International Journal of Computer Engineering and Applications, Volume IX, Issue X, Oct. 15
www.ijcea.com ISSN 2321-3469

INDEX BASED IMPROVED K-MEANS ALGORITHM FOR TEXT CLUSTERING

Swati. D. Shejale[1], Prof. S. B. Vanjale[2]

[1]*M.Tech. Student, Department of Computer Engineering, BVDUCOE Pune, India*

[2]*Ph.D. Research Scholar, Department of Computer Engineering, BVDUCOE Pune, India*

ABSTRACT:

Clustering is the group of data this group of data belongs to the same class. The data in one cluster are similar to each other while data in one cluster are dissimilar to data in another cluster. K-means is most popular partition based clustering technique because of its speed and simplicity. In K-mean clustering technique, the classification is done very quickly, smoothly and efficiently. There are two limitations in basic K-means algorithm, one is the output of K-means algorithm clustering depends upon the initial cluster centers, in this initial cluster centers are chosen randomly and second limitation is the number of clusters are required as input. Because of the requirement of domain knowledge the value of number of clusters K cannot be easy to predict. In this approach we have devised a technique where we eliminate the selection of initial center and user input of number of clusters. To achieve this goal an Erasure encoding based indexing mechanism will be used.

Keywords: Data mining, clustering, k-means clustering algorithm, indexing, compressed indexing.

[1] INTRODUCTION

The huge amount of data is available in industries (IT environment). Analyzing and extracting useful information from this huge amount of data is necessary because these data is useless if it is not converted into the useful information. Data mining is the process of extracting information from huge amount of data set or mining the knowledge from data. Need of data mining: In the field of Information technology, the huge data is available, so there is need to convert these huge amount of data into useful information. This useful information can be used for variety of applications such as market analysis, fraud detection, customer retention, production, control, science exploration etc. The idea deal with the mining knowledge from data is data mining. Some data mining techniques are Frequent Pattern Matching, Association Rule analysis, classification, Clustering [1].

The process of creating indexes for collection of records is called as indexing. Text indexing is the basic problem in computer science, where task is to index the specific text from given documents. The indexing mechanism is simple mechanism but it will not be simple mechanism

when it will be highly compressed and size will be small. This will reduce memory consumption and computational cost also.

Clustering is he group of data, group of data belong to the same class. The data in one cluster are similar to each other while data in one cluster are dissimilar to data in another cluster. This grouping process is called as clustering. Clustering analysis is greatly used in many fields including classifying documents on the web for discover distinct groups in their customer basis, in outlier detection application such as detection of credit card fraud [2].

One of the most popular techniques of clustering is K-mean. K-means is a partition based clustering technique. It is most popular because of its speed and simplicity. The high quality clusters are produced by good clustering methods, the good clustering methods such as intra-cluster (class) and inter-cluster. The similarity is high in intra cluster method and similarity is low in inter-cluster methods. The cluster quality is depends on similarity measure and hidden patterns.

K-mean is most popular partition based clustering technique of its speed and simplicity. K-means is distance based partitioning technique and another widely used distance based partitioning technique is K-medoid clustering algorithm. In K-mean clustering technique, the classification is done very quickly and smoothly, efficiently. There are two limitations in basic k-means algorithm, one is the output of K-mean clustering algorithm depends upon the initial cluster centers, in this the initial cluster centers are chosen randomly and second limitation is the number of clusters are required as input. Because of requirement of domain knowledge the value of number of cluster K cannot be easy to predict. So one of main limitation excluded in proposed system by modifying K-means algorithm which requires number of clusters K is input. So there is no need to define number of clusters as input. And the indexing algorithm is used for indexing purpose which is helpful to reduce memory consumption and computational cost.

[2] BASIC K-MEANS

K-means is used for solving eminent clustering problems and is one of the simplest and easy unsupervised learning algorithm. Classification is done with given data sets into certain number of clusters. First have to initialize cluster centers K. The points in the dataset which has minimum distance from cluster centers, assign these points to that cluster center. Repeat procedure until all points in each cluster are at the minimum distance from their cluster centers.

