ABSTRACT

The need for clustering of documents is high in applications like document summarization, information retrieval etc. Huge collections of documents are pilling every day. It is really challenging to efficiently cluster given text documents. It is evident that clustering needs to be performed with best preprocessing and analysing techniques with respect to preserving the order of sequence of words and concept of words in the documents. In order to best understand the concepts in a document which further helps in clustering the document and putting it into the most appropriate cluster, it is essential to represent the document in a semantic representation. Semantic representation preserves the meaning of words in a document. There are many algorithms and approaches used till date which have their own merits and demerits. The algorithms used for word vectors here is "Word2Vec-Skip grams Model", a word vector model is a neural network which generates a 100 dimension word vector i.e. a vector of 100 dimensions for each word, and the document is represented by computing a feature vector. A feature vector is calculated by using the word vectors by applying the min max method, min max method which is used summarizes all the vectors of the document into a single feature vector and for clustering is "k means". It is used for clustering the documents.

CONTENTS

Abstract **i**

List of Figures **iv**

List of Tables **v**

CHAPTER-1. INTRODUCTION **1**

1.1 Clustering 2
1.1.1 Traditional Clustering 2
1.1.1.1 Partitioning clustering 3
1.1.1.2 Hierarchical clustering 6
1.1.1.3 Fuzzy clustering 7
1.1.1.4 Density-based clustering 7
1.1.1.5 Model-based clustering 9
1.1.2 Semantic Clustering 9
1.2 Document Representation Techniques 13
1.2.1 One-Hot 14
1.2.2 Word bag model (Bag of Word, BOW) 15
1.2.3 Bag of nn-grams 15
1.2.4 Vector Space Model (VSM) 16
1.2.5 N-gram 18
1.2.6 Co-occurrence matrix 20
1.2.7 Word Embeddings 22
1.2.8 Doc2vec (Para2Vec) 23
1.2.9 Word2Vec 23
1.2.10 CBOW(Continuous Bag of Words) 24
1.2.11 SKIP-GRAM MODEL 25
1.2.12 LSA- Latent semantic analysis 26
1.2.13 LDA- latent Dirichlet allocation 27
1.2.14 Probabilistic latent semantic analysis(PLSA) 28
1.2.15 LSI-Latent semantic indexing 28
1.2.16 Semantic hashing 28
1.3 Evaluation Measures 29
1.3.1 Silhouette Coefficient 29
1.3.2 Calinski Harabasz score 29
1.3.3 Davies Bouldin Index 30
1.4 Motivation of Work 31
1.5 Problem Statement 31
1.6 Organization of the thesis 32

CHAPTER- 2. LITERATURE SURVEY **33**

2.1 Introduction 34
2.2 Existing methods for document clustering 34
2.2.1 Diagonal Dominance in kernel document clustering 34
2.2.2 Document embedding with paragraph vectors 34
2.2.3etecting Topics by Clustering Word Vectors 35

2.2.4 Number of topics optimisation clustering approach 35
2.2.5 Document clustering: TF-IDF approach 35
2.2.6 search result clustering method using name entities 35
2.2.7 Supervised clustering with support vector machines 36

CHAPTER-3. METHODOLOGY 37

3.1 System Architecture 38
3.2 Proposed System 40
3.2.1 Preprocessing 40
3.2.1.1 Dataset 40
3.2.1.2 Tokenization 40
3.2.1.3 Stop words Removal 40
3.2.1.4 Stemming 40
3.2.2 Word vector generation 41
3.2.3 Feature Vector Generation 42
3.2.4 K-Means and Principal component analysis 43
3.2.4.1 K-means 43
3.2.4.2 PCA 44

CHAPTER-4. EXPERIMENTAL ANALYSIS AND RESULTS 46

4.1 System Configuration 47
4.1.1 Software Requirements 47
4.1.2 Hardware Requirement 47
4.2 Sample Code 47
4.3 Screen shots 59
4.4 Experimental Analysis 62

