ISSN:0975-9646

# Efficient Clustering of Text Document Using spherical K-means algorithm

A. Ramana Lakshmi, $_{\text{M.Tech, (Ph.D)}}$ #1, V.Balakrishna, $_{\text{(M.Tech)}}$ *2

#Associate professor, Department of CSE, PVPSIT, Kanuru, India
*Student, Department of CSE, PVPSIT, Kanuru, India

**Abstract -- The problem of text clustering arises in many application domains such as the web applications, network applications, and other digital collections. Enormously increasing amounts of text data in the substance of these large collections has led to a fascination in creating scalable and effective mining algorithms. Clustering is especially useful for organizing documents to enhance retrieval and support browsing. Many text documents has text data along with other auxiliary attributes, that are also known as side information or Meta information. That side information may be useful for clustering purpose or may be harmful as it has noisy attributes. The spherical k-means algorithm, i.e., the k-means algorithm with cosine similarity, is a popular method for clustering high-dimensional text data. In this algorithm, each document as well as each cluster mean is represented as a high-dimensional unit-length vector.**

**Keywords-- Clustering, k-means, spherical k-means, distance.**

## I. INTRODUCTION

Clustering is a popular strategy for implementing parallel processing applications because it enables companies to leverage the investment already made in PCs and workstations. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data[4].

Efficient content mining is an important task for Web communities. Most of existing information is stored on the Web. Its electronic format makes it easy searchable by means of informational technologies. Such a way Web becomes an electronics repository of human knowledge. As the information on the Web has no fixed structure the web users has to rely on dynamic, learning based methods to get efficient access to needed information.

Achievement of better efficiency in retrieval of relevant information from an explosive collection of data is challenging. In this context, a process called document clustering can be used for easier information access. The goal of document clustering is to discover the natural grouping(s) of a set of patterns, points, objects or documents. Objects that are in the same cluster are similar among themselves and dissimilar to the objects belonging to other clusters. The purpose of document clustering is to meet human interests in information searching and understanding. The challenging problems of document clustering are big volume, high dimensionality and complex semantics. Our motive in the present paper is to extract particular domain of work from a huge collection of documents using K-Means and K-Medoids clustering algorithm and to obtain best clusters which later on can be used for document summarizations.

## II. K-MEANS CLUSTERING

Suppose a data set, D, contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, $C_{1,}\ldots C_k$, that is, $C_i$ D and $C_i \cap Cj = \emptyset$; for($1 \leq i, j \leq k$). An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters [3]. This is, the objective function aims for high intracluster similarity and low intercluster similarity.

$$E = \sum_{i=1}^{k} \sum_{p \epsilon c_i} \mathrm{dist}(p, c_i)^2$$

### Distance Metrics:

In order to measure the similarity or regularity among the data-items, distance metrics plays a very important role. It is necessary to identify, in what manner the data are inter-related, how various data dissimilar or similar with each other and what measures are considered for their comparison. The main purpose of metric calculation in specific problem is to obtain an appropriate distance /similarity function.

Euclidean distance computes the root of square difference between co-ordinates of pair of objects.

$$Dist_{x,y} = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$$

## III.SPHERICAL K-MEANS

The spherical k-means algorithm" of is a simple fixed-point heuristic for minimizing $\sum_i(1-\cos(x_i, p_{c(i)}))$ which iterates between computing optimal cluster ids for fixed prototypes and computing optimal prototypes for fixed cluster ids. A centroid-based partitioning technique uses the centroid of a cluster, $C_i$, to represent that cluster. Conceptually, the centroid of a cluster is its centre point [4]. The centroid can be defined in various ways such as by the mean or medoid of the objects (or points) assigned to the cluster.

