ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/73986, November 2015* ISSN (Online) : 0974-5645

An Integrated Harmony Search Method for Text Clustering using a Constraint based Approach

S. Siamala Devi[1*] and A. Shanmugam[2]

[1]Sri Krishna College of Technology, Coimbatore - 641008, Tamil Nadu, India; siamalamagesh@gmail.com
[2]SNS College of Technology, Coimbatore - 641035, Tamil Nadu, India; dras_bit@yahoo.com

Abstract

Background: An application get a global reach only if it is web based. Such types of applications are found existing in large. Storing and retrieval of information is always challenging task. Retrieving relevant data from high dimensional data is always very significant and complicated as well. Data mining plays a major role in the information retrieval process. **Method:** Grouping of data makes information retrieval easier. Clustering is one of the most important data mining techniques for grouping the data. Document clustering partitions the entire data into number of groups, where the data in each group should have large degree of resemblance. **Findings:** K-means algorithm is one of the most important portioning based algorithms which is easy to implement. Due to its time complexity, K-means can be hybridized with Harmony Search Method(HSM). HSM is a new meta-heuristic optimization method which imitates the music improvisation process. The various methodologies like Term Frequency-Inverse Document Frequency (TF-IDF), Coverage Factor (CF), Concept Factorization, Constrained based clustering have been applied on the same dataset to cluster the documents. A comparison has been made among all the above methodologies and an experimental result shows that constraint based clustering method has produced efficient clusters and it outperforms the other three methods. This constraint based clustering helps the input documents to be clustered in an effective way.

Keywords: Concept Factorization, Constraint Based Clustering, Coverage Factor, Harmony Search Method

1. Introduction

Data mining is a process of extracting the useful data from large amount of data. Now a days the web databases are congested and grouping of data is very essential. In this aspect, data mining has become dominant tool and this has been used in almost every applications such as marketing, scam detection etc. There are large range of data mining techniques such as Association Rule mining, Prediction, Clustering, Classification, Outlier analysis etc.

Document clustering[1] has also been used to routinely produce hierarchical clusters of documents. Now a days, the partitioning clustering methods are well suited for clustering bulky documents. In all terms, clustering is a process of grouping the data objects. Document/Text clustering accepts the input as set of documents. Those input documents are grouped based on its similarity. The documents in same cluster should show evidence of large degree of likeness, whereas, the documents in different clusters should show very little degree of resemblance. The input documents should undergo several preprocessing[2] steps like tokenization, stop word removal and stemming.

Tokenization[3] is a process of splitting a stream of wording up into meaningful elements.

Stop Word Removal[3] is the method of removing words like is, was, the, are, they, etc.

Stemming[3] is a way of reducing the stemmed words to its root word. For example, the words like references, referred, referring can be reduced to its root word 'refer'.

After all these preprocessing[2]steps, a set of words can be acquired and those words are called as features or terms.

* *Author for correspondence*

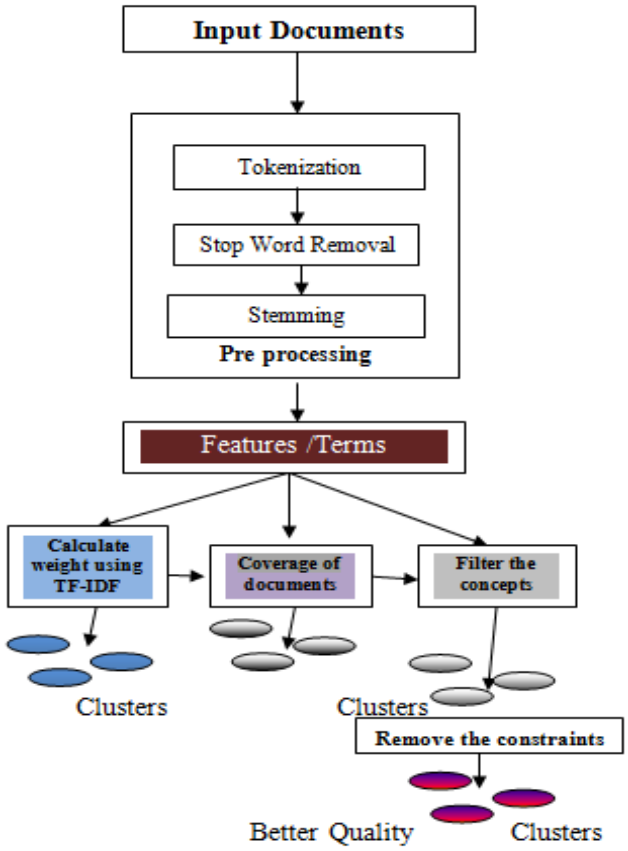

Figure 1. Overview of the clustering methodologies.

