OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

CLUSTERING OF TEXTUAL DATA BY USING K-MEANS TECHNIQUE

Miss. Geeta G. Dayalani

P.G. Student, Computer Science & Engineering , Everest Educational Society's Group of Institutions, Aurangabad, Maharashtra, India.

Abstract: To store the textual data and various documents, the usage of electronic media is widespread. To retrieve the important information from the large document collection of unstructured data is a very difficult and time consuming task. It is easier to find the relevant documents from a huge data collection only when the data collected is in ordered form or the data is classified by certain group or category. Still the problem persists to find the best grouping technique. This paper concentrates on the implementation technique of k-means clustering algorithm. K-means technique is used here to cluster the unlabeled data or the text document collection that is highly unstructured. It begins with the representative model of the unstructured data and finally generating the set of sorted clusters as a result. Furthermore, the results can be refined by analyzing the sorted set of clusters.

Keywords: *K Means, Vector Space Model (VSM), Euclidean Distance, Text Clustering, Residual sum of square, TF-IDF, HAC.*

I INTRODUCTION

1.1 Introduction:

Clustering is a process in which the data is divided into the group of similar item sets or objects which results in the simplification of data. That is, the grouping is made of similar as well as the dissimilar objects. Text clustering is also indicated as document clustering. It is one of the applications of cluster analysis used for textual documents. To use the clustering algorithms to maximum following things are important:

•*The* representation of an object,

• A distance or the similarity measure between the objects.

Clustering algorithms can be classified as:

1. *Flat clustering**:* This method generates a number of clusters that are flat without any definite structure which can relate the clusters to one another. It is also called as exclusive clustering.

2. *Hierarchical clustering*: This method generates a hierarchy of n number of clusters.

3. Hard clustering: This method allows each object to be a part of only one cluster at a time.

4. Soft clustering: This method allows partitioning the document over all the existing clusters. In this assignment, a document follows the fractional membership.

Algorithms used for Clustering are:

Agglomerative (Hierarchical clustering) K-Means that id Flat clustering and Hard clustering EM Algorithm that is Flat clustering and Soft clustering K-Means algorithm and Hierarchical Agglomerative Clustering (HAC) are being used for text clustering in a forthright way. It uses TF-IDF-weighted vectors, cosine similarity and normalized one. The k-means algorithm is used for n number of points in n-dimensional vectored space for the purpose of text clustering.

K-means is very prominent clustering algorithm. The main objective of flat K-means basically is to reduce the average squared distance of the objects from the centers of clusters, where a center of cluster can be defined by the centroid μ of objects within a C cluster:

$$\vec{\mu}(C) = \frac{1}{|C|}\sum_{\vec{x}\in C}\vec{x}$$

Residual Sum of Squares (RSS) is a measure of how the centroids or mean represent the members of the clusters, which is given by

$$\mathrm{RSS}_i = \sum_{\vec{x}\in C_i} \| \vec{x} - \vec{\mu}(C_i) \|^2$$

$$\mathrm{RSS} = \sum_{i=1}^{K} RSS_i$$

1.2 Necessity of Text Clustering:

Clustering is highly needed in data mining because:
1) To deal with huge databases, we need scalable clustering algorithms.
2) The algorithms should be highly capable to deal with any kind of data which may be numerical data, binary data or the categorical data.
3) The algorithms should be design is such a way that they are irrespective of the shape of cluster.
4) The algorithm should be capable of handling the high dimensional data as well in addition to the low dimensional data.
5) The algorithm should be capable of handling the noisy data because noisy data can result in degraded clusters.

1.3 Objective of Textual Clustering:

The main objective of clustering the text is to distribute the disorganized set of objects or data into the created clusters or groups. Grouping is done in a way that the objects with the same cluster are too similar and the objects within different clusters are too dissimilar.

