Extracting IEEE LOM 9.4 using Memetic Algorithm for Textual Learning Objects

Siddhartha Kumar Arjaria
Deptt. of Comp. Sc. & Engg.
MANIT Bhopal (M.P)

Deepak Singh Tomar
Deptt. of Comp. Sc. & Engg.
MANIT Bhopal (M.P)

Devshri Roy
Deptt. of Comp. Sc. & Engg.
MANIT Bhopal (M.P)

ABSTRACT

With the growing number of learning objects & their increasing use for learning, it is required to search the desired learning material in the least time. So it is needed to have the quality learning object. The quality of learning objects means they are tagged with correct & complete metadata. IEEE LOM set up a standard for uniformity of Meta data values. The IEEE LOM has 9 different categories for metadata. These categories are used to describe the learning object when filled with metadata values and in turn they are useful searching and understanding the learning object without actually opens it. IEEE LOM 9 category belongs to the classification and under this category, a important subcategory of IEEE LOM 9.4 lies which actually belongs to keywords of learning objects. This paper uses the memetic algorithm based approach to extract the keywords for each learning objects of different classes. The correct keywords strongly reflect the learning object class and in turn it is very useful for user's point of view to decide whether this is the learning object he is searching for.

General Terms

Computational Intelligence, Memetic Algorithm, E-learning

Keywords

Learning Objects, E-Learning, Memetic Algorithm, IEEE LOM.

1. INTRODUCTION

Due to wide availability of learning content in the Internet, web based learning becomes an important subject of research. Learning objects are defined as any entity digital or non digital which can be used reused or referenced during technology supported learning including computer based training system [1]. In other words the learning object should be quality learning object means sufficiently correct & complete Meta data is tagged with it so that it can be reused and searched appropriately. A common metadata standard is necessary, as in the absence of standard the different learning objects has different set of metadata information leading. So following a standard leading towards the uniformity in learning object tagged metadata.

Several metadata standards are used for description of learning objects like Dublin Core metadata initiative (DCMI, http://www.dublincore.org/),SCORM Metadata (http://www.adlnet.gov/scorm), Advance Distributed Learning Initiative (http://www.adlnet.org), IEEE Learning Object Metadata (http://ltsc.ieee.org) etc.

Many learning object repositories (LOR) are available which stores learning materials which helps in delivering good quality learning materials relevant to the student’s requirement. Ariadne [http://www.ariadne-eu.org/], EdNA [http://www.edna.edu.au/edna/page1.html], & Merlot [http://www.merlot.org/] are examples of some of LORs.

Table 1. IEEE 9.0 category

IEEE 9.0	Classification	This category describes where the learning object falls within a particular classification system.
9.2	Taxon Path	A taxonomic path in a specific classification system. Each succeeding level is a refinement in the definition of the preceding level.
9.2.1	Source	The name of the classification system. This data element may use any recognized "official" taxonomy or any user-defined taxonomy.
9.2.2	Taxon	A taxon is a node that has a defined label or term. A taxon may also have an alphanumeric designation or identifier for standardized reference. An ordered list of taxons creates a taxonomic path.
9.2.2.1	Id	The identifier of the taxon, such as a number or letter combination provided by the source of the taxonomy.
9.2.2.2	Entry	The textual label of the taxon.
9.3	Description	Description of the learning object relative Classification
9.4	Keyword	Keywords and phrases descriptive of the learning object relative to the stated Classification

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 Rep. (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.

Greenberg.j [2] identifies four different people which may be involved in metadata tagging. Professional metadata creator, technical metadata creators, content creator and community enthusiasts so that annotation when done manually is a labor intensive, time consuming and costly activity. And sometimes the tagging is not done satisfactorily correctly.

Keyword is important metadata specified in IEEELOM 9.4. This metadata is important for the classification purpose and in turns searching purpose. Correctly tagged keywords with learning object improve the accessibility and quality. The Learning Object contains hundreds of thousands words with it. All of them are not treated as keywords. The keywords of learning object are dedicated words of the class in which the Learning Object belongs. The paper uses the Memetic algorithm based technique for extracting the keywords.

