Article can be accessed online at http://www.publishingindia.com

AN IMPROVED BISECTING K-MEANS ALGORITHM FOR TEXT DOCUMENT CLUSTERING

Janani Balakumar*, S. Vijayarani**

**Bharathiar University, Coimbatore, Tamil Nadu, India*
Email: janani.sengodi@gmail.com
***Assistant Professor, Bharathiar University, Coimbatore Tamil Nadu, India.*
E-mail: vijimohan_2000@yahoo.com

Abstract *Cluster analysis is an unsupervised learning approach that aims to group the objects into different groups or clusters. So that each cluster can contain similar objects with respect to any predefined condition. Text document clustering is the important technique of text mining in efficiently organizing the large volume of documents into a small number of significant clusters. The main objective of this research work is to cluster the collection of documents into related groups based on the contents of the particular documents. In order to perform this clustering task, this research work makes use of two existing algorithms, namely K-means and Bisecting K-means algorithm, and also this research work proposes a new clustering algorithm namely Enhanced-Bisecting K-means algorithm. From the experimental results it is observed that the proposed algorithm gives the better clustering accuracy than other algorithms.*

Keywords: *Text Mining, Text Document Clustering,K-means, Bisecting K-means, Enhanced Bisecting K-means*

Introduction

Clustering is the unsupervised technique that groups the data object into various groups or clusters so that objects within the similar cluster has a greater degree of similarity than the objects in different cluster [1]. The objects are enormously distinct to objects in other groups or clusters. The cluster is an ordered list of information which has familiar features. Cluster exploration can be accomplished through finding similarities between information. The similarities can also be located by using evaluating the characteristics. A good clustering process can able to produce high ability clusters.

Text clustering plays a vital role in the text mining process. Text clustering is used to group the similar documents that to form a comprehensible cluster, whereas documents that are different have separated into different clusters [2]. Accurate clustering involves a detailed definition of the closeness between a pair of objects, in terms of either the pairwise similarity or distance.

This paper organized as follows, section II explains the related work and section III presents the methodology of this research work. Experimental results are given in Section IV and Section V describes the conclusion of this research work.

Related Works

In [3] a simple and effective variant of K-means, bisecting K-means, where centroids are updated incrementally, was introduced. It creates better clusters than those fashioned by regular K-means. Bisecting K-means has a linear time complexity. The authors have first compared three agglomerative hierarchical techniques, namely Intra-Cluster Similarity Technique (IST), Centroid Similarity Technique (CST), and UPGMA. Experimental result shows that UPGMA is the best hierarchical algorithm when compared with K-means and bisecting K-means. Bisecting k-means is proving to be superior to UPGMA and regular k-means. The better performance of bisecting K-means is because of production of quite uniform size clusters.

In [4] principal component analysis (PCA) and linear transformation is used for dimensionality reduction and feature extraction. Then initial centroid is calculated. After the reduced dataset, the K-Means clustering algorithm was applied to the reduced data set.

In [5] proposed a web fuzzy clustering model. In their work the experimental result of web fuzzy clustering in web user clustering shows the probability of web fuzzy clustering in web usage mining.

In [6] presents a hybrid clustering algorithm that incorporates divisive and agglomerative hierarchical clustering algorithm. They have used bisecting K-means for divisive clustering and UPGMA for agglomerative clustering. First, they have cluster the number of documents using bisecting K-means algorithm with the value of K, which is greater than the total number of clusters, K. Next, they have calculated the centroids of K clusters attained from the previous step. Experimental results describes that the proposed method gives the better accuracy than the standard bisect K-means algorithm.

Methodology

The main objective of this research work is to cluster the group of documents by analyzing the contents of the documents using document clustering algorithms. In order to perform this task, this research work uses two existing algorithms; they are K-means and Bisecting K-means Algorithm, and also this research work proposes a new clustering algorithm namely Enhanced-Bisecting K-means algorithm. The performance factors are cluster purity, cluster entropy, F-measure, precision and recall. Figure 1 shows the architecture of this research work.

