A Hybrid Algorithm for Document Clustering Using Optimized Kernel Matrix and Unsupervised Constraints

S. Siamala Devi, M. Deva Priya, P. Anitha Rajakumari, R. Kanmani, G. Poorani, S. Padmavathi, and G. Niveditha

Abstract Document clustering plays a dominant role in data mining, and grouping of data makes information retrieval easier. Significant information can be mined from a collection of documents by clustering them effectively. Several researches that concentrate on clustering documents are available in the literature. In the former works, document clustering is performed by using methodologies such as Term Weight-based Hybridized Harmony K-Means (TW-HHKM) and Coverage Factor-based Hybridized Harmony K-Means (CF-HHKM) searches. Clustering is normally performed using K-means algorithm, and cluster centroids are optimally found by using Harmony Search Algorithm (HSA). The main challenge faced by the existing methods is the reduced accuracy as unrelated documents may be grouped together. To overcome this problem, Novel Feature Weighting and Feature Selection-based Hybrid Scheme for Document Clustering (NFW-FS-HSDC) with optimized and unsupervised constraint kernel matrix K-means and Harmony Search Method (HSM) is introduced for accurate clustering of documents. The weights of the data instance and softness parameter decide the performance of spherical kernel K-means clustering. The parameter values for different datasets are identified after many trials. The kernel parameters are determined by applying Particle Swarm Optimization (PSO)-based technique. The proposed NFW-FS-HSDC algorithm

S. Siamala Devi · M. Deva Priya (✉) · G. Poorani · S. Padmavathi · G. Niveditha
Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India
e-mail: s.siamaladevi@skct.edu.in; m.devapriya@skct.edu.in; poorani.g@skct.edu.in; padmavathi.s@skct.edu.in; niveditha.g@skct.edu.in

P. Anitha Rajakumari
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ghaziabad, Uttar Pradesh, India
e-mail: anitharp@srmist.edu.in

R. Kanmani
Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu, India
e-mail: r.kanmani@skct.edu.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

A. Haldorai et al. (eds.), *3rd EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-78750-9_1

optimizes the kernel matrix based on the feature ranges of the dataset by utilizing PSO. The exploratory tests are conducted on Newsgroup and Text REtrieval Conference (TREC) dataset, and it is obvious that the proposed NFW-FS-HSDC offers better Precision, Recall, F-Measure, Average Distance to Document Centroid (ADDC), Entropy, Overall Similarity and Cluster Purity.

Keywords Document clustering · Unsupervised constraints · Kernel matrix · Harmony Search method

1 Introduction

Data mining is a technique used for finding the interesting and hidden patterns from the given data. It plays a dominant role in all the fields, possibly to convert the data into valuable information [1]. It is generally used in a huge variety of profiling practices such as scientific discovery, fraud detection, surveillance and marketing. Data mining is also utilized to find out the patterns in a dataset but is frequently applied only on data samples. Generally, if the samples in the dataset are represented in a better way, then data mining becomes more effective. However, if the samples are not good, then the process becomes ineffective. As more number of unstructured web pages, commercial reports and scientific publications are increasingly available with the Information Retrieval (IR) systems and on the Internet, high-quality document clustering process becomes indispensable and significant in several applications such as web data management, data mining, search engine and even more.

1.1 Tasks of Data Mining

Traditional data mining tasks are divided into four classes:

- **Classification:** It organizes data into predefined groups. For instance, an email program may be classified as legitimate or spam. Commonly, algorithms such as Neural Networks (NN), Naive Bayesian Classification (NBC), Decision Tree Learning (DTL) and nearest neighbour methods are included for many real-time applications.
- **Clustering:** It is a procedure to find similar data and cluster them into groups using similarity measures. But, the groups are not predefined. So, various procedures and parameters are used to group similar data and identify outliers [2].
- **Regression:** It is a method for assessing the relationship among variables. It involves several mechanisms for demonstrating and examining variables with an emphasis on the connection between a variable and one or more independent

variables. This is a technique used to decrease error while organizing data using regression function [3].

- **Association Rule Mining (ARM):** This is a technique used to find the association among objects. For instance, a superstore may gather knowledge on client-buying practices. By applying ARM, the superstore can decide the products that are often bought together. This is often typically remarked as market basket analysis [4].

1.2 Clustering

Clustering is the procedure of considering the similarity that exists among the data to divide them and form groups with identical data objects. It deals with forming a group based on the likeness existing between data items. The data objects within the cluster have the highest similarity with huge deviation among clusters [5]. The inter-cluster and intra-cluster distances are measured using distinctive distance measures. Probably, the most primary variation between clustering and classification process is that clustering is an unsupervised learning system, whereas classification is a supervised one. For clustering, there is no need for a special mechanism to group the data items. Clustering is also known as automatic classification; but it is deceptive as it is fashioned without prior knowledge of classes. However, for classification, the classes of data are predefined. Clustering is a procedure that finds the cluster member based on the qualities and characteristics of data. On the other hand, in classification, classifiers find association among classes in the training dataset. In the dataset, data items are perfectly labelled by a specialist, and the trained classifier learns the relation among data and classes. Thus, the learned models are used to label data.

2 Related Work

Article clustering plays a dominant role in organizing documents in article publication organizations. The issues in arranging data are not clearly perceived and needs an expressive substance spotting of the produced clusters. To overcome the total inadequacy, an intensification of K-means algorithm referred to as Weighted-K-means (W-K-means) is proposed by Tran et al. [6]. W-K-means performs grouping by considering the broader meaning and then produces constructive labels for consequential groups. It offers 10 times better results when compared to the K-means algorithm by yielding more intra-cluster resemblance and less inter-cluster resemblance.

In partitioning-based document clustering algorithm, clustering results mainly depends on the number of clusters that is given as input to the algorithm. If the number of clusters is not known, then the results of clustering will be affected.