CHAPTER-5. CONCLUSION AND FUTURE SCOPE 65

5.1 Conclusion 66
5.2 Future Scope 66

APPENDICES 67

A. List of stopwords 67
B. Porter Stemmer Algorithm 68

REFERENCES 71

LIST OF FIGURES

Fig. No.	NAME	PageNo.
1.1	Partition Clustering Algorithms	3
1.2	Example of Dendogram	7
1.3	DBSCAN	8
1.4	Document Representation Techniques	13
3.1	System Architecture	39
3.2	Word Vectors of a Document stored in a row in word in one cell and vector in the cell to its right.	42
3.3	Feature Vector of a Document stored in a row	43
3.4	Visualization of raw data before PCA	44
3.5	Visualization of raw data after PCA	44
4.1	Screenshot of News Paper Articles	59
4.2	Screenshot of Extracting Words From Documents	59
4.3	Screenshot of Vectors of each word	60
4.4	Screenshot of Words and its corresponding vectors	60
4.5	Screenshot of 100 dimension vectors	61
4.6	Screenshot of Feature Vectors	61
4.7	Screenshot of output(Clusters)	62
4.8	Graph of No. of unique words Vs Time required to train the words	63
4.9	Silhouette Score Analysis of K-means	63
4.10	The raw data	64
4.11	Graphical representation of clusters formed by K-means	64

CHAPTER-1: INTRODUCTION

Text clustering is the application of cluster analysis to text-based documents. It is an efficient analysis technique used in the domain of the text mining to arrange a huge unorganised text documents into a subset of coherent clusters. Documents which are similar belong to the same cluster, whereas the documents which are dissimilar belong to different clusters. Clustering is unsupervised; it creates the clusters depending upon the pattern. Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what are the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.

1.1 Clustering

Basically, clustering involves grouping data with respect to their similarities. It is primarily concerned with distance measures and clustering algorithms which calculate the difference between data and divide them semantically. Clustering methods are used to identify groups of similar objects in a multivariate data sets collected from fields such as marketing, bio-medical and geo-spatial. The following are types of clustering

- Traditional Clustering
- Semantic Clustering

1.1.1 Traditional Clustering

They are different types of clustering methods, including:

- Partitioning methods
- Hierarchical clustering
- Fuzzy clustering
- Density-based clustering
- Model-based clustering

1.1.1.1 Partitioning clustering

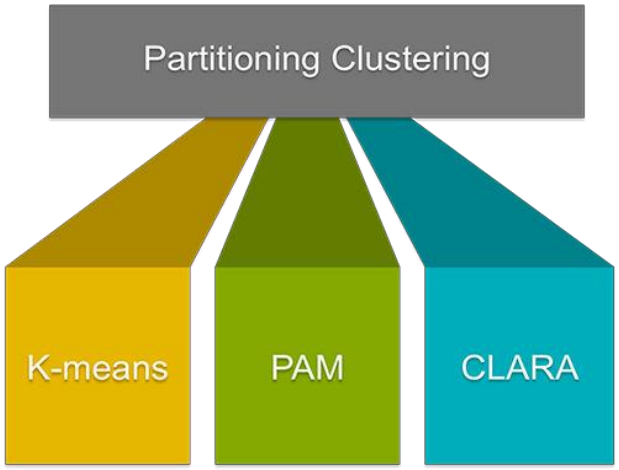

Fig: 1.1 Partition Clustering Algorithms

Partitional clustering (or partitioning clustering) are clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. The algorithms require the analyst to specify the number of clusters to be generated.

This course describes the commonly used partitional clustering, including:

- K-means clustering (MacQueen 1967), in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to anomalous data points and outliers.
- K-medoids clustering or PAM (Partitioning Around Medoids, Kaufman & Rousseeuw, 1990), in which, each cluster is represented by one of the objects in the cluster. PAM is less sensitive to outliers compared to k-means.
- CLARA algorithm (Clustering Large Applications), which is an extension to PAM adapted for large data sets.

