ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Hybrid Algorithm for Document Clustrering Using Concept Factorization

Siamala Devi S [*1], Dr. A. Shanmugam[2]

[*1] Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, India

[2] Professor, Department of Electronics and Communication Engineering, SNS College of Technology, Coimbatore, India

Abstract

Massive amount of assorted information is available on the web. Clustering is one of the techniques to deal with huge amount of information. Clustering partitions a data set into groups where data objects in each group should exhibit large measure of resemblance. Objects with high resemblance measure should be placed in a cluster (intra cluster). Resemblance between the objects of different clusters should be less (inter cluster). The most commonly used partitioning-based clustering algorithm, the K-means algorithm, is more suitable for bulky datasets. K-means algorithm is simple, straightforward, easy to implement and works well in many applications. K means algorithm has the limitation of generating local optimal solution. Harmony Search Method (HSM) is a new meta- heuristic optimization method which imitates the music improvisation process. HSM has been a successful technique in a wide variety of optimization problem. Better results can be obtained by hybridizing K-means with HSM. In conventional clustering methods, Term Frequency and Inverse Document Frequency(TF-IDF) of a feature can be calculated and the documents are clustered. In, the projected work an effort has been made to apply the concept factorization method for document clustering problem, to find optimal clusters in sufficient amount of time.

Keywords: Harmony Search Method, Term Frequency, Inverse Document Frequency, Concept factorization.

Introduction

In current situation, unlimited number of webpages, reports , documents are available in the internet. Clustering plays a major role in grouping the documents. In general, Clustering involves dividing a set of documents into a specified number of groups. The documents within each group should have maximum similarity while the similarity among different clusters should be minimized. Some of the more familiar clustering methods are: partitioning algorithms based on dividing entire data into dissimilar groups, hierarchical methods, density and grid based clustering, some graph based methods and etc.

Clustering algorithms can be broadly classified into two categories: hierarchal and partitional algorithms. On the other hand, in recent years the partitioning clustering methods are well suited for clustering a large document dataset due to their relatively low computational requirements. The time complexity of the partitioning techniques is almost linear, which makes them widely used. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a document. More recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a user's query. Document clustering has also been used to automatically generate hierarchical clusters of documents. The best known method in partitioning clustering is K-means algorithm.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Although K-means algorithm is simple, straightforward and easy to implement, it suffers from some major drawbacks that make it inappropriate for many applications.

In general, documents are preprocessed before the clustering process takes place. The pre processing steps are as follows

i. Tokenization is a process of breaking a stream of text up into meaningful elements. This is useful both in linguistics (where it is also known as "Text segmentation"), and in computer science, where it forms part of lexical analysis.

ii. Stop Word Removal is the process of removing the words such as "is, was, a, an, and, the etc..." in a document.

iii. Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. For example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", "fish", and "fisher" to the root word, "fish".

After all the preprocessing steps, we will be obtaining number of words, and those are denoted as features or terms.

Literature Review

Measuring the similarity between the documents is always an important task. Because a feature may appear in more than one document; a feature may appear in only one document; a feature may not appear in any document. A novel similarity measure between two documents is proposed by Yung-Shen et al.,[1]. The presence or absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity degree increases when the numbers of presence-absence feature pairs decreases. The same concept can also be applied to measure the similarity between two document sets.

Taiping et al.,[3] proposed new spectral clustering method called correlation preserving indexing (CPI). This similarity measure is more suitable for detecting the intrinsic geometrical structure of the document space than Euclidean distance. This method maximizes the correlation between the documents in local patches and minimizes correlation between documents outside these patches; this yields good generalization capacity- effectively deal with datas of very large size

Now-a-days web is overcrowded 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 of standard K-means algorithm called W-K Means is proposed by Christos & Vassilis, [2]. W-k means initially enriches the clustering process itself by utilizing hypernyms 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.

In general, text clustering is based on a term, either word or phrase. Term Frequency captures the importance of a feature only within a document. But, sometimes two different terms can have same frequency but one term contributes more to the meaning of its sentence than the other term. To overcome this, new concept- based mining model is proposed by Shady et.al., [5] is used. This includes sentence based concept analysis, document based concept analysis, corpus based concept analysis, concept based analysis algorithm. All these analyzes terms on multilevel i.e., based on sentence, document, corpus level instead of analyzing document only.

A new dynamic method for document clustering is based on Genetic Algorithm (GA) .K-means is greedy algorithm, which is sensitive to the choice of cluster center. Genetic Algorithm is a global convergence algorithm, which can find the best cluster centers easily. This concept is explained by Xiang et al.,[6]

In order to overcome the disadvantages of classical K-means clustering algorithm, the concept of K-means clustering method is combined with PSO algorithm which is called the clustering algorithm based on Particle Swarm Optimization Algorithm which was implemented by Zhenkui, [10]. It used the global optimization of PSO algorithm to make up the

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

shortage of the clustering method. This algorithm is more effective and promising.