### Cosine Similarity:

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle [2]. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

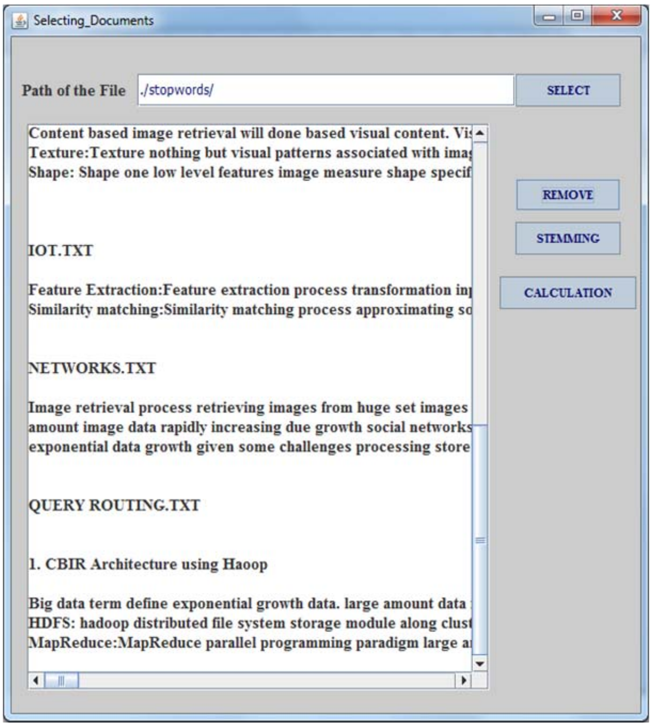
## IV. ALGORITHM

Step 1: Read the dataset with 'n' number of documents.
Step 2: Remove stop words from each document.
Step 3: Apply stemming to reduce inflectional forms.
Step 4: Compute term frequency of each document.
Step 5: Apply k-Means algorithm.
Step 6: Find Euclidian distance between each cluster.
Step 7: Apply spherical K-Means algorithm based on centroids.
Step 8: Find purity factor.
Step 9: Query search for results.

## V. IMPLEMENTED WORK

### a) Stopword removal:

Collected documents contain some unnecessary words by which dimensionality of a document will be increased; we should remove those words to get proper result. Pronoun, adverb, preposition etc. which are used constantly throughout in a document has to be removed.



Path of the File ./stopwords/

Content based image retrieval will done based visual content. Vi
Texture:Texture nothing but visual patterns associated with ima
Shape: Shape one low level features image measure shape specif

IOT.TXT

Feature Extraction:Feature extraction process transformation in
Similarity matching:Similarity matching process approximating sc

NETWORKS.TXT

Image retrieval process retrieving images from huge set images
amount image data rapidly increasing due growth social networks
exponential data growth given some challenges processing store

QUERY ROUTING.TXT

1. CBIR Architecture using Haoop

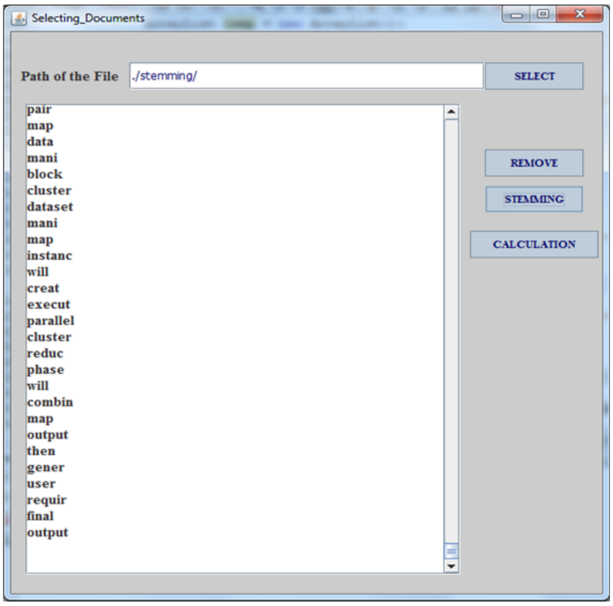
Big data term define exponential growth data. large amount data
HDFS: hadoop distributed file system storage module along clust
MapReduce:MapReduce parallel programming paradigm large a

Fig.1. Removing stopwords from each cluster

Fig.1. shows removing of stopwords from the text files.

