International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

TEXT DOCUMENT CLUSTERING USING K-MEANS ALGORITHM

Dr. A. Sudha Ramkumar[1], R.Nethravathy[2]

[1]*Assistant Professor, Sri Kanyaka Parameswari Arts and Science College for Women, Chennai, India*
[2] *PG Student, Sri Kanyaka Parameswari Arts and Science College for Women, Chennai, India*

Abstract - *Text document Clustering is the process of gathering relevant information into cluster. A cluster is specially designed for storing and analyzing the huge amount of text documents. There are several algorithms for clustering the large set of information from the text documents. In this paper, K-Means clustering algorithm is carried out to cluster the text documents. Document term matrix is constructed using the documents and all the unique words of documents. This matrix is highly sparse and it introduces complexity in clustering process. Dimension reduction techniques can be used to reduce the dimension of the document term matrix which intern reduces the complexity of clustering algorithm. In this paper, text documents are clustered using three dimension reduction (DR) techniques and it is compared with K-Means clustering algorithm. BBCSports dataset has been used for the experiment K-Means clustering using dimension reduction outperforms the K-Means clustering algorithm is proved through the experimental results.*

Key Words: Document Clustering, K-Means, Dimension Reduction, Confusion Matrix, Preprocessing

1. INTRODUCTION

1.1 TEXT DOCUMENT CLUSTERING

Text document clustering is the process of grouping a similar set of documents into clusters. Text clustering is accomplished by representing the documents as a set of features as indexes associated with numerical weights. The goal is always to cluster the given text documents, such that they get clustered based on the similarity measures with a reasonable accuracy. During text clustering, the documents need to be preprocessed before analyzing the data. The dimensions of the vector that represent the documents need to be reduced.

Text document clustering is generally considered to be a centralized process. Text document clustering may be used for different tasks, such as grouping similar documents and analyze, discovering meaningful implicit subjects across all documents. Using similarity measures, the document term matrix is constructed in traditional clustering methods. Since each document contains different terms, the dimension of this document term matrix is very high and sparse in nature. Because of this high dimensionality, the clustering process yields irrelevant results.

Text document clustering is used for partitioning a collection of text documents into similar clusters based on the distance or similarity measure. Document clustering groups similar documents to form a coherent cluster. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting. In document clustering, similarity is typically computed using associations and commonalities among features, where features are typically words and phrases. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard coefficient, Euclidean distance and Pearson Correlation Coefficient.

The aim of text document clustering is to group the documents. The documents should have high intra-cluster similarity and low inter-cluster similarity. The intra-cluster similarity is the documents within the cluster, the documents are closely related with each whereas the inter-cluster similarity is nothing but between the clusters, the documents are different with each other.

Clustering is the most common form of unsupervised learning and is a major tool in a number of applications in many fields of business and science. According to the Pankaj jajoo [1], the clustering is used for the following,

• **Finding Similar Documents:** This feature is often used when the user has spotted one "good" document in a search result and wants more-like-this. The interesting property here is that clustering is able to discover documents that are conceptually alike in contrast to search-based approaches that are only able to discover whether the documents share many of the same words.

•**Organizing Large Document Collections:** Document retrieval focuses on finding documents relevant to a particular query, but it fails to solve the problem of making sense of a large number of uncategorized documents.

• **Duplicate Content Detection:** In many applications, there is a need to find duplicates in a large number of documents. Clustering is employed for plagiarism detection, grouping of related news stories and to reorder search results rankings (to assure higher diversity among the topmost documents). Note that in such applications the description of clusters is rarely needed.

• **Recommendation System:** In this application, a user recommending articles based on the articles the user has already read. Clustering of the articles makes it possible in real time and improves the clustering quality.

• **Search Optimization:** Clustering helps a lot in improving the quality and efficiency of search engines as the user query can be first compared to the clusters instead of comparing it directly to the documents and the search results can also be arranged easily.