Algorithm 1: Basic K-means

Let D be the set of data and K be the number of clusters.

1. Randomly select k initial cluster centers from set of data D
2. Compute distance between each point and cluster centers by Euclidean distance

Euclidean Distance,

$$d_{euc}(x, y) = \sum_{i=i} \sqrt{(x_i - y_i)^-} \qquad (1)$$

3. Assign calculate the new cluster center
4. Again find the distance between each data point and new calculated cluster centers.
5. If no data point was reassigned then stop the process otherwise repeat from step 4.[1]

[3] PROPOSED SYSTEM

In our approach of Text Clustering we have devised a two step mechanism where the whole text of the document will be first indexed and then apply modified K-mean algorithm on the indexes for removing dependency on the K. The indexing mechanism we apply will not be a simple indexing mechanism rather it will be highly compressed and thus the index size will be small. This will not only reduce the memory consumption for storing the indexes but also reduce the computational cost while performing the clustering of the relevant data.

The main components of system architecture are analyzing documents, feature extraction, compressed indexing, clustering documents. The working of these components is described below

- Analyze Document
- Feature Extraction
- Compressed Indexing
- Clustering

Documents: In this stage, the system analyzes the input documents, analyzing the documents also called as document preprocessing. In document [reprocessing first find the stop words in the input documents and then remove them. For this process we used the stop word Removal Algorithm. Simply he stop word means common words such as “a”, “the”, “and”, “is” etc

Feature Extraction: Feature extraction is the process of determining set of features which are most simply and meaningfully express and the documents which is important for analysis and clustering. When two or more documents have same meaning then this documents are combined which has grammatical similar meaning such as “play”, “playing”, “played” etc (streaming).

Compressed Indexing: The word compress generally means “abbreviate or wrap”. Similarly the meaning of compressed indexing is wrapping the large number of documents into minimum number of documents. The compressed indexing is done after indexed text. And output will be the suitable number of clusters distributed properly.

The improved K-means with indexing algorithm is as follows:

Algorithm 2: Improved K-means

Input: Nmber of documeents A1,A2,….Am.
M is number of documents.

Output: Compressed text index.
Number of clusters distributed properly.

Method:

1. Create cluster center using K-means clustering algorithm in n-dimensional vector space.
2. Similarity of document is measure using cosine similarity.

3. Document cluster is prepared and grouping of similar type of text document is done here

K is initial cluster center

Document collection is docment corpus

4. K initial centroids are selected and randomly attact one object t each centroid.

5. Avoid repeation of random number, if same no is generated more than once same document is added to the next custer so avoid uing HasSet collecton.

6. Initialize the result set for next iteration

7. Generate unique random number and also ensure the generated random number lies with in a range of total number of documents

"uniqRand"

"k"

"docCount"

8. Initialize the result cluster centroid for the next iteration, that holds the result to be returned

"centroid"

"count"

9. Check the stopping condition for the iteration, if centroid do not move their position it means it meets the condition or it the global counter exist its predefined limit (minimum iteration threshold) then iteration terminates

"prevClusterCenter"

"newClusterCenter"

10. 1 equal to centroid has moved and 0 equal to centroid do not moved its position

11. If index contain 1

Then returns index of closest cluster centroid

12. Returns index of closest cluster centroid

13. If document is similar then assign the document to the lowest index cluster center to avoid the long loop

14. Reposition the centroid

15. Reassign new calculated mean on each cluster center, it indicates the reposition of centroid

16. Find Residual Sum of Square it measures how well a cluster centroid represents the member of their cluster

17. Use Residual Sum of Square value as stopping condition of k-means algorithm when decreases value falls below a threshold t for small t, terminate the algorithm.

[4] EXPERIMENTAL RESULTS

The k-means need to input nuber of clusters k and output of K-means totally depends on th initial cluster centers, initial cluster centers are chosen radomly. In proposed system we removed

this lmitations successfully so that the efficiency of the system increased. The proposed algorithm is more efficient because it does not require number of cluster K as input.

