ISSN:0975-9646

Clustering of Documents using Particle Swarm Optimization and Semantics Information

Sunita Sarkar[1], Arindam Roy[2], Bipul Syam Purkayastha[3]

[1,2,3]Department of Computer Science, Assam University
Silchar, Assam, India

Abstract. **With the ever increasing volume of information, document clustering is used for automatic document organization so as to yield relevant information in an expeditious manner. Document clustering is an automatic grouping of text documents into clusters so that documents within a cluster have similar concepts. Representation of document is a very important step in any Information Retrieval (IR) system. In traditional document representation methods, the feature vector representing the document is constructed from the frequency count of document terms. But traditional document representation methods can not identify semantically related terms. In this paper, we present a semantic document clustering method that uses Universal Networking Language(UNL) and Particle Swarm Optimization(PSO). We generate feature vectors using UNL. The hybrid PSO+K-means algorithm is used to cluster the documents. Some experiments are performed to compare efficiency of the UNL method with the traditional term frequency based method. The results obtained show that the PSO-based clustering method using the UNL performs better than the term frequency based Method.**

Keywords—*Universal Networking Language; Document clustering; Particle Swarm Optimization; K-means.*

I. INTRODUCTION

Document clustering is a fundamental operation used in unsupervised document organization, automatic topic extraction, and information retrieval[1]. Document clustering can be defined as grouping of text documents into clusters so that documents within a cluster have similar concepts. The representation of documents is an important issue in Information Retrieval (IR) system. The representation should be as compact as possible in order to allow efficient processing of large document collections, yet it should contain all of the relevant information. The vector space model is a widely used method for document representation in information retrieval. In this model, each document is represented by a feature vector. The unique terms occurring in the whole document collection are identified as the attributes (or features) of the feature vector[2]. Binary method, tf (term frequency) method [4], and tf-idf (inverse document) method [3] etc. are different term weighting methods may be used in the vector space model. The "bag of words" feature representation is used in most cases. But such a "bag of words" feature representation is not able to reflect the semantic content of a document because of the synonym problem and polysemy problem. For instance, having "intelligent" in one document and "brilliant " in another document does not contribute to the similarity measure among these two documents. Two terms with a close semantic relation and two other terms with no semantic relation are both treated in the same way. This ignorance about semantics could reduce clustering quality result.

In this paper, we present a semantic text document clustering approach that using Universal Networking Language and Hybrid PSO+K-means algorithm. We generate feature vectors using Universal Networking Language (UNL) and then clustering of the documents using a hybrid clustering technique. UNL captures the semantics relation between words in a sentence. The UNL represents the document in the form of a semantic graph with universal words as nodes and the semantic relation between them as links.

The rest of this paper is organized as following; In section 2, we discuss about the document representation. Related work is discussed and presented in section 3. Section 4 is devoted to UNL and feature vector generation using UNL. In section 5 hybrid PSO-Kmeans algorithm is presented. In section 6, we will give out the experimental results. Finally, conclusion and future work are given in section 7.

II. DOCUMENT REPRESENTATION

Document preprocessing is the first step to represent the input documents collection into vector space model. Preprocessing is a very important and essential phase in an effective document clustering. Document preprocessing tasks involve tokenization, removal of stop words, stemming and term weighting .

A. Document Preprocessing

Document preprocessing consists of the following steps:

1) *Tokenization.*

Tokenization is the very first step involved in most text processing tasks. A tokenizer separates a text into a set of component elements called tokens. The simplest tokenization method is splitting the text according to blanks and punctuation marks.

2) *Stop Words*

In written text, some words are very common and have no additional meaning to the actual content of the text, and has little or nothing to say about the text itself. Prepositions, conjunctions, nouns and articles are examples of such words(stop words). We can save a lot of processing time and working memory after removal of stop words.