The amount of available data and collected is greater than the human capability to analyze and extract knowledge from it. To help in the analysis and to efficiently and to automatically extract knowledge, new techniques and new algorithms need to be developed. Two new data clustering approaches are introduced using the Particle Swarm Optimization (PSO) Esmin et al.,[7]. In this PSO can be used to find centroids of a user specified number of clusters. The data clustering PSO algorithm using the original fitness function is evaluated on well known data set.

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. Forsati et al.,[8] explains this a novel clustering algorithms based on Harmony Search (HS) optimization method that deals with web document clustering

Clustering is currently one of the most crucial techniques for dealing with massive amount of heterogeneous information on the web. Most commonly used partitioning-based clustering algorithm, the K-means algorithm, is more suitable for large datasets. However, the K-means algorithm can generate a local optimal solution. Forsati et al.,[9] demonstrates a Novel harmony search clustering algorithms that deal with documents clustering based on harmony search optimization method. A pure harmony search based clustering algorithm that finds near global optimal clusters within a reasonable time. Contrary to the localized searching of the K-means algorithm, the harmony search clustering algorithm performs a globalize search in the entire solution space. Then harmony clustering is integrated with the K-means algorithm in three ways to achieve better clustering.

The Clustering algorithm based on object function resolves the clustering problem into optimization problem, thereby it becomes to the main investigatory stream. But it has some shortcomings such as its sensitivity to initial condition, and it is easy to fall in local peak. To overcome these deficiencies, Zhang et al.,[12] applies an ant colony optimization algorithm is to clustering analysis and a novel clustering based on an improved ant colony optimization algorithm. Experimental analysis shows that this method is faster and more efficient to convergence upon the optimal value in the whole field.

Data clustering is an unsupervised task that can generate different shapes of clusters for a particular type of dataset. Hence choosing an algorithm for a particular type of dataset is a difficult problem. A choice of clustering algorithm is entered by a comparison between three techniques such as, K-means, Self Organizing Map (SOM) and Density Based Spatial Clustering. Comparison is performed on the basis of cluster quality index. Dehuri et al.,[11] represents a density based clustering algorithm is preferable, if the clusters are of arbitrary shape. K-means or SOM is preferred if the clusters are of hyper-spherical.

Methodology

Term Frequency- Inverse Document Frequency(TF-IDF)

This is a originally a term weighting scheme developed for information retrieval (as a ranking function for search engines results), that has also found good use in document classification and clustering.

```
mln--0.0
dlr--0.0
ct--0.0
share--1.3862943611198906
reuter--0.0
acq--2.1972245773362196
chemlawn--43.0322530597096
chem--41.845385723286604
rise--3.2188758248682006
hope--3.912023005428146
for--0.0
higher--5.545177444479562
bid--8.317766166719343
chemlawn--43.0322530597096
corp--0.6931471805599453
chem--41.845385723286604
attract--3.912023005428146
higher--5.545177444479562
bid--8.317766166719343
dlr--0.0
per--1.3862943611198906
share--4.1588830833596715
offer--11.090354888959125
wast--27.38416103799702
```

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Figure 1: Weight of the features

Weight of a feature can be calculated using the formula,

$$W_{ij}=\text{TF-IDF}(i,j)$$

$$\text{TF-IDF}(i,j)= tf(i,j).\ (\log\frac{N}{df(j)})$$

tf (i, j) the frequency of feature j in a document di, N is the number of documents in the whole collection, and df (j) is the number of documents where feature j appears

Term frequency denotes the number of times a particular term has occurred in the document. The following figure shows the weight of features of input documents.

Concept Factorization

Deng et al., [4] explains Concept factorization models, in which each cluster is represented as a linear combination of the data points, and each data point as a linear combination of the cluster centers. 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.

