used such as course, a case etc.) and the *Context URL description*.

The Education Network Australia (<http://www.edna.edu.au/>) online supports and promotes the benefits of Internet for learning education and training in Australia. It provides a database of learning resources useful for teaching and learning. One can search 20,000 educational resources for schools and for higher education. It offers standard and advanced search facilities. The EdNA advanced search provides the facility of searching on categories like *Adult and Community Education (ACE)*, *Vocational Education and Training (VET)*, *General References*, *Higher Education*, *Educational Organizations*, *School Educations* or all the above categories. Apart from the EdNA online repository a user can search from other repositories like *Government Education Portal*, *ABC online*, *Cultural and Recreation Portal*, *Gateway to Educational Materials (GEM)*, *Multimedia Educational Recourses for Learning and Online Teaching (MERLOT)*, *Picture Australia*, *UNESCO/NCVER database (VOCED)*, *VLORN (Vocational education and training learning object repository)*. The EdNA Metadata Standard is based on the Dublin Core Metadata Standard. Consistent with the extensibility principles of Dublin Core, the EdNA Metadata Standard V1.1 includes additional elements and element qualifiers, for the specific application to the Australian education domain and to support the operational requirements of EdNA Online. The additional metadata elements are *Audience* (a category of user for whom the resource is intended), *Approver* (email of a person or organization approving the item for inclusion in EdNA Online), *Category-Code* (a numerical code derived from the database tables which support the EdNA Online browse categories), *Entered* (data item was entered as an entry in the online item database, used for management purposes), *Indexing* (to what extent should EdNA online spidering software follow links from this page), *Review* (a third party review of the resource), *Reviewer* (name of the person and/or organization or authority affiliated for reviewing), *Version* (version of the EDNA Metadata Standard applied).

The iLumina (<http://www.ilumina-dlib.org>) is a digital library of sharable undergraduate learning materials for chemistry, biology, physics, mathematics, and computer science. It provides the facility of simple search, advanced search and browsing of learning materials. The search can be done by giving *keyword*, *subject*, *author*, *title*, *journal title*, *author/title*, *ISBN/ISSN numbers*. A user can also fill options like *type of material*, *language*, *year*, *sort* etc. The search results are sorted according to the date or the title or relevance. The learning resources in iLumina are cataloged in the *Machine-Readable Cataloging* (<http://www.loc.gov/marc/marcdocz.html>) metadata formats, which capture both technical and education-specific information about each resource. NSDL metadata standard consists of Dublin Core set of 15 basic elements, their associated element refinements plus the three IEEE LOM elements recommended by the DC Education Working Group.

The goal of the LearnAlberta Online Curriculam Repository (<http://www.learnalberta.ca/>) is to create a collection of learning object repositories in the field of education with access through a set of linked portals. It contains learning materials for students of various grade levels ranging from kindergarten to grade 12. A user can search learning materials by giving keywords or can browse learning resources using grades.

Campus Alberta repository of educational objects (<http://www.ucalgary.ca/commons/careo/>) is a learning object repository that holds links to learning objects, as well as some learning objects themselves. The entry page displays the newest and most popular learning objects. Users can search learning materials by giving keywords or can perform advanced search by giving extra fields like *title*, *discipline*, *technical format*, *learning resource type* etc. Users can also browse learning objects based on the discipline. A personal profile give access to a workspace (My Objects) with bookmarks. A user can access a history of objects that had been downloaded by him.

The LydiaLearn (<http://www.lydialearn.com/>) is a commercial network of learning object repositories. It has a central site (Lydia Global Repository or LGR) where users can search for learning materials after registering. The metadata for each learning object describes ownership and price, and users can use Lydia's transaction basket to buy learning objects.

The summary of the different available learning object repositories with a comparative discussion of their features and facilities available is given in Table 2 & 3.

Table 2 Comparison of learning object repositories based on the metadata standards used for annotating learning materials

Learning Object Repository	Features		
	Meta-data Standard	Metadata annotation	Document Repository
ARIADNE	IEEE LOM	Manual and Automatic Metadata Generation	Links
SMETE	IEEE LOM	Manual	Links
MERLOT	IEEE LOM	Manual	Links
HEAL	IEEE LOM	Manual	Links
EdnA	Dublin Core profile	Manual	Links
iLumina	IEEE LOM	Manual	Document Repository
Learn-Alberta	IEEE LOM	Manual	Document Repository + Links
CAREO	IEEE LOM	Manual	Document Repository
Lydia	IEEE LOM Profile (SCROM)	Manual	Document Repository

Table 3 Comparison based on the searching and browsing facilities in different subject domains provided by the Learning Object Repositories

LOR	Facilities			
	Simple Search	Advanced Search	Browsing	Subject Domain
ARIADNE	Keyword search	Document title ,Usage right,Author's name, Main Concept	No	Science
SMETE	Keyword search	Keyword, Learning resource type ,Grade, Title, Author, Collection, Publication year	Browse by discipline	Science, Mathematics, Engineering, Technology
MERLOT	Keyword search	Subject, Subcategory, Material type , URL, Description, Primary Audience, Technical format, Language, Author's name	Browse by discipline	Science, Engineering, Business, Justice, Music, Health Science
HEAL	Keyword search	Learning resource type, Title, Description, Contributors, Medical subject heading term	Browse by Medical subject heading or by collection	Health Science
EdnA	Keyword search	Adult & community education, Vocational education & training, General references, Higher education, Educational organizations, School education	Browse by discipline	Education
iLumina	Keyword, Subject, Author, Journal title, ISBN/ISSN number	Type of material, Year, Language	Browse by discipline, subject, topic	Science, Mathematics, Engineering & Technology
Learn-Alberta	Keyword search	N/A	Browse by grades	Mathematics, physical Education, Science, Social Studies, Language, Fine Arts, CTS, CALM
CAREO	Keyword search	Title, Description, Keyword, Discipline, Technical format, Learning resource type, Intended user role	Browse by discipline	Science, Engineering, Business, Justice, Music, Health Science

4. METADATA ANNOTATION

High quality metadata is essential for reusability and for effective retrieval of learning objects. Metadata can be generated in two ways: Manual and Automatic.

