Automatic Metadata Tagging of Learning Resource Types

Tripti Malviya[1] Prof. Devshri Roy [2]
[1] M. E. Student, Information Technology
[2] Associate Professor, Information Technology
[1,2] Maulana Azad National Institute of Technology, Bhopal (M.P.) , India

Abstract – Metadata is data about data. IEEE learning object metadata is a standard metadata. For relevant retrieval ofa learning material, it is tagged with metadata and kept in the learning object repository. Generally metadata tagging is done manually. Manual metadata tagging is a time consuming and tedious job. To avoid this drawback, we have worked on automatic metadata tagging. We have presented an automatic way to determine a set of metadata from the IEEE LOM 5.2 i.e. the type of learning resource type.In this work, we have mainly determined the *narrative text*, *experiment type* and *experiment type* learning resources. We have developed different pattern bases for different type of learning resources. Pattern matching algorithms with various rules have been developed for extracting the type of learning resource type. Experimental results are shown to depict the accuracy achieved.

Keywords: Learning Object Metadata, Learning Object Type, text classification

I. INTRODUCTION

We live in an age of e-learning where a student can learn from online information. The online information is growing in the form of document repositories and databases. The growth is evident to rise in proportion of the usage. And retrieving what one need is more like finding needle in a haystack! Determination of the relevance of information is one of the major obstacles today. One of the ways to solve the above problem is to tag the documents with different metadata information and keep them in the document repository. Many learning object repositories like MERLOT (http://www.merlot.org/merlot/index.htm) iLumina (http://www.ilumina-dlib.org) etc. are available where documents are tagged with standard IEEE learning object metadata which helps in retrieving the relevant documents. In most of the learning object repositories the metadata tagging of documents is done manually. The quality of metadata tagging depends totally on the manual metadata tagger. There is need to automate the process of metadata tagging. Referring to the need, we have developed an algorithm which automatically extract metadata from documents and automate the process of metadata tagging. In this work, we have extracted a set of metadata from the IEEE 5.2 learning resource type[1]. We have mainly concentrated on extracting *narrative text type, experiment type* and *exercise type* documents.

II. IEEE METADATA 5.2 LEARNING RESOURCE TYPE

The IEEE LOM standard specification (http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf) specifies a standard for learning object metadata [1]. It specifies a conceptual data schema that defines the structure of a metadata instance for a learning object [6]. The IEEE LOM specification consists of nine categories. The fifth category IEEE 5.0 is the educational type. In that 5.2 is the learning resource type that is shown in detail in Table 1.

IEEE LOM 5.2	Learning Resource Type	Specific kind of learning object. The most dominant type shall be the first. – exercise – simulation – questionnaire – diagram – figure – graph – index – slide – table – narrative text – exam – experiment – problem statement – self assessment – lecture

Table. 1 Types of Learning Resources

The IEEE 5.2 learning resource type metadata is the most important and useful metadata for e-learners [11].Different educational psychologists have proposed different learning theories to reach the educational objective of a student. For example in 1968, Ausubelhas proposed a learning sequence consisting of four learning phases. The learning phases are *advance organizer*, *progressive differentiation*, *practice*, and *integrating*. Bloom identified six phases namely, *knowledge, comprehension, application, analysis, synthesis* and *evaluation*.

Generally any course curriculum is designed considering different learning theories so that a student can achieve his educational objectives [10]. A student first gain general knowledge about a topic.

To learn the topic in depth, he learns the narrative text or the explanatory type documents of that topic. To gain practical knowledge, he performs experiments on the 2 same topic. To evaluate himself or to check that he has achieved his educational objective, he solve exercises on that topic.

To achieve the educational objective, an e-learner searches these documents from the World Wide Webwhich is a very time consuming job. No search engine provides a particular type of document on a given topic directly in its search result. There is need that documents are categorized according to different resource type and tagged with this information. The tagged documents are kept in the document repository, and then searching of these different types of documents will be very easy.

We have worked on automatic categorizing of documents according to IEEE 5.2 learning resource type [12] [13] [14]. At present work, we categorize documents into three categories

A. *Narrative text or explanatory type*

B. *Exercise type*

C. *Experiment type.*

This paper is divided in the following sections: Section 3 describes the proposed tool for identification of learning resource type. In Section 4performance evaluation of the proposed tool is given. Finally, section 5 concludes our work.

III. TOOL FOR IDENTIFICATION OF LEARNING RESOURCE TYPE

A. *Pattern Base*

Most of the commercially available systems rely upon Boolean logic and exact matching [3]. In such environments, documents are represented by a list of significant terms without taking into account the different contribution of each term to the document characterization. A term either applies or does not apply to a document. The same is true for the query which consists of single search terms, denoting concepts, combined together via Boolean operators [5]. In response, documents are retrieved when the attached index terms, regarded as secondary keys, match the query specification *exactly*. Thus the document collection is simply partitioned into two sets: documents that satisfy the query and documents that do not. In the proposed work, the importance of each matched pattern is also considered. Different weights are given to different patterns. The position of the pattern in the document is also considered. The following section discusses about the pattern base identified for different type of documents [4].

