Expert Systems With Applications 134 (2019) 192–200

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization

R. Janani*, Dr. S. Vijayarani

Department of Computer Science, Bharathiar University, Coimbatore, India

ARTICLE INFO

Article history:
Received 23 November 2018
Revised 22 May 2019
Accepted 23 May 2019
Available online 24 May 2019

Keywords:
Text mining
Information retrieval
Text clustering
Spectral clustering
Optimization techniques
SK-means
Expectation-Maximization
Particle Swarm Optimization
SCPSO

ABSTRACT

Document clustering is a gathering of textual content documents into groups or clusters. The main aim is to cluster the documents, which are internally logical but considerably different from each other. It is a crucial process used in information retrieval, information extraction and document organization. In recent years, the spectral clustering is widely applied in the field of machine learning as an innovative clustering technique. This research work proposes a novel Spectral Clustering algorithm with Particle Swarm Optimization (SCPSO) to improve the text document clustering. By considering global and local optimization function, the randomization is carried out with the initial population. This research work aims at combining the spectral clustering with swarm optimization to deal with the huge volume of text documents. The proposed algorithm SCPSO is examined with the benchmark database against the other existing approaches. The proposed algorithm SCPSO is compared with the Spherical K-means, Expectation Maximization Method (EM) and standard PSO Algorithm. The concluding results show that the proposed SCPSO algorithm yields better clustering accuracy than other clustering techniques.

1. Introduction

Document clustering is a process of partitioning a pool of documents into distinctive clusters based on the content similarity. The group of clusters which contains the unstructured format also handles the same format (Mohammed, Yusof, & Husni, 2015a,b). It has been extensively used for effective navigation, organization, extraction, summarization and retrieval of the huge volume of text documents (Abraham, Das, & Konar, 2006). At first, the document clustering is used to improve the accuracy of information retrieval system. The document clustering is a proficient way of discovering the nearest neighbour of a particular document within the document collections. Nowadays, the clustering technique is applied in surfing a collection of documents and to normalize the results which are given by the search engine based on the user's query. Text clustering is used to extract the applicable features and denotes the features in meaningful ways (Mohammed, Yusof, & Husni, 2015a,b). The documents in text mining system are represented as high dimensional documents with complex semantics. General applications of document clustering are automatic topic extraction, document organization and information retrieval. However, a lot of important research work has been done in the area of text clustering, which needs the determination to improve and enhance the quality of text clustering process. The proposed work is projected in that particular direction.

Spectral clustering is an innovative clustering technique used for partitioning the graph matrix (Kumar and Daumé, 2011). It has the capability to recognize the non-convex distribution when compared with the standard clustering algorithms (Ng, Jordan, & Weiss, 2002; Yogatama and Tanaka-Ishii, 2009). It makes the undirected, weighted graph based on the data objects, to achieve the optimal clustering results by considering the Eigen values and Eigen vectors which are associated with the graph. Spectral clustering is applied in the fields of machine learning, text mining, speech recognition, very large-scale integration design, web classification and image processing.

1.1. Motivation

In recent days, the volume of information is growing extremely in size. There is a great quality of data in the form of unstructured nature which is to be located in our data repositories. Users will be able to download the documents which will be residing their system drives or different folders. Sometimes the similar content will be stored in different places. For example, if the user wants to retrieve the "Data Mining" related documents from their system;

* Corresponding author.
E-mail address: janani.sengodi@gmail.com (R. Janani).

https://doi.org/10.1016/j.eswa.2019.05.030
0957-4174/

they have to search each and every folder manually and also have to go through the full content of the documents, because of improper file names. In this situation the search task become tedious and the search time also will increase.

In order to overcome these problems, a formulation of text clustering approach that enables the grouping of related documents based on the content is applied.

1.2. Contribution

Text document clustering is a basic process used in information retrieval, automatic topic extraction and document organization. High quality clustering algorithms plays a vital role efficiently in organizing, summarizing and navigating the unstructured documents. The main contributions of this research work is as follows,

- This research work proposes a novel approach to text clustering, by grouping documents into clusters based on the content.
- The proposed approach including the Swarm Intelligence algorithm, Particle Swarm Optimization (PSO) with the Spectral Clustering algorithm has been enforced and estimated for text clustering.

This paper deals with iterative algorithms with a stochastic approach for improving text document clustering accuracy. The rest of this paper is organized as follows: Section 2 illustrates the various related work for spectral clustering with optimization methods. Section 3 describes the methods of document clustering, spectral clustering, optimization techniques and its functions. Section 4 presents the proposed SCPSO algorithm for improving text document clustering. Section 5 shows the performance measures for evaluating the text clustering. Section 6 clarifies the observations of the experiments carried out on different data sets, results comparison with existing methods and the implementation. As a final point, Section 7 discusses the conclusion of this paper and recommends for the future improvement.

2. Related works

Among the researchers in the area of text document clustering, some of the researchers utilize the advanced and insistent techniques for document clustering. Document clustering is according to partition of documents into distinct clusters based on the similarity of the particular document content. Document clustering is the simple, inventive method and this process is influential in the area of text mining (Shahnaz, Berry, Pauca, Plemmons, 2006). The common clustering algorithm k-means always gives the local optimal solution for the particular problem. Hence, to achieve the global optimal solution this basic algorithm was used by the Swarm Intelligence (Karaboga and Ozturk, 2011). Swarm Intelligent algorithms are the combined intelligence of the simple agents (Bonabeau, Marco, Dorigo, & Theraulaz, 1999). These optimization techniques are developed to discover the global optimal or local optimal solutions for a particular problem (Karaboga and Ozturk, 2011).