2. RELATED WORK

Whatever be the mode of learning conventional or e-learning, the importance of keyword is ubiquitous. Key word extracted will determine whether the given document matches the interest of reader or not .The mutual information used by Steier and Belew [3] to finds two-word key phrases. Munoz [4] uses an algorithm, based on Adaptive Resonance Theory (ART) neural networks. A large list of phrases produced, causes low F-measure. Ohsawa, et al. [5], develop a method Key Graph using the clustering to find which words in a document are representative of it. Turney [6] develop GenEx algorithm for extracting the keywords of learning object. A Naive Bayes method is used on the same document collection as used by turney by Frank [7], with improved results observed. Matsuo and Ishizuka [8] to extract keywords from a single document by using word co-occurrence. Ercan & Cicekli [9] proposed that lexical chains can be useful to trace the significant words. They build decision tress using the C4.5 to determine the suitability of given word as a key word. Coursey et al.[10],present several methods Union, intersection and LCS method for automatic keyword extraction and evaluate them on a collection of learning objects of an undergraduate history course.

3. EXTRACTING IEEE LOM 9.4

3.1 Feature selection

To process the textual learning object it is needed to convert it into a feature vector. The bag of words method is used most commonly, where each term in the textual learning object is treated as a feature. A document can contain thousand of terms. Therefore, the feature set will be of high dimensionality and is difficult to maintain. Due to high dimensionality it causes over fitting. It also takes more processing time to identify the class of the unknown sample. It is required to select the most significant features from the text document to reduce the dimension of the feature space vector.

To overcome the problem of over fitting the first step is to remove stop words which have no important information in it and occur in almost each document repeatedly. The preprocessing steps will targeted to remove these common words like 'the', 'to', 'and', 'a', 'an' and so on. Then after stemming, count number of times each term occurred in document. The terms & it's frequency in documents is stored in matrix form. Where each row represents a document and the indices of columns represent the terms. The value stored at the intersection of row & column is the frequency of that term in that document. So the documents appeared in the following way after this

$$D_I = \{(Term_{I1}, Frq_{I1}), (Term_{I2}, Frq_{I2}), \ldots\ldots (Term_{In}, Frq_{In})$$

This representation is known as term-frequency representation where each term is considered as a feature. Although the size of raw document is reduced by great extent. But all these terms are not considered as the strong representative of that learning object e.g. the keywords of that learning object. The keywords are important metadata of learning object which will come under the category of IEEE LOM 9.0 that is dedicated for classification. The good set of keywords for learning object is the glimpses of the learning object itself.

3.2 K-Nearest Neighbor

KNN makes its prediction based on the K training patterns that are closest to the unlabelled testing sample. A testing sample is classified by a majority vote of its nearest neighbors. The testing sample being assigned to the class most common amongst its K nearest neighbors where K is a positive integer. Let the number of training documents:
Tr= {Tr1, Tr2, Tr3----------Tr n}
Let the number of testing documents:
Te= {Te1, Te2, Te3----------Te m}
The training and testing documents are represented in the vector space model.
Let the test document Te be represented in the vector space
Te = ($a_{1,1}$, $a_{1,2}$, $a_{1,3}$---$a_{1,k}$).
Let the training document Tr $_l$ be represented in the vector space Tr $_l$1 =($b_{l,1}$, $b_{l,2}$, $b_{l,3}$--------$b_{l,k}$).
Where $a_{1,i}, b_{l,i}$ (1≤i≤k) is the frequency of term i of testing & training documents respectively.
The distance between the training and testing document is defined as

$$d(Te, Tr_{\,l}) = \sqrt{\sum_{i=1}^{k} \left(a_{1,i} - b_{l,i}\right)^2}$$

In KNN the algorithm calculate the distance between testing to each training sample. Now majority votes of K training samples in the neighbors of testing sample decide the class of testing sample.

3.3 Memetic Algorithm

Memetic comes from the term 'meme'[12]. Memetic Algorithms (MAs) are population-based metaheuristics composed of an population-based global search and a set of local search algorithms. It is inspired by Neo-Darwinian's principles of natural evolution and Dawkins' notion of a meme defined as a unit of cultural evolution that is capable of local refinements. [12].

