Seminars in Medical Writing and Education. 2024; 3:68
doi: 10.56294/mw202468

ORIGINAL

Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques

Exploración de documentos científicos mediante técnicas de segmentación basadas en aprendizaje no supervisado

Mohamed Cherradi[1] ✉, Anass El Haddadi[1]

[1]Data Science and Competetive Intelligence Team (DSCI), ENSAH. Abdelmalek Essaâdi University (UAE). Tetouan, Morocco.

Cite as: Cherradi M, El Haddadi A. Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques. Seminars in Medical Writing and Education. 2024; 3:68. https://doi.org/10.56294/mw202468

Submitted: 30-10-2023 **Revised:** 26-01-2024 **Accepted:** 31-03-2024 **Published:** 01-04-2024

Editor: Dr. José Alejandro Rodríguez-Pérez

ABSTRACT

Navigating the extensive landscape of scientific literature presents a significant challenge, prompting the development of innovative methodologies for efficient exploration. Our study introduces a pioneering approach for unsupervised segmentation, aimed at revealing thematic trends within articles and enhancing the accessibility of scientific knowledge. Leveraging three prominent clustering algorithms—K-Means, Hierarchical Agglomerative, and DBSCAN—we demonstrate their proficiency in generating meaningful clusters, validated through assessment metrics including Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. Methodologically, comprehensive web scraping of scientific databases, coupled with thorough data cleaning and preprocessing, forms the foundation of our approach. The efficacy of our methodology in accurately identifying scientific domains and uncovering interdisciplinary connections underscores its potential to revolutionize the exploration of scientific publications. Future endeavors will further explore alternative unsupervised algorithms and extend the methodology to diverse data sources, fostering continuous innovation in scientific knowledge organization.

Keywords: Document Clustering; Information Retrieval; K-Means; CAH; DBSCAN.

RESUMEN

Navegar por el extenso panorama de la literatura científica supone un reto importante, que impulsa el desarrollo de metodologías innovadoras para una exploración eficiente. Nuestro estudio presenta un enfoque pionero para la segmentación no supervisada, cuyo objetivo es revelar tendencias temáticas dentro de los artículos y mejorar la accesibilidad del conocimiento científico. Aprovechando tres destacados algoritmos de agrupación -K-Means, aglomerativo jerárquico y DBSCAN-, demostramos su capacidad para generar agrupaciones significativas, validadas mediante métricas de evaluación como la puntuación Silhouette, el índice Calinski-Harabasz y el índice Davies-Bouldin. Desde el punto de vista metodológico, la base de nuestro enfoque es el análisis exhaustivo de las bases de datos científicas, junto con una limpieza y preprocesamiento minuciosos de los datos. La eficacia de nuestra metodología para identificar con precisión los ámbitos científicos y descubrir conexiones interdisciplinarias subraya su potencial para revolucionar la exploración de las publicaciones científicas. En el futuro exploraremos algoritmos no supervisados alternativos y ampliaremos la metodología a diversas fuentes de datos, fomentando la innovación continua en la organización del conocimiento científico.

Palabras clave: Agrupación de Documentos; Recuperación de Información; K-means; CAH; DBSCAN.

INTRODUCCIÓN

In the realm of scientific literature, the proliferation of documents within scientific databases has led to an overwhelming volume of data. With this surge in information, extracting meaningful insights from heterogeneous datasets remains a challenging task. Particularly challenging is the segmentation of these datasets to uncover relevant trends and patterns. Our research is situated within this context, focusing specifically on the application of clustering techniques based on unsupervised learning algorithms to scientific documents. As the number of documents continues to grow exponentially, it remains imperative to develop efficient methods for organizing and extracting insights from this wealth of information. By delving into the realm of unsupervised learning and clustering, we aim to contribute to the development of strategies that enable researchers, academic institutions, and decision-makers with actionable insights to navigate and extract valuable knowledge from large and diverse scientific databases effectively.

Within this paper, we present a structured approach aimed at addressing a specific research challenge within the domain of scientific inquiry. In the contemporary era of digital information, the proliferation of scientific literature has presented challenges in effectively extracting meaningful insights.[1] Our focus shifts towards precisely defining the specific research problem at hand within the realm of scientific literature analysis. This entails identifying potential limitations in existing methodologies for analyzing scientific documents, exploring new avenues for interpretation, and delving into unresolved inquiries regarding the discovery of latent thematic connections. Emphasizing the importance of this targeted problem or question is crucial to ensure its recognition and significance within the broader scientific discourse.

