Asian Journal of Information Technology 15 (22): 4512-4521, 2016
ISSN: 1682-3915

Hybridization of K-Means and Harmony Search Based on Optimized Kernel Matrix and Unsupervised Constraints

[1]S. Siamala Devi and [2]A. Shanmugam
[1]Department of Computer Science and Engineering, Sri Krishna College of
Technology, Coimbatore, Tamil Nadu, India
[2]Department of Electronics and Communication Engineering, SNS College
of Technology, Coimbatore, Tamil Nadu, India

Abstract: Clustering is one of the effective techniques that separate the data into meaningful groups. Feature selection is an important concept to enhance efficiency in clustering process. Existing work presented a method called hybridization of K-means algorithm and Harmony Search Method (HSM) for clustering the documents. In this method, concept factorization is used to extract the meanings to cluster the documents. But it needs to improve clustering accuracy in the document clustering process. In this manuscript, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. Spherical kernel is utilized as the higher order kernel that is higher rate of computation. Furthermore, the weight of each concept is calculated and select as the weighted features. The problem in this method is poor generalization performance so it needs to select optimal kernel matrix. So, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) is presented to select the optimal value of kernel matrix. In this method, kernel set is to chosen accurately to improve clustering performance but the accuracy is less. Furthermore Unsupervised Constrained based Hybrid Clustering (UC-HC) to improve the clustering performance. In this method, data are extracted by identifying an assignment that rises similarity score between strings and informs to the constraints. Experimental result compares methods such as KWC, PSO-OKMS and UC-HC to measure the clustering accuracy. The proposed UC-HC method shows high accuracy when compared to KWC and PSO-OKMS methods.

Key words: Document clustering, K-means algorithm, harmony search, kernel function, particle swarm optimization

INTRODUCTION

Due to the enormous growth in the world wide web, many of the research areas are focused on how the information is organized in a way that will make it easier to the users to identify the information accurately in the web (Hammouda and Kamel, 2004). In the web, the information is presented as the text documents and some of the web document processing systems depends on the text data mining methods. In the text mining, clustering methods are used to find the similar text documents into groups so that achieve high intra-cluster similarity and low inter-cluster similarity. The text document clustering aims to separate the documents into clusters in which every cluster denotes some topic that is disparate from other clusters. If the text mining is applied to the web domain it is called as web mining. Generally, there are three types of web mining web structure mining, web usage mining and web content mining. The applications of document clustering are: Retrieval of clustered documents is comprehensible to the users and effectual retrieval of information by considering the relevant subsets rather than the whole documents.

Document clustering has also been used to automatically make hierarchical clusters of documents. The best known method in partitioning clustering is K-means algorithm. Although K-means algorithm is straightforward, uncomplicated and easy to develop it suffers from some major disadvantages that make it unsuitable for many applications. K-means algorithm is simple, easy to develop and successfully used in different applications. In the K-means method, the number of clusters should be denoted at the first step. But the problem in the K-means method is there is restriction in producing local optimal solution. Harmony Search Method (HSM) is a new optimization method that imitates

Corresponding Author: S. Siamala Devi, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India

the music invention process. This method is utilized in different applications for an optimization problem. The integration of K-means and HSM provides better results. Generally, in the clustering methods term frequency and inverse document frequency for a feature is computed and based on this clustering is performed. But, the drawback in this method is only terms are considered. So, the concept factorization method is used to cluster the documents so that the optimal clusters are identified in possible amount of time. But it needs to improve the clustering accuracy.

In this study, Kernel and Weighted feature based Clustering (KWC) method is presented for clustering the documents. Spherical Kernel is used as the higher order kernel that has high computation rate. The spherical kernels are similar to the circular kernel. Both kernels have a linear behavior at the origin which is also true for the exponential kernel. In addition to that weighted features are taken based on the weight of the concept and selected the features. The concept is extracted by using the word Net-based clustering method where first the concepts are identified as a set of terms and relationship between the synonyms. But the disadvantage in this method is poor generalization performance so it requires selecting the optimal kernel matrix. So, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO OKMS) is introduced to select the optimal value of kernel matrix. The kernel set is selected effectively to enhance the clustering performance. In the PSO algorithm, the set of particles are considered as the kernel matrix and compute the fitness value. The fitness value is computed by the objective function. The objective function is to minimize the function of Kernel matrix.

