I.J. Education and Management Engineering, **2018, 4, 10-18**
Published Online July 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.04.02

Available online at http://www.mecs-press.net/ijeme

Exploration of Various Clustering Algorithms for Text Mining

Neha Garg[a], R.K. Gupta[b]

[a] *Department of CS&A, ITM University, Gwalior, India*
[b] *Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior, India*

Received: 15 December 2017; Accepted: 20 March 2018; Published: 08 July 2018

Abstract

Due to the current encroachments in technology and also sharp lessening of storage cost, huge extents of documents are being put away in repositories for future references. At the same time, it is time consuming as well as costly to recover the user intrigued documents, out of these gigantic accumulations. Searching of documents can be made more efficient and effective if documents are clustered on the premise of their contents. This article uncovers a comprehensive discussion on various clustering algorithm used in text mining alongside their merits, demerits and comparisons. Further, author has likewise examined the key challenges of clustering algorithms being used for effective clustering of documents.

Index Terms: Text Mining, Document Clustering, Partitioning Algorithms, Hierarchical Algorithms.

1. Introduction

With rapid development in technology and in addition sharp lessening in the storage cost, it has turned out to be conceivable to store extensive number of text files for future references. At the same time, it is time consuming as well as costly to recover the document of interest out of these huge accumulations.

Clustering is a convenient text mining technique that subdivides the documents into desired number of clusters, so that documents in same gathering have higher similarity in contrast with documents having a place with various gathering [1,2]. The persistence of clustering the documents is to introduce an order by grouping them. At a point, when a gathering is sorted out into clusters, it is easier to recover the required documents.

There are various requirements of a good clustering in text mining are-

– Clustering methods have able to pact with various types of attributes like numerical, categorical

* Corresponding author: Neha Garg
E-mail address: nehagarg179@gmail.com

attributes etc.

- Clustering algorithms would be highly scalable to deal with enormous document databases.
- In the cluster analysis, there is a less necessity of the input parameters because the clustering results can be delicate to input parameters.
- Clustering algorithms can deal with outliers or noisy data.
- There is a need of incremental clustering algorithms to consolidate the recently embedded data in the database.

The Fig 1 illustrates the general overview of text mining process which begins with gathering text documents from different sources after that preprocessing is done for cleaning or formatting the documents using stemming and stop words removal techniques. Subsequently, preprocessed documents are epitomizing as Vector Space Model (VSM). The VSM results present the features matrix of documents (where documents are in rows and the terms weight are in columns), which gives as contribution to the clustering algorithm for gathering the set of documents into significant clusters.

The calculation of a term weight wt_{ij} of term t_j and document d_i will be as follows:

$$wt_{ij}=tf_{ij}*\log\frac{n}{df_j} \quad (1)$$

where tf_{ij} denotes the term weight, n represents the entire documents in archives and df_j defines the what number of documents have term t exists.

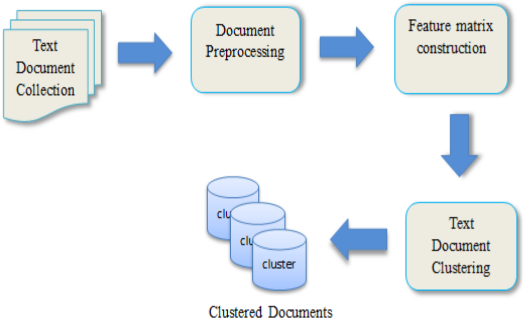

Fig.1. Text Mining Process

There are diverse gathering techniques are portrayed that composed into various classes. This paper discusses two noteworthy classifications: Partitioning Clustering and Hierarchical Clustering. The further zones in this paper uncover the past work that has been done on the clustering algorithms alongside merits, demerits and comparison of the algorithms. The author likewise discusses the key challenges of the algorithms being used in text mining for effective clustering of extensive number of documents.

2. Partitioning Method

Partitioning Clustering strategy segment the database D of N documents into a preordained number of clusters. The most generally used partitioning clustering strategies are K-means, K-medoid, CLARA and CLARANS.