In light of the significant challenge posed by the extensive volume of scientific documents housed within databases like Scopus and Web of Science. The process for an effective methodology for segmentation and analysis is crucial.[2,3,4,5] Introducing a pioneering approach distinguished by its efficiency and efficacy, our research proposal responds to this pressing need. By integrating a diverse set of unsupervised clustering algorithms, including k-means, hierarchical clustering, and DBSCAN, our methodology offers a robust framework capable of addressing the intricate complexities inherent in scientific document datasets. What truly distinguishes our proposal is its rigorous evaluation process, which carefully assesses performance using key metrics such as the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. This comprehensive evaluation not only ensures the reliability and effectiveness of our approach but also underscores its originality and novelty within the field. Through our research contribution, we aspire to propel advancements in the domain of scientific document analysis, providing insights that facilitate more precise and efficient segmentation of scientific literature.

The reminder of the paper is organized as follows: Section 2 delves into a comprehensive review of related works, providing insights into existing methodologies and approaches in the field of scientific document analysis and clustering techniques. In Section 3, we present our research methodology, outlining the detailed steps involved in implementing unsupervised clustering algorithms and evaluating their performance metrics. Following this, Section 4 presents the results of our experiments and initiates discussions on the implications of our findings. Finally, in Section 5, we offer concluding remarks and outline future perspectives for further research endeavors in the domain of scientific document analysis and clustering methodologies.

Related Works

In this section, we present a selective review of recent works closely related to our research topic of clustering techniques applied to scientific document analysis. Our focus is on identifying state of the art methodologies and key contributions that inform the context of our study.

One notable study by Wang and Koopman[6] introduced a novel approach to clustering scientific articles based on semantic similarity, leveraging advanced natural language processing techniques. Their methodology not only demonstrated promising results in accurately identifying thematic trends within large document collections but also highlighted the potential for automated knowledge discovery in scientific literature. This innovative aspect of their work is particularly intriguing as it suggests the possibility of leveraging machine learning for more efficient and insightful literature reviews, potentially revolutionizing the way researchers access and digest vast amounts of scholarly information. However, a limitation of their approach was the computational complexity associated with processing extensive datasets, underscoring the need for scalable solutions to accommodate the ever-expanding volume of scientific publications.

Similarly, Shetty and Singh[7] proposed a hierarchical clustering method tailored specifically for scientific literature analysis survey. Their approach showcased efficiency in handling large-scale datasets and provided interpretable clusters representing distinct research themes. Furthermore, their methodology incorporated domain-specific knowledge to enhance the interpretability of clustering results, facilitating deeper insights into emerging research trends and interdisciplinary connections. This integration of domain expertise with computational techniques holds promise for more nuanced analyses and insightful interpretations of scientific literature, potentially paving the way for interdisciplinary collaborations and breakthrough discoveries. Despite

https://doi.org/10.56294/mw202468

its effectiveness, their methodology lacked robustness when dealing with noisy or unstructured text data, highlighting the ongoing challenge of data preprocessing in scientific document analysis.

In addition, recent advancements in machine learning techniques, such as deep learning-based clustering algorithms proposed by Rezaul et al.[8], have garnered significant attention in the field. Their approach, which integrates neural network architectures with unsupervised learning, offers enhanced capability in capturing complex patterns within scientific document datasets. Moreover, the interpretability of their model's latent representations opens avenues for novel discoveries and hypothesis generation in scientific research. The potential of deep learning methods to uncover hidden patterns and relationships in scientific literature is particularly captivating, offering new opportunities for knowledge synthesis and discovery. However, challenges remain in interpreting the black-box nature of deep learning models and ensuring reproducibility across different domains, emphasizing the importance of transparent and interpretable clustering methodologies.

Overall, while these works represent significant advancements in the domain of scientific document clustering, each approach has its respective advantages and limitations. By positioning our research within the context of these state-of-the-art methodologies, we aim to build upon existing foundations and contribute novel insights that address current challenges in scientific document analysis. Through the integration of innovative clustering techniques and rigorous evaluation methodologies, our research endeavors to provide actionable insights for researchers, academic institutions, and decision-makers, ultimately enhancing the accessibility and usability of scientific knowledge on a global scale.