2. RELATED WORK

Anna Huang et al., [2] 2008, Compared and analyzed the effectiveness of similarity measures in partitional clustering for text document datasets which uses the standard K-Means algorithm. The experiment shows the use of K-Means algorithm and five similarity measures that have been most commonly used in text clustering.

Charu C.Aggarwal and ChengXiangZhai [9] 2012, provide a detailed survey of the problem of text clustering. In this, they provide the key challenges of the clustering, as it applies to the text domain and the key methods used for text clustering.

Bin Tang et al.,[8] 2005, Comparing four Dimension Reduction Techniques for text document clustering problem, using five benchmark data sets.
A.Anil Kumar and S.Chandrasekhar [4] 2008, provide the detailed concept of preprocessing and comparing the dimension reduction techniques.

Rakesh Chandra Balabantaray et al., [6] 2013, provide the complete process of clustering and provide the comparison of K-Means and K-Medoids algorithms.

Twinkle savdas and jasmin jha, [7] 2015, provide the system to categorize the text documents and form a cluster with the electronic data.

Pankaj Jajoo [1] 2008, provide the approaches, the first approach is improvement of graph partitioning techniques used document clustering. And the second approach is that the words clustered first and then the word cluster used to cluster the documents. This reduces the noise in data and thus improves the quality of the clusters.

D. Sailaja et al., [3] 2015, provide an overview of pre-processing text clustering methods, introduce an effective digital text analysis strategy using E-mail dataset.

ShouvikSachdeva and BhupendraKastore [5] 2014, they used the "Bag of words model" to represents each document and compare the representations using various similarity measures.

3. METHODOLOGY

The text document clustering using K-Means clustering algorithm uses the following methodology. The Methodology contains the six phases. These phases are Data collection, preprocessing, document term matrix, K-Means clustering, DR techniques and cluster evaluation.

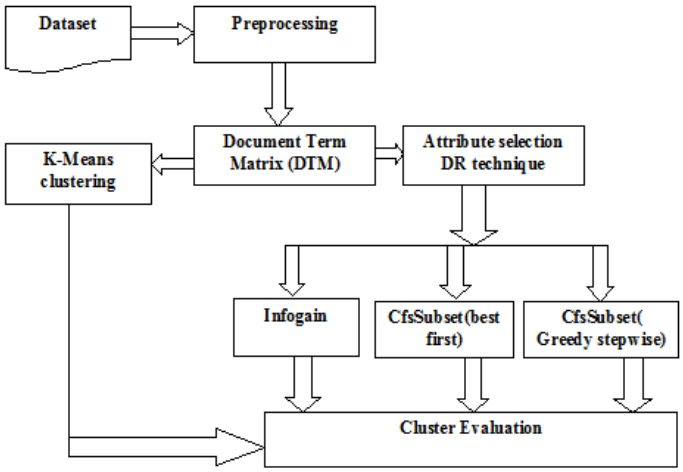

Figure 1: Methodology of K-Means with DR technique

Document collection: BBCSports dataset is downloaded from the BBC website. BBCSports consists of 737 text documents. BBCSports contains five classes such as, Athletics, cricket, football, rugby and tennis class. These five classes are combined together in the BBCSports dataset. 101 text documents are in athletics class, 124 text documents are in cricket class, 265 text documents are in football class, 147 text documents are in rugby class and 100 text documents are in tennis class. These 737 text documents are used for K-Means clustering as well as K-Means using Dimension Reduction techniques.

Preprocessing is used for extracting information from unstructured data. A dataset consists of massive volume of text documents which is collected from heterogeneous sources of text documents. For the efficient preprocessing of text documents the following techniques are used. There are tokenization, stopword removal, and stemming. Tokenization is the first step of analyses. The main use of tokenization is identifying the meaningful keywords. Stopword removal is reduces the text data and improves the system performance. Stopwords are the words like "also", "and", "or", "can", "this" which occurs frequently but are meaningless. Stemming is the process of reducing derived words into their base or root word. For example, jumping, jumped, jumps must be reduced into its common root "jump".

