ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

A Hybrid Approach for Improving Text Document Clustering

[1] Mrs. R.Janani, [2] Dr. S.Vijayarani
[1] PhD Research Scholar, [2] Assistant Professor
[1][2] Dept. of CSE, Bharathiar University, Coimbatore.

Abstract;- **Text document clustering is the process of distributing documents into similar groups called clusters, in order that documents within a cluster have great affinity in comparison to other documents in different clusters. It has been taken into consideration intensified due to the fact of its substantial applicability in various areas like information retrieval, web mining and search engines like Google. It is determining the similarity among documents and based on the similarity it will organize the documents together. Fast and greatness of text document clustering algorithms perform a vital role in dramatically navigating, encapsulating, and establishing the information. The clustering algorithms can only generate the optimal solution. A global optimal solution can be attained by applying high-geared and high-quality optimization. The main objective of this research work is to grouping the documents based on their contents and also to improve the cluster accuracy based on the content of the documents. In order to perform this task this research work uses the existing algorithms DBSCAN and PSO. A hybrid algorithm which is a combination of PSO and DBSCAN algorithm is also proposed. The outcome of this research work is the identification of cluster of documents which has the same contents.**

Keywords: Document Clustering, preprocessing, DBSCAN, PSO, Hybrid Algorithm

I. INTRODUCTION

Document clustering is the subset of data clustering, which encompasses the perceptions from the fields of information retrieval, natural language processing and machine learning. It systematizes the collection of documents into one kind of groups, called as clusters, anywhere the documents in each cluster has mutual properties allowing to distinct similarity measure. The high quality document clustering algorithms play a major role to effectively handle, review, and organize the documents. Clustering can achieve moreover disjoint or overlapping partitions. In an overlapping partition, it is feasible for a document to seem in numerous clusters. In disjoint clustering, each document seems in precisely only one cluster [1]. Document clustering may be divided into two most important subcategories, hard clustering and soft clustering. Hard clustering calculates the hard assignment of the particular document to a cluster. Soft clustering is divided into partitioning, hierarchical and frequent item set based clustering [1]. These clustering algorithms can only give the optimal solution to those clustering problems. So, the optimization algorithms are used to acquire the excessive high-quality outcomes.
Optimization algorithms may be both deterministic and stochastic in essentiality. Prior techniques to solve optimization issues necessitate massive computational determinations, which generally tend to fail because the problem size increases. This is the inspiration for retaining bio inspired stochastic optimization algorithms as computationally efficient replacements to deterministic method [2]. Recently, the interest in the application of nature inspired algorithms has grown due to various motives which include the generation of population of solutions and explicit memory of previously visited solutions.

This paper is organized as follows, section II explains the related work and section III presents the methodology of this research work. Experimental results are given in Section IV and section V describes the conclusion of this research work.

II. RELATED WORKS

In [3] authors have presented LF-DBSCAN algorithm to accomplish the actual un-uniform data set clustering. The comprehensive constraint for the traditional DBSCAN algorithm specifies to poor cluster multi-density dataset, as well as the high-dimensional data processing result is not reasonable enough. This paper presents a LF-DBSCAN algorithm to attain effective non-uniform data set clustering. But LF-DBSCAN algorithm quiet it has more for enhancements, such as the need to subjectively regulate the constraints MinPts and k. It was found that the value k impacts the results of clustering effects, how to further diminish the influence of k is the next issue to be studied.

In [4] they presented the AntClass algorithm is a new algorithm applying ant colony clustering algorithm to analyze the clusters. To attack the deliberate speed of the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

AntClass algorithm, a new algorithm named DBAntCluster was proposed. Firstly, the extraordinary density clusters is gotten in the dataset by using DBSCAN algorithm, and then these high density clusters are distributed in the grid panel as a superior kind of data object with other single data objects in the dataset.
In [5] their study, a text document clustering algorithm established by using PSO algorithm. In the PSO clustering algorithm, the clustering effectiveness can be categorized into two phases: the global searching phase and the local filtering phase. The global searching phase assurances, each particle examines extensively enough to cover the entire problem space. The refining phase creates all particles converge to the ultimate when a particle extents the locality of the optimal solution. For a huge volume of dataset, conservative PSO can conduct a globalized examining for the optimal clustering, but entails the number of iteration and computation than the K-means algorithm does.