Agrawal and Phatak [7] have proposed a novel algorithm which generates quantity of clusters automatically for any unknown textual content dataset and clusters the records adequately based on the cosine similarity that exists between them. The system is capable of generating clusters for unknown data. But, when documents of the same category are to be clustered, there is a problem of over-clustering.

The fundamental task of document clustering is to evaluate the common similarities among the documents. The new method for calculating the likeness between documents is given by Jun et al. [8]. Many features are implanted in this symmetric measure. The dissimilarity between the presence and absence of features is essential rather than the variance that exists between the values related to a current feature. The similarity rises due to the decrease in the change among values. Moreover, the contribution of change is typically scaled. The likeness drops when the amount of presence or absence aspects increases. A feature that is not present does not contribute to the similarity. The propounded measure is applied to estimate the likeness among different units of documents. The measure is applied to a number of textual content applications together with single-labelled classification, multi-labelled classification, K-means and hierarchical agglomerative clustering for which the outcome acquired reveal the efficiency of the propounded similarity measure.

Every document dataset has its own features and choosing the right clustering algorithm that can manipulate all types of clusters is challenging. Clustering algorithms have their unique methods for finding the number of clusters. They may be hybridized to analyse the issues of single algorithms and increase performance. Sadeghian and Nezamabadi-pour [9] have propounded Gravitational Ensemble Clustering (GEC), a clustering approach that utilizes ensemble and gravitational clustering together for efficient grouping in contrast to the individual approaches such as K-Means algorithm and two hierarchical clustering methods namely, Single Link (SL) and Average Link (AL). The gravitational clustering based on Newton's gravitation law and ensemble clustering depends on two functions namely, generation and consensus function. These techniques offer better clustering accuracy, but the major shortcoming is the challenge faced in the optimization of continuous space.

Devi and Shanmugam [10] have proposed an effective ranking approach to rank the documents based on TF-IDF of the keywords in the documents. This reduces the communication overheads as needless documents are not downloaded. Furthermore, an effective query randomization scheme is propounded wherein multiple queries concerning the same search terms seem to be distinctive.

Fan et al. [11] have proposed a method that takes the advantages of Invasive Weed Optimization (IWO) algorithm, low complexity of K-means and hybridizing IWO with K-means. Different data experiments are carried out with various documents in different categories. The results show that IWO-K-means algorithm offers better clustering accuracy. The system finds optimal clusters but involves more convergence time.

Habibi and Popescu-Belis [12] have proposed a technique that considers an exact type of Just-In-Time (JIT) retrieval mechanism for conversational environments. This helps the users to identify the know-how requirements. Modelling the customer's know-how wishes by means of deriving implicit queries from brief dialog fragments is focused. These queries are established on sets of key terms extracted from the dialog. A clustering process is proposed to divide the set of keywords into smaller subsets independent of topics, thus constituting implicit queries. To improve performance of supervisory information clustering, a Semi-Supervised Concept Factorization (SSCF) is propounded by Lu et al. [13]. Pairwise constraints are incorporated into CF as reward and penalty terms. These terms assure that the data points in a cluster in the original space remain in the same cluster even in the transformed space.

Abualigah et al. [14] have propounded a scheme for feature selection called Feature Selection method by using PSO algorithm for Text Clustering (FSPSOTC) that generates a new subcategory of informative text features. This subgroup of features offers better performance involving less computational time. Investigations are performed for six standard text datasets with numerous features. These datasets are usually used in text clustering. The results reveal that this method enhances the efficiency of the text clustering scheme by dealing with a new category of informative features.

Handa et al. [15] have proposed a method that considers the association between keywords to cluster documents. The documents within the appropriate cluster are searched instead of the whole dataset.

2.1 Clustering Methods

The hybridization of K-means and HSM algorithm offers better clustering of document. Forsati et al. [16] have used TF-IDF as a feature for document clustering, that is, term-weighting scheme based on TF-IDF is used. In many documents, certain terms may appear just one or two times. Hence, the TF of words cannot bring out the actual words which lead to creation of tiny clusters in a document cluster. To overcome this, Coverage Factor (CF) feature is used in the proposed Novel Feature Weighting and Feature Selection-based Hybrid Scheme for Document Clustering (NFW-FS-HSDC).

Probably, K-means is the most widely used partitioning-based clustering algorithm which is appropriate for enormous datasets. It is self-effacing, simple, easy to implement and can be applied to a wide variety of applications. In this algorithm, the number of clusters must be specified in the preliminary step itself. It faces a challenge in producing local solution due to the selection of initial cluster centre randomly. Devi and Shanmugam [17] have proposed a Harmonic Search Method (HSM) which is a new meta-heuristic optimization procedure that emulates the tune improvisation process. It has been successful in handling the optimization crisis. In the hybridizing HSM with K-means, HSM is efficient in finding the best initial cluster centre for K-means to provide better results.

3 Hybridization of K-Means and Harmony Search Method (HSM)

K-Means algorithm when applied to document clustering always performs local search. The result or solution is always dependent on the previous step. The K-means algorithm uses arbitrarily selected data as centroid of every cluster and these values change in every iteration. The solution depends on the initial arbitrarily produced centroid values, whereas HSM is excellent in discovering the centroid values, but takes extra time to converge. As a way to overcome the above points of K-means and HSM, a hybrid algorithm that mixes both the ideas can offer effective outcome. Hybridization of K-means and HSM will also yield better results when compared to K-means algorithm [18].

The steps involved in HSM are detailed below.