The MA has memes similar to chromosomes. Meme is term of philosophy intended as unit of cultural transmission. In it the fittest idea of society remain unchanged while the weak idea constantly disappear and replaced by more fittest idea[13].The traditional evolutionary algorithm like GA start the evolution from the random initial solution while the MA start with the quality solution initially by applying the local search on chromosomes.

Elite_group is the variable that contains the features whose presence causes the improvement of accuracy.

Non_Elite_group is the variable that contains the features whose presence causes the degradation of accuracy.

Initially the value of Elite_group & Non_Elite_group is null. Elite_group & Non_Elite_group updates during each occurrence of local search. The memetic algorithm starts with the generation of population. The population of memetic is known as chromosomes. The random population of chromosomes of fixed size containing the binary values 0 & 1at each bit position is generated\. The 0 indicates the absence of that feature from chromosome while 1 indicates the presence of that feature in chromosome.

Chromosome 1	10100010001 11 10
Chromosome 2	11100111101 00 10
Chromosome N	00010110101 00 11

Fig 1: Population of chromosomes

The algorithm starts with the random initial population of chromosomes and then for each chromosome performs the local search. The local search is the method that observes the set of solutions repeatedly near the current solution and replaces the solution with better solution. Local search is applied on some randomly selected features by changing the feature status from absent to present in current chromosome. If randomly generated feature value is already 1 in chromosome, target another feature that is absent in the chromosome and make it present .The accuracy new chromosome is calculated to judge the fitness. Now one of the three scenarios can happen

- If new_accuracy is greater than the initial_pre_accuracy or pre_accuracy, the feature is added in Elite_group
- If new_accuracy is less than initial_pre_accuracy, the feature is added to Non_Elite_group
- If new_accuracy equal pre_accuracy, no entry in Elite_group & Non_Elite_group is done.

On termination of local search for chromosome the local optima is achieved. Once the local search on the initial population performed, each chromosome is replaced by the chromosome with local optima value.

After initialization the algorithm iterates up to predefined number of iterations by selecting the crossover or the mutation operation randomly. If crossover operation is selected two chromosomes from the initial population is selected randomly to give birth the two new offspring by setting the crossover points. Two point crossover operation is applied. Again the local search operation on both the offspring is applied. CP1, CP2 indicates two crossover points

					CP1					CP2		
1	0	0	0	1	0	0	0	1	1	1	1	P_1
1	0	0	1	1	1	1	0	1	0	0	1	P_2
1	0	0	1	1	1	1	0	1	1	1	1	OF_1
1	0	0	0	1	0	0	0	1	0	0	1	OF_2

Fig 2: Two point Crossover Operation

P_1 & P_2 are two parents to perform the crossover. OF_1 & OF_2 are two offspring's produced as a result of two point crossover applies over parents.

The mutation is performed by flipping the bit position from 0 to 1 and 1 to 0 to be mutated. The mutation rate will decide the number of positions to be mutated. A good mutation rate should not be too high or too low.

1	0	0	0	1	0	0	0	1	1	1	1	Parent
0	1	0	1	1	0	0	1	0	0	0	1	M1
1	**1**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**M2**

Fig 3: Mutation Operation

Here in the M_1 is mutation chromosome and M_2 is mutated chromosome Here the mutation chromosome indicates the bit position to be flipped in parents. The 1 in mutation chromosome will show the bit positions in parent to be flipped resulted in mutated chromosome. Now apply the local search on mutated chromosome. This will update the Elite_group & Non_Elite_group.

