Documents Clustering based on Similarity Measures using Global Keyword Vector Generation

Harneet Kaur[1], Rupinder Kaur[2]
[1, 2]DIET, Kharar, Punjab, INDIA

Abstract - Clustering in documents remains always an interesting and challenging task. Earlier, the clustering algorithms were based on k-means clustering, the top most favourite algorithm, but with some limitation on no. of clusters and centre oriented. However, the present need of the clustering is multi-view point based document clustering. A K-means and cosine term based clustering is fair enough once the no. of documents is less. However, if the no. of documents increases and data as well, k-means and cosine based clustering becomes time consuming even being centred oriented. If multi view points are made basis for clustering, then the clustering problem will be never ending process. In the presented algorithm, the documents are scanned for their texts material. Each word in the text is given a unique flag identifier and based on the flags, a word histogram is created. Further, a dictionary support is obtained for the words synonyms in order to find the similar word and their synonyms in other documents for clustering purposes.

Keywords: *Clustering, Text Mining*

I. INTRODUCTION

Document clustering is an important activity while working on data extraction from the stack of documents. This may be understood by the example of extracting skill set information from a large no. of resumes for an employment service providing agency, it is a very common practice for employment service agencies to extract the skill set information form the candidate's resumes. Normally, the agencies have their own format of resume and the information extraction is easier in that case. However, when the resumes are called on from open channel, then the resumes are not in uniform format and skill set information retrieval is very tedious task in that case.

Very often, the task of information retrieval from resume is outsourced to the persons working from home based PCs as option of online earning from home. But, this again be slow process and the reliability is poor if a new person is assigned the job. For, it is very common need of a document clustering and information extraction from the set of documents using an efficient algorithm. The documents can be clustered based on similarity and dissimilarity. One of the most similarity measure approach is to find the Euclidean distance between the document vectors. The document vector may include the similar words frequency, similar sentence frequency and citations.

II. RELATED WORKS

A novel multi view analysis is proposed for document clustering. A fair performance is achieved using the modified k-means clustering along with some adaptive features that are document centred oriented. Further, the documents are clustered to a good extent when a cosine term based approach is combined along with k-means clustering. K-means clustering has been remained the favourite choice among all the existing clustering algorithms [1]. The documents are clustered based on similarity as well as dissimilarity approach. The results are validated based on empirical study and theoretical analysis. A comparison of k-means, cosine terms based and presented approach is given in details here [2].

A novel clustering approach is discussed in this paper based on similarity measures. Euclidean distance based documents vector distance is computed and based on minimum Euclidean distance between documents vectors, the documents are clustered [3]. Assessment of similarity and dissimilarity is done based on keywords and their frequency. Different documents keywords at=re placed in respective vectors. The vectors are aligned based on most similar words and then the Euclidean distance analysis is done to get the Euclidean distance and in turn similarity or dissimilarity [4].

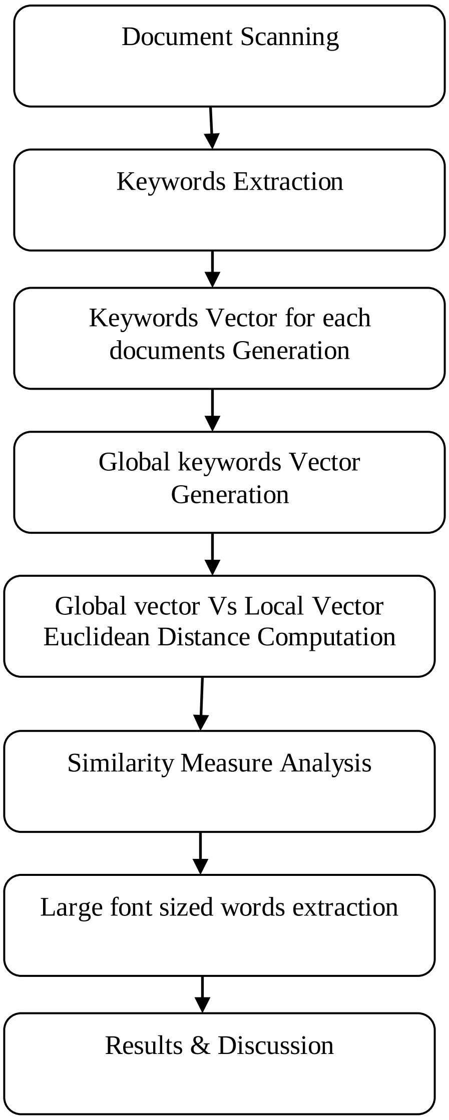
The documents under scanner are structured based on font style other than the normal font style of the document. It has been observed that the key words are normally given special effects once in a while in the document. And may be in normal font in the entire document. The keywords are extracted based on the font style and stored in a array of keywords. A single array of keywords is generated for all the documents and compiled for its uniqueness. Now the entire array is compared with the set of documents under scanner. Based on some criteria, the documents are flagged to the cluster [5]. A hierarchical divisive based clustering approach is presented in this paper. The documents keywords are arranged based on their frequency in the documents. This approach has proved to be an accurate approach while document clustering [6].