In [6] research clustering is done for optimization of Kmeans clustering algorithm using PBO algorithm. It gives better result than the previous algorithm and this can be concluded on the bases of some parameter like recall, precision and f-measure. This can give better results for the conceptual clustering. Experiments proved that our algorithm greatly improves the effect of text clustering, and then provides a more forceful and better support for efficient text retrieval.

In [7] they have presented a hybrid evolutionary optimization algorithm to resolve the clustering problems. The algorithm is established on a permutation of the ant colony optimization and the simulated galvanizing. In their proposed algorithm, they have used the simulated galvanizing algorithm as a local inquirer for each and every colony. To estimate the effectiveness of the hybrid algorithm, it is correlated to other stochastic algorithms, the standard ACO, SA and k-means algorithms on numerous distinguished real life data sets. The investigational results indicated that the proposed optimization algorithm is similar to the other algorithms in terms of function metrics and standard abnormalities.

III. METHODOLOGY

The main objective of this research work is to improve the cluster accuracy based on the content of the documents. In order to perform this task this research work uses DBSCAN, PSO and hybrid algorithms. The performance factors are cluster purity, Precision, Recall and F-measure. Figure 1 shows the architecture of this research work.

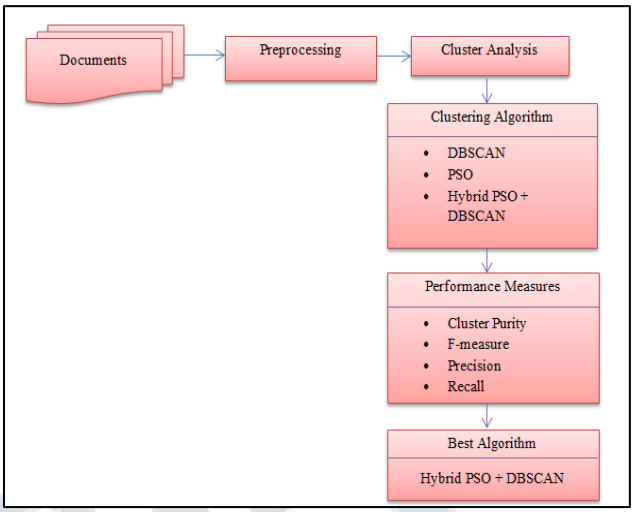

Figure. 1: System Architecture

A. Dataset

The data sets used in this research are Reuter's dataset, 20 newsgroup dataset and TREC Text Retrieval Conference dataset. These have been broadly used for evaluating feature selection techniques, classification and clustering. In order to make certain diversity, the datasets are from different sources, some containing newspaper articles, some containing newsgroup posts and the remaining being academic papers. The detailed summary of the dataset is given in Table 1.

B. Preprocessing

Text preprocessing is a significant task in text mining, information retrieval (IR) and Natural Language Processing (NLP). In text mining, text preprocessing is used for extracting non-trivial knowledge from unstructured text data. Before clustering the documents, the preprocessing techniques have been applied to the specific dataset to reduce the size of the dataset. These processes will increase the efficiency of the text based document clustering system. In this research work, stemming, stop word removal, numbers and punctuation removal techniques are used.

TABLE 1: SUMMARY OF DATASET

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

Dataset	Source	Number of Documents	Number of classes	Number of Words
re0	Reuters	1504	13	11465
re1	Reuters	1657	25	3758
20news	20news group	18828	20	28553
tr41	TREC	878	10	7454
la1	TREC	3204	6	31472

C. DBSCAN

DBSCAN is the leading density based clustering algorithm. This algorithm was offered by Ester et al. in 1996, and it was intended to cluster data of arbitrary shapes in the occurrence of noise in spatial and non-spatial high dimensional databases [3]. The DBSCAN (Density-based spatial clustering of applications with noise) algorithm can be used to recognize the clusters in large spatial datasets by perceiving at the local density of database features, using one input parameter. Moreover, the users get an impression on which parameter value that would be suitable. Accordingly, minimal knowledge of the domain is crucial.

