A survey on document clustering and tagging system

[1]Monika Yadav, [2]Dr Nayan S Patel

[1]Assistant Professor, [2]Associate Professor
[1]BCA, BBA-ITM, PGDCA & M.Sc(Data Science) Department,
[1]C P Patel & F H Shah Commerce College (Autonomous), Anand, India
[1]yadavmonika2314@gmail.com , [2]nspatel99@gmail.com

***Abstract*— A document clustering and tagging system serves as an intelligent solution for managing extensive collections of text by automatically grouping related documents and assigning them suitable tags. The system employs various machine learning strategies, including hierarchical and k-means clustering, along with deep learning models, to evaluate and compare the semantic content of documents. Techniques from natural language processing are used to identify main topics and create descriptive metadata tags, which support efficient organization and quick retrieval of information. By implementing this approach, organizations can greatly improve the process of finding information, streamline the management of knowledge, and enhance decision-making by making large datasets more accessible and easier to search.**

***Index Terms*— Document clustering, Tagging system, Machine learning, Information retrieval.**

I. INTRODUCTION

The rapid expansion of digital content has necessitated efficient techniques for organizing and retrieving information from large document repositories. Traditional manual categorization is both time-consuming and impractical for handling vast amounts of textual data. A document clustering and tagging system offers an automated solution by leveraging machine learning and Natural Language Processing (NLP) to streamline information retrieval and knowledge management. Document clustering groups similar documents based on their semantic relationships using techniques such as hierarchical clustering, k-means clustering, and deep learning-based approaches. This process helps in discovering patterns and correlations within large document sets, making it an essential tool in fields like academic research, business intelligence, and legal document analysis.

Complementary to clustering, document tagging assigns metadata labels or keywords to individual documents, enhancing search ability and classification. NLP-based techniques, including named entity recognition (NER), topic modeling, and keyword extraction, facilitate automatic tagging, reducing the need for manual annotation. Together, document clustering and tagging contribute to efficient information retrieval, better knowledge organization, and improved decision-making across various domains, including healthcare, news categorization, and enterprise content management. The integration of these techniques represents a significant advancement in automated text processing, offering a scalable and intelligent approach to managing unstructured textual data.

II. PROPOSED METHODOLOGY

A well-structured document clustering and tagging system integrates preprocessing, feature extraction, clustering techniques, and evaluation metrics. This approach enables meaningful grouping of documents while ensuring effective tagging for better retrieval and analysis.

1. Preprocessing and Text Refinement

Tokenization: Splitting text into discrete words or tokens.
Stop Word Removal: Filtering out frequently occurring but non-informative words like "is," "the," and "a."
Stemming & Lemmatization: Standardizing word forms by reducing them to their root versions, such as "running" to "run."

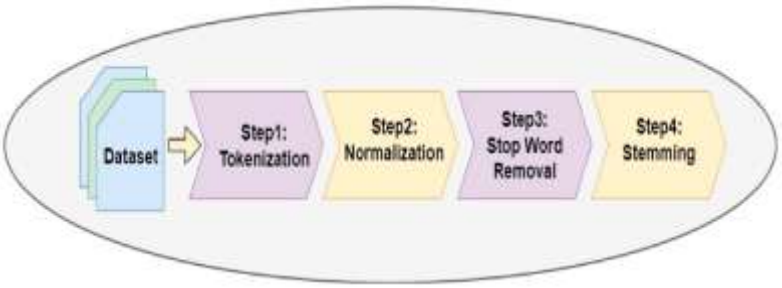

Figure 1.1

2. Feature Extraction for Document Representation

TF-IDF Scoring: Assigning importance to words based on their occurrence within a document and their rarity across the entire dataset.
Word Embeddings _ (Word2Vec, GloVe, BERT): Encoding words as dense vectors in a semantic space to reflect their meaning.
Document-Term Matrix: Structuring documents as matrix rows and terms as columns, where each entry represents term frequency or TF-IDF weight.

3. Clustering Approaches

K-Means Algorithm: Assigning documents to distinct clusters based on their similarity to a central representative.
Hierarchical Clustering: Organizing documents into tree-like structures for scalable exploration at different levels.