Document term matrix or term document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document

term matrix, rows correspond to documents in the collection and columns correspond to terms.

K-Means clustering algorithm: The k-means algorithm takes the input parameter k, and partitions a set of n-objects into K-clusters so that the resulting intra-cluster similarity is high whereas the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in the cluster, which can be viewed as the clusters center of gravity. The syntactical similarity of the terms is calculated by using Euclidean distance.

Steps of K-Means clustering

1. Select k observations as initial cluster centroids (seeds).
2. Assign each observation to the cluster that has the closest centroids (for example, in Euclidean sense).
3. When all the observations have been assigned, recalculate the positions of the k centroids.
4. Repeat until the cluster centroids no longer change.

A **confusion matrix** is a table that is often used to describe the performance of a clustering model on a set of text document for which the true values are known. It allows the visualization of the performance of an algorithm and in unsupervised learning it is called as matching matrix and is shown in the figure 2. In the confusion matrix, all the diagonal elements are true positives and it is the relevant document to that particular class, where as the number of documents retrieved are True Negatives (TN) and True Positives (TP).

ACTUAL CLASS		PREDICTED CLASS	
		POSITIVE	NEGATIVE
	POSITIVE	TP	FN
	NEGATIVE	FP	TN

Figure 2: Confusion matrix

DR techniques, The large numbers of attributes are selected from the dataset, the Document term matrix is high dimensional sparse matrix. Attribute selection dimension reduction method is used to reduce the dimension of the matrix. The selected attributes are analyzed and reduced by filter based attribute selection method. The InfoGain feature selection method selects the features from the original set of attributes based on ranking of attributes. This reduced feature set is applied to the K-Means clustering method. The results of K-Means as well as K-Means with Attribute Selection DR method are validated using the evaluation metrics.

Infogain (IG) DR technique, Ingogain selects many items in pure feature sets of text documents. Infogain (IG) is an effective Feature Selection method and is widely used in text document. Infogain does not concern the relation between a certain feature word and certain class, but treat all classes in training set as a whole. And the importance of a certain word is measured by calculating the information amount that each class takes.

Cfssubset (CSS) DR technique, In CfsSubset, values of subsets are correlate highly with the class value and low correlation with each other. It is used to evaluate the worthy of attributes subset by considering the individual predictive ability of each attribute along with the degree of redundancy between them. Attribute Subsets that are highly correlated with the class while having low intercorrelation are preferred.

The Search Method is the structured way in which the search space of possible attribute subsets is navigated based on the subset evaluation. Baseline methods include Random Search and Exhaustive Search, although graph search algorithms are popular such as Best First Search.

Attribute evaluation method is:

- **BestFirst**: Uses a best-first search strategy to navigate attribute subsets.
- **GreedyStepWise**: Uses a forward (additive) or backward (subtractive) step-wise strategy to navigate attribute subsets.

Cluster quality evaluation

K-means clustering is applied on the dataset and a class to clusters evaluation method of WEKA tool is used. It generates on output in the form of confusion matrix R.E-Benchs (2018).

4. EVALUATION METRICS

The evaluation metrics used in this paper are precision, recall, f-measure and accuracy.

PRECISION

This measure retrieves the number of correct text documents out of the number of total text documents made by the system.

$$\text{Precision} = \frac{\text{Number of relevant documents retrieved (TP)}}{\text{Number of documents retrieved (TP+FP)}} \quad \ldots (1)$$

RECALL

This measure retrieves the number of correct text documents made by the system, out of the number of all possible text documents.

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved (TP)}}{\text{Number of relevant documents (TP+FN)}} \quad \ldots (2)$$

ACCURACY

The accuracy of a measurement is how close a result comes to the true value.Systemetic error or Inaccuracy is quantified by the average difference (bias) between a set of measurements obtained with the test method with a reference value or values obtained with a reference method.

