A Particle Swarm Optimization based fuzzy c means approach for efficient web document clustering

P Jaganathan [#1] , S Jaiganesh [#2]

[#] Department of Computer Applications,
PSNA College of Engineering and Technology,
Dindigul, India.
[1] jaganathodc@gmail.com
[2] jaiganesh.dgl@gmail.com

Abstract— **There is a need to organize a large set of documents into categories through clustering so as to facilitate searching and finding the relevant information on the web with large number of documents becomes easier and quicker. Hence we need more efficient clustering algorithms for organizing documents. Clustering on large text dataset can be effectively done using partitional clustering algorithms. The Fuzzy C-means algorithm is the most suitable partitional clustering approach for handling large dataset with respect to execution time. This paper introduces a new Hybrid Particle Swarm Optimization method that combines the best features of PSO and fuzzy C-means algorithms for efficient web document clustering. We have tested this hybrid PSO algorithm on various text document collections. The document range varies from 512 to 1639 in the dataset and the terms ranges from 12367 to 19851. Based on the experimental results our proposed PSOFCM approach performs better clustering than other method.**

Keyword- Document clustering, PSO, Partitional clustering, Vector Space Model, Fuzzy C-means

I. Introduction

Clustering is a popular technique in data mining deals with the process of grouping a set of objects into clusters so that objects within the same cluster are similar to each other but are dissimilar to objects in other clusters [21]. Document clustering is an important operation that helps in effective organization of documents so as to enable to retrieve documents very effectively. The clustering techniques are broadly classified into partitioning clustering and hierarchical clustering. The partitioning technique divides the given group of documents into well defined and unique clusters. The hierarchical clustering builds a hierarchy of clusters, showing the relations between individual members and merging clusters of data based on similarity. The partitional clustering technique is the most appropriate for clustering a large document dataset. Fuzzy C-means technique is considered as an appropriate partitioning algorithm for clustering a large data vectors into a well assumed and specified number of clusters [24]. This approach tends to fixate on local optima near the initial cluster centers, which are assigned randomly. Thus, many researchers have presented heuristic clustering algorithms to overcome this problem.

Another clustering method Fuzzy C-Mean (FCM) is a better performing clustering algorithm when compared to K-mean as cluster boundaries are no more hard boundaries, but it is also dependent upon clustering centre initialization [6]. We can obtain better initial clustering centroids using some other techniques and thus it may lead Fuzzy C-means algorithm to perform better in the process of finding the best possible clustering centers by the way of improving the clustering centroid [2]. The Particle Swarm Optimization (PSO) algorithm is a heuristic technique that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality [4, 11, 17]. The better initial cluster centroid can be obtained using PSO. We propose PSO combined with Fuzzy C-means algorithm that produces a quick and better clustering and also it tries to avoid in getting fascinated towards a confined best possible outcome. The experimental outcome shows that the new proposed PSOFCM algorithm can produce the best outcome in minimum number of iterations when compared to Fuzzy C-means and PSO algorithms individually.

The remainder of this paper is arranged as follows: In Section 2, we describe clustering problem, metrics for finding the similarity between documents and representation of documents in clustering algorithms. In Section 3, we describe the functionality of PSO algorithm. The working principle of PSO with Fuzzy C-means algorithm is discussed in Section 4. The experimental setup and outcomes of the PSOFCM algorithm are provided in Section 5. In section 6, conclusion is given.

II. Clustering Problem

Document Clustering is used by the computer to group documents into meaningful groups. In many of the cluster algorithms, the dataset to be clustered is characterized as a set of vectors $X=\{x_1, x_2, \ldots., x_n\}$, where x_i corresponds to an object and is referred as feature vector. Every object is characterized with set of well defined and accurate attributes. Vector Space Model (VSM) is a statistical model for representing text document as vectors [8]. Each dimension corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. The term weight value specifies the importance of the term in a document. In the classic vector space model proposed by Salton, Wong and Yang the term specific weights in the document vectors are products of local and global parameters. The model is known as term frequency-inverse document frequency model [8]. Equation (1) gives the weight of term *i* in document *j*:

$$w_{ji} = tf_{ji} * idf_{ji} = tf_{ji} * log_2(\frac{n}{df_{ji}}) \qquad (1)$$

where tf_{ji} refers the number of presence of term *i* in document *j*; df_{ji} refers the term frequency in the group of documents; and *n* refers the total number of documents in the dataset.