K-means algorithm is one of the traditional algorithms that are highly suitable for clustering. But it has disadvantages like local optima problem and more number of iterations. In order to make this algorithm more effective, it can be hybridized with Harmony Search Method[4]. It is a Meta heuristic algorithm. In this, each decision variable generates a value for finding a global optimum solution. Based on the extracted terms, weight of features[4,5] can be calculated. It can be refined by applying the CF[3,6], concept factorization[7] and the set of clusters can be obtained as anoutput. The features can be still filtered based on the concepts and quality clusters can be obtained. Finally, the constraints like document based constraints and word-based constraints[8] are removed and thence high quality clusters can be produced. This hybrid algorithm has been applied on same dataset with different methodologies like Term Frequency-Inverse Document Frequency[9], Coverage Factor, Concept Factorization, Constrained based clustering.

2. Related Works

Automatic text categorization[6] has been extensively used both in the natural language process and in the organization and management of information. This categorization has gained a prominent status in information systems and the data mining field due to the increased availability of documents in digital formats and the ensuing need to organize them. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clusteringwas investigated for improving the precision or recall[4] in information retrieval systems and as an efficient way of finding the nearest neighbours of a document. Targeting useful and relevant information on the World Wide Web is a topical and highly complicated research area.

Clustering techniques have been applied to categorize documents on web and extracting knowledge from the web and this novel clustering algorithms based on Harmony Search (HS)[10,2] optimization method that deals with web document clustering. Today web is over crowded with news articles. So, document clustering is mandatory to organize the data. The problems in organizing the data are synonymy, vagueness and lack of a descriptive content marking of the generated clusters. To overcome all the drawbacks, an enhancement ofstandard K-means algorithm called W-K Means is proposed. W-k means[9] initially enriches the clustering process itself byutilizing hyponyms and then it generates useful labels for the resulting clusters. 10-times improvement had been obtained over the standard k-means algorithm in terms of high intra-cluster similarity and low inter-cluster similarity[11]. Document clustering has also been used to automatically generate hierarchical clusters of documents Harmony Search Method (HSM)[4] is new meta-heuristic optimization methods duplicate the music inventiveness practice where musicians make up their instruments' pitches penetrating[13] for a perfect state of synchronization. Harmony search[12] algorithm had been very victorious in a large array of optimization problems. The weight of the features can be used to cluster the documents.

Automatic web page classification[6] is always considered as the difficult task. Employing some clustering techniques might be a solution for this. Feature selection mechanism,

using tree structure mechanism called DC tree[6] makes the clustering procedure easier and less perceptive to insertion order. Documents are grouped based on a method called Coverage Factor[3]. However, as an optimization algorithm, it easily leads to local optimal clusters[14]. To overcome this Short coming, this paper introduces a hybrid approach of which combines the fuzzy c-means[14,16] and harmony search algorithms for clustering of text document.

3. Algorithms

3.1 K-Means Algorithm[10,11]

It is one of the most important partitioning based algorithms. This algorithm follows very simple and easy steps to cluster the documents. The very significant part is, the number of clusters should be mentioned in the first step itself. After specifying the clusters, a data object will be placed in each cluster. And that data object acts as centroid of the cluster. Euclidean distance[13] measure is used to find the distance between upcoming data object and the centroid value.

If the obtained value is very closer to the centroid value of cluster[16], the data item will be placed in the cluster[1]. Whenever a data object is placed in a cluster, distance measure will be calculated and the centroid values[11] will be keep on changing successive iterations.

The steps of K-means algorithms for clustering is as follows

- Initialize the number of clusters K.
- Place the data objects in K clusters and these objects act as centroids[10].
- Assign the remaining data objects to the K clusters based on the neighboring centroid values.
- Neighboring centroid values can be obtained based on Euclidean distance measure.
- Repeat the step 3 until there is no change in centroid values.

The Euclidean Distance[1,5] is to find the distance between two objects or clusters, taking into account direction and magnitude.

$$dE(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_1 - y_1)^2} \quad (1)$$

where x_i is termed as centroid value and y denoted the data object.