Literature review: Guan *et al.* (2012) suggested Nonnegative Matrix Factorization (NMF) that is a powerful matrix decomposition method which approximates a nonnegative matrix by computing the product of two low-rank nonnegative matrix factors. But the problem in this method is slow convergence rate and less stability. So, a novel effective NeNMF method is suggested to concurrently solve the main problems. This method uses the Nesterov's optimal gradient method in which it alternatively optimizes one factor whereas other factor is permanent.

Lian *et al.* (2004) suggested a hierarchical algorithm (S-GRACE) to group the documents according to the structural information in the data. This structure graph is used that is a computationally effectual distance metric defined between the documents and the group of documents. This metric is used for the clustering method that is high efficient when compared to the other methods. Hammouda and Kamel (2009) presented a hierarchically distributed Peer-to-Peer (HP2PC) architecture and clustering method. This method is based on the multilayer overlay network of peer neighborhoods. The super nodes are act like neighborhoods that can be grouped to make the higher level neighborhoods. Skabar and Abdalgader (2013) presented the novel fuzzy clustering algorithm which utilizes the graph representation of the data and operates on the expectation-maximization framework.

Shehata *et al.* (2010) proposed concept-based mining model for finding the significant matching concepts between documents based on the semantics of their sentences. The similarity between the documents is also calculated based on the concept-based similarity measure. Gu presented the new semisupervised spectral clustering method for clustering over the LC similarities including with must-link and cannot-link constraints. Trappey *et al.* (2009) proposed the fuzzy ontological knowledge document clustering for matching the suitable document clusters for the given patents based on their derived ontological semantic webs. Cruz and Hruschka (2013) presented an approach based on document clustering algorithms to forensic analysis of computer seized in police investigations. In addition, two relative validity indexes are utilized for automatically estimating the number of clusters.

MATERIALS AND METHODS

Hybridization of K-means and harmony search based on concept based, kernel and weighted feature based clustering

Concept based weight feature selection: A document is denoted as a feature vector d = $(tf_{t1,}...tf_{ti})$ in which tf_t represents the accurate frequency of the term t in the document t∈D where D represents the set of documents. The word net based clustering method is presented in which the concepts are identified in the documents that have identity or synonym association. Then, the concept frequencies are computed as follows:

$$Cf_c = \sum tf_m \tag{2}$$

$$t_m \in r(c) \tag{3}$$

In this equation, r(C) denotes the set of different terms for the document that belongs to the concept C. If the three terms t_1, t_2, t_3 have the term frequencies tf_{t1}, tf_{t2}, tft_3 correspondingly and the terms have similar meaning and corresponds to the concept then, $Cf_{c1} = tf_{t1}+tf_{t2}+tf_{t3}$.

The word net provides the list of synonyms according to a term. The ordering is based on the most frequently used terms and it has shown that using the first synset as recognized concept for a term so that the clustering performance. The weight of each concept C in document d is computed as:

$$W_c = Cf_c \times idf_c \qquad (3)$$

In this equation, idf_c represents the inverted document frequency of concept C by computing the how many documents in which the concept emerges. Finally, a document d is denoted as a vector of weights:

$$d = \left(Wc_1,...,Wc_i\right) \qquad (4)$$

Most of the document clustering methods uses the similarity measure between the documents to group the similar documents. The cosine similarity measure is most frequently used similarity measure for the clustering process:

$$\text{Similarity}\left(d_1,d_2\right) = \text{Cosin e}\left(d_1,d_2\right) = \frac{\left(d_1,d_2\right)}{\left\|d_1\right\|.\left\|d_2\right\|} \qquad (5)$$

In this equation, denotes the vector dot product and || denotes the length of a vector.

Spherical Kernel based clustering: Let $X = \{x_1, x_2, ...x_n\}$ is the input dataset that includes n data points in which $x_i \in \Re^d$, C denotes the number of clusters and $K \in \Re^{n\times n}$ represents the kernel matrix with $K_{ij} = k(x_i,x_j)$:

$$K(x_i,x_j) = 1 - \frac{\frac{3}{2\left\|x_i - x_j\right|}}{\sigma} + \qquad (6)$$
$$1/2(x_i - x_j\left|\,\right|/\sigma)$$

In this equation, k(.) represents the spherical kernel function. denotes the constant scalar. Let $\mathcal{H}_k$ denotes the Reproducing Kernel Hilbert Space (RKBS) endowed by the kernel function k(.) and$//_{\mathcal{H}k}$ denotes the functional norm for $\mathcal{H}_k$. The intent of kernel k-means is to reduce the clustering error. The clustering error is defined as the sum of squared distances between the data points and the center of the cluster to which the point is allocated. The kernel k-means problem can be represented as the following optimization problem:

$$\min_{U\in p} \max_{\{Ck(.)\in H_k\}} C_{k=1} \sum_{K=1}^{C} \sum_{j=1}^{n} Uk_i \big| \qquad (7)$$
$$Ck(.) - k(x_i)^2 H_k^2$$