The DBSCAN can also regulate what information would be categorized as noise or outliers. In spite of this, its working method is fast and scales very well with the size of the database almost linearly. By using the density circulation of nodes in the database, DBSCAN can group these nodes into distinct clusters that describe the different classes. DBSCAN can find clusters of arbitrary shape [4]. But, clusters that lie close to each other tend to fit into the identical class. The important parameters are ε (eps) and the minimum number of points essential to custom a dense region (P_m).

Algorithm 1: DBSCAN algorithm

Input: C the number of clusters, N is Neighbors and RQ is Range query
DBS(D, ε, P_m)
Step 1: Initialize C = 0
Step 2: for each unvisited point P in dataset D
Step 3: mark P as visited
Step 4: N = regionQuery (P, ε)
Step 5: if sizeof (N) < P_m
Step 6: mark P as NOISE
Step 7: else
Step 8: C = next cluster
Step 9: RQ (P, N, C, ε, P_m)

Function RQ (P, N, C, ε, P_m)

Step 1: Add P to cluster C
Step 2: for each point P in N
Step 3: if P is not visited
Step 4: Then mark P as visited
Step 5: N = regionQuery (P, ε)
Step 6: if sizeof (N) >= P_m
Step 7: N = N joined with nearest neighbor N_1
Step 8: if P is not yet member of any cluster
Step 9: add P to cluster C

D. Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is a computational intelligence oriented, speculative, population - based global optimization procedure suggested by Kennedy and Eberhart in 1995 [8]. The main objective of this algorithm is to find out the appropriate centroids of clusters. In this algorithm the high dimensional documents are modeled as the global problem space using document vector space model. The document which contains the word represents the unique dimensional of the global problem space.

In PSO, the term "particles" symbolizes to population participants, which are less capacity and are subject to velocities and accelerations on the way to a superior behavior mode [9]. Every particle in the swarm infers an optimal solution for document clustering problem. Respectively every particle will maintain a matrix M_i = $\{C_1,C_2\ldots.C_i\,C_k\}$ where C_i denotes the i^{th} cluster vector and k denotes the number of cluster [10, 11]. Each and every iteration the particle updates its position and velocity as Equation 1,

$$p_{j+1}^{i} = p_j^i + v_{j+1}^i$$
$$v_{j+1}^{i} = v_j^i + s_1 r_1 (bp_j^i - p_j^i) + s_2 r_2 (bp_j^i - p_j^i) \quad (1)$$

where, p_j^i represents Particle position, v_j^i represents Particle velocity, bp_j^i represents Best position, s_1,s_2 represents cognitive and social parameters and r_1, r_2 are random numbers between 0 and 1

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

Algorithm 2: PSO algorithm

Step 1: Initialize the population
Step 2: For I = 1 to population size N
Step 3: Estimate the fitness function F for all particles.
Step 4: If F is better than the best one pbest
Then current value = new pbest.
Step 5: End for
Step 6: Select the particle with pbest of the entire particle as a gbest
Step 7: For I = 1 to population size N
Step 8: Calculate the velocity and update the position of all the particles using (1)
Step 9: Repeat steps 2 to 5 till a stopping criterion is encountered.

E. Proposed Hybrid Algorithm

In the hybrid algorithm which consist two important implementation; PSO and DBSCAN. The hybrid algorithm first implements the PSO algorithm to discover the permanency factor which is close to the optimal solution by global search and concurrently evade high computation time [12]. In this case PSO clustering is dismissed as soon as the maximum number of iterations is exceeded. The result of the PSO algorithm is then used as preliminary density based vectors in the DBSCAN algorithm [13].

The DBSCAN algorithm is then accomplished until the determined number of iterations is extended. This algorithm inclines to converge faster than the PSO, but frequently with a less accurate clustering [14, 15] and PSO can find the behavior of globalized searching for the optimal clustering, but necessitates additional iterations and calculation than the DBSCAN algorithm. The hybrid PSO algorithm combines the advantage of both the algorithms: globalized searching of the PSO algorithm yet the quick merging concerning the DBSCAN algorithm.
This hybrid algorithm reduces the computational cost when compared to the existing algorithm. This approach met the optimal solution within a few generations. It was also perceived that the basic clustering established selection algorithm assured the selection of cluster with the ideal solution in each generation.