In the case of manual metadata generation, the maintainer of the learning object repository generates metadata manually. Sometimes the author of the learning material submits the metadata, which is then

assessed and organized by the maintainers. The learning materials in most of the learning object repositories discussed in section 2.2 are manually annotated. The manual annotation often results in a very high quality metadata but is a very time consuming and labor intensive activity. It is advantageous to describe a constantly changing and evolving learning materials with some degree of automation (Downes 2004; Duval et al. 2004; Simon et al. 2004).

Automatic metadata generation depends on machine processing. The advantage of automatic metadata generation is that an automated tool can discover much more data much more quickly than humans. *Metadata harvesting* and *metadata extraction* has been identified as two methods of automatic metadata generation. *Metadata harvesting* is the process of automatically collecting resource metadata already embedded in or associated with a resource. The harvested metadata is originally produced by humans or by semi-automatic processes supported by the software. *Metadata extraction* is the process of automatically pulling metadata from the resource's content. The resource's content is mined to produce the structured standard metadata.

Automatic metadata generation from learning materials is an upcoming and challenging area of research. Some work has been done on automatic metadata generation, which is discussed below.

DC-dot (<http://www.ukoln.ac.uk/metadata/dcdot/>) is a metadata generator developed by UKOLN (UK Office for Library and Information Networking) based at the University of Bath. The DC-dot is open source and can be redistributed or modified under the terms of the GNU General Public License as published by the Free Software Foundation. The DC-dot produces Dublin Core metadata and can format output according to the number of different metadata schemas like USMARC, RDF, and IMS etc. Metadata creation with the DC-dot is initiated by submitting a URL. It copies resource *identifier* metadata from the web browser's *address prompt*, and harvests *title*, *keywords*, *description*, and *type* metadata from the resource's META tags. If resource META tags are absent, DC-dot automatically generates *keywords* by analyzing anchors (hyperlinked concepts) and presentation encoding such as bolding and font size. It also generates *type*, *format* and *date* metadata automatically.

Han et al. (Han, Giles and Manavoglu 2003) proposes a machine learning method using support vector machine for automatic metadata extraction of Dublin Core Metadata. Sometimes the metadata standard does not meet the requirements of a particular learning system and requires local extensions and modifications. Han et al. have extended the Dublin Core metadata and included additional metadata *author's affiliation*, *author's address*, *author's email*, *publication number* and *thesis type*. These additional metadata helps in building unified services for heterogeneous digital libraries, while at the same time enabling sophisticated querying of the databases and facilitating construction of the semantic web. The reported metadata extraction results are based on the experiments conducted on research papers. Most of the information like author's name, affiliations, address, and the title are collected from the header of the research paper. The header consists of all the words from the beginning of the paper up to either the first section, usually the introduction, or to the end of the first page, whichever occurs first. They have illustrated the dominance of SVM based metadata extraction algorithm over Hidden Markov Model based systems. They have also introduced a method for extracting individual names from the list of authors within the same network and present a document extraction method using SVM classification, combining chunk identification. A new feature extraction method and an iterative line classification process using contextual information are also introduced in their work.

The work by Jenkins C. and Inman D. (Jenkins and Inman 2000) propose a technique for automatically generating qualified Dublin Core metadata on a web server using a Java Servlet. The metadata is structured using the Resource Description Framework (<http://www.w3.org/RDF/>) and expressed in Extensible Markup Language (<http://www.w3.org/XML/>). The description covers ten out of fifteen standard metadata. The metadata elements are *title*, *creator*, *subject*, *description*, *publisher*, *date*, *format*, *identifier*, *relation*, and *rights*. When the URL of a document is passed to the servlet, it harvests the *title*, *date*, and *format* from the html tags. The metadata *description* is represented by an abstract, which are the first 25 words found in the body of the document. The metadata *subject* is represented by a series of keywords. The keywords are extracted by parsing the actual content. The metadata *relation* is used to represent a resource that is hyper-linked from the current resource.

Another work by Li y., Zhu Q., and Cao Y. (Li, Zhu and Cao 2004) also automatically generates the qualified Dublin Core metadata from web pages. It extracts total 10 metadata elements from web pages. The 9 elements *title*, *creator*, *description*, *publisher*, *date*, *format*, *identifier*, *relation* and *rights* are generated by the same techniques as used by Jenkins C. and Inman D. The *subject* element is obtained using neural network. The *subject* element of a resource is a term weight vector in multidimensional space, which represents the whole content of the resource. In the vector, each word is assigned a weight, which represents its degree of importance. The principal component analysis (PCA) technology of neural network is used to select the effective terms to represent a resource.

In the above few paragraphs, we have discussed the work on the automatic extraction of Dublin Core metadata useful in general purpose applications. There is a need to extract attributes describing the pedagogical perspective of a document to cope with educational concerns. The IEEE LOM standard contains attributes describing the pedagogic characteristics of the document. The work has been started on automatic extraction of the educational category metadata elements of the IEEE LOM specification.

Jovanovic et al. (Jovanovic, Gasevic and Devedzic 2006) present an ontology-based approach for automatic annotation of learning objects based on IEEE LOM. In their work, the metadata elements *title*, *description*, *unique identifier*, *subject* and *pedagogic role* are automatically generated from the learning object. They mainly used content mining algorithms and certain heuristics for determining these metadata elements to annotate the learning content. They have annotated documents only in slide format. The whole document forms the *learning object* and the different slides of a document forms the *content object*. In their work, they focus on smaller units for reusability on a finer scale and annotated each slides (content objects) of a document along with the whole document (learning object). The metadata element *title* of the content objects is simply extracted from the title of the slides. If the slides do not have the title, then those instances of content objects do not assigned the *title* element. The *subject* annotation of content objects depends on the author's information. The *subject* annotation of each of the content objects (or slides) is based on domain ontology and derived from the subject metadata provided by the author during submission of the learning object. In their work, they automatically generate the pedagogic roles like *example*, *summary* and *references*. To infer the pedagogic role of the learning content, they opted heuristic based approach. To identify the pedagogic role, they observed the presence of some specific terms along with some patterns. If the title of the slide contains terms like *summary*, *conclusion* and the body of the slide is structured in the form of a list, the slide is annotated with pedagogic type *summary*. Similarly in the case of *reference* type, the learning content contains terms like references, reference list, bibliography etc. The *description* of the whole learning object is generated by the combining the different attributes like *type*, *title*, *subject creator*, and *date* etc.