More than one document are made as reference for k-means clustering. This approach has been shown to be an accurate one while clustering the no. of documents [7].

III. ALGORITHM

Data mining is that the method of extracting or mining information from great deal of information .It's Associate in analytic method designed to explore giant amounts of information in search of consistent patterns and systematic relationships between variables and to validate the findings by the detected patterns to new subsets of information. It is often viewed as a result of natural evolution in development of Functionalities like data assortment, information creation, information management, information analysis. It is the process where intelligent methods are applied in order to extract data patterns from databases, data warehouses, or other information repositories Clusters are often thought of the foremost necessary unsupervised learning problem, thus as Each different drawback of this sort, it deals with finding a structure in an exceedingly assortment of unlabelled information.

A cluster is so a set of objects that are coherent internally, however clearly dissimilar to the objects to different clusters. A loose definition of cluster could also be “the methodology of organizing objects into groups whose members’ are similar in some way”. Documents to be clustered are arranged in a folder and scanned for the text contained in Process. By processing the documents, we can get the child files which are linked to document file. Histogram displays the no of documents by showing the similarity range between 0 to 1.Clusters formed by considering similarity of the documents. Similarity is calculated between the keyword tags between two files. Result is displayed as a bar chart which axis has similarity between file to file.


Document Scanning
Keywords Extraction
Keywords Vector for each documents Generation
Global keywords Vector Generation
Global vector Vs Local Vector Euclidean Distance Computation
Similarity Measure Analysis
Large font sized words extraction
Results & Discussion

The documents are scanned in MATLAB using the text read function. The text retrieved from the text document is made a unique word set using the unique command in matlab. This way, each word now appears once in the document. This helps in tagging each word with identifier. Now the frequency of each word is computed form the original scanned document thereby giving the word histogram. Similarly all the documents under clustering process are scanned and word histogram is computed. Further, the words having the most frequency in the histogram are clustered in a global keyword vector as well against each document keyword vector.

The documents keyword vector are compared with the global keyword vector and based on Euclidean distance of each document keyword vector to that of the global keyword vector, the documents are arranged in either in ascending or descending order. The lowest Euclidean distance document keyword vectors are the clusters of same nature.

IV. Conclusion

The presented work is basically a review work and an approach is presented that will be taken up in the project to develop the algorithm. The algorithm is initially targeted to work on text documents only. Later on, the algorithm may be optimized for table and image contents as well. Further, the image text contents may also be targeted as there is great demand for clustering of images based on scene text detection.

V. References

[1] Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan, “Clustering with Multiviewpoint-Based Similarity Measure”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 6, JUNE 2012

[2] S. Sesha Sai Priya, 2K. Rajini Kumari, “The Clustering with Multi-Viewpoint based Similarity Measure”, IJCST Vol. 3, Iss ue 1, Spl. 5, Jan. - March 2012

[3] Gaddam Saidi Reddy1, Dr. R. V. Krishnaiah, “Clustering Algorithm with a Novel Similarity Measure”, ISSN: 2278-0661 Volume 4, Issue 6 (Sep-Oct. 2012), PP 37-42

[4] 1S. Chandrasekhar, 2K. Sasidhar, 3M.Vajralu, “Study and Analysis of Multi - viewpoint clustering with similarity measures”, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 10, October 2012

[5] Aggadi Gnanesh1, M.Sudhir Kuma, “An Advance towards Standard Utilities for Document Clustering”, International Journal of Computer and Electronics Research [Volume 2, Issue 4, August 2013]

[6] B.Amuthajanaki1, K.Jayalakshmi2, “A HIERARCHICAL DIVISIVE CLUSTERING BASED MULTI-VIEW POINT SIMILARITY MEASURE FOR DOCUMENT CLUSTERING”, Volume 2, No.8, August 2013, International Journal of Advances in Computer Science and Technology

[7] Annavazula Mrinalini, A.Rama Mohan, “Implementation of Multi View point method for similarity Measure in clustering the documents”, Volume 2, Issue 1, January 2014 International Journal of Advance Research in Computer Science and Management Studies

She is pursuing her M.Tech. in CSE from DIET, Kharar, Punjab India. Her field of interest is in software engineering and efforts estimation.