A Survey of Text Document Clustering by using Clustering Techniques

[1]T. Elavarasi* ,[2] Dr. R. Nagarajan
[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India
[2]Associate Professor, Department of Computer and Information Science, Annamalai University, Annamalai Nagar, Tamil Nadu, India
***Corresponding Author**
E-mail: arasithirugnanam@gmail.com

Abstract - *Clustering is one of the best important unsupervised data analysis technique, which divides data objects into clusters based on similarity and summarization of datasets. Clustering has been studied and applied in many different fields, including pattern recognition, advanced data mining, computational data science and Machine learning, information retrieval. This research focused on text document which are containing of similarities word. The combination of two algorithm methods, improved k-means and traditional k-means algorithm use to improving quality of initial cluster centres.*

***Key Words*: Text Clustering, K-means, Clustering Text Document, Text similarity**.

1. INTRODUCTION

Clustering is important data analysis technique, which divides data objects into clusters based on similarity and each cluster contains objects that are similar to other objects within same cluster [6]. Now a days there are many data on internet is dramatically increasing every single day by bay, clustering is considered an important data mining technique in categorizing, summarizing, classifying text documents. The data mining is extracting meaningful information or data from large datasets, the data mining techniques contains many fields like text mining, information extraction, document organization, information retrieval. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information.

Data clustering refers to an unsupervised learning technique, which offers refined and more abstract views to the inherent structure of a data set by partitioning it into a number of disjoint or overlapping (fuzzy) groups. Clustering refers to the natural grouping of the data objects in such a way that the objects in the same group are similar with respect to the objects present in the other groups. Document clustering is an important research direction in text mining, which aims to apply clustering algorithm on the textual data such that text documents can be organized, summarized and retrieved in an efficient way [6]. There are broadly three types of clustering, namely, Hierarchal clustering, Density based clustering, and Partition based clustering.

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. There are two types of method hierarchical clustering as Divisive and Agglomerative. The Divisive method is top-down clustering method and the observation to single cluster and then partition the cluster to two least similar clusters. The Agglomerative method is bottom-up clustering method and then compute the similarity between each of the clusters and join the two most similar clusters. The partitional clustering algorithm obtain k clusters of a set of data point without any hierarchical structure. Each cluster contains at least one object and each object belongs to exactly one cluster. Clustering methods used to classify observation, within data set, into multiple group based on their similarity. Partitional clustering algorithm contains algorithm like k-means, k-medoids or PAM **(**partitioning around medoids**)** etc.

The procedure of synthesizing the information by analyzing the relations, the patterns, and the rules among textual data - semi-structured or unstructured text. Why Text Mining? Massive amount of new information being create 80-90% of all data is held in various unstructured formats Useful information can be derived from this unstructured data.

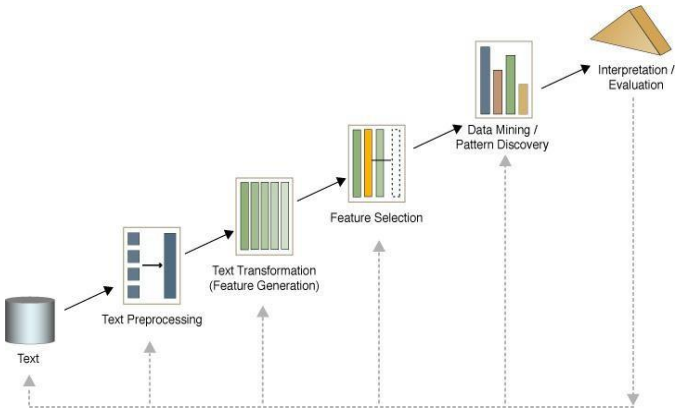

Figure 1.1: Text Mining Process

1.1 PARTITIONAL CLUSTERING ALGORITHMS

Partitional clustering algorithms divide a dataset of n objects into k clusters such that each partition represents a particular cluster. In this scheme, clusters are formed to optimize the chosen criterion, such as minimizing the sum of squared error [1]. In partitional clustering algorithms, all possible partitions are enumerated to achieve a global optimum. The enumeration of all possible partitions is not a computationally feasible task. Hence, heuristic methods have been frequently applied in partitional clustering. These heuristic approaches include, K-means, K-modes and K medoids algorithms. In the following subsections, important

partitional clustering algorithms (namely, K means, K means++ and K-medoids algorithms) utilized in the empirical analysis are briefly introduced [6].