3.2 HSM

HSM is a new, meta-heuristic algorithm[10,12]. It has very simple concept with minimum constraints. It has been applied on many real world problems and optimization problems[13].

The steps of HSM are as follows[10]

Step 1: Initialize the optimization problem and algorithm parameters[12].
Step 2: Initialize the harmony memory (HM).
Step 3: Improvise a new harmony from the HM.
Step 4: Update the HM.
Step 5: Repeat Steps 3 and 4 until the optimized solution is satisfied.

3.3 Hybridization of K-means and HSM

However K-means algorithm performs localized searching whereas HSM performs globalized searching for finding optimal solutions. In [12] on the other side, K-means is very good at finding hopeful areas in the search space but HSM is not as good as K-means. In this regard, K-means is hybridized with HSM to obtain effective results[4,13].

3.3.1 Hybrid Algorithmfor Document Clustering[4]

- /*producing initial clusters*/
  run the harmony search clustering process for the maximum number of iterations.
- Select the best vector from the Harmony memory with highestfitness.
- Calculate cluster centroids and setas the initial centroid vectors of K-means.
- /*refining the clustering*/
  runK-means process until maximum number ofiterations is reached.
- Set A[i][j] to one if the document d_i is assignedto cluster j.
- Return A.

4. Methodologies

The following methodologies were applied on the same hybrid algorithm and the results were compared.

4.1 TF-IDF[3–5]

Documents can be clustered based on the weight of features[5,9] or terms in a document. The weighted mechanism uses TF_IDF. Its expansion is Term Frequency-

-Inverse Document Frequency. This methodology is used to identify the significance level of a word in a document. TF-counts the number of occurrences of a feature in a document[9]. IDF-can be calculated using the log value of total number of documents and document frequency value. The following Figure 2 represents the screen shot shows the weight values of extracted features.

```
mln--0.0
dlr--0.0
ct--0.0
share--1.3862943611198906
reuter--0.0
acq--2.1972245773362196
chemlawn--43.032530597096
chem--41.84538572328660 4
rise--3.2188758248682006
hope--3.9120230054281 46
for--0.0
higher--5.545177444479562
bid--8.317766166719343
chemlawn--43.032530597096
corp--0.6931471805599453
chem--41.84538572328660 4
attract--3.9120230054281 46
higher--5.545177444479562
bid--8.317766166719343
dlr--0.0
per--1.3862943611198906
share--4.158883083359671 5
offer--11.090354888959125
wast--27.38416103799702
```

Figure 2. Weight of the features extracted.

It is calculated as [4,3]

$$Wij=TF\text{-}IDF(i,j) \quad (2)$$

$$TF\text{-}IDF(i,j)= tf(i,j).\left(\log\frac{N}{df(i,j)}\right)^{5} \quad (3)$$

tf (i, j) the number of occurrences of feature j in a document di, N is the total number of documents considered as input, and df (j) is the number of documents where feature j appears.

4.2 Coverage Factor[3]

The exposure of the features is defined as the percentage of the documents containing at least one of the features of the feature extracted[3,6]. This method balances the trade-off between coverage and the number of features used for document representation. Coverage factor is mainly used to get reasonable number of clusters. Because large number of clusters complicates the searching process. On the other hand, minimum number of clusters leads to overhead. This method also makes use of coverage threshold[6] to guarantee that the features selected have ample coverage. Fix the number of clusters as T/k, Where T is total number of clusters and parameter k is used to set an estimated number on cluster size. The steps of CF are as follows[6]:

- Select sample set of documents with size n from the dataset.
- Retrieve the set of words that materialize at least once in a document.
- Calculate the document frequency vale.
- Set the k value as upper and lower bound values.
- Check whether the coverage of the features is larger than threshold value or not.
- If larger, stop the process.
- If not, decrement lower and increment upper by 1.

4.3 Concept Factorization[7,17]

In this model, each cluster is represented as a linear combination of the data points, and each data point as a linear combination of the cluster centers[7].