2.1. K-means Method

The K-means algorithm (MacQueen, 1967) is a centroid based methodology in which the similarity of groups is measured in concern to the mean estimations of the documents in a group. The k-means algorithm begins with the k number of clusters and dataset containing N documents as inputs. The procedure of k-means algorithm can be clarified in the following steps:

- To start with select the k documents from the dataset to form k beginning centroid clusters.
- For each of residual documents, discover the cluster whose centroid is most comparative in perspective of the distance between the documents and mean of cluster.
- For every cluster, ascertain the new mean by taking the average of points in a cluster.
- The process is processed until there is no disparity in the assignment of documents to clusters.
- Finally, we give a list of clusters as output.

This final list of clusters are relies upon preliminary selection of cluster seeds. Along these lines, by taking the ideal selection of centroids, the implementation of clustering algorithm will progress. The k-means algorithm is easy to see likewise it is relatively efficient and scalable however there a few constraints in this technique which is somewhat difficult to take out. There are various methods have proposed by the authors to defeats the insufficiency of the k-means algorithm.

In 2000, Michael Steinbach [3] presents an assessment of agglomerative hierarchical clustering approach and k-means (‘standard’ k-means and ‘bisecting’ k-means) method for document clustering. The nature of groups is assessed by the three measures: entropy, overall similarity and F-measures. Last comparisons exhibits that the bisecting k-means performs superior to standard k-means and hierarchical technique.

In 2013, Vikas Kumar Sihag [4] introduced a graph based technique with figure the underlying group seeds for seeds for the k-means algorithm. In this technique, at first discover the edge with least weight and erase it from network for identify a group structure. Presently, figure the centrality of each node based on cohesiveness and dissimilarity value and the node having high centrality is taken as starting centroids for k-means. Text consequences of the F-measures shows the graph based technique gives better outcomes regarding accuracy in examination with existing method.

In 2014, Sunita Sarkar [5] have proposed a hybrid PSO + k-means algorithm and compared it with the k-means and PSO algorithm. The hybrid approach at first executes the PSO algorithm and the consequences of the PSO algorithm is then utilized as initial centroid of the k-means. The procedure of k-means is then executed until most variety of iterations is reached. Last comparisons of inter cluster similarity and intra cluster similarity proves that the hybrid algorithm performs superior to both PSO and k-means for text clustering.

Later, in 2015, S. Jaiganesh [6] characterizes a suitable comparability measures for the k-means to discover the similarity amongst the documents. The author uses the five similarity measure such as Euclidean Distance, Pearson Correlation and Jaccard Coefficient, Kullback-Leibler Divergence and Cosine Similarity for assessing the performance of these comparability measures with k-means algorithm. The quality of clusters of five different text document datasets was assessed using the purity and entropy measure. An experimental result proved that the cosine measure is the best comparability measure for the k-means calculation.

There are various requirements of good clustering which were not completely handled in the k-means algorithm. So there are some other partitioning clustering algorithms has presented.

2.2. K-medoid Method

The k-medoid method (Kaufman, L. and Rousseeuw, 1987) is a representative object based technique. PAM (Partitioning Around Medoids) was the k-medoid algorithm which has clarified in the accompanying steps:

- First pick the k documents from dataset D as delegate objects.
- Now the rest of the documents are appointed to the closest illustrative documents in view of the distance measure.

- At each phase, a swapping between a representative object and a non-representative objects is made in view of the cost function. And the process repeats until there is no modification in the clusters comes about.

The advantage of k-medoids over k-means is that it is more robust than k-means because it is less effect by outliers. As similar to k-means, PAM clustering results are also additionally relies upon the determination of beginning k set of medoids and it cannot works over gigantic amount of data.

To overwhelm the drawback of k-medoid method as it is not scalable for taking care of the gigantic amount of data, the author [7] utilized an idea of BIRCH algorithm by taking the Clustering feature (CF) in the k-medoid method. Presently as in the phase-1 of the BIRCH algorithm, sampled the data in the CF-tree at that point perform clustering of CF in leaf nodes in view of the k-medoids method. An exploratory outcome gives the better nature of clusters in contrast with k-medoids results.

