Review Article

Novel Journal of Applied Sciences Research

Comprehensive Survey of Document Clustering Methods: Exploring Traditional, Hybrid, and Meta-Heuristic Approaches

Qazi Waqas Khan[1]*, and Fazila Malik[2]

[1]Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea
[2]Department of Computer Science, Iqra University Islamabad, Islamabad 44000, Pakistan

***Corresponding author:** Qazi Waqas Khan, Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea.

*These authors contributed equally

Submitted: 31 July 2024 **Accepted:** 01 August 2024 **Published:** 02 August 2024

***Citation:** Qazi Waqas Khan, and Fazila Malik (2024) Comprehensive Survey of Document Clustering Methods: Exploring Traditional, Hybrid, and Meta-Heuristic Approaches. Nov Joun of Appl Sci Res 1(4), 01-03.*

Abstract

Document Clustering is a process of combining the data into k groups based on similarity. In the literature, many document clustering methods cluster documents based on their similarity. There is a need for a comprehensive survey where we discuss the details of existing clustering methods. This survey paper discussed existing document clustering methods, such as k means and other hybrid and meta-heuristic-based clustering methods. The existing literature suggests a hybrid and meta-heuristic-based method enhanced the performance of document clustering.

Keywords: K-means Clustering, Hybrid Algorithms, Meta-heuristic Approaches, Document Clustering, Optimization Problems.

Introduction

Clustering categorizes a population N data point into K subgroups so that data points in one group are more similar to those in other groups. The higher the resemblance inside a group and the more significant the variance between groups, the better or more definite the clustering. It is a method that converts vast amounts of data into understandable information. With reduced data dimensions, we efficiently minimize the time a computer takes to collect the requested information. Clustering is an unsupervised learning approach with no class labels [1]. The primary advantage of unsupervised learning is solving problems that humans may find impossible due to limited capacity or a lack of equality.

The fundamental goal of clustering is dividing data into reasonable groupings based on similarity. Clustering helps to define the internal structure of data and is also useful for exploring data [2]. Clustering methods can be applied to detect anomalous behaviour, such as segmenting customers on their buying patterns and reducing large datasets into a smaller number of related categories.

Clustering is evaluated using intra-cluster and inter-cluster distance. Intra-cluster distance is the distance between data points within a cluster. This distance should be small if there is good clustering. Data points in different clusters are separated by inter-cluster distance. The inter-class distance should be large if there is good clustering [3]. Clustering methods are classified into two major categories, Hierarchical and Partitional clustering. There are a variety of subtypes and algorithms for identifying clusters within each type [4]. This study discussed the details of the existing document clustering method.

Existing Methods for Document Clustering

K-means algorithm is heuristic and capable of finding clusters in polynomial time. But optimal or best clusters are not guaranteed due to their inherent drawbacks.

By improving the initial cluster center, Caiquan, X. et al. [5] suggested an improved K-means text clustering algorithm. This algorithm major goal is to identify the first cluster centers based on the density parameters of dataset, ensuring that the initial cluster centers are consistent. This approach helps to improve text clustering results by removing the sensitivity of K-means algorithm to the initial cluster center. Experiments are conducted on five different categories of Chinese corpus data (politics, art, economics, sports, and environment). Two measures are used to determine the efficiency of the clustering method: accuracy and recall. Experiment findings indicate that the improved K-means method can enhance text clustering stability and accuracy. The suggested approach takes a long time to execute and has a greater time complexity than the k-mean algorithm because calculating the distance between all data points is required to find the initial cluster center. The problem of determining the value of k has not been solved.

For effective initial seed selection, Kumar, K. et al. proposed RDBI (Robust Density-Based Initialization) approach for improving K-means filtering algorithm [6-10]. Using this approach the seed points are found in dense portions of the dataset, which are recognized by representing the data in kd-tree. Experiments are executed on a variety of synthetic and real data sets (Image Segmentation, Pen Digits, Letter Recognition, Shuttle, and Poker Hand) from the UCI Machine Learning Repository. This approach can handle large datasets because the complexity of the algorithm is linearly proportional to a number of features. With this new technique, K-means filtering method is improved for high-dimensional data and clusters of undifferentiated centers. The average distance computation and running time are used to examine a method's performance. The pre-identification of k-values in the k-mean method is not addressed [11-14].

