Computational Intelligence Methods for Clustering of Sense Tagged Nepali Documents

Sunita Sarkar[1], Arindam Roy[2], Bipul Syam Purkayastha[3]
[1,2,3] *(Department of Computer Science , Assam University ,India)*

Abstract: *This paper presents a method using hybridization of self organizing map (SOM), particle swarm optimization(PSO) and k-means clustering algorithm for document clustering. Document representation is an important step for clustering purposes. The common way of represent a text is bag of words approach. This approach is simple but has two drawbacks viz. synonymy and polysemy which arise because of the ambiguity of the words and the lack of information about the relations between the words. To avoid the drawbacks of bag of words approach words are tagged with senses in WordNet in this paper. Sense tagging of words provide exact senses of words. Feature vectors are generated using sense tagged documents and clustering is carried out using proposed hybrid SOM+PSO+K-means algorithm. In the proposed algorithm initially SOM is applied to the feature vectors to produce the prototypes and then K-means clustering algorithm is applied to cluster the prototypes. Particle Swarm Optimization algorithm is used to find the initial centroid for K-means algorithm. Text documents in Nepali language are used to test the hybrid SOM+PSO+K-means clustering algorithm.*

Keywords: *Computational Intelligence, Sense tagging, Self organization map, Particle swarm optimization.*

I. Introduction

Computational intelligence (CI) is a set of nature-inspired computational methodologies and approaches to address complex real-world problems. Paradigms that comprise CI techniques are Neural networks, evolutionary computing, swarm intelligence and fuzzy systems. Computational Intelligence methods have been successfully applied to many fields such as diagnosis of diseases, speech recognition, data mining, composing music, image processing, forecasting, robot control, credit approval, classification, pattern recognition, planning game strategies, compression, combinatorial optimization, fault diagnosis, clustering, scheduling, and time series approximation. control systems, gear transmission and braking systems in vehicles, controlling lifts, home appliances, controlling traffic signals, and many others[1]. This paper concerns with the application of CI methods in document clustering.

Document clustering is the process of grouping/dividing a set of documents into subsets (called clusters) so that the documents are similar to one another within the cluster and are dissimilar to documents in other clusters. Vector space model with bag of words is the common approach for representing a text. This approach suffers from two drawbacks viz. several words can have same meanings (synonymy) and same words can have multiple meanings (polysemy). In this paper an attempt has been made to handle these issues by tagging the words with senses. Given a word and its possible senses, as defined in a knowledge base, sense tagging is the process of assigning the most appropriate senses to the words in the corpus within a given context where sense can be defined as semantic value (content) of a word when compared to other words; i.e. when it is part of a group or set of related words[2]. When words are sense tagged, the most appropriate senses are attached to the words. In this work feature vectors are generated using sense tagged document corpus and clustering is done by a hybrid SOM+PSO+K-means clustering algorithm.

Self organizing map[3] is an artificial neural network and have been successfully applied to document clustering. The SOM is an algorithm used to visualize and interpret large high-dimensional data sets. It is an unsupervised learning algorithm. It produces a set of prototype vectors representing the data set and carries out a topology preserving projection of the prototypes from the n-dimensional input space onto a low-dimensional grid.

PSO is a computational intelligence technique first introduced by Kennedy and Eberhart in 1995[4] . 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[5].

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:

$$x_i=(o_{i1},...,o_{ij},...,o_{iNc}) \qquad (1)$$

DOI: 10.9790/0661-17158389

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.

$$f = \frac{\sum_{i=1}^{N_c}\left\{\frac{\sum_{j=1}^{P_i} d(o_i, m_{ij})}{P_i}\right\}}{N_c} \tag{2}$$

where m_{ij} denotes the j^{th} data 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 data vector m_{ij} and the cluster centroid o_i; P_i stands for the number of dataset, which belongs to cluster C_i and N_c stands for the number of clusters.

In this paper a hybrid SOM+PSO+K-means algorithms is presented for document clustering. Several experiments were performed to analyze the performance of clustering algorithm on sense-tagged Nepali document corpus and Nepali document corpus without sense-tagging.

