Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 06, pp.33236-33239, June, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

K-MEANS CLUSTERING ALGORITHM ON TEXTUAL DATA WITH IMPROVED INITIAL CENTER

*Anupam Goel, Shashank Kumar and Dharmveer Singh Rajput

Jaypee Institute of Information Technology, Noida, India

ARTICLE INFO

Article History:
Received 24th March, 2016
Received in revised form
23rd April, 2016
Accepted 09th May, 2016
Published online 30th June, 2016

Key words:

Clustering, Data Mining, Data partitioning, Initial cluster centers, K-means clustering algorithm.
Cluster analysis.

ABSTRACT

In this modern era, a lot of data is available to all the organisations so the real task is to gather the useful information from the data. There is a need to develop a robust technique that can resolve this problem. Clustering is a technique in data mining that groups together the similar data items into cluster. The data in a cluster is more similar to each other than data in the other cluster. There are different types of clustering algorithms like K–means (partitioning based), greedy based, hierarchical based algorithms, density based. Clustering has a wide range of application such as text mining information retrieval, Business analytics, data analysis, machine learning etc. In this we are using K-means clustering, firstly we have discuss how K-means algorithm is implemented then we have discuss the advantages and disadvantages of K-means and then finally we have selected one of the shortcoming of K-means algorithm. Then we have proposed are own method to improve this shortcoming and then we applied this algorithm on textual dataset. Finally we get the result of our algorithm.

Citation: Anupam Goel, Shashank Kumar and Dharmveer Singh Rajput, 2016. "K-means clustering algorithm on textual data with improved initial center", *International Journal of Current Research*, 8, (06), 33236-33239.

INTRODUCTION

In data analysis the major problem is to separate the similar type of data and group this similar type of data in one cluster. In data analysis it is practically impossible to analyse each and every data point so we analyse the data by forming the clusters. Cluster analysis is widely used and primary data analysis method for practical application. Clustering is a technique in data mining that groups together the similar data items into cluster. The data in a cluster is more similar to each other than data in the other cluster. There are different types of clustering algorithms like K –means (partitioning based), greedy based, hierarchical based algorithms, density based. A lot of research has been made to improve clustering. K-means is one of the most widely use clustering algorithm. There are various practical applications of clustering like in Business and Marketing, Bioinformatics, Image Processing, Plagiarism etc. In this paper we are applying K-means algorithm on textual data set. K-means algorithm is very sensitive to the initial centers so we have to choose these centers very carefully, hence we are choosing these centres using a technique rather than choosing them randomly so there won't be any outlier or empty cluster in our dataset.

Then we are implementing this on textual data

The K-means algorithm

K-means is an unsupervised learning algorithm and it is one of the most widely use clustering algorithm. In K- means we randomly select k points as initial cluster centers. Then we put all other object in the nearest cluster by calculating distance between the object and the cluster center. After that K clusters are formed then we again calculate the centroid of each cluster, then again repeat this step of forming cluster by putting these object in the nearest cluster. We keep on repeating these step until cluster centers do not change.

The algorithm is composed of the following steps:

1. First select K points from the dataset, these points represent initial group centroids.
2. Assign each object to the cluster that is closest to the centroid.
3. When we have placed all the objects, then we recalculate the position of centroid of each cluster.
4. Repeat Steps 2 and 3 until the centroids no longer change. This produces a separation of the objects into cluster from which the metric to be minimized can be calculated.