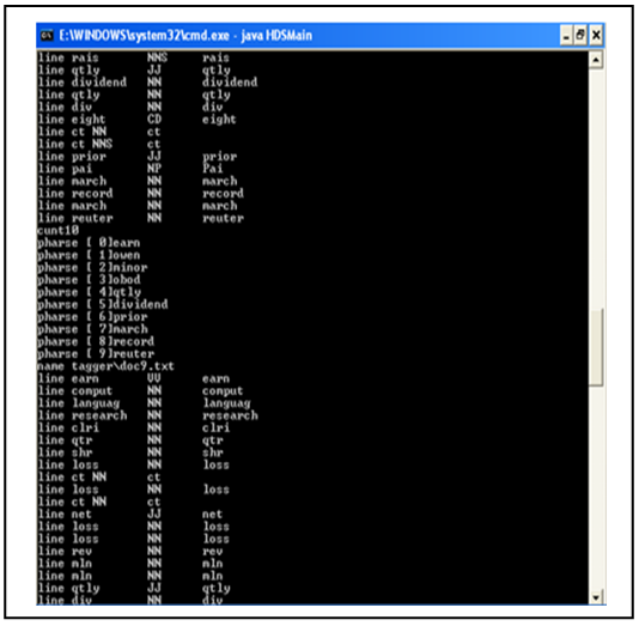

Figure 3. Concept factorization values.

The data clustering is then accomplished by computing the two sets of linear coefficients, which is carried out by finding the nonnegative solution that minimizes the reconstruction error of the data points. That is instead of considering, coverage factor the concepts[17] in the

documents was considered and clustering process takes place based on the concepts in the documents. Figure 3 shows the screen shot of concept factorization process[17] for the given input documents. The words can be categorized as Noun, Adverb, Adjective etc.The core point is that when documents are clustered based on term frequency and CF, the least significant terms might be considered. If the clustering is based on concepts, the clusters with higher relevance can be formed.

4.4 Constraints based Clustering[8]

This type of clustering eliminates both document-based and word-based constraints. When dealing with text clustering, one of the most essential distance measures is document similarity[8]. Since document similarity is often determined by word correspondence, the semantic associations between words may affect document clustering results. Document restraints constructed based on human annotations are difficult to obtain. The semantic distance of two words can be computed based on their relationships. After eliminating unsupervised constraints, better text feature matrix can be obtained[8].

5. Experimental Results

The performance of all these methods are calculated based on Precision[14,15], Recall[14,15], F-Measure[14], Entropy[19], Overall Similarity[18] and Cluster Purity[16].

5.1 Precision[2, 3,15]

Precision refers to the closeness of two or more measurements to each other. The precision values of different clustering methods have been shown in Figure 4.

$$\text{Precision}=\frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}} \quad (4)$$

5.2 Recall[2,3,15]

This is also called as sensitivity. The formula to calculate recall value is

$$\text{Recall}=\frac{\text{Number of relevenant documents retrieved}}{\text{Total number of relevant documents}} \quad (5)$$

The improvement in Recall values has been demonstrated in Figure 5.

5.3 F-Measure[2,3]

Using Precision and Recall, the F-measure value can be calculated. The formula is,

$$\text{F-Measure}=\frac{2.\text{Precision}^{*}\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

The Precision, Recall, F-Measure values of different clustering methods are mentioned in Table 1.

Table 1. Comparison of precision recall and F-measure values of clustering methods

Methods	Precision	Recall	F-Measure
TF-IDF	0.469	0.512	0.489
Coverage Factor	0.51	0.569	0.569
Concept Factorization	0.606	0.630	0.618
Constraint based	0.848	0.882	0.864

The development in Recall values has been demonstrated in Figure 6.

5.4 Entropy[16,19,20]

The Entropy of a cluster can be defined as the degree to which each cluster consists of objects of a single class. The entropy values of different clustering methods are given in Table 2. The entropy of a cluster j is calculated using the standard formula,

$$e_j = -\sum_{i=1}^{L} p_{ij} < ogp_{ij} \quad (7)$$

Where L is the number of classes and p_{ij} denotes the Probability that a member of cluster j belongs to class i. The total entropy[20] 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,

$$e = \sum_{j=1}^{k} \frac{\beta_j}{n} e_j \quad (8)$$

Table 2. Entropy values of clustering methods

Methods	Entropy
TF-IDF	0.865
Coverage Factor	0.623
Concept Factorization	0.549
Constraint based	0.482

Lower the entropy values, better the clustering results.

5.5 Overall Similarity[18]

The overall similarity between and is determined by taking average over all the viewpoints not belonging to cluster. The overall similarity values of different clustering methods are denoted in Table 3.

