SCIENCE SET
OPEN ACCESS PUBLISHERS

Research Article

Novel Journal of Applied Sciences Research

Enhancing Clustering of News20 Dataset Using Cosine Similarity and K-Means: An Evaluation of Performance Metrics

Qazi Waqas Khan*

Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea

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

Submitted: 11 July 2024 **Accepted:** 15 July 2024 **Published:** 22 July 2024

Citation: *Qazi Waqas Khan (2024) Enhancing Clustering of News20 Dataset Using Cosine Similarity and K-Means: An Evaluation of Performance Metrics. Nov Joun of Appl Sci Res 1(4), 01-05.*

Abstract

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 fundamental goal of clustering is dividing data into reasonable groupings based on similarity. Clustering helps define and explore the internal structure of data. Clustering methods can be applied to detect abnormal behavior, segment customers on their buying patterns, and reduce large datasets into fewer related categories. This study used the cosine similarity with the K means clustering method to cluster a news20 dataset. The performance of a proposed system is evaluated using the homogeneity, completeness, V-measures, adjusted rand index, and silhouette coefficient metrics. The experimental findings of a proposed method show the proposed method achieved better performance for clustering of a News 20 dataset.

Introduction

Data is growing at an incredible rate [1]. We live in a time where technology is advancing quickly, allowing for a considerable amount of information to be generated through online platforms like blogs and social media [2]. Every day, an enormous amount of data, equivalent to 2.5 quintillion bytes, is produced [3]. The past two years alone have contributed to 90% of all the data in the world [4]. With so much data available, finding the information that interests us has become challenging. To address this challenge, a field called "searching for useful data" has emerged in the Big Data world [5]. This field focuses on finding ways to effectively search for and extract valuable information from massive data [6]. It uses advanced techniques like algorithms, data analysis, and machine learning to discover patterns and insights that can be helpful for different purposes [7]. The following objectives have been taken for current research work. To analyze various existing document clustering techniques, propose a framework for efficient document clustering using different variations with the k-means algorithm, implement the proposed framework, and measure its performance.

If the objective is to reveal and explore the hidden patterns in data, clustering becomes an exploratory method. However, if the generated clusters facilitate other machine learning or data mining tasks, clustering is used in the pre-processing stage. Clustering is a method that can be applied to some data mining and machine learning tasks like network analysis, pattern classification, pattern recognition, information retrieval, image segmentation, etc [8]. It can be applied in a pre-processing stage or for an exploratory task [9]. Although the number of clustering algorithms presented so far is relatively large, improving the clustering quality is still possible. Document clustering is complex because of the complexity of words and their relationships [10]. The following are the two main challenges for enhancing clustering quality: semantically disambiguated terms are used to represent texts containing ambiguous words, and after that, clustering using an efficient method can be carried out as the next step [11]. There are several clustering techniques to choose from. These techniques may be broken down into a few categories: density-based, hierarchical, and partitioning [12]. When using distance-based approaches for partitioning, the points are clustered by how similar they are. This class of algorithms creates one-level partitioning and non-overlapping spherical clusters. K-medoid and k-means are two commonly used partitioning algorithms. Opposing this, the hierarchical method works by partitioning the data into different levels, which appears and works like a hierarchy. The clustering method helps with better summarization and data visualization. In this work we used the k

means algorithm with cosine similarity on the publicly available data set of News 20.

Proposed System

Document clustering is the automated clustering of text documents into groups with high similarity values among documents within a cluster but dissimilarity values across other clusters. Search engines, information retrieval, web mining, and topological analysis are only a few of its wide range of applications. The proposed system applies different phases, and the outcomes are presented. The performance measures of the clusters have also been explored.

Document clustering may also be applied to create partition groupings of documents. To successfully search and retrieve data in DMS ("Document Management Systems"), metadata set for the documents must be generated with sufficient details. However, one metadata collection is insufficient for entire document management systems. This is because different document types require distinct properties to be properly identified.

Proposed Architecture

Clustering techniques are used in document clustering to create groups of similar documents. However, clustering methods could not be directly used to datasets comprising documents. There needed to be a series of stages before applying the clustering approach to a text document. Modules such as extractor, document reader, pre-processor, and VSM creator are used to accomplish these processes. The following diagram, Figure 1, depicts the whole document clustering process, including various components.