Roy et. al (Roy, Sarkar and Ghose 2008) has worked on automatic extraction of a subset of IEEE learning object metadata. In their work, they have automatically extracted the following metadata elements: *identifier*, *description*, *format*, *size*, *topic* and *type of the learning resources*. Different learners require different learning content depending on their learning style (Papanikolaou, 2002). The *type of learning resources* is a very important metadata, which helps in retrieving a specific type of learning material according to the learner's requirement. In their work, Roy et. al. classifies the learning materials into three categories *narrative text*, *experiment*, *exercise* (a subset of IEEE LOM 5.2 specification). The *narrative text* type documents generally contain definitions, statement of laws or facts about concepts. The definition and uses of narrative text type documents are discussed in the CanCore guidelines (<http://www.cancore.ca/en/help/44.html>). The category *exercise* includes the documents containing exercises, numerical problems, questions etc. For the better understanding of any theory, law or principle, students perform experiments. The *experiment type* documents contain instructions and discussion on experiments. They have used machine learning approach to classify the learning materials. They have identified some of the surface level features of the text. The feature set consists of a set of specific verbs, trigger words, phrases and special characters. *Narrative text type* documents contain discussion about a concept or concepts. Therefore verbs like "define", "known", "state", "described", "explained", "discussed", "illustrated" etc. are frequently found in *Narrative text type* documents. *Experiment type* of 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 play an important role in describing documents. Documents belonging to the category *experiment* usually contain words like “*introduction*”, “*objective*”, “*results*”, “*goal*” etc. *Exercise* 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. *Exercise* type documents may contain interrogative sentences which can be identified as they end with question mark. These character level cues are important and used for classification. They have used the above discussed features to classify the documents into different types using neural network.

Kris Cardinaels et al. (*Cardinaels, Meire and Duval 2005*) developed a framework for automatic metadata generation of IEEE Learning Object Metadata as a web service. They tried to generate a metadata set that contains all the mandatory elements defined in the ARIADNE application profile. Metadata elements are *document type, package size, publication date, creation date, operating system type, access right, main discipline, language, format, title, and author’s detail*. *Author’s detail* includes his *postal code, affiliation, city, telephone, department, and email*. They proposed an idea of deriving metadata from two different sources. The first source is the learning object itself; the second is the context in which the learning object is used. Metadata derived from the object itself is obtained by the content analysis, such as keyword extraction, language classification and so on. The contexts typically are learning management systems in which the learning objects are deployed. A learning object context provides the extra information about the learning object that can be used to define the metadata. The proposed framework for automatic metadata generation consists of two major groups of classes, namely Context-based indexers and Object-based indexers. As discussed above, the Context-based indexers use a context to generate metadata. When an object is used in a specific context and the data about that specific context are available, then these data gives the information for annotation of the object. The Object-based indexers generate metadata based on the learning object itself, isolated from any other learning object or learning management system. A Metadata-Merger combines the results of the different indexers into one set of metadata.

Dehors et al. (*Dehor, Faron-Zucker, Stromboni et al. 2005*) have proposed a methodology for semi-automatic annotation of learning resources based on the document layout features. They assume that every course is based on a learning or pedagogical model, which includes some pedagogical strategy. So, first the author is asked to explicit the pedagogical strategy for his/her course. The annotation task begins by interviewing the author of the document to determine the relations between the employed presentational model of the existing document and how this model supports the envisioned educational strategy. Once, this model is defined, a phase of content re-authoring takes place to ensure that the employed visual features are compliant to the established instructional model. Only then it is possible to automatically identify and annotate content units according to their pedagogical role. The employed pedagogical ontology is generated on the fly and includes concepts that formalize elements of a content author’s specific pedagogical strategy. Although this approach tends to be more precise in recognition of instructional roles of content units, but it requires more human effort: interviewing the author and content re-authoring.

Work has been done on the semantic annotation of web documents with general meta information. The KIM platform (*Popov, Kiryakov, Kirilov et al. 2003*) provides a novel knowledge and information management infrastructure and services for automatic semantic annotation, indexing, and retrieval of documents. The KIM platform is based on the PROTON ontology (<http://proton.semanticweb.org/>) and a knowledge base providing extensive coverage of entities of general importance. They worked on the automatic semantic annotation on general type of Meta information such as *organization, person, date, location, percent, and money*.

Piggy Bank (*Huynh, Mazzocchi and Karger 2005*) is a tool integrated into the contemporary web browser that lets users to collect information from various websites, presents them in ontology based format and annotate them with metadata. It invokes the screen scrapers to re-structure information within web pages into semantic web format (RDF format). Semantic Bank is a repository of RDF triples to which a community of PiggyBank users can contribute and share the information they have collected.

We have also worked on the automatic metadata annotation of learning materials. The automatic metadata annotation tool developed for annotating learning materials with metadata *topic* is discussed in the next section.

5. AUTOMATIC METADATA ANNOTATION TOOL

Most of the work done on automatic extraction concentrated on generating IEEE general category (1, IEEE LOM specification), Meta Metadata (3, IEEE LOM specification) and technical category metadata (4, IEEE LOM specification) like *identification number*, *creator of the learning material*, *publisher*, *format*, *date* etc. (Han 2003; Cardinaels et al. 2005; Li et al. 2004). Some work has been done on automatic extraction of IEEE *educational category metadata*. Roy et. al. have worked on automatically annotating the learning materials with the IEEE educational category metadata (LOM no. 5.2) *type of learning resource* (Roy, Sarkar and Ghsoe 2008).

The IEEE classification category metadata (9, IEEE LOM specification) are also very important and useful metadata for retrieving learning materials for e-learners. The attribute *taxon path* (9.2, IEEE LOM specification) indicates the taxonomic path with respect to the topic tree in a subject domain and gives the topic of the document. The attribute *taxon* (9.2.2, IEEE LOM specification) is a node in the topic taxonomy that has a defined label or term (which gives the name of the topic). An ordered list of taxons creates a taxonomic path.