3) *Stemming*

Stemming is the next process after stop word removal. Stemmers try to identify the stem of a raw word in a text to reduce all such similar words to a common form, making the statistical data more useful. In the process of stemming, the commoner morphological and inflexional endings are removed from words in English. For example, the phrases analysis, analyzer, and analyzing all have the stem form analy. Porter stemmer [6] and Lovins stemmer [5] are the two most widely used stemmers.

B. Document Representation

A Document is represented by a set of keywords/ terms extracted from the document. Vector-space model introduced by Salton et al.[8] is the widely used keyword based model to represent textual data .Vector Space Model uses the concepts of linear algebra to address the problem of representing and comparing textual data[7]. In this model a document d is represented as a document vector [wt_0 ,wt_1 , . . .wt_n], where t_0, t_1, . . . t_n is a set of words of a given document and wt_i is the weight (importance) of term t_i to document d. The importance of word t_i in a document d is represented by its value wt_i to that document. Given a set of documents, their document vectors can be put together to form a matrix called a term-document matrix
In the vector space model, the most widely-used weighting scheme is TF*IDF, which is the combination of the term frequency (TF) and the inverse document frequency (IDF). TF*IDF is mathematically written as

$Wij= tf_{i,j}* \log (N / df_i)$
Where w_{ij} is the weight of the term i in document j,
$tf_{i,j}$= number of occurrences of term i in document j.
N is the total number of documents in the corpus, df_i is the number of documents containing the term i.

III. RELATED WORK

Text documents clustering can be challenging due to complex linguistics properties of the text documents. Most of clustering techniques are based on traditional bag of words to represent the documents. In such document representation, ambiguity, synonymy and semantic similarities may not be captured using traditional text mining techniques that are based on words and/or phrases frequencies in the text. Gad and Mohamed S. Kamel [11] proposed a semantic similarity based model to capture the semantic of the text. The proposed model in conjunction with lexical ontology solves the synonyms and hypernyms problems. It utilizes WordNet as an ontology and uses the adapted Lesk algorithm to examine and extract the relationships between terms. The proposed model reflects the relationships by the semantic weighs added to the term frequency weight to represent the semantic similarity between terms.

In [9,10] authors introduced conceptual features in text representation. A concept feature is an aggregation of a few words that describe the same high level concept, for example, dog and cat describing the concept animal. They proposed three methods to include concept features in VSM, namely, (i) adding concept features to the term space (i.e., term+concept); (ii) replacing the related terms with concept features and (iii) reducing the VSM to only concept features For text clustering, experimental results [10] showed that only the term+concept representation improved clustering performance.

Sridevi and Nagaveni [12] showed that combination of ontology and optimization improve the clustering performance. They proposed a ontology similarity measure to identify the importance of the concepts in the document. Ontology similarity measures is defined using wordnet synsets and the particle swarm optimization is used to cluster the document.

In [15] authors proposed a semantic text document clustering approach based on the WordNet lexical categories and Self Organizing Map (SOM) neural network. The proposed approach generates documents vectors using the lexical category mapping of WordNet after preprocessing the input documents.

Liping Jing et al[13] proposed a new similarity measure combining the edge-counting technique and the position weighting method to compute the similarity of two terms from an ontology hierarchy. They modified the VSM model by readjusting term weights in the document vectors based on its relationships with other terms co-occurring in the document. They applied three different clustering algorithms, bisecting k-means, feature weighting k-means and a hierarchical clustering algorithm to cluster text data represented in knowledge based VSM

B. Choudhary and P.Bhattacharyya [16], described a new method for generating feature vectors, using UNL link and UNL relation method. UNL captures the semantic relations between the words in a sentence. The UNL presents the document in the form of a semantic graph with universal words as nodes and the semantic relation between them as links. The clustering method applied to the feature vectors is the Kohonen Self Organizing Maps (SOM). Experiments show that UNL method for feature vector generation tends to perform better than the term frequency based method. Chirag Shah et al.[17] describes a method for document vector construction where the semantic relation between words in a sentence are taken into consideration. To demonstrate that this method scores over other conventional methods viz tf ,tf-idf etc, the authors have used the mutual information concept from information theory which has been called the goodness of document vectors.