DBSCAN (Density-Based Spatial Clustering): Detecting clusters based on regions of high document density, allowing identification of arbitrary-shaped clusters.
Graph-Based Clustering: Mapping documents as nodes in a network and applying graph algorithms to form clusters.

4. Tagging Mechanism

Cluster Labeling: Extracting keywords or representative terms from clusters using TF-IDF or topic modeling techniques.
Automated Tagging: Assigning tags to documents based on their cluster association or content characteristics.

5. Evaluation Metrics for Cluster Validation

Precision: Measuring the proportion of correctly clustered documents within assigned groups.
Recall: Assessing how effectively the system retrieves relevant documents within a cluster.
F1-Score: Balancing precision and recall for a comprehensive clustering performance measure.
Silhouette Score: Evaluating how well documents fit within their clusters compared to others.
Davies-Bouldin Index: Quantifying the separation between clusters to determine clustering effectiveness.

6. Continuous Optimization & User Interaction

User Feedback Loop: Incorporating user insights to enhance clustering accuracy and refine tagging strategies.
Interactive Clustering: Allowing users to explore and fine-tune clusters based on domain-specific needs.
Multiple Perspectives: Offering different views, such as keyword-based and document-based exploration, for improved usability.

7. Key Considerations for System Design

Dataset Characteristics: Selecting algorithms suited to document scale, language, and domain.
Computational Efficiency: Ensuring that chosen methods can handle large-scale datasets effectively.
Scalability: Structuring the system for dynamic expansion as document volumes grow.
Explainability: Providing clarity on why specific documents belong to particular clusters.

By integrating these methodologies and refining processes based on real-world data, an effective document clustering and tagging system can be developed to support applications in fields like information retrieval, knowledge management, and personalized recommendations.

III. Literature Review

1. Hybrid Optimization and Nature-Inspired Methods

Researchers have turned to nature-inspired optimization to overcome the well-known limitations of classical clustering algorithms. For example, Abualigah, L., & Almotairi, K. H. "[1]" proposed an advanced dynamic evolutionary clustering approach by enhancing the Aquila optimizer with operators drawn from arithmetic optimization and differential evolution. Their method not only improved convergence but also produced robust clusters on both standard and text benchmark datasets. Similarly, Bezdan, T., Stoean, C., Naamany, A. al, Bacanin, N., Rashid, T. A., Zivkovic, M., & Venkatachalam, K. "[2]" introduced a hybrid fruit-fly optimization algorithm merged with K-means, capitalizing on swarm intelligence to refine document groups. In yet another study, Malik, F., Khan, S., Rizwan, A., Atteia, G., & Samee, N. A. "[11]" developed a clustering framework based on the Black Hole algorithm that efficiently searches the large solution space of text collections. Further, a comprehensive analysis by Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hussien, A. G., Khasawneh, A. M., Alshinwan, M., & Houssein, E. H. "[17]" reviewed nature-inspired algorithms—including Harmony Search, Genetic Algorithm, PSO, and others—for text document clustering, underscoring the potential of hybrid methods in obtaining high-quality partitions even in high-dimensional settings.

2. Enhanced Partitioning Approaches (K-Means and EM Variants)

Several works focus on improving traditional partition-based methods. Duo, Zhang, and Hao L "[5]" improved the K-means algorithm by incorporating subject feature vectors to refine the initialization of cluster centers, thereby boosting the quality of clusters. Priyanka Dayal "[18]" proposed an "Entropy-k-means" method that uses entropy-regularized terms to automatically select the optimal number of clusters and reduce noise by eliminating irrelevant features. In a similar vein, studies by Naeem and Wumaier "[19]" and by Xiaoli Wang, Ying Li, Meihong Wang, ZiXiang Yang and Huailin Dong "[25]" examine methods for choosing the correct value of k and modifying the iterative centroid update, respectively, to enhance clustering performance. Thangaraj M and Ponmani K "[28]" further advanced the field by designing an expectation maximization–based clustering algorithm tailored for e-content analysis that yields higher accuracy compared to standard EM and K-means approaches.