Nepali is an Indo-Aryan language spoken by approximately 45 million people in Nepal, where it is the language of government and the medium of much education, and also in neighboring countries (India, Bhutan and Myanmar). Nepali is written in the Devanagari alphabet. It is written phonetically, that is, the sounds correspond almost exactly to the written letters. Nepali has many loanwords from Arabic and Persian languages, as well as some Hindi and English borrowings[6].

The rest of this paper is organized as follows: Related work is discussed in section II. Section III provides a description of generation of document vectors. In section IV clustering algorithm is discussed. Section V discusses the experimental results and Section VI concludes the paper.

II. Related Works

Many clustering techniques have been developed and successfully applied for clustering of text documents. Cui and Potok [7] proposed a hybrid PSO + K -Means algorithm for clustering of text documents. Document vectors were generated using tf-idf method. They applied three different clustering algorithms namely K-Means, PSO and Hybrid PSO+ K-Means to the generated document vectors. The result showed that the proposed Hybrid PSO+ K-Means algorithm achieves higher clustering quality than other two clustering algorithms.

Lo[8] applied Self-Organizing Map (SOM) algorithm to cluster Chinese botanical documents. Vector Space Model with bag of words approach was used to represent each botanical document.

Xinwu[9] presented a text clustering algorithm combining K-means algorithm and SOM. In the proposed approach clustering is typically carried out in two stages: first, the data set is clustered using the SOMs to obtain a group of output weights which. the number of SOM network's output nodes equals to the number of the texts' categories. The results obtained from SOM are used to initialize K-means algorithm's cluster centers and implement K-means algorithm to cluster the text sets.

Sridevi and Nagaveni,[10] 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 [11]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.

Yang and Hodges [12] used sense disambiguation method to construct feature vector for document representation. In this system, words are first mapped to word senses using a semantic relatedness based word sense disambiguation algorithm. Then these senses are used to construct the feature vector to represent the documents. Two different sense representation methods, namely, senseno and offset, are used. For clustering four algorithms are applied namely K-Means, Buckshot, HAC and bisecting k-means.

Meshrif et.al [13] have developed a system that utilizes Kohonen's Self Organizing Map to cluster Arabic textual documents containing information about different types of crimes. The system used the correlation or dependency relationship between some intransitive verbs and some prepositions to recognize the type of crime.

III. Generation of document vectors

Vector space model with bag of words is the simple and common way of representing a text. In this method words are used as the component of the document vector. The limitations of this approach is that it cannot capture ambiguity, synonymy, semantic relations between words. For example suppose a sentence, "A bear can bear very cold temperatures", the word bear has frequency count of 2. In the above stated example sentence the word "bear" has two different senses, viz., first bear means an animal and meaning of second bear is tolerate. Hence, the problem with VSM using bag of words approach is that it finds the frequency count of a word whereas for a proper generation of document vectors the frequency count of the senses of a word is required. In this work rather than bag of words senses has been used for generation of document vectors. For example consider the sentences, "the fly can fly" and "A bear can bear very cold temperatures" which appear in a document, the vector corresponding to these sentences considering words as component of the vector is shown in the Table 1

Table 1: Document Vector

	Fly	Cold	Bear	Temperature
D1	2	0	0	0
D2	0	1	2	1

If the above example sentences are sense tagged then they will look like "fly_02192818 can fly_01944262" and the A bear_02134305 can bear_00670017 very cold_14168983 temperatures_05018974. Vectors corresponding to these sense tagged sentences where we have considered synset id as the component of the vectors are shown in Table 2

Table 2: Document Vector with senses

	Fly_ 02192818	Fly_ 01944262	Cold_ 14168983	Bear_ 02134305	Bear_ 00670017	Temperature_ 05018974
D1	1	1	0	0	0	0
D2	0	0	1	1	1	1