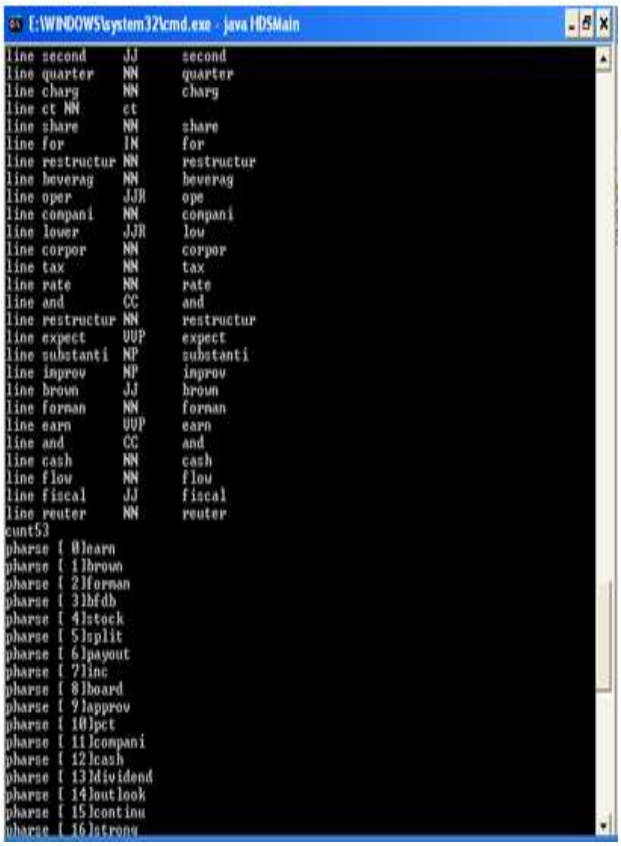


Figure 2: Concept Factorization Process

That is instead of considering, coverage factor the concepts in the documents was considered and clustering process takes place based on the concepts in the documents.

Figure 2. Shows the screen shot of concept factorization process for the given input documents

Experimental results shows that high quality clusters can be obtained when the documents are clustered based on the concepts in it.

Algorithms

K-Means Algorithm

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids shoud be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

a. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
b. Assign each object to the group that has the closest centroid.
c. When all objects have been assigned, recalculate the positions of the K centroids.
d. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Harmony Search Method

It is a meta heuristic algorithm[8]. In this, each decision variable generates a value for finding a global optimum solution. The HS algorithm initializes the Harmony Memory (HM) with randomly generated solutions. The number of solutions stored in the HM is defined by the Harmony Memory Size(HMS). Then iteratively a

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

new solution is created as follows. Each decision variable is generated either on memory consideration and a possible additional modification, or on random selection. The parameters that are used in the generation process of a new solution are called Harmony Memory Considering Rate (HMCR) and Pitch Adjusting Rate (PAR). Each decision variable is set to the value of the corresponding variable of one of the solutions in the HM with a probability of HMCR, and an additional modification of this value is performed with a probability of PAR. Otherwise (with a probability of 1- HMCR), the decision variable is set to a random value. After a new solution has been created, it is evaluated and compared to the worst solution in the HM. If its objective value is better than that of the worst solution, it replaces the worst solution in the HM. This process is repeated, until a termination criterion is fulfilled.

a. Initialize the HM with HMS randomly generated solutions
b. repeat
c. Create a new solution in the following way
d. for all decision variables do
e. With probability HMCR use a value of one of the solutions in the harmony memory
f. and additionally change this value slightly with probability PAR
g. Otherwise (with probability 1-HMCR) use a random value for this decision variable
h. end for
i. if the new solution is better than the worst solution in the harmony memory then
j. Replace the worst solution by the new one
k. end if
l. until Termination criterion is fulfilled
m. return The best solution in the harmony memory

Hybridization of K-Means and Harmony Search Method

Using K-means algorithm for document clustering performs localized searching. That is, the solution obtained is generally related to the solution obtained in the previous step. Since because K-means algorithm uses randomly generated values as centroids of each clusters and the centroid values keeps on changing at every iteration. So, the final solution depends on initial randomly generated centroid values. Whereas, Harmony Search Method is good at finding the centroid values. But it takes more time to converge. In order to overcome the above aspects of K-means and HSM, a hybrid algorithm [9] that combines both the ideas can produce effective results. Hybridization of K-means and HSM can be done in three different ways.

Experimental Results

The algorithms are compared based on 2 factors: Quality and speed of convergence. The performance of the proposed method for clustering documents was calculated using Precision, Recall and F-measure. [8][9]

The precision value can be calculated using the following formula,

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved}}{\text{Total number of retrieved documents}}$$

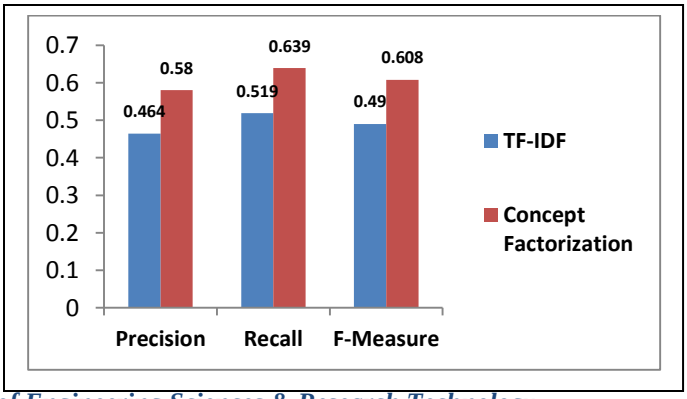
The Recall and F-Measure values can be calculated using the following formula,