Table 1: Literature reviews about initial centroids selection

R.No	Author/ Date	Problem Focus area and Methodology	Dataset	Evaluation Measure	Results strength	Limitations
[5]	Caiquan X et.al 2016	Improved K-means for initial cluster center optimization	Political, Art, Economy, Sport and En-vironment	Precision and recall	According to the reviewed literature, improved K-means method can en-hance text clustering sta-bility and accuracy	K-value de-termination is not being addressed.
[6]	Kumar, K. et al 2017	Robust Densi-ty-Based Initiali-za-tion approach for K-mean filtering	Image Seg-men-tation, Pen Digits, Letter Recog-ni-tion, Shut-tle, and Poker Hand	Average dis-tance compu-ta-tion and run-ning time	With this new technique, K-means filtering method is improved for clusters of undifferentiated centers and high-dimensional data	Pre-identifica-tion of k-value in is not addressed
[7]	Lakshmi, R et.al 2019	DIC-DOC k-means algorithm	Webkb and Reuters 8	Entropy, purity, and F-measure	Proposed approach DIC-DOC k-means algorithm performs better	Identifica-tion of k-values is not addressed
[8]	Rajini kanth, T et.al 2017	Algorithm based on k-means and fuzzy similarity measures for document clus-tering	Reuters 8	Silhouette score	Proposed parameter (Gaussian membership) improves the performance of standard k-means clus-tering	The number of clusters k is done by user manual-ly

Above table 2 shows a number of challenges in literature which are as follows:

- Traditional clustering algorithms perform clustering in full dimension space, thus their performance degrades with the increase in dimensions. This problem is termed as the “curse of dimensionality”.
- Distance measure loses its significance as there is inherent sparsity in data space. The Euclidean distance between farthest point and closest point decreases with an increase in dimensions.
- With increase in dimensions, number of clusters grows exponentially. Finding optimal clusters in high dimensional data is NP-hard problem.
- All dimensions may not be important for all clusters. Clusters may exist in various subsets of dimensions i.e. subspaces. Hence finding clusters in different subspaces of high dimensional data is a challenging problem.

Table 2: Literature review of hybridization of algorithms

.N Q	Author/ Date	Problem Focus area and Methodology	Dataset	Evaluation Measure	Results strength	Limitations
9	Chouhan R. et al	Combined PSO and K-means	BBC Sports, FOX, BBC, and CNN	Cohesion, entropy, and separation	According to the reviewed literature, the proposed approach outperforms that the traditional K-means	Determining the number of clusters remains a challenge
10	Lakshmi, K. et al	Crow Search Algorithm with K-means	Glass, Breast Cancer, CMC, Iris, Wine, and Haberman's Survival	Purity, Rand Index, Recall, F-Measure Precision	The CSAK method outperforms than the K-means, K-means++, Genetic K-means, and PSOK-means algorithms in test experiments	There is a need to automatically decide the number of clusters