Step 1: (i) Initialize the possible solutions as harmony and algorithm parameters.
(ii) Calculate the fitness function for document clustering using Average Distance of Documents to the Cluster Centroid (ADDC)
Step 2: Initialize the Harmony Memory (HM) with arbitrarily produced centroid using the fitness function of K-means
Step 3: Improve HM by adjusting the selected centroid
Step 4: If the adjusted centroid has better fitness, then update the HM
Step 5: If all the adjusted data of HM have same fitness value, then stop the search

Initially, the parameters of the HS algorithm which include Harmony Memory Measurements (HMMs), Harmony Memory Allowing Rate (HMAR), Pitch Tuning Rate (PTR) and the Quantity of Improvisations (QI) or the stopping criterion are unique. HMM is a reminiscence location wherein the solution vectors (sets of resolution variables) are saved. The HMAR and PTR are involved in the reinforcement of the solution vector. In HM initialization, the HM matrix is made with the arbitrarily produced resolution vectors similar to the HMM. Generating a new harmony is known as ‘improvisation’.

3.1 Term Weight-Based Hybrid Scheme for Document Clustering (TW-HSDC)

In Term Weight-based Hybrid Scheme for Document Clustering (TW-HSDC), TF-IDF is used as a feature for document clustering. TF denotes the number of times a distinctive term appears in the document. Vector Space Model (VSM) is commonly utilized in document clustering. In the VSM, each report is represented using a feature vector. Traditionally, each characteristic resembles a key word or phrase that appears within the collection of documents. Every entry in the vector stores

Algorithm1: Hybridization of K-means and HSM

```
Input:
D = $\{d_1, d_2, \dots \dots d_n\}$ // set of n documents
$\{f_1, f_2, \dots \dots f_m\}$ - Features of a document
K - Number of clusters
HMM - Harmony Memory Measurement
HMAR - Harmony Memory Allowing Rate
$PTR_{min}$ - Minimum Pitch Tuning Rate
$PTR_{max}$ - Maximum Pitch Tuning Rate
MI - Maximum Number of Improvisation
NHM - New Harmony Memory
Output:
Clustered documents
// Find the initial centroid for clustering documents using HSM
// Cluster documents using K-means Clustering

HSM:
  Initialize HM with random solutions
  Assess the fitness of solutions in HM
  Improvise NHM as follows:
  for each i $\in$[1,N] do
      if U(0, 1) $\in$ HMAR then
          $x_i$= $x^j_i$ where j ~ U (1,…, HMM)
          if(U(0, 1) $\leq$ $PTR_t$) then
              $x_i$=$x_k^{best}$ , where best is the index of the harmony in HM
          else
              Perform arbitrary selection
          end if
      end if
  end for
  Assess the fitness of new solutions in NHM
  Update HM with new solutions until maximum improvisation is attained
  Choose a solution from the HM that has best fitness
  Find the cluster centroids for the newly generated solution

K-means
Input:
  D = $\{d_1, d_2, \dots \dots d_n\}$ // Set of documents
  C = $\{c_1, c_2, \dots \dots c_n\}$ // Set of centroids
  Calculate the distance between the data point and the cluster centers
  for each data point '$d_i$', find the closest centroid '$c_j$' and assign '$d_i$' to cluster '$c_j$'
      Set ClusterID[i]=j; // j - ID of the closest cluster
      Set NearestDist[i]=d($d_i$, $c_j$)
      for each cluster recalculate the centroid
          Repeat
              if each data point '$d_i$' enters a cluster
                  Compute the distance between the data points and the centroid of the closest cluster
                  Place the data point in a cluster when the distance between the centroid of a cluster and data point is
                  less than the distance between any other cluster center
              else
                  for every centroid
                      Estimate the distance d($d_i$, $c_j$)
                      Allocate '$d_i$' to the cluster with adjoining centroid '$c_j$'
                      Set ClusterID[$d_i$]=j;
                      Set NearestDist[$d_i$]= d($d_i$, $c_j$);
                  end for
              end if
              Re-compute the centroids
          Until there is no cluster updations
      end for
  end for
```

the numerical weight for the respective feature of the document. The report is represented by using a vector of weights with ‘n’ elements.

$$d_i = \left(w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots . w_{in}\right) \tag{1}$$

where

w_{ij} – frequency of feature ‘j’ in document ‘i’
n – number of features

It is a term-weighting scheme designed for know-how retrieval (as ranking in search engines like Google) IS used to find the right use of features in record classification and clustering. Weight of a characteristic can be found using the ensuing formula

$$w_{ij} = \mathrm{TF} - \mathrm{IDF} = \mathrm{TF}\,(i, j)\,.\log \frac{N}{df(j)} \tag{2}$$

where

TF(i, j) – frequency of feature ‘j’ in a document ‘d_i’
N – number of documents in the collection
$df(j)$ – number of records in which feature ‘j’ appears

The algorithm TW-HSDC uses TF of features to cluster the documents. As a result, only low-level clusters are formed. The TF-IDF model does not consider words which appear once or twice in the document. There are chances for the words to be used for defining the category of documents. The significance of these words cannot be identified in TF.

3.2 Coverage Factor-Based Hybrid Scheme for Document Clustering (CF-HSDC)

In Coverage Factor-based Hybrid Scheme for Document Clustering (CF-HSDC), the hybridization of K-means and HSM uses the Coverage Factor (CF) as a feature for document clustering. The drawbacks of the TF-based document clustering are eliminated by using the CF. The coverage of the features is outlined as the percentage of files containing a minimum possibility of an important term.