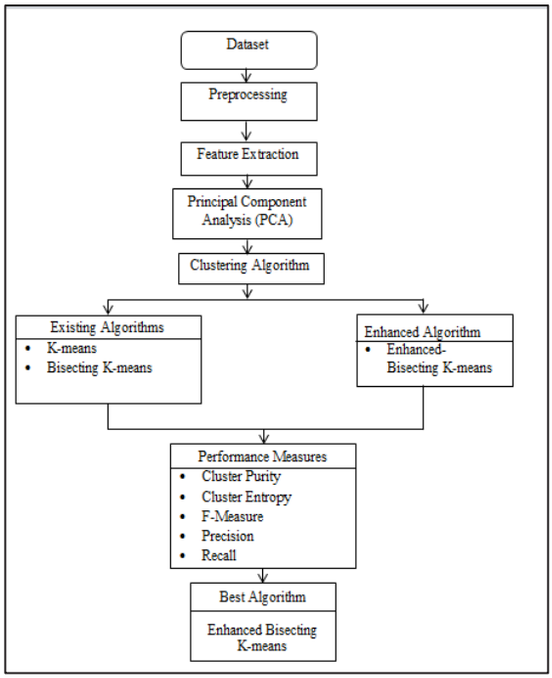

Fig. 1: System Architecture

A. Dataset

For the practical experiment the datasets can be collected from reuters dataset, 20 news group dataset, Text Retrieval Conference (TREC) dataset. These have been extensively used for evaluating feature selection techniques, classification and clustering. In order to make sure the multiplicity, the datasets are from different sources, some containing newspaper articles, some containing news group posts and the remaining being academic papers. The detailed summary of the dataset is given in Table 1.

B. Preprocessing

Document preprocessing is a crucial task in text mining, Natural Language Processing (NLP) and information retrieval (IR). In text mining, the document preprocessing is used for mining substantial knowledge from unstructured text data [7]. Before clustering the documents, the preprocessing techniques were applied to the specific dataset to moderate the size of the dataset. It will increase the efficiency of the particular document clustering system. In this research work, stemming, stop word removal, numbers and punctuation removal techniques are used.

C. Feature Extraction

Feature extraction incorporates to reduce the amount of resources required to describe a huge set of data. Feature extraction is a normal time period for strategies of building combinations of the variables to get around these problems while, despite the fact that describing the data with suitable accuracy. In this research work, the modified Principal component analysis (PCA) algorithm used for feature extraction [7].

Modified Principal Component Analysis

Principal component analysis or PCA is a typically used technique for dimensionality reduction. It was invented by Karl Pearson. The essential aim of PCA is to reduce the dimensionality of data prior to a few different statistical processing. The dimensionality reduction is correlated with the unsupervised clustering [6]. PCA has also been mounted in the applications of image compression, face recognition and to find patterns in high dimensional data.

The primary goal of utilizing PCA is a dimensionality reduction of data and to discover new crucial variables. An approach to finding or extract key accessories from trendy data influences the division of clusters. Important element evaluation is one such technique builds a linear combination of a collection of vectors that could exceedingly describe the variety of data [8].

Algorithm 1: Modified Principal Component Analysis

1. X: Original Data matrix, T: Transformation, Y: Centered data matrix

2. $Y=(y_1,y_2 \ldots y_n)$ where $y_i = x_i - x\square$. $x\square$ represents centered data matrix where $x\square = \sum x_i/n$.

3. Find the covariance matrix YYT.

4. Select the first p principal components having largest variances.

5. Form transformation matrix W consisting of p principal components.

6. Calculate the variance using the diagonal elements of D.

7. Sort the variances in descending order.

8. Find the reduced dataset D.

9. Partition the D data points into a K number of clusters using K-means; where K is the preferred number of clusters.

D. K-means Algorithm

K-means Clustering is an easy and empirical data analysis technique. This is a non-hierarchical technique of grouping objects collectively [8]. This clustering aims to partitioning the number of objects into k clusters in which all observation belongs to the cluster with nearest mean value.

In k-means algorithm, first choose the k initial centroids, where k is a user given parameter, referred to as the number of clusters. Each and every point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster [9]. The centroid of each cluster is then updated based on the points assigned to the particular cluster. Repeat those update steps until no point changes clusters or until the centroids remain the equal.

Algorithm 2: K-means

1: Initialize *k* centroids

2: repeat

3: for all objects in input do

4: Assign each element to its closest centroid

5: end for

6: for all centroids do

7: Compute the mean of the assigned points. This mean now becomes the new centroid

8: end for

9: until all centroids remains an unchanged or other termination criterion

E. Bisecting K- means

The bisecting k-means algorithm is a forthright extension of the k-means algorithm. The basic concept of bisecting k-means algorithm; to attain the K clusters, cut up the set of all points into two clusters, pick one of these clusters to split and so on, until K clusters has been fashioned [6].

It is the effective technique to reduce the dimensionality of featured vector for classification and clustering. There are various ways to select which cluster to split [10]. Here use the criterion primarily based on both size and the lowest sum of squared error.