There are mainly four types of partitioning algorithm includes as K-Mean Algorithm, K-Mediod Algorithm i.e. PAM (Portioning Around Medoids), CLARA and CLARANS.

K-M ean Algorithm: K-Mean is first developed by James MacQueen in 1967.A cluster is represented by its centroid, which is usually the mean of points within a cluster. The objective function used for k-means is the sum of discrepancies between a point and its centroid expressed through appropriate distance.

They have convex shapes clusters. Procedure of K-Mean:-

a) The technique requires arbitrary selection of choose k objects from D as the initial centres, where k is the number of clusters and D is the data set containing n objects.

b) Repeat the first step.

c) Reassign each object to the cluster to which object is most similar. It is based on the mean value of the objects in the cluster.

d.) Calculate the mean value of the objects for each cluster.

e) Until no change

Advantages of K-Mean:

1. If the variables are large, then K-Means most of the time computationally faster than hierarchical clustering methods.

2. K-Means produces tighter clusters than Hierarchical Clustering Method.

Disadvantages of K-Means Partition Algorithm:

1. It is difficult to predict the K Value.

2. More difficulty in comparing quality of cluster.

3. K-Means Algorithm does not work well with global clusters.

K-M ediod Algorithm: Partition around Medoids (PAM) is developed by Kaufman and Rousseeuw in 1987. It is based on classical partitioning process of clustering the algorithm selects k-medoids initially and then swaps the medoids object with non mediod thereby improving the quality of cluster. This method is comparatively robust than K-Mean particularly in the context of _noise‘or _outlier‘. K-Medoids can be defined as that object of a cluster, instead of taking the mean value of the object in a cluster according to reference point. K-Medoids can find the most centrally located point in the given dataset. Procedure of K-Medoids:-

Input:

- K: The number of clusters
- D: A data set containing n objects

Output:

- A set of K clusters Method:

The following steps are recommended by Tagaram Soni Madhulatha

1. The algorithm begins with arbitrary selection of the K objects as mediod points out of n data points (n>K).

2. After selection of the K mediod points, associate each data object in the given data set to most similar mediod.

3. Randomly select non-mediod object O.

4. Compute total cost S of swapping initial mediod object O.

5. If S>0, swap initial medoids with the new one.

6. Repeat steps until there is no change in the medoids.

Advantages of K-Medoids:

1) It is simple to understand and easy to implement.

2) K-Medoids Algorithm is fast and converges in a fixed number of steps.

3)Partition Around Medoids (PAM) algorithm is less sensitive to outliers than other partitioning algorithms.

Disadvantages of K-Medoids:

1) K-Medoids is more costly than K-Means Method because of its time complexity.

2) It does not scale well for large datasets.

3) Results and total run time depends upon initial partitions

CLARA (Clustering for Large Application)

CLARA means clustering large applications and has been developed by Kaufman and Rousseeuw in 1990. This partitioning algorithm has come into effect to solve the problem of Partition around Medoids (PAM).CLARA extends their K-Medoids approach for large number of object. This technique selects arbitrarily the data using PAM. According to Raymond T. Ng and Jiawei Han the following steps are performed in case of CLARA as given by the authors.

1) Draw a sample of 40+2k objects randomly from the entire data set, and call Algorithm PAM to find k medoids of the sample.

2) For each of the object determine the specific K medoids which is similar to the given object (Oj).

3) Calculate the average dissimilarity of the clustering thus obtained. If the value thus obtained is less than the present minimum we can use it and retained the K-Medoids found in the second step as best of medoids.

4) We can repeat the steps for next iteration.

Advantages of CLARA:

1) CLARA Algorithm deals with larger data sets than PAM (Partition around Medoids).

