CHAPTER 3

PREPROCESSING AND ANNOTATING OF LEARNING OBJECTS WITH NEW MATCHING SCORE BASED ON THE PRESENCE OF CONCEPTS

3.1 INTRODUCTION

This chapter provides a description of finding the percentage level of the concepts present in the learning objects and the role of WordNet in annotating the documents. The implementation of semantic web demands distributed availability of semantic annotations for all the documents posted on the web. Authoring tools are used to manage the manual annotation and it also supplies an integrated environment for concurrently authoring and annotating the text. Manual annotation needs a great deal of human intervention to create an effective semantic metadata and hence, it is believed to be a time-consuming process. The need of representing different views of a data source for multiple users is also to be considered while annotating the documents and it can be beneficial to support the demands of different kind of users. For example, annotations can help the vision-impaired user by providing faster navigation through a website (Yesilada et al. 2003) and normal users can use annotations of the same document to provide a descriptive view of a domain. Thus, different kind of ontologies needs to be created in the same document.

Automated annotation provides support for scalability and reduces the complexity involved in annotating new documents. The semantic annotation platforms (SAPs) can be classified based on the type of annotation
technique used. The pattern based annotation technique (Brin et al. 1998) and machine learning based annotation technique are two primary divisions of SAPs. Sometimes, both the techniques work together to improve the strengths and to compensate the weakness of the annotation techniques. This process is called as a multi-strategy annotation. Here the pattern based SAPs can execute the pattern discovery with the preliminary set of entities and the corpus is scanned to discover the patterns in which the entities exist. This pattern discovery is typically deployed in identifying the fresh entities along with the new patterns. Likewise, more entities are identified by repeatedly continuing the process. The ontology also plays a critical role in annotating the document and to provide a semantic description.

WordNet is one of the famous examples of ontology widely used for experimental evaluations. Although not explicitly designed as ontology, WordNet largely fits into the ontology definitions. The WordNet database organizes simple words and multi-word expressions of different syntactic categories into the synonym sets (Synsets). Synsets represent an underlying concept and link these through semantic relations (Scott et al. 1998). Its hierarchical structure is not necessarily a tree structure. It may also be a directed acyclic graph, possibly linking concepts to multiple super concepts at the same time.

The retrieval of educational resources on the web and extracting the metadata relevant to the document is a challenging task. Normally, educational resources were classified and annotated using a set of properties such as content categories, course title, etc. But still their metadata failed to reflect the content of the textual segments. To sum up in the context of learning information retrieval, there exist systems in favor of manual or semi-automatic annotation. Due to the time consumption and frequent need for
human intervention, the manual annotation will not be an interesting task for users. When the annotation is automatic, metadata relative to the whole learning document is extracted and they are automatically annotated with semantic metadata relative to their learning categories.

Concept-based information retrieval is an important IR approach that purports to tackle these problems differently. Instead of representing only keywords, concept-based IR represents both documents and queries using semantic concepts and performs the retrieval process. This approach has the promise that representing documents and interrogations using high-level concepts will result in a recovery model that is less dependent on the specific terms used. Such a model could yield matches even when the same notion is identified by different terms in the query and target documents, therefore improving the synonymy problem. Similarly, if the concepts are chosen for ambiguous words from both in the query and in the documents, non-relevant documents that were retrieved with the bag of words approach could be eliminated from the results. This leads in improving the polysemy problem and increasing precision.

In a keyword-based approach, the user searches for the desired information by giving a query. The query is represented as a set of words, which ranges from a single word to a collection of words. The input query is parsed into an unstructured set of keywords and it is matched against an inverted index that links keywords with the documents in which they appear. Hence, the documents that contain the occurrence of keywords that match the input query are retrieved and ranking algorithms are employed to place the pages based on the relevant content of the user. The problem with this approach is that no attempt is made to identify the meaning of the query and to compare it with the concept of the documents.
When a learner searches for learning materials by giving a keyword based on the domain interest in the search engine, the search query returns many irrelevant documents. The major drawback of this approach is that same keyword may belong to many different topics and have a different meaning, but it cannot select the documents, which contain similar meaning. If the topic of a learning material is identified and annotated with the learning content, it aids in the retrieval of LOs according to the learner’s interested topic. Therefore, there is a clear need to develop new annotated systems that can be taken into consideration.