In this equation, $U = (u_1,..u_C)^T$ denotes the cluster membership matrix, $C_k(.)\in H_k$, $k\in[C]$, represents the cluster centers and domain $P = \{U\in\{0, 1\}^{C\times n}$ in which 1 denotes a vector of all. The two normalized versions of U is to introduced. Let $n_k = u^T_k 1$ denotes the number of data points allocated to the kth cluster. The cluster membership matrix is verified as follows:

$$C_k(.) = \sum_{i=1}^{n} \hat{U}_{ki} K(x_i), k \in \left[C\right] \qquad (8)$$

Based on the concept based weighted features and kernel matrix the clustering is performed. The clustering is accomplished by the Hybridization of K-means and Harmony Search method.

Hybridization of k-Means and Harmony search clustering method: In this study the hybridization clustering algorithm is presented. In the hybridization clustering process, the K-means algorithm and Harmony Search (HS) approach is used. The HS algorithm has high power and K-means algorithm has high speed. In this hybrid algorithm there are two modules: HS module and K-means module. The HS module identifies the region of the optimum and K-means take responsible to identify the optimum centroids.

Algorithm 1:

Hybridization of K-Means and Harmony Search based on Concept based, Kernel and weighted feature based Clustering algorithm

Input: Set of Documents

Output: Clustering Process

Given a set of Documents D and document D is denoted as a feature vector $d = (tf_{t1},...tf_{ti})$// tf_t= frequency of the term t in the document

Concept for a term is identified by using the word net tool

The concept frequencies are computed as follows:

$$Cf_c = \sum tf_m;\ t_m \in r(c)$$

// r(c) denotes the set of different terms for the document that belongs to the concept C.

Weight of each concept C in document d is computed as:

$$W_c = Cf_c \times idf_c$$

where, idf_c represents the inverted document frequency of concept C.

Document d is denoted as a vector of weights:

$$d = (Wc_1,...Wc_i)$$

Similarity is identified as:

$$\text{Similarity} = (d1,d2) = \text{Cosin e}(d1,d2) = \frac{(d1.d2)}{\|d1|.|d2|}$$

```
where represents the vector dot product and || denotes the length of a vector.
 // Spherical Kernel matrix
 The kernel matrix is represented as:

   $K(x_i,x_j)=1-\frac{\frac{3}{2|x_i-x_j|}}{\sigma}+1/2(x_i-x_j|\,|/\sigma)$

where, k(.) represents the spherical kernel function. σ denotes the constant scalar:
 Clustering process by hybridization of K-means and HS
 Produce initial clusters
 Run the HS process
  Select the best vector
   Calculate cluster centroids to set as the initial centroid:

   $C_{kj}=\frac{\sum_{i=1}^{n}(\alpha ki)d_{ij}}{\sum_{i=1}^{n}\alpha_{ki}}$

 Select Vectors of K-means
  /*refine the cluster*/
  Run K-means process
  Set A[i][j] to $d_i$ of cluster j
  Return A
```

Optimization of Kernel matrix using PSO: In this study, the kernel matrix is optimized by using the particle swarm optimization algorithm. Let K be the convex set of positive semi definite kernel functions. A standard example is the group of all affine combinations of given positive semi definite kernel functions. K_1, K_2, ...K_p:

$$K=\left\{K:x\times x\to\mathbb{R}\,\middle|\,K=\sum_{i=1}^{p}\theta_iK_i,1^T\theta=1,\theta\geq 0\right. \quad (9)$$

where, 1 denotes the vector of all ones and $\theta\geq 0$ means $\theta_i\geq 0$, i = 1, ...p. Often the kernels K_i are selected to satisfy the normalization constraint. In this method, PSO is used for finding the optimized kernel matrix. In the optimization algorithm, each particle has set of kernel matrices which begin with random initialization of particle's position and velocity. In the swarm, every particle has two specifications: a position that represents the suggested location and a velocity represents the speed of moving. The particle in the swarm negotiates over the complete search space and memorizes the finest position found. The communication is takes place between the particles so that they regulate their locations and velocities based on solutions discovered by others. The position of the particle is scored by the fitness. The fitness is computed by the objective functions. The objective function is to reduce the Clustering Error (CE):