IV. UNIVERSAL NETWORKING LANGUAGE

The UNL [19] has been defined as a digital Meta language for describing, summarizing, refining, storing and disseminating information in a machine independent and human language neutral form. It represents information, i.e. meaning, sentence by sentence. Each sentence is represented as a hypergraph, where nodes represent concepts and arcs represent relation between concepts[18]. This hyper-graph is also represented as a set of directed binary relations between the pair of concepts present in the sentence. Concepts are represented as character strings called Universal Words (UWs). UNL is composed of three elements: Universal Words, Relation and Attribute.

Universal Words (UWs): Word knowledge is expressed by Universal Words which are language independent. UWs constitute the UNL vocabulary and the syntactic and semantic units that are combined according to the UNL laws to form UNL expressions. They are tagged using restrictions describing the sense of the word in the current context..

The following is the syntax of description of UWs in context-free grammar (CFG):

<UW> ::= <Head Word> [<Constraint List>]
<Head Word>::=<character>
<Constraint List> ::= "("<Constraint>[","<Constraint>]… ")"

Head Word is an English word/compound word/phrase/sentence that is interpreted as a label for a set of concepts. UW's are used to index the UNL knowledge base (UNLKB).For example: drink, eat, dog etc.

The Constraint List restricts the range of the concept that a Basic UW represents. Each restricted UW represents a more specific concept, or subset of concepts. For example:
state(equ>nation) : denotes nation.
state(icl>situation) : kind of situation
state(icl>government) : kind of government

Relation Labels: Conceptual knowledge is captured by the relationship between Universal Words (UWs) through a set of UNL relations.

For example, John is reading a Novel is described in the UNL expression as,

[UNL]
agt(read(icl>do) @entry.@present.@progress, John(iof>person))
obj(read(icl>do) @entry.@present.@progress, novel(icl>book))
[/UNL]

where, *agt* means the agent and *obj* means object. The terms *read(icl>do)*, *novel(icl>book)* and *John(iof>person)* are the UWs denoting concepts.

Attribute Labels: UNL attributes capture speaker's view, aspect, time of events et. For instance, in the above example, the attribute *@entry* denotes the main predicate of the sentence, *@present* denotes the present tense, *@progress denote an event* is in progress.

An UNL expression can also be represented as a graph. For example, the UNL expressions and the UNL graph for the sentence, *Ram went to China from India by aeroplane to attend a conference,* are shown in Fig. 1

{unl}
agt(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@past,Ram(icl>person))
plt(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@past,China(iof>asian_country>thing))
plf(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@past, India(iof>asian_country>thing))
met(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@past,aeroplane(icl>heavier-than air_craft>thingequ>airplane))
obj:01(attend(icl>go_to>do,agt>person,obj>place).@entry, conference(icl>meeting>thing).@indef)
pur(go(icl>move>do,plt>place,plf>place,agt>thing).@entry.@past,:01)
{/unl}