In the example sentences stated above the word fly and bear have two senses. In the first sentence first 'fly' is an two- winged insect and second 'fly' is travel through the air. In the second sentence first bear is an animal and second bear is tolerate. The term based method ignores the fact that words may be ambiguous and generate vectors by considering only the frequency count of the words. We can see from the example that vector generated by the term based method both the words fly has been considered as a single term and has frequency count of 2.Similarly the word bear has frequency count of 2. But in the vector generated from sense tagged sentences they have find different places in the vector because they have different synset ids. The synset id of first fly is 02192818 and the synset id of second fly is 01944262 and the frequency count of synset id_02192818 and synset id_01944262 is 1. Similarly both bears have find different places in the vector because they have different synset ids. Once the feature vectors are completed in this way, weights are assigned to each word/ synset id across the corpus using TF*IDF method [14], which is the combination of the term/sense frequency (TF), and the inverse document frequency (IDF). Terms which are not present in the WordNet are not considered as element of the feature vector.

Table 3, 4,5 and 6 presents some sample Napali text documents and their corresponding sense tagged Nepali text documents.

Table 3 Sample Napali text document1

वर्तमान युगमा टेक्नोकल्चरको व्यापकता बढिरहेछ। यस व्यापक विश्व बजारमा प्रतिदिन बढिरहेको तक्निकी र भौतिक संसाधनले सबैलाई अधीनस्थ गरेको छ।

Table 4. Sample Napali text document2

हेलिकप्टरमा त बसमा जस्तो भिड भइहाल्दो रैनछ नि त परैबड (बाट) लइन (लाइन) लगाएर लाँदो रैछ’।उनीहरूले जीवनकालमा सधैँ बस भिड भएको देखेका छन्। बसमा चढ्नुभन्दा अगाडि पैसा उठाउने एकदेखि दुई घण्टा पर्खाउने।

Table 5. Sample sense tagged Napali text document1

वर्तमान_12784 युग_128346मा टेक्नोकल्चरको व्यापकता_13365 बढिरहेछ_210794। यस व्यापक_4895 विश्व बजार_15303मा प्रतिदिन_36459 बढिरहेको तक्निकी र भौतिक_42443 संसाधन_110146ले सबै_46962लाई अधीनस्थ_415356 गरेको छ_210794।

Table 6. Sample sense tagged Napali text document2

हेलिकप्टर_117822मा त बस_19610मा जस्तो_45655 भिड_15499 भइहाल्दो रैनछ_210794 नि त परैबड (बाट) लइन (लाइन_11878) लगाएर लाँदो रैछ’। उनीहरूले जीवनकाल_13665मा सधैँ_37161 बस_19610 भिड_15499 भएको_41469 देखेका छन्। बस_19610मा चढ्नुभन्दा अगाडि_38219 पैसा_1859 उठाउने एक_12929देखि दुई_49167 घण्टा_15014 पर्खाउने।