$$\text{Fitness=Min (CE)} \quad (10)$$

The clustering error is computed as:

$$L(U,\xi)=tr(K)+\sum_{k=1}^{C}L_k(U,\xi) \quad (11)$$

In this equation, ξ = 1 in which 1 is a vector of all ones, implies that the entire data points are selected for constructing the subspace H_b that is equivalent to the kernel based clustering by using the full kernel matrix. Based on the fitness value, the particle is quantified as a good solution. During the execution of the PSO algorithm, the best fitness value is considered as the individual best fitness value. Comparing the entire particles in the swarm, the best fitness value is called global fitness value.

Algorithm 2: PSO algorithm for optimization of Kernel matrix:

```
 Initialize N number of particles with set of tasks and allocate the resources randomly, a position of particle is denoted by $X_i$ and velocity is denoted as $V_i$
   P best represents the best well-known position best signifies the best position of the entire swarm
   Particle position is initialized as $X_i$
   For every particle i=1, 2… N
    Compute the fitness value for each particle
     // Fitness computation
  Fitness = Min (CE)
 Clustering error is computed as:

   $L(U,\xi)=tr(K)+\sum_{k=1}^{C}L_k(U,\xi)$

  If the fitness value is higher than the
  Set the present value as the new pBest
  Until a termination criterion is met
 Select the particle with best fitness value of all particles as the gbest
   // Calculation of particle velocity
    Update particle position and velocity
   $x_i(t+1) = x_i(t)+v_i(t+1)$
     Until some stopping condition is met
```

Unsupervised constrained based hybridization of K means and harmony search: In this study, additional semantic constraints are considered for document clustering. Generally the word constraints are considered.

Document constraint: The document constraints are constructed according to the human annotations that are highly complicated to acquire. In this research, new methods are used to derive "good but imperfect" constraints using information routinely mined from either the content of a document.

Word constraints: In this section, the information in word net is leveraged to create the word constraints. Particularly, the semantic distance between the two words is computed according to the association in word net. While, the word must-links are constructed according to the semantic distances. If the distance between two words is less than a threshold, the word must-link is added. Additional lexical information is used to achieve high clustering efficiency in the clustering algorithm. Furthermore while word knowledge can be relocated to the document side during co-clustering with extra word constraints it is probable to enhance the document clustering as well.

Algorithm 3: Unsupervised constrained based hybridization of K-means and harmony search

```
Input: Romanization Table T: $C_s \neg C_t$, constraints C, source words $V_s$ and target words $V_t$:
1. Initialize model
Let $W:C_s \times C_t \neg R$ be a weight vector
Initialize W using T by using the process:

    $\forall (C_s, C_t), (C_s, C_t), \in T \Rightarrow W((C_s, C_t) = 0$
    $\forall (C_s, C_t), (C_s, C_t), \in T \Rightarrow W((C_s, C_t) = -1$
    $\forall (C_s), W(C_s) = -1, \forall (C_t), W(C_t) = -1$

 Constraints for unsupervised learning
while not converged
 $\forall_{vs} \in V_s, v_t \in V_t$, use C $\alpha$nd W to generate a representation $F(v_s, v_t)$
 $\forall_{vs} \in V_s$ find the top ranking transliteration pair by solving $v_t^* = \text{argm}\alpha\text{x}v_t$ score $(F(vs, v_t^*))$
 $D = \{(+, F(v_s, v_t^*)/\forall vs \in V_s\}$
  Clustering process by hybridization of K-means and HS
   Produce initial clusters
  Run the HS
   Select the best vector
  Calculate cluster centroids using and set as the initial centroid:

    $C_{kj} = \frac{\sum_{i=j}^{n} (\alpha_{ki}) d_{ij}}{\sum_{i=1}^{n} \alpha_{ki}}$

 Vectors of K-means
  Run K-means process
 Set A[i][j] $d_i$ is assigned to cluster j
  Return A
```

RESULTS AND DISCUSSION

Experiments were performed on 20 newsgroups (or NG20) and Reuter's data sets. The proposed algorithm is compared with the other challenging algorithms under similar experimental setting. The experimental results of KWC, PSO-OKMS and UC-HCon NG20 data set are obtained when the number of nearest neighbors is set to seven or eight. For Reuter's data set, the number of nearest neighbors used for KWC, PSO-OKMS and UC-HC varies from 3-24. In all experiments, the proposed algorithm performs better than or competitively with other algorithms. The details of experiments can be described as follows.