The CF feature considers the significance of terms. CF-HSDC calculates the document frequency of words and performs word selection based on the threshold values. The CF focusses on the importance of less frequent terms by considering the percentage of documents in the dataset containing these words. The coverage of terms is calculated based on the number of input documents taken for clustering and the number of clusters. Based on the number of clusters and percentage of term availability, the lower and upper ranges of coverage are automatically decided.

```
Algorithm 2: Coverage Factor
    Step 1: Arbitrarily choose a subgroup of documents of size ‘m’ from the whole set
            of documents
    Step 2: Mine the terms that appear atleast once in the documents
    Step 3: Eradicate stop words and join words with the same root using stemming
    Step 4: Find the Document Frequency (DF) of words that are mined in the former
            step
    Step 5: Set lower=K and upper=K
    Step 6: Choose words with DF that are in the range ‘lower’ to ‘upper’
    Step 7: If the word coverage is more than the pre-defined threshold, then Stop.
            Else Set lower=lower-1 and upper=upper+1 and goto previous step
```

Here ‘K’ is the cluster size. The number of clusters is almost equal to ‘N/K’. As the document set is typically huge, it is ineffective to perform feature extraction on the whole set of documents. To extract good features, a set of sample documents is randomly selected. A stop word list is generated which includes the meaningless words to be removed from the documents. Stemming is applied to combine similar words (Example: Learning, Learned, Learns can be represented as Learn). Since feature reduction leads to a decrease of clustering time, steps 4 to 6 try to minimalize the number of features and get sensible attention for the features. For example, the user wants the resultant cluster to comprise of ‘K’ documents. Initially, features with DF equal to ‘K’ (Step 4) are chosen. The range {lower, upper} is increased recurrently in step 6. This is to ensure sufficient coverage for the resulting feature set. The exact number of clusters is not equal to ‘N/K’. The proposed method makes use of a coverage threshold to guarantee that the chosen feature has adequate coverage of all important terms that occur less frequently.

4 Novel Feature Weighting and Feature Selection-Based Hybrid Scheme for Document Clustering (NFW-FS-HSDC)

The proposed Novel Feature Weighting and Feature Selection-based Hybrid Scheme for Document Clustering (NFW-FS-HSDC) includes the following steps.

4.1 Optimized Kernel Matrix K-Means Algorithm

In this method, the parameters including weight, softness, number of clusters of the general kernel function are tuned using PSO to improve discrimination in the higher dimensional space. Thus, optimized kernel is formed for improving the accuracy of document clustering. In this method, PSO is used for finding the optimized kernel matrix by identifying optimal weight and the membership values of each object. In every iteration, each particle has a set of kernel matrices that are derived from data objects along with the updated weight and membership matrix. The optimized kernels are very useful for calculating the similarity among heterogeneous features. An objective function '$J(P)$' for fine-tuning the parameters of vector 'P' is defined using PSO:

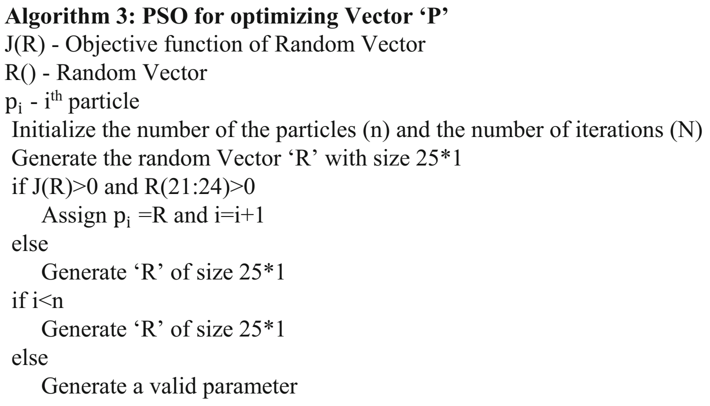
$$J(P) = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n_i} d\,(j, i, i)}{\sum_{i=1}^{c} \sum_{j=1, j \neq i}^{c} \sum_{k=1}^{n_i} d\,(k, i, j)} \tag{3}$$

$d(j, i, i)$ – Distance between successive ith vector in jth cluster
$d(k, i, j)$ – Distance between ith vector in jth cluster in total number of clusters

The distance between the ith vector in the jth cluster (u_{ij}) with the centroid of the qth cluster in the higher dimensional space is given by

$$d\,(i, j, q) = k\left(u_{ij}, u_{ij}\right) - \left(\frac{2}{n_q}\right) \sum_{k=1}^{n_q} k\left(u_{ij}, u_{kq}\right) + \left(\frac{1}{n_q^2}\right) \sum_{k=1}^{n_q} \sum_{l=1}^{n_q} k\left(u_{kq}, u_{lq}\right) \tag{4}$$

$d(i, j, q)$ – Distance between the ith vector in the jth cluster (u_{ij}) with the centroid of the qth cluster
$k(u_{ij}, u_{ij})$ – ith vector in jth cluster
u_{ij} – ith vector in jth cluster
n_q – Centroid of cluster 'q'
$k(u_{ij}, u_{kq})$ – ith vector in jth cluster; kth vector in qth cluster
n_q^2 – Centroid of cluster 'q'
$k(u_{kq}, u_{lq})$ – kth vector in qth cluster; lth vector in qth cluster



Algorithm 3: PSO for optimizing Vector 'P'

```
J(R) - Objective function of Random Vector
R() - Random Vector
$p_i$ - $i^{th}$ particle
 Initialize the number of the particles (n) and the number of iterations (N)
 Generate the random Vector 'R' with size 25*1
 if J(R)>0 and R(21:24)>0
     Assign $p_i$ =R and i=i+1
 else
     Generate 'R' of size 25*1
 if i<n
     Generate 'R' of size 25*1
 else
     Generate a valid parameter
```

Thus, optimization maximizes the discrimination in the higher dimensional space and increases the accuracy of the clustering process.

4.2 Unsupervised Constraint Kernel Matrix K-Means

Unsupervised constraint clustering is performed by the machine itself. It includes two types of constraints, namely, word and document constraints.