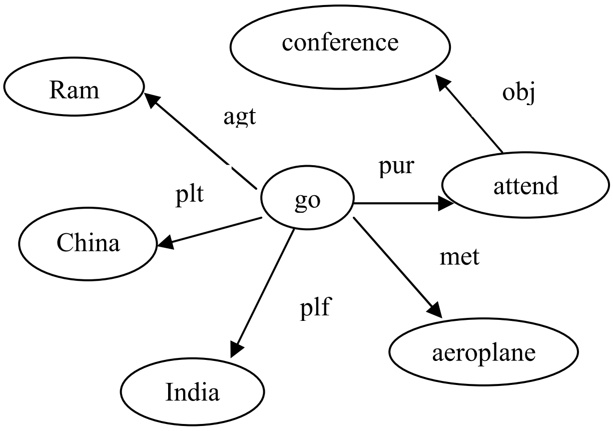

Fig.1.UNL expression and UNL graph

A. Generation of Document vector using UNL Links

UNL is an Interlingua representation of a document, which represents information in the form of semantic graphs with hyper-nodes. Each of the nodes represents an unambiguous concept and edges represent the relationship that they form among themselves. In the UNL link method, Universal Words are used as the components of the document vector[16]. Since each UW is represented in an unambiguous manner, multiple words in the document get automatically differentiated, thereby producing correct frequency count.

For example in the sentence, *A bear can bear very cold temperatures* The word bear has two different senses, viz., bear(icl>animal) and bear(icl>tolerate).

Hence, the frequency count of 2 is wrong for this word. They find different places in the document vector. After this, each component of the document vector which represents a different universal word (i.e., a concept) is assigned the number of links incident on the node, considering the graph to be undirected. The weight of each node in the graph is determined by the number of links to the node. When a UW is not present in the UNL graph of the document then 0 is written in its position.

The process of construction of the document vector from the UNL representation is described [16]:

1. Parse the UNL document to construct the UNL graph.
2. For each UW in the UNL graph count the links to other UWs from it.
3. Construct the feature vector by merging the counts got from step two.
4. Output the feature vector.

For example, consider the following two UNL graph given in figure 2 and 3 as given documents.

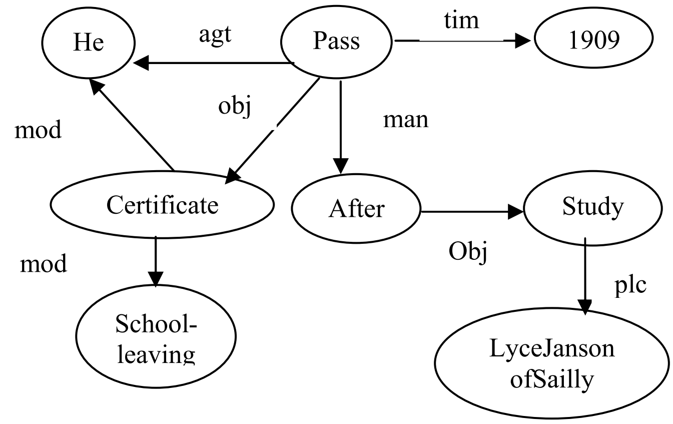

Fig.2. UNL graph for the sentence "after studying at the LyceeJanson of Sailly, he passed his school-leaving certificate in 1909"

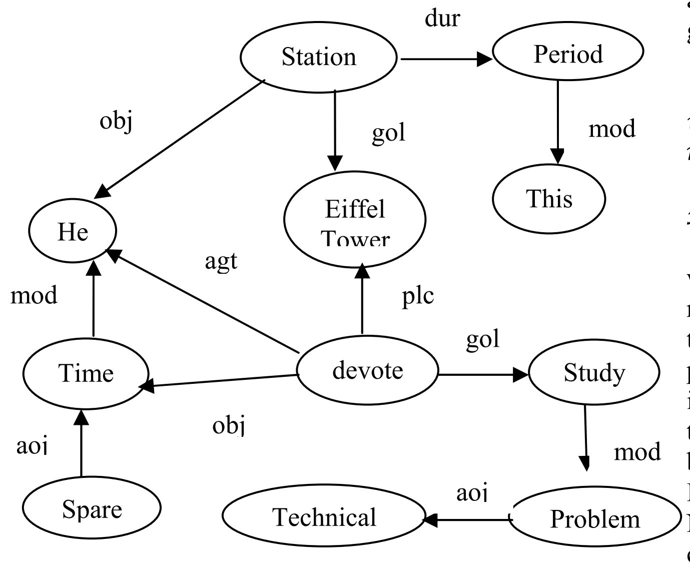

Fig.3.UNL graph for the sentence "during this period he was stationed at the Eiffel Tower, where he devoted his spare time to the study of technical problems"