II LITERATURE SURVEY

Variety of works which relates to clustering of text using various data mining procedures have driven this study. Vector space model [1] is an algebraic model that is used for representing the textual documents as vectors of identifiers. The tf-idf weighting scheme, can be used, where tf indicates the term frequency and idf represents the inverse document frequency. The common terms are removed by using the tf-idf scheme**.** In most of the already existing text clustering algorithms, text documents are generally revealed by the vector space model. In the model, each document is considered as a vector in the term-space and is represented by the following term frequency (TF) vector:

Hierarchical Agglomerative clustering and K-means are the two clustering techniques that are used widely for document clustering. Hierarchical clustering algorithms are used in the clustering for entries of various types like textual data input, multidimensional numeric data as well as categorical data items. For the process of searching, the combination of diverse hierarchical clustering algorithms and the method of HAC are mainly helpful to hold different methods of searching because it normally creates tree-like hierarchical structure. The effectiveness of this technique is bettering the efficiency of searching over a sequential scanning method. The main theme of the algorithm is to combine the documents based on their similarity into the number or variety of clusters. Single Linkage, Average Linkage and also the Complete Linkage Clustering are the various clustering techniques which are agglomerative.

Hierarchical clustering is regarded as the better quality approach for clustering, but is limited due to its quadratic time complexity.

EM (Expectation–Maximization) Algorithm:

EM algorithm is a method that is iterative of discovering the maximum-likelihood estimates of the parameters of a crucial distribution from already provided data set whenever the data is usually not complete or has certain missing values.
The EM algorithm has many presenting properties, some of which are:
1. It is arithmetically stable with every iteration of EM by raising the likelihood within the data.
2. Under general conditions, it has reliable global convergence.
3. It is simply implemented, analytically and computationally. In particular, it is generally simple to program and requires very small storage space.
4. The cost per iteration is too low, which can offset the larger number of iterations need for the EM algorithm when compared to other procedures that are competing for clustering.
5. It can also be used to provide the estimates of any missing data.

III SYSTEM DEVELOPMENT

The problem here is to cluster together all the set of documents that contains the similar data in one cluster and dissimilar data in the another cluster which is solved by K-Means algorithm.

Firstly the documents are collected from a huge corpus data set. Thereafter, we extract a single word for preprocessing like removing any stop word. Then the words are shown in vector form using their respective weights, and the weights are measured based on the difference of term frequency, and also the inverse term frequency. Highest frequency words and lowest frequency words are removed for clustering efficiently. Terms that are used in vector are regarded as the keywords used for clustering. Then, the weights of words are sorted in declining order, to minimum the keywords for clustering various documents. Finally, the documents are clustered using the similarity measure and k-means clustering algorithm. Figure 1 shows the system architecture.

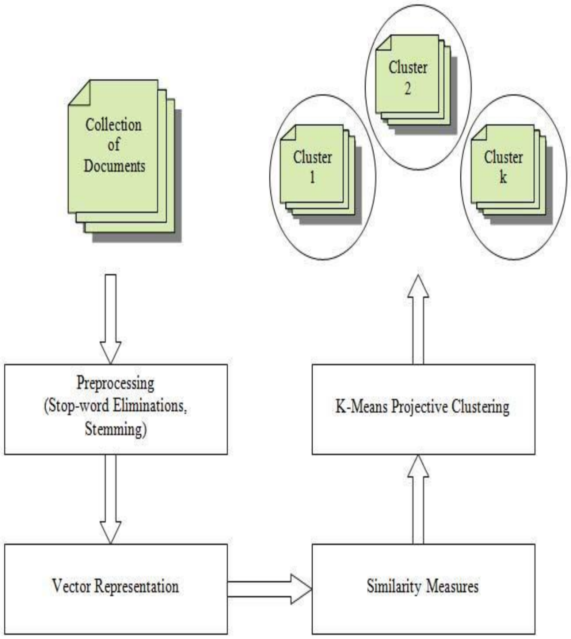
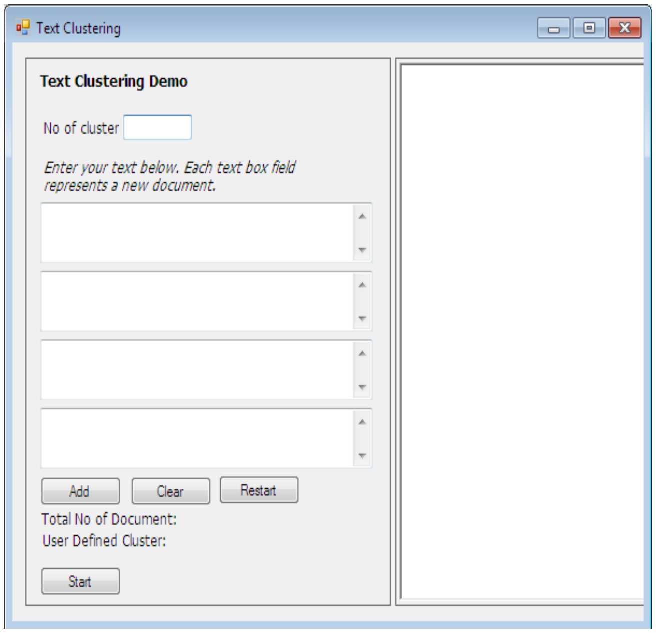