The concept of combining the methods of multiple clustering algorithms to make a hybrid method, beats the downsides of the algorithms likewise gives the better clustering results. The author [8] upheld this idea in a way that it proposed an efficient density based k-medoids clustering algorithm which defeats the downsides of both DBSCAN and k-medoids algorithm. The author additionally measures the accuracy of the algorithms by utilizing the Rand measure and the outcomes appeared in the table and graphs performed well with the proposed method conversely with other two.

In 2015, Monica Jha [9] utilized the idea of k-medoids algorithm for clustering of documents which is additionally utilized for summarization. The author gives an idea of clustering the documents by considers the process of text mining. To begin with performing the documents preprocessing, after that documents are exemplified as a Vector Space Model which gives the matrix of documents where columns signifies the documents and rows signifies the term weights of corresponding documents. After the formation of bunches, following stage is document summarization which gives a significant summary of each document.

2.3. CLARA Algorithm

Clustering LARge Application method **(**Kaufmann & Rousseeuw, 1990) deals with large dataset. Instead of taking the entire dataset, CLARA first draws a numerous samples of the dataset and then applies PAM concept on each sample to get the ideal set of medoids and the outcome gives the clustered documents.

The key benefit of this method over k-medoid is that it handles larger datasets. However, the disadvantage of this method is the proficiency of CLARA relies upon the sample size. In this way, for upgrade the quality and adaptability of the algorithm, the randomized CLARA was introduced.

2.4. CLARANS Algorithm

Clustering Large Applications based upon RANdomized Search method (Ng & Han, 1994) is similar to PAM and CLARA however it draws a sample with some randomness in every step of the search. It starts with PAM and haphazardly selects the pairs for swapping at the present state. CLARANS is more effective method than the previous medoid based method however it experiences a few weaknesses that the clustering result are very sensitive to input order and it does not deal with high dimensional data.

In 2011, P. Murugavel [10] discuss the PAM, CLARA, CLARANS clustering algorithm with k-medoid distance based method for outlier detection. The algorithm initially performs clustering using any one of three. After getting the clustering results, outliers can be handled effectively based on the ADMP and threshold value. Last outcomes show that the CLARANS algorithm improves the accuracy and time efficiency in contrast with CLARA and PAM.

The author likewise performs modification on the architecture that was considered by author in the forgoing paper. In this architecture, author introduced a new algorithm ECLARANS. The Enhanced CLARANS algorithm is a new partitioning algorithm which has used to improve the accuracy of outliers with a choice of proper arbitrary points. Final chart shows that the new proposed method detects more outliers than the existing methods [11].

3. Idea towards the Document Clustering using Partitioning Algorithm

The k-means clustering algorithm is a simple method, the author characterizes the several ways for making a technique for selecting initial centroids of clusters. The underlying centroid of clusters influences the quality of k-means clustering. In this way, heuristic approach is used to improve the quality of beginning seeds of clusters that are picked in the clustering procedure. For example pick the most likely or more similar documents for the initial k clusters representatives instead of picked in random fashion. Additionally, sampling the set of documents in view of some threshold or some other approach can improve the clustering results. Other strategy that additionally utilizes the combination of partitioning and hierarchical clustering in which the hierarchical method figures the initial set of seeds and the outcomes of this technique is gives as input to the further process.

By taking the advantage of k-medoids as it is less impact by outliers, use this concept in text mining which gives more optimize clustering results by detecting some of the documents as outlier on the basis of some parameters. Also focus on improving the scalability and precision of the k-medoid algorithm in text mining.

As in 2014, author introduced the concept of hybrid algorithm which gives the preferable outcomes over the current method similarly utilizes this idea for making a combination of different clustering method which may performs superior to basic algorithms. Likewise uses the various optimization techniques with existing algorithm to get more optimize results of text documents.

The author in 2011 proposed an ECLARANS algorithm for improving the accuracy of outliers, can further improvement on this algorithm by trying to reduce the time complexity of the algorithm with the goal that it is more effective for used as clustering of documents.

