Research Article

TEXTUAL DOCUMENTS CLUSTERING BASED ON GLOBAL KEYWORD VECTOR GENERATION AND EUCLIDEAN DISTANCE

Harneet Kaur and *Rupinder Kaur
DIET, Kharar, Punjab, INDIA
**Author for Correspondence*

ABSTRACT

Documents clustering are an important task in digital data base of textual documents. Manual clustering of text documents is very tedious and time consuming and labour intensive as well. The keywords vector set determine the vicinity of documents among each other in a given data base. In the presented algorithm, the keywords are extracted from the documents based on the font style, frequency of words and their synonyms. The same step is iterated for each document in the data base and a collective matrix vector of keywords is generated. Now the different columns of the matrix represents individual documents keywords, The Euclidean distance among each column and based on minimum or threshold Euclidean distance, the documents are clustered into different clusters. This gives an adaptive approach in computing the no. of clusters as the no. of clusters is not made as input the system algorithm,

Keywords: *Clustering, Text Mining*

INTRODUCTION

All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. The major difference between a traditional dissimilarity/similarity measure and ours is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects, assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Clustering is a process of grouping a set of physical or abstract objects into classes of similar objects and is a most interesting concept of data mining in which it is defined as a collection of data objects that are similar to one another. Purpose of Clustering is to catch fundamental structures in data and classify them into meaningful subgroup for additional analysis. Many of the clustering algorithms have been published every year and can be proposed for different research fields. They were developed by using various techniques and approaches. But according to the recent study k-means has been one of the top most data mining algorithms presently. For many of the practitioners k-means is the favourite algorithm in their related fields to use. Even though it is a top most algorithm, it has a few basic drawbacks when clusters are of differing sizes, densities and non-globular shape. Irrespective of the drawbacks is simplicity, understandability, and scalability is the main reasons that made the algorithm popular.

Related Works

K-means clustering has been remained the favourite choice among all the existing clustering algorithms (Duc *et al.*, 2012). 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 (Sesha and Rajini, 2012). Euclidean distance based documents vector distance is computed and based on minimum Euclidean distance between documents vectors, the documents are clustered (Gaddam and Krishnaiah, 2012). 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 (Chandrasekhar *et al.*, 2012). The entire array is compared with the set of documents under scanner. Based on some criteria, the documents are flagged to the cluster (Aggadi and Sudhir, 2013). The documents keywords are arranged based on their frequency in the documents. This approach has proved to be an accurate approach while document clustering (Amuthajanaki and Jayalakshmi, 2013). More than one document is made as reference for k-means clustering. This approach has been shown to be an accurate one while clustering the no. of documents (Annavazula and Rama, 2014).

Algorithm

Keywords can be considered as condensed version of documents and short forms of their summaries. Keyword extraction is a significant technique for number of text mining related tasks such as document retrieval, webpage retrieval, document clustering and summarization. The main aim of keyword extraction is to extract the keywords with respect to their relevance in the text.

The proposed work is based on extraction of key words based on documents word's data base arranged in an array. A histogram of each word is extracted by incrementing the histogram count against each word array index.

The words histogram method basically removes the unnecessary words from the search like prepositions, articles verbs etc. These are the words that have the maximum count and need to be removed from the search space. Now the document is remained with some words that are specific to the document matter. The document may further be filtered out by the title of the document and the words appearing the title are given prior importance.

Document Scanning

The input document is scanned using the matlab command textread. The textread command scans the document under study for its textual content including punctuation marks. The white space is eliminated while reading the document. Thereby, just having the textual matter in the document under study.

Words Array Generation

The text read command captures the textual content in an array of word. The array formed is row matrix of size (Nx1), where N is the total number of words in the given document.

W = {'This', 'is', 'an', 'example', 'of'', 'keyword', ... },

Elimination of Normal Words

Normal words are a part of natural language that does not have so much meaning in a retrieval system. The reason that normal-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing normal words reduces the dimensionality of term space. The most common words are in text documents are prepositions, articles, and pro-nouns etc that does not provide the meaning of the documents. These words are treated as normal/stop words.

The document under study contains many words that never can be the keywords. 'am' 'are' etc. These are the language supporting words like 'this', 'that', 'is', 'am' and 'are' etc. A file has been created containing these words and can be appended any time if a word seems to be under the category of normal words. Therefore, this file keeps on enriching with time. The word array is compared with this normal word file and the normal word are eliminated from the main word file. The resultant file is the normal file that contains now the candidate keywords in the document.