- Word constraints depend on the word classifications learned from auxiliary corpus. The must-links are added if two words are close to each other semantically. Named entity identification and overlapped named entities are used to frame constraints for documents. The constraints are obtained by calculating the semantic distance from synonyms in WordNet.
- In case of document constraints, the must-link constraints are built from the connected entities such as person, location, organization, etc.

5 Results and Discussion

The proposed NFW-FS-HSDC offers better Precision, Recall, F-Measure, Average Distance to Document Centroid (ADDC), Entropy, Overall Similarity and Cluster Purity for Newsgroup and Text REtrieval Conference (TREC) datasets. From these representations, it is evident that the propounded scheme outperforms the existing methodologies in fine-tuned clusters. Hence, the effort and price of using physically labelled constraints are reduced by implementing the unsupervised constraints.

5.1 Performance of the Proposed NFW-FS-HSDC Scheme for Newsgroup Dataset

The clustering results of existing and proposed approaches for Newsgroup dataset are clearly given in Table 1.

For Newsgroup dataset, the proposed NFW-FS-HSDC scheme offers 5.6% and 4.2% better Precision in contrast to the existing TF-HSDC and CF-HSDC schemes. Similarly, the propounded NFW-FS-HSDC scheme offers 5% and 3.7% better Recall when compared to the existing TF-HSDC and CF-HSDC schemes.

Table 1 Results of document clustering for Newsgroup dataset

	Methods		
Parameters	TF-HSDC	CF-HSDC	NFW-FS-HSDC
Precision	90.05	91.23	95.07
Recall	90.39	91.53	94.94
F-Measure	90.22	91.38	95.01
ADDC	73.17	66.73	60.34
Entropy	86.6	84.7	79.33
Overall Similarity	84.37	85.21	92.81
Cluster Purity	86.63	87.92	94.46

Table 2 Results of document clustering for TREC dataset

	Methods		
Parameters	TF-HSDC	CF-HSDC	NFW-FS-HSDC
Precision	88.93	90.41	93.98
Recall	89.27	89.89	94
F-Measure	89.1	90.15	93.99
ADDC	70.81	64.33	57.7
Entropy	82.82	81.03	77.99
Overall Similarity	81.81	86.39	89.88
Cluster Purity	86.74	88.17	92.07

The benchmarked TF-HSDC and CF-HSDC schemes offer 5.3% and 4% reduced F-Measure in contrast to the proposed NFW-FS-HSDC scheme. Similarly, the benchmarked TF-HSDC and CF-HSDC schemes offer 21.3% and 10.6% reduced ADDC when compared to the proposed NFW-FS-HSDC scheme.

The proposed NFW-FS-HSDC scheme offers 9.2% and 6.8% better Entropy in contrast to the existing TF-HSDC and CF-HSDC schemes. The proposed NFW-FS-HSDC scheme offers 10% and 8.9% improved Overall Similarity in contrast to the existing TF-HSDC and CF-HSDC schemes. Similarly, the benchmarked TF-HSDC and CF-HSDC schemes offer 9% and 7.4% reduced Cluster Purity when compared to the proposed NFW-FS-HSDC scheme.

5.2 *Performance of the Proposed NFW-FS-HSDC Scheme for TREC Dataset*

Similarly, the results of existing and proposed approaches for TREC dataset are shown in Table 2.

For TREC dataset, the proposed NFW-FS-HSDC scheme offers 5.7% and 3.9% better Precision in contrast to the existing TF-HSDC and CF-HSDC schemes.

Similarly, the propounded NFW-FS-HSDC scheme offers 5.3% and 4.6% better Recall when compared to the existing TF-HSDC and CF-HSDC schemes. The benchmarked TF-HSDC and CF-HSDC schemes offer 5.5% and 4.3% reduced F-Measure in contrast to the proposed NFW-FS-HSDC scheme. Similarly, the benchmarked TF-HSDC and CF-HSDC schemes offer 22.7% and 11.5% reduced ADDC when compared to the proposed NFW-FS-HSDC scheme.

The proposed NFW-FS-HSDC scheme offers 6.2% and 3.9% better Entropy in contrast to the existing TF-HSDC and CF-HSDC schemes. The proposed NFW-FS-HSDC scheme offers 9.9% and 4% improved Overall Similarity in contrast to the existing TF-HSDC and CF-HSDC schemes. Similarly, the benchmarked TF-HSDC and CF-HSDC schemes offer 6.1% and 4.4% reduced Cluster Purity when compared to the proposed NFW-FS-HSDC scheme.

It is seen that Precision (Fig. 1), Recall (Fig. 2), F-Measure (Fig. 3), ADDC (Fig. 4), Entropy (Fig. 5), Overall Similarity (Fig. 6) and Cluster purity (Fig. 7) for Newsgroup and TREC datasets of the proposed NFW-FS-HSDC scheme are better in contrast to the benchmarked TF-HSDC and CF-HSDC schemes considered for investigation.