IV. Hybrid SOM+PSO+K-MEANS Clustering Algorithm

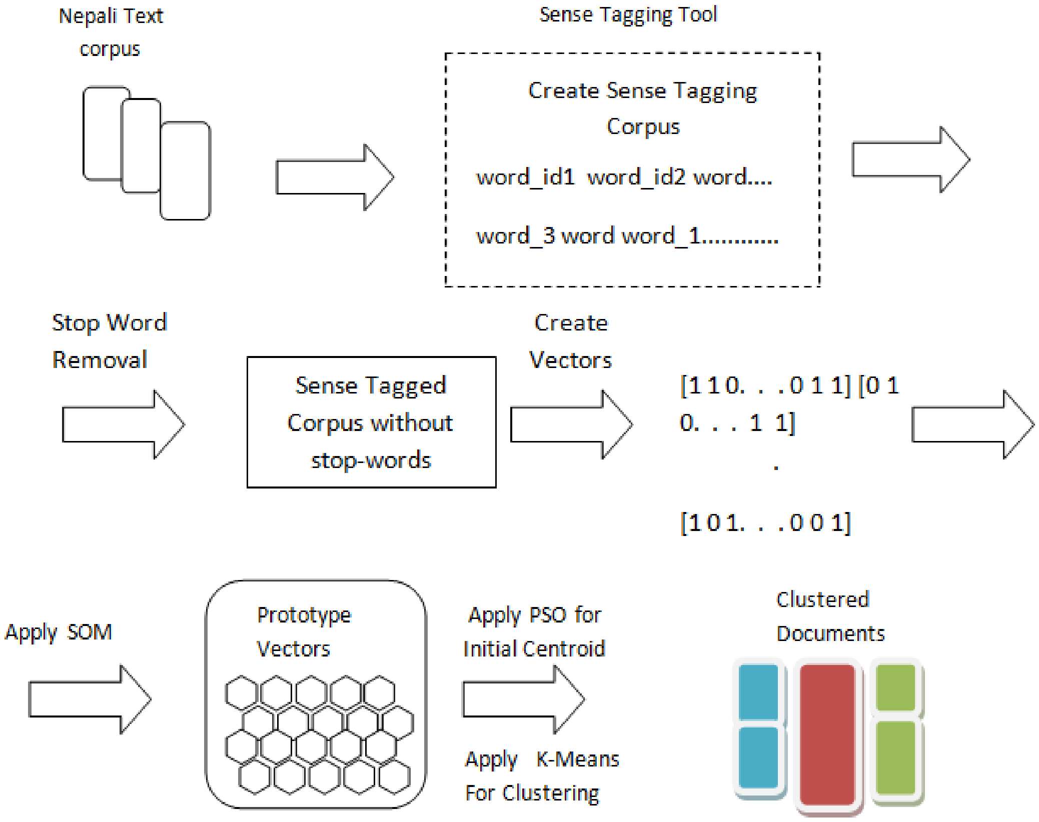
The proposed hybrid SOM+PSO+k-Means clustering algorithm consists of three phases as described below:

Phase I: In this phase Self Organizing Map Algorithm is applied on the input vectors to produce the prototype vectors which can be interpreted as “protoclusters. The number of prototype vectors are much larger than the expected number of clusters. The prototypes are grouped in the next step to form the actual clusters.

Phase II: In this phase PSO algorithm is executed on the prototypes vectors to find clusters' centroid for K-menas algorithm.

Phase III: The cluster centres obtained in Phase II are used in phase III. K-Means algorithm is applied in this phase to generate the actual clusters . The result from PSO which is found in phase II is used as the initial seed of the K-means algorithm. Phase III converges quickly when the centroids from Phase II are used.

The overall process of clustering is described in figure 1.



Nepali Text corpus
Sense Tagging Tool
Create Sense Tagging Corpus
word_id1 word_id2 word....
word_3 word word_1...........
Stop Word Removal
Sense Tagged Corpus without stop-words
Create Vectors
[1 1 0. . .0 1 1] [0 1 0. . . 1 1]
.
[1 0 1. . .0 0 1]
Apply SOM
Prototype Vectors
Apply PSO for Initial Centroid
Apply K-Means For Clustering
Clustered Documents

Fig.1 Overall Process of Clustering

The complete algorithm is now presented next.

Phase I Self Organizing Map

1). Each node's weights are initialized.

2). A vector is chosen at random from the set of training data and presented to the network.

3). Every node in the network is examined to calculate which ones' weights are most like the input vector. The winning node is commonly known as the Best Matching Unit (BMU).

$$c = \arg\min_i \|x - w_i\| , \quad for\ i = 1, 2, \cdots, n$$

n is the total number of grid neurons.

4). The radius of the neighborhood of the BMU is calculated. This value starts large. Typically it is set to be the radius of the network, diminishing each time-step.

5). Any nodes found within the radius of the BMU, are adjusted to make them more like the input vector. The closer a node is to the BMU, the more its weights are altered.

$$w_i(t+1) = w_i(t) + \alpha(t)\ h_{ci}\ (t)[x(t) - w_i(t)]$$

where

t time;

$\alpha(t)$ adaptation coefficient;

$h_{ci}(t)$ neighborhood kernel centered on the winner unit:

$$h_{ci}(t) = exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)$$

where r_c and r_i are positions of neurons c and i on the SOM grid. Both $\alpha(t)$ and $\sigma(t)$ decrease monotonically with time.
6). Repeat step 2 for N iterations.