Algorithm 3: DBPSO algorithm

Input : C the number of clusters, N is Neighbors and RQ is Range query
DBS(D, ε, P_m)
Step 1: Initialize C = 0 and population for optimization
Step 2: For I = 1 to population size N
Step 3: for each unvisited point P in dataset D
Step 3: mark P as visited
Step 4: N = regionQuery (P, ε)
Step 5: if sizeof (N) < P_m
Step 6: mark P as NOISE
Step 3: Estimate the fitness function F for all particles
Step 4: If > pbest then
Current value = new pbest.
Step 5: End for
Step 7: else
Step 3: Select the particle with pbest of the entire particle as a gbest
Step 8: Calculate the velocity and update the position of all the particles using (1)
Step 8: Assign C = next cluster
Step 9: Returnε, P_m and clusters

IV. RESULTS AND DISCUSSION

In the experiments, we realistic the PSO, DBSCAN and hybrid clustering algorithm on five different document datasets. The number of documents in the dataset varies from 878 to 18828. The final result demonstrates that the hybrid algorithm has produced the better clustering results than the PSO and DBSCAN algorithms.

In order to perform this task, the performance factors are cluster purity, Precision, Recall and F-measure [16, 17].

Purity: Higher value of cluster purity indicates better cluster predictive discrimination.

Precision: It is considered as the fraction of pairs properly situate in the same cluster.

Recall: It is computed as the part of actual pairs that were recognized.

F-measure: It is the harmonic mean of precision and recall.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

Table 2 illustrates the total number of documents are grouped into two clusters. Figure 2 shows the number of documents within the cluster1 and cluster 2.

Table 2: Number of Documents within the Clusters

Dataset	Cluster 1	Cluster 2
re0	370	1134
re1	813	844
20news	6276	12552
tr41	439	439
la1	1068	2136

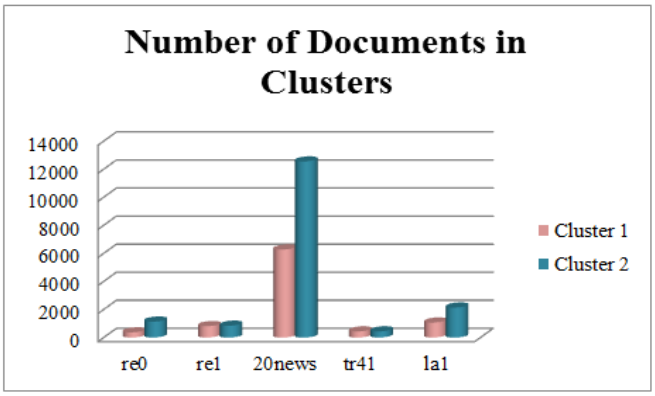

Figure. 2. Number of Documents within the clusters

Table 3 describes the cluster purity of all the five datasets. Figure 3 explains the purity for three algorithms.

Table 3: Cluster Purity

Dataset	DBSCAN	PSO	PSO + DBSCAN
re0	0.31	0.47	0.61
re1	0.59	0.61	0.75
20news	0.42	0.50	0.85
tr41	0.36	0.52	0.79
la1	0.44	0.57	0.74

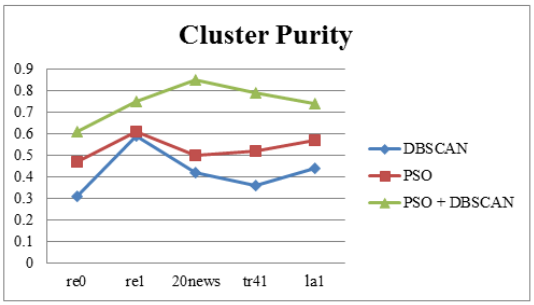

Figure. 3. Cluster Purity

Table 4 describes the precision of all the five datasets. Figure 4 explains the precision for three algorithms.