The experimental result of index based improved K-means algorithm for text clustering is as follows:

First we input the number of documents

Travel
Different
Traveling
Document
Place
Study
Text
Learn
Placed
Grasp

Table2. Input number of documents

Then add it, after adding these documents the index will be created using counting, locating and extracting.

1.	Travel, traveling
2.	Different
3.	Document
4.	Place, placed
5.	Study, learn, grasp
6.	Text

Table3. Output Indexed documents

Now secondly apply improved k-means on indexed document and clustering output is

Cluster1	Different, Document, Place, Study, Text
Cluster2	Travel

Table4. Output clusters

If there is no change in clustering then those will be accurate clusters so the number of iterations performed till documents are stopped moving. Recalculating the centroids until no improvement in accurcy.

[5] CONCLUSION

There are two limitations in basic K-means algorithm, one is the output of K- mean clustering algorithm depends upon the initial cluster centers, in this the initial cluster centers are chosen randomly and second limitation is the number of clusters are required as input. Because of the requirement of domain knowledge the value of number of clusters K can not be easy to predict. In this approach we have devised a technique where we eliminate the selection of initial center and user input of number of clusters. To achieve this goal an Erasure encoding based indexing mechanism will be used. Our work in this paper is limited to categorical data set. In our future work we will try to extend this algorithm for mixed data set i.e. numeric as well as categorical. Also work efficiency of this algorithm can be improved by checking different methods to calculate initial centroid.

REFERENCES

[1] Anupama Chadha, Suresh Kumar, "An Improved K-Means Clustering Algorithm:A Step Forward for Removal of Dependency on K" International Conference on Reliability, Optimization and Information Technology - ICROIT 2014, India, Feb 6-8 2014, pages 136-140, 2014 IEEE.

[2] B M Ahamed Shafeeq, K S Hareesha, " Dynamic Clustering of Data with Modified K-Means Algorithm", International Conference on Information and Computer Networks (ICICN 2012), IPCSIT, vol. 27,pages 221-225, 2012.

[3] Charu C. Aggarwal, ChengXiang Zhai, "A SURVEY OF TEXT CLUSTERING ALGORITHMS".

[4] Paolo Ferragina,Rodrigo Gonzalez, Gonzalo Navarro , Rossano Venturini, "Compressed Text Indexes:From Theory to Practice!".

[5] Kehar Singh, Dimple Malik, Naveen Sharma, "Evolving limitations in K-means algorithm in data mining and their removal", IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011 ISSN (Online): 2230-7893.

[6] N C Chauhan, Shalini S Singh, "K-means v/s K-medoids: A Comparative Study", National Conference on Recent Trends in Engineering & Technology, 13-14 May 2011.

[7] Moshe Lewenstein, Yakov Nekrich, Je_rey Scott Vitte, "Space-Efficient String Indexing for Wildcard Pattern Matching".

[8] Manjot Kaur, Navjot Kaur, "Web Document Clustering Approaches sing K-Means Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X.

[9] Kusum Bharti, Shweta Jain, Sanyam Shukla, "Fuzzy K-mean Clustering Via For Intrusiion Detection System", Kusum Bharti et al. / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (4) , 2010, 315-318.

[10] Kohei Arai, Ali Ridho Barakha, "Hierarchical K-means: An algorithm for centroids initialization for K-means", Reports of the Faculty of Science and Engineering, Saga University, vol. 36, issue. 1, pages 25-31, 2007.

[11] M. AI Dauod, "A New Algorithm for Cluster Initialization", World Academy of Science, Engineering and Technology, issue 4 ,2007.

[12] Mohamed Abubaker, Wesam Ashour, "Efficient Data Clustering Algorithms: Improvements over K-means", International Journal of Intelligent Systems and d Applications, vol. 5,issue 3, pages 37-49, 2013.