In this research work the Meta Heuristic optimization technique is used to attain the global optimal solution. The Particle Swarm Optimization (PSO) algorithm was framed by Kennedy and Eberhart in the year 1995. The ultimate purpose of this algorithm is to have whole particles which come across the optima in a high dimensional data volume (Wang, Shi, Hong, 2010). Spectral clustering is a dominant clustering technique using a graph matrix (Ding et al., 2001). A spectral based method with genetic algorithms (Menéndez and Camacho, 2015) was used to analyze large volume of data analysis. This technique was applied in various applications along with information retrieval, which leads in bringing out successful solution for the clustering problem (Kamvar et al., 2003).

A deep analysis of spectral clustering methods is specified in Ng et al. (2002) an associated method termed Modularity Eigen Map which is used to retrieve the structured features from the document. Spectral Clustering is actually powerful (Kamvar et al., 2003) but needs to solve the Eigen Value problem of the Laplacian Matrix converted from the similarity matrix corresponding to the given data set. The major disadvantage of the spectral clustering is to handle a large volume of documents. This algorithm exemplifies document collections is an undirected graph (Shi and Malik, 2000). The standard functions used are normalized cut (Shi and Malik, 2000), ratio cut (Chan, Schlag, & Zien, 1994), average association (Shi and Malik, 2000) and min-max cut (Ding et al., 2001). The self-tuning clustering algorithm was proposed and it spontaneously evaluated the local parameters of each document data point based on the data distribution.

Later a few researchers have presented the development and improvement of this algorithm to enhance its performance. They proposed a Spectral Clustering algorithm to handle the multiple views of the document corpus (Kumar and Daumé, 2011). An innovative Non-negative Matrix Factorization (NMF) used as affinity matrix for document clustering was proposed, which implements the non-negativity constraint and orthogonal constraints concurrently (Bao, Tang, Li, Zhang, & Ye, 2008). A new spectral algorithm was proposed with the combination of Nystrom Spectral Clustering and Genetic Algorithm. It causes the ordinal encoding in the genetic approach with the spectral extension (Menéndez and Camacho, 2015).

3. Methods

The main aim of the document clustering is to classify the documents into groups or clusters based on the content similarity of the particular document.

3.1. Document preprocessing

The document processing is the significant phase to represent the documents for efficient document clustering (Oliveira, and Seok, 2005). In this research work the preprocessing techniques such as tokenization, stop word removal and stemming are used. Tokenization is the method of separating a stream of text content into words, terms, symbols or certain other expressive features called tokens. The list of tokens goes into an input for further processing which includes parsing or text mining. Most of the words in the documents occur very often, but they are basically meaningless words as they are used to connect the words well organized to form a sentence. Generally, it is assumed that stop words, which does not denote content or context of text documents. Stemming is the method of combining the different types of a word into a typical illustration, the stem (Porter, 1980).

3.2. Similarity measures

Before clustering the documents, the similarity measure between the documents should be determined. There are two extended ways which are being used to measure the correspondence among two documents.

3.2.1. Cosine similarity

When the documents are signified as a term vectors the correspondence between two documents that relate to the correlation among the vectors. This can be calculated as the cosine angle between the vectors. So, this similarity measure can be defined as cosine similarity. This similarity measure is the most popular

method which is used in the field of text mining such as, information retrieval, document classification and document clustering (Karypis, Kumar, & Steinbach, 2000). Consider the documents d_1 and d_2 the cosine similarity measure is defined as

$$CSim(d_1, d_2) = \frac{d_1.d_2}{|d_1|\ X\ |d_2|} \tag{1}$$

where the $d_1$and d_2 are the multidimensional vectors above the set of terms $TS = \{T_1,T_2,\ldots, T_n\}$. From this, every dimension denotes a term along with its weight between documents, which is non-negative. So, the similarity measure is non-negative and bounded within {0, 1}. A significant property of this similarity measure does not depend on the length of the document.

3.2.2. Euclidean distance

Euclidean distance is a typical metric for many kinds of data analytical problems. It is also the normal distance between two arguments and so it can be measured in a multi-dimensional space. Euclidean distance is broadly used in document clustering and classification (Karypis et al., 2000). Let the documents be d_1 and d_2 the Euclidean distance of these two documents is defined as

$$EDist(d_1, d_2) = |d_1 - d_2| \tag{2}$$

where d_1 and d_2 are the multidimensional vectors above the set of terms $TS = \{T_1,T_2,\ldots, T_n\}$.

3.3. Document representation

In this research work we use the Term Frequency- Inverse Document Frequency (TF-IDF) vector space model (Karypis et al., 2000). Consider the given documents $D = \{d_1,d_2,\ldots, d_n\}$. and the term $t = \{t_1,t_2,\ldots, t_n\}$ occur in document $d_1,d_2,\ldots, d_n$ the raw count is denoted by $r_{t,d}$. The TF is defined as

$$TF(t, d) = \log\left(1 + r_{t,d}\right) \tag{3}$$

Let N be the total number of documents in the document collection, the IDF is defined as

$$IDF(t, d) = log\frac{N}{|d \in D : t \in d|} \tag{4}$$

So the TF-IDF is computed as

$$TFIDF(t, d, D) = TF(t, d).IDF(t, D) \tag{5}$$

3.4. Maximum likelihood estimation

To estimate the maximum likelihood, the generalized EM algorithm is used in this research work. It is a mixture model which is used to perform cluster analysis. Various distribution metrics are used to estimate the likelihood such as Bernoulli, Gaussian, Multinomial, Von-Misses Fisher, Poisson, Binomial, Beta etc., The EM algorithm is broader than normal distribution. Hence to find out the probability of large volume of distribution, the Gaussian distribution is used in this research work. The probability of Gaussian distribution at the point of x is considered as

$$\text{probablity}(x_i|\Theta) = \frac{1}{\sqrt{2\Pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6}$$