The curriculum is a set of topics. In the school syllabus there are several subjects like Physics, Chemistry, Biology, History etc. Each subject consists of several topics. Each topic may again be divided into several subtopics. A portion of the syllabus for the subject Physics is shown in Figure 1 in hierarchical structure. The same can be represented in tree structure as shown in Figure 2. The subject *Physics* consists of many topics like *Optics*, *Kinematics*, *Electricity* etc. These topics in turn consist of many subtopics. The topic *Optics* contains many subtopics like *Spherical mirror*, *Lens* etc. This hierarchical structure can be represented by the topic taxonomy.

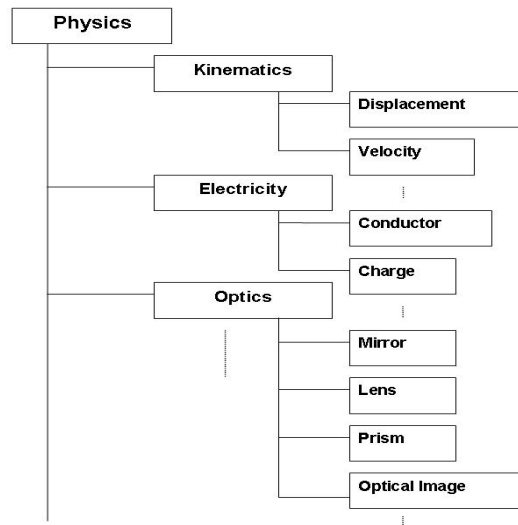


Figure 1 A portion of the syllabus for the subject Physics

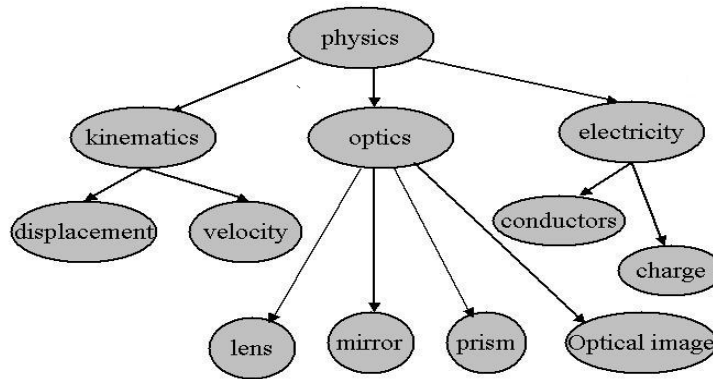


Figure 2 A portion of the syllabus shown in Figure 1 is represented in tree structure

The e-learners' requirement is generally given in terms of topics in a syllabus. It is expected that a student's interest is to learn the topics of the syllabus of his grade level. For learning a topic of the syllabus, a learner needs learning materials on that topic. A E-learner can retrieve learning materials either from the World Wide Web or from the learning object repository or can browse the topic tree to access learning materials on the topics of her interest. The learning material retrieval system needs to identify the learning materials belonging to a topic according to the learner's requirement. If the learning materials in the repository are semantically tagged with the metadata *topic*, then it becomes easier to search and identify documents according to the learner's interest. We want to automatically classify learning materials into different topics as given in the syllabus, so that a learner can search or navigate documents on topics in the topic taxonomy according to the curriculum requirement. In other words, we want to automatically identify the texons (9.2.2, IEEE specification) in the topic Taxonomy.

Researchers have carried out the work on automatic generation of topics from web documents in the past. Haruechaiyasak et al (*Haruechaiyasak, Shyu, Chen et al. 2002*) proposed a method of automatically classifying the web documents into a set of categories using the fuzzy association. The fuzzy association is used to capture the relationships among different index terms or keywords in the document. Each pair of words has an associated value to distinguish itself from the others. Therefore, the ambiguity in the word usage is avoided. They showed that the result with this approach is better than the result obtained with the vector space model.

Gelbukh et al (*Gelbukh, Sidorov and Guzman-Arenas 1999*) have given a method of document classification on a hierarchical dictionary of topics. The hierarchical links in the dictionary are supplied with the weights that are used for detecting the main topics of a document. The dictionary consists of two major parts, the vocabulary and the hierarchical structure. The vocabulary contains keywords. The hierarchical structure represents the topics. The links in the hierarchy have different weights. These weights give the strength of the relationship of the keywords to the given topics. For example, the word *Italy* belongs to the topic *Europe*, thus, the weight of this link is 1. On the other hand, the word *London* can refer to a city in *England* but with much less probability, in *Canada*, consequently the weight for the link between *London* and *Canada* is very less. To obtain the topics of documents, the keywords in a document are compared with hierarchical dictionary of topics.

There are some research works, where ontology of the domain has been used for automatic topic identification. In the work of ontology based automatic annotation of learning content (*Jovanovic Gasevic and Devedzic, 2006*), Jovanovic et al. annotate the documents with *subject (topic)* using the domain ontology. They annotate the documents, which are available in slide format. The whole document is the *learning object* and the different slides of the document form the *content object*. Initially, the author

provides the *subject (topic)* of the *learning object* during submission. They generate the metadata elements of the *content objects* based on the *subject (topic)* of the *learning object* provided by the author. The annotation of the different *content objects* (or slides) is done by looking at the related concepts of the *subject* of the *learning object*. The annotation of the *content objects* depends on the subject of *learning object*. This method fails to annotate the *content objects*, if the subject of the *learning object* is not available. A major limitation of their work of subject identification is that it needs author's supplied information.

We have used machine learning approaches to identify the topic of a learning material. Machine learning method is a dominant approach in example-based classification and is used by many researchers (*Li et al. 2004; Bot et al. 2004; Haruechaiyasak et al. 2002*) for topic identification of web documents. The approach used in our work is example based where a classifier is trained with a set of example documents and then the classifier is used to identify the topic of a document based on the examples.

The annotation tool developed by us for semantic tagging of learning materials with metadata is discussed in Section 5.