### b) Stemming:

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing.



Path of the File ./stemming/

pair
map
data
mani
block
cluster
dataset
mani
map
instanc
will
creat
execut
parallel
cluster
reduc
phase
will
combin
map
output
then
gener
user
requir
final
output

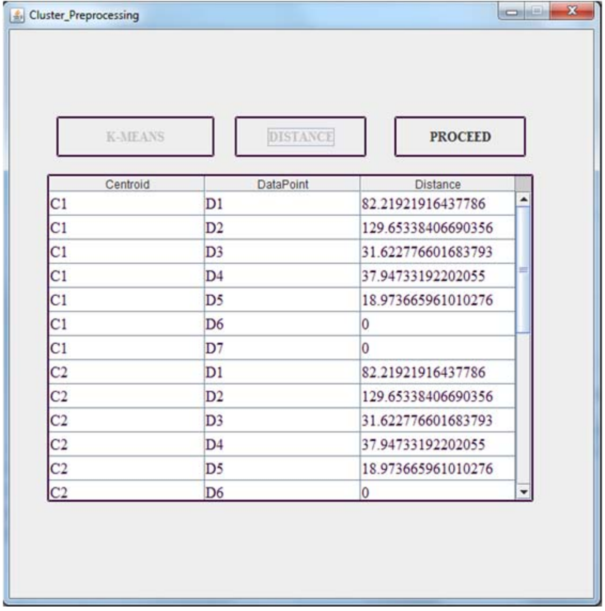
Fig.2. Apply stemming to reduce inflectional forms

Fig.2. shows applying stemming to reduce inflectional forms from the text files.

### c) K-means:

Suppose a data set, D, contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, $C_1, \ldots C_k$, that is, $C_i \subset D$ and $C_i \cap Cj = \emptyset$; for$(1 \leq i, j \leq k)$. An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is, the objective function aims for high intracluster similarity and low intercluster similarity.

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} \mathrm{dist}(p, c_i)^2$$



| Centroid | DataPoint | Distance |
|---|---|---|
| C1 | D1 | 82.21921916437786 |
| C1 | D2 | 129.65338406690356 |
| C1 | D3 | 31.622776601683793 |
| C1 | D4 | 37.94733192202055 |
| C1 | D5 | 18.973665961010276 |
| C1 | D6 | 0 |
| C1 | D7 | 0 |
| C2 | D1 | 82.21921916437786 |
| C2 | D2 | 129.65338406690356 |
| C2 | D3 | 31.622776601683793 |
| C2 | D4 | 37.94733192202055 |
| C2 | D5 | 18.973665961010276 |
| C2 | D6 | 0 |

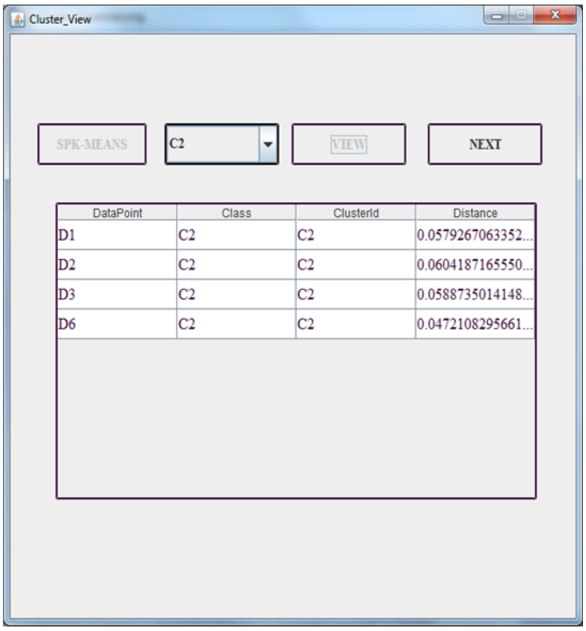
Fig.3. calculating k-means with distance

Fig.3. shows calculating Euclidian distance for clusters.