Algorithm 3: Bisecting K-means

1: Input: K - Number of Clusters, D - Number of Documents. Initialize the list of clusters (Ls(c)) to contain the cluster consisting of all points

2: Repeat

3: Select the cluster from the Ls(c)

4: For i= 1 to K do

5: Bisect the number of clusters using basic k-means Bisect(c)

6: End for

7: select the two clusters from Ls(c) with the lowest total sum of the squared error (SSE)

8: add Bisect(c) ->Ls(c)

9: until the Ls(c) contains K clusters

F. Enhanced Bisecting K-means (EBKM)

The enhanced bisecting K-means method is a form of a hierarchical clustering method using k-means. The algorithm starts by setting all the documents in a single cluster. It divides the original cluster into two clusters by using K-Means. It makes the cluster, which has highest intra cluster similarity as solid and recursively split the other cluster into two more clusters using K-means and continue this until the preferred number of clusters are created.

Algorithm 4: Enhanced Bisecting K-means (EBKM)

Input: K: Number of clusters, D: Top N documents obtained by vector space similarity

Output: K clusters

1: put all the N documents in a single cluster C

2: for i=1 to K-1 do

3: for j=1 to ITER do

4: Use K-means to split the C into two sub-clusters, C1 and C2

5: if (intra-cluster similarity (C1)> intra-cluster similarity (C2))

6: make cluster C1 as permanent

7: C = C2

8: else

9: make cluster C2 as permanent

10: C = C1

11: end if

12: end for

13: end for

14: end

The important part is which cluster to choose for splitting. And there are different ways to proceed; here it is based on the intra cluster similarity.

$$\text{Sim}(X) = \sum_{d \in x} \text{cosine}(d, c)$$

Where d is a document in the cluster, X, and c is the centroid of cluster X.

Results and Discussion

In order to perform this task, the performance factors are cluster purity, cluster entropy, Precision, Recall and F-measure. In this research work, the documents are grouped into two clusters.

Purity: Given the true clustering assignment of the subjects, this function calculates the cluster purity index comparing with clustering assignment determined by integrative NMF algorithms. Higher value of cluster purity indicates better cluster predictive discrimination [11].

Entropy: Given the true clustering assignment of the subjects, this function calculates a cluster entropy index comparing with clustering assignment determined by integrative NMF algorithms [12]. Smaller values of cluster entropy indicate better cluster predictive discrimination.

Precision: It is calculated as the fraction of pairs correctly put in the same cluster.

Recall: It is calculated as the fraction of actual pairs that were identified.

F-measure: It is the harmonic mean of precision and recall.

Table 1: Summary of Datasets

Dataset	Source	Number of Documents	Number of Classes	Number of Words
re0	Reuters	1504	13	11465
re1	Reuters	1657	25	3758
20 news	20 news groups	18828	20	28553
tr41	TREC	878	10	7454
la1	TREC	3204	6	31472

Table 2 illustrates the total number of documents are grouped into two clusters. Fig 2 shows the number of documents within the cluster 1 and cluster 2.

Table 2: Number of Documents within the Clusters

Dataset	Documents Within the Cluster using K-means		Documents Within the Cluster using Bisecting K-means		Documents Within the Cluster using Enhanced Bisecting K-means	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
re0	376	1128	379	1125	391	1113
re1	813	844	827	830	831	826
20news	6276	12552	6301	12527	6329	12499
tr41	439	439	428	450	461	417
la1	1068	2136	1086	2118	1103	2101

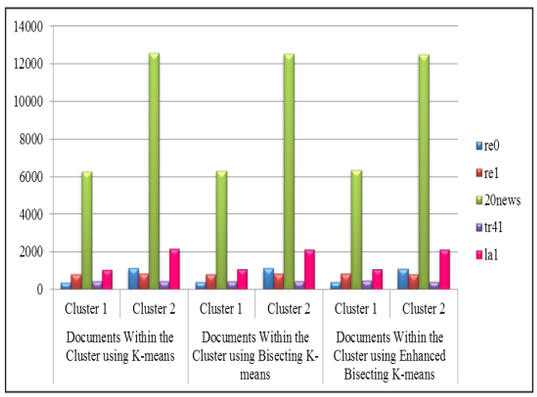

Fig. 2. Number of Documents within the Clusters

Table 3 describes the average of cluster purity for two clusters of all the data sets. Figure 3 explains the purity for three algorithms.