A set of m points are generated from the Gaussian distribution with one dimensional space. Assume that the generated points are independent and the probability of these points is the product of individual probabilities. The probability is considered as,

$$\text{probablity}(x_i|\Theta) = \prod_{i-1}^{m} \frac{1}{\sqrt{2\Pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

3.5. Spectral Clustering algorithm

Spectral Document Clustering has developed in recent times as a widespread clustering technique, which motivated the emerging criterion functions and the developing algorithm to produce further accurate clusters (Ng et al., 2002 and Ailem, Role, & Nadif, 2017). It uses the Eigen Vectors of the graph matrices which is derived from the documents. This algorithm is based on the concept of the weighted undirected graph. It demonstrates the collection of documents, i.e. document corpus $D = \{d_1,d_2,\ldots\ldots, d_n\}$. as an undirected graph $G(V_s,E_s,M_a)$ where V_s is a Vertex Set, E_s denotes the Edge Set and M_a denotes the Graph Affinity Matrix. Each vertex $Vi \in V_s$ signifies the ith document and edge $(i, j) \in E_s$ is allocated to an affinity score (Bach, and Jordan, 2004). This algorithm consists of the following steps.

3.5.1. Similarity graph

The spectral clustering method is established on the similarity graph, where, there are a numerous way to transform a set of documents $\{d_1,d_2,\ldots, d_n\}$. from document collections with the pair wise similarity ps_{ij} or distance wise similarity $ds_{ij,}$ into a graph (Abdulsahib, 2015; Ailem, Role, & Nadif, 2015). These similarities can be expressed in the following methods:

- $\in$ – Neighborhood Graph: Connects all the points whose pair wise distance is smaller than $\in$
- K- Nearest Neighbor Graph: Connects the vertex V_i to V_j if V_j is the nearest neighbor of V_i
- Fully connected Graph: Connects all the points with the positive similarity

3.5.2. Graph Laplacian Matrix

The most important part of spectral clustering is the graph Laplacian Matrices. If there is no unique resolution then the matrix is precisely called as graph Laplacian (Chan, et al., 1994; Ding et al., 2001). The frequently used graph Laplacian types are as follows,

- Un-normalized Graph Laplacian: It defines the matrix as

$$LM = D - W \tag{8}$$

- Normalized Graph Laplacian: There are two matrices which are called normalized graph Laplacian, when the matrices are thoroughly associated with each other and are defined as

$$\begin{aligned} LM_{sm} &= D^{-1/2} LM\ D^{-1/2} \\ &= I - D^{-1/2} W\ D^{-1/2} \end{aligned} \tag{9}$$

- Normalized Graph Laplacian is related to Random Walks and it is defined as

$$\begin{aligned} L_{rw} &= D^{-1/2} LM \\ &= I - D^{-1/2} W \end{aligned} \tag{10}$$

where the LM denotes the Laplacian Matrix, W is the Weighted Graph with the Weight Matrix, where $W_{ij} = W_{ji} = \geq 0$. In this formula, I is the Identical Matrix, D is the Diagonal Matrix, LM_{sm} is a Symmetric Matrix and L_{rw} is closely connected to Random Walk (Bach, and Jordan, 2004).

3.5.3. Eigen Vectors

The Eigen Vectors of this matrix is deliberated as the points and the clustering algorithms are applied over them to express the proper clusters (Menéndez and Camacho, 2015).

Algorithm. Spectral Clustering.

Input: Similarity Matrix $S \in R^{n \times n}$, the number of clusters K. Let W be the weighted matrix
Output: K number of clusters
Step 1: Build the similarity matrix using the way which is described in Section 3.2.1
Step 2: Compute the un-normalized Laplacian using Eq. (8)
Step 3: Estimate the cluster K with the Eigen Vectors V_1,V_2,..., V_k of LM
Step 4: Let $V \in R^{n \times n}$ which contains the vectors V_1,V_2,..., V_k as a column
Step 5: Let $Y_i \in R^k$ be the vector corresponding to the ith row of V
Step 6: Group the points Y_i in R^k with the clustering algorithm into {K_1,K_2,..., K_n}

3.6. Particle Swarm Optimization

Particle Swarm Optimization is a global stochastic optimization technique for incessant methods and it was described by Eberhart and Kennedy in 1995. The K-means algorithm is suitable for the initial clustering conditions, which can cause this algorithm to converge upon suboptimal solutions. But, the PSO algorithm is less sensitive for the initial conditions because of its population based nature. Hence, the PSO is more likely to find the near optimal solution. The ultimate purpose of this algorithm is to have whole particles come across the optima in a high dimensional data volume (Dhillon, Guan, & Kulis, 2004). The location of the particle in the high dimensional problem space (Foong and Yong, 2016), epitomize to explain a particular problem. While a particle travels from one location to a new location, another solution will be generated. These solutions are assessed by a fitness function which provides the optimal solution. Each particle will recall its recent coordinates and its velocity which shows the particle movement speed along with the dimensions of a problem space, from which the finest fitness value is established. The best value is coupled with its neighbor's best value, impacts the movement of every particle over the problem space.