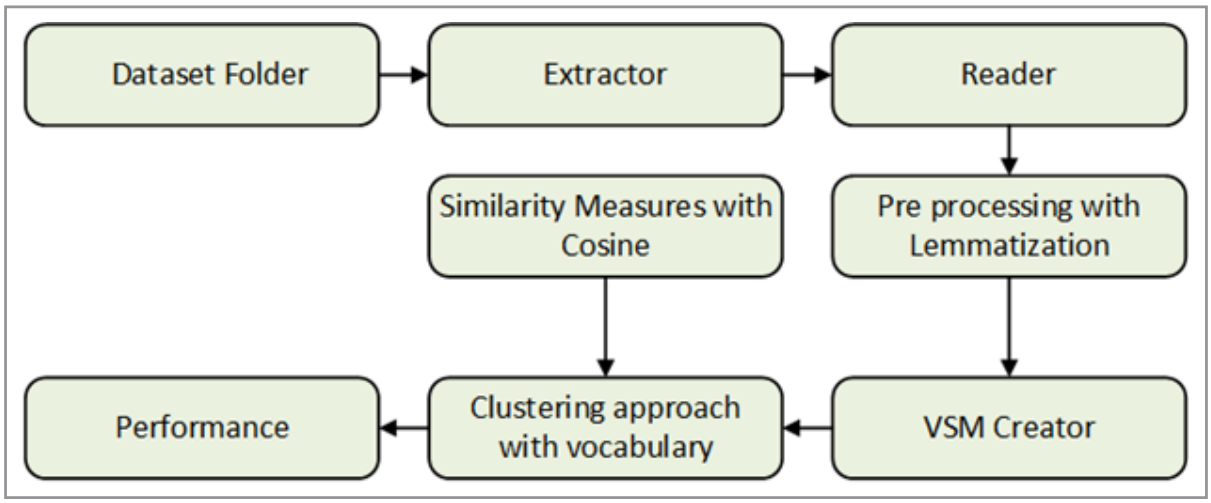

Figure 1: Architecture of Proposed System

- **Dataset Folder:** The dataset folder contained the different classes' folder of the input dataset and the categories of documents.
- **Extractor:** In the step of extractor, read the dataset and give the various classes, and every class maintains a separate document.
- **Reader:** Every document has several lines, and each line maintains the group of words. So, the reader reads every document line by line and gives the various words of these lines.
- **Preprocessing with Lemmatization:** The proposed system uses Lemmatization instead of stemming. Lemmatization is the process of converting a word to its base form. It considers the context and converts the word to its meaningful base form. If we lemmatize the word 'Caring,' it would return 'Care.'
- **VSM Creator:** The vector space model converts documents with words to numerical forms with the help of a document matrix.
- **Similarity Measures with Cosine:** In this method, documents are represented as term vectors. The comparability of two documents relates to the relationship of the vectors, which is measured as the cosine of the edge between vectors. So, in the proposed system, the cosine similarity measure is replaced by Euclidean distance. In the chapter, I discussed all similarity measures in detail.
- **Clustering Approach with Vocabulary:** In proposed clustering Approach use vocabulary and divided the words of vocabulary into k parts. Where k= no. of clusters shown as figure 2
- **Performance:** In this phase, evaluate the results of the proposed system in terms of given parameters.