5.1 Architecture of the Annotation Tool

The overall architecture of the annotation tool is shown in Figure 3. It consists of the modules *Document collector*, *Metadata extractor module* and the *metadata annotated learning object repository*.

5.1.1 Document Collector

It provides the input interface through which the documents can be inserted. The input interface is very simple. It has two input buttons *Document Submission* and *Web Based Submission*. The *Document Submission* input accepts documents from contributors or Authors. The contributors/authors need not to do any kind of manual entry while uploading. *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.

5.1.2 Metadata Extractor Module

The submitted document is sent to the automatic metadata extractor module of the system. The metadata is automatically extracted from the document.

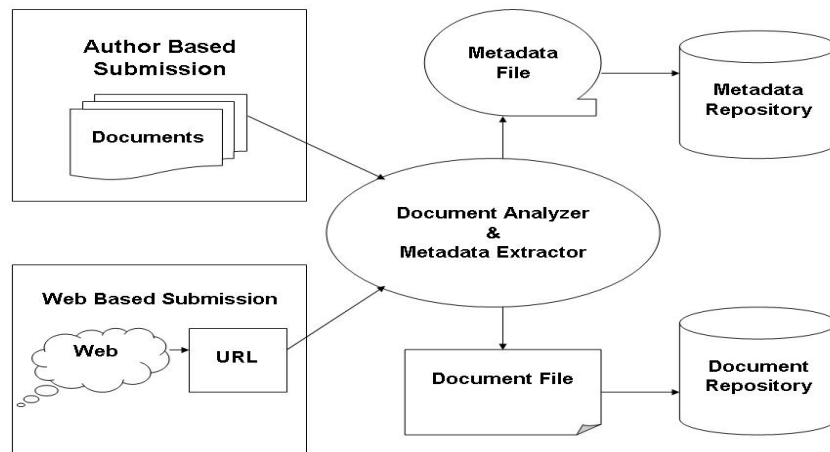


Figure 3 Architecture of the annotation tool

We have used the probabilistic neural network (PNN) to obtain the topic of a document. Probabilistic Neural Networks (PNN) with Gaussian functions has been used to design the modular network structure. The architecture of a typical PNN is shown in Figure 4. The PNN architecture is composed of many interconnected neurons organized in successive layers. The PNN has a 3-layer feed-forward structure.

Pattern layer: When an input is presented, the first layer computes distances from the input vector to the training vectors, and produces a vector whose elements indicate how close the input is to a training vector. This layer assigns one node for each of the training pattern. There are two parameters associated with each pattern node.

$w_i \rightarrow$ the centres with dimension $N \times 1$

$\Sigma_i \rightarrow$ the covariance matrix $N \times N$

where,

$N \rightarrow$ is the dimension of the input vector or the number of features

The output of each of the pattern nodes is given as:

$$v_i = \exp \left\{ -(\mathbf{x} - \mathbf{w}_i)^T \Sigma_i^{-1} (\mathbf{x} - \mathbf{w}_i) \right\}, i = 1, 2, \dots, M$$

$M \rightarrow$ the number of training patterns

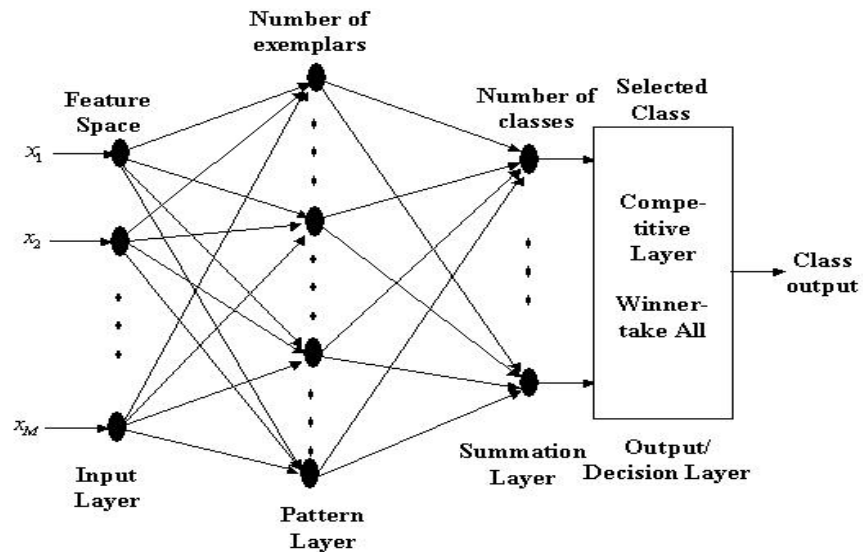


Figure 4. Probabilistic neural network architecture

Summation layer: The second layer sums these contributions for each class of inputs to produce as its net output vector of probabilities. The number of nodes in this layer is the number of classes. Each of these nodes receives an input from each of the pattern nodes through a set of weights. The output of this layer is given as:

$$o_j = \sum_{k=1}^M \mathbf{w}_{jk}^{(s)} v_k, \quad j = 1, 2, \dots, L$$

$L \rightarrow$ the number of classes

$\mathbf{w}_{jk}^{(s)} \rightarrow$ the weight associated with the k^{th} pattern node to j^{th} summation node

$o_j \rightarrow$ is the output of the j^{th} summation node

Decision layer: Finally, a complete transfer function on the output of the second layer picks up the maximum of these probabilities and produces a 1 for that class and a 0 for the other classes.

We have used the MATLAB neural network toolbox to create a probabilistic neural network. The feature set used for the automatic identification of the topic of the document is discussed below.

Feature Set: The documents on a topic contain discussion of several concepts. For example, a document on the topic *spherical mirror* may contain several of the following set of concepts like *Reflection, pole, concave surface, convex surface, focus, normal, focal plane, principal axis, focal length, concave mirror, convex mirror* etc. The concepts are domain specific and therefore unambiguous. The feature set contains the concepts present in the document.