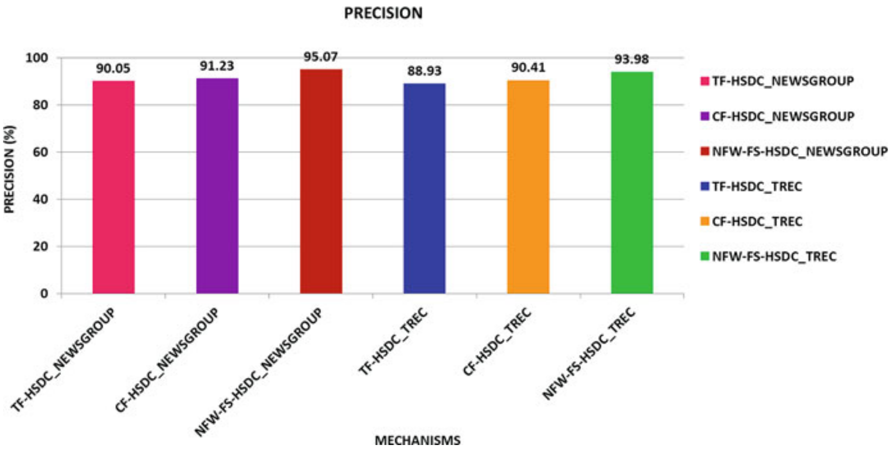


Fig. 1 Precision of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

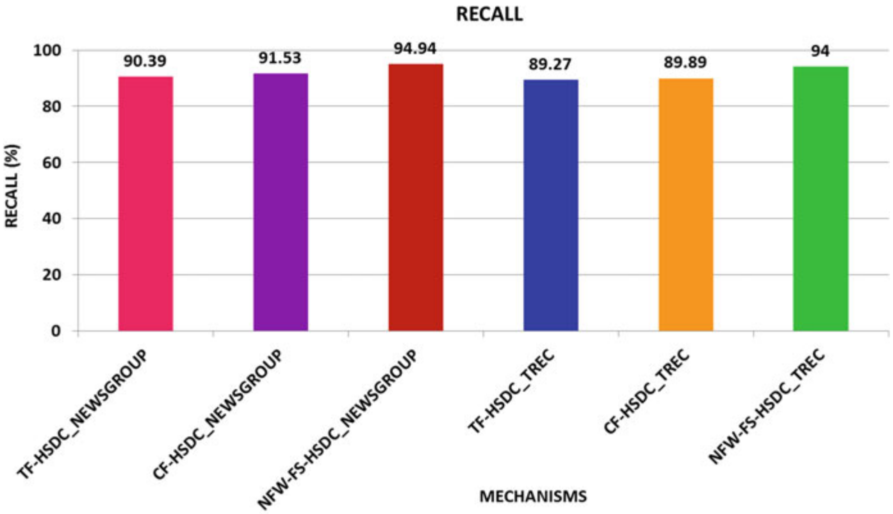


Fig. 2 Recall of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

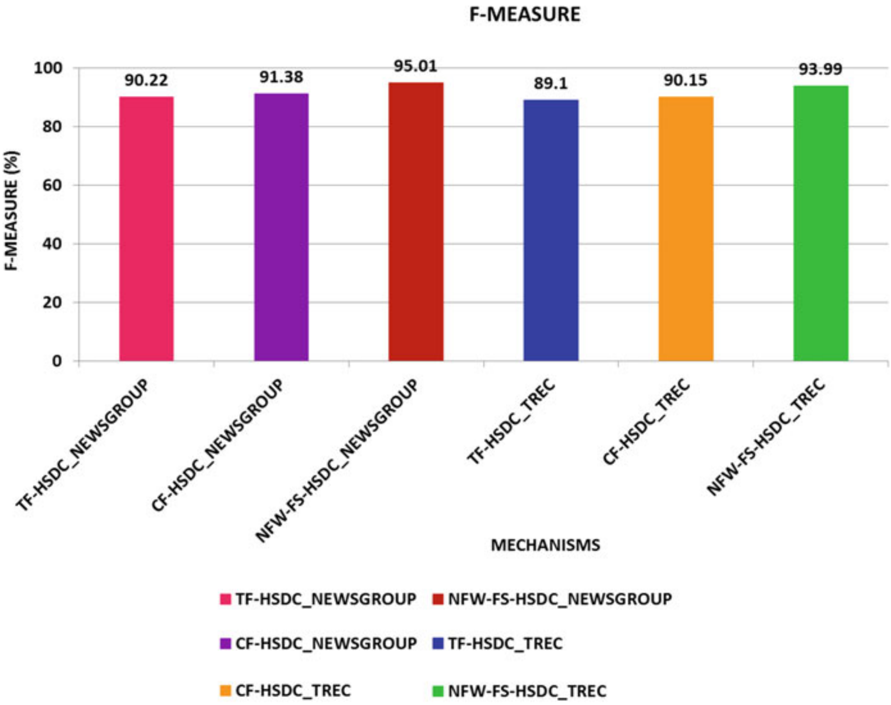


Fig. 3 F-Measure of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