Concept based annotation aims at representing the content of the text document and learning object by means of the metadata. The important rule is to identify the concept of the learning objects either by creating domain ontologies or by using other linguistic resources such as WordNet. For this aim, representative keywords are first identified in each document, using classical indexing techniques. Then the content of the learning object is checked for the presence of concepts, identified, weighted according to their concept frequency distribution and relatedness to other concepts in the learning object. The entire process of the proposed work can be summed up in two pieces:

- Identification of concept supported by LOs.
- Concept weighting in LOs.

3.2 ANNOTATIONS AND METADATA

A large number of documents available on the web may become a cause for the information overload problems that will have an impact on the users’ sensible decision making to identify the appropriate documents. Document annotation provides a solution to solve such type of problems and
it is the process of providing an additional description to the text documents. In general, annotations are the metadata, which provides additional information about an existing piece of data. Automatic metadata generation depends on machine processing and the advantage of automatic metadata generation is that it can discover much more data much more quickly than humans. The process of automatically extracting the content of the resource and to mine it to create the structured metadata called as metadata extraction.

The annotation of learning documents also helps in obtaining structured information from an unstructured collection of text documents. For example, consider a learning object containing information regarding data mining. By semantic annotation, only the most important information about the particular document is made available in a structured format. This could be the important concepts supported by the learning object and it helps in reducing the complexity of document searching and retrieval process. Annotations facilitate the task of finding the relevant learning objects, assist the reader to quickly overview, and understand the document.

3.3 ALGORITHM FOR ANNOTATING THE DOCUMENTS BASED ON NEW MATCHING SCORE

The users generally tend to retrieve the relevant learning objects from the database by entering a keyword. The proposed algorithm is designed to annotate the learning objects, which are apparently relevant to the input given by defining a new matching score. The new matching score illustrating the various concepts of the learning objects will be associated with its metadata and it will be very useful during the retrieval process. Two types of concepts are involved in the algorithm, namely master concepts and derived concepts. Keywords, Index terms, frequent keyword occurrence are taken as
measures to identify the concept of the learning object. The proposed architecture of the annotation model is depicted in Figure 3.1.

Figure 3.1 Annotation Model
The step by step process of the algorithm is listed below:

- Read the document and remove the stop words
- Read the contents of the document to find the frequent keyword and the concepts of the document.
- Identify the index terms from the document
- Apply the master concept in word net to form derived concepts
- Compare the derived concepts and contents from the documents and create the concept matrix based on contents.
- Compare the derived concept, index terms of the documents, and create the concept matrix based on index terms.
- Find the importance measure for the concept matrix
- Apply the matrix reduction technique after finding the importance measure
- Find the resultant matrix after applying the matrix reduction technique
- Identify the presence of concept in a document using the resultant matrix
- Associate the value with the metadata of the document

3.3.1 Pre-Processing

Pre-processing phase transforms the data into a format that can be more effective and easily processed by the algorithms. It helps in removing the irrelevant/redundant data present in the documents, resulting in improving
the quality of the documents. Hence, the original text documents are converted into simple algorithm understandable word format. Initially, text documents are processed as a normal string and then the sequence of string terms is divided into the simple tokenized list of strings. Pre-processing is used to characterize the document process and convert the content of a document into a sequence of terms like words or phrases. Two processes namely stop word elimination and stemming is used to remove the unwanted entities and words from the sequence of words in an efficient manner.

3.3.2 Stop Word Removal and Stemming Process

Generally, the search engines remove the commonly used words or the stop words from the keyword phrase to give the most pertinent result. In addition to that, the input text documents have great quantities of semantically correct as well as basic functional standard stop words.