3. Semantic-Based and Ontology-Driven Clustering

Because text inherently carries meaning beyond simple word frequency, several researchers have turned to semantic information. Cozzolino and Ferraro "[3]" offer a comprehensive review of document clustering that includes discussion of graph-, prototype-, and model-based approaches while emphasizing the need for methods that capture conceptual similarity. Urkude and Pandey "[15]" propose a density-based clustering method that leverages domain ontology and WordNet to weight features; this integration of background knowledge helps reduce dimensionality while preserving semantic relationships. In the educational sphere, Sara Alaee and Fattaneh Taghiyareh "[27]" developed an ontology-based clustering organizer specifically for e-learning documents, thereby ensuring that the clusters carry meaningful labels and capture overlapping topics—which is crucial when documents span multiple subject areas.

4. Deep Learning and Representation Learning Methods

Recent advances in deep learning have also been applied to document clustering. Subakti, A., Murfi, H., & Hariadi, N. "[14]" investigated the role of pre-trained language models such as BERT in transforming text into dense, context-aware embeddings. Their work showed that these representations—when combined with standard clustering algorithms—yield improved performance compared to conventional statistical features like TF-IDF. In another study, Wang, B., Liu, W., Lin, Z., Hu, X., Wei, J., & Liu, C. "[22]" proposed a deep representation learning algorithm that merges neural embedding techniques with traditional clustering, demonstrating how rich semantic representations lead to more coherent clusters in high-dimensional settings.

5. Distributed and Scalable Document Clustering

Given the exponential growth of digital text, scalability is a key concern. Zamora, J., Allende-Cid, H., & Mendoza, M. "[20]" introduced a distributed clustering algorithm that partitions large-scale text collections across multiple nodes; by leveraging distributed processing, their approach maintains high clustering quality while reducing processing time. Neepa Shah, Sunitha Mahajan "[26]"conducted a scalability analysis of a semantic-based distributed document clustering algorithm, revealing that using semantic cues not only improves cluster quality but also enhances the algorithm's scalability in multi-node environments.

6. Additional Approaches and Applications

Other works have targeted specific applications or incorporated additional processing steps. Dubey Shivkishan "[4]" focused on refining word sense disambiguation by applying multilevel clustering techniques, which enable the model to capture nuanced semantic differences that are often lost in conventional clustering. Devi, S. A., & Siva Kumar, S. "[9]" proposed a hybrid framework that couples document feature extraction with clustering-based classification; this framework is particularly useful for managing large document sets where intra-cluster variations are high. Lin, Z., Laska, E., & Siegel, C. "[10]" introduced an iterative clustering algorithm that repeatedly refines a proximity matrix to converge upon more stable clusters. Furthermore, practical applications such as information retrieval have been enhanced by integrating clustering with search optimization techniques, as illustrated by Inje, B., Nagwanshi, K. K., & Rambola, R. K. "[8]".

Summary Table 1.1 for Literature Review

References	Approach / Methodology	Key Contributions	Significance / Remarks
1. Abualigah & Almotairi (2022)	Improved Aquila optimizer combined with Arithmetic Optimization and Differential Evolution	Developed a dynamic evolutionary clustering algorithm that overcomes local optima and improves search balance	Robust performance on both standard data and text document benchmarks
2. Bezdan et al. (2021)	Hybrid Fruit-Fly Optimization integrated with K-means	Uses swarm intelligence to refine cluster partitions and improve text clustering accuracy	Demonstrates improved robustness on benchmark text datasets
3. Cozzolino & Ferraro (2022)	Comprehensive review of document clustering methods	Examines various clustering paradigms (graph-, prototype-, hierarchical) and discusses document representation	Provides guidance for selecting clustering methods in different contexts
4. Dubey (2024)	Multilevel clustering for word sense disambiguation	Introduces a clustering framework that captures semantic subtleties in context, enhancing disambiguation tasks	Improves clarity in semantic interpretation of ambiguous words
5. Duo et al. (2021)	K-means clustering using subject feature vectors	Enhances initial centroid selection using domain-specific features	Yields improved clustering accuracy in text document applications