In this context, the current position is P_i, the best position is bp_i and the velocity is V_i, then the ith particle is signified by its position designated as $p_i = \{p_{i1}, p_{i2}, \ldots, p_{in}\}$. The typical Particle Swarm Optimization technique will refurbish the velocity and the position of each and every particle as stated below,

$$v_{in}(I+1) = W.v_{in}(I) + c_1.rand(\ ).(bp_{in} - x_{in}) + c_2.rand(\ ).(bp_{in} - x_{in}) \quad (11)$$

$$x_{in}(I+1) = v_{in}(I+1) + x_{in}(I) \quad (12)$$

where $rand()$ is random number which lies in {0, 1}, c_1 and c_2 are the positive constants, bp_{in} is the best position set up by the ith particle respectively. The count of the iteration is denoted by I and the weight W is rapidly diminishing during the iterations. The weight function W is used to stabilize the local and global search (Foong and Yong, 2016).

Algorithm. Particle Swarm Optimization Clustering (PSO).

Step 1: Initialize each particle with K cluster centers, d is the distance vector
Step 2: For iteration count $I = 1$ to expected_number_of_iterations {
Step 3: For all particle p_i and the entire pattern P_i in the corpus {
Step 4: Estimate the Euclidean distance Eq. (2) of P_i with all cluster centroids
Step5: Set P_i to the current cluster which has the closest centroid to P_i
Step 6: Estimate each particle's fitness function
$f = Min \sum_{j=1}^{K} \sum_{i=1}^{n} d[v_i, x_{in}]$

(continued on next page)

if $f(v_i) < pb_i$ *then*
$pb_i \leftarrow f(v_i)$
if $f(v_i) < gb_i$ *then*
$gb_i \leftarrow f(v_i)$
} End for
Step 7: Find the local best and global best position of each particle.
Step8: Updating the particle position and particle velocity by using Eqs. (11) and (12) until to reach maximum iteration or no change in global best position.
} End for

In the PSO document clustering, the high dimensional document vector space is demonstrated as the problem space in the swarm optimization. Each and every term within the document denotes the problem space with a single dimension. In this context, every document vector is often delineated as a part within the problem space. The swarm which contains a single particle signifies the possible solution for the text document clustering. Hence, the swarm denotes a variety of clustering solutions for the collection of documents. So, each particle in the swarm conserves a form of matrix

$$P_i = (K_1, K_2, \ldots, K_i, K_n) \quad (13)$$

where K_i denotes the ith cluster centroid and K_n represents the cluster number. At every emphasis, the particle changes the centroid position over the problem space as indicated by its own particular experience and neighbour particles (Fig. 1).

4. SCPSO algorithm

The proposed SCPSO is used to improve the accuracy of text document clustering. It will reduce the global convergence, computational complexity and it can handle the number of objective functions. The SCPSO is a spectral based clustering method which uses the particle swarm optimization algorithm. The main goal of this algorithm is to improve the text clustering accuracy. It has the following three important steps:

4.1. Similarity graph generation

The similarity function is to be applied to the document dataset and combining all the points with each other. This will accomplish the similarity graph matrix.

4.2. Particle swarm optimization

This step is used to find out the optimal solution. The probable solutions which are called particles, fly over the global problem space by subsequent current optimum particles. Each particle observes its matches in the global problem space, associated with the fitness value. This value is referred to as the pbest. As soon as the particle precedes all the population as its neighbors, this value is denoted as global best or gbest. The particle swarm optimization concept comprises of, at every step, altering the velocity of each particle on the way to its pbest and best locations.

4.3. Clustering methods

The greatest accuracy solution is preferred as a solution of the clustering algorithm and the numbers of documents that are dispensed to the particular cluster leading to the preferred solution.

The proposed SCPSO algorithm is explained in a step by step process as follows. First, it will choose the document which belongs to the document corpus D. There is a necessity to retrieve the Eigen vectors from the range in order to the particular document. Let S_m be the Similarity Matrix and it is well-defined by using the Eq. (1). Then calculate the next similarity matrix between

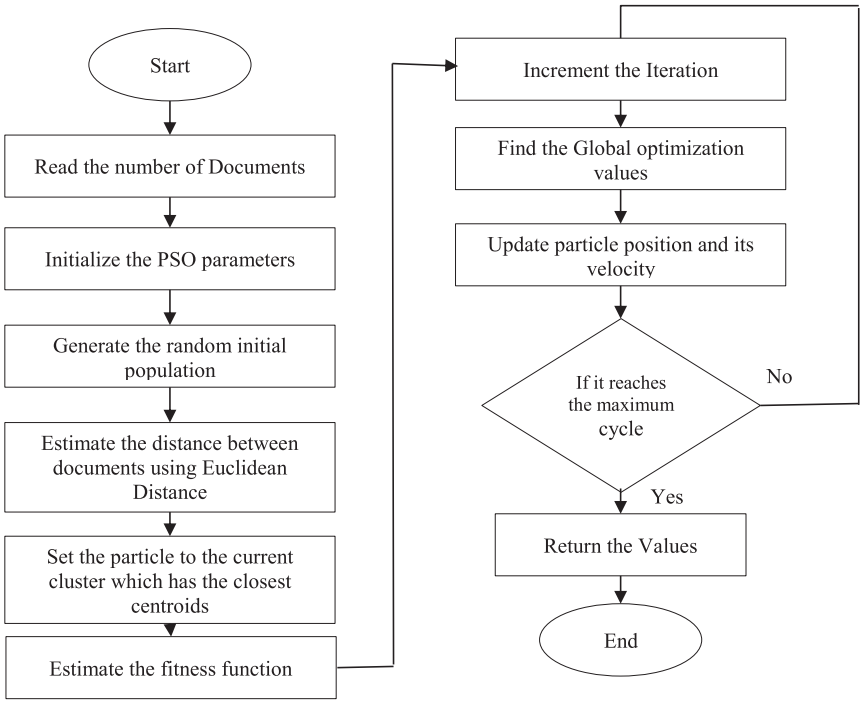

Fig. 1. Flow chart for PSO document clustering.

the first similarity and the remaining documents in the document corpus (Step 1 to Step 3).