To obtain the concepts present in a document, the text of each of the submitted document is tokenized. Each token is compared with the concept ontology. The concept ontology keeps different concepts of different topics and relationship between them. We have developed concept ontology for three domain Physics, Biology and Geography (*Bhowmick, Roy, Sarkar et. al., 2007*) containing total of 3400 concepts. After comparing with the concept ontology, matched concepts are extracted from the document and added into the list of concepts. The number of occurrences of each concept in the document i.e. the frequency of each concept in the document is found. The frequency is normalized with respect to the number of words present in the document. Specifically, the importance of a concept is proportional to the normalized frequency of the concept in each document. The normalized frequency gives the weight of the concept. The concept-weight vector of the document is used as a feature vector for the PNN classifier. The classifier is trained using feature vectors and tested for identifying the topic of the document. The experimental results are presented and discussed in Section 6.

5.1.3 Metadata annotated learning object repository

The automatically extracted metadata is stored in the metadata file. Metadata is expressed in a semantic web language in RDF format in the metadata file. It is compliant with IEEE LOM RDF binding specification so that it can be used by any learning management system. Each learning object is represented by a RDF-triplet of form $\langle R, P, V \rangle$ where R is the RDF resource of the learning object. The metadata field is represented by a property P of R and V is the value of P.

6. RESULTS AND DISCUSSIONS

To present the performance output of the classifier, we have considered a part of a topic tree of subject Physics shown in Figure 2. The root node of the tree is the subject *physics*. The child nodes of the root of the tree are the chapters of physics such as *kinematics, optics, electricity* etc. Each chapter node has many child nodes, which represent the topics of that chapter in the topic tree. A document can belong to one or more topics that are to be identified using the classifier.

The experiment for topic identification has been carried out to test the performance of the classifier. For experimentation, we have collected documents on different topics namely *lens, mirror, optical image, prism, telescope, refraction*. For each topic, 40 documents are collected (It is very difficult to collect sufficient number of documents of the same topic) and total 240 feature vectors are created. Out of total 240 vectors, 120 vectors are used for training and the rest 120 vectors are used for testing the classifier. The experimental results are given in the Table 4.

Table 4 classification of the documents based on topics

Input		Output (Number of documents)						
Topic name	No. of documents	Lens	Mirror	Optical Image	Telescope	Law of refraction	Prism	Avg. Precision
Lens	20	12	2	4	-	4		55%
Mirror	20	-	10	6	-	-	4	
Optical image	20	5	1	12	-	-	2	
Telescope	20	7		3	10			
Law of refraction	20	-	-	-	-	10	10	
Prism	20	6				4	12	

The average precision obtained is 55% and is very poor. We tried to find out the reason of the poor performance of the classifier. For that we have tested the classifier performances for two different cases with less number of classes.

Case 1: In the first case, only two topics are taken. The topics are belonging to the same parent or the sibling nodes in the topic tree. For example, we are looking at the classification task to classify the documents into the topic *lens* and the *mirror* from the chapter *optics* as shown in Figure 2. Feature distribution in documents of the topic *mirror* and *lens* are shown in Figure 5 and 6. We find that there are many common concepts such as *principal axis*, *focus*, *center of curvature* etc present in the documents belonging to the topics *lens* and *mirror*. These common concepts mislead the classifier and leads to incorrect classification. Precision of the classifier is 82%. Experimental results are given in the Table 5.

Table 5 Classifier output for identification of topics belonging to the same parent

Manual Observation		Classifier output						Precision
Lens	Mirror	Lens			Mirror			
No. of documents	No. of documents	Correct	False Positive	False Negative	Correct	False Positive	False Negative	
25	25	20	4	5	21	5	4	82%

Distribution of features

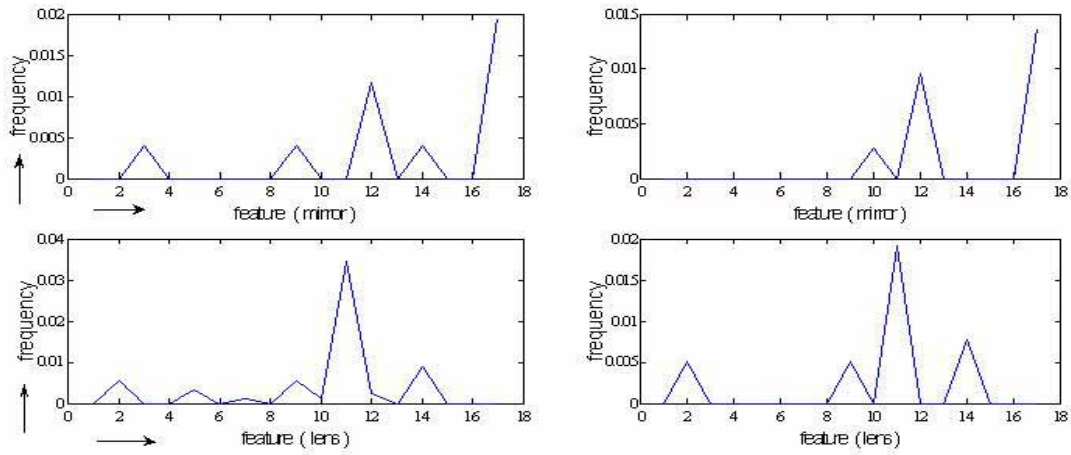


Figure 5 Features distribution in documents belonging to the topics lens and mirror