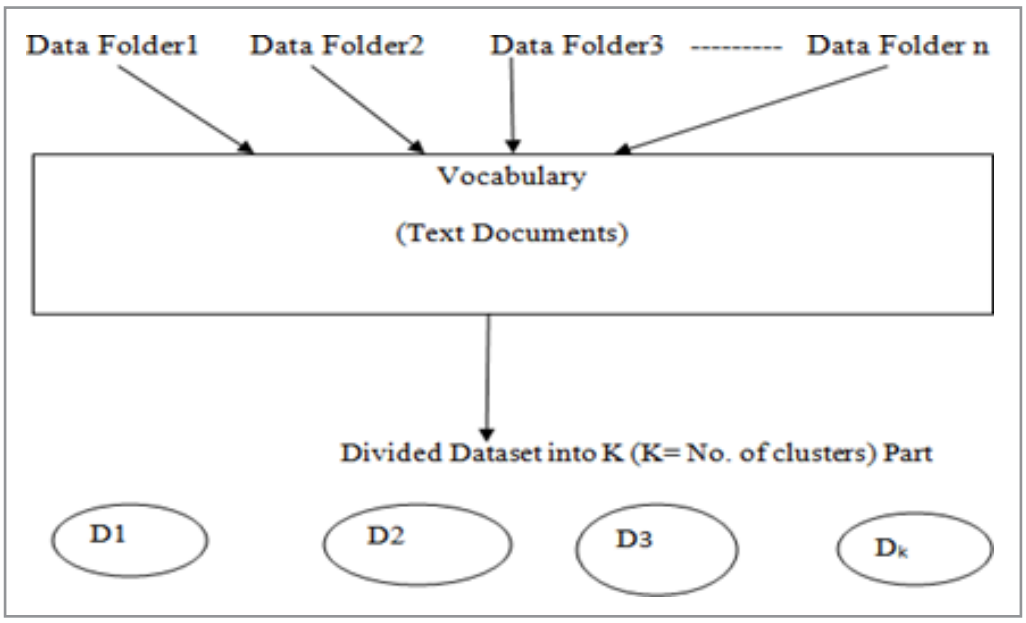

Figure 2: Representation of Vocabulary

Flow Chart for Proposed System

The flow chart of the proposed system is shown in Figure 3. This flow chart represented the different phases that work for outcome of research work. Input the dataset according to vocabulary and convert the textual data into numerical data that's help for clustering approach.

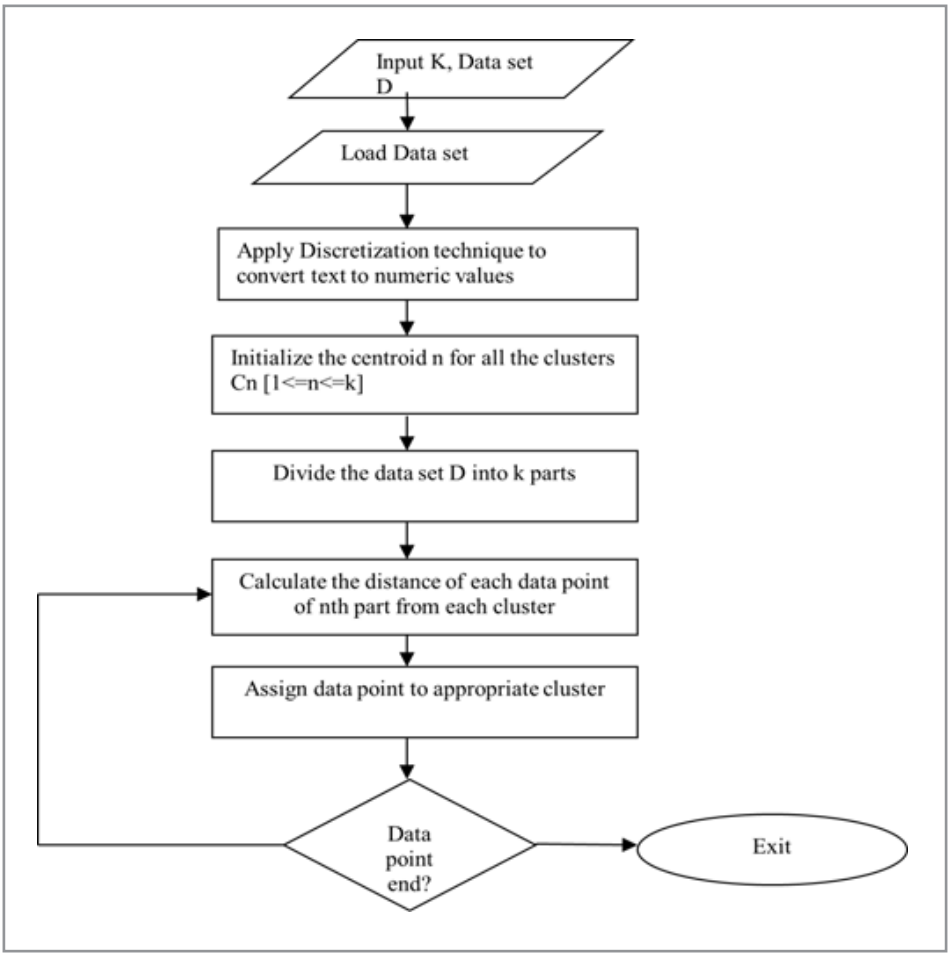

Figure 3: Flow Chart for Proposed System

Performance Metrics

Evaluate how well the clustering works by comparing the groups produced by the clustering approach to known classes. However, if one clustering approach performs better than other clustering approach on many of these measures. Performance measures include homogeneity, completeness, V-measures, adjusted rand index, and silhouette coefficient.

