Text Document Clustering Using K-means Algorithm with Its Analysis and Implementation

Jaya Zade¹, Dr. G. R. Bamnote², Prof. P. K. Agrawal³
¹P.G. Student, Department of computer science & Engineering, P.R.M.I.T.R. Badnera, Amravati, Maharashtra, India
²,³Professor, Department of computer science & Engineering, P.R.M.I.T.R. Badnera, Amravati, Maharashtra, India.

Abstract: With the huge increase of information in daily life, it is essential to manage that information. So with the help of advance technology that will reduce the storage costs by using electronic media for strong textual document and information. But it has become very difficult to bring together all the relevant information just in time and as we know that people has always shortage of time or we may say that people always need everything in quick time. Therefore, clustering is discovered which bunch up relevant information in a cluster. There are various algorithm for clustering information out of which in this paper we discuss the implementation of k-means clustering algorithm for clustering unstructured text starting with the representation of unstructured text and reaching the resulting set of clusters together with its strength and weakness which will help us to compare it to other algorithm to choose correct one.

Keyword- clustering, k-means, cluster, document summarization

1. Introduction

Today as we know in every industry almost all document paper has their electronic storage copies. As compare to manual storage this method is safer and occupies much smaller space and also this electronic file has very easy and quick access. But because of increase in number of electronic document, it is very hard to handle that it is hard to organize, analyze and manage the electronic document efficiently by putting some manual effort. So this comes to a challenge for efficient and effective organization of a text in document automatically [1]. And therefore this increases the demand of the tool that can be used for analyze and discover useful information from document. Solution for this problem is to use data mining technique and usage of data mining technique on text document is called as text mining or text data mining which already increase the interest of many researchers in the field of research [2].

Text mining is roughly equivalent to text analysis where it derives high quality of information from the text. This text mining actually involves analysis of text such as information retrieval, lexical analysis to study word frequency distributions, pattern recognition, information extraction. The main aim of this technique is to turn text in useful data. Text mining technique is used in various application such as security application, biomedical application, record management, software application, online media application, marketing application, sentiment analysis, academic application etc. All these application are based on one common task that is extracting high quality information from the text document. Document clustering is field of data mining which automatically arranged useful data into group where data in category are similar to each other and dissimilar to other category of document [3].

Clustering can be known as one of the most important unsupervised learning problem. In general, the clustering is defined as the process of organizing objects into groups where its members are similar in some way⁷. Therefore, cluster is a collection of objects which are similar internally, but clearly dissimilar to the objects belonging to other clusters. In simple cluster is refer as group of similar kind of objects and it very useful for organizing document to improve the browsing retrivation and support. Clustering is also defined as a process of partitioning a set of data or objects into a set of meaningful sub-classes which is called as a clusters. Clustering is very helpful to user in understand the natural grouping or structure in a data set and it is used as either standalone tool to get better insight into data distribution or in pre-processing step of other algorithm. Good clustering will produce high quality result where intra class similarity is high and inter class similarity is low. The quality of a clustering result is depending on both the similarity measure used by the method and its implementation and also it is measured by its ability to discover some or all of the hidden patterns. Clustering has wide application in pattern recognition, image processing, document classification, economic science, marketing, insurance, land use and many more [3].

Document clustering is broadly classified in two categories that are hard clustering and soft clustering where hard clustering is the type clustering
which compute the hard assignment of a document to a cluster i.e., each document is assigned to exactly one cluster and they giving a set of disjoint clusters, therefore hard clustering is also called as disjoint clustering. And soft clustering is the type clustering in which it computes the soft assignment i.e., each document is allowed to appear in multiple clusters and thus it will generate a set of overlapping clusters and therefore it is also known as overlapping clustering. Soft clustering is further categorized into partitional clustering and hierarchical clustering where partitioning clustering allocate documents into a fixed number of non-empty clusters while hierarchical clustering is to build dendrogram, a hierarchical tree of clusters, whose leaf nodes represent the subset of a document collection. The most well-known partitioning methods are the K-means and its variants [1].