METHOD

Proposed Methodological Blueprint for Advanced Scientific Document Clustering

The research methodology presented in this study delineates a comprehensive framework for exploring scientific documents through segmentation using unsupervised learning techniques. This innovative approach integrates state-of-the-art unsupervised learning algorithms with thorough data preprocessing and evaluation steps to ensure the reliability and robustness of the segmentation process. By leveraging advanced machine learning techniques, our methodology provides a unique perspective on analyzing scientific literature. Furthermore, the proposed methodology, illustrated in Figure 1, offers a visual representation of the sequential steps involved in the segmentation process. This visual aid clarifies the intricate interplay between data collection, preprocessing, algorithmic execution, and evaluation, thereby enhancing the clarity and comprehensibility of our approach.

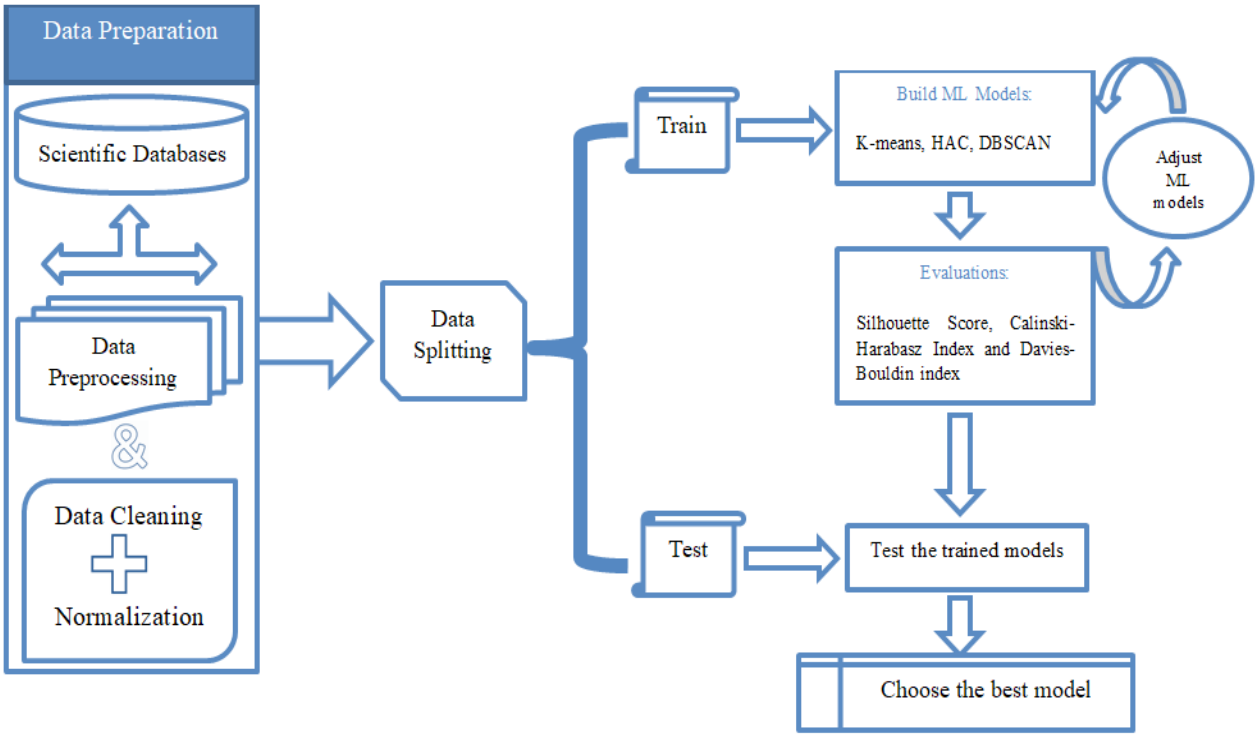

Figure 1. Flowchart of the proposed methodologies.

The proposed blueprint figure encapsulates a series of carefully crafted steps aimed at unraveling the complex landscape of scientific literature through segmentation via unsupervised learning techniques. Each step in this methodological process is designed to complement the preceding one, promoting a comprehensive

approach towards document analysis. The following details the several steps:

- **Data Collection**: Embarking on our exploration, we strategically harness scientific databases as the cornerstone of our research endeavor. These databases serve as central hubs where researchers disseminate their scientific articles and findings, offering a vast repository of knowledge across diverse scientific domains. Through meticulous techniques such as web scraping and crawling, we gather a rich and varied collection of scientific documents from esteemed databases like Scopus or Web of Science. This deliberate approach ensures the acquisition of a comprehensive dataset essential for our subsequent in-depth analysis. By leveraging cutting-edge methods, we curate a dataset that encapsulates the breadth and depth of scientific literature, laying a robust foundation for our research pursuits.
- **Data Preparation**: Following data collection, the obtained documents undergo preprocessing steps to ensure uniformity and consistency. This involves a series of essential data preparation operations, including stemming, lemmatization, and the removal of stopwords and irrelevant characters. Additionally, text normalization techniques are applied to standardize the textual data, ensuring coherence across the dataset. Furthermore, the documents are tokenized to break them down into individual words or phrases, facilitating subsequent analysis. Moreover, techniques such as part-of-speech tagging and named entity recognition may be employed to identify and categorize specific linguistic elements within the text. Through these significant data preparation operations, we enhance the quality and usability of the dataset, laying a solid foundation for effective analysis and interpretation.
- **Split Data for Training and Testing**: Splitting the dataset into distinct training and testing subsets is a critical step in our methodology, enabling robust model development and evaluation. This partitioning ensures that the machine learning model is trained on a subset of the data, allowing it to learn from patterns and relationships present within the dataset. Subsequently, the model is tested on unseen data from the testing set to evaluate its ability to generalize to new, unseen instances. This rigorous evaluation process assesses the model's performance under real-world conditions, providing insights into its effectiveness and generalization capabilities. By employing this systematic approach to data splitting, we ensure the reliability and robustness of our machine learning models, setting the stage for accurate and meaningful analysis of scientific documents.
- **Build Unsupervised Learning Algorithms**: Building upon the preprocessed dataset, we proceed to develop unsupervised learning models, leveraging a diverse array of cutting-edge algorithms tailored to the task at hand. Among these, prominent methods such as k-means clustering, hierarchical clustering, and DBSCAN are prominently featured. These algorithms autonomously traverse the data landscape, discerning underlying patterns and structures, thereby facilitating the segmentation of scientific documents into cohesive groups. Additionally, advanced techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) may be employed to further refine the representation of the data, enhancing the efficacy of the segmentation process. Through the judicious selection and fine-tuning of these algorithms, we aim to unlock latent insights buried within the vast expanse of scientific literature, paving the way for nuanced analysis and interpretation.
- **Evaluation of Trained Model**: Evaluating the trained model is a pivotal phase in our methodology, where the performance of each unsupervised learning algorithm is thoroughly examined using a range of relevant metrics. These metrics, including the silhouette score, Davies-Bouldin index, and completeness score, serve as objective measures to gauge the effectiveness and suitability of each algorithm for the task at hand. The evaluation process goes beyond mere accuracy, providing nuanced insights into the algorithms' ability to capture meaningful patterns and structures within the data. By leveraging these comprehensive metrics, we gain a deeper understanding of the strengths and limitations of each algorithm, enabling informed decisions regarding their selection and optimization. This rigorous evaluation ensures the robustness and reliability of our segmentation approach, laying the groundwork for meaningful analysis and interpretation of scientific documents.
- **Adjust the Model**: Refining the model is a pivotal step in our iterative process, where we carefully analyze the evaluation results to make necessary adjustments. Drawing insights from the evaluation metrics, we fine-tune the parameters and configurations of the unsupervised learning algorithms. This iterative refinement aims to optimize the model's performance and augment its capability to accurately segment scientific documents. By iteratively adjusting the model based on evaluation feedback, we strive to achieve the highest levels of precision and effectiveness in document segmentation, ensuring that our approach remains adaptive and responsive to the nuances of the dataset.
- **Choose the Best Model**: In the culmination of our methodology, we carefully assess the performance of each model based on the evaluation metrics to discern the optimal solution. The model that emerges as the top performer, as determined by its superior performance across the evaluation criteria, is deemed the most suitable segmentation framework. This selection process ensures that we leverage the most effective and reliable model to delineate scientific documents into coherent groups. As the cornerstone

https://doi.org/10.56294/mw202468

of our research methodology, this chosen model lays the foundation for extracting nuanced insights from the segmented scientific literature. It represents the culmination of our rigorous approach and embodies our commitment to delivering robust and meaningful results.

By adhering to this carefully crafted research methodology, we aim to provide a robust framework for the exploration and segmentation of scientific documents using unsupervised learning techniques. Through each step of the process, we strive to achieve our overarching goal of uncovering hidden patterns and trends within the vast landscape of scientific literature.

