Volume 2, Issue 6 June 2016 ISSN NO: 2454 - 7514

North Asian International Research Journal Consortium

North Asian International Research Journal

Of

Science, Engineering and Information Technology

IRJIF IMPACT FACTOR: 3.01

Welcome to NAIRJC

ISSN NO: 2454 -7514

North Asian International Research Journal of Science, Engineering & Information Technology is a research journal, published monthly in English, Hindi. All research papers submitted to the journal will be double-blind peer reviewed referred by members of the editorial board. Readers will include investigator in Universities, Research Institutes Government and Industry with research interest in the general subjects

Editorial Board

Address: - Dr. Ashak Hussain Malik House No. 221 Gangoo, Pulwama, Jammu and Kashmir, India - 192301, Cell: 09086405302, 09906662570, Ph. No: 01933-212815,
Email: nairjc5@gmail.com, nairjc@nairjc.com, info@nairjc.com Website: www.nairjc.com

IRJIF IMPACT FACTOR: 3.01

CLUSTERING OF TEXT FILES USING DISTANCE FORMULAE WITH SEARCH OPTIMIZATION

[1] VARSHA ANDHALE, [2] PRAJKTA VARHADE, [3] KOMAL KOLI, & [4] ROSHNI GAWANDE
Nutan Maharashtra Institute of Engineering and Technology, Talegaon Dabhade, Department of Computer Engineering , Savitribai Phule Pune University, India

***Abstract*—This paper presents the results of an experimental study of some common document clustering techniques. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. previously, document clustering was invented for improving the mining in information from database and as an easy way of finding the closer data of a document. Clustering algorithms are normally used for scanning the data, where there is very less information rare about the records. This technique can be applied in many application of computer forensics. In today's world we have large amount of digital data of text resources over internet and a digital library, organizing this data has become a prior need. Clustering is a technique which organizes large number of data objects into small number of similar data objects groups.**

Keywords*—*Clutering Techniques; Forensic;Text Mining.

I. INTRODUCTION

The procedure of clustering has been learned deeply in the database and statistics literature in the context of a wide variation in data mining work. The clustering problem is defined to be that of finding groups of same objects in the database. The common properties between the different data objects is calculated with the use of a similarity function. The technique of clustering can be very helpful in the data mining domain, where the objects to be clusters can be of different which having commons such as documents, paragraphs, sentences or terms. Clustering is especially useful for arranging documents to increase text mining and support searching and browsing. Among clustering formulations that are based on minimizing a formal objective function, perhaps the most algorithm used and learned is k-means clustering. . Measures such as Cosine similarity have been suggested and mostly used for distance measure and have proven to be successful. In this experiment, we do deep study on two similarity measures namely Euclidean Measure, Cosine Similarity using K-means Algorithm.

A. K-Means Algorithm:

K-Means is one of the easiest unsupervised learning algorithms are the solution of the well known clustering problem. The objective function of K- means is to select minimum average squared distance of objects from their centroid, where a document of file is a center defined as the *center* μ of the data objects in a *cluster K*. The result of K-means is greatly dependent on the initial first selection of center. Even with deterministic process, different starting states lead to different results.

B. Hierarchical Algorithm:

Hierarchical algorithm generates a nested sequence of parts, with a singular, all including cluster at the high level and single clusters of separate points at the low level. Each intermediate level can be viewed as joining two clusters from the next bottom level (or splitting a cluster from the next highest level). The result of a hierarchical clustering technique can be displayed graphically as tree, called a dendogram. This tree graphically shows the combining process and the intermediate clusters. There are two basic types to generating a hierarchical clustering.

a) Agglomerative: Start with the individual point of clusters and, at each step, concat the most similar or nearest pair of clusters. This must know a definition of cluster equality or distance.

b) Division: Start with one cluster, all including cluster and, at each level, break a cluster until only single clusters of individual points kept remain as it is. In this matter, we need to take decision, at each level, which cluster to split down and how to perform the division.

II. ARCHITECTURE PROCESS

This is the flow of system

A. Collection of Data:

We have allowed few types of files to be cluster by our system. We cannot cluster media files by our system but we are trying to do neglect media files if those give as input by mistake. We can recognize media files from extension and put them in separate space.

We can cluster pdf, txt, doc, ppt and only textual files.

B. Pre-Processing

We have removed stop words, i.e., common words such as “a”, “are”, “do”, “and”, “for”, “to”, “but”, “this”,”that”, “The”, “Be”, Etc.