Here, the order of words is pass, certificate, after, he, 1909, school-leaving, study,LyceeJanson of Sailly, station, Eiffel tower, period, devote, this, time, study, spare, problem and technical. The vectors corresponding to the graphs are

V1 =<4,3,2,2,1,1,2,1,0,0,0,0,0,0,0,0,0,0>
V2=<0,0,0,3,0,0,0,0,3,2,2,4,1,3,2,1,2,1>

The weight of each word is calculated using the method described above.

V. Hybrid Particle Swarm Optimization And K-Means Clustering Algorithm

A. Particle Swarm Optimization

PSO is a evolutionary computation technique first introduced by Kennedy and Eberhart in 1995 [20]. PSO is a population-based stochastic search algorithm which is modeled after the social behavior of a bird flock. In the context of PSO, a swarm refers to a number of potential solutions to the optimization problem, where each potential solution is referred to as a particle. The aim of the PSO is to find the particle position that results in the best evaluation of a given fitness (objective) function [23].

Each individual in the particle swarm is composed of three D-dimensional vectors, where D is the dimensionality of the search space. These are the current position x_i the previous best position p_i, and the velocity v_i [21] The i^{th} particle is represented by a position denoted as $x_i = (x_{i1}, x_{i2},. . . , x_{iD})$. In a PSO system, each particle flows through the multidimensional search space, adjusting its position in search space according to its own experience and that of neighboring particles. To evolve towards an optimal solution a particle uses a combination of the best position realized by itself and the best position realized by its neighbours. The standard PSO method updates the velocity and position of each particle according to the equations given below.

$$v_{id}(t+1) = \omega \cdot v_{id}(t) + c_1 \cdot rand() \cdot (p_{id} - x_{id}) + c_2 \cdot rand() \cdot (p_{gd} - x_{gd}) \qquad (1)$$

$$x_{id}(t+1) = v_{id}(t+1) + x_{id}(t) \qquad (2)$$

where c_1 and c_2 are two positive acceleration constants, rand() is a uniform random number in (0, 1), p_{id} and p_{gd} are the best positions found so far by the i^{th} particle and all the particles respectively, t is the iteration count and ω is an inertia weight which is usually, linearly decreasing during the iterations. The inertia weight ω plays a role of balancing the local and global search.
In the context of clustering, a single particle represents the N_c cluster centroid vectors. That is, each particle x_i is constructed as follows:

$$x_i=(o_{i1},...,o_{ij},...,o_{iNc}) \qquad (3)$$

Where o_{ij} refers to the j^{th} cluster centroid vector of the i^{th} particle in cluster C_{ij}. Therefore, a swarm represents a number of candidate clusters for the current data vectors. The fitness of particles is measured using the equation given below.

$$\phi = \frac{\sum_{i=0}^{N_c}\{\frac{\sum_{j=0}^{P_i} d(o_i, m_{ij})}{P_i}\}}{N_c} \qquad (3)$$

where m_{ij} denotes the j^{th} document vector, which belongs to cluster i; o_i is the centroid vector of the i^{th} cluster; $d(o_i, m_{ij})$ is the distance between document m_{ij} and the cluster centroid o_i;P_i stands for the number of documents, which belongs to cluster C_i; and N_c stands for the number of clusters.