Figure 1: System Development

IV PERFORMANCE ANALYSIS

Execution details of the system:



Text Clustering Demo

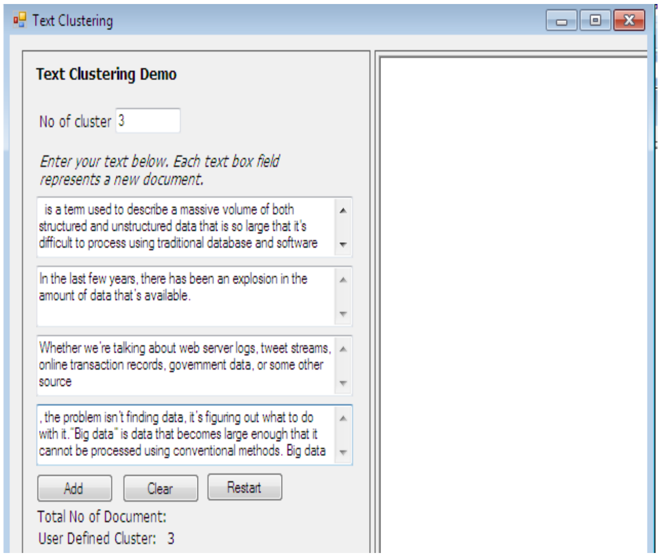
No of cluster

Enter your text below. Each text box field represents a new document.

Total No of Document:

User Defined Cluster:

The above Screen is the home screen for text clustering. By using this, we can cluster all the similar text together in a single cluster.



Text Clustering Demo

No of cluster 3

Enter your text below. Each text box field represents a new document.

is a term used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software

In the last few years, there has been an explosion in the amount of data that's available.

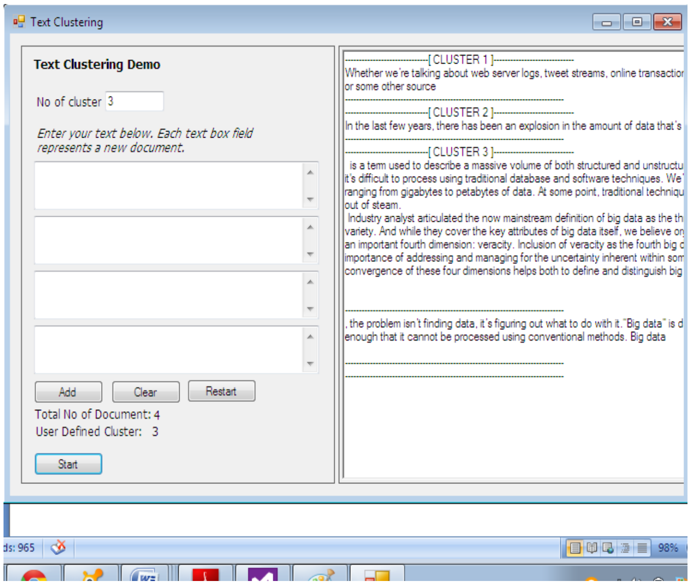
Whether we're talking about web server logs, tweet streams, online transaction records, government data, or some other source

, the problem isn't finding data, it's figuring out what to do with it."Big data" is data that becomes large enough that it cannot be processed using conventional methods. Big data

Total No of Document:

User Defined Cluster: 3

Initially, Number of clusters is needed to be entered and the text is entered in a document to cluster together. Once the data is entered, click on add to add it to the document.



Text Clustering Demo

No of cluster 3

Enter your text below. Each text box field represents a new document.