1) *Pattern base for narrative text type documents:*

Explain, interpret, outline, discuss, distinguish, predict, restate, translate, complete, examine, clarify, affirms, conform, use, conclusion, one of, most, basically, generally, Can be classified as, called as, consists of, deal with, defined as, depicted as,, described as, explained as etc.

2) *Pattern base for experiment type documents:*

Experimental,laboratory,manual,observation,estimate,lab,practical,experiment,prelab,objective,method,references,error,conclusion,procedure,technique,principle,requirement,illustrate,goal,law,theorem,configure,establish,study,protocols,troubleshoot,design,following,application,throughput,check,theory,description,instruction,hypothesis,report,title,abstract,sample,apparatus,parameter,device,equipment,install,program,discussion,tools,command,default,implement,simulation,analysis,activity,function,execution,peripherals,code,output,input,setting,print,operation,information,tool,save,develop,user,wap,schedule,guidelines,product,relationship,frequency,constant,measure,peak,criteria,show,determine,draw,purpose,test,safety,precaution,diagram,note,adjust,reading,quantity,hints,graph etc.

3) *Pattern base for exercise type documents:*

Who, what, where, when, why, how, which, whom, whose, how many, how much, how come, show that, what kind, to whom, Do you etc.

There are similar such pattern found in different learning resources which are collected by carefully analyzing the data.

B. *Workflow*

The tool developed for identification of learning resource type is shown in Figure 1.
The work flow is as follows:

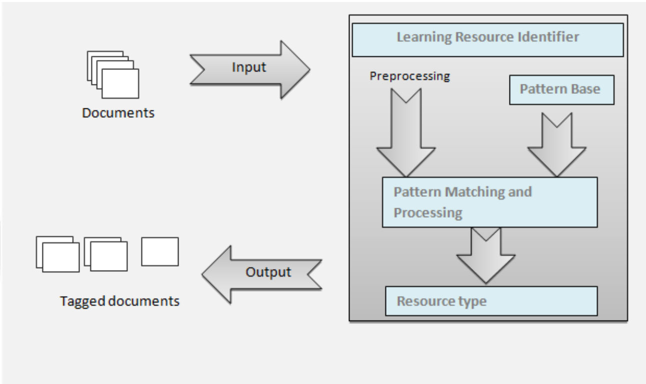

Fig. 1 Tool for identification of learning resource types

- Each document is processed line by line, finding the possible matches with each of the pattern base.
- Depending on the significance of the pattern in a sentence, the weight of each sentence is incremented to contribute to its final type of the document Different rules are need to be followed to assign the weight.
- A cumulative sum of the weights is taken for each category for a document.
- The frequency of the most frequently occurring words which are from the pattern base, in comparison to the total number of terms are checked
- Depending on the frequency count the document is tagged as narrative text, exercise, experimental.

If simply patterns are matched in sentences, the classification accuracy is low. For example the question words like who, which also occur in narrative text and experiment type text. So it is needed to make some rules for assigning weights to the sentences on pattern matching. Some of the rules are discussed below.

Rule 1: If a document contains words like '*Experiment*' ,'*lab*', '*report*' are given higher weight for a Experiment type of document. And then, it is further checked for some more important terms to be found in the experimental texts.

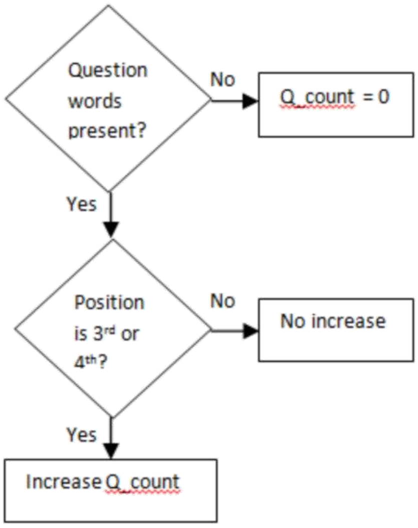

Fig.2 Rule 1

Rule 2: Simply finding question words ('*which*', '*why'* etc.), which normally also occur in descriptive or experimental texts, we have increased the weight if the question word occurs in the sentence within a range of first four to five words. This is done to cover the variety of the formatting (Ex. Ques. 1. (a) what is a process?) and also as per the rules of grammar of the language.

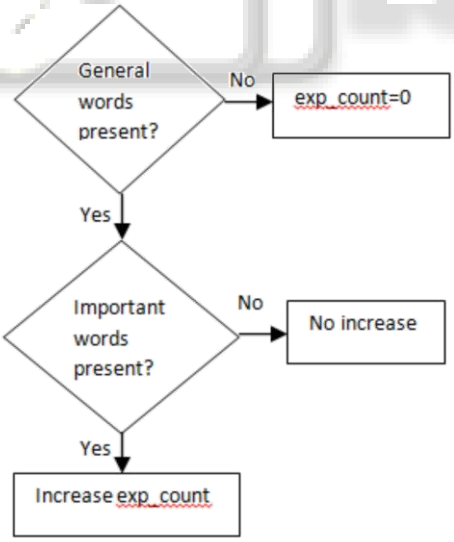

Fig. 3 Rule 2

These rules have been applied in order to make use of the structural pattern used in English language to get better results. Similar formulation of perception process of human mind may add complexity to the model and may lead to better results, i.e. modeling of how to we perceive or differentiate an exercise type document from an experiment or a narrative type.