Next there is a need to construct the Laplacian matrix. So we need to express the diagonal matrix D whose (i, i) element is the sum of the *i*th row of sample document S. The diagonal matrix can be restructured as,

$$D = \begin{pmatrix} D_{i1} & 0 \\ 0 & D_{i2} \end{pmatrix} \tag{14}$$

which is based on the Eq. (9).

$$LM = D^{-1/2}\ W\ D^{-1/2}$$

By setting the diagonal matrix,

$$D_{i,j} = \sum_{j=1}^{N} W a_{i,j} \tag{15}$$

From this find out the large Eigen Value of laplacian matrix (Step 4 to Step 6).

Initialize each food particle in the optimization phase. For each particle we have to calculate the fitness value. The $x_{i,k}$ represents the *n*th document vector which belongs to the cluster k.

$$Fitness = Min \sum_{j=1}^{K} \sum_{i=1}^{n} d\left[v_i, x_{i,k}\right] \tag{16}$$

Then estimate the particle velocity of each particle by using this formula (Step 7 to Step 13),

$$Pbest(t+1) = \begin{Bmatrix} pbest_i(t)\ if\ f(x_i(t+1)) \geq f(pbest_i(t)) \\ x_i(t+1)\ if\ f(x_i(t+1)) < f(pbest_i(t)) \end{Bmatrix} \tag{17}$$

Update the current position by using

$$Gbest(t+1) = \arg\ min_i f(pbest_i(t+1))\ 1 \leq i\ \leq N$$

Select the best individual of this search and assign the particular documents into the matrix Y (Step 14 to Step 19).

Algorithm. SCPSO.

Input: $D = \{d_1, d_2, \ldots\ldots, d_n\}$ and the number of clusters C
Output: Best cluster C
Step 1: Choose the sample documents $S = \{s_1, s_2, \ldots s_n\} \in D = \{d_1, d_2, \ldots\ldots, d_n\}$ where $n_1 < < n$
Step 2: Since the graph similarity $S \in R^{nXn}$ defined by the $S_1 = similarity(s_i, s_j)$
Step 3: Estimate the similarity matrix S_2 designed by the similarities between the elements of S_1 and the remaining documents in the corpus.
Step 4: Define the diagonal matrix D whose (*i*, *i*) element is the sum of the *i*th row of S.
Step 5: Construct the Laplacian matrix $LM = D^{-1/2} W D^{-1/2}$
Step 6: Find the largest Eigen Vectors of Laplacian matrix *LM* and construct the matrix $V = V_1, V_2, \ldots, V_k \in R^{nXk}$
Step 7: Form the matrix Y from V by renormalizing each row of V to have unit length.
Step 8: Initialize each particle with random position and velocity
Step 9: For each particle calculates the fitness value
$f = Min \sum_{j=1}^{K} \sum_{i=1}^{n} d[v_i, x_{i,k}]$
Step 10: If the *fitness value* > *pbest*
Set *pbest* = *current fitness value*
Step 11: If the *pbest* > *gbest*
Set *gbest* = *pbest*
Step 12: Calculate the particle velocity for each particle
$Pbest(t+1) = \{^{pbest_i(t)\ if\ f(x_i(t+1)) \geq f(pbest_i(t))}_{x_i(t+1)\ if\ f(x_i(t+1)) < f(pbest_i(t))}\}$
Step 13: Update the current position and velocity of the particle
$Gbest(t+1) = \arg\ min_i f(pbest_i(t+1)) 1 \leq i \leq N$
Step 14: End For
Step 15: Select the best individual of this search
Step 16: Assign the documents into Matrix Y with its centroids to the cluster C
Step 17: until it reaches the maximum iteration or minimum error criteria
Step 18: Return the clustered documents

5. Implementation details

5.1. Experimental setup

All the experiments are carried out on a 2.00 GHz Intel CPU with 1GB of memory and running on windows 8. We implement the algorithm to acquire the accurate cluster of documents and verify the success of clustering.

Table 1
Dataset description.

Dataset name	Characteristics			
	Number of documents	Number of words	Number of clusters (*K*)	Sparsity (%)
Reuters	8203	18,914	15	99.41
20Newsgroup	8759	29,584	15	99.82
TDT2	9394	36,771	15	99.64

The proposed algorithm was experimented with the benchmark datasets, which is used to report the problems faced while clustering the text documents. Further, the proficiency of the proposed SCPSO algorithm has been verified by comparing with various clustering techniques, namely, SK-means, Expectation Maximization (EM) method and standard Particle Swarm Optimization (PSO) algorithm. The objective functions such as, Cluster Purity, entropy, mutual information metric, precision, recall and f-measure are deliberated for achieving the global optimal solution.

5.2. Datasets

In this experimentation, we have used a total number of three different data sets with fifteen semantic categories. For all the datasets, we applied a preprocessing technique which is explained in the above section. The summary of dataset is given in Table 1. It describes the number of documents, words, clusters and how the documents are scattered in the dataset.