4. Hierarchical Method

A Hierarchical Clustering method makes a hierarchical decomposition of the documents in a dataset by either merging or splitting method. It makes a hierarchy of clusters based on the two approaches:

- Agglomerative (Bottom-up) approach
- Divisive (Top-down) approach

Agglomerative approach begins with all documents in a different group or cluster and the sets of clusters are converged into a bigger clusters and the procedure to be proceeded until the point when some specific end condition are come to. Divisive approach is an inverse approach of Agglomerative approach in which begins with all documents in one cluster and parts the cluster into little ones until the point when some end condition comes.

The main advantage of this method over partitioning method is that they needn’t bother with k number of clusters as an input and these are suitable for huge datasets. Regardless, there is an issue as for the assurance of merge and split points because the quality of clusters are relies upon the merge and split choices. The various hierarchical algorithms are examined in following section to conquer these challenges.

4.1. BIRCH Algorithm

Balanced Iterative Reducing and Clustering Using Hierarchies (Zhang, Ramakrishnan and Livny, 1996) is referred as agglomerative clustering algorithm. It was intended to cluster the large volume of data by proposing an idea of clustering feature (CF) and CF tree used to abstract the cluster representations. A clustering feature tree is called a height-balanced tree which stores the clustering features and it has two parameters: branching factor and the threshold. Initially, BIRCH scans the database for creating a initial CF tree. After that applies the appropriate clustering algorithm to cluster the leaf nodes of CF tree.

The different inadequacies of this algorithm are that it is not performing well if clusters are not spherical in shape and the clustering results are heavily depending on threshold value. The problem with single threshold is that if threshold value is very less, at that point huge size clusters are split into little size clusters and if threshold value is increase, then clusters are converged into bigger ones and furthermore absorb noisy points [12].

In 2014, Nidal Ismael [13] proposed a concept of using numerous thresholds instead of single threshold which overcomes the drawback of BIRCH algorithm. The author presents an enhanced CF tree with the expansion of threshold parameter for each CF entry independently. In the proposed method the process of scan the data points is same as BIRCH. However the difference is that if the CF entry not absorb the data point then update the threshold by modifying factor. The author performs the experiments on the real and artificial dataset and the outcomes demonstrates the size of the CF tree will be less with improved multi threshold BIRCH as contrast with basic BIRCH which increases the efficiency of the BIRCH algorithm.

In 2015, Mamta Gupta [14] present an altered BIRCH algorithm in which initially take the mean of all points of document. In this method, author uses a jaccard measure to figure the distance between mean and all points and this process repeats for all the points after that apply the BIRCH algorithm. The authors performs comparisons of k-means and modified BIRCH on different documents size, shows that proposed method is more efficient than the existing approaches for clustering of documents.

4.2. CURE Algorithm

Clustering Using Representatives (S. Guha, R. Rastogi & K. Shim, 1998) is an agglomerative hierarchical clustering algorithm for larger databases that endeavours to utilize a balance between two methodologies i.e. all-points and centroid. CURE maintains an arrangement of well-scattered points as representative points of each sub cluster instead of one point- centroid [15]. CURE is a sampling based method that can recognize both spherical and non-spherical shaped clusters. It handles larger size databases efficiently and the algorithm is robust for handling the outliers.

4.3. ROCK Algorithm

Robust Clustering with Links (S. Guha, R. Rastogi & K. Shim, 1998) is an agglomerative hierarchical algorithm for categorical attributes. ROCK takes the idea of links for defining similarity of documents. The no. of common links between documents defined as the no. of common neighbors between the documents. The algorithm begins with each tuples is in own cluster and two nearest clusters are converged until the point when the required clusters are gotten.The benefits of ROCK over other clustering algorithm is that, it is outstanding amongst other suited technique for grouping categorical data and the clustering results comes out with the more significant groups and outliers can be taken care adequately with disposing of those points having few or no neighbors. By taking these points, the authors perform changes over the algorithm for better clustering quality, high accuracy and less time taken when perform clustering.