The most common words in text documents are represented by means of articles, prepositions and pronouns, etc., which will not have any relevance to the depiction of the concept present in the learning objects (Boubekeur et al. 2007) and are called as stop words. They are removed from the documents because those words are not treated as keywords in data mining applications. Since the stop word makes the text documents very heavier to process and less important for analysis, this acts as a first step of the pre-processing stage. Removing stop words helps in reducing the dimensionality of term space.

Stop word removal can be associated with two processes, namely queries and text documents. Thus, it helps in improving the performance of the search process. The stop word removal is done while parsing a document to gain the information about the content and removing the commonly used
words that have less significant meaning than the keywords. List of stop words is stored in the database and compared with the content of the learning objects. If a stop word is found, it is removed and the remaining terms are extracted.

After removing the stopword in text documents, the stemming process is applied to reduce the number of unique terms and words from their stem or root form. This process is also referred as the morphological root of the word. If the text documents contain sentences with no stop words, it will be processed as a single common term. Consider, for the input word “Integration System”, the stemming process is applied by means of stemming algorithms, it converts different words (“integration”, “integrations”, “integrative”, “integrated” and “integrating”) into a similar canonical form (“integrate”). This acts as a standard procedure used to eliminate common morphological, inflectional endings and suffixes. In addition to this, HTML tags and non-alpha characters are removed from the input data. The stemming process is applied to the data or the document given as input by the user and as a result, the morphological terms are removed. The result of this process will be based on the resultant stem that acts as input to important resultant matrix detection.

3.3.3 Identification of Concepts from Learning Objects

In the cognitive aspect of the globe, there exists the presumption that the significance of a text (word) depends on conceptual relationships to objects in the world rather than to linguistic or contextual relations found in texts or dictionaries. A new generation information retrieval model is drawn from this perspective called as concept-based information retrieval model. This section demonstrates the techniques used to identify the concepts supported by the learning objects, to extract higher dependency of a keyword
from the documents and to find the relationship between the keywords. Mutual Information is applied to find the exact matching solution during the retrieval process and the keyword that shows higher dependency in the document amid others is considered as a concept.

Text data usually comprises strings of characters, which are transformed into a representation suitable for learning. In the feature space representation, the sequences of characters of text documents are represented as a sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Learning Objects are also classified under text documents and to identify the concept of the learning objects a two-step process is followed:

- Identify the most frequent word associated with the documents
- Identify the relationship between most frequent words with other words and use mutual information to indicate the concept associated with the document

Two main techniques used for parsing the documents are:

- Window-based modeling
- Corpus-dependency based modeling.

In window-based models, word co-occurrences within a window of a given size determine the text corpus, where the window, simply spans a number of words occurring in instances of a target word. For corpus dependency-based models, word co-occurrences in a particular syntactic relation determine the text corpus with a target word. Window based corpus technique outperforms the corpus-based technique if the window size is selected accurately (Douwe Kiela & Stephen Clark 2014). Hence, windows
based technique is chosen to parse the learning object with the sentence size of three. The extraction of concepts from the learning objects is initiated by finding the frequency occurrence of each word in the document.

The frequent occurrence of a word is calculated as the ratio between the total occurrences of a particular word in the document to the total number of words present in the document. The selection of the keyword to the concept is done by finding the interrelationship amid the keyword and other keywords. The bond amid the two keywords is obtained by finding the probability of occurrence of the keyword. Conditional probability is used to discover the relation between the keywords and the positive point-wise mutual information is used to extract the concept. The keyword that shows higher dependency amid others is considered as a concept. Examination of this technique shows that more dependency would extract the concept more from the text corpora.