Exploration of Unsupervised Machine Learning Techniques

Our research delves into the exploration of unsupervised machine learning techniques, where we investigate innovative methods to uncover patterns and structures within complex datasets. Through significant experimentation and analysis, we aim to harness the power of unsupervised learning algorithms such as (a) k-means clustering, (b) Hierarchical Agglomerative clustering, and (c) DBSCAN to segment and categorize scientific documents effectively. By leveraging these advanced techniques, we strive to unravel hidden insights and relationships within vast repositories of scientific literature, ultimately advancing our understanding of complex phenomena and facilitating informed decision-making in various domains.

The K-Means Clustering Approach

In the realm of our research proposal, situated within the domain of Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques, the K-means clustering approach assumes paramount importance. This algorithm serves as a cornerstone in our endeavor to dissect and comprehend the intricate fabric of scientific literature.[9] By employing K-means clustering, we aim to disentangle the dense web of information present in scientific documents, facilitating their segmentation into coherent clusters. This segmentation enables us to unveil latent patterns, themes, and relationships embedded within the corpus, thereby advancing our understanding of the underlying knowledge domain. Figure 2 depicts a blueprint illustrating the sequential steps involved in implementing the K-means clustering approach, providing a visual representation of our methodology and underscoring its pivotal role in achieving our research objectives.

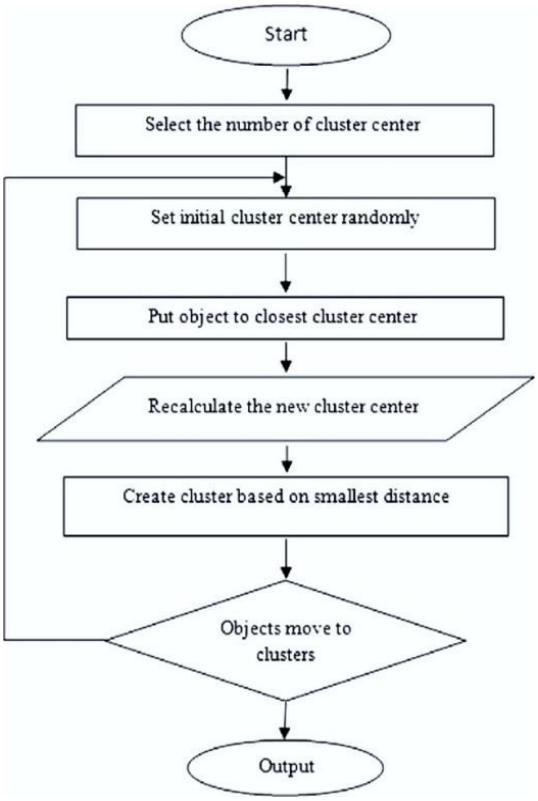

Figure 2. Flowchart of k-means clustering Algorithm.

Hierarchical Agglomerative Clustering Approach

Within the framework of our research proposals, entrenched in the Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques, the hierarchical agglomerative clustering

method emerges as a key component. Unlike K-means clustering, hierarchical agglomerative clustering adopts a bottom-up approach, iteratively merging similar documents into increasingly larger clusters.[10] This hierarchical structure provides a more nuanced understanding of the relationships between documents, allowing for the identification of nested themes and subtopics within the scientific literature. By leveraging hierarchical agglomerative clustering, we seek to unravel the complex interplay of concepts and ideas present in scientific documents, facilitating a comprehensive exploration of the knowledge landscape. Figure 3 illustrates a blueprint delineating the sequential steps involved in implementing hierarchical agglomerative clustering, offering a visual representation of our methodology and underscoring its significance in advancing our research objectives.

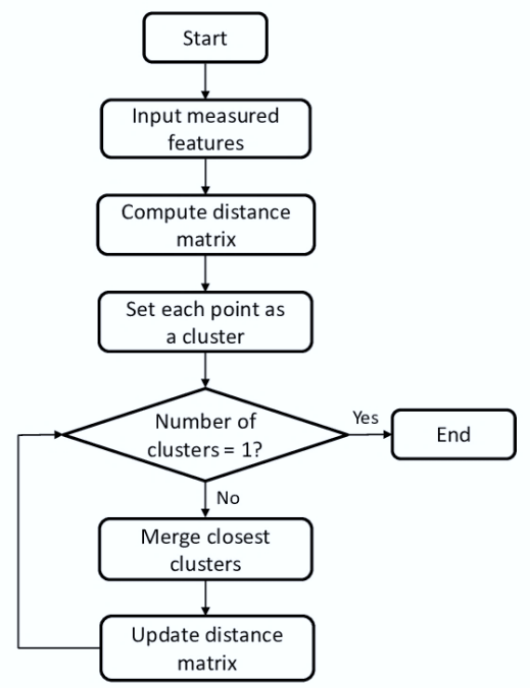

Figure 3. Flowchart of HAC Algorithm.