In 2006, Shaoxu Song [16] proposed an improved ROCK (IROCK) algorithm which uses link weight overlaps instead of link count to measure correlation of text documents. The author performs examination of improved ROCK with ROCK and k-means on two different dataset with different number of clusters and the outcomes shows that the execution of ROCK and k-Means are very compatible, however the IROCK performs superior to the next two methods.

In 2010, Rizwan Ahmad [18] proposed an algorithm named upgraded ROCK (EROCK) for text clustering in which they use the cosine measure instead of jaccard coefficient and try to improve the storage efficiency by using the concept of adjacency list instead of sparse matrix for keeping up the link information between the neighboring clusters. The proposed approach (EROCK) makes use of entire dataset for clustering instead of draws random sample from the dataset as in ROCK algorithm. Last examination of the two algorithms that runs on different document sizes shows that EROCK performs better with taken less time by enhancing the similarity measure and storage technique.

In 2010, Qiongbing Zhang [17] proposed a ROCK algorithm based on genetic optimization. The author defines a similarity function which is used all through the clustering process. In the ROCK algorithm, some marginal points may not be clustered but in the final step of the GE-ROCK algorithm, those points are also being clustered by comparing each point with every cluster based on goodness measure. A comparative table of ROCK and GE-ROCK on votes datasets defines that the GE-ROCK gives more optimized outcomes.

5. Idea towards the Document Clustering using Hierarchical Algorithm

The Hierarchical clustering algorithm is suitable for handling the numerical as well as categorical dataset. If the documents are clustered on the basis of BIRCH algorithm by taking the concept of multiple thresholds as defined by the author, it may improve the efficiency of the algorithm. In future we shall also improve policies of accuracy and memory of BIRCH algorithm.

The author defines some improvement on the ROCK algorithm for getting better clustering results in the same way further advancement would have been done in the algorithm for improving the running time of the ROCK algorithm. As in 2010, the author introduced the genetic ROCK method with several user defined threshold in the proposed approach. Similarly, use some other parameters in the clustering method for improving the performance of the algorithm in text mining.

6. Comparative Study of the Various Clustering Algorithms

Table 1 shows the comparison of different partitioning and hierarchical clustering algorithms.

Table 1. Comparative Table of Various Clustering Algorithms

S.No.	**Comparison Parameter**	**K-means**	**K-medoid**	**CLARA**	**CLARANS**	**BIRCH**	**CURE**	**ROCK**
1.	**Input Parameter**	k no. of clusters with a given dataset	k no. of clusters with a given dataset	Database of documents and works on sample of dataset	Dataset, maxneighbor, numlocal	Scans the database to build an initial CF tree	Random sampling is done	K no. of clusters with set of n sampled documents
2.	**Cluster Shape**	Spherical	Arbitrary	Arbitrary	Arbitrary	Spherical	Arbitrary	Arbitrary
3.	**Handling Large Dataset**	No	No	Yes	Yes	Yes	Yes	Yes
4.	**Handling Outlier**	Sensitive to outliers	Less impact by outliers	Handle outlier	Outliers can be handled	Handle outlier	Handle outlier	Handle outlier
5.	**Type of dataset**	Numerical	Numerical	Numerical	Numerical	Numerical	Numerical , Nominal	Categorical
6.	**Capability to undertake HD data**	No	No	No	No	No	Yes	No
7.	**Complexity**	O(nkt)	$O(k(n-k)^2)$	$O(ks^2+k(n-k))$	$O(n^2)$	O(n)	$O(n^2)$	$O(n^2+nm_m m_a +n^2 logn)$

7. Conclusions and Future Work

In this paper, we have studied various partitioning and hierarchical clustering algorithms for text data, each promoting a new idea. A good clustering of text data relies upon appropriate feature selection and furthermore best choice of the algorithm that suited with dataset. Algorithm based on random sampling turned out to be a proficient approach for huge dataset. It is valuable to extract some features from the database, with the objective that these features convey satisfactory data for clustering. The principle reason behind this paper is to give a concise overview on various clustering algorithms and gives an idea of proficiently usage these clustering algorithms in text mining. As a future work, implement noble and heuristic strategies as a change over existing algorithms for enhancing the execution of document clustering algorithms

References

[1] Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd Ed., 2006.