The process of Mutual Information consists of the following steps: For two terms, term 1 and term 2, the relation finding terms can be expressed as

$P(Term_1: Term_2) = \frac{P(Term_1 \mid Term_2)}{P(Term_2)}$, for terms 1 and 2 and D represents the documents in the corpus. The calculation of the mutual information as in equation (3.1).

$$Ml(Term_1: Term_2) = \frac{P(Term_1 \mid Term_2)}{P(Term_2)} \text{, term}_1, \text{term}_2 \in D$$

(3.1)
In addition to this, the function $P\left(\frac{\text{Term}_1}{\text{Term}_2}\right)$ i.e. the conditional probability of each word from the document is represented by equation (3.2).

\[
\text{Conditional Probability} \left( \frac{\text{Term}_1}{\text{Term}_2} \right) = \frac{P(\text{term}_2 \cap \text{term}_1)}{P(\text{term}_1)}
\]

(3.2)

$P(\text{term}_2 \cap \text{term}_1)$ represents the probability of occurrence of both the key words. Since the results of the mutual information produced values ranging from positive to negative values, Positive Pointwise Mutual Information (PPMI) is applied to remove the negative values. The PPMI is calculated by equation (3.3).

\[
\text{PPMI}_{\text{term}_1 : \text{term}_2} = \max(\log_2 \frac{P_w}{P_{w|s}P_s} , 0)
\]

(3.3)

where $P_w$ is the probability that the word $w_i$ occurs in the sentence with respect to the entire document, $P_{w|s}$ is the probability of word $w_i$ in the entire documents and $P_s$ is the probability of a sentence in the entire document. From the above equation (3.2), the connectivity between the keywords is found corresponding to the conditional probability. Observations from equations (3.2) and (3.3) consider two extreme cases:

i. If the keyword shows higher dependency amid others, then it is considered as a concept.

$$\text{PPMI}(\text{Term}_1 : \text{Term}_2) = (\text{term}_1) \in D$$

(3.4)
The probability of the above result decides whether the mutual information is high or low for a document. That is, a term with low probability has a larger dependency and term with high probability has a lower dependency.

ii. Higher dependency would extract the more concepts from the input text corpora.

The positive value of PPMI maximizes when term1 and term2 are perfectly associated. Once the mutual information process is completed, the end result of this process constitutes terms showing higher dependency indicating the concept associated with the learning objects. This is used to generate the concept matrix.

3.3.4 Concept Extraction from Learning Objects

Once the concept of the learning object is identified, it is compared with that of the master concept and the derived concept of the learning object. The two concepts involved in the process are:

- Master concept
- Derived concepts or derived master concepts

The master concepts are the concepts which act as keywords and here it is represented by the word identified by means of PPMI. Since the identified master concept may belong to different topics and have a different meaning; the semantics of the master concept is to be essentially analyzed using semantic resources. The derived master concepts are the concepts, which are to be derived from the master concept, by referring WordNet. Derived master concepts contain synonyms of each word present in the master concept and it is combined with each other to form multiple concepts.
WordNet is used both as a lexical database and a semantic resource to form the derived concept. Thereafter, the master concept and the derived master concept are to be checked with the contents of each learning object. The objective of the semantic matching of the master concept is to extend the set of syntactic matches with semantically related concepts (Uren & et al. 2006). The proposed work takes the master concept as an input concept and traverses the structure of WordNet to discover semantically related concepts.

For instance, if the master concept is given as “Data Retrieval” it is applied separately in the WordNet, i.e. “Data” as a single word and “Retrieval” as a single word. The synonym of the master concept is fetched from the WordNet separately for each word respectively. Once the synonyms of both the words are derived from the WordNet it is combined with the synonym of one word with the synonym of another resulting in the derived master concept. In the derived master concepts, the keyword information retrieval is repeated because the master concept has to be included with the derived concepts.

3.3.5 Concept matrix Generation

Once the derived concepts are identified by means of master concepts, concept matrix is created for all the master concepts identified for the learning object. Hence, the presence of the concepts of a learning object can be identified; both by means of contents and index terms. Two types of concept matrix are created:

- Based on the presence of contents
- Based on the presence of index terms
Concept matrix is created separately for both the processes and merged later to improvise the accuracy of the annotation associated with the learning objects.