### d) Spherical k-means

The spherical k-means algorithm" of is a simple fixed-point heuristic for minimizing, which iterates between computing optimal cluster ids for fixed prototypes and computing optimal prototypes for fixed cluster ids. A centroid-based partitioning technique uses the centroid of a cluster, $C_i$, to represent that cluster. Conceptually, the centroid of a cluster is its centre point. The centroid can be defined in various ways such as by the mean or medoid of the objects (or points) assigned to the cluster.

$$\sum_i(1-\cos(x_i, p_{c(i)}))$$



| DataPoint | Class | ClusterId | Distance |
|---|---|---|---|
| D1 | C2 | C2 | 0.0579267063352... |
| D2 | C2 | C2 | 0.0604187165550... |
| D3 | C2 | C2 | 0.0588735014148... |
| D6 | C2 | C2 | 0.0472108295661... |

Fig.4. calculating spherical k-means clustering

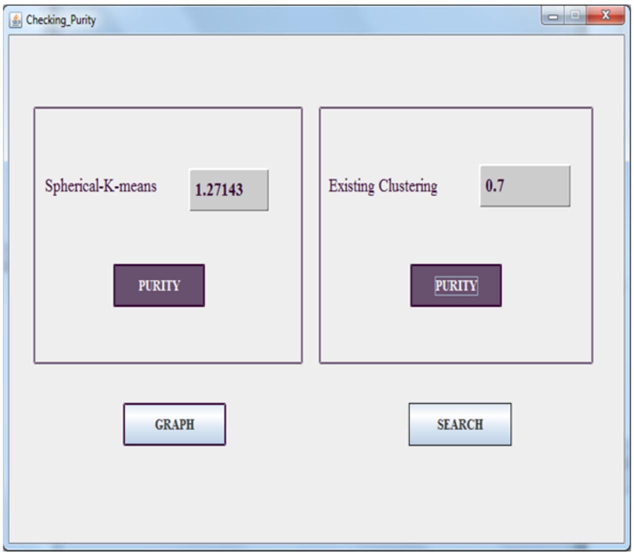
Fig.4. shows performing spherical k-means clustering with cosine similarity measurement.

### e) Purity

Purity is a one of very important validation measure to determine the cluster quality. Based on the purity factor efficiency of a cluster can evaluate. Here, the cluster efficiency is evaluated between standard k-means and spherical k-means [1]. The purity factor formula has given below

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N}\sum_k \max_j |\omega_k \cap c_j|$$

Where $\Omega = \{w_1, w_2, \ldots, w_k\}$is the set of clusters and $\mathbb{C}=\{c_1,c_2,\ldots,c_j\}$ is the set of classes. We interpret $w_k$ as the set of documents in $w_k$ and $c_j$ as the set of documents in $c_j$ in Equation.



Spherical-K-means 1.27143

Existing Clustering 0.7

Fig.5. calculating purity factor

Fig.5. shows the calculating purity factor between k-means clustering and spherical k-means clustering.

### f) Bar graph:

Pictorial representation of purity measure of k-means and spherical k-means.