- Reuters −21,578 was collected from the Reuters Newswire in the year 1987. The documents in this corpus were categorized by personnel from Reuters and Carnegie group Ltd. This document corpus contains 135 categories with a total number of 21,578 documents.
- 20 Newsgroup document dataset was collected from 20 different types of newsgroups and the document corpus contains 20 categories with approximately 20,000 numbers of documents.
- TDT2 is a task for topic detection and topic tracking. This document corpus contains 96 categories with 64527number of documents from voice of America World News, American Broadcasting Company (ABC) World News, CNN headline News and AP World stream.

5.3. Performance measures

In this research work, the clustering performance is assessed by associating the acquired label of every document with that delivered by the collection of documents. The performance measures are Accuracy, Normalized Mutual Information and Adjusted Rand Index which leads to discover the best clustering accuracy.

5.3.1. Accuracy

The Accuracy is equivalent to the ratio of accurate matching pair number to the total matching pair number. A true positive (TP) result allocates two related documents in the same cluster; a true negative (TN) result allocates two dissimilar documents to different clusters. A (FP) result assigns two dissimilar documents to the same cluster. A (FN) result assigns two similar documents to different clusters (Oliveira, and Seok, 2005).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

5.3.2. Normalized mutual information (NMI)

Normalized mutual information is an external measure for defining the good quality of clustering. This method is normalized hence, we can compare the normalized mutual information among various numbers of clusters. Let K be the set of clusters, C be the class label, H(.) is the entropy and *I*(*C: K*) is the mutual information between C and K. Then mutual information is calculated by using,

$$I(C:K) = H(C) - H((C|K)) \tag{19}$$

$$NMI\ (C,K) = \frac{2\ X\ I\ (C:K)}{[H(C) + H(K)]} \tag{20}$$

5.3.3. Adjusted Rand index (ARI)

The adjusted Rand index adopts the global hyper geometric distribution as the model of randomness, i.e., the U and V partitions are selected randomly such that the number of objects in the classes and clusters are fixed. Let n_{ij}be the number of objects that are in both class u_i and cluster v_j. Let $n_i$and $n_j$be the number of objects in class $u_i$and cluster v_j respectively (Yeung, Fraley, Murua, Raftery, & Ruzzo, 2001). Then the ARI is defined as follows,

$$ARI = \frac{\sum_{i,j}\binom{n_{i,j}}{2} - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \left[\sum_i \binom{n_i}{2}\sum_j \binom{n_j}{2}\right] \Big/ \binom{n}{2}} \tag{21}$$

6. Numerical experiments

In this section, numerous experiments were implemented to illustrate the effectiveness of our proposed clustering algorithm. Three standard collection of documents with different categories were used in this research work: Reuters −21,578, 20newsgroup and TDT2. We compared our proposed algorithm with widely used clustering algorithm and optimization technique. Table 2 summarizes the performance of the compared algorithms in terms of Accuracy (ACC), Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI), over all datasets. The results were averaged for the 15 clusters. From this, we observe that the proposed SCPSO perform noticeably better than the other existing algorithms such as, SK-means in terms of Acc, NMI and ARI, in almost all cases.

Note that, the PSO algorithm performs slightly better than the other variants in almost all the situations in terms of all the performance measures. Furthermore, we observed that the proposed SCPSO algorithm reached good performance when compared to the existing algorithms. In Fig. 2, the accuracy was compared for all the

Table 2
Performance measures comparison.

Dataset	Performance measures	SK-means	EM	PSO	SCPSO
Reuters	ACC	0.612	0.679	0.765	0.842
	NMI	0.645	0.713	0.804	0.884
	ARI	0.607	0.648	0.721	0.811
20News group	ACC	0.675	0.714	0.728	0.832
	NMI	0.705	0.738	0.755	0.861
	ARI	0.628	0.694	0.713	0.806
TDT2	ACC	0.674	0.699	0.762	0.850
	NMI	0.709	0.732	0.806	0.894
	ARI	0.637	0.681	0.738	0.820

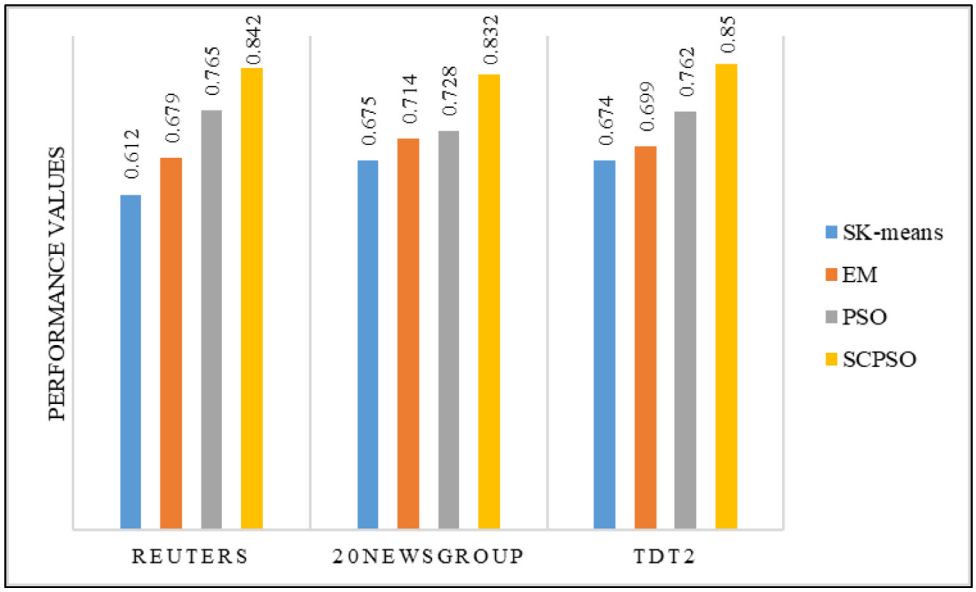

Fig. 2. Accuracy comparison.