DBSCAN clustering Approach

In the realm of our research proposals, positioned within the domain of Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm assumes a pivotal role. Unlike K-means clustering and Hierarchical Agglomerative Clustering (HAC), DBSCAN identifies clusters based on the density of data points within a specified radius.

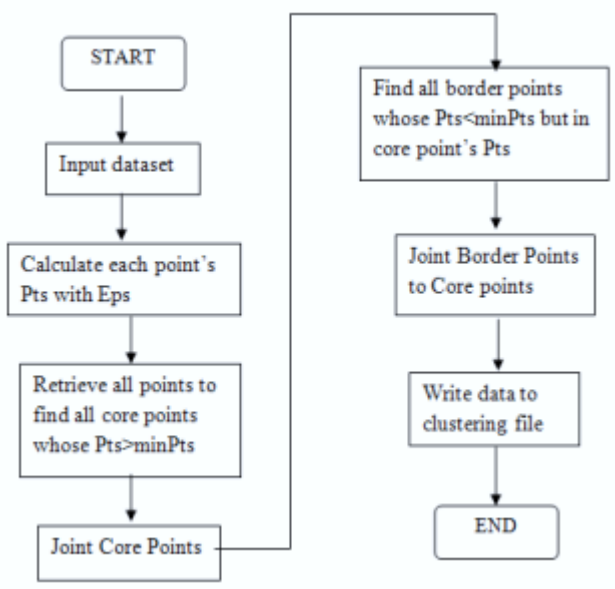

Figure 4. Flowchart of DBSCAN Algorithm.

This enables the detection of arbitrary-shaped clusters and effectively handles noise and outliers present in the dataset.[11] By leveraging DBSCAN, we aim to uncover intricate patterns and structures within scientific documents, offering a nuanced understanding of the underlying themes and relationships. Figure 4 depicts a blueprint outlining the sequential steps involved in implementing DBSCAN, providing a visual representation of our methodology and emphasizing its significance in advancing our research objectives.

RESULTS AND DISCUSSIONS

In this pivotal section, we delve into the outcomes yielded by the segmentation of our scientific document's dataset. Employing three distinct clustering algorithms – K-means, Hierarchical Agglomerative Clustering (HAC), and DBSCAN – we aimed to harness their unique advantages in exploring diverse facets of our data. Through rigorous analysis, we unravel insights and patterns embedded within the dataset, providing a comprehensive understanding of its underlying structure. Figure 5 serves as a visual representation, illustrating the clusters generated by each algorithm and highlighting their distinctive characteristics. Notably, the utilization of K-means and HAC revealed the presence of three discernible clusters, each delineating distinct trends and patterns within the data. Conversely, DBSCAN identified two distinct clusters, shedding light on different aspects of the dataset. Through significant examination and comparison of these results, we examine the effectiveness of each algorithm in uncovering hidden structures within our dataset, paving the way for meaningful discussions and implications.

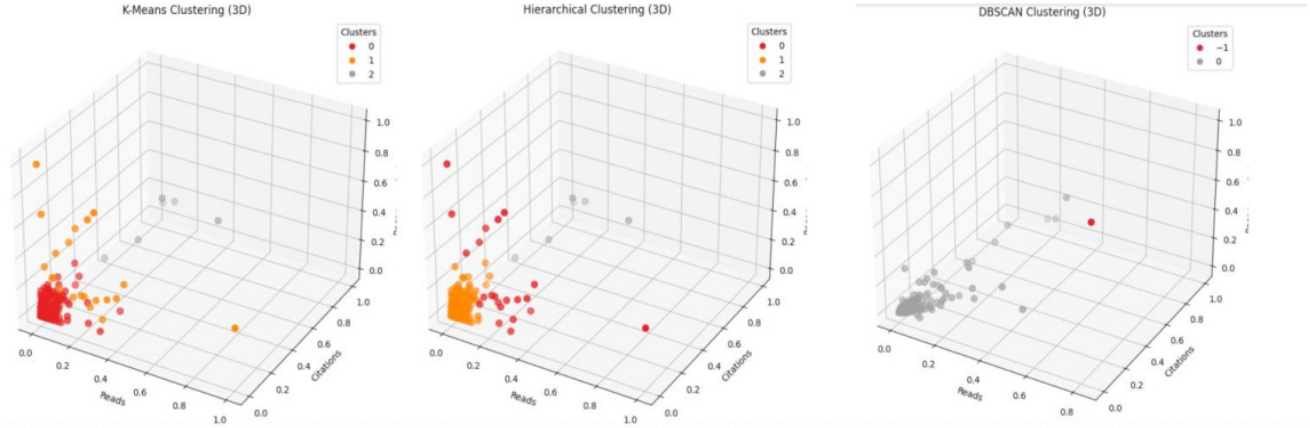

Figure 5. Cluster visualization using different algorithms.

