www.ijarcsse.com

Volume 5, Issue 5, May 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Effective and Efficient Algorithm for Document Clustering

Latika
Department of Computer Science and Applications
Kurukshetra University, Kurukshetra,
Haryana, India

Abstract- This paper proposes an effective and efficient algorithm for clustering text documents. This algorithm is formulated by using the concept of well known k-means algorithm. The standard k-means algorithm suffers from the problem of random initialization of initial cluster centers. The proposed algorithm eliminates this problem by introducing a new approach for selection of initial cluster centroids. Several experiments are conducted on mini_newsgroups dataset to measure the performance of proposed algorithm and the results obtained are very promising when compared to two other algorithms: k-means and enhanced k-means.

Keywords- Document clustering; Initial centroids; F-measure; K-means; Vocabulary; VSM;

I. INTRODUCTION

With the advancement of internet technology and less expensive hardware, there is a tremendous growth in the volume of electronic documents available on the web. For proper organization and summarization of these documents, various techniques have been approached. Document Clustering is one of those techniques. Document clustering categorizes documents into various groups called clusters such that documents in one cluster are similar to each other and documents in different clusters are dissimilar. By organizing similar documents together, a large collection of documents can be easily browsed, organized and summarized [5]. Document clustering also plays important role in various fields like knowledge discovery, business applications, and biological taxonomy etc.

Many algorithms are available in the literature for performing data clustering [1].Out of these, two major categories of algorithms are commonly used for document clustering are: partitioning algorithms and hierarchical algorithms. Partitioning clustering algorithm divides the documents into fixed partitions, where each partition represents a cluster. Hierarchical clustering creates a hierarchical decomposition of the given documents. Partitioning algorithms generally perform better than hierarchical algorithms [3][4]. The most widely used partitioning algorithm is k-means. Although k-means algorithm is simple, but it is quite sensitive to the selection of initial cluster centers (centroids) [7] i.e. the results of clustering greatly depend on the position of initial cluster centers which are randomly chosen in standard k-means algorithm. Several methods have been provided to improve the results of standard k-means algorithms. This paper presents an improved algorithm which is formulated by using the concept of k-means algorithm. This algorithm introduces a new mechanism for selection of initial cluster centers.

This paper is organized as: Section II presents a brief survey of various techniques used for document clustering so far. Section III provides the procedure and algorithms used for document clustering. Experimental setup and Experimental results are shown in Section IV and Section V. Finally, the conclusion of this paper and future work is given in Section VI.

II. LITERATURE REVIEW

Document clustering has been widely studied in computer science literature due to its importance in various fields. Various approaches have been used for categorizing similar documents together. A brief overview of these approaches is given below.

Ying Zhao and George Karypis [3] performed comparison of six partitional and nine hierarchical agglomerative algorithms. The experimental results show that partitional algorithms produce better hierarchical clustering solutions than agglomerative methods. Michael Steinbach et al. [4] performed comparison of k-means, bisecting k-means and UPGMA. Experimental results show that bisecting k-means is better than k-means and UPGMA and UPGMA is best among three agglomerative techniques. Mushfeq-Us-Saleheen Shameem and Raihana Ferdous [6] proposed an efficient k-means algorithm for document clustering. In the proposed method Jaccard distance measure is used with k-means algorithm for finding most dissimilar k documents for centroids of k clusters. Ramiz M. Aliguliyev [5] shows that assignment of weights to documents improves clustering results. GA based approach has been used for document clustering. This paper has presented twelve criterion functions, six of these are weighted and six are unweighted. Proposed criterion functions have been compared with k-means algorithm and the results of weighted criterion functions are better than k-means and unweighted criterion functions. This paper also proposed an adjusted cosine similarity measure that outperforms than cosine similarity measure. O.H ODUKOYA et al. [8] proposed an improved algorithm for document clustering. The proposed algorithm is a variation of k-means algorithm that introduces a new initialization method for selection of initial centers for k-means algorithm.