A. K-means algorithm

K-means algorithm is a very popular clustering algorithm owing to its easy implementation, simplicity and efficiency. It takes the number of clusters (*k*) as the input parameter. It initiates with clusters. Each of the remaining objects are assigned to the clusters with the closest centres based on similarity. The algorithm continues to compute new mean of clusters until the stopping criterion. The algorithm gives promising results on clusters with well-aligned and compact shapes. It is an efficient clustering method with scalability. However, it may suffer from several shortcomings, such as being highly dependent on the position of randomly selected initial cluster centres [6].

B. K-means++ algorithm

As indicated in advance, the performance of K-means algorithm has been greatly affected by randomly selected initial cluster centres. Hence, K-means++ algorithm employs a randomized seeding mechanism to obtain more consistent clustering results. In the randomized seeding mechanism, a heuristic function is employed so that the initial centres for K means algorithm are carefully selected. The heuristic function utilizes a probabilistic distribution, which is obtained from the distance of data points to already selected initial centres [6].

2. METHODS

1) Similarity measures

Before clustering the documents, the similarity measure between the documents should be determined. There are two ex- tended ways which are being used to measure the correspondence among two documents [3].

2) Euclidean distance

Euclidean distance is a typical metric for many kinds of data analytical problems. It is also the normal distance between two arguments and so it can be measured in a multi-dimensional space Let the documents be d 1 and d 2 the Euclidean distance of these two documents is defined as *EDist* (d 1, d 2) = | d 1 −d 2 | [3].

3) Clustering Validation Techniques

1) Purity

Purity is an evaluation measure of how pure is a cluster with regard to the dominant class in that cluster. Purity is then computed based on the percentage of all objects of dominant classes for each cluster with regard to the number of all objects [2].

$$purity = \frac{1}{N}\sum_{k} max_j|\omega_k \cap c_j|$$

Where N is the number of all objects, k is the number of clusters, k is the dominant class, and cj is the real class (ground truth). The largest the value of purity the better clustering with maximum value of one if the dominant class of a cluster represents all objects in that cluster. [2].

2) F-measure

This measure is the harmonic mean of both recall and precision. Recall represents the fraction of documents of one category in one cluster out of all documents of that category. Whereas precision is the fraction of documents of one category in one cluster out of all documents in that cluster. Note from such definitions that values of precision and recall in isolation will not give a correct indication of the quality of clustering for several reasons found in the literature, therefore a combination of the two makes sense when appear in one measure, viz., the F-measure. To compute recall, precision and F-measure, then confusion matrix is usually used which is composed of four values [2].

	Same cluster	Different cluster
Similar documents	True Positive (TP)	False Negative (FN)
Different documents	False Positive (FP)	True Negative (TN)

- TP: indicates that the two documents are similar and belong to the same cluster.
- FN: indicates that the two documents are similar and belong to different clusters.
- FP: indicates that the two documents are different and belong to the same cluster.
- TN: indicates that the two documents are different and belong to different clusters.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

3) Performance Evaluation

- TP -True positive demonstrates allocates two related documents in same cluster.
- FP –False positive demonstrates assigns two dissimilar documents to same cluster.
- FN –False negative demonstrates assign two similar documents to different clusters.

- TN- True negative demonstrates allocates two dissimilar documents to different clusters.

Accuracy = (TP+TN) / (TP+TN+FP+FN)

3. RELATED WORK

The classical K-means algorithm selects K initial cluster centres randomly and calculates text similarity by Euclidean distance. The improved algorithm is based on the maximum distance method to pre-process and select K initial cluster centres, and the KL divergence is used to calculate the similarity and as per experiment on both algorithm final output is to improve time consumption and the improve time consumption of improved k-means algorithm is lower than traditional k-means algorithm[11].

The clustering on Arabic document with different types of datasets like BBC news, CNN news and the create VSM (vector space model) model and apply on TF-IDF weighting and normalize results. The after all process applying clustering algorithm is k-means and combined algorithm with LDA to validate cluster and to compare with clustering validation techniques like purity, entropy, F-measure and evaluation index like rand index, Jaccard index. The new method with new dataset uses which contain large data and the combined method in terms of improve clustering quality for Arabic text documents and the result shows that purity is **0.933** compared to **0.82** K-means algorithm [2].

The k-means++ algorithm apply on web services topic vectors on topic model documents. The similarity calculation between services is based on the TF-IDF and LDA is also similarity calculation between services is based on their probability topic vectors. There are many clustering measurements of methods like recall, precision, purity, entropy but the LDA-PK of F-measure is 0.7499 are best result than other evaluation index [7].