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad \ldots (3)$$

F-MEASURE

This measure is a combination of the precision and recall measures used in machine learning.

$$\text{F-Measure} = \frac{2*(\text{Precision}*\text{recall})}{(\text{Precision}+\text{recall})} \quad \ldots (4)$$

RESULTS AND DISCUSSION

In this paper, comparison of K-Means clustering and K-Means clustering with DR technique using the BBCSports dataset, which has five classes such as athletics, cricket, football, rugby and tennis has been proposed.

In the following table, it is clear that K-Means with infogain (IG) DR technique is more effective than the K-Means clustering without dimension reduction techniques. The K-Means with infogain DR technique has 97.8% precision, 96.4% recall, 96.7%accuracy and 97% F-measure.

Table 1: Comparison of K-Means and K-Means with DR techniques

Evaluation Matrix	**Precision**	**Recall**	**Accuracy**	**F-Measure**
K-Means	89.6	85.8	86.2	87.6
K-Means(CSS bestfirst)	95	94.2	94	94.5
K-Means(CSS greedy)	95.6	96.4	95.5	95.9
K-means(IG)	97.8	96.4	96.7	97

The following figure.3 shows the effectiveness of K-Means with infogain DR techniques over the K-Means clustering without the DR technique.

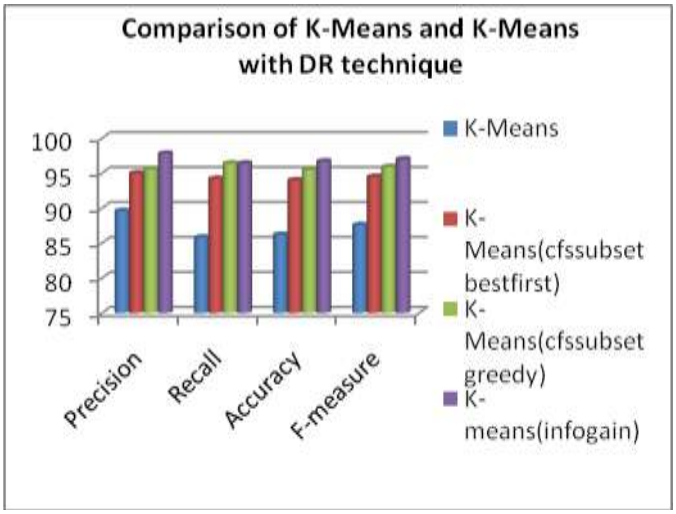

Figure 3: Comparison of K-Means and K-Means with DR techniques

5. CONCLUSION

The main aim of text document clustering is to grouping the similar documents into a cluster. This paper discusses comparison of K-Means clustering and K-Means clustering with DR techniques. The K-Means clustering with DR techniques improves the clustering quality significantly. When compared to K-Means clustering. The experimental results of K-Means and K-Means using DR techniques clustering algorithm using evaluating measures such as, precision, recall, accuracy and f-measure has been discussed in this paper.

REFERENCES

1. Pankaj Jajoo, Document Clustering, Indian Institute of Technology Kharagpur, 2008.
2. Anna and Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, Similarity Measures for Text Document Clustering, 2008.
3. D Sailaja et al, International Journal of Computer Science and Information Technologies (IJCSIT), An Overview of Pre-Processing Text Clustering Methods, 2015.
4. Anil Kumar and S.Chandrasekar, Dept of CSE, Sri Sivani College of Engineering, India. Text data preprocessing and dimensionality reduction for document clustering 2012.

5. ShouvikSachdeva and BhupendraKastore, Indian Institute of Technology, Kanpur, Document Clustering: Similarity Measures, 2014.
6. Rakesh Chandra Balabantaray et al, Document Clustering using K-Means and K-Medoids, 2013.
7. Twinkle Savdas and Jasmine Jha, Document Cluster Mining On Text Document, 2015.
8. Bin Tang et al., Comparing dimension reduction technique for document clustering, 2005.
9. Charu C.Aggarwal and ChengXiangZhai, A Survey Of Text Clustering Algorithms, 2012.
10. R.E Banchas, Text Mining with MATLAB", Springer, 2012.