The formula for calculating overall similarity values is given below:

$$sim(Di,Dj) = \frac{2\sum_{h=1}^{n} d_{ih} d_{jh}}{\sum_{h=1}^{n} d_{ih}^2 + \sum_{h=1}^{n} d_{jh}^2} \quad (9)$$

Table 3. Overall Similarity values of clustering methods

Methods	Overall Similarity
TF-IDF	0.493
Coverage Factor	0.546
Concept Factorization	0.628
Constraint based	0.871

The overall similarity of different clustering methods is shown in Figure 7.

5.6 Cluster Purity[16]

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:

$$\prod_{simple}(C,W) = \frac{1}{N}\sum_{i}^{K} \arg\max_{j}(n_j^k) \quad (10)$$

The cluster purity values of different clustering methods has been mentioned in Table 4.

Table 4. Cluster purity values of clustering methods

Methods	Cluster Purity
TF-IDF	0.508
Coverage Factor	0.552
Concept Factorization	0.67
Constraint based	0.882

The upgradation of different clustering methods are shown in Figure 8.

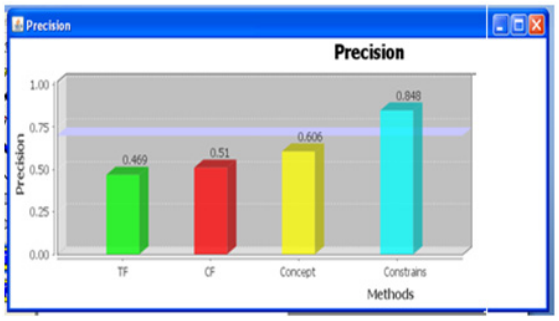

Figure 4. Precision values.

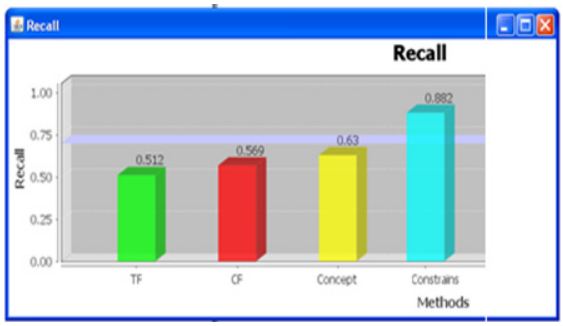

Figure 5. Recall values.

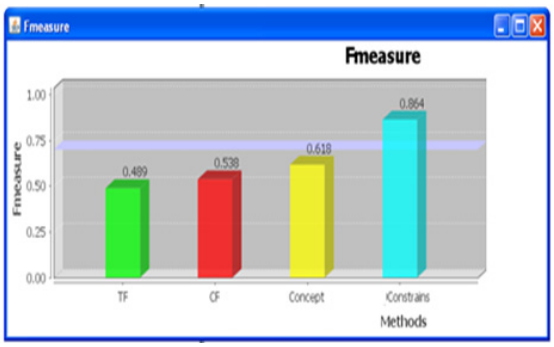

Figure 6. FMeasure values.

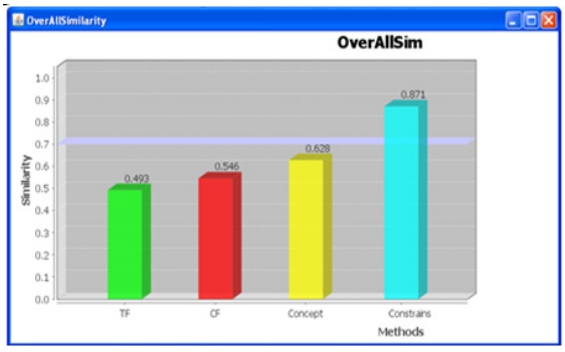

Figure 7. Overall similarity values.

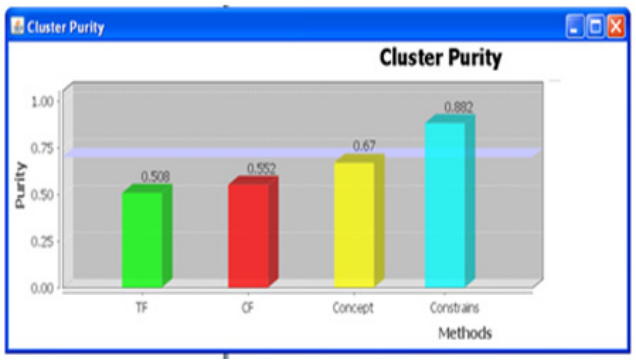

Figure 8. Cluster purity values.