11	Yogesh G. et al	Enhanced fuzzy PSO-based clustering method with K-harmonic means (EFPSOKHM) for clustering	Numeric (Cancer, Iris, Wine, CMC, Glass) and text datasets (CACM and CISI)	F-measure, Inter-cluster distance Fitness value Intra-cluster distance, Runtime	The proposed approach (EFPSOKHM) is more stable and produces more consistent results than other evolutionary-based clustering algorithms	K-value identification is not handling. The suggested method takes longer to run than standard clustering methods like K-means
12	Abualigah LM et al	Hybrid KH (Krill herd) algorithm named (MHKHA)	Nine text standard datasets	Convergence behavior, accuracy, recall, precision, and F-measure	The MHKHA produced the best results for all datasets	Determining the number of clusters remains a challenge, require the number of clusters to be predetermined
13	Yong Wang et al	Hybrid approach based on Hybrid Rice Optimization algorithm and k-means for cluster centers	Wine, Seeds, and Iris	Running time	The presented algorithm saves running time while increasing algorithm efficiency	The pre-determination of k-value is not addressed
14	Thangarasu M et al	Hybrid algorithm namely DPPSOK which is based on PSO, k-means	http://trec.nist.gov	Normalized Euclidian distance	DPPSOK performs better than traditional k-means and PSOK algorithms	The k-value identification issue is not resolved

The studies mentioned in Table 2 prove the suitability of nature-inspired algorithms (NIA) with heuristic algorithms for finding near-optimal solutions for hard problems. These algorithms have gained immense popularity to solve complex optimization problems where an actual solution does not exist. Due to their immense usefulness, large number of nature-inspired algorithms have been developed and classified into various categories.

Conclusion

This review highlights the significant advancements in document clustering methods, particularly the effectiveness of hybrid and meta-heuristic approaches in addressing complex optimization problems. The discussed algorithms show promise in improving clustering accuracy and stability. However, challenges such as the determination of the number of clusters and handling high002D-dimensional data persist, necessitating further research and development in this area.

References

1. Mahmoud A Mahdi, Khalid M Hosny, Ibrahim Elhenawy (2021) Scalable Clustering Algorithms for Big data: A Review. IEEE Access Corpus ID: 235383925
2. Dhote CA, Anuradha D Thakare, and Shruti M Chaudhari (2013) Data clustering using particle swarm optimization and bee algorithm. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). IEEE.
3. Vijay Kumar, Jitender Kumar Chhabra, Dinesh Kumar (2014) Performance evaluation of distance metrics in the clustering algorithms. INFOCOMP Journal of Computer Science 13: 38-52.
4. Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, et al. (2017) A review of clustering techniques and developments. Neurocomputing 267: 664-681.
5. Caiquan Xiong, Zhen Hua, Ke Lv, Xuan Li (2016) An Improved K-means text clustering algorithm By Optimizing initial cluster centers. 7th International Conference on Cloud Computing and Big Data (CCBD). IEEE.
6. Mahesh Kumar K, Rama Mohan Reddy A (2017) An efficient k-means clustering filtering algorithm using density based initial cluster centers. Information Sciences 418: 286-301.
7. Lakshmi R, Baskar S (2019) DIC-DOC-K-means: dissimilarity-based Initial Centroid selection for DOCument clustering using K-means for improving the effectiveness of text document clustering. Journal of Information Science 45: 818-832.
8. Rajinikanth T, Suresh Reddy G (2017) A soft similarity measure for k-means based high dimensional document clustering. IADIS International Journal on Computer Science & Information Systems 12.
9. Rashmi Chouhan, Anuradha Purohit (2018) An approach for document clustering using PSO and K-means algorithm. 2nd International Conference on Inventive Systems and Control (ICISC). IEEE.
10. Lakshmi K, Karthikeyani Visalakshi N, Shanthi S (2019) Data clustering using k-means based on crow search algorithm. Sādhanā 11: 1-12.
11. Yogesh Guptha, Ashish Saini (2019) A new swarm-based efficient data clustering approach using KHM and fuzzy logic. Soft Computing 23: 145-162.
12. Laith Mohammad Abualigah, Ahamad Tajudin Khader, Essam Said Hanandeh (2018) A combination of objective functions and hybrid krill herd algorithm for text document clustering analysis. Engineering Applications of Artificial Intelligence 73: 111-125.
13. Xiaoyan Wang, Yanping Bai (2016) A modified minmax-means algorithm based on pso. Computational intelligence and neuroscience.
14. Thangarasu M, Hannah Inbarani H (2016) DPPSOK Algorithm for Document Clustering. 201-207.