Now after performing crossover or mutation operation, examine the accuracy of offspring chromosomes (icross) or the mutated chromosome (imut). If icross or imut is better than the worst chromosome present in initial population then replace the worst chromosome with the icross or imut and continue the process for all the iteration . Global optima are achieved in the end of process with the important keyword list that affects the accuracy most. The memetic based keywords selection algorithm is as follows

Pseudo-code for a MA

```
Begin;
Randomly generate initial population of chromosomes of same size;
Elite_group={ }
Non_Elite_group={ }
Apply local search for each chromosomes
For j = 1 to max_iter
Randomly selection crossover or mutation
If crossover;
Select two parents ichromo1 and ichromo2 at random;
Generate offspring icross= crossover (ichromo1 and ichromo2);
Do local search for each offspring (icross);
Else if mutation;
Select a chromosome i at random;
Generate an offspring imut = mutation (i);
Do local search for each offspring (imut);
End if ;
If icross or imut is better than the worst Chromosome then
Replace worst chromosome by icross or imut;
Next j;
Best features stored in Elite_group
End;
```

Pseudo code for local search

```
Begin;
Randomly generate generations for local search
ran= random number between 5 to 20
k=length(chromosome)/ran
loc_pos=Generate k random numbers between 1 to length(chromosome)
calculate accuracy using KNN classifier of given chromosome i
initial_ pre_ accuracy = accuracy of original_chromosome
pre_ accuracy = initial_ pre_ accuracy
For M=1 to loc_pos
If chromosome's feature is present in chromosome i or feature is in Non_Elite_group then
Randomly select another feature
Make the feature present in chromosome i
Accuracy using KNN classifier
Else
Accuracy using KNN classifier
End If
If new_ accuracy > pre_ accuracy then
Replace the original chromosome with modified_chromosome
pre_accuracy=new_accuracy
Add the feature in Elite_group of feature
```

```
Else if (new_ accuracy < pre_ accuracy) & (new_ accuracy > initial_ pre_ accuracy ) then
Add the feature in Elite_group of feature
Else
Add the feature in Non_Elite_group
Retain the original chromosome;
End If;
Next M;
End;
```

Each learning object is represented in vector space form(<t1,f1>,<t2,f2>…………..<tn,fn>) where t1,t2,….tn are the terms appeared in that learning object and f1,f2……..fn are the frequencies of these terms. Not all the terms are so important to be used as the dedicated keywords of that learning object. During the process of memetic algorithm, Elite_group is maintained, which is used to contain that features that causes the increase in overall accuracy. The features that appeared in the learning object and at the same time they are also appeared in the Elite_group is regarded as the best representing keywords of that learning object. The good feature improves the searching time.

4. THE CORPORA

Data is collected from the publically available Cornell university library (www.in.arXiv.org). In this library, manually tagged documents of different subjects are available. Paper performed experiments on total 9 categories, of which 5 belong to astro physics & 4 categories belong to high energy physics. 70 & 100 documents of each category are used for training & testing purpose respectively.

Table 2. Topics of Astro & High Energy Physics

Class of Astro physics	Class of High Energy Physics
Cosmology	High energy Physics-Experiment
Earth & Planetry	High Energy physics-theory
Galaxy	High Energy Physics-Lattices
High Energy Astro	High Energy Physics-Phenomenal
Solar & Steller	

5. EVALUATION

Accuracy achieved at each iteration is used to evaluate the performance of algorithm. The accuracy is calculated as number of test samples correctly classified by the algorithm. The fig.4 shows the performance of algorithm for 15 chromosomes for 20 iterations. The algorithm optimized the solution space in terms of accuracy.

Table 3. Glimpses of output for two classes

Class of LO	Some Keywords
Cosmology	cosmology, dark, flux, emission, density, convolute, consid, spectrum, anticommut, antineutrino ,antiparticle ,antiproton, antiquark,antisymmetr, appelquist amplitude
High Energy Physics-Ph.	action ,boson ,boundary ,charge ,compute condit, coupl , fermion , gaug , gener, hep, hole, horizon, jhep, momentum, paramet, space time, state, supersymmetr, gravity

The best representative keywords of any learning object is the find by the intersection of Elite_group and the terms appeared in the document's vector space form. Maximum accuracy achieved with data is 76%. This is due to the fact that documents extracted from Cornell university site are research articles and are not topic centered articles. The area of these articles covered many topics.