Phase II Particle Swarm Optimization

```
(1) Initialize each particle with K random cluster centroid vectors.
(2) For t= 1 to $t_{max}$ do
a)For each particle i do:
b)For each data vector $m_p$ do
 (i) Calculate the distance d $(m_p,o_{ij})$, to all cluster centroids $C_{ij}$
 (ii) Assign each data vector to the closest centroid vector.
 (iii) Calculate the fitness value based on equation (2).
c) Update the global best and local best positions
d) Update the cluster centroids using equations (3) and (4)
```

$$v_{id}(t+1)=\omega.v_{id}(t)+c_1.rand().(p_{id}-x_{id})+c_2.rand().(p_{gd}-x_{gd}) \quad (3)$$

$$x_{id}(t+1)=v_{id}(t+1)+x_{id}(t) \quad (4)$$

Where t_{max} is the maximum number of iterations.

Phase III K-means Algorithm

15. Inheriting cluster centroid vectors from Phase II.
16. Assigning each prototype vector to the closest cluster centroids.
17. Update the cluster centers in each cluster using equation (5)

$$C_j = \frac{1}{n_j}\sum_{\forall d_j \in S_j} d_j \quad (5)$$

where d_j denotes the document vectors that belong to cluster S_j; c_j stands for the centroid vector; n_j is the number of document vectors belong to cluster S_j.
18. Repeat the steps 16 and 17 until there are no more changes in the values of the centroids.

V. Experimental Results

To demonstrate the performance of the proposed hybrid SOM+PSO + K-means clustering algorithm an evaluation study was carried out. Comparison with other clustering algorithms viz. K-means, SOM + K-means and PSO + K-means was conducted on sense tagged Nepali text corpus and Nepali text corpus without sense tagging . Nepali document dataset has been collected from Technology Development for Indian Language website [15].The corpus in Nepali language provides data from different domains such as literature, science, media, art etc. Sense tagging of the whole document corpus was done using sense marking tool. In this study the quality of the clustering is measured using Intra-cluster similarities and Inter-cluster similarity. A large average intracluster value indicates that the vectors within a category are grouped tightly together. A small average intercluster value indicates that the individual clusters are far apart. The results obtained are shown in table 7.

Table 7 Performance Comparison Of K-Means, Hybrid Pso+K-Means, Hybrid Som+K-Means And Hybrid Som+Pso+K-Means Algorithms

No. of documents	Dimension	Method	Clustering Algorithm	Intracluster	Intercluster
100	3174	Word based	K-Means	3.6403	181.8389
			PSO+Kmeans	3.6657	77.7284
			SOM+K-means	1.6121	.6507
			SOM+PSO+K-Means	1.6033	.6507
	2435	Sense Based	K-Means	4.1226	112.5602
			PSO-Kmeans	4.0470	77.4100
			SOM-K-means	2.3611	1.0084
			SOM+PSO+K-Means	2.3682	1.0293

The bar graph of the comparison of the parameters is plotted for the different algorithm (k-means, hybrid PSO+K-means, SOM+K-Means and hybrid SOM+PSO+K-means").