Table 4: Precision

Dataset	DBSCAN	PSO	PSO + DBSCAN
re0	23.78	32.18	41.87
re1	14.89	25.89	32.78
20news	26.78	39.47	48.01
tr41	26.48	39.87	47.89
la1	39.78	48.47	54.48

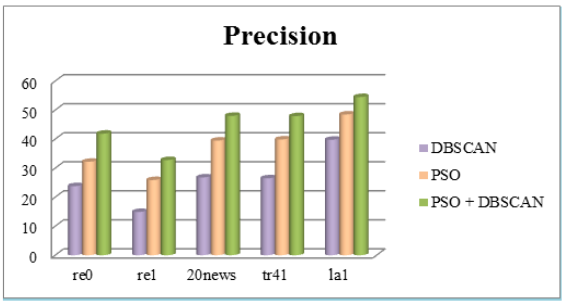

Figure 4: Precision

Table 5 describes the recall of all the five datasets. Figure 5 explains the recall for three algorithms.

Table 5: Recall

Dataset	DBSCAN	PSO	PSO + DBSCAN
re0	21.32	32.18	41.87
re1	36.12	48.21	50.12
20news	24.15	49.12	58.47
tr41	18.36	19.12	27.15
la1	20.46	28.01	36.09

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 12, December 2017

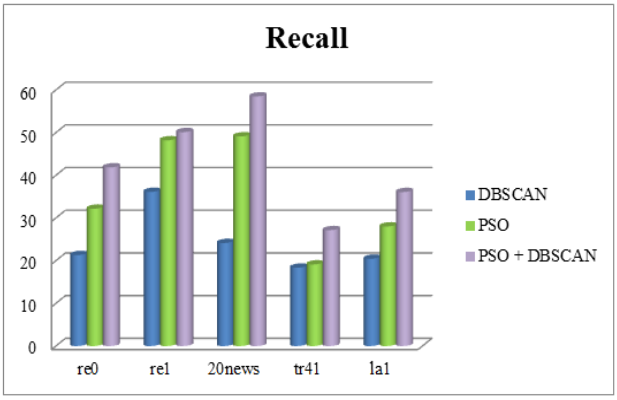

Figure 5: Recall

Table 6 describes the F-Measure of all the five datasets. Figure 6 explains the F-Measure for three algorithms.

Table 6: F-Measure

Dataset	DBSCAN	PSO	PSO + DBSCAN
re0	30.74	40.78	49.12
re1	46.01	46.32	49.58
20news	41.25	49.35	52.14
tr41	52.14	59.38	67.15
la1	52.17	53.17	59.78

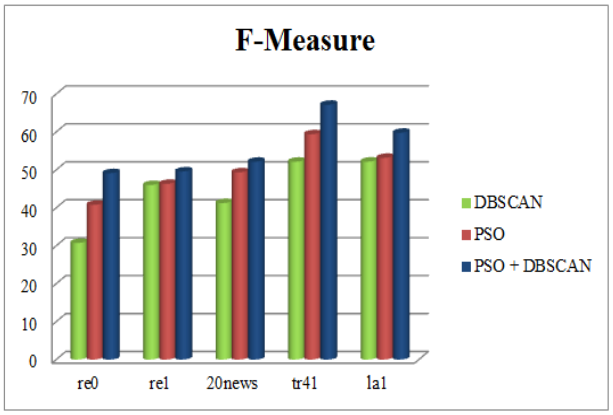

Figure. 6. F-Measure

Table 7 describes the sample clustered documents with its related topics. In this we use 20newsgroup dataset for the sample clustered documents.

Table 7: Sample Clustered Documents of 20newsgroup

Cluster results: Keywords and sample documents	Related topic
Capacity, Voltage, Socket, Wire 1. The heat capacity of air is less than that of concrete, dirt, or wood, so it heats faster. 2. Then wipe the socket instead of spaying the stuff directly from the can.	Electronics
Cycle, Motor, Light, Tire, Machine 1. The scary bit about this is the non-availability of rear lights at all. 2. I was wondering why there was all the pointless waffle about motorcycles.	Motor Cycle
Decoration, Audio, Video, Color 1. My friend is a interior decor designer. 2. How do you access the extra video memory on a video board.	Graphics

V. CONCLUSION

Document clustering is an essential process used in information retrieval, efficient document organization and automatic topic extraction. In this research work, it analyses the performance measures of existing and hybrid clustering algorithm for text clustering. From this analysis, the hybrid clustering algorithm gives the best accuracy for this benchmark data set. The evolutionary algorithms are between the greatest dominant algorithms for optimization which is standard to have a broad impact on forthcoming generation computing. In future, the most efficient algorithms have to be developed for clustering all types of documents.