To gain a comprehensive understanding of the formed clusters, we will explore each one's profile. Cluster 0 provides insights into machine learning, demonstrating a notably positive sentiment and encompassing discussions on algorithms, applications, predictions, data analysis, and challenges. Drawing from a diverse array of international publications with varied authorship, the research within this cluster garners significant interest, as evidenced by high citation rates and recommendations. Moreover, it fosters active engagement from readers, underscoring its significance and impact in the field. In Cluster 1, we encounter a neutral sentiment, offering a diverse range of topics and perspectives spanning education, numerical impact, and machine learning applications. Touching upon models, data analysis, and advancements in quantum materials, this cluster features authors from diverse backgrounds and locations, contributing to a rich tapestry of ideas. Despite its lower level of research interest and citation rates, the cluster fosters moderate reader engagement, representing an area ripe for exploration. Lastly, Cluster 2 delves into exploring AI techniques and opportunities, marked by a moderately positive sentiment. It encompasses discussions on AI models, serverless approaches, and business intelligence applications, predominantly authored by individuals with publications originating from varied sources. While exhibiting moderate research interest and citation rates, the cluster garners higher reader engagement due to its promising opportunities. Despite the absence of a DOI and featuring varied publication locations, including unknown ones, its content holds significant potential for driving innovation and exploration in the field.

In examining the profiles of the clusters derived primarily from scientific factors such as citations, recommendations, and reads, we gain valuable insights into the characteristics and significance of each cluster. Cluster 2, characterized by high citation rates and recommendations, represents a group of articles that have garnered considerable attention and acclaim within the scientific community. These articles likely contribute significantly to their respective fields and are deemed influential by peers. Cluster 0, with moderate citation rates and recommendations, indicates a level of interest and relevance but may not have achieved the same level of recognition as Cluster 2. Finally, Cluster 1, with lower citation rates, suggests articles that may be relatively less prominent or impactful within their fields. Table 1 provides a detailed breakdown of these clusters

based on various scientific factors, offering a comprehensive overview of their profiles and implications.

Table 1. Profiling of Clusters Based on Scientific Factors

Cluster	Citations	Recommendations	Reads
0	0,026316	0,337662	0,266017
1	0,006259	0,025277	0,018998
2	0,684211	0,266667	0,220733

In our evaluation of the clusters produced by the algorithms under scrutiny, we employed a comprehensive array of widely recognized performance metrics. These metrics provided invaluable insights into the quality and coherence of the clusters, enabling us to gauge the efficacy of each algorithm in segmenting the scientific documents dataset. Key among the metrics utilized were the Silhouette Score, Calinski-Harabasz Index and Davies-Bouldin index, each offering distinct perspectives on clustering performance. Turning our attention to specific performance metrics, the Silhouette Score served as a barometer of cluster quality: higher scores indicated superior cluster formation. Both DBSCAN and K-Means demonstrated proficiency in creating meaningful clusters, while Hierarchical Agglomerative Clustering also exhibited commendable performance. The Calinski-Harabasz Score (CH Score) corroborated the separation between clusters, with K-Means achieving the highest score, indicative of distinct cluster formation. Hierarchical Agglomerative Clustering also showcased robust performance in this regard. The Davies-Bouldin Score (DB Score) scrutinized inter-cluster dissimilarity, with DBSCAN securing the lowest score, suggesting the formation of relatively unique clusters. K-Means and Hierarchical Agglomerative Clustering similarly demonstrated strong performance. Therefore, while DBSCAN may be preferred for clear and unique clusters, K-Means remains a viable option for achieving satisfactory results with simpler methods. Through a significant analysis of these performance measures, we gained a comprehensive understanding of the strengths and limitations of each clustering algorithm in our study. Notably, Table 2 presents a detailed overview of the metrics assessment, providing a comprehensive summary of the clustering performance based on various evaluation criteria.