Clustering is a popular approach of automatically finding classes, concepts, or groups of patterns. Thus, a similarity metric has to be defined in any text document analysis. There are two prominent ways to calculate the similarity. The Minkowski distances based formula for finding the similarity between document $\mathrm{m_p}$ and document $\mathrm{m_j}$ is given by equation (2):

$$\mathrm{D_n}(m_p, m_j) = \left(\sum_{i=1}^{\mathrm{d_m}} \left|\mathrm{m_{i,p}} - \mathrm{m_{i,j}}\right|^{\mathrm{n}}\right)^{\frac{1}{n}} \qquad (2)$$

The Minkowski distance is a metric on Euclidean space which can be considered as a generalization of both Euclidean distance and the Manhattan distance. When the value of n is 2, Euclidean distance(ED) is obtained. Here Normalized Euclidean distance is used as the similarity measure for the documents, m_p and m_j. The distance measure is calculated using the equation (3):

$$d(m_p, m_j) = \sqrt{\sum_{k=1}^{d_m} \left(m_{pk} - m_{jk}\right)^2 / d_m} \qquad (3)$$

where m_p and m_j are document vectors; d_m refers the dimension number of the vector space; m_{pk} and m_{jk} refers the weight values in dimension *k* for the documents m_p and m_j. Cosine correlation (CC) is another frequently utilized similarity metric and is given by equation (4):

$$\cos(m_p\,, m_j) = \frac{m_p^t m_j}{|m_p||m_j|} \qquad (4)$$

where $m_p^t m_j$ refers the intersection of the two document vectors.

III. PSO Algorithm

Particle Swarm Optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behaviour of bird flocking. It uses a number of agents (particles) that constitute a swarm moving around in the search space looking for the best solution. A swarm is a disorganized collection (population) of moving individuals that tend to cluster together while each individual seems to be moving in a random direction. It uses a number of elements (particles) that constitute swarm moving around in search space looking for best solution. The direction and velocity of every element moving in the space will be changed with every generation of movement. Both the individual experience of each element, P_{id} and its adjacent element's experience, P_{gd} together manipulate the movement of every element. The $rand_1$ and $rand_2$ are random values which are utilized to make sure that elements travel in the region of the search space prior to converging to the best possible solution. During the search process, the particle successively adjusts its position according to two features, namely their personal best position and the global best position. The c1 and c2 values are used to manage the P_{id} and P_{gd} values. Every particle is updated by P_{id} and P_{gd} during its iterations. The velocity and location of *ith* element changes using to the following equations:

$$v_{id} = w * \ v_{id} + c_1 * rand_1 * (p_{id} - x_{id}) + c_2 * \ rand_2 * (p_{gd} - x_{id}) \qquad (5a)$$

$$x_{id} = x_{id} + v_{id} \qquad (5b)$$

where *w* refers the inertia weight factor**;** p_{id} is the location of the element that realizes the local best value; p_{gd} is the location of the elements that realizes a overall best value; c_1 and c_2 are called as learning factors; *d* refers the dimension of the search domain; $rand_1$, $rand_2$ refers arbitrary values generated between 0 and 1.