The PSO Clustering algorithm can be summarized as:

```
(1) Initially, each particle randomly select k different document vectors from the document collection as the initial cluster centroid vectors.

(2) For t= 1 to $t_{max}$ do
   a)For each particle i do:
   b)For each document vector $m_p$ do
      (i) Calculate the distance d $(m_p,o_{ij})$, to all cluster
          centroids $C_{ij}$
      (ii) Assign each document vector to the closest centroid vector.
      (iii) Calculate the fitness value based on equation 4.
   c) Update the global best and local best positions
   d) Update the cluster centroids using equations (1)
      and (2)
      Where $t_{max}$ is the maximum number of iterations.
```

B. K-Means Algorithmn

In K-means algorithm data vectors are grouped into predefined number of clusters. At the beginning the centroids of the predefined clusters are initialized randomly. The dimensions of the centroids are same as the dimension of data vectors. Each data object is assigned to the cluster based on the similarity between the data object and the cluster centroid. The reassignment procedure is repeated until the fixed iteration number, or the cluster result does not change after a certain number of iterations.

.

The K-means algorithm is summarized as

1. Randomly initialize the N_c cluster centroid vectors

2. Repeat

(a) For each data vector, assign the vector to the class with the closest centroid vector,

(b) Recalculate the cluster centroid vectors, using

$$C_j = \frac{1}{N}\sum_{\forall d_{j \in S_j}} D_j \qquad (4)$$

until a stopping criterion is satisfied

C. Hybrid Particle Swarm Optimization+K-Means Algorithm

In the hybrid PSO clustering algorithm, the multi-dimensional document vector space is modeled as a problem space. Each term in the document dataset represents one dimension of the problem space[1].In the hybrid PSO algorithm [1], the algorithm includes two modules, the PSO module and the K-means module. The hybrid algorithm first executes PSO clustering algorithm to find points close to the optimal solution by global search and simultaneously avoid high computation time. In this case PSO clustering is terminated when the maximum number of iterations is exceeded. The result of the PSO algorithm is then used as initial centroid vectors of the K-means algorithm. The K-means algorithm is then executed until maximum number of iterations is reached. The K-means algorithm tends to converge faster (after less function evaluations) than the PSO, but usually with a less accurate clustering [13] and PSO can conduct a globalized searching for the optimal clustering, but requires more iteration numbers and computation than the K-means algorithm. The hybrid PSO algorithm combines the advantage of both the algorithms: globalized searching of the PSO algorithm and the fast convergence of the K-means algorithm.

VI. EXPERIMENTAL RESULT

UNL data is collected from UNL site [22]. For the term frequency method in those document datasets, stop words are removed completely and different forms of a word are reduced to one canonical form by using Porter's algorithm [6].

In this paper, we created the document vector by UNL link method proposed in [16], TF and TFIDF method. The hybrid PSO algorithm is applied to document vector created by term frequency method and UNL link method. The size of the vector created by TF method was 534 and the size of the vector created by UNL link method was*561*.The cosine correlation measure is used as the similarity metrics in this algorithm. In the PSO clustering algorithm, we choose 10 particles, the inertia weight ***w*** is initially set as 0.72 and the acceleration coefficient constants ***c1*** and ***c2*** are set as1.49. In this study the quality of the clustering is measured according to the following two criteria:

Iintra-cluster similarities, i.e. the distance between data vectors within a cluster, where the objective is to maximize the intra-cluster similarity;

Inter-cluster similarity, i.e. the distance between the centroids of the clusters, where the objective is to minimize the similarity between clusters.

These two objectives respectively correspond to crisp, compact clusters that are well separated. Both intra-cluster similarity and inter-cluster similarity are internal quality measures.

Accuracy of clustering is given by[16],

Accuracy = No. of documents correctly clustered / Total no. of documents

The results obtained are shown in table 1.

Method	Intra cluster	Inter cluster	Accuracy
TF	.1500	.3688	.6428
TF-IDF	.6949	.6270	.8571
UNL	.7442	.7340	.9763

Table 1. Results

CONCLUSION

In this paper document vectors are created by term frequency method and UNL link method and Hybrid PSO algorithm is applied to cluster the documents. UNL method uses the semantic information present in the form of relations between words in sentences. In this paper the performance evaluation of term frequency method and UNL link method has been studied. The performances of the UNL link method were compared with the tradition

TF and TFIDF method. The result shown in Table 1 depicted betterperformance for UNL link method than term frequency methods.