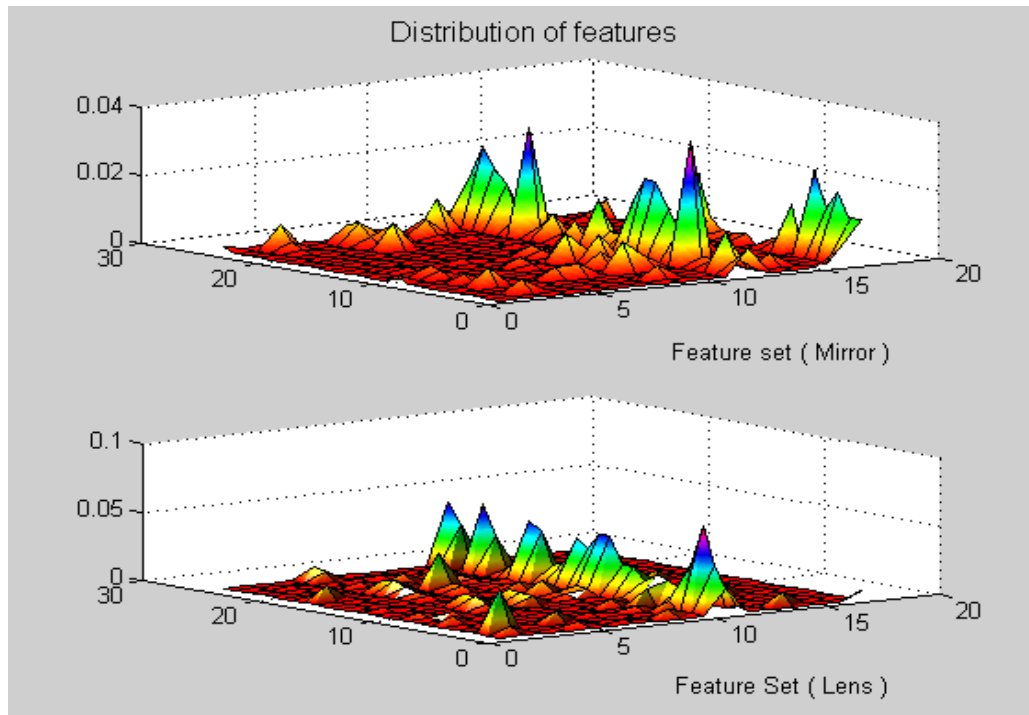


Figure 6 Distribution of features in document sets

Case 2: In the second case, the classifier is used to identify the topic of those documents, which are not the sibling nodes (do not belong to the same parent) in the topic tree. For example, the documents of the topic *lens* (parent node *optics*) and the topic *velocity* (parent node *kinematics*) as shown in Figure 7 are considered for classification. The x-axis represents the feature set and the y-axis represents the frequency of occurrences of the features in a document. The feature distribution in few documents of the topic *lens* and the topic *velocity* are shown in Figure 7. Since the topics are chosen from different chapters, the documents have very few concepts in common.

Distribution of features

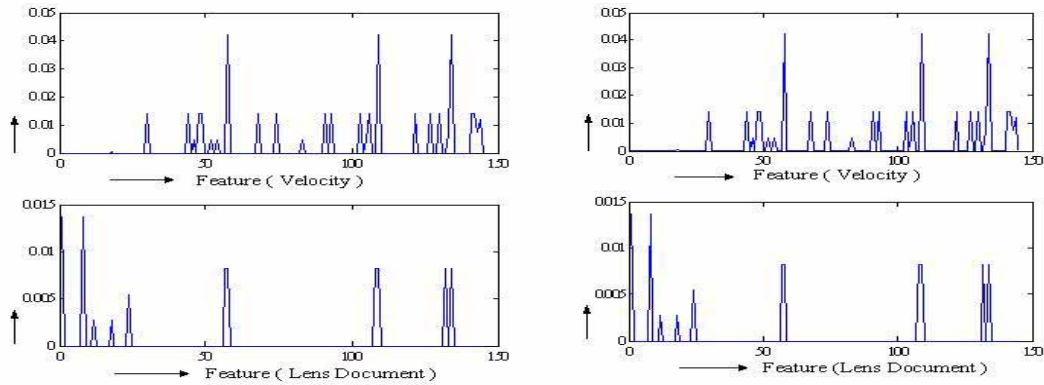


Figure 7 Features distribution in documents belonging to the topics lens and velocity

The experiment is carried out on 100 documents. We have categorized 50 documents into the topic *velocity* and 50 documents into the topic *lens* by manual observation. Out of 100 documents, 50 documents (25 documents from the topic *velocity* and 25 documents from the topic *lens*) are used to train the classifier and the rest 50 documents are used to test the classifier performance. The observation is mentioned in Table 6. The precision of the classifier is 90 %.

Table 6 Classifier output for identification of topics belonging to different parents

Manual Observation		Classifier output						Precision
Velocity	Lens	Velocity			Lens			
No. of documents	No. of documents	Correct	False Positive	False Negative	Correct	False Positive	False Negative	90%
25	25	25	5	0	20	0	5	

From the observations of the above two cases, we find that classifier accuracy is good for the topics which contain different concepts. The classifier accuracy reduces on presence of the same overlapping concepts. For example, the concepts used to explain *optical image* are “*magnification*”, “*image formation by lens*”,

“*image formation by mirror*” etc. So a document explaining “*image formation by lens*” can also belong to the topic *lens*. Similarly a document which deals with “*telescope*” can also belongs to the topic *lens* and *image*. We find that a document can belong to more than one topic. To get a fairly good classifier performance, the classifier should be trained with a large set of documents.

7. CONCLUSION

A comparative study of different available metadata standards and the learning object repositories have been presented in this chapter. It addresses the need of metadata annotation for efficient retrieval of learning materials from learning object repositories. The advantages of annotating the documents with some standard metadata for making them reusable and interoperable between different learning systems have been discussed. The web is large source of good quality learning materials. The ability of annotating the learning materials automatically will enable the learning systems to harness the resources available in the Web and build the learning object repository automatically.

We have explored the feasibility of automatic annotation of learning materials with metadata. This facilitates the creation of an e-Learning repository for storing these annotated learning materials, which can be used by different learning management systems or tutoring systems. The idea is to make use of learning materials from various sources such as from the web, from other repositories or from authors for developing high quality learning materials with specific standard metadata information. The learning materials are reusable and interoperable between different learning management systems.

We have discussed the developed automatic annotation tool for semantic tagging of learning materials with the standard IEEE LOM 9.2.2 *texon* (we have used term *topic*) of a taxonomy belonging to the classification category. Automatic annotation of learning materials is a difficult task. However, the use of machine learning approach is making this difficult task possible. The probabilistic neural network classifier is used to obtain the *topic* of a document. Documents on various topics of school level syllabus are collected from various sources to obtain the feature vector to train the classifier. To get a fairly good classification performance, it is required to train the classifier with a large number of feature vectors. Collection of large number of documents on the same topic is very tedious job, but once the classifier is trained, it can be used for identifying the topic of the document.