The '*w*' gives the essential range to the swarm by altering the momentum of elements to avoid the non movement of elements at the confined solution. Eberhart and Shi [7] proved that the advancement of search effectiveness is possible by steadily diminishing the value of w from a higher value to a lower value. Equation (5a) requires each element to record its current coordinate X_{id} and its velocity V_{id} that specifies the speed of its movement along the dimensions in a problem space and best fitness values are calculated using the coordinates P_{id} and P_{gd} . The fundamental idea of this algorithm is to create a swarm of particles which moves in the problem domain searching for their goal, the place which best suits their needs given by a fitness function and it is calculated using the following equation (6):

$$P_i(t+1) = \begin{cases} P_i(t) & f(X_i(t+1) \leq f(X_i(t)) \\ X(t+1) & f\big(X_i(t+1)\big) > f(X_i(t)) \end{cases} \tag{6}$$

where *f* refers the fitness function; $P_i\ (t)$ denotes the best fitness values and t refers the generation step.

This analysis helps to implement PSO method on the clustering solution. A globalized search can be performed using the PSO clustering algorithm in the whole solution space contrary to Fuzzy C-means method [4, 17]. As document dataset is very large, PSO requires much more iteration to converge to the best possible [23]. The Fuzzy C-means clustering approach is considered as a better proficient algorithm for handling the huge collection of documents particularly with respect to execution time [1]. The Fuzzy C-means approach has a tendency to congregate faster than PSO. This leads to the confusion of selecting an appropriate method for clustering huge text document collection. On these circumstances, we introduce an algorithm combining PSO and Fuzzy C-means called as "PSOFCM" algorithm.

IV. PROPOSED PSOFCM ALGORITHM

In our proposed PSOFCM algorithm, the multi-dimensional document vector space is modeled as a problem space. Every word in the text dataset indicates a dimension in problem domain. Every vector is represented as a dot in the problem domain. In PSO component, the position of the particle is the aggregate of center position of each cluster. This proposed PSOFCM method consists of two components: The PSO component and improved Fuzzy C-means component. During first phase, the PSO component is processed specified number of times to find the local solution. These are fed into Fuzzy C-means component to find better final solution.

A. PSO Component

PSO algorithm seeks to get the best possible solution in the search domain. But still, the search method is not carried out entirely arbitrarily. Each particle maintains a matrix $X_i = (C_1, C_2, \ldots, C_i, .., C_k)$, where C_i represents the *ith* cluster centroid vector and *k* is the cluster number. Each particle's movement is the composition of a velocity and two randomly-weighted influences. The two randomly-weight influences are individuality, or the tendency to return to its best previous position, and sociality, or the tendency to move towards its neighborhood's best previous position. A problem-specific fitness function is employed to determine the next search step. The fitness function is given below:

$$f = \frac{\sum_{i=1}^{N_c}\left\{\frac{\sum_{j=1}^{P_i} d(o_i, m_{ij})}{P_i}\right\}}{N_c} \tag{7}$$

where m_{ij} refers the *jth* document vector of cluster *i*; O_i means centroid vector of *ith* cluster; $d(o_i, m_{ij})$ is the distance between m_{ij} and centroid $O_{i.}$; P_i refers the document number of cluster C_i; N_c refers the cluster number.

The PSO component is:

(1) Every particle arbitrarily selects k number of vectors from the text dataset as the centroid vectors.

(2) Perform the following for every element:

(a) Every vector should be assigned to the nearest centroid vector.

(b) Compute the fitness value using fitness function.

(c) Update equation (5a) and (5b) to produce next solution with the recent values of velocity and particle recent position.

(3) Perform step (2) again and again till any of the following conditions is fulfilled.

(a) The number of iterations performed has reached maximum.

(b) The mean change in centroid vectors is negligible.

(c)

B. *Fuzzy C-means Component*

In fuzzy C-means each cluster is represented by a cluster prototype and the membership degree of a document to each cluster depends on the distance between the document and each cluster prototype. The closest the document is to a cluster prototype, the greater is the membership degree of the document in the cluster. The final outcome of PSO component is fed into Fuzzy C-means as the first input and will keep on carry out the best possible centroids to produce ultimate result. The steps followed in FCM Algorithm are given below:

(1) Initialize the fuzzy partition matrix, U=[uij] matrix, U(0)

(2) At k-step: calculate the centre’ s vectors C (k) = [cj] with U (k).

(3) Update U (k),U (k+1) .

(4) If || U (k+1) - U (k) ||< then STOP; otherwise return to 2.