REFERENCES

[1] Nicholas O. Andrews, Edward A. Fox, "Recent Developments in Document Clustering", October 16, 2007.

[2] S.J Nanda, G. Panda, "A Survey on nature inspired meta heuristic algorithm for partition clustering" Swarm

and Evolutionary Computation, Elsevier, Vol. 16, pp. 1-18, 2014.

[3] Fang Yuankang, Huang Zhiqiu, Luo Yuping, Ye Zan and Liu Ying "Research on Improve DBSCAN Algorithm Based On Ant Clustering" The Open Automation and Control Systems Journal, 2014, 6, 1076-1084

[4] Shang L., et al., "A New Ant Colony Algorithm based on DBSCAN, Proceedings of 2004 International Conference on Machine Learning and Cybernetics, pp. 1491 – 1496, 2004.

[5] Xiaohui Cui, Thomas E. Potok, Paul Palathingal "Document Clustering using Particle Swarm Optimization" Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE

[6] Manpreet KaurÀ and Navpreet KaurÀ "Text Clustering using PBO algorithm for Analysis and Optimization" International Journal of Current Engineering and Technology, E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

[7].Taher, N., et al., An Efficient Hybrid Evolutionary Optimization Algorithm based on PSO and SA for Clustering, Journal of Shejiang University – Science A, Vol. 10, No. 4, pp. 512 – 519, 2009.

[8] Ananthy Christy, Perianayagam Ajay-D-Vimal Raj, Sanjeevikumar Padmanaban, Rajasekar Selvamuthukumaran, Ahmet H.Ertas, "A bio-inspired novel optimization technique for reactive power flow", Engineering Science and Technology an International Journal, Vol. 19, pp. 1682-1692, 2016.

[9] K.Aparana and Mydhili K.Nair, "Enhancement of k-means algorithm using ACO as optimization technique on high dimensional data" 2014 international conference on Electronics and Communication Systems (ICECS) IEEE, pp. 1-5, 2014.

[10] A.M. Aibinu, H.Bello Salau, Najeeb Arthur Rahman, M.N. Nwohu, C.M. Akachukwu, "A novel Clustering based Genetic Algorithm for route optimization", Engineering Science and Technology an International Journal, Vol. 19, pp. 2022– 2034, 2016.

[11] Cui, X. et al., Document Clustering using Particle Swarm Optimization, Proceedings of IEEE Swarm Intelligence Symposium, pp. 185 – 191, 2005.

[12] Hamed Nikbakht, Hamid Mirvaziri, "A new clustering approach based on K-means and Krill Herd algorithm" 23rd Iranian Conference on Electrical Engineering, IEEE, 2015

[13] Chen, et al., HDACC: A New Heuristic Density based Ant Colony Clustering Algorithm, Proceedings of IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 397 – 400, 2004.

[14] Wu, B., et al., CSIM: a Document Clustering Algorithm based on Swarm Intelligence, Proceedings of 2002 Congress on Evolutionary Computation, pp. 477 – 482, 2002.

[15] Upeka Premaratne , Jagath Samarabandu, and Tarlochan Sidhu, ―A New Biologically Inspired Optimization Algorithm‖,Fourth International Conference on Industrial and Information Systems, ICIIS 2009,28-31 December 2009, Sri Lanka.

[16] Henal Parmar, Bhailal Limbasiya,"A Review on Genetic Algorithm-based Text Clustering Technique "International Journal of Advance Research in Computer Science and Management Studies , Volume 3, Issue 2, February 2015

[17] Sunita Sarkar , Arindam Roy and B. S. Purkayastha "A Comparative Analysis of Particle Swarm Optimization and K-means Algorithm For Text Clustering Using Nepali Wordnet" International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014