Total No of Document: 4

User Defined Cluster: 3

```
---------------------------[ CLUSTER 1 ]---------------------------
Whether we're talking about web server logs, tweet streams, online transactio
or some other source
-----------------------------------------------------------------------
---------------------------[ CLUSTER 2 ]---------------------------
In the last few years, there has been an explosion in the amount of data that's
-----------------------------------------------------------------------
---------------------------[ CLUSTER 3 ]---------------------------
 is a term used to describe a massive volume of both structured and unstructu
it's difficult to process using traditional database and software techniques. We'
ranging from gigabytes to petabytes of data. At some point, traditional techniqu
out of steam.
 Industry analyst articulated the now mainstream definition of big data as the th
variety. And while they cover the key attributes of big data itself, we believe on
an important fourth dimension: veracity. Inclusion of veracity as the fourth big d
importance of addressing and managing for the uncertainty inherent within som
convergence of these four dimensions helps both to define and distinguish big


-----------------------------------------------------------------------
, the problem isn't finding data, it's figuring out what to do with it."Big data" is d
enough that it cannot be processed using conventional methods. Big data

-----------------------------------------------------------------------
-----------------------------------------------------------------------
```

Once the document is entered, click on start to begin the clustering process. The clustered text with similar properties by using cosine rule is grouped together and displayed on the given blank screen.

V RESULT ANALYSIS

Implementation is the process of executing a plan or design to reach at the desired output.

Document Representation

Vector space model is the one of the efficient methods of representing documents as vectors using the term frequency weighting scheme as mentioned in Section 2.3.

The entire collection of dataset from the XML file is represented as vectors using the Vector space model.

Clustering Using K-means Algorithm

Input:

k: the n number of clusters,

Output:

A set of *k number of* clusters.

Method:

Step 1: Choose *k* numbers of clusters .

Step 2: Choose C*k* centroids randomly as the initial centers of the clusters chosen.

Step 3: Repeat

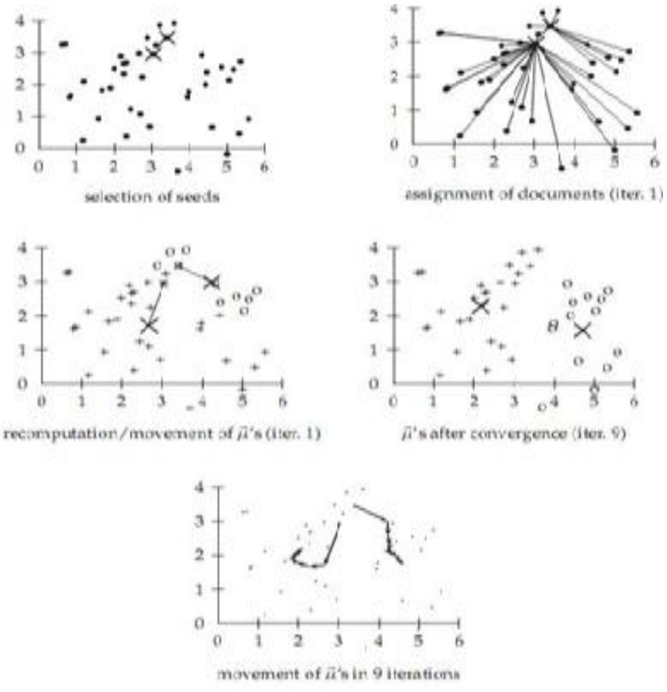
3.1: Assign each object to their most closest cluster center using Euclidean distance formula.

3.2: Compute the new cluster center by calculating the mean or centroid points.

Step 4: Until

4.1: No change in center of cluster OR

4.2: No object changes its respective clusters.



selection of seeds

assignment of documents (iter. 1)

recomputation/movement of $\vec{\mu}$'s (iter. 1)

$\vec{\mu}$'s after convergence (iter. 9)

movement of $\vec{\mu}$'s in 9 iterations

.

Termination Condition:

Any one of the termination condition can be applied from the following:

- The defined number of iterations has been finished.
- Between the iterations assignment of data to clusters should not change.
- Centroids do not alter between the iterations.
- Terminates when the decrease in RSS lapses below a threshold *t*

Performance Comparison:

Result Statement:

The results obtained from the entire process, which, in fact are the clusters of similar text, are presented in the console and also on a sample web page. The document shows clusters of text with similar data together.