Table 3: Average Cluster Purity

Dataset	K-means	Bisecting K-means	Enhanced Bisecting K-means (EBKM)
re0	0.41	0.47	0.61
re1	0.59	0.61	0.73
20 news	0.43	0.48	0.59
tr41	0.29	0.52	0.76
la1	0.33	0.55	0.69

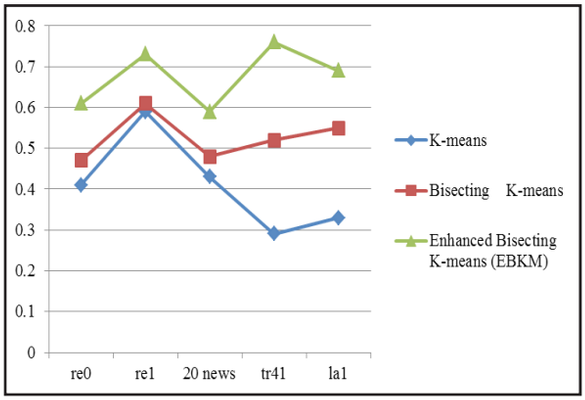

Fig. 3. Cluster Purity

Table 4 describes the average of the cluster entropy for two clusters of all the datasets. Figure 4 explains the entropy for three algorithms.

Table 4: Average Cluster Entropy

Dataset	K-means	Bisecting K-means	Enhanced Bisecting K-means (EBKM)
re0	0.50	0.47	0.31
re1	0.65	0.61	0.45
20 news	0.59	0.43	0.21
tr41	0.78	0.69	0.40
la1	0.76	0.51	0.51

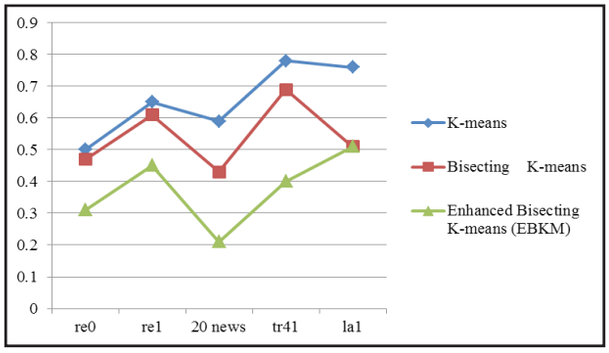

Fig. 4. Cluster Entropy

Table 5 describes the average of precision for two clusters of all the datasets. Figure 5 explains the precision for three algorithms.

Table 5: Average Precision

Dataset	K-means	Bisecting K-means	Enhanced Bisecting K-means (EBKM)
re0	29.48	38.32	56.18
re1	36.89	36.94	41.98
20 news	28.36	48.36	59.26
tr41	23.15	32.11	59.36
la1	14.36	29.13	39.15

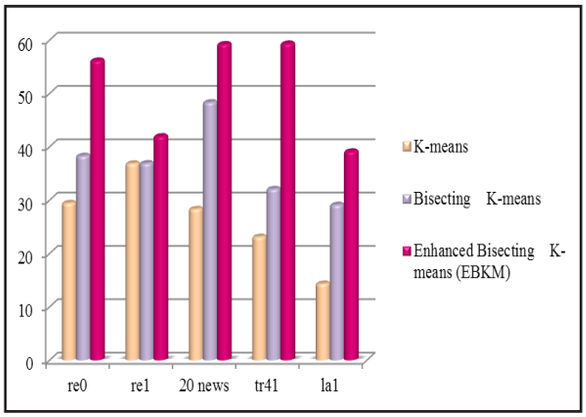

Fig. 5. Precision

Table 6 describes the average of recall for two clusters of all the datasets. Figure 6 explains the recall for three algorithms.

Table 6: Average Recall

Dataset	K-means	Bisecting K-means	Enhanced Bisecting K-means (EBKM)
re0	31.46	45.36	49.26
re1	30.36	36.25	48.78
20 news	41.23	58.26	69.36
tr41	45.98	58.63	59.78
la1	48.26	59.56	62.19

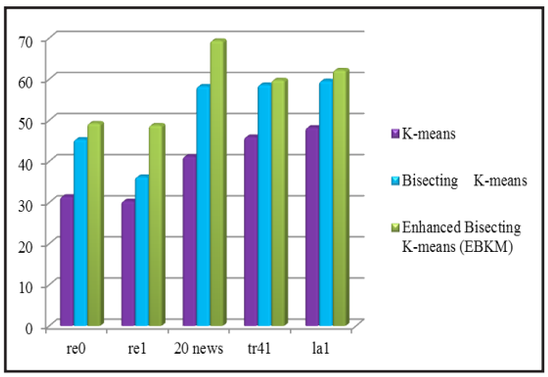

Fig. 6. Recall

Table 7 describes the average of F-Measure for two clusters of all the datasets. Figure 7 explains the F-Measure for three algorithms.