PSOFCM brings the facility of large-scale finding of the PSO method and the quick congregation of the Fuzzy C-means method. The PSO component is utilized during early phase to facilitate identifying the neighborhood of the best possible result. The outcome of PSO is utilized as the input to the Fuzzy C-means algorithm to generate the best possible solution.

V. EXPERIMENTS AND RESULTS

A. *Datasets*

The datasets of large text document collections are downloaded from http://trec.nist.gov/data.html to study the performance of Fuzzy C-means, PSO, PSOK and proposed PSOFCM algorithms. Every document vector is normalized to a unit length to diminish the influence of the length variations of various documents. We are using four different datasets and each dataset differs from other by number of documents and number of words as shown in Table 1.

B. *Experimental Setup*

The Fuzzy C-means, PSO and PSOFCM algorithms are tested with different collection of text documents. We have used two similarity measures namely, Euclidian distance and cosine correlation for all the methods taken for study.

TABLE I. TEXT DOCUMENT DATASETS

Data	**Number of documents**	**Number of terms**
Dataset1	814	16429
Dataset2	713	15804
Dataset3	512	12367
Dataset4	1639	19851

The Fuzzy C-means algorithm is conventionally viewed as unsupervised clustering algorithm. The Fuzzy C-means makes quick congregation when compare to other methods. The efficiency of this algorithm can be enhanced by giving the initial centroid obtained from some other effective methods. The Fuzzy C-means algorithm, PSO, PSOK and proposed PSOFCM algorithms are tested with four different datasets. From our observation it is clear that fuzzy C-means method can congregate to a solution in minimum number of iterations if it is applied to large text collection.

The PSO algorithm simply uses the objective function to evaluate its candidate solutions, and operates upon the resultant fitness values. The value 0.72 is assigned to the inertia weight w and the value 1.49 is assigned to c1 and c2, termed as acceleration or learning factors represent the weighing of the stochastic terms that pull each particle towards local best and global best respectively, on the basis of the experimental outcomes produced by researchers Shi and Eberhart. To ensure globalized search the inertia weight is kept constant for all algorithms.

C. *Results*

Though the fitness function is used for fitness value calculation in PSO, it is also used for analyzing the excellence of the clusters formed. ADDCC is the value represents the mean distance between documents and the cluster centroid. If the ADDCC value is lower, it means that the final outcome is more compact. Table 2 and 3 shows the investigational outcomes of all the five methods taken for study. The ADDCC values at each iteration are noted for all the five algorithms to show the congregation attribute of all the methods. The ADDCC value calculated using Euclidian distance measure is displayed in Table 2 and ADDCC value calculated using Cosine Correlation measure is displayed in Table 3.

TABLE II. PERFORMANCE BASED ON EUCLIDIAN DISTANCE

ED	ADDCC Value			
	PSO	**FCM**	**PSOK**	**PSOFCM**
Dataset1	6.748	7.247	4.556	4.149
Dataset2	6.318	7.036	4.824	4.211
Dataset3	4.169	4.489	2.550	2.549
Dataset4	9.309	9.091	6.024	6.003

TABLE III. PERFORMANCE BASED ON COSINE CORRELATION

CC	ADDCC Value			
	PSO	**FCM**	**PSOK**	**PSOFCM**
Dataset1	10.638	8.989	9.083	7.681
Dataset2	9.689	8.065	8.126	7.685
Dataset3	5.761	5.088	5.078	4.351
Dataset4	12.860	10.219	10.363	9.538