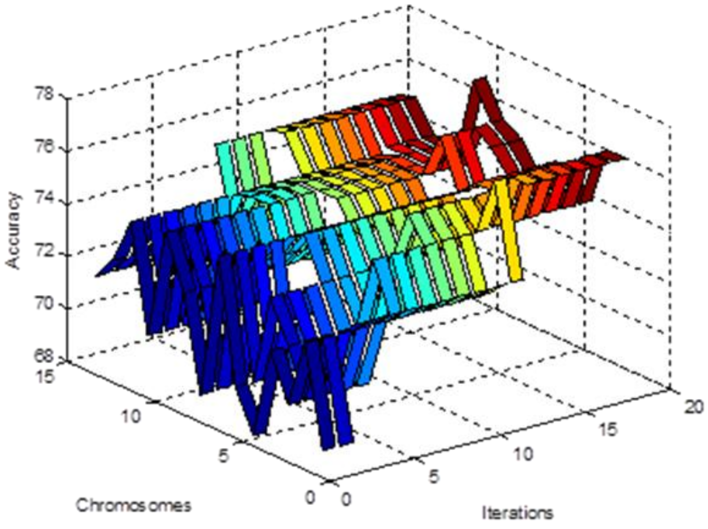

Fig 4: Performance of memetic algorithm in each iteration

6. CONCLUSION

Learning Object should be able to searched, evaluated by different users. A Learning object with more & correct information tagged with it augments the reusability, which is the main thought behind the learning content management. IEEE LOM has 60 elements divided in to 9 different categories. Each category has its own importance. The metadata items like author name, date etc is very easy to tag. The Keywords discussed in category 9 states about the important terms of that learning object reflecting the content & in turn class of learning object. As always be a dedicated set of keywords about any topic make searching fast, easy & correct, ultimately results in increasing quality of learning object.

Memetic algorithm an evolutionary technique, for optimized classification result is being discussed in this paper. The accuracy of classification indicates the quality of keywords used for classification. The dedicated keywords results in optimized classifier result. The dedicated set of keywords for learning objects of each class has extracted. The classifier results show about 76 % accuracy is achieved. This is due to the fact that documents extracted from Cornell university site are research articles and are not topic centered articles. The area of these articles covered many topics.

7. REFERENCES

[1] IEEE. IEEE Std 1484.12.1-2002, IEEE Standard for Learning Object Metadata. http://www.ieee.org/.

[2] Greenberg, J. 2003. Metadata Generation: Processes, People and Tools. Bulletin of American Society for Information Science and Technology, 29(2) (Dec.2003), 16-19.

[3] Steier, A. M., and Belew, R. K. 1993. Exporting phrases: A statistical analysis of topical language. In R. Casey and B. Croft, editors, Second Symposium on Document Analysis and Information Retrieval, 179-190.

[4] Muñoz, A. 1996. Compound key word generation from document databases using a hierarchical clustering ART model. Intelligent Data Analysis.1 (1)(1996),25-48.

[5] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida 1998. Keygraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Proceedings of the Advances in Digital Libraries Conference, ADL, IEEE Computer Society.

[6] Turney, P. 1999. Learning to extract key phrases from text. Technical report, National Research Council, Institute for Information Technology.

[7] Frank,E., Paynther, G.W. , Witten, I. H., Gutwin, C. & Nevil-Manning, C.G. 1999. Domain-specific key phrase extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence.

[8] Y. Matsuo and M. Ishizuka 2004. Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools. 13(1) (2004), 157-169.

[9] G. Ercan and I. Cicekli 2007. Using lexical chains for keyword extraction. Information Processing & Management. 43(6)(2007), 1705–1714.

[10] Coursey, K.H., Mihalcea R., Moen,W.E. 2008.Automatic keyword extraction for learning object repositories. In Proceedings of the American Society for Information Science and Technology

[11] Dawkins, R 1976. The Selfish Gene. Clarendon Press, Oxford.

[12] Yew-Soon Ong, Meng-Hiot Lim, Ning Zhu, and Kok-Wai Wong 2006. Classification of Adaptive Memetic Algorithms:A Comparative Study. IEEE Transactions On Systems, Man, and Cybernetics—Part B: Cybernetics. 36(1)(February 2006), 141-152.

[13] F. Neri et al. (Eds.) 2012. Handbook of Memetic Algorithms, Studies in Computational Intelligence, Springer. SCI 379, 43–52.