We have used stemming to gain original word Thus; all the words sharing the same stem are considered to be the same word. For example, words “wanted”, “wanting”, and “wants” are stemmed to “want”.

C. Performing Clustering

I our clustering algorithms we have arranged documents and files or we can say we represented them using the vector-space model.

We are also using Cosine Similarity and Euclidean Distance Formulae for clustering.

D. Post-Processing

It is for searching after clustering. To check whether the result is pure or not, How much time taken by system to perform clustering, It checks the limitations of the system.

Details about Formulas:

1. In Vector space model each document, d, is considered to be a vector, d, in the term-space (set of files "words"). In its simplest meaning, each file or document is represented by the (TF) vector, dtf = (tf1, tf2, …, tfn), where TF is the term frequency of the specific term in the document.

   In this case, we use this technique of this model that weight of each keyword is dependent on its inverse document frequency (IDF) in the document collection.

   TFIDF= (No. of presence of term T in this document D) * log((total No. of documents)/(No. of documents with mention of term t)).

2. Cosine similarity:

   It is a measure of similarity between two vectors of an inside product space that calculates the cosine of the angle between them.

3. Euclidean Distance:

   There are two documents DA and DB displayed by their term vectors TA and TB respectively, the Euclidean distance is given simply as:

   DE(−⟶ ta,−⟶ tb) = (Pm t=1|wt,a −wt,b|2) 1 2

III. FUTURE SCOPE

We are working on efficiency of time complexity. We are clustering files of size near about 10-20KB but we wanted to cluster files of size in MBs.

Then this system can be use as backend application for any Data Mining Applications. We wonder if such application could use for email system. We are having different views in current email system like social, primary etc. but we can cluster emails so we can recognize which emails are important to keep and which are not useful to remove or we can archive some emails.

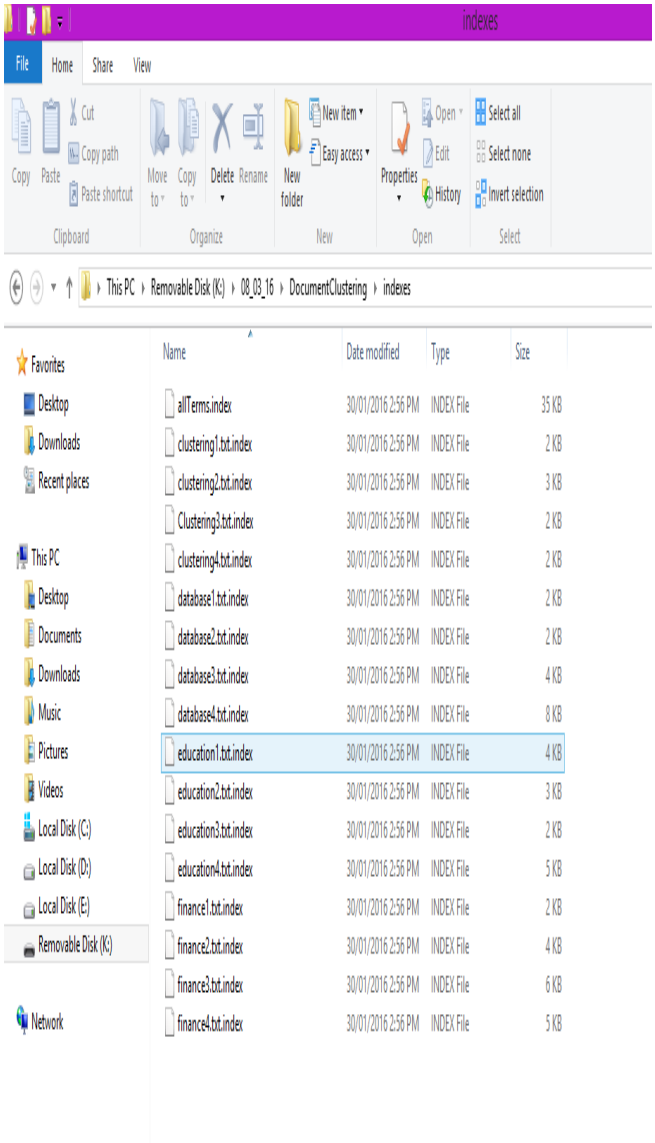
Searching or Investigation purpose accomplished by this clustering system.

A. Issues Occurred:

- Time required more for large size files.
- We cannot cluster media files.
- We are using distance based formulae so not sure about purity of clusters.