Experimental Results

As per the evaluation of cluster quality, Figure 4 shows the different parameters like homogeneity, completeness, V-measure, ARI, and silhouette for k-means, k-means with Lemmatization, k-means with cosine similarity, and the proposed system on cluster size 10. As per representation, the proposed system gives a better result than others because of the probability of the proposed system being better in all parameters than other approaches.

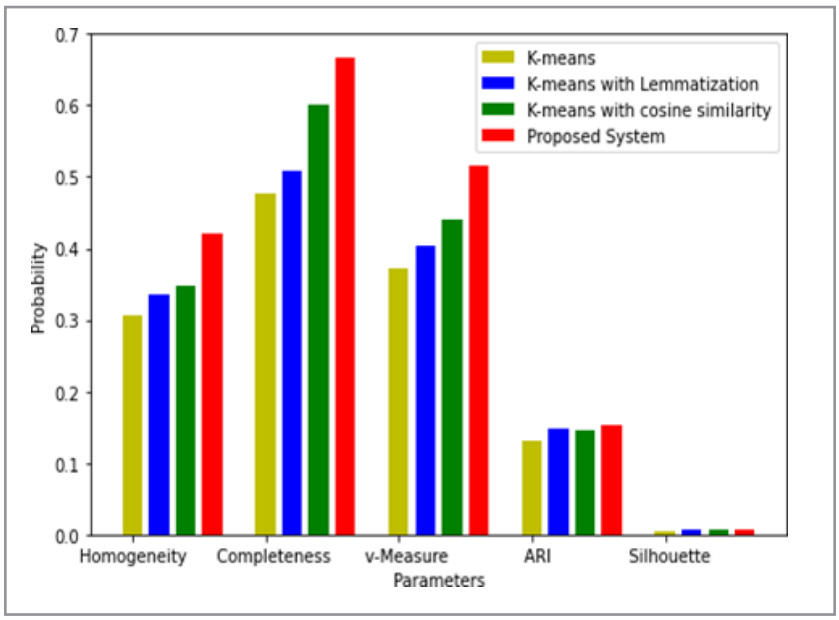

Figure 4: Performance with K=10

The proposed system implemented different cluster sizes to give the parameter values. Figure 5 shows the outcome of the proposed system for cluster size 15 with various parameters. This shows a remarkable outcome in comparison with k-means and others because the size of the cluster has increased.

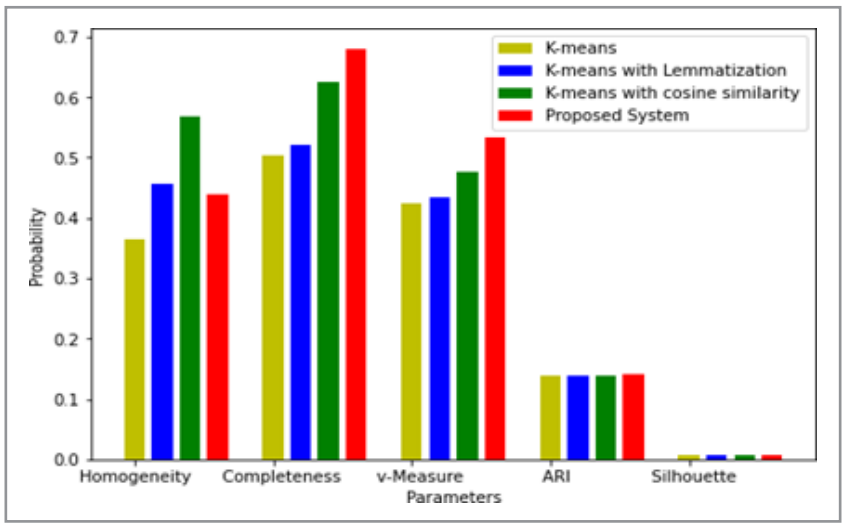

Figure 5: Performance with K=15

The output of the proposed system when the cluster size becomes 20 is shown in Figure 6. The proposed system gives makeable results compared to other approaches in all parameters because the probability has to balance the cluster when the cluster size varies in increasing order.