M. Arshad [9] proposed a new clustering algorithm Kea bisecting K-means which is based on KEA and Bisecting k-means algorithm. The proposed method uses KEA (Automatic Key-phrase Extraction) for extracting several key phrases from documents and Bisecting k-means algorithm for performing clustering on these extracted key phrases. Fathi H. Saad et al. [10] performed comparison of various hierarchical agglomerative algorithms. The comparison of algorithms has included six criterion functions that can be divided into four groups (internal, external, graph based and hybrid) and three selection schemes (single-link, complete link and UPGMA). Cosine similarity measure has been used for measuring similarity. Leena. H. Patil and Mohammed Atique [11] presented a novel approach for feature selection method tf-idf. Three different term weighting schemes tf-idf, tf-df, and tf 2 have been used and the terms having weight higher than a pre-specified threshold are selected as key features. Tf-idf feature selection scheme is found to be better among all. Sunghae Jun et al. [12] presented a method to overcome the problem of sparseness in document clustering. To remove sparsity in document term matrix SVD-PCA is used. K-means clustering based on support vector clustering and Silhouette measure is used for performing clustering operation. Yinglong Ma et al. [13] presented a three phase approach to document clustering. In first phase the best topic model is determined, in second phase initial clustering centers are obtained using k-means++ and in last phase k-means is applied for document clustering.

III. METHODOLOGY

Document clustering uses clustering algorithms to make clusters of similar documents. But clustering algorithms cannot be directly applied to datasets containing documents. We need to follow several steps before applying any clustering method on text documents. In this work, these steps are implemented using various modules like extractor, document reader, preprocessor and VSM creator. The whole document clustering process including these modules is shown in figure 1 and a brief description is also given.

Extractor: The input to this module is a dataset folder that contains various topic folders and these folders further contain documents. Extractor module reads dataset folder and returns various topics in it and also, returns documents in each topic. For e.g. well known 20_newsgroups dataset folder contains 20 categories and each category contains 1000 documents. Extractor module will read this folder and will return 20 topics and documents in these topics.

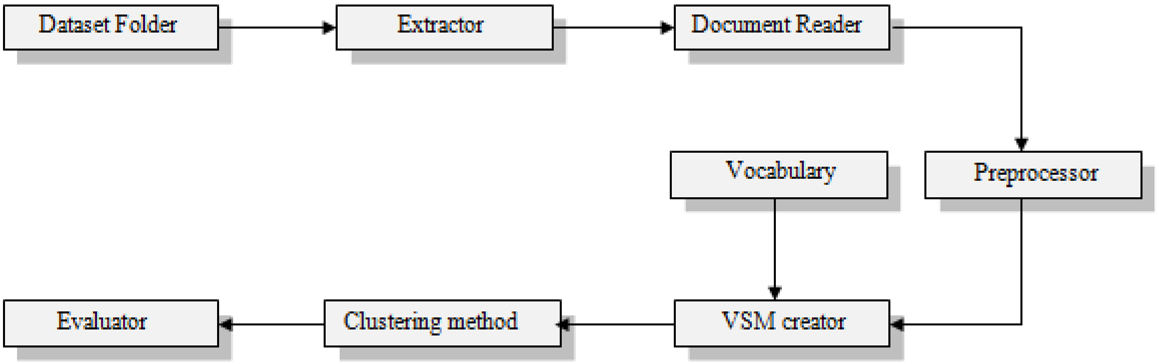

Fig 1.Various Modules used in Document clustering

Document Reader: Each document consists of several lines and each line further consists of several words. Document reader module reads each document line by line and returns various words in these lines. These words after extraction are called tokens. These tokens may contain some words which are insignificant from clustering purpose so, we apply pre-processing on these tokens.

Preprocessor: This module takes various tokens returned by document reader as input and produces a set of new tokens as output which is suitable for clustering. Pre-processing of tokens is done using following steps:

- All the special symbols, punctuations and digits are removed. Upper case letters are converted to lower case letters.
- All the stop words are removed. Stop words are those frequently words which are insignificant because they do not carry any meaning from clustering point of view. For e.g. the, by, for etc. For stop words removal, a list of stop words is created and then each token is compared with this list. If token is a stop word then it is removed else it is kept for further processing.
- Word stemming is performed. In word stemming, all the words are reduced to their stem word. For e.g. all words like ‘looks’, ‘looked’ and ‘looking’ are reduced to their stem word ‘look’. For performing stemming of words Porter’s algorithm is used.

Vocabulary: After pre-processing, the unique words from the pre-processed words are extracted. But the main problem is that even a moderate size collection consists of thousands of unique words. All of these words do not have discriminating power. So, the main idea is to remove all the words that have very low frequency and are irrelevant. Vocabulary contains frequent and relevant words of each category. Each document has only a subset of these words.

VSM creator: To apply clustering algorithm, documents needs to be represented in some mathematical form. VSM creator module represents documents in vector space model. In vector space model n documents with m words are represented as n×m document by word matrix. Under VSM, each document has different weights for different words as shown in figure 2.