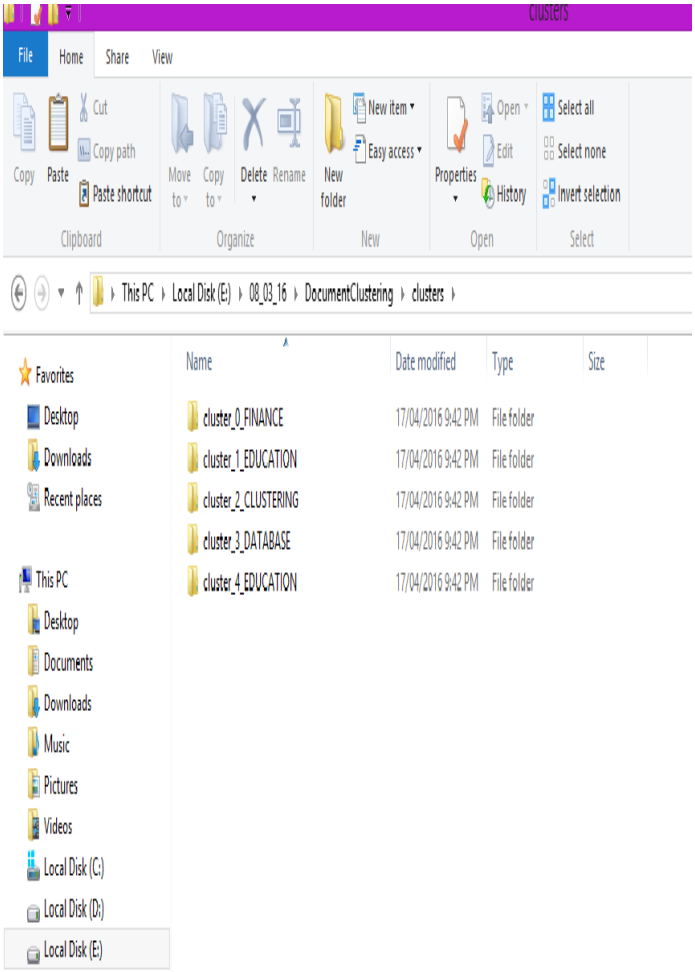
IV. RESULT

A. *Index Created After Pre-Processing*



Indexes files are created after pre processing these index files are nothing but keyword found in file and weight of files is stored in vector space model. We have used cosine similarity to count distance between two files and Euclidean distance to count distance between two clusters.

B. *Clusters Created by System*



We have done clustering technique based on distance between two files. We can see in second diagram there are 5 clusters. Zero cluster having files of Finance, first cluster is having files of Education, second cluster is having files of clustering, third cluster is having files of database, fourth cluster is having files related to Database, and last cluster is again having files of Finance but these files have different content than previous one. We have labeled cluster so we can easily search for required file.

V. CONCLUSION

Hence we implemented text clustering using distance formulas. We are still working on its issues like time complexity and purity of result. This system useful for searching and investigation of computer files. We will improve this system to make more applicable to every other big data mining systems.

REFERENCES

[1] Michael Steinbach, George Karypis, Vipin Kumar.”A Comparison of Document Clustering Techniques”, Department of Computer Science and Engineering, University of Minnesota Technical Report #00-034.

[2] Luís Filipe da, Cruz Nassif, Eduardo Raul Hruschka,” Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection”, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL.8, NO.1, JANUARY 2013

[3] Priyanka kasote, Shruti singh, Ms. Chhaya Varade, ” Document Clustering for Forensic Analysis via Cosine Similarity”, IJCSET

[4] Bhagyashree Umale, Prof. Nilav M, ”Survey on Document Clustering Approach for Forensics Analysis”, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3335-3338.

[5] Joaquim Silva, Jo˜aoMexia, Agra Coelho, Gabriel Lopes “Document Clustering and Cluster Topic Extraction in Multilingual Corpora”, FCT/Universidade Novade Lisboa 2725 Monteda Caparica, Portugal gpl@di.fct.unl.pt

Publish Research Article

Dear Sir/Mam,

We invite unpublished Research Paper,Summary of Research Project,Theses,Books and Book Review for publication.

Address:- Dr. Ashak Hussain Malik House No-221, Gangoo Pulwama - 192301 Jammu & Kashmir, India
Cell: 09086405302, 09906662570,
Ph No: 01933212815
Email:- nairjc5@gmail.com, nairjc@nairjc.com , info@nairjc.com
Website: www.nairjc.com