2. Review

Throughout these years, there has been a lot of work carried out on document clustering using k-means by various researchers employing different means. In beginning, the researchers worked by using the simple k-means algorithm and then in later years, various modifications were executed. The term k-means was at first used by James MacQueen in 1967 and the standard algorithm was first proposed by Stuart Lloyd in 1957 which is known as technique for pulse-code modulation, though it wasn't published outside of Bell Labs until 1982. In 1965, E. W. Forgy published essentially the same method, which is why it is sometimes referred to as Lloyd-Forgy [5]. The traditional K-means method produces exactly k different clusters, which have most possible distinction (Bradley and Fayyad, 1998). The usage of K-mean clustering and Vector Space model was employed by using the text data by considering it as high dimensional. It was said that the time taken for entire clustering process was linear in the size of document collection [Indrajit S. Dhillon et.al. 2001]. Some of the researchers found an effective technique for K-means clustering which proves that principal components are the continuous solutions to the discrete cluster membership [Chris Ding et.al.2004].

Recently, work on the performance of the partition clustering techniques in terms of complex data objects and comparative study of the cluster algorithm for corresponding data and proximity measure for specific objective function based on K-means and EM Algorithms was executed. Comparison and evaluation clustering algorithms with multiple data sets, like text, business, and stock market data was performed. Comparative study of clustering algorithms identified one or more problematic factors such as high dimensionality, efficiency, scalability with data size, sensitivity to noise in the data. [Satheelaxmi.G et.al.2012]. Work by integrating the constraints into the trace formulation of the sum of square Euclidean distance function of K-means by combining criterion function transformation into trace maximization was optimized by eigen-decomposition [Guobiao Hu et.al.2008]. Plentiful times work on K-means algorithm on WEKA model had been implemented in the past which in turn has improved the WEKA tool set. [Sapna Jain et.al.2010]. Previously, work based on several datasets, including synthetic and real data, show that the proposed algorithm may reduce the number of distance calculations by a factor of more than a thousand times when compared to existing algorithms while producing clusters of comparable quality was carried out [Maria Camila N. et.al.2006]. Proposal of a new language model, to simultaneously cluster and sum up at the same time has been implemented in past. The method implies good document clustering method with more meaningful interpretation and a better document summarization method taking the document context information into consideration [Wang, Shenghuo Zhu et.al.2008] [6].

3. K-means Algorithm

k-means algorithm is one of the simplest unsupervised learning algorithms that is used for solving most of the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set into a certain number of clusters. The main idea is to define k centre, one for each cluster. Basically this k-means algorithm is known as most efficient in clustering large data sets and the main aim of this algorithm is to partition the set of objects according to their attributes or features into number of k-clusters where this k-cluster may be user defined or predefined. The centre of these cluster which is also called as centroid of cluster is closely related to the all the other object in terms of some similarity. k-means algorithm is coming under the flat clustering and hard clustering method where they assign each document or object member of exactly one cluster and they create set of clusters without any explicit structure that would relate clusters to each other [4].

Basically this k-means algorithm assign each point of cluster whose center are nearest and the center of this cluster is the average of all the points of the clusters but to choose centroid of the all points in the cluster is very tough task because different location causes different result and therefore said that the center should be placed in cunning way. Suppose a data set D contains n objects in Euclidean space and the partitioning methods distribute the objects in D into k clusters such as C1, C2, C3, C4, C5, C6, C7, C8, C9, C10, where Ci is belonging to D and the intersection between Ci,
Cj is null for 1 ≤ i and j ≤ k. The main objective function of this k-means algorithm is to use partitioning quality that is object within the cluster are similar to one another and dissimilar to other cluster and main goal is to increase the intracluster similarity and decrease intercluster similarity. This k-means algorithm is also called as centroid based partitioning technique which uses centroid of the cluster, basically centroid of cluster is center point of the cluster where it is defined by the mean or medoids of the objects or points assigned to the cluster [7]

K-means algorithm firstly define the centroid of the cluster where it randomly selects the k cluster in the dataset D where each of this is initially represent a cluster mean or center. The other remaining objects are assigned to the cluster to which it is most similar which depends on the Euclidean distance between the object and centroid of the cluster. The k-means algorithm after that in each iteration improves within cluster variation that is for every cluster it computes the new mean using the cluster of the previous iteration. All the object here in this technique are then reassign using the updated mean as new cluster centers and in this technique iteration is continue until the assignment is stable that is the cluster formed in the previous iteration round are same as the current round [7]. According to the this the flow of this k-means algorithm is given by the following flowchart.