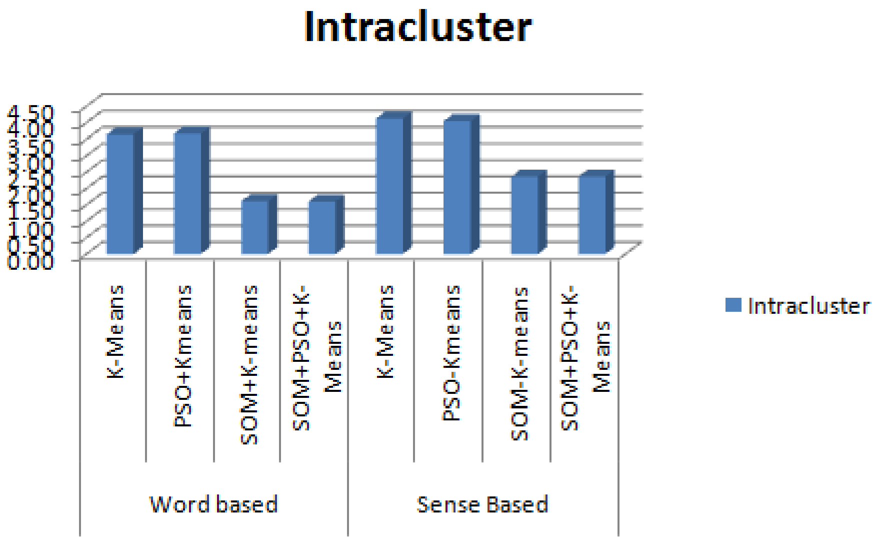


Fig2: Intra cluster similarity using various algorithms

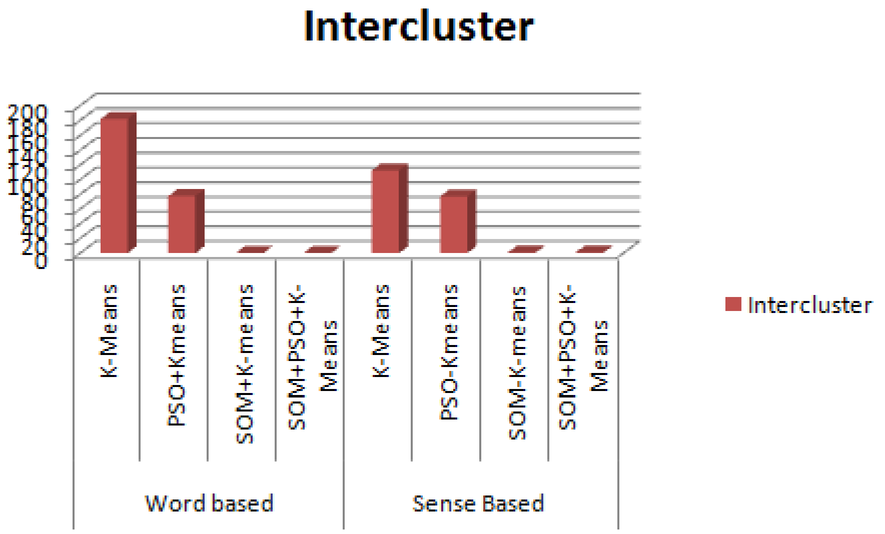


Fig3: Inter cluster similarity using various algorithms

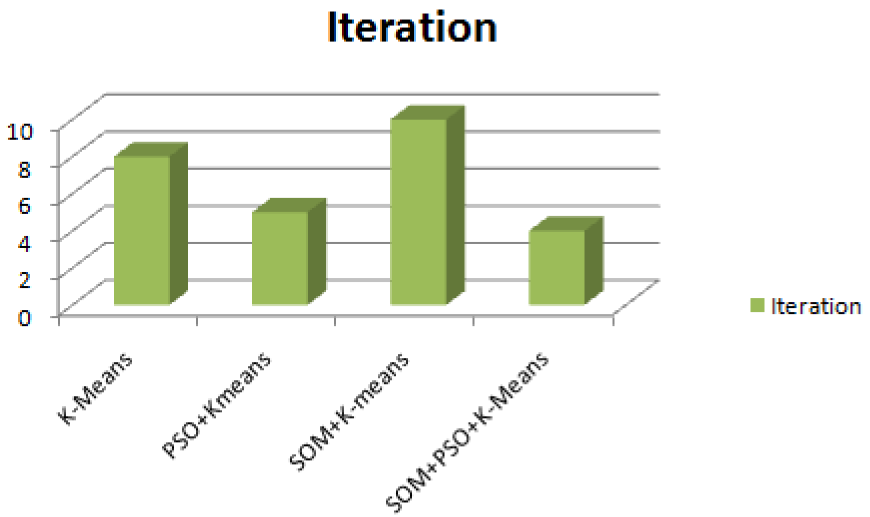


Fig 4: Number of iteration

From the experimental results it is observed that hybrid SOM+PSO+K-means algorithm generates better result compared to K-means, hybrid PSO+K-means, SOM+K-means in terms of both intercluster similarity and number of iterations. Table 1 shows that hybrid SOM+PSO+K-means algorithm with sense based performs better than the hybrid PSO clustering algorithm with bag of words representation.