The proposed approach of PAM++ algorithm is compared with other partitional clustering algorithm, such as K-means and k-means++ on text document clustering and evaluated of F-measure. At the end of result improving the performance of PAM algorithm on text document clustering [6].

The using spectral clustering with particle swarm optimization to improve text document clustering with large data sets as to compare with other algorithms. The similarity measures use for distance measures like Euclidean, cosine, document representation, maximum likelihood estimation to improve cluster quality and accuracy. At the end of result SCPSO getting better performance result as to another algorithm constant 0.82 accuracy [3].

4. PROPOSED WORK

As far as research is done so far, small data set is used in K-means algorithm which restrict its applicability to traditional K-means algorithm. Accuracy varies according to initial cluster medoids. Sensitive to noise and outliers, so a small number of such data can substantially influence the mean value. Provides local optimum solution.

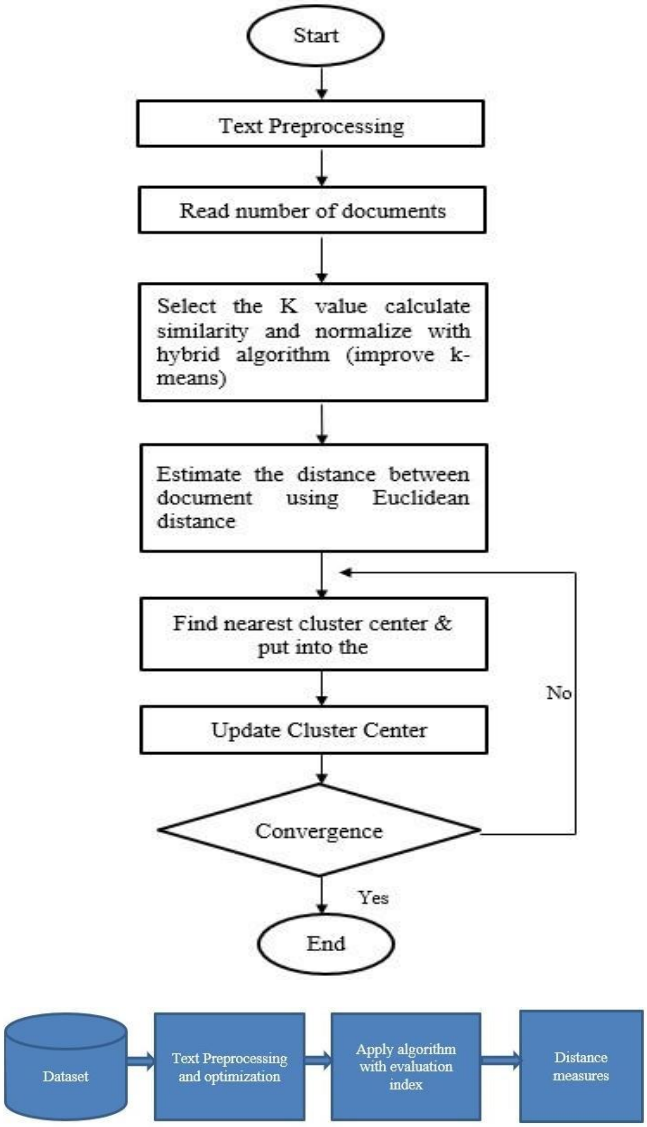

Figure 1.2 Block diagram of proposed method

As shown above the proposed model is divided in to four stages:

Stage 1 Dataset: text document dataset is taken as input data.

Stage 2 Preprocessing and optimization: Here the profiling is done by optimizing the dataset by reducing the noise and normalizing the values to reduce unusual semantic of data.

Stage 3 Apply algorithm with evaluation index-: the applying large dataset and optimizing data, after algorithm applying with indexes like rand index, NMI etc. and distance measures with Euclidean distance, Manhattan distance etc.

Stage 4 Distance measure: to measure distance of datasets of data with Euclidean distance and improve cluster quality.

5. CONCLUSION

The proposed system will definitely help in improving the text document clustering of k-medoids algorithm by increasing its accuracy and improve efficiency capability by reducing the unusual data. The existing systems are focused on using large data sets with improved k-means algorithm only, whereas the proposed system is working with combined stacking approach which is quite advantageous for improving the accuracy. The Standardized Euclidean distance method as dissimilarity measure, compute the distance between every pair of all objects and root mean square error is distance measure which is sum of the squared differences between pairs of measurement.