### 3.3.5.1 Concept matrix generation based on contents

This section introduces the generation of concept matrix based on contents, using the basic concepts of master concept and derived concept. In general, a concept matrix based on contents can be expressed as a Master Concept ($C_1$) [derived master concepts (M) matrix × documents (D) matrix] i.e., $C_1 [M \times D]$. Each matrix row expresses the documents that can be defined for several learning objects. For a sample of ‘n’ documents ($D_1, D_2, \ldots, D_n$), the derived concept is given by ($CB_1, CB_2, \ldots \ldots, CB_n$). After discovering the derived concept from the master concepts, the creation of concept matrix is pursued separately for all the master concepts which are identified for the contents of the documents. The method for creating the concept matrix based on the contents of the documents is given in equation (3.5).

$$C_m D_n C = \frac{R}{TC}$$

(3.5)

Where ‘m’ denotes the number of Concepts, ‘n’ denotes the number of documents, ‘R’ denotes the repeated concept and TC represents the total number of words in a particular document. Based on the design, concept matrix is expressed as the ratio of derived concepts to the total number of concepts in the document. The matrix is represented by having derived concept in rows and the number of documents in columns. It contains the values of each derived concepts and their corresponding documents.
3.3.5.2 Concept matrix generation based on Index terms

Each learning object is described by a set of representative keywords called as index terms. An index term is just a document word whose semantics help in recalling the master theme of the documents. The characteristics of index terms are:

i. Index terms are used to index and summarize the text file contents.

ii. Index terms are mainly nouns because nouns have meant by themselves and hence, their semantics are easier to identify and to comprehend.

iii. Adjectives, adverbs and connectives are less useful as index terms because they work mainly as complements.

iv. Index terms can comprise a word, phrase, or an alphanumerical term.

v. Created by analyzing the document either manually with subject indexing or automatically with automatic indexing by means of algorithmic processes in a faster fashion

Many journals and databases provide access to index terms made by authors of the articles being published or represented. People looking for research papers on their domain knowledge will look for terms relating to the corresponding topic of their interest. Index terms otherwise known as keywords or descriptions or tags will help people look up the subject in a worldwide index of articles. Authors of articles are usually asked for such terms while publishing the article. If appropriate index terms or keywords are applied by the authors, it makes the retrieval process much easier as it will be more relevant to the user demand. The relative quality of indexer-provided
index terms and author provided index terms are of interest to research in data retrieval. The quality of both kinds of indexing terms depends, of course, on the qualifications of the provider. In general authors have difficulties providing indexing terms that characterize the document relative to other documents in the database. Learning objects also can be represented by index terms and the keywords given by the author forms an integral part of literature.

The journal transactions that comprise associated keywords are classified as index terms which play a major role in identifying the relevant concept of the learning objects. The index terms are calculated based on the equation (3.6).

\[
C_m D_n I = \frac{RW}{NWC}
\]

(3.6)

where RW denotes the occurrence of the index terms in the learning objects and NWC denotes total number of words in the learning object. This formula explains that it is the ratio of the related words in the index terms to the number of words present in the concept. Similar to the concept matrix of contents, here the concept matrix contains the values of each derived concept of a respective learning object related to the index terms. In the same way, the concept matrix is calculated for all the master concepts that are relevant to the index terms of the learning objects.

3.3.6 Matrix Reduction

Importance measure is interpreted as the ratio of the concept matrix value of a derived concept in a particular document to the sum of the
values of documents for the same derived concept. In order to minimize the matrix as simple with the help of a number of derived concepts present in the learning objects, importance measure is calculated. In this process, \( CB_j \) represents the number of derived concepts and intuitively \( D_n \) is the number of documents which is used to calculate the result of importance measure with the minimal loss of information and that is denoted in equation (3.7).

\[
\text{Importance Measure} = \frac{CB_j(D_n)}{\sum_{i=1}^{n} D_i}
\]