REFERENCE

1. Cui X, Potok T.E, Palathingal P (2005) Document Clustering using Particle Swarm Optimization. IEEE
2. Wang Y and Hodges J(2005) Document Clustering using Compound Words. ICAI'05 - The International Conference on Artificial Intelligence.
3. Salton G(1971), The SMART Retrieval System – Experiments in Automatic Document Retrieval. New Jersy, Englewood Cliffs: Prentice Hall Inc.
4. Salton G and. Buckley C (1988) Term-Weighting Approach in Automatic Text Retrieval. Information Processing & management, vol. 24, no. 5, , pp. 513-523.
5. Lovins J.B (1968) Development of a Stemming Algorithm. Mechanical Translation andComputational Linguistics, vol. 11, pp. 22-31.
6. Porter M.F (1980) An Algorithm for Suffix Stripping. Program, vol. 14, no. 3, pp. 130-137
7. Weiss D, Descriptive Clustering as a Method for Exploring Text Collections, Ph.D Thesis.
8. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18:613– 620
9. Hotho A, Staab S, Stumme G (2003)Wordnet improves text document clustering. In: Proceedings of the semantic web workshop at 26th annual international ACM SIGIR conference, Toronto, Canada
10. Hotho A, Bloehdorn S (2004) Text classification by boosting weak learners based on terms and concepts.In Proceedings of the proceedings of the IEEE internal conference on data mining, Brighton, UK, pp 72-79
11. Walaa K. Gad and Mohamed S. Kamel (2009) Enhancing Text Clustering Performance Using Semantic Similarity, ICEIS, LNBIP 24, pp. 325–335, 200
12. Sridevi.U. K. and Nagaveni. N(2011) Semantically Enhanced Document Clustering Based on PSO Algorithm European Journal of Scientific Research, ISSN 1450-216X Vol.57 No.3, pp.485-493
13. Jing L, Michael K. Ng , Huang J. Z.(2010) Knowledge-based vector space model for text clustering KnowlInfSyst 25:35–55
14. Song L, Ma J, Yan P, Lian L and Zhang D (2008). "Clustering deep web databases semantically", Proceedings of the 4th Asia information retrieval conference on Information retrieval technology, pp. 365-376.
15. Gharib T.F, Fouad M. M, Mashat A, Bidawi1 I(2012) Self Organizing Map -based Document Clustering Using WordNet Ontologies. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2
16. Choudhary B and Bhattacharyya P(2002). Text clustering using semantics. In The Eleventh International World Wide Web Conference, 2002.
17. Shah C and Bhattacharyya P (2002) Constructing Better Docuement Vectors Using Universal Networking Language. International Conference on Knowledge Based Computer Systems (KBCS 2002), Mumbai, India.
18. Ali Md. N. Y, Ripon. S and Allayear S.M (2012)UNL Based Bangla Natural Text Conversion –Predicate Preserving Parser Approach IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 3.
19. Uchida, H. and M. Zhu, The Universal Networking Language (UNL) Specification Version 3.0, 1998, United Nations University: Tokyo, Japan.
20. Kennedy, J. and Eberhart R.C.(1995) Particle Swarm Optimization. Proc. IEEE, International Conference on Neural Networks. Piscataway. Vol. 4, pp 1942-1948,
21. Sarkar S, Roy A, Purkayastha B S(2013) Application of Particle Swarm Optimization in data clustering : A survey. International Journal Of Computer Applications (0975- 8887) Volume 65- No.25.
22. http://www.undl.org/unldoc/bb.htm
23. Satapathy S.C, B. Rao N. VSSV P, Murthy JVR, Prasad R. P.V.G.D(2007) A Comparative Analysis of Unsupervised K-means, PSO and Self- Organizing PSO for Image Clustering. International Conference on Computational Intelligence and Multimedia Applications.