References

- Advance Distributed Learning Initiative. <http://www.adlnet.org> (accessed December 30, 2009)
- ARIADNE. <http://www.ariadne-eu.org> (accessed December 30, 2009)
- Bhowmick P.K., Bhowmick S., Roy Devshri, Sarkar S. and Basu A. (2007). *Sahayika: A Framework for Participatory Authoring of Knowledge Structures for Education Domain*. *Proceedings of the 2nd IEEE/ACM International Conference on Information and Communication Technologies and Development, Bangalore, December 15-16, 2007*, pp. 1-11..
- Bot, R. S., Wu Y. B., Chen, X., Li Q. (2004). *A Hybrid Classifier Approach for Web Retrieved Documents Classification*. *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC' 04)*, pp. 326-330.
- CanCore guidelines. <http://www.cancore.ca/en/help/44.html> (accessed December 30, 2009)
- CanCore. <http://www.cancore.ca/> (accessed December 30, 2009)
- Cardinaels, K., Meire, M., Duval, E., (2005). *Automating Metadata Generation: the Simple Indexing Interface*, *Proceedings of the 14th International Conference on World Wide Web Committee (IW3C2), WWW 2005, may 10-14, Chiba, Japan*.
- CAREO. *Campus Alberta Repository of Educational Objects* <http://www.ucalgary.ca/commons/careo/> (accessed December 30, 2009)

DC-dot. <http://www.ukoln.ac.uk/metadata/dcdot/> (accessed December 30, 2009)

DCMI. <http://dublincore.org/> (accessed December 30, 2009)

Dehors, S., Faron-Zucker, C., Stromboni, J., Giboin, A. (2005). *Semi-automated Semantic Annotation of Learning Resources by Identifying Layout Features*. Workshop on Applications of Semantic Web Technologies for E-learning, July 18, Amsterdam, The Netherlands.

Downes, S. (2004). *Resource Profiles*. *Journal of Interactive Media in Education, Special Issue on the Educational Semantic Web*, vol. 5.

Duval, E., Hodgins, W. (2004). *Metadata matters*. *Proceedings of International Conference on Metadata and Dublin Core Specification, DC-2004*, Shanghai, China.

EdNA. The Education Network Australia, <http://www.edna.edu.au/> (accessed December 30, 2009)

Gelbukh, A. F., Sidorov, G., Guzman-Arenas, A. (1999). *Document Comparison with a Weighted Topic Hierarchy*, 10th International Workshop on Database & Expert Systems Applications (*Proceedings IEEE Computer Society*), Florence, Italy, 1-3 September, pp. 566-570.

Han, H. C., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.A. (2003). *Automatic document metadata extraction using support vector machines*. *Proceedings of the third ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 37 - 48.

Haruechaiyasak C., Shyu M., Chen S., Li X. (2002). *Web Document Classification Based on Fuzzy Association*, *Proceedings of the 26th annual international computer software and applications conference (COMPSAC'02)*, pp. 487-492

Health education assets library. HEAL, <http://www.healcentral.org> (accessed December 30, 2009)

Huynh, D., Mazzocchi, S., Karger, D. (2005). *PiggyBank: Experience the Semantic Web inside your Web browser*. *Proceedings of the 4th International Semantic Web Conference*, Galway, Ireland, pp. 413-430.

IEEE LOM. <http://ltsc.ieee.org/wg12/index.html> (accessed December 30, 2009)

iLumina. <http://www.ilumina-dlib.org> (accessed December 30, 2009)

IMS: Standard for Learning objects. <http://www.imsglobal.org/> (accessed December 30, 2009)

Jenkins, C., Inman, D. (2000). *Server-Side Automatic Metadata Generation using Qualified Dublin Core and RDF*. *Kyoto, International conference on digital libraries: research and practice*, pp. 262 – 269.

Jovanovic, J., Gasevic, D., Devedzic, V. (2006b). *Automating Semantic Annotation to Enable Learning Content Adaptation*. *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2006)*, June 21-23, Dublin, Ireland, pp. 151-160.

LC Action Plan. <http://lcweb.loc.gov/catdir/bibcontrol/actionplan.pdf> (accessed December 30, 2009)

LearnAlberta Online Curriculum Repository. <http://www.learnalberta.ca/> (accessed December 30, 2009)

Li, Y., Zhu, Q., Cao, Y. (2004). *Automatic metadata generation based on Neural Network*. *Proceedings of the 13th International Conference on Information Security, ACM International Conference Proceeding Series*, vol. 185, pp.192 –197.

LydiaLearn. <http://www.lydialearn.com/> (accessed December 30, 2009)

MARC. Machine-readable cataloging, <http://www.loc.gov/marc/marcdocz.html> (accessed December 30, 2009)

MERLOT, *Multimedia Educational Resources for Learning and Online Teaching*. <http://www.merlot.org> (accessed December 30, 2009)

Ochoa, X., Cardinaels, K., Meire, M., & Duval, E. (2005). *Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories*. *World Conference on*

Educational Multimedia, Hypermedia & Telecommunications, EDMEDIA 2005, Montreal, Canada, pp. 1407-1414.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M. (2003). *KIM- Semantic annotation platform. Proceedings of the 2nd International Semantic Web Conference, Florida, pp. 834-849.*

PROTON ontology. <http://proton.semanticweb.org/> (accessed December 30, 2009)

RDF, Resource Description Framework. <http://www.w3.org/RDF/> (accessed December 30, 2009)

Roy Devshri, Sarkar Sudeshna, Ghose Sujoy. (2008). "Automatic Extraction of Pedagogic Metadata from Learning Content" *International Journal of Artificial Intelligence in Education, Vol.18, No. 2. pp 97-118.*

Simon, B., Dolog, P., Miklos, Z., Olmedilla, D., Michael, S. (2004). *Conceptualizing Smart Spaces for Learning. Journal of Interactive Media in Education, Special Issue on the Educational Semantic Web, Vol. 3(03), pp. 22-26.*

SMETE, The National Science, Mathematics, Engineering, and Technology Education Digital Library, <http://www.smete.org> (accessed December 30, 2009)

XML, Extensible Markup Language. <http://www.w3.org/XML/> (accessed December 30, 2009)