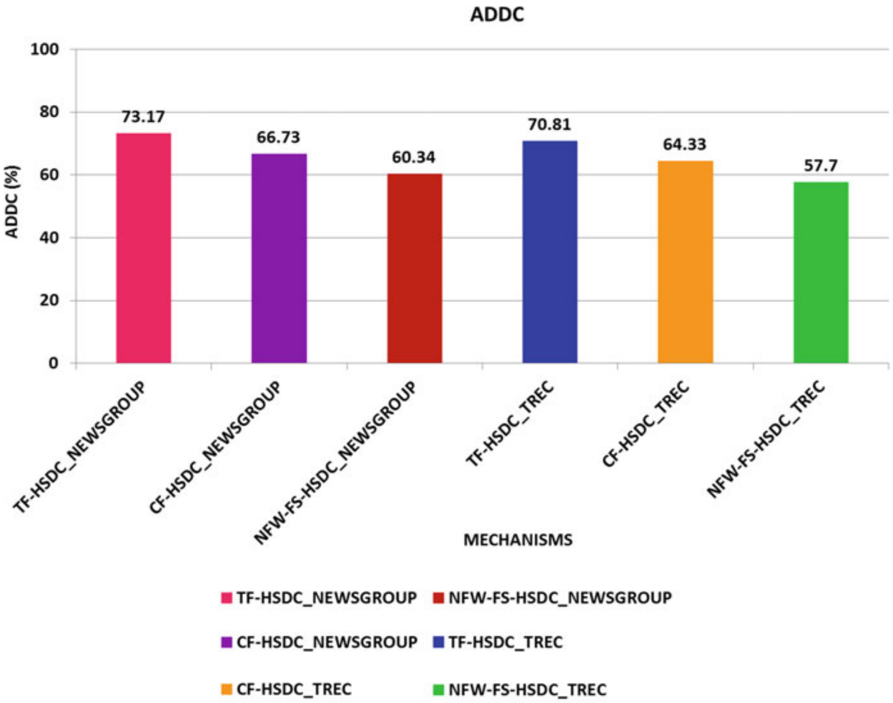


Fig. 4 ADDC of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

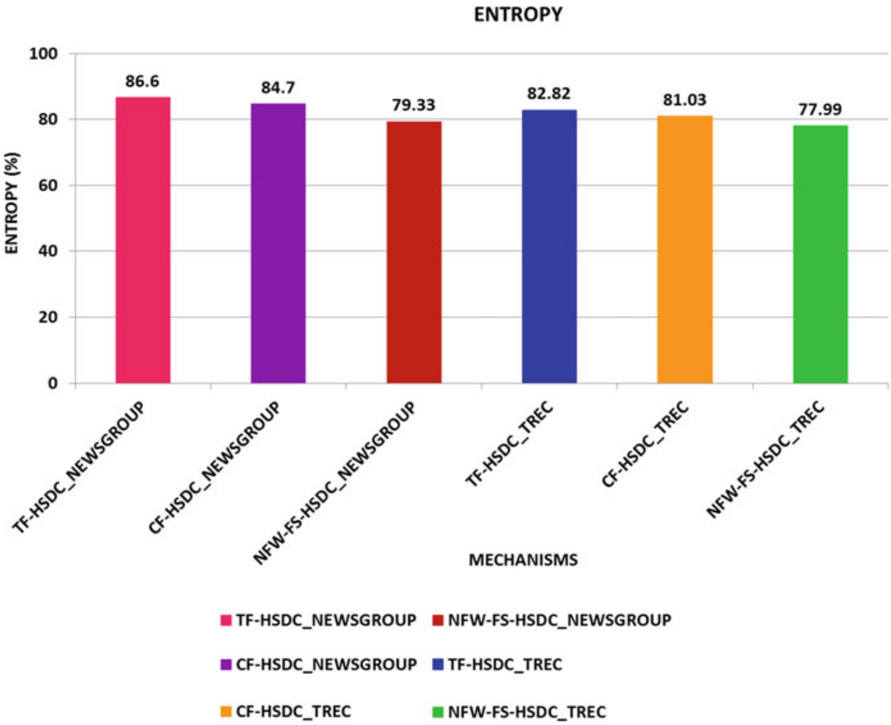


Fig. 5 Entropy of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

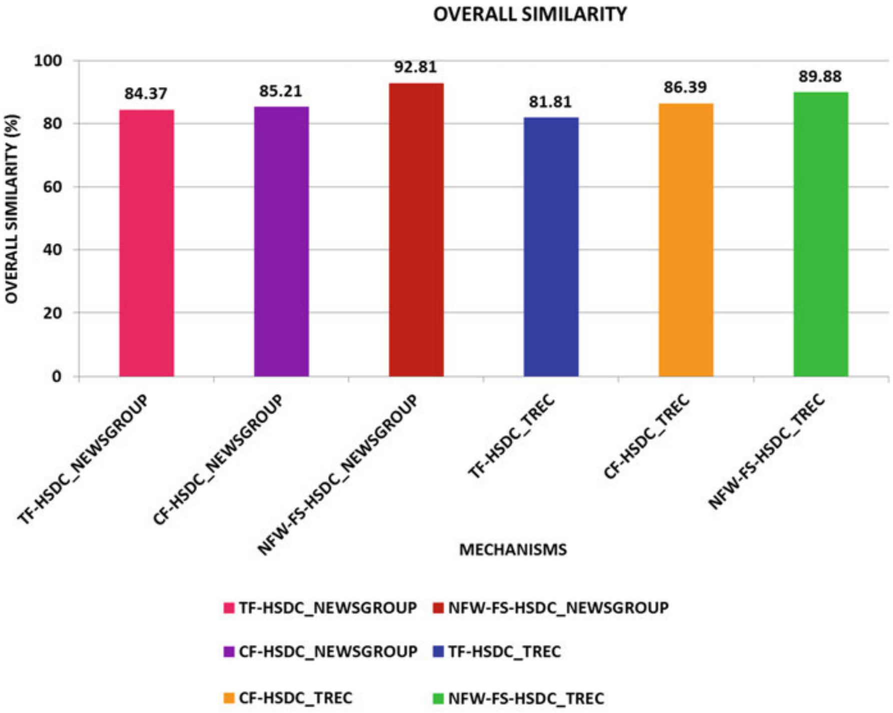


Fig. 6 Overall Similarity of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