C. *Algorithm*

```
Begin:
For each document in the directory
    For each sentence in the document
        if (question words present)
            if (question word present at position three or four from beginning)
                Increase q_count;
        if (descriptive words present)
                Increase d_count;
        if (exp_general words present)
            if(exp_imp words present)
                Increase e_count;
    q=q+q_count;
    d=d+d_count;
    e=e+ e_count;
    End // inner for
The biggest percentage match count is assigned as type to the document
End// outer for
End
```

Fig. 4 Algorithm

IV. PERFORMANCE

The performance of the developed tool is checked by using calculating recall and precision. These measures tell about the quality of the retrieved document.*Precision* is the total number of relevant documents retrieved over the total number of documents retrieved. Recall is the number of relevant document retrieved over the total number of relevant documents in the collection. The mathematical formulation of these measures is as follows:

$$\text{Precision} = \frac{t_p}{t_p + f_p} \quad (1)$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \quad (2)$$

Where,

tp = true positive, number of assessment where system and human expert agree for a label assignment.

fp= false positive, number of assessments where labels assigned by the system does not agree with expert assignment.

fn= false negative, number of assessments where system failed to assign as they were given by the expert.

We took a dataset of 150 documents. It contains 57 narrative text types, 34 Exercise type and 59 Experiment type of learning resource type documents. The developed algorithm is tested on this data set. The final results were checked with the manually classified documents.

Type of learning document	**No. of document**	**Precision**	**Recall**
Narrative text	57	68.96	35.08
Exercise	34	47.45	82.35
Experiment	59	55.84	72.88

Table. 2 Result

V. CONCLUSION

Metadata tagging is one of the most important areas of research. The advancements in this area can be attributed to a number of factors such as:

It can be used to improve the accuracy of information retrieval process as in searching[7].

As with the proliferating information database size, it necessitates a quick and reliable way to handle it. Manual work may lead time and may not be efficient too in the case of large data size.

It also increases the machine processabilty of the information base as it enable advanced applications to find learning objects without human intervention.

We have classified three types of learning resource document in relevance to the pattern occurring in them .In order to improve the accuracy different rules are applied with pattern matching. This kind of metadata is very useful for managing large repositories of online document resources and useful for the students. There exists a variety of such resource types which may also be considered for the metadata tagging as a future extension to this work.

REFERENCES

[1] IEEE LTSC, IEEE Learning Technology Standards Committee , http://www.ieeeltsc.org:8080/Plone
[2] http://dublincore.org/specifications/
[3] Sonal Jain, Jyoti Pareek , Automatic Identification of Learning Object Type ,in the International Journal of Computer Applications (0975 – 8887) Volume 42– No.7, March 2012
[4] Devshri Roy, Sudeshna Sarkar, Sujoy Ghose, Automatic Extraction of Pedagogic Metadata from Learning Content Konstantin Nosov. Automatic Generation of Metadata for Learning Objects. Master's Thesis Project Description, Master of Science in Media Technology 30 ECTS Department of Computer Science and Media Technology Gjøvik University College, 2011
[5] Jehad Najjar, Stefaan Ternier, Erik Duval. Learning Object Metadata: Opportunities and Challenges for the Middle East and North Africa
[6] Dragan Gaševiand Marek Hatala. Ontology mappings to improve learning resource search
[7] http://www.wsdot.wa.gov/NR/rdonlyres/70DDD702-C97D-4C12-B850-D2A59E6576/0/CVII.pdf accessed on 9 december,2012
[8] Devshri Roy, Sudeshna Sarkar, Sujoy Ghose.A Comparative Study of Learning Object metadata, Learning Material Repositories, Metadata
[9] Annotation & an Automatic Metadata Annotation Tool. Advances in Semantic Computing(Eds. Joshi, Boley & Akerkar), Vol. 2, pp 103 – 126, 2010 Chapter 6
[10] http://edutechwiki.unige.ch/en/Learning_Object_Metadata_Standard accessed on 21st April 2013.
[11] Carsten Ullrich, German Research Center for Artificial Intelligence. The Learning-Resource-Type is Dead, Long Live the Learning-Resource-Type! *Learning Objects and Learning Designs* 1(1), 7-15.
[12] IEEE P1484.12.3, Draft 8 Draft Standard for Learning Technology— Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata, Draft P1484.12.3/D8dated 16 February 2005.Sponsored by Learning Technology Standards Committee of the IEEE Computer Society
[13] 1484.12.1IEEE Standard for Learning Object Metadata, IEEE Computer Society. Sponsored by the Learning Technology Standards Committee