***Corresponding author: Anupam Goel,**
Jaypee Institute of Information Technology, Noida, India.

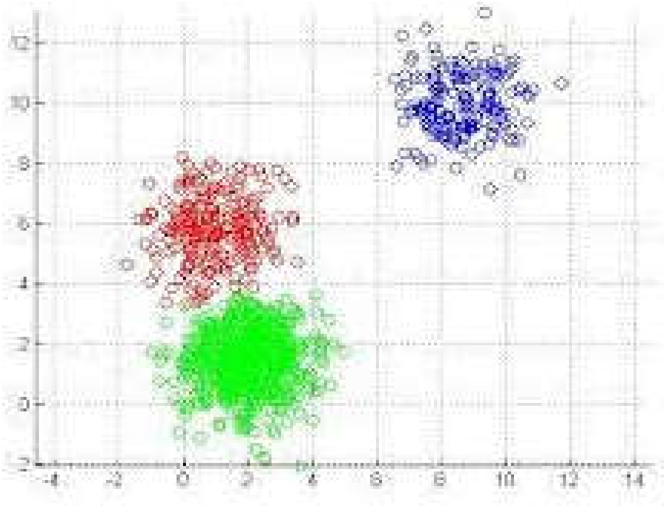


Fig. 1. Data points in Two Dimensional Space

Related work

In k-means algorithm it is very crucial to select initial K centroids of the cluster because if we get any outlier or a faraway point as a centroid then there are chances that this cluster would remain empty and we won't get accurate result. Many researchers have found various techniques and method to select initial cluster centers to increase the efficiency and accuracy of K-means. In traditional K-means algorithm the distance of each object is calculated from each cluster center then that object is assign to the nearest cluster, so the required computational time depends on the number of objects, number of initial centers and the number of iteration. (Fahim *et al.*, 2006) A researcher Fahim proposed a method for assigning objects to the cluster so that we can get better result for our algorithm. (Abdul Nazeer and Sebastian, 2009) Abdul Nazeer, this researcher proposed two method for increasing the efficiency and accuracy of K- means algorithm. First proposed method help in selecting initial K centroids and second method help us in accurately assigning the objects to the nearest cluster. These two algorithms reduces the time complexity and number of iterations, so it increases the efficiency of the K-means algorithm. (Yuan *et al.*, 2004) Fang Yuan is one more researcher who proposed a method in selecting initial cluster centers. If we select the centers randomly then there are chances of not getting result accurately, so he proposed his own method for selecting these cluster centers and this can affect the output of K-means algorithm. Rather selecting the centers randomly he prefer to select the centers symmetrically to get better result. (Chen Zhang and Shixiong Xia, 2009) Zhang Chen *et al.* proposed an algorithm for selecting initial centroids of the cluster and he avoided randomly selection of the initial cluster centers. A lot of research work has been made on Clustering (K-means) improvement. Initially inappropriate choice of no of cluster (Pham *et al.*, 2004) and inappropriate center selection take more iteration and sometimes produces empty clusters. (Ball and Hall) firstly produces the approach of centroid initialization in 1967. After initializing the initial centers we came across another problem for selection of initial centers. In K-means algorithm it is very crucial to have refine initial cluster centers. And many methods were proposed in the literature to find initial centers which improve both the accuracy and efficiency of the k-means clustering

Proposed algorithm

Document Formation

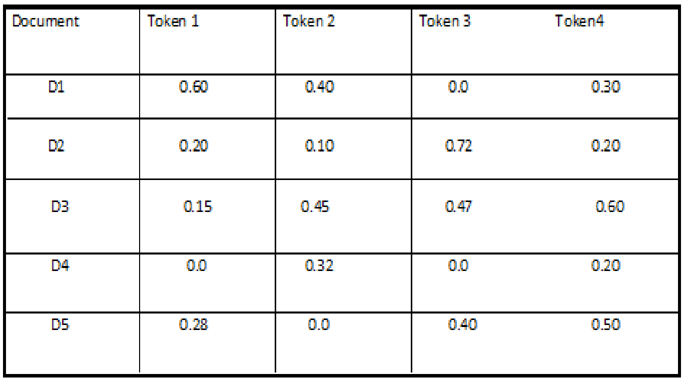
Firstly we select the data for clustering we gather different type of textual dataset for clustering and form different documents of gathered data.

Tokens Formation

After the document formation tokens are generated from each document. We remove stop words punctuation, numbers from tokens. Each token of every document has its own part of speech.
Many part of speech taggers have been developed in natural language processing. We use these classifiers for part of speech tagging of each token.