Table 7: Average f-Measure

Dataset	K-means	Bisecting K-means	Enhanced Bisecting K-means (EBKM)
re0	21.3	36.23	49.36
re1	36.21	39.65	49.34
20 news	36.12	42.36	59.87
tr41	36.23	39.84	42.15
la1	20.36	36.98	51.36

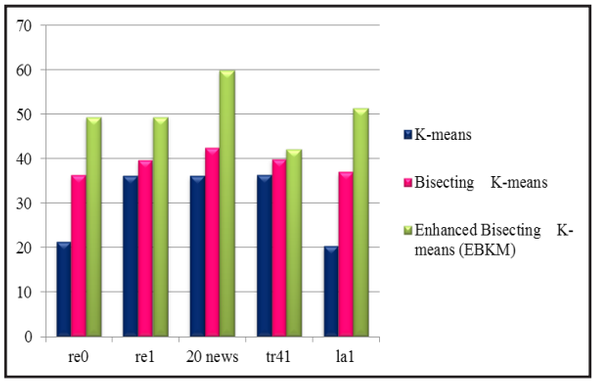

Fig. 7. F-Measure

Conclusion

Document clustering is a necessary process used in information retrieval, unsupervised document organization and automatic topic extraction. In this research work, it analyses the performance measures of existing and enhanced clustering algorithm for document clustering. The performance metrics are cluster accuracy, precision, recall, f-measure, cluster purity and cluster entropy. From this analysis, in existing algorithms the Bisecting K-means clustering algorithm gives the best accuracy for all types of data set. From the existing and enhanced algorithms, Enhanced Bisecting K-means (EBKM) algorithm gives the best accuracy. In future, the most efficient algorithms have to be developed for clustering all types of documents.

References

Steinbach, M., Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. *Proceedings of Knowledge Discovery and Data Mining (KDD) Workshop Text Mining*.

Baghel, R., & Dhir, R. (2010). A frequent concepts based document clustering algorithm. *International Journal of Computer Applications*, July, 4(5), 6-12.

Li, Y., Lv, X., Liu, Y., & Shi, S. (2010). Research on text clustering based on concept weight. *4th International Conference on Genetic and Evolutionary Computing*.

Napoleon, D., & Pavalakodi, S. (2011). A new method for dimensionality reduction using k-means clustering algorithm for high dimensional data set. *International Journal of Computer Applications*, January, 13(7), 41-46.

Liu, M., He, Y., & Hu, H. (2004). Web fuzzy clustering and its applications in web usage mining. *Proceedings of 8th International Symposium on Future Software Technology*.

Katariya, N. P., & Chaudhari, M. S. (2015). Bisecting k-means algorithm for text clustering. *International Journal of Advanced Research in Computer Science and Software Engineering*, Februrary, 5(2), 221-223.

Uncu, O., Gruver, W. A., Kotak, D. B., Sabaz, D., Alibhai, Z., & Ng, C. (2006). GRIDBSCAN: Grid density-based spatial clustering of applications with noise. *IEEE International Conference on Systems, Man, and Cybernetics*, October 8-11, Taipei, Taiwan.

Han, J., & Kambr, M. (2001). *Data Mining: Concepts and Techniques*. Hand Book. Beijing: Higher Education Press.

Thangamani, M., & Thangaraj, P. (2010). Ontology based fuzzy document clustering scheme. *Modern Applied Science*, July, 4(7), 148-156.

Jayabharathy, J., Kanmani, S., & Parveen, A. (2011). Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature.

Beil, F., Ester, M., Xu, X. (2002). Frequent term-based text clustering. ACM 1-58113-567-X/02/0007.

Deng, J., Hu, J. L., Chi, H., & Wu, J. (2010). An improved fuzzy clustering method for text mining. *2nd International Conference on Networks Security, Wireless Communications and Trusted Computing*.

Hamzah, A., Susanto, A., Soesianto, F., & Istyanto, J. E. (2007). Concept based text document clustering. *Proceedings of International Conference on Electrical Engineering and Informatics*, Indonesia June.

Ji, J., Chan, T. Y. T., & Zhao, Q. (2009). Fast document clustering based on weighted comparative advantage. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics* San Antonio, TX, USA - October.