Reuters: The numerical results are evaluated for the existing and the proposed method. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) is presented to select the optimal value of kernel matrix. In addition to that, Unsupervised Constrained based Hybrid Clustering (UC-HC) is presented to improve the clustering performance. The performance is evaluated in terms of precision, Recall, F-Measure and retrieval accuracy.

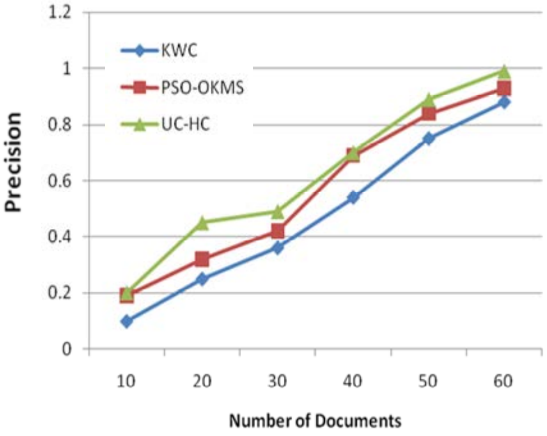

Fig. 1: Precision

Precision: Precision value is evaluated according to the retrieval of documents at true positive prediction, false positive:

$$\text{Precision} = \frac{\begin{matrix}\text{Number of relevent}\\ \text{documents retrieved}\end{matrix}}{\begin{matrix}\text{Total number of}\\ \text{retrieved documents}\end{matrix}}$$

Figure 1 shows the precision rate for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis precision is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is presented to improve the clustering performance. When compared to the existing system there is high precision rate in the proposed UC-HC system.

Table 1 shows the precision rate comparison for the existing and proposed system. If the number of document is 60 the precision rate in the existing KWC method is 0.88 for PSO-OKMS method is 0.93 and for UC-HC method is 0.99.

Table 1: Precision

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.10	0.19	0.20
20	0.25	0.32	0.45
30	0.36	0.42	0.49
40	0.54	0.69	0.70
50	0.75	0.84	0.89
60	0.88	0.93	0.99

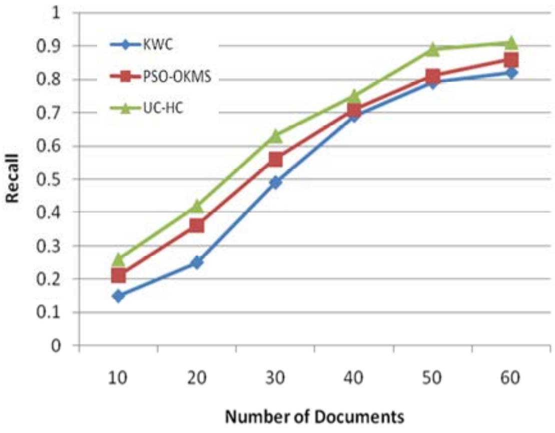

Fig. 2: Recall

Recall: Recall value is evaluated according to the retrieval of documents at true positive prediction, false negative:

$$\text{Recall} = \frac{\text{Number of relevent documents retrieved}}{\text{Total number of retrieved documents}}$$

Figure 2 shows the recall rate for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis recall is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is presented to improve the clustering performance. When compared to the existing system there is high recall rate in the proposed UC-HC system.

Table 2 shows the recall rate comparison for the existing and proposed system. If the number of document is 60, the recall rate in the existing KWC method is 0.82, for PSO-OKMS is 0.86 and for UC-HC method is 0.91.

Table 2: Recall

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.15	0.21	0.26
20	0.25	0.36	0.42
30	0.49	0.56	0.63
40	0.69	0.71	0.75
50	0.79	0.81	0.89
60	0.82	0.86	0.91

Table 3: F-measure

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.10	0.15	0.23
20	0.21	0.35	0.39
30	0.42	0.46	0.57
40	0.52	0.59	0.62
50	0.68	0.74	0.79
60	0.87	0.92	0.99