We remove the tokens having preposition, conjunction, interjection, pronoun as their tags. A dictionary containing all these remaining tokens is formed. This dictionary helps to calculate the frequency of all tokens. The tokens reside inside their own documents.

Document token matrix:



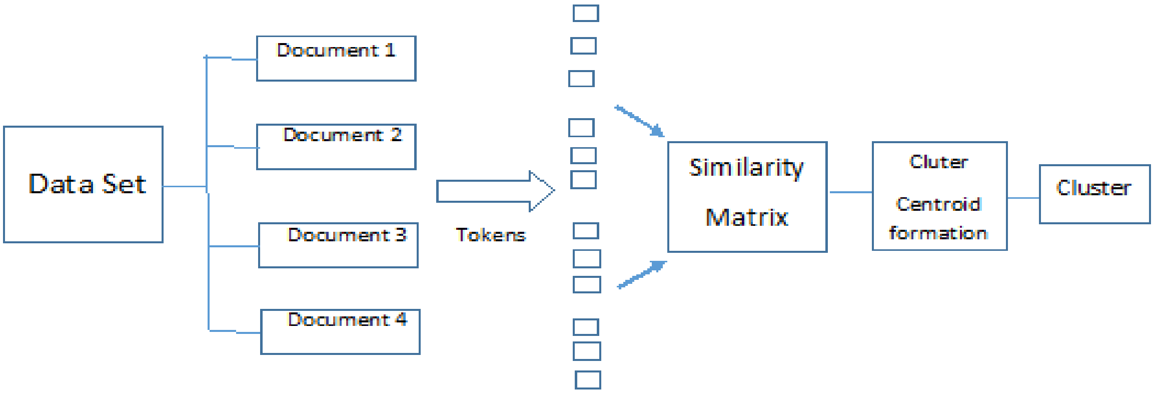
Document	Token 1	Token 2	Token 3	Token4
D1	0.60	0.40	0.0	0.30
D2	0.20	0.10	0.72	0.20
D3	0.15	0.45	0.47	0.60
D4	0.0	0.32	0.0	0.20
D5	0.28	0.0	0.40	0.50

We further choose all total unique tokens from all documents. We put these tokens as attribute of our matrix and all documents as their rows. We map frequency of each token in each document.

Document Token Matrix

Similarity matrix:

We form similarity matrix between documents by using the cosine similarity formula:
Cos(d1,d2)={vector(d1)dotproductvector(d2)}/mod(d1) mod(d2)
Suppose we have n documents (d1,d2,d3,d4……….dn).
Then we get n*n matrix (m(n)(n)).In this each value of matrix denotes the similarity between two documents.
Initial k-centers:



The main task in k-means clustering algorithm is to identify the initial k centers. In classical k-means clustering algorithm we randomly select the initial centers. Although it work well but sometime the results are not satisfactory because we get empty clusters. We propose a method to select initial k centers. We choose a set X of k initial centers from the available document set (d1,d2,d3,…… dn).
k is the no of clusters.
N is the no of documents.

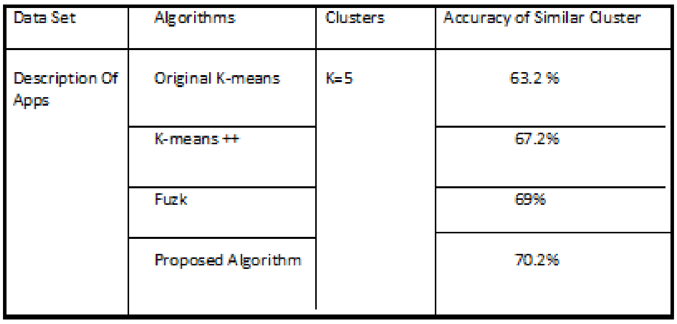
Step-1: Calculate the Similarity of each document with all other documents and add these distances as Similar_sum(i).
Step-2: Select the maximum value of Similar_sum and then find the highest density point dh.
Step-3: Add dh to X as the first centroids.
Step-4: For every document di set p(di) to be the distance between di and the nearest point in X.
Step-5: Find y as the sum of distances of first m/k nearest points from the dh.
Step-6: Find the unique integer I so that
Step7: $p(d1)^2+p(d2)^2+……..+p(di)^2>=y>p(d1)^2+p(d2)^2+……..+p(di-1)^2$.
Step-8: Add di to C.
Step-9: Repeat steps 4-7 till desired no of clusters not come.