Disadvantages of CLARA:

1) The efficient performance of CLARA depends upon the size of dataset.

2).A biased sample data may result into misleading and poor clustering of whole datasets.

1.1.1.2 Hierarchical clustering

Hierarchical clustering is an alternative approach to partitioning clustering for identifying groups in the dataset. It does not require pre-specifying the number of clusters to be generated.

In contrast to partitional clustering, the hierarchical clustering does not require to pre-specify the number of clusters to be produced.

Hierarchical clustering can be subdivided into two types:

Agglomerative clustering in which, each observation is initially considered as a cluster of its own (leaf). Then, the most similar clusters are successively merged until there is just one single big cluster (root).

Advantages:

1) No apriori information about the number of clusters required.

2) Easy to implement and gives best result in some cases.

Disadvantages:

1) Algorithm can never undo what was done previously.

2) Time complexity of at least $O(n^2 \log n)$ is required, where *'n'* is the number of data points.

3) Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the following:

4) No objective function is directly minimized

5) Sometimes it is difficult to identify the correct number of clusters by the dendogram.

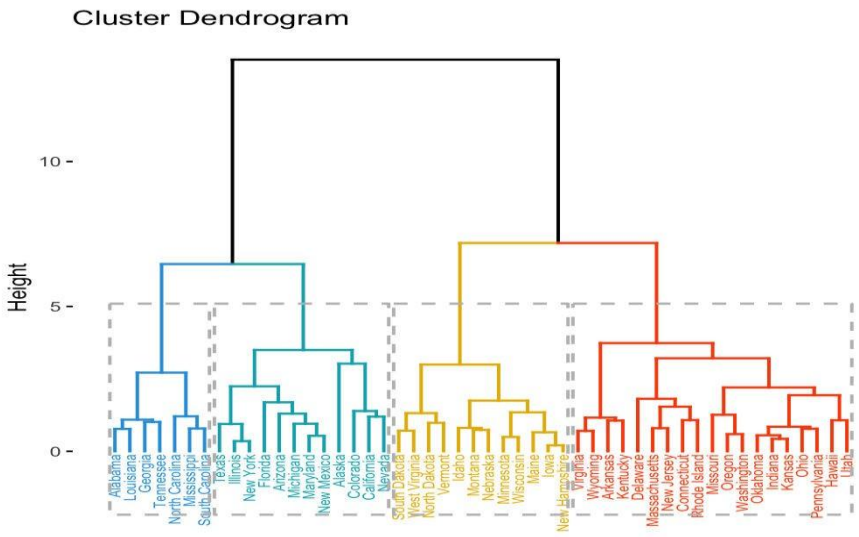

Fig: 1.2 Example of Dendogram

1.1.1.3 Fuzzy clustering

The fuzzy clustering is considered as soft clustering; in which each element has a probability of belonging to each cluster. In other words, each element has a set of membership coefficients corresponding to the degree of being in a given cluster.

This is different from k-means and k-medoids clustering, where each object is affected exactly to one cluster. K-means and k-medoids clustering are known as hard or non-fuzzy clustering.

In fuzzy clustering, points close to the center of a cluster may be in the cluster to a higher degree than points in the edge of a cluster. The degree, to which an element belongs to a given cluster, is a numerical value varying from 0 to 1.

The fuzzy c-means (FCM) algorithm is one of the most widely used fuzzy clustering algorithms. The centroid of a cluster is calculated as the mean of all points, weighted by their degree of belonging to the cluster.

1.1.1.4 DBSCAN (Density-based clustering)

DBSCAN (Density-Based Spatial Clustering and Application with Noise), is a density-based clustering algorithm (Ester et al. 1996), which can be used to identify clusters of any shape in a data set containing noise and outliers.

The basic idea behind the density-based clustering approach is derived from a human intuitive clustering method. For instance, by looking at the figure below, one can easily identify four clusters along with several points of noise, because of the differences in the density of points.