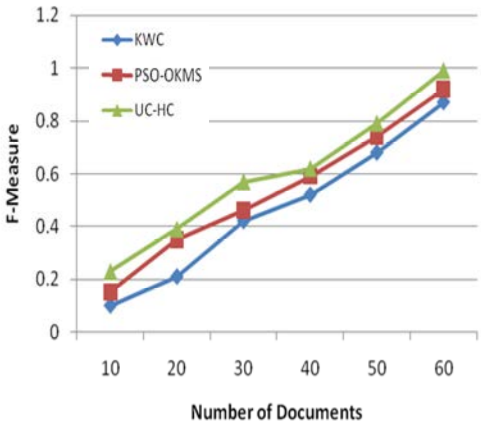

Fig. 3: F-measure

F-measure: F-measure is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct results divided by the number of all returned results and r is the number of correct results divided by the number of results that should have been returned. The F-measure score can be interpreted as a weighted average of the precision and recall where an F_1 score reaches its best value at 1 and worst score at 0:

$$\text{F-measure} = 2.\text{Precision.recall} / (\text{precision} + \text{recall})$$

Figure 3 shows the F-measure for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis F-Measure is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is presented to improve the clustering performance. When compared to the existing system there is high F-measure rate in the proposed UC-HC system.

Table 3 shows the F-Measure comparison for the existing and proposed system. If the number of document is 60, the F-Measure in the existing KWC method is 0.87, for PSO-OKMS method is 0.92 and for UC-HC method is 0.99.

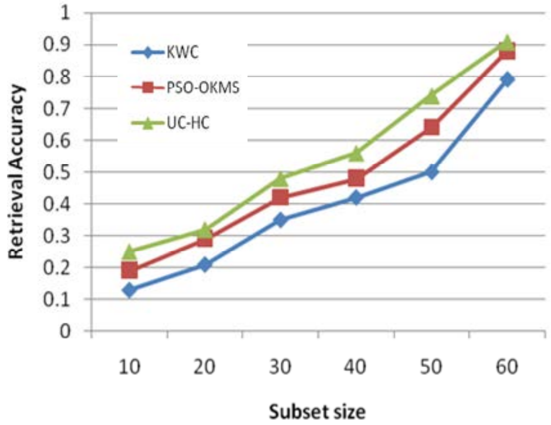

Fig. 4: Retrieval accuracy

Table 4: Retrieval accuracy

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.13	0.19	0.25
20	0.21	0.29	0.32
30	0.35	0.42	0.48
40	0.42	0.48	0.56
50	0.5	0.64	0.74
60	0.79	0.88	0.91

Retrieval accuracy: Retrieval accuracy is defined as the accurate retrieval of documents. Retrieval accuracy is evaluated as:

$$\text{Accuracy} = \frac{(\text{True positivie} + \text{True negative})}{(\text{True positivie} + \text{True negative} + \text{False positive} + \text{False negative})}$$

Figure 4 shows the retrieval accuracy for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis retrieval accuracy is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC)is presented to improve the clustering performance. When compared to the existing system there is high retrieval accuracy in the proposed UC-HC system.

Table 4 shows the retrieval accuracy for the existing and proposed system. If the number of document is 60, the retrieval accuracy in the existing KWC method is 0.87, for PSO-OKMS method is 0.92 and for UC-HC method is 0.99.

ADDC: ADDC is defined as the average distance of documents to the cluster centroid. The equation used for ADDC is as follows:

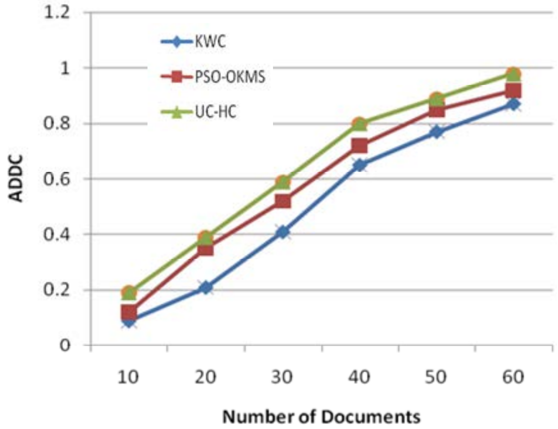

Fig. 5: ADDC

Table 5: ADDC

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.09	0.12	0.19
20	0.21	0.35	0.39
30	0.41	0.52	0.59
40	0.65	0.72	0.8
50	0.77	0.85	0.89
60	0.87	0.92	0.98

$$f = \frac{\sum_{i=1}^{k}\{\frac{\sum_{j=1}^{ni} D(c_i, d_{ij})}{ni}}{K}$$

Where:

K = The number of clusters
n_i = The number of documents in cluster I
D = Distance function
d_{ij} = The jth document of cluster it

Figure 5 shows the ADDC for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis ADDC is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents.