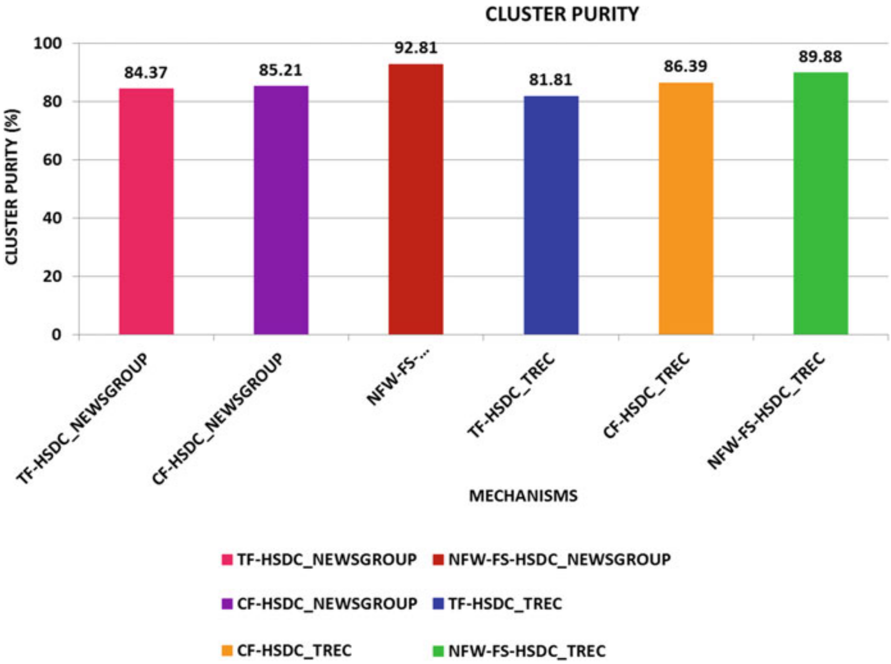


Fig. 7 Cluster Purity of the TF-HSDC, CF-HSDC and NFW-FS-HSDC schemes for the Newsgroup and TREC dataset

6 Conclusion

Different techniques such as Term Frequency-based Hybrid Scheme for Document Clustering (TF-HSDC), Coverage Factor-based Hybrid Scheme for Document Clustering (CF-HSDC) and Novel Feature Weighting and Feature Selection-based Hybrid Scheme for Document Clustering (NFW-FS-HSDC) are proposed in this Chapter to facilitate efficient document clustering. It is important to observe that the K-means algorithm has the limitation of generating local optimal solution, whereas Harmony Search Method (HSM) takes more time to converge. In the TF-HSDC algorithm, clusters are produced based on only TF features count. In CF-HSDC, CF does not guarantee that all documents are considered for the clustering process. In the proposed NFW-FS-HSDC, a single viewpoint is used to cluster documents. The effectiveness of single viewpoint clustering mainly depends on the appropriateness of the similarity measure. The experimental results prove that the NFW-FS-HSDC outperforms the other two clustering methods in terms of Precision, Recall, F-Measure, ADDC, Entropy, Overall Similarity and Cluster Purity for Newsgroup and TREC datasets.

References

1. Rubiano, S. M. M., & Garcia, J. A. D. (2016). Analysis of data mining techniques for constructing a predictive model for academic performance. *IEEE Latin America Transactions, 14*(6), 2783–2788.
2. Aouad, L. M., Le-Khac, N. A., & Kechadi, T. M. (2007). Lightweight clustering technique for distributed data mining applications. In *Industrial conference on data mining* (pp. 120–134). Berlin/Heidelberg: Springer.
3. Gaobo, C., & Xiufang, C. (2011). Combining partial least squares regression and least squares support vector machine for data mining. In *IEEE international conference on E-business and E-government* (pp. 1–4).
4. Bouziri, A., Latiri, C., Gaussier, E., & Belhareth, Y. (2015). Learning query expansion from association rules between terms. In *7th international joint conference on knowledge discovery, knowledge engineering and knowledge management* (Vol. 1, pp. 525–530).
5. Kumar, N., Verma, V., & Saxena, V. (2013). Cluster analysis in data mining using k-means method. *International Journal of Computer Applications, 76*(12), 11–14.
6. Tran, T. N., Wehrens, R., & Buydens, L. M. (2006). KNN-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics & Data Analysis, 51*(2), 513–525.
7. Agrawal, R., & Phatak, M. (2013). A novel algorithm for automatic document clustering. In *3rd IEEE international advance computing conference* (pp. 877–882).
8. Jun, S., Park, S. S., & Jang, D. S. (2014). Document clustering method using dimension reduction and support vector clustering to overcome sparseness. *Expert Systems with Applications, 41*(7), 3204–3212.
9. Sadeghian, A. H., & Nezamabadi-pour, H. (2015). Document clustering using gravitational ensemble clustering. In *International symposium on IEEE artificial intelligence and signal processing* (pp. 240–245).
10. Devi, S. S., & Shanmugam, A. (2015). An integrated harmony search method for text clustering using a constraint based approach. *Indian Journal of Science and Technology, 8*(29), 1–7.
11. Fan, C., Zhang, T., Yang, Z., & Wang, L. (2015, August). A text clustering algorithm hybriding invasive weed optimization with K-means. In *12th IEEE international conference on ubiquitous intelligence and computing and 12th IEEE international conference on autonomic and trusted computing and 15th IEEE international conference on scalable computing and communications and its associated workshops* (pp. 1333–1338).
12. Habibi, M., & Popescu-Belis, A. (2015). Keyword extraction and clustering for document recommendation in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(4), 746–759.
13. Lu, M., Zhao, X. J., Zhang, L., & Li, F. Z. (2016). Semi-supervised concept factorization for document clustering. *Information Sciences, 331*, 86–98.
14. Abualigah, L. M., Khader, A. T., & Hanandeh, E. S. (2018). A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *Journal of Computational Science, 25*, 456–466.
15. Handa, R., Krishna, C. R., & Aggarwal, N. (2019). Document clustering for efficient and secure information retrieval from cloud. *Concurrency and Computation: Practice and Experience, 31*(15), e5127.
16. Forsati, R., Meybodi, M., Mahdavi, M., & Neiat, A. (2008). Hybridization of k-means and harmony search methods for web page clustering. *In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 1*, 329–335.
17. Devi, S. S., & Shanmugam, A. (2014). Hybridization of K-means and harmony search method for text clustering using concept factorization. *International Journal of Advanced Research in Computer Engineering & Technology, 3*(8), 2685–2689.
18. Devi, S. S., & Shanmugam, A. (2016). Hybridized harmony search method for text clustering using concept factorization. *International Journal of Advanced Computer Technology, 2016*, 320–327.