Fig. 3. NMI comparison.

three datasets. From this, we infer that the SCPSO algorithm performs well when compared to other variants. When compared to the datasets accuracy, the Reuters and TDT2 are almost same with slight modification when compared to the 20Newsgroup dataset. In Reuters dataset, the SCPSO algorithm increases its performance by 6% in terms of accuracy and 7% for the 20Newsgroup dataset. In TDT2 dataset, the accuracy was increased over existing techniques by 7%.

Fig. 3, the NMI comparison was shown. From this we inferred that, the proposed clustering algorithm yields better performance for all the instances. Compared to existing techniques, PSO algorithm performs well. Compared to PSO the SCPSO gives better results. The NMI result is high when compared with the TDT2 dataset and it is slightly differing from the Reuters dataset. Relatively, there is a 7% of increase in the NMI for all the datasets with a little difference.

The ARI comparison is shown in Fig. 4. It is observed that, the SCPSO algorithm performs well with other variants of document clustering techniques. In Reuters dataset, the ARI value is increased by 6% from other algorithms. In 20Newsgroup dataset, the proposed algorithms is increased by 6% in terms of ARI and 7% is increased for the TDT2 dataset. On the whole, the proposed SCPSO algorithm yields better performance in terms of ARI.

Table 3 reveals the comparison of t-test values for all the datasets with proposed and existing techniques. This test shows that the SCPSO algorithm outperforms in all the cases in terms of ACC, NMI and ARI. The effectiveness of proposed algorithm is slightly increased when compared to the PSO clustering algorithm. The correction is a multiplicative factor depending on the total sample size, the cluster size, and the intra class correlation p. The corrected t-statistic has a student's t-distribution with reduced degrees of freedom. The corrected statistic reduces to the t-statistic computed by ignoring clustering when $p = 0$. It reduces to the t-statistic computed using cluster means when $p = 1$. If $0 < p < 1$it lies between these two values.

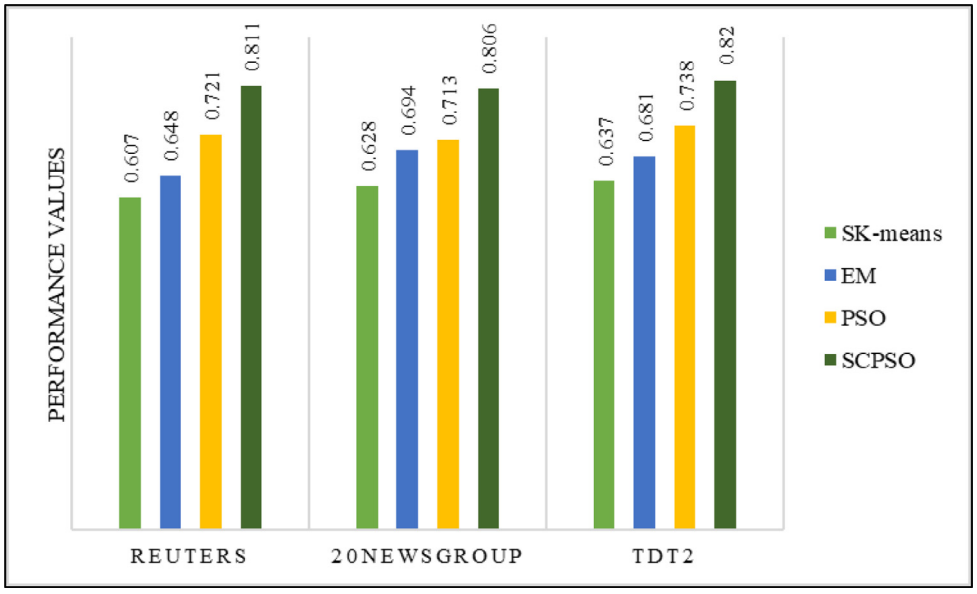

Fig. 4. ARI comparison.

Table 3
Comparison of *t*-test values.

Dataset	Performance measures	SK-means	EM	PSO	SCPSO	*P*-value
Reuters	ACC	0.63 ± 0.002	0.68 ± 0.004	0.77 ± 0.001	0.84 ± 0.005	≤0.001
	NMI	0.65 ± 0.001	0.71 ± 0.01	0.80 ± 0.004	0.88 ± 0.009	≤0.001
	ARI	0.61 ± 0.002	0.65 ± 0.005	0.72 ± 0.008	0.81 ± 0.004	≤0.001
20News	ACC	0.66 ± 0.003	0.71 ± 0.002	0.73 ± 0.05	0.83 ± 0.02	≤0.001
group	NMI	0.70 ± 0.001	0.74 ± 0.001	0.75 ± 0.005	0.86 ± 0.005	≤0.001
	ARI	0.63 ± 0.004	0.70 ± 0.008	0.71 ± 0.004	0.81 ± 0.006	≤0.001
TDT2	ACC	0.65 ± 0.002	0.70 ± 0.04	0.76 ± 0.006	0.85 ± 0.001	≤0.001
	NMI	0.71 ± 0.005	0.73 ± 0.004	0.81 ± 0.007	0.89 ± 0.01	≤0.001
	ARI	0.64 ± 0.004	0.68 ± 0.006	0.74 ± 0.004	0.82 ± 0.006	≤0.001

The SCPSO algorithm increases by 6% in terms of accuracy, 8% increase in terms of NMI and 9% increase in term of ARI when compared to PSO clustering algorithm in Reuters dataset. In 20Newsgroup dataset, there is 7% increase in accuracy, 9% increase in terms of NMI and 9% increase in ARI when compared to the PSO clustering algorithm. The 6% of increase in accuracy, 8% increase in NMI and in terms of ARI, there is a 9% of increased results in SCPSO. Hence the proposed algorithm yields better performance in terms of clustering the documents based on its contents.