In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is presented to improve the clustering performance. When compared to the existing system, there is high ADDC in the proposed UC-HC system.

Table 5 shows the ADDC for the existing and proposed system. If the number of document is 60, the ADDC in the existing KWC method is 0.87, for PSO-OKMS method is 0.92 and for UC-HC method is 0.98.

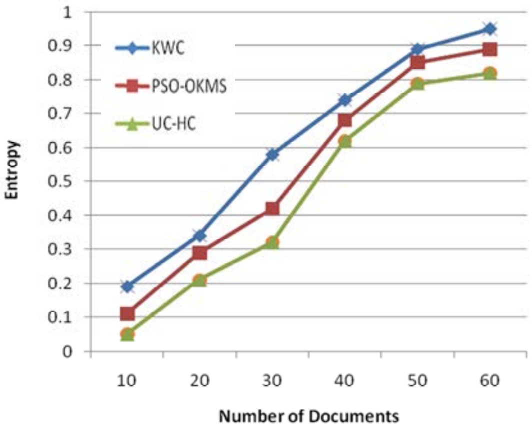

Fig. 6: Entropy

Entropy: The entropy of a cluster can be defined as the degree to which each cluster consists of objects of a single class. The entropy of a cluster j is calculated using the standard equation:

$$e_j = -\sum_{i=1}^{L} p_{ij} \log p_{ij}$$

Where:

L = The number of classes

p_{ij} = The probability that a member of cluster j belongs to class i

The total entropy of the overall clustering result is defined to be the weighted sum of the individual entropy value of each cluster. The total entropy e is defined as equation:

$$e = \sum_{j=1}^{k} \frac{\beta_j}{n} e_j$$

Where:

k = The number of clusters

n = The total number of documents in the corpus

In general, the better clustering result is given by the smaller entropy value. Fig. 6 shows the Entropy for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis Entropy is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is presented to improve the clustering performance. When compared to the existing system there is less entropy in the proposed UC-HC system.

Table 6 shows the Entropy for the existing and proposed system. If the number of document is 60, the retrieval accuracy in the existing KWC method is 0.95, for PSO-OKMS method is 0.89 and for UC-HC method is 0.82.

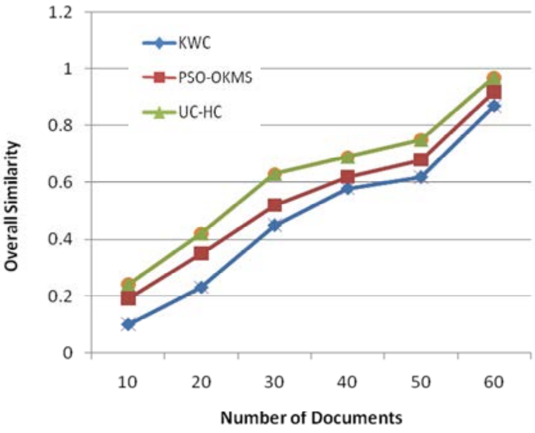

Fig. 7: Overall similarity

Table 6: Entropy

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.19	0.11	0.05
20	0.34	0.29	0.21
30	0.58	0.42	0.32
40	0.74	0.68	0.62
50	0.89	0.85	0.79
60	0.95	0.89	0.82

Table 7: Overall similarity

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.1	0.19	0.24
20	0.23	0.35	0.42
30	0.45	0.52	0.63
40	0.58	0.62	0.69
50	0.62	0.68	0.75
60	0.87	0.92	0.97

Overall similarity: The overall similarity between and is determined by taking average over all the viewpoints not belonging to cluster. Fig. 7 shows the overall similarity for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis overall similarity is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is high Overall Similarity in the proposed UC-HC system.

Table 7 shows the Overall Similarity for the existing and proposed system. If the number of document is 60, the Overall Similarity in the existing KWC method is 0.87, for PSO-OKMS method is 0.92 and for UC-HC method is 0.97.