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

$$\text{F} - \text{Measure} = \frac{2.\,\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 1: Summary Description of Clusters with input documents=50

Evaluation Measures	K- Means + Harmony Search Method	
	TF-IDF	Concept Factorization
Precision	0.464	0.580
Recall	0.519	0.639
F-Measure	0.490	0.608



ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Figure 3. **Evaluation measure values for TF-IDF, and Concept Factorization with input documents=50**

The input documents are retrieved from Reuters dataset. A sample of 50 documents has been retrieved and the cluster values are represented in the Table 1. Experimental results shows that concept factorization method produces better results and it outperforms the TF-IDF

Conclusion

In this paper, we have described a hybridization of K-means and HSM with two different methodologies. K-means algorithm is simple, it has many drawbacks and the most important one is, it suffers from local optima problem. The Harmony Search Method overcomes the problem of local optima but it takes time to converge. So, Hybridization of K-means and HSM was implemented in order to overcome the above mentioned disadvantages. Clustering documents based on the weight of the features was implemented using TF-IDF method. Better quality clusters may not be obtained if we consider only frequency of a term. But high quality clusters can be obtained if the clustering process takes place based on the concepts. So, the method called Concept Factorization is used to cluster the documents. The metrics like Precision, Recall and F-Measure are used to evaluate the performance of the clusters. Experimental results prove that concept factorization outperforms the results of Term Frequency – Inverse Document Frequency.

References

1. *Yung-Shen L, Jung-Yi J, Shie-Jue L, "A Similarity Measure for Text Classification and Clustering", in IEEE Transactions on Knowledge and Data Engineering,2013*
2. *Christos B, Vassilis T, "A clustering technique for news articles using WordNet", in ELSEVIER-Knowledge-Based Systems , July 2012*
3. *Taiping Z, Yuan Y T, Bin F, Yong X, "Document Clustering in Correlation Similarity Measure Space",in IEEE Transactions on Knowledge and Data Engineering, VOL. 24, NO. 6, 2012*
4. *Deng C, Xiaofei H,Jiawei H ,"Locally Consistent Concept Factorization for Document Clustering", in IEEE Transactions on Knowledge and Data Engineering, VOL.23, NO.6, 2011*
5. *Shady S, Fakhri K, Mohamed S,''An Efficient Concept-Based Mining Model for Enhancing Text Clustering",in IEEE Transactions on Knowledge and Data Engineering, VOL. 22, NO. 10, 2010*
6. *Xiang J W., Huai L, Hong S.Y., Ning S, "Application of Genetic Algorithm in Document Clustering",International Conference on Information Technology and Computer Science, 2009*
7. *Esmin A.A.A , Pereira D.L. , Araaujo F.P.A, "Study of Different Approaches to Clustering Data by Using the Particle Swarm Optimization Algorithm" in IEEE Congress on Evolutionary Computation, 2008*
8. *Forsati R, Mahdavi M, Kangavari M, Safarkhani B, "Web Page Clustering using Harmony Search Optimization". International Conference on Information Technology,2008*
9. *Forsati R, Meybodi M.R., Mahdavi M, Neiat A.G., "Hybridization of K-means and Harmony Search Methods for Web Page Clustering",inIEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008a*
10. *Zhenkui P, Xia H, Jinfeng H, "'The Clustering Algorithm Based on Particle Swarm Optimization Algorithm", in International Conference on Intelligent Computation Technology and Automation, 2008*
11. *Dehuri S, Mohapatra C, Ghosh A, Mail R ," "A Comparative Study of Clustering Algorithms",Information Technology Journal, 2008*
12. *Zhang X, Peng H, Zheng Q , "A Novel Ant Colony Optimization Algorithm for Clustering". ICSP Proceedings, 2006*
13. *Wai-Chiu Wong, Ada Wai-Chee Fu , "Incremental Document Clustering for Web Page Classification", IEEE International Conference on Information Society, 2000*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Author Biblography

	**Ms. S. Siamala Devi,**completed her B.E. CSE and M.E. CSE in Bannari Amman Institute of Technology,Tamilnadu, India. She is having 4 years of teaching experience. Currently she is working as an Assistant Professor in CSE department at Sri Krishna College of Technology, Coimbatore. She is pursuing her Ph.D in Anna University, Chennai. Her research area is data mining. **Email:** siamalamagesh@gmail.com
	**Dr. A. Shanmugam**, is working as Professor & Dean/ ECE at SNS College of Technology. He is having more than 35 years of teaching experience. He has published more than 120 papers in National and International Journals. He is a member of various professional bodies. He has guided more than 16 Ph.Ds and guiding 15 Ph.Ds in the field of computer communication networks He is having long list of achievements and awards.