VI. Conclusion

In this paper a hybrid SOM+PSO+K-means algorithm is presented for clustering of document dataset. The algorithm is executed on vectors generated from sense tagged Nepali document dataset where synset ids are considered as component of the vectors. By sense tagging exact senses of words in the context are tagged with words as synset id is unique for each word for each sense. In the experiments K-means and PSO+K-means were executed both on direct document vector and on prototype vectors generated by SOM. Experimental results indicate that using hybrid algorithm that is execution of PSO+K-means on prototype vectors can produce compact clustering result compared to K-means, PSO+K-means and SOM+K-means.

References

[1]. A.P. Engebrecht, Computational Intelligence An Introduction (John Wiley & Sons, Ltd, 2007)

[2]. S. Urooj, S. Shams, S. Hussain, F. Adeeba, Sense Tagged CLE Urdu Digest Corpus, Proc. Conf. on Language and Technology, Karachi, 2014

[3]. T. Kohonen, : Self-organized formation of topologically correct feature maps. Biol.Cybern. 43 59–69 (1982)

[4]. J. Kennedy and R.C. Eberhart, “Particle Swarm Optimization,” Proc. IEEE, International Conference on Neural Networks.

[5]. Piscataway. Vol. 4, pp 1942-1948,1995

[6]. S.C. Satapathy, N. VSSV P B. Rao, JVR. Murthy, R. P.V.G.D. Prasad, “A Comparative Analysis of Unsupervised K-means, PSO and Self- Organizing PSO for Image Clustering,”Proc. International Conference on Computational Intelligence and Multimedia Applications,2007.

[7]. A. Roy, S. Sarkar, B. S. Purkayastha, “A Proposed Nepali Synset Entry and Extraction Tool,” Proc. 6th Global wordnet conference, Matsue,Japan,2012.

[8]. X. Cui, T.E. Potok, Document Clustering Analysis Based on Hybrid PSO+ K-Means Algorithm, Journal of Computer Sciences (Special Issue): 27-33, 2005 ISSN 1549-3636

[9]. H. Lo, C.C. Lin, R-J. Fang, C. Lee, and Y-C. Weng, Chinese Document Clustering Using Self Organizing Map Based on Botanical Document Warehouse, proceedings European Computing Conference, Lecture Notes in Electrical Engineering 28.

[10]. U.K. Sridevi and N. Nagaveni, Semantically Enhanced Document Clustering Based on PSO Algorithm, European Journal of Scientific Research, ISSN 1450-216X Vol.57 No.3, pp.485-493,2011

[11]. T.F Gharib, M.M Fouad, A. Mashat, I. Bidawi1, Self Organizing Map based Document Clustering Using WordNet Ontologie. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 2, 2012

[12]. L. Xinwu, Research on Text Clustering Algorithm Based on K_means and SOM International Symposium on Intelligent Information Technology Application Workshops, 2008.13

[13]. Y. Wang, J. Hodges, Document clustering with semantic analysis, Proc. 39th Annual hawaii International Conference on System Sciences, HICSS, Vol. 03, pp.54.3, 2006

[14]. M. Alruily, A. Ayesh, A. Al-Marghilani, Using Self Organizing Map to Cluster Arabic Crime Documents, Proc. International Multiconference on Computer Science and Information Technology, pp. 357–363, ISBN 978-83-60810-27-9 ISSN 1896-7094

[15]. G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18:613–620.

[16]. http://tdil-dc.in.