**Flowchart of the k-means algorithm**

The k-means algorithm for partitioning where each cluster’s centre represents mean value of the objects in the cluster and generally the input of the k-means algorithm are number of cluster represent by k and a data set containing n object represent by D where output of k-means algorithm is a set of k clusters. The k-means algorithm are given as follows.

1. Choose k number of cluster to be determined
2. Choose k objects randomly as the initial cluster centre
3. Repeat
   i. Assign each object to their closest cluster
   ii. Compute new cluster that is calculate mean points
4. Until
   i. No change on the cluster centres
   ii. No object changes its cluster.

**Fig. Clustering of set of objects using k-means algorithm**

To under the implementation of the k-means algorithm consider a set of objects as shown in following fig. and consider k=3 that is the user would like the objects to be partitioned into three clusters with its center of the cluster shown by symbol ‘+’. Each object is assigned to a cluster based on the cluster center to which it is the nearest. Such a distribution forms silhouettes encircled by dotted curves, as shown in figure2(c). Next, the cluster centers are updated. That is, the mean value of each cluster is recalculated based on the current objects in the cluster. Using the new cluster centers, the objects are redistributed to the clusters based on which cluster center is the nearest. Such a redistribution forms new silhouettes encircled by dashed curves, as shown in Figure2 (b). This process iterates, leading to Figure2 (c). The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as iterative relocation. Eventually, no reassignment of the objects in any cluster occurs and so the process terminates. The resulting clusters are returned by the clustering process.

**4. Strengths of k-means Algorithm**

K-means algorithm is the most popular clustering algorithm. K-means algorithm is simple, easily understandable algorithm and at the same time this algorithm is also fast and robust which is the one of the plus point of this algorithm over other clustering algorithm. K-means algorithm gives the best result when data set is distinct or well separated from each other. In k-means algorithm items are automatically assign to cluster and it is very efficient algorithm whose time complexity is given as O(tkn)
where $n$ is the number of data points, $k$ is the number of clusters and $t$ is the number of iterations. Since both $k$ and $t$ are small, k-means is considered a linear algorithm. In practice we notice that this k-means algorithm is performing well on text.

5. Problem of the k-means method

k-means algorithm at the same time has to face some of the problem in which first one is this k-means algorithm is sensitive to outlier which means that since an object with an extremely large value may be substantially distort the distribution of the data. Because outliers are data points that are very far away from other data points and it could be errors in the data recording or some special data points with very different values. The second one is use of Exclusive Assignment that is if there are two highly overlapping data then k-means will not be able to resolve that there are two clusters. K-means algorithm is failed to applicable not only for categorical data set but also for non-linear data set and at the same time this algorithm is unable to handle noisy data and outlier.

6. Conclusion

In this paper I discussed the concept of the text mining and document clustering. Here in this paper I also discuss and analyze the k-means algorithm together with its pitfall and strength which helps us to find the best clustering algorithm over other algorithm. As we discuss the k-means algorithm is typically used for finding quick solution as opposed to other clustering algorithm. Moreover, huge document size also gives an advantage to the K-means method because this improves the similarity measure. Hence, it has been ascertained in the end that the best cluster is obtained using K-means algorithm which can later be used for multi-document summarization. The summarization helps the user in saving their time by providing just the key points of a document which represents the category of the document because of which the user can easily retrieve the document they want to access.

7. References


[2] Mrs.S.C. Punitha1 and Dr.M. Punithavalli, “A COMPARATIVE STUDY TO FIND A SUITABLE METHOD FOR TEXT DOCUMENT CLUSTERING”,