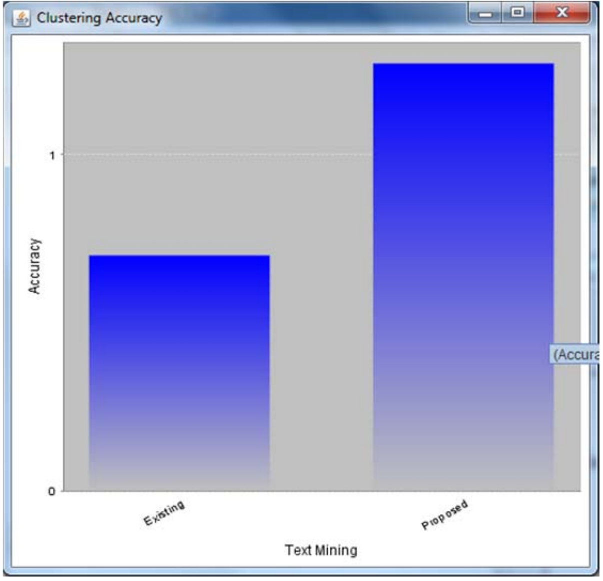
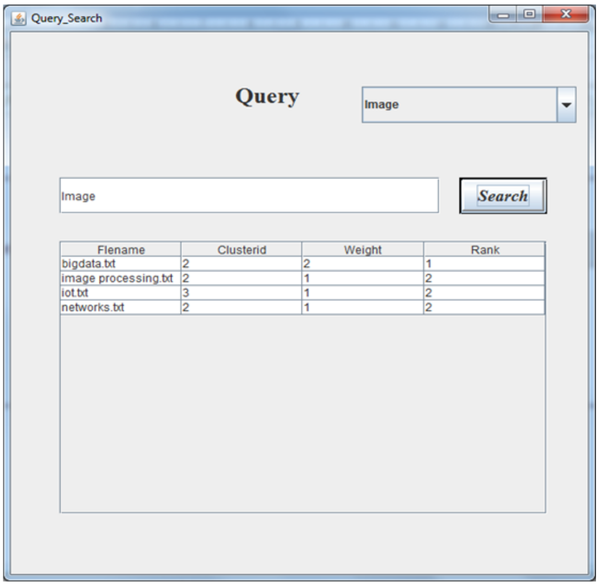


Fig.6. bar graph for purity factors

Fig.6. shows the bar graph for purity factors of k-means and spherical k-means.

### g) Query search

Searching each word from database in query search the results will be displayed with clustered and weight rank of the term and the documents having the search term.



| Flename | Clusterid | Weight | Rank |
|---|---|---|---|
| bigdata.txt | 2 | 2 | 1 |
| image processing.txt | 2 | 1 | 2 |
| iot.txt | 3 | 1 | 2 |
| networks.txt | 2 | 1 | 2 |

Fig.7. performing query search from database

Fig.7. shows the query search results of each term in the given dataset.

## CONCLUSION

Spherical k-means clustering is a central technique for addressing current data analysis challenges, especially in the context of large collections of text documents. Solving spherical k-means clustering problems corresponds to finding optimal group memberships employing the cosine similarity measure. By finding purity factor for both standard k-means and spherical k-means we concluding that spherical k-means is more efficient for grouping large context datasets.

## ACKNOWLEDGEMENT

I sincerely thanks to Mrs. A.Ramanalakshmi, Associate professor, department of C.S.E, PVPSIT for supporting to complete this project successfully. We are also grateful to our college staff for their assistance ship both technical and nontechnical.

## REFERENCES

[1] satya chaitanya sripada, dr. m.sreenivasa rao, “comparison of purity and entropy of k-means clustering and fuzzy c means clustering”, IJCSE, Vol. 2 No. 3 Jun-Jul 2011.

[2] Kurt Hornik, Ingo Feinerer, Martin Kober, Christian Buchta, “Spherical k-Means Clustering”, JSTATSOFT, September 2012, Volume 50, Issue 10.

[3] Rakesh Chandra Balabantaray, Chandrali Sarma, Monica Jha, “Document Clustering using K-Means and K-Medoids”, IJKBCS, Volume 1 Issue 1 June 2013.

[4] Jiawei Han, Micheline Kamber,Jian Pei, “Data Mining Concepts and Techniques”

[5] Third Edition MORGAN KAUPHANN.

[6] Sachin Shinde, Bharat Tidke, “Knowledge Discovery for research Documents using Improved K-means Technique”, ACM September-2015

[7] Subamanikandan A, Arulmurugan R, “On the Use of Side Information for Text Mining using Clustering and Classification Techniques-A Survey” , Volume-3, Issue-11 November, 2014

[8] JinHuaXu, HongLiu, “Web User Clustering Analysis based on KMeans Algorithm”, International Conference on Information, Networking and Automation (ICINA) 2010

[9] http://www.webopedia.com/TERM/C/clustering.html

[10] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/