REFERENCES

1. Afzal, M. and Kumar, S., 2019, February. Text Document Clustering: Issues and Challenges. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Comicon) (pp. 263-268) IEEE.

2. Ahlawat, M. and Hegazi, M., 2018. Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. IEEE Access, 6, pp.42740-42749.

3. Janani, R. and Vijayarani, S., 2019. Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. Expert Systems with Applications, 134, pp.192-200.

4. Jin, C.X. and Bai, Q.C., 2016, June. Text clustering algorithm based on the graph structures of semantic word co-occurrence. In 2016 International Conference on Information System and Artificial Intelligence (ISAI) (pp. 497-502). IEEE.

5. Madaan, V. and Kumar, R., 2018, October. An Improved Approach for Web Document Clustering. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) (pp. 435-440). IEEE.

6. Onan, A., 2017, October. A K-medoids based clustering scheme with an application to document clustering. In 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 354-359). IEEE.

7. Shi, M., Liu, J., Cao, B., Wen, Y. and Zhang, X., 2018, July. A prior knowledge based approach to improving accuracy of Web services clustering. In 2018 IEEE International Conference on Services Computing (SCC) (pp. 1-8). IEEE.

8. Wang, B., Yin, J., Hua, Q., Wu, Z. and Cao, J., 2016, August. Parallelizing k-means-based clustering on spark. In 2016 International Conference on Advanced Cloud and Big Data (CBD) (pp. 31-36). IEEE.

9. Wang, X., Li, Y., Wang, M., Yang, Z. and Dong, H., 2018, October. An Improved K_Means Algorithm for Document Clustering Based on Knowledge Graphs. In 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-5). IEEE.

10. Zhang, H., Guo, X., Ye, L. and Li, S., 2018, December. Marrying K-means with Evidence Accumulation in Clustering Analysis. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 2050-2056). IEEE.

11. Huan, Z., Pengzhou, Z. and Zeyang, G., 2018, June. K-means Text Dynamic Clustering Algorithm Based on KL Divergence. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (pp. 659-663). IEEE.

12. Xinwu, L., 2009, June. Research on Text Clustering Algorithm Based on Improved K_means. In 2009 ETP International Conference on Future Computer and Communication (pp. 19-22). IEEE.

13. Song, W., Liang, J.Z. and Park, S.C., 2014. Fuzzy control GA with a novel hybrid semantic similarity strategy for text clustering. Information Sciences, 273, pp.156-170.

14. Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H. and Zhang, G., 2018. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding Journal of Informatics, 12(4), pp.1099-1117.

15. Ailem, M., Role, F. and Nadif, M., 2015, October. Co-clustering document-term matrices by direct maximization of graph modularity. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1807-1810). ACM.

16. Ailem, M., Role, F. and Nadif, M., 2017. Sparse poisson latent block model for document clustering. IEEE Transactions on Knowledge and Data Engineering, 29(7), pp.1563-1576.

17. Liao, K., Liu, G., Xiao, L. and Liu, C., 2013. A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval. Knowledge-Based Systems, 49, pp.123-133.

18. Onan, A., Bulut, H. and Korukoglu, S., 2017. An improved ant algorithm with LDA-based representation for text document clustering. Journal of Information Science, 43(2), pp.275-292.

19. Rossi, R.G., Marcacini, R.M. and Rezende, S.O., 2013. Benchmarking text collections for classification and clustering tasks. Institute of Mathematics and Computer Sciences, University of Sao Paulo.

20. Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.

21. Park, H.S. and Jun, C.H., 2009. A simple and fast algorithm for K-medoids clustering. Expert systems with applications, 36(2), pp.3336-3341.

22. Neha, D. and Vidyavathi, B.M., 2015. A survey on applications of data mining using clustering techniques. International Journal of Computer Applications, 126(2).

23. Djenouri, Y., Belhadi, A. and Belkebir, R., 2018. Bees swarm optimization guided by data mining techniques for document information retrieval. Expert Systems with Applications, 94, pp.126-136.

24. Reddy, G.S., Rajinikanth, T.V. and Rao, A.A., 2014, February. A frequent term based text clustering approach using novel similarity measure. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 495-499). IEEE.

25. Chen, Y. and Sun, P., 2018, August, an Optimized K-Means Algorithm Based on FSTVM. In 2018 International Conference on Virtual Reality and Intelligent Systems [ICVRIS] (pp. 363-366). IEEE.