(3.7)

where ‘j’ varies from the first concept to till the end of the concept count. Matrix reduction depends on the importance measure of the document associated with two things, namely the index terms\((W_i)\) and concepts\((W_c)\). After finding the importance measure for all the derived concepts with its respective documents, matrix obtained needs to be reduced into a row vector to show the similarity between every concept and every document in the test set by the following formulae:

\[
\text{CMR} = C_1 + \frac{1}{2} \sum_{i=1}^{n} \text{CB}_i
\]

(3.8)

The above stated equation (3.8) covers each document separately. If the first document is considered, then the \( C_1 \) value of the first document is to be taken and the \( \text{CB}_i \) denotes the derived concept values of the first document. Hence, the concept matrix is reduced based on the contents and index terms of the documents after applying the importance measure.
3.3.7 Final Resultant Matrix

The final resultant matrix is formed by combining the reduced concept matrix based on the contents of the documents and the reduced concept matrix based on the index terms of the documents. The final resultant matrix is identified through formula (3.9) stated subsequently.

\[
\text{FRM} = \frac{W_c \times \text{CMR}_c + W_i \times \text{CMR}_i}{W_c + W_i}
\]

(3.9)

Here, FRM is the final resultant matrix, \(W_c\) is the content based weight value, \(\text{CMR}_c\) is the content based matrix reduction value, \(W_i\) is the index term based weight value and \(\text{CMR}_i\) is the index term based matrix reduction value. The final resultant matrix contains the value of the master concept with respect to various documents and based on the values present in the resultant matrix a document can be associated with its concept by means of its metadata.

Finally, all the learning objects that are given as input will be associated with metadata that contains values relevant to the content of it. The search engine would display the documents, which have the value of given concept or keyword as high. If the learning objects are classified based on the values of the concepts present in it, i.e. the document has a certain value of the first concept, second concept, etc., it would be easier for retrieval.
3.4 RESULTS AND DISCUSSION

3.4.1 Dataset Description- Introduction IEEE and ACM Database

IEEEExplore is the most widely examined transactional database used for research purposes and document retrieval. It comprises an enormous collection of publications belonging to various predefined categories and subcategories. The Association for Computing Machinery Digital Library (ACMDL) database collection has become a popular, comprehensive collection of full-text articles and bibliographic records for performing research in text applications such as document retrieval and clustering.

Research scholars across the world will be relying on these articles to explore the information and these documents are associated with the metadata. The meta-data available for each document includes various information domain area, publisher information, ISBN information, accession number and so on. The published part of each document consists of an abstract, author, references, citation, keywords, metrics and similar articles section. Hence, the published article or document is classified as learning objects. Annotating the learning objects of IEEE transaction set and the ACM transaction dataset will help the researchers to identify their relevant information. 500 documents belonging to various categories of the IEEE and ACM transactions are chosen for evaluating the performance of the annotation algorithm. The annotation methodology is implemented in Java using the NetBeans IDE. The learning objects of the dataset are taken from the following concept categories.

i. Data Mining

ii. Convergence Stability

iii. Error Analysis
Figure 3.2 Concept identification values in the learning objects

Figure 3.2 illustrates the experimental results of the proposed annotation technique. It shows the presence of the master concept “Error Analysis” in the learning objects taken for evaluation and it has more concept value when compared with other master concepts in the first learning object. In the second and the third learning objects, also the master concept “Error Analysis” has more values compared to the other master concepts, which are chosen for the demonstration in the dataset. The tree structure of an XML document for the metadata is shown in Figure 3.3
3.4.2 Conclusion

The proposed technique is useful to find the level of presence of the learning object concepts that will help the readers/learners associated with the Web. If documents or learning objects need to be found by means of search engine based on a keyword or concept, it is essential to display the relevant materials, which are more related to that concept given as input. The search engine would display the learning objects that comprise the value of a given concept or keyword constituted with a higher priority. Once the learning objects are annotated based on the presence of concepts, it can be retrieved to the user based upon the annotated concept values present in the metadata. It helps in effective retrieval process and in satisfying the learners’ need to a larger extent.