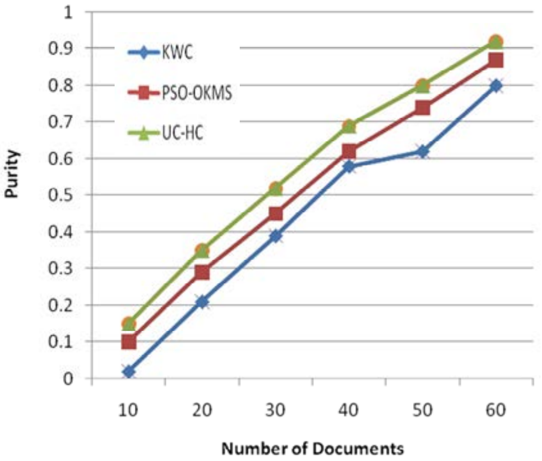

Fig. 8: Purity

Table 8: Purity

No. of documents	KWC	PSO-OKMS	UC-HC
10	0.02	0.1	0.15
20	0.21	0.29	0.35
30	0.39	0.45	0.52
40	0.58	0.62	0.69
50	0.62	0.74	0.8
60	0.81	0.87	0.92

Cluster purity: Purity is a one of very primary validation measure to determine the cluster quality. The concept of purity of the clusters is very important. The purity evaluates the quality of the clusters according to the labeled samples available. A cluster is considered pure if it contains labeled objects from one and only one class. Inversely, a cluster is considered as impure if it contains labeled objects from many different classes. Then, the purity can be defined as equation:

$$\prod_{simple}(C, W) = \frac{1}{N}\sum_{i}^{K} \alpha rgmax(n^{i}{}_{j})$$

This evaluation of the purity consists in estimating the percentage of labeled objects of the majority class in each cluster for all the clustering. It takes its value in [0; 1], 1 indicating that all clusters are pure, i.e., they contain only labeled objects of one class. Fig. 8 shows the purity for the existing and proposed system. In the X-axis subset size is taken. In the Y-axis purity is taken. In the existing method, Kernel and Weighted feature based Clustering (KWC) method is presented to cluster the documents. In the proposed system, Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) and Unsupervised Constrained based Hybrid Clustering (UC-HC) is high purityin the proposed UC-HC system. Table 8 shows the Entropy for the existing and proposed system. If the number of document is 60, the entropy in the existing KWC method is 0.81, for PSO-OKMS method is 0.87 and for UC-HC method is 0.92.

CONCLUSION

In the presented research, Kernel and Weighted feature based Clustering (KWC) method is presented is used to cluster the documents. The disadvantage in this method is poor generalization performance so it needs to select the optimal kernel matrix. So it is necessary to select the optimal kernel matrix. So, the Particle Swarm Optimization (PSO) based Optimal Kernel Matrix Selection (PSO-OKMS) is introduced to choose the best kernel matrix. In addition to that Unsupervised Constrained based Hybrid Clustering (UC-HC) to improve the clustering performance. In this method, features are extracted by recognizing an assignment that increases the similarity score between the two strings and conforms to the constraints. The results show that the proposed UC-HC method shows high clustering accuracy when compared to the KWC and PSO-OKMS methods.

RECOMMENDATIONS

For future work to enhance the clustering accuracy Multi-view with multi-level clustering multi viewpoint is used which is based similarity measure and two related clustering methods.

REFERENCES

Cruz, N.L.F.D. and E.R. Hruschka, 2013. Document clustering for forensic analysis: An approach for improving computer inspection. IEEE. Trans. Inf. Forensics Secur., 8: 46-54.

Guan, N., D. Tao, Z. Luo and B. Yuan, 2012. NeNMF: An optimal gradient method for nonnegative matrix factorization. IEEE. Trans. Signal Process., 60: 2882-2898.

Hammouda, K.M. and M.S. Kamel, 2004. Efficient phrase-based document indexing for web document clustering. IEEE. Trans. Knowl. Data Eng., 16: 1279-1296.

Hammouda, K.M. and M.S. Kamel, 2009. Hierarchically distributed peer-to-peer document clustering and cluster summarization. IEEE. Trans. Knowl. Data Eng., 21: 681-698.

Lian, W., D.W. Cheung, N. Mamoulis and S.M. Yiu, 2004. An efficient and scalable algorithm for clustering XML documents by structure. IEEE Trans. Knowledge Data Eng., 16: 82-96.

Shehata, S., F. Karray and M. Kamel, 2010. An efficient concept-based mining model for enhancing text clustering. IEEE Trans. Knowl. Data Eng., 22: 1360-1371.

Skabar, A. and K. Abdalgader, 2013. Clustering sentence-level text using a novel fuzzy relational clustering algorithm. IEEE Trans. Knowledge Data Eng., 25: 62-75.

Trappey, A.J., C.V. Trappey, F.C. Hsu and D.W. Hsiao, 2009. A fuzzy ontological knowledge document clustering methodology. IEEE. Trans. Syst. Man Cybern. Part B., 39: 806-814.