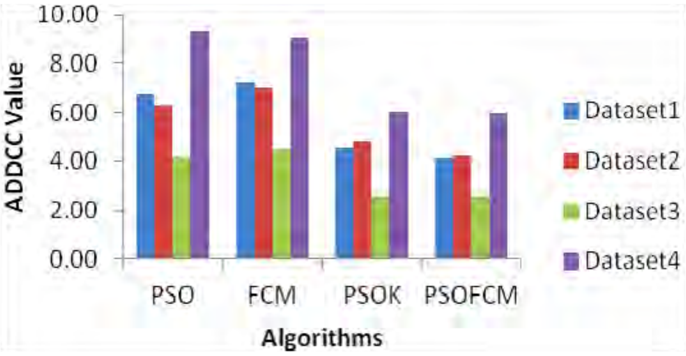

Fig.1.Performance based on Euclidian distance

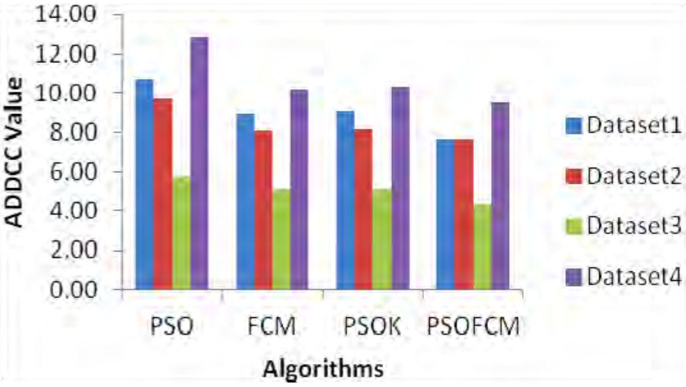

Fig.2. Performance based on Cosine Correlation

The lowest ADDCC value was produced by PSOFCM clustering approach. The simple PSO method produces higher ADDCC value and it means that this method alone may not be used for producing better optimal solution. Our proposed new method PSOFCM is performing well with both the similarity measures and it produces the lowest ADDCC value for all the datasets.

Fig. 1 shows the pictorial representation of performance of different methods on four given different datasets using the similarity measure Euclidian distance. *Fig. 1* shows the pictorial representation of performance of different methods on four given different datasets using the similarity measure cosine correlation.

From *Fig.1 and Fig.2*, it is clear that FCM and PSOK approaches have almost the same functionality. As the initial processing is done with the help of Fuzzy C-means clustering algorithm, the functionality and outcome is also seems to be uniform in all these three methods when it is applicable with different datasets. But when we use PSO for initial processing it performs better. Both PSO and PSOFCM perform similar initially but PSOFCM produces better ADDCC value as it takes the input from PSO and optimization is done with Fuzzy C-means. The PSOFCM method's efficiency considerably increases. The PSOFCM method has same congregation approach as like PSO because during the first 25 iterations, the PSO and the PSOFCM methods carry out the exact PSO procedure. During the second part of execution, the value of ADDCC is getting reduced from higher value to a lower value.

D. Discussion

Some researchers identified that most of the methods uses Fuzzy C-means for producing basic solution as an input for methods to get best possible solution. But few researchers also suggested an idea of enhancing the process of PSO method by giving the input taken from the output of Fuzzy C-means method. Some researchers also experimented with minimum elements and maximum number of execution. But our investigational outcome displayed in Tables 2 & 3 shows that the simple PSO method has not proved any development in the case of huge text collection. We have done this experiment with the document set having more than 5000 terms. From our experimental results it is very clear that our proposed PSOFCM method produces the least ADDCC value for every dataset taken for our study under both Euclidian measure and cosine correlation measure. The outcome of each method is shown in Tables 2 and 3 for each dataset.

VI. CONCLUSION

In this work, an algorithm called "PSOFCM" is formulated. It is clear that the Fuzzy C-means clustering method converges quicker than PSO method and at the same time this method generally fascinates towards a confined best possible solution. The efficient comprehensive findings of PSO method and the capacity of quickest convergence of Fuzzy C-means is clubbed together in our proposed PSOFCM method. This helps to reduce the demerits of Fuzzy C-means and PSO algorithm individually. This proposed method consists of two parts: PSO component and Fuzzy C-means component. The PSO part is processed during the beginning phase to identify the area of best possible solution by a comprehensive search and also to keep away from unbearable and repetitive calculation. The outcome of the PSO part is well utilized in the Fuzzy C-means part for refinement and producing the global best solution using Fuzzy C-means algorithm. The experimental result shows that our proposed PSOFCM method produces better clustering results than other approaches.