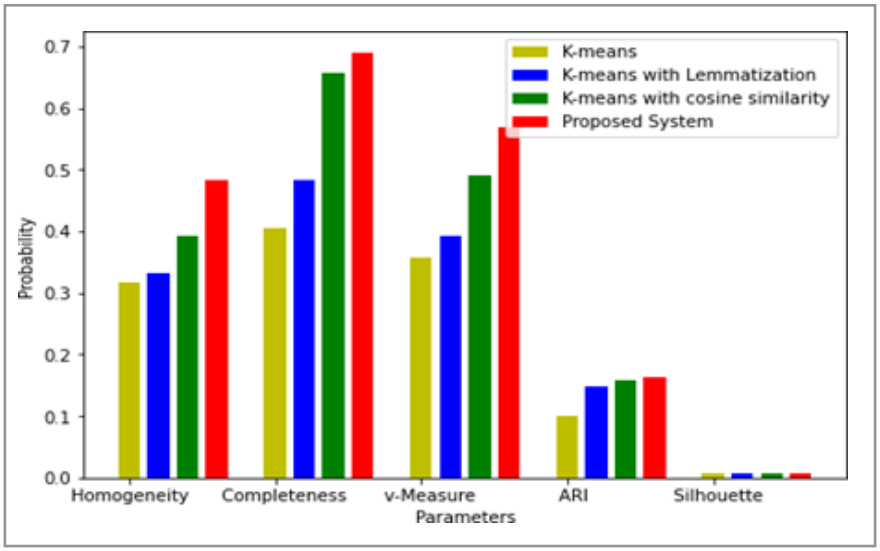

Figure 6: Performance with K=20

Conclusion

The proposed system is used for clustering where the segmentation or division of the data is appropriate. Appropriate segmentation would be assigned based on the number of clusters. This measurement may impact the number of computer resources needed since clustering is frequently followed by further processing of the individual cluster data. So, the proposed system is the best way to form clusters. These different parameters represent the output of the proposed system, which is remarkably different from the existing system. As per the document clustering application, the proposed system gives up performance. This gives wings to the various applications of document clustering to form clusters with optimum resource usage.

References

1. Danny Dorling (2020) Slowdown: The end of the great acceleration—and why it's good for the planet, the economy, and our lives. Yale University Press, 2020.
2. Anastasiia Bessarab, Olha Mitchuk, Anna Baranetska, Natalia Kodatska, Olha Kvasnytsia, et al. (2021) Social networks as a phenomenon of the information society. Journal of Optimization in Industrial Engineering Special Issue 14: 17-24.
3. Sharmila, Dhananjay Kumar, Pramod Kumar, Alaknanda Ashok (2020) Introduction to multimedia big data computing for IoT. Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions 3-36.
4. Rosamond L Naylor, Ronald W Hardy, Alejandro H Buschmann, Simon R Bush, Ling Cao, et al. (2021) A 20-year retrospective review of global aquaculture. Nature 591: 551-563.
5. Michael Mattioli (2017) The data-pooling problem. Berkeley Technology Law Journal 32: 179-236.
6. Amina Adadi (2021) A survey on data-efficient algorithms in big data era. Journal of Big Data 8: 24.
7. Mohammad Mustafa Taye (2023) Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. Computers 12: 91.
8. Shahneela Pitafi, Toni Anwar, Zubair Sharif (2023) A taxonomy of machine learning clustering algorithms, challenges, and future realms. Applied sciences 13: 3529.
9. Aitdaoud, Mohammed, Abdelwahed Namir, Mohammed Talbi (2023) A New Pre-Processing Approach Based on Clustering Users Traces According to their Learning Styles in Moodle LMS. International Journal of Emerging Technologies in Learning 18: 226-242.
10. Abiodun M Ikotun, Absalom E Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming (2023) K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences 622: 178-210.
11. Abdo Ababor Abafogi (2023) Enhanced word sense disambiguation algorithm for Afaan Oromoo. International Journal of Information Engineering and Electronic Business 14: 41.
12. Xingcheng Ran, Yue Xi, Yonggang Lu, Xiangwen Wang, Zhenyu Lu (2023) Comprehensive survey on hierarchical clustering algorithms and the recent developments. Artificial Intelligence Review 56: 8219-8264.