Testing:

By creating many documents and entering the text data files for different types of data items and then run project for all created clusters repeatedly and obtain proper correct result.

Tests are the individual tests specified in a test plan document. Each test is typically described by

- An initial system state.
- A set of actions to be performed.
- The expected results of the test.
- K-means is a heuristic method of constructing clusters of the documents. There is no guaranteed output but the execution time is really fast. The K-means algorithm constructs the clusters of the documents based on their minimal distances to the centroid of the cluster. The output depends on optimal calculation of centroids of the clusters.

Algorithm	Initial Centroids	Accuracy (%)	Time taken (ms)
k-means algorithm (executed 7 times with randomly selected initial centroids)	5.1, 3.5, 1.4, 0.2 4.3, 3, 1.1, 0.1 6.6, 2.9, 4.6, 1.3	52.6	71
	7, 3.2, 4.7, 1.4 6.7, 3.1, 4.4, 1.4 5.1, 3.5, 1.4, 0.2	88.7	69
	7, 3.2, 4.7, 1.4 6.7, 3.1, 4.4, 1.4 7.4, 2.8, 6.1, 1.9	89.3	70
	7.4, 2.8, 6.1, 1.9 6, 3, 4.8, 1.8 6.7, 3.1, 4.4, 1.4	89.3	72
	5.1, 3.5, 1.4, 0.2 4.3, 3, 1.1, 0.1 6, 3, 4.8, 1.8	52.7	70
	6, 3, 4.8, 1.8 5.8, 2.7, 5.1, 1.9 5.1, 3.5, 1.4, 0.2	89.3	72
	5.1, 3.5, 1.4, 0.2 7, 3.2, 4.7, 1.4 6.3, 3.3, 6, 2.5	89.3	71
Mean value	-	78.7	70.7

V CONCLUSION

K-means algorithm is a simple yet popular method for clustering analysis.In this method, the quality of the sorted clusters depends heavily on the initial seeds that is centroids. These are selected in a random fashion. This method is very sensitive to the initial starting points and it does not promise to produce the unique results for clustering. Finally the proposed method mainly focuses on the less similarity based clustering to find the initial seed cluster centers efficiently. This method tries to minimize the time complexity to a greater extent. Text clustering is an important component of document organization and management. But, clustering of documents according to semantic features is a challenging problem in text data mining. Due to rapid increase in digital copies of data, scalability is also an issue. Further, semantics can be added to increase the quality of document clustering.

. The result produced would have been more accurate if the combined method of clustering such as hybrid clustering based on partition clustering and hierarchical clustering methods had been used. For future work is to plan to incorporate artificial intelligence technique with this method to improve the quality and accuracy of the system**.** In future, categorization methods can be added to render results and user interface more manageable and user friendly.

REFERENCES

[1] Athman Bouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering Volume 8, No. 2, April 1996.

[2] onner, R., On Some Clustering Techniques. IBM journal of research and development,
8:22-32, 1964.

[3] Fraley C. and Raftery A.E., “How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster
Analysis”, Technical Report No. 329. Department of Statistics University of Washington, 1998.

[4] Guha, S., Rastogi, R. and Shim, K. CURE: An efficient clustering algorithm for large databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 73-84, New York, 1998.

[5] Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[6] Hartigan, J. A. Clustering algorithms. John Wiley and Sons., 1975. Huang, Z., Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3), 1998.

[7] Hoppner F. , Klawonn F., Kruse R., Runkler T., Fuzzy Cluster Analysis,Wiley, 2000.

[8] Tyron R. C. and Bailey D.E. Cluster Analysis. McGraw-Hill, 1970 Jain, 2010] JAIN ANIL K.,Data Clustering: 50 Years Beyond K-Means [Jain & Dubes, 1988] JAIN, ANIL K., & DUBES, RICHARD C. 1988. Algorithms for clustering data. Prentice Hall.term-based text clustering, *ACM KDD Conference*, 2002.

[9]. *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, _o. 1, 2010* Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance.

[10]. Kmeans algorithm - Wikipedia, the free encyclopedia.