REFERENCES

[1] Al-Sultan, K. S. and Khan, M. M. Computational experience on four algorithms for the hard clustering problem. Pattern Recognition Letters, Vol.17, No. 3, pp: 295–308, 1996.

[2] Anderberg, M. R., Cluster Analysis for Applications. Academic Press, New York, NY, 1973.

[3] Berkhin, P., Survey of clustering data mining techniques. Accrue Software Research Paper, 2002.

[4] Carlisle, A. and Dozier, G. An Off-The-Shelf PSO, Proceedings of the 2001 Workshop on Particle Swarm Optimization, pp. 1-6, Indianapolis, IN, 2001.

[5] Cios K., Pedrycs W., Swiniarski R.,. Data Mining – Methods for Knowledge Discovery, Kluwer Academic Publishers., 1998

[6] Cui X., Potok T. E., Document Clustering using Particle Swarm Optimization, IEEE Swarm Intelligence Symposium 2005, Pasadena, California.

[7] Eberhart, R.C., and Shi, Y., Comparing Inertia Weights and Constriction Factors in Particle Swarm Optimization, 2000 Congress on Evolutionary Computing, vol. 1, pp. 84-88.

[8] Everitt, B., Cluster Analysis. 2nd Edition. Halsted Press, New York. 1980.

[9] Jain A. K., Murty M. N., and Flynn P. J.,. Data Clustering: A Review, ACM Computing Survey, Vol. 31, No. 3, pp. 264-323, 1999.

[10] Hartigan, J. A.. Clustering Algorithms. John Wiley and Sons, Inc., New York, NY., 1975.

[11] Kennedy J., Eberhart R. C. and Shi Y.,. Swarm Intelligence, Morgan Kaufmann, New York, 2001.

[12] Omran, M., Salman, A. and Engelbrecht, A. P., Image classification using particle swarm optimization. Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning 2002 (SEAL 2002), Singapore. pp. 370-374.

[13] Porter, M.F., An Algorithm for Suffix Stripping. Program, Vol.14 No. 3, pp. 130-137, 2002.

[14] Salton G. Automatic Text Processing. Addison-Wesley, 1989.

[15] Salton G. and Buckley C., Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24 (5): pp. 513-523, 1988.

[16] Selim, S. Z. And Ismail, M. A. Fuzzy C-means type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Mach. Intell. 6, 81–87, 1984.

[17] Shi, Y. H., Eberhart, R. C., Parameter Selection in Particle Swarm Optimization, The 7th Annual Conference on Evolutionary Programming, San Diego, CA., 1998.

[18] Steinbach M., Karypis G., Kumar V.,A Comparison of Document Clustering Techniques. TextMining Workshop, KDD, 2000.

[19] TREC. Text Retrieval Conference. http://trec.nist.gov., 1999.

[20] Van D. M., Engelbrecht, A. P. Data clustering using particle swarm optimization. Proceedings of IEEE Congress on Evolutionary Computation (CEC 2003), Canbella, Australia. pp. 215-220, 2003.

[21] Zhao Y. and Karypis G., Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering, Machine Learning, 55 (3): pp. 311-331, 2004.

[22] Tunchan cura, A particle swarm optimization approach to clustering, Expert Systems with Applications, Vol: 39 pp.1582–1588., 2012.

[23] Xiaohui Cui, Thomas E. Potok, Document Clustering Analysis Based on Hybrid PSO+Fuzzy C-means Algorithm, Computational Data Analytical Group,2006.

[24] Jaganathan P and Jaiganesh, An improved k-means algorithm combined with particle swarm optimization approach for efficient web document clustering, IEEE eXplore, 2013.

[25] Sunita Sarkar, Arindam Roy, Bipul Shyam Purkayastha, Application of Particle Swarm Optimization in Data Clustering: A Survey, International Journal of Computer Applications (0975 – 8887) Volume 65– No.25, March 2013.