E = {'This', 'That', 'is', 'am', 'are', ……}

The candidate keywords array C = W – E

Unique Word Array Formation

The candidate C array contains the words irrespective of their frequency. i. e. a word may appear no. of times in the candidate array. Therefore, a unique array is generated so that the unique flags may be attributed to each word. This will help in making the histogram of the unique words in the document.

U = Candidate Keywords appearing only once

Words Histogram Extraction

Now, the unique key words are scanned over the entire original document for computing their frequency. A histogram is plotted between word identifier and their frequency. This gives the more close idea about the possible keywords in the document.

Max. Freq. Words Extraction

A threshold is selected for extraction of words that have frequencies equal to or more than the threshold are stored in the semi-final keyword array.

Title Words Histogram Extraction

Title words are now scanned over the entire document for their frequency and stored in the semi-final array of possible keywords.

Bold, Italic and Large Font Words Extraction

Similarly bold, italic ad underlined words are extracted and stored in the semi-final keywords array. The special effect words are given more weightage in the keywords weights table so as to given more preference while ascertaining weights for back propagation neural network.

No. of Sentence Containing the Candidate Keywords

No. of sentences are obtained that contain the possible candidate keywords. This gives a valid justification for the candidate word to be a keyword for the document under study.

RESULTS AND DISCUSSION

Results

Different domain documents are used for test purposes as follows:
Domain-1: News Clips on Swachh Bharat Abhiyan
Domain-2: News Clips on NAMO US Visit
Domain-3: Text Mining Abstracts
Each document size was restricted to 500 words for testing purpose. Ten documents from each domain were collected in a data base and the discussed algorithm was tested taking on each of them one by one in a loop condition. The results obtained are compiled in following table:

Domain	No. of Key Words	Common Keywords Vector Set
Domain-1	21	
Domain-2	23	63
Domain-3	19	

Now three test documents from each set are made as input to the algorithm and clustered. The results are summarised as follows:

Domain	Keywords	Cluster Assigned
Domain-1	15	Domain-1
Domain-2	20	Domain-2
Domain-3	17	Domain-3

Conclusion

The documents clustered with the discussed approach have been found to be very accurate and are very similar as clustered manually i.e. ideally for cross check purposes. Further, a text summary may also be generated based on the extracted keywords and then the documents may be clustered in second tier i,e. Based on keywords as that extracted from the computed text summary. That gives very refined documents clusters. The length of the keywords or keyword vector set may be controlled by analysing the words histogram. However, keeping very close frequency threshold may result in less no. of documents and the clustering may be unfruitful. However, relaxing so much the word frequency may result in false clustering. Therefore, there is trade off between the clustering degree and the keywords vector set.

REFERENCES

Aggadi Gnanesh and Sudhir Kumar M (2013). An Advance towards Standard Utilities for Document Clustering, *International Journal of Computer and Electronics Research* **2**(4) 447-449.

Amuthajanaki B and Jayalakshmi K (2013). A hierarchical divisive clustering based multi-view point similarity measure for document clustering. *International Journal of Advances in Computer Science and Technology* **2** 155-159.

Annavazula Mrinalini and Rama Mohan A (2014). Implementation of Multi View point method for similarity Measure in clustering the documents. *International Journal of Advance Research in Computer Science and Management Studies* **2**(1) 200-205.

Chandrasekhar S, Sasidhar K and Vajralu M (2012). Study and Analysis of Multi - viewpoint clustering with similarity measures, *International Journal of Emerging Technology and Advanced Engineering*, Available: www.ijetae.com, ISSN 2250-2459 **2** 606-609.

Duc Thang Nguyen, Lihui Chen, Senior Member IEEE and Chee Keong Chan (2012). Clustering with Multiviewpoint-Based Similarity Measure, *IEEE Transactions on Knowledge and Data Engineering* **24**(988-1001).

Gaddam Saidi Reddy and Krishnaiah RV (2012). Clustering Algorithm with a Novel Similarity Measure, ISSN 2278-0661 **4**(6) 37-42.

Sesha Sai Priya S and Rajini Kumari K (2012). The Clustering with Multi-Viewpoint based Similarity Measure, *IJCST* **3**(1Spl.) 880-882.