Cluster formation:

We check the similarity of each document with these cluster centers and initialize the cluster centers with documents having higher similarity. Assume that m documents are distributed uniformly to k (number of clusters) clusters then each cluster is expected to contain m/k documents.

Experiment Result

To test this algorithm we crawl the description of 100 different app from the google play and form the document of each app description. Our main motive was to group the similar type of app in same cluster. We select 20 tokens of maximum frequency from each document. Then we check the similarity of the documents and form the cluster. We initially select 5 clusters. The performance of algorithm to form the cluster is compared with other existing algorithm like original K-means, K-means++. The comparison results are relevant



Data Set	Algorithms	Clusters	Accuracy of Similar Cluster
Description Of Apps	Original K-means	K=5	63.2 %
	K-means ++		67.2%
	Fuzk		69%
	Proposed Algorithm		70.2%

Output result

C1: app1, app4, app5 ……….
C2: app6, app10,
app13……….. C3:app2,
app3, app7…………….
C4:app11, app12,
app15…………. C5:app17,
app19, app21………….

Conclusion

One of the most popular clustering algorithm is k- means clustering algorithm, in this method we have selected the initial centroids by our well define technique and not randomly to increase the efficiency of the clustering algorithm. In this we have applied K- means algorithm on textual data. In this we first tokenize the words of all document and then form the similarity matrix. Then we have form clusters of document. We have tested our algorithm on 100 app description and group together similar type of app in one cluster. This can also help us in finding malicious app within a cluster. Our future plan is to increase the efficiency of selecting initial K centroids to get better result and also to improve the algorithm for calculating the similarity matrix of the document. This would improve the cluster formation. We would also try to check our algorithm on well define dataset to check the efficiency of our algorithm.

REFERENCES

Abdul Nazeer K. A. and M. P. Sebastian, “Improving the accuracy and efficiency of the k-means clustering algorithm,” in *International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of* *the World Congress on Engineering (WCE-2009),* Vol 1, July 2009, London, UK.

Bhattacharya A. and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles," *bioinformatics*, Vol. 24, pp. 1359-1366, 2008.

Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," *in Second International Workshop on Knowledge Discovery and Data Mining* (WKDD), pp. 790-792, 2009.

Deelers S. and S. Auwatanamongkol, "Enhancing K- Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," *International Journal of Computer Science,* Vol. 2, Number 4.

Diggle Data. 2010. Available: http://lib.stat.cmu.edu/ datasets/diggle

Elmasri, Navathe, Somayajulu, Gupta, *Fundamentals of Database Systems,* Pearson Education, First edition, 2006. (11) (2010) The UCI Repository website. (Online). Available: http://archive.ics.uci.edu/

Fahim A. M., A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," *journal of Zhejiang University,* 10(7): 16261633, 2006.

Height-Weight Data, 2010. (Online). Available: http://www.disabledworld.com/artman/publish/height-weight-teens.shtml

Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," *department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University,* Vol. 36, No.1, 2007.

Margaret H Dunham, Data *Mining-Introductory and Advanced Concepts,* Pearson Education, 2006.

Mc Queen J, "Some methods for classification and analysis of multivariate observations," *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, (1): 281–297, 1967.

Yuan F., Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," *proceedings of the 3rd International Conference on Machine Learning and Cybernetics,* pp. 26-29, August 2004.

****