Table 2. Comparative Analysis of Clustering Performance Metrics

Metrics assessments	K-means	HAC	DBSCAN
Silhouette Score	0,83954762	0,83558884	0,9104094
Calinski-Harabasz Index	403,4611394	375,476685	80,5134969
Davies-Bouldin Index	0,856120275	0,89936581	0,06095107

CONCLUSION

In conclusion, our investigation showcases the remarkable potential of unsupervised learning methodologies in dissecting the intricate landscape of scientific literature. Through the adept application of clustering algorithms such as K-means, Hierarchical Agglomerative Clustering, and DBSCAN, we have adeptly categorized scholarly documents into cohesive clusters, unraveling nuanced thematic threads within diverse research domains. Our meticulous scrutiny, bolstered by an array of comprehensive evaluation metrics, offers profound insights into the efficacy and nuances of each algorithm. Looking forward, the horizon holds promise for further exploration, potentially integrating advanced features and cutting-edge machine learning models to refine the precision and granularity of document segmentation. Moreover, the envisaged development of interactive visualization tools promises to revolutionize scholarly exploration, offering intuitive pathways to decipher clustered documents and unravel emerging research vistas. In essence, our study contributes significantly to the evolving discourse on unsupervised learning-driven analyses in scientific literature, heralding a new era of knowledge discovery and scholarly discourse.

BIBLIOGRAPHIC REFERENCES

1. Afzali, M., & Kumar, S. (2019). Text Document Clustering : Issues and Challenges. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 263-268. https://doi.org/10.1109/COMITCon.2019.8862247

2. Cozzolino, I., & Ferraro, M. (2022). Document clustering. Wiley Interdisciplinary Reviews: Computational Statistics, 14. https://doi.org/10.1002/wics.1588

3. Mishra, S., Saini, N., Saha, S., & Bhattacharyya, P. (2022). Scientific document summarization in multi-objective clustering framework. Applied Intelligence, 52, 1-24. https://doi.org/10.1007/s10489-021-02376-5

4. Jalal, A., & Ali, B. (2021). Text documents clustering using data mining techniques. International Journal of Electrical and Computer Engineering, 11, 664-670. https://doi.org/10.11591/ijece.v11i1.pp664-670

5. Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences, 9. https://doi.org/10.1186/s13673-019-0192-7

6. Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. Scientometrics, 111(2), 1017-1031.

7. Shetty, P., & Singh, S. (2021). Hierarchical Clustering : A Survey. International Journal of Applied Research, 7, 178-181. https://doi.org/10.22271/allresearch.2021.v7.i4c.8484

8. Karim, R., Beyan, O., Zappa, A., Costa, I., Rebholz-Schuhman, D., Cochez, M., & Decker, S. (2020). Deep learning-based clustering approaches for bioinformatics. Briefings in bioinformatics, 22. https://doi.org/10.1093/bib/bbz170

9. Ikotun, A., Ezugwu, A., Abualigah, L., Abuhaija, B., & Heming, J. (2022). K-means Clustering Algorithms : A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. Information Sciences, 622. https://doi.org/10.1016/j.ins.2022.11.139

10. Lubis, R., Huang, J.-P., Wang, P.-C., Khoifin, K., Elvina, Y., & Kusumaningtyas, D. (2023). Agglomerative Hierarchical Clustering (AHC) Method for Data Mining Sales Product Clustering. Building of Informatics, Technology and Science (BITS), Volume 5, 285-294. https://doi.org/10.47065/bits.v5i1.3569

11. Bushra, A., & Yi, G. (2021). Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. IEEE Access, PP, 1-1. https://doi.org/10.1109/ACCESS.2021.3089036

FUNDING

None

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Mohamed Cherradi, Anass El Haddadi.
Data curation: Mohamed Cherradi, Anass El Haddadi.
Formal analysis: Mohamed Cherradi, Anass El Haddadi.
Fund acquisition: Mohamed Cherradi, Anass El Haddadi.
Research: Mohamed Cherradi, Anass El Haddadi.
Methodology: Mohamed Cherradi, Anass El Haddadi.
Project administration: Mohamed Cherradi, Anass El Haddadi.
Resources: Mohamed Cherradi, Anass El Haddadi.
Software: Mohamed Cherradi, Anass El Haddadi.
Supervision: Mohamed Cherradi, Anass El Haddadi.
Validation: Mohamed Cherradi, Anass El Haddadi.
Display: Mohamed Cherradi, Anass El Haddadi.
Writing – original draft: Mohamed Cherradi, Anass El Haddadi.
Writing – review and editing: Mohamed Cherradi, Anass El Haddadi.