[2] Konchady, M. (2006). Text Mining Application Programming, Programming Series Charles River Media (2006).

[3] Steinbach, M., Karypis, G., Kumar, V. (2000). A Comparison of document clustering techniques. Technical report, Department of Computer Science and Engineering, University of Minnesota.

[4] Sihag, V. K., Kumar, S. (2013). Graph based Text Document Clustering by Detecting Initial Centroids for k-means, *International Journal of Computer Applications*, 62(19), Jan 2013.

[5] Sarkar, S., Roy, A., Purkayastha, B. S. (2014). A Comparative Analysis of Particle Swarm Optimization and K-means Algorithm For Text Clustering Using Nepali Wordnet, *International Journal on Natural Language Computing*, 3(3), June 2014.

[6] Jaiganesh, S., Jaganathan, P. (2015). An Appropriate Similarity Measure for K-Means Algorithm in Clustering Web Documents, *International Journal for Scientific Research & Development*, 3(2), 2015.

[7] Cao, D., Yang, B. (2010). An improved k-medoids clustering algorithm, *International Conference of* Computer and Automation Engineering, 3, pp. 132 – 135, 2010.

[8] Pratap, R., Vani, K. S., Devi, J. R., Rao, K. N. (2011). An Efficient Density based Improved K-Medoids Clustering algorithm, *International Journal of Advanced Computer Science and Applications*, 2(6), 2011.

[9] Jha, M. (2015). Document clustering using k-medoids, *International Journal on Advanced Computer Theory and Engineering*, 4(1), 2015.

[10] Murugavel, P., Punithavalli, M. (2011). Improved Hybrid Clustering and Distance-Based Technique for Outlier Removal, *International Journal on Computer Science and Engineering*, 3(1), Jan 2011.

[11] Vijayarani, S., Nithya, S. (2011). An Efficient Clustering Algorithm for Outlier Detection, *International Journal of Computer Applications*, 32(7), October 2011.

[12] Zhang, T., Ramakrishnan, R., Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases, *In: Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 103 – 114, 1996.

[13] Ismael, N., Alzaalan, M., Ashour, W. (2014). Improved Multi Threshold Birch Clustering Algorithm, *International Journal of Artificial Intelligence and Applications for Smart Devices*, 2(1), Feb 2014.

[14] Gupta, M., Rajavat, A. (2015). Time Improving Policy of Text Clustering Algorithm by Reducing Computational Overheads, *International Journal of Computer Applications*, 123(5), August 2015.

[15] Pujari, A. K. (2007). Data Mining Techniques, 3rd Ed., Universities Press, 2007.

[16] Song, S., Li, C. (2006). Improved ROCK for Text Clustering Using Asymmetric Proximity, SOFSEM, LNCS 3831, pp. 501 – 510, Jan 2006.

[17] Zhang, Q., Ding, L., Zhang, S. (2010). A Genetic Evolutionary ROCK Algorithm, *International*

Conference on Computer Application and System Modeling, 12, pp. 347 – 351, Nov 2010.

[18] Ahmad, R., Khanum, A. (2010). Document Topic Generation in Text Mining by Using Cluster Analysis with EROCK, *International Journal of Computer Science and Security*, 4(2), May 2010.

Authors' Profiles

NEHA GARG is an Assistant Professor in Department of Computer Science and Applications at ITM University, Gwalior. She has received her M.Tech from Madhav Institute of Technology and Science, Gwalior. Her research interests include Data Mining.

R. K. GUPTA is a Professor and Head in Department of Computer Science and Information Technology at Madhav Institute of Technology and Science, Gwalior. He has received his Ph.D. from ABV- IIITM Gwalior and M.Tech from IIT Delhi. His research interests include Data Mining.

How to cite this paper: Neha Garg, R.K. Gupta,"Exploration of Various Clustering Algorithms for Text Mining", International Journal of Education and Management Engineering(IJEME), Vol.8, No.4, pp.10-18, 2018.DOI: 10.5815/ijeme.2018.04.02