7. Conclusion and future direction

Nowadays, the document clustering and classification problem is an open issue for researchers in the area of text mining and information retrieval. The ultimate purpose of this study is to attain the goal of assessing evolutionary algorithms and discovering ways to improve their performance to achieve the optimal solution. This research work proposed a new algorithm Spectral Clustering with Particle Swarm Optimization (SCPSO) to improve the document clustering accuracy and it leads the result on the way to an optimal solution. The Spectral Clustering algorithm has better results when compared to the existing clustering methods. Compared to spectral clustering, the proposed SCPSO algorithm yields better results to improve the accuracy of text clustering. The proposed algorithm attains best results even when volume of documents is high. The execution time and computational complexity are the leading deal of the execution as it gambles on the number of iterations, parameter selection, particle initialization, etc. The execution time is calculated for existing and proposed algorithms. In future, this method can be implemented on a multi-core CPU. It can also be joined or stretched to any other evolutionary algorithms to acquire the outstanding optimal outcomes. To attain the accomplished results of text document clustering, distinct intentions may be introduced. Furthermore, this work will be motivated by the enhancements that can be applied to the spectral and optimization algorithms.

Conflict of Interest

None.

References

Abdulsahib, A. K. (2015). *Graph based text representation for document clustering*. Universiti Utara Malaysia.

Abraham, A., Das, S., & Konar, A. (2006). Document clustering using differential evolution. In *2006 IEEE international conference on evolutionary computation, July* (pp. 1784–1791). IEEE.

Ailem, M., Role, F., & Nadif, M. (2015). Co-clustering document-term matrices by direct maximization of graph modularity. In *Proceedings of the 24th ACM international on conference on information and knowledge management, October* (pp. 1807–1810). ACM.

Ailem, M., Role, F., & Nadif, M. (2017). Sparse poisson latent block model for document clustering. *IEEE Transactions on Knowledge and Data Engineering, 29*(7), 1563–1576.

Bach, F. R., & Jordan, M. I. (2004). Learning spectral clustering. In *Advances in neural information processing systems* (pp. 305–312).

Bao, L., Tang, S., Li, J., Zhang, Y., & Ye, W. P. (2008). Document clustering based on spectral clustering and non-negative matrix factorization. In *International conference on industrial, engineering and other applications of applied intelligent systems, June* (pp. 149–158). Berlin, Heidelberg: Springer.

Bonabeau, E., Marco, D. D. R. D. F., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems*: No. 1. Oxford University Press.

Chan, P. K., Schlag, M. D., & Zien, J. Y. (1994). Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 13*(9), 1088–1096.

Dhillon, I. S., Guan, Y., & Kulis, B. (2004). *A unified view of kernel k-means, spectral clustering and graph cuts*. Computer Science Department, University of Texas at Austin.

Ding, C. H., He, X., Zha, H., Gu, M., & Simon, H. D. (2001). A min-max cut algorithm for graph partitioning and data clustering. In *Proceedings 2001 IEEE international conference on data mining* (pp. 107–114). IEEE.

Foong, O. M., & Yong, S. P. (2016). Swarm LSA-PSO clustering model in text summarization. *International Journal of Advances in Soft Computing and its Applications, 8*(3), 88–99.

Kamvar, K., Sepandar, S., Klein, K., Dan, D., Manning, M., & Christopher, C. (2003). Spectral learning. *International joint conference of artificial intelligence, April*. Stanford InfoLab.

Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied soft computing, 11*(1), 652–657.

Karypis, M. S. G., Kumar, V., & Steinbach, M. (2000). A comparison of document clustering techniques. *TextMining workshop at KDD2000 (May 2000), August*.

Kumar, A., & Daumé, H. (2011). A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 393–400).

Menéndez, H. D., & Camacho, D. (2015). GANY: A genetic spectral-based clustering algorithm for large data analysis. In *2015 IEEE congress on evolutionary computation (CEC), May* (pp. 640–647). IEEE.

Mohammed, A. J., Yusof, Y., & Husni, H. (2015a). Document clustering based on firefly algorithm. *Journal of Computer Science, 11*(3), 453 (5).

Mohammed, A. J., Yusof, Y., & Husni, H. (2015b). Determining number of clusters using firefly algorithm with cluster merging for text clustering. In *International visual informatics conference, November* (pp. 14–24). Cham: Springer.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems, 14*, 849–856.

Oliveira, S., & Seok, S. C. (2005). A multi-level approach for document clustering. In *International conference on computational science, May* (pp. 204–211). Springer.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Shahnaz, F., Berry, M. W., Pauca, V. P., & Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing Management, 42*(2), 373–386.

Shi, J., & Malik, J. (2000). *Normalized cuts and image segmentation Departmental Papers (CIS)*, 107.

Wang, L. G., Shi, Q. H., & Hong, Y. (2010). Hybrid optimization algorithm of PSO and AFSA. *Computer engineering, 36*(5), 176–178.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics, 17*(10), 977–987.

Yogatama, D., & Tanaka-Ishii, K. (2009). Multilingual spectral clustering using document similarity propagation. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 2-volume 2, August* (pp. 871–879). Association for Computational Linguistics.