6. Conclusion and Future Work

Obtaining high quality clusters for text clustering is the objective of this paper. K-means is often used algorithm for document clustering. Due to its drawbacks, K-means algorithm is hybridized with Harmony Search Method (HSM). Four different methodologies were applied on the same hybrid algorithm and experimental results proved that constraints based clustering will produce better clusters when compared to other three methodologies considered for this assessment. Nearly, seven different performance measures have been applied to evaluate the cluster quality. It can be further enhanced by considering meaning, hyponyms and hyper-hyponyms of the extracted features.

7. References

1. Abarna R, Pradeepa S. A hybrid approach for extracting web information. Indian Journal of Science and Technology. 2015 Aug; 8(17):1–6.
2. Cobos C, Andrade J, Constain W, Mendoza M, Leon E. Web document clustering based on global-best harmony search, K-means, frequent term sets and Bayesian information criterion. IEEE Congress on Evolutionary Computation;Barcelona. 2010 Jul 18-23. p. 1–8.
3. Devi SD, Shanmugam A. Hybridization of K-means and harmony search method for text clustering using concept factorization. International Journal of Advanced Research in Computer Engineering and Technology. 2014 Aug; 3(8):2685–9.
4. Forsati R, Meybodi M, Mahdavi M, Neiat A. Hybridization of K-means and harmony search methods for web page clustering. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology; Italy. 2009. p. 329–35.
5. Zhang T, Tang YY, Fang B, Xiang Y. Document clustering in correlation similarity measure space. IEEE Transactions on Knowledge and Data Engineering. 2012 Jun; 24(6):1002–13.
6. Wong WC, Fu AWC. Incremental document clustering for web page classification. Department of Science and Engineering. The Chinese University of Hong-Kong; 2002. p. 101–10.
7. Cai D, He X, Han J. Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering. 2011 Jun; 23(6):902–13.
8. Song Y, Pan S, Liu S, Wei F, Zhou MX, Qian W. Constrained text coclustering with supervised and unsupervised constraints. IEEE Transactions on Knowledge and Data Engineering. 2013 Jun; 25(6):1227–39.
9. Li CH, Yang JC, Park SC. Text categorization algorithms using semantic approaches. Corpus-Based Thesaurus and Wordnet. Expert Systems with Applications. 2012 Jan; 39(1):765–72.
10. Cobos C, Leon E, Mendoza M. A harmony search algorithm for clustering with feature selection. Scientific Electronic Library Online. 2010 Sep; 55:153–64.
11. Huang R, Yu G, Wang Z, Zhang J, Shi L. Dirichlet process mixture model for Document clustering with Feature Partition. IEEE Transactions on Knowledge and Data Engineering. 2013 Aug; 25(8):1748–59.
12. Weyland D. A rigorous analysis of the harmony search algorithm - How the research community can be misled by a novel methodology. International Journal of Applied Metaheuristic Computing. 2010 Apr; 1(2):50–60.
13. Forsati R, Mahdavi M, Shamsfard M, Meybodi MR. Efficient stochastic algorithms for document clustering. Information Sciences. 2013 Jan; 220:269–91.
14. Kang J, Zhang W. Combination of Fuzzy C-means and harmony search algorithms for clustering of text document. Journal of Computational Information Systems. 2011; 7(6):5980–6.
15. Nagaraj R, Thiagarasu V. Correlation similarity measure based document clustering with directed ridge regression. Indian Journal of Science and Technology. 2014 May; 7(5):692–7.
16. Skabar A, Abdalgader K. Clustering sentence-level text using a novel fuzzy relational clustering algorithm. IEEE Transactions on Knowledge and Data Engineering. 2013 Jan; 25(1):62–75.
17. Shehata S, Karray F, Kamel MS. An efficient concept-based mining model for enhancing text clustering. IEEE Transactions on Knowledge and Data Engineering. 2010 Oct; 22(10):1360–71.
18. Lin YS, Jiang JY, Lee SJ. Asimilarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering. 2014 Jul; 26(7):1575–90. DOI: 10.1109/tkde.2013.
19. Chung CH, Dai BR. A fragment-based iterative consensus clustering algorithm with a robust similarity. Knowledge Information System. Springer. 2014 Dec; 41(3):591–609.
20. Li T, Ma S, Ogihara M. Entropy-based criterion in categorical clustering. Proceedings of the 21st International Conference on Machine Learning; 2004. p. 68.