In this work, Term frequency (TF) is used for finding out weights of each word in each document. Term frequency is how many times a term occurs in a document. It is a measure of significance of a term in a document. Assumption is that more frequently used words are significant. Under TF weights are represented as:

$W_{ij} = f_{ij}$

where f_{ij} : Frequency of jth word in ith document.

	Word 1	Word 2	. . .	Word m
Document 1	$W_{1,1}$	$W_{1,2}$		$W_{1,m}$
Document 2	$W_{2,1}$	$W_{2,2}$		$W_{2,m}$
.				
Document n	$W_{n,1}$	$W_{n,2}$		$W_{n,m}$

Fig 2.Document by word matrix

Clustering Method: Clustering method is the clustering algorithm which organizes similar documents into clusters. Document clustering algorithms take VSM as input and provide clusters of documents as output. In this work, three clustering algorithms are used for performing clustering of documents.

Algorithm 1 (Basic k-means): Most popular and widely used algorithm for document clustering is k-means algorithm. Various steps involved in k-means algorithm are shown in figure 3. Here, n data vectors are n document vectors of VSM.

Input:

D: a set of n data vectors

k: number of clusters

Output:

A set containing k clusters

Steps:

- Arbitrarily choose k data vectors from n for determining initial cluster centers and assign them to c_j $(1 \leq j \leq k)$
- Calculate the distance of each data vector in D to each cluster center in C.
- Assign each data vector to the cluster to which the data vector is most similar i.e. having minimum distance.
- Update the cluster centers by calculating the mean of data vectors assigned to a cluster.
- Repeat above three steps until stopping criteria is met.

Stopping criteria:

Number of iterations or change in the position of cluster centers in consecutive iterations.

Fig 3.Various steps involved in Algorithm 1

Although, k-means is simple and basic algorithm for cluster analysis but there are some problems [7]. The main problem with k-means is its random initialization method. Random initialization of initial cluster centers may cause solution to converge to local optimal. It produces different clusters for different set of values of initial centroids. The quality of results depend on the selection of initial cluster centers i.e. poor selection of initial centroids will result in poor clustering results and better selection will provide improved results. Due to this reason the random initialization method of k-means is replaced in Algorithm 2 and Algorithm 3.

Algorithm 2 (Enhanced k-means): Algorithm 2 is an enhancement of standard k-means algorithm. This enhanced method is based on the method proposed by [2]. In this algorithm, the dataset is divided into k subsets called cluster and then cluster centers are initialized by the average of each cluster. Various steps involved in algorithm 2 are shown in figure 4.

Input:

D: a set of n data vectors

k: number of clusters

Output:

A set containing k clusters

Steps:

- Determine the size S of cluster by using round (n/k).
- Create k arrays C_i ($1 \le i \le k$)
- Transfer S data vectors from input set D to each C_i until all data vectors are removed from D.
- Compute the average Av_i of each cluster C_i ($1 \le i \le k$) to determine initial cluster centers.
- Find out distance Dt of each data vector in cluster C_i from its own cluster center Av_i ($1 \le i \le k$).
- Find out the distance Ds of each data vector in V from each cluster center Av_i ($1 \le i \le k$).
- If $Dt \le \min(Ds)$ then data vector stay in same cluster else data vector will move to that cluster to which it is most similar or having less distance.
- Recalculate the cluster centers & move data vectors until no change.

Fig 4.Various steps involved in Algorithm 2

Algorithm 3 (Proposed): Algorithm 3 introduces a new approach for determining initial cluster centers. In this approach Vocabulary V is divided into k parts where each part contains the features or words for each category. After that for each cluster center C_i, weights are assigned to ith part of vocabulary ($1 \le i \le k$). Different steps involved in algorithm 3 are shown in figure 5.

Input:

D: a set of n data vectors

k: number of clusters

V: Vocabulary containing m words

Output:

A set containing k clusters

Steps:

- Organize the m words of vocabulary into k parts.
- For each C_i, assign weights to the ith part of vocabulary ($1 \le i \le k$).
- Consider these C_i ($1 \le i \le k$) as initial cluster centers.
- Calculate the distance from each data vector in D to each cluster center in C_i ($1 \le i \le k$) .
- Assign each data vector to the center to which it is most similar or having minimum distance.

Fig 5.Various steps involved in Algorithm 3

Evaluator: This module calculates the result of three different clustering algorithms in terms of F-measure, Time complexity and Number of iterations.

F-measure: F-measure is harmonic mean of precision and recall. Precision can be defined as ratio of number of relevant documents to the number of total documents retrieved. Recall can be defined as the ratio of number of relevant documents retrieved to the total number of relevant documents. Consider a class i and cluster j, then

$$\text{Prec } (i,j) \;=\frac{n\ (i,j)}{\text{nj}}$$

$$\text{Recall } (i,j) \;=\frac{n\ (i,j)}{\text{ni}}$$

Where n (i, j) represents number of documents of class i present in cluster j. ni represents number of documents in class i. nj represents number of documents in cluster j. Then, F-measure of class i with respect to cluster j is given as:

$$\text{F } (i,j) \;= \frac{2\ *\ \text{prec(i, j)}\ *\ \text{recall(i, j)}}{(\text{prec(i, j)}\ +\ \text{recall(i, j)})}$$

The average F-measure of is given as:

$$\text{F} = \frac{1}{\text{k}} \sum_{i=1}^{\text{k}} \text{F } (i,j)\ \mid 1 \leq \text{j} \leq \text{k}\}$$

IV. EXPERIMENTAL SETUP

All the experiments are performed using MATLAB 7.8.0 on a laptop computer with following key configuration: window vista, Intel (R) Pentium (R) Dual CPU @ 2.16 GHz, and 2 GB RAM. All the three clustering algorithms are evaluated by using mini_newsgroups dataset [15]. Mini_newsgroups is a subset of popular dataset 20_newsgroups. Mini_newsgroups contains 20 newsgropus categories where each category contains 100 documents. This experiment is performed on the three categories of this dataset: alt.atheism, comp.graphics, and comp.os.ms-windows.misc.

A list of 456 stop words is used for stop words removal. Porter's algorithm that is used for performing stemming of words is taken from [14]. For clustering algorithm 3 the value of weight that is assigned to each part of vocabulary is taken as 1.

V. EXPERIMENTAL RESULTS

All the three clustering algorithms for document clustering are implemented by using above set up. Performance of all the three algorithms is observed and compared. The evaluation measures that are used for evaluation of clustering algorithms are: Execution time of algorithm, Number of iterations, F-measure, and Time complexity. The experiment is conducted several times for three categories of mini_newsgroups dataset. In each experiment, the results are computed and then the average value of these results is presented as final results. Table 1 shows the final results obtained for three clustering algorithms.

Table I. Final results obtained for three solutions

	Exec. Time	No. of iteration	F-measure	Time comp.
Algorithm 1	0.2450	11	0.1720	0.2587
Algorithm 2	0.1532	6	0.5175	0.1641
Algorithm 3	0.0255	1	0.8163	0.0308

The graphical representation of comparison among three clustering algorithms on the basis of different evaluations measures is shown in the following figures.

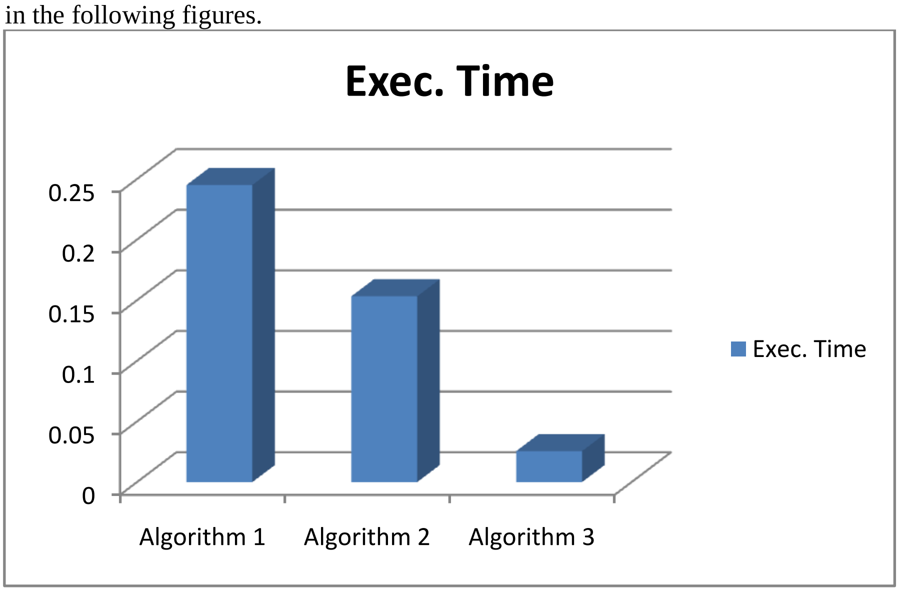

Fig 6.Execution time comparison among three clustering algorithms

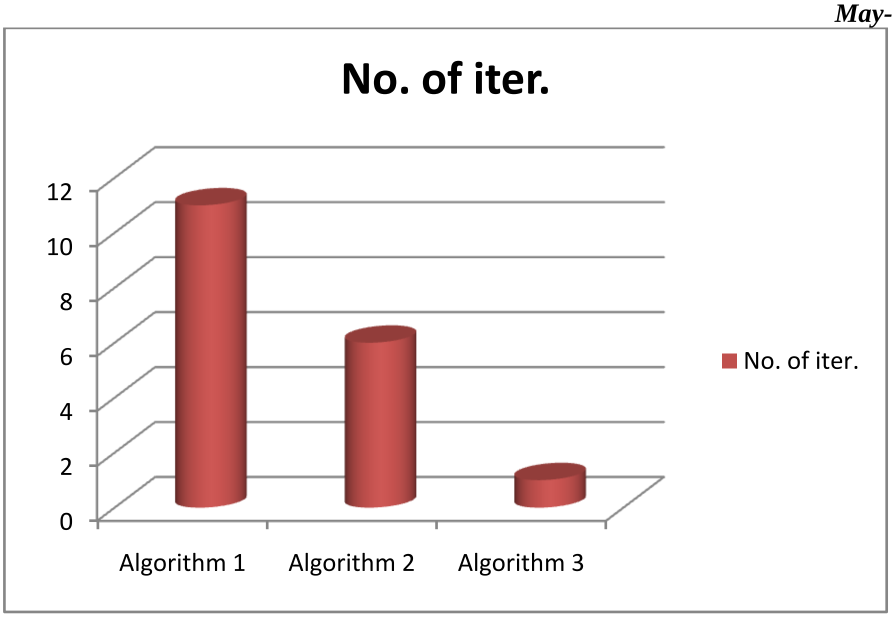

Fig 7.Number of iterations comparison among three clustering algorithms

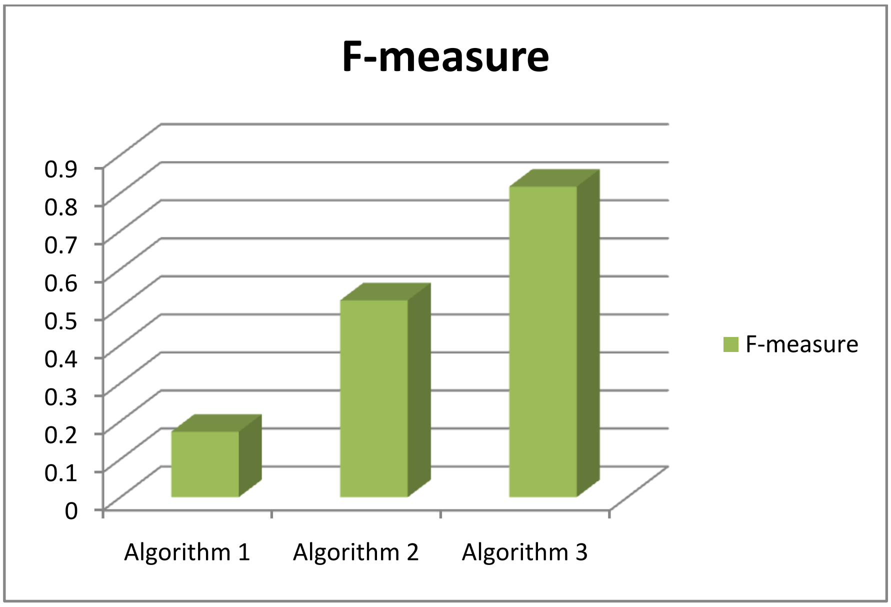

Fig 8.F-measure comparison among three clustering algorithms

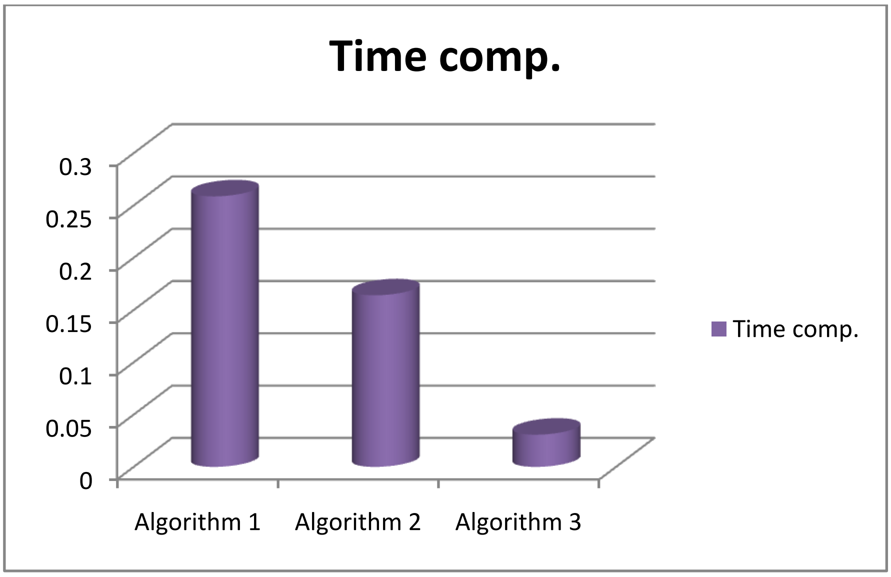

Fig 9.Time complexity comparison among three clustering algorithms

VI. CONCLUSION & FUTURE WORK

This paper proposes a new clustering algorithm for performing clustering on documents. This algorithm uses a novel approach for selection of initial cluster centers or centroids which will be further used for clustering. To measure the effectiveness and efficiency of proposed algorithm, it is compared with two other algorithms: Standard k-means and Enhanced k-means. The experimental results show that the proposed clustering algorithm outperforms than the other two algorithms in terms of execution time, f-measure, No of iterations and time complexity. In future, the experiment can be conducted for large volume of datasets. New clustering approaches can also be explored to further improve the clustering effectiveness.

REFERENCES

[1] J. Han and M. Kamber, *"Data Mining concepts and techniques"*, Morgan Kaufmann Publishers, Second edition, 2009.

[2] Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, *"Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity"*, Middle-East Journal of Scientific Research, Vol.12, No.7, pp. 959-963, 2012.

[3] Ying Zhao and George Karypis, *"Evaluation of Hierarchical Clustering Algorithms for Document Datasets"*, Technical Report, Jun. 2002.

[4] Michael Steinbach , George Karypis, and Vipin Kumar, *"A comparison of document clustering techniques"*, In KDD Workshop on Text Mining, 2002.

[5] Ramiz M. Aliguliyev, *"Clustering of document collection – A weighting approach"*, Expert Systems with Applications 36, pp. 7904-7916, 2009.

[6] Mushfeq-Us-Saleheen Shameem and Raihana Ferdous, *"An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering"*, IEEE, 2009.

[7] P. Prabhu and N. Anabazhagan, *"Improving the performance of K-Means clustering for High Dimensional Data Set"*, International Journal on Computer Science and Engineering, Vol. 3, No.6, June 2011.

[8] O.H ODUKOYA, G.A. ADEROUNMU AND E.R. ADAGUNODO, *"An Improved Data Clustering Algorithm for Mining Web Documents"*, IEEE, 2010.

[9] M. Arshad, *"Implementation of Kea-Keyphrase Extraction Algorithm By Using Bisecting k-means Clustering Technique For Large And Dynamic Data Set"*, International Journal of Advanced Technology & Engineering Research, Vol.2, Issue 2, Mar.2012.

[10] Fathi H. Saad, Omer I. E. Mohamed and Rafa E. Al-Qutaish, *"Comparison of Hierarchical Agglomerative Algorithms for Clustering Medical Documents"*, International Journal of Software Engineering & Applications, Vol.3, No.3, May 2012.

[11] Leena. H. Patil and Mohammed Atique, *"A Novel Approach for Feature Selection Method TF-IDF in Document Clustering"*, 3rd IEEE International Advance Computing Conference (IACC), 2013.

[12] Sunghae Jun, Sang-Sung Park and Dong-Sik Jang, *"Document clustering method using dimension reduction and support vector clustering to overcome sparseness"*, Expert Systems with Applications 41, pp. 3204–3212, 2014.

[13] Yinglong Ma, Yao Wang and Beihong Jin, *"A three-phase approach to document clustering based on topic significance degree"*, Expert Systems with Applications 41, pp. 8203–8210, 2014

[14] http://tartarus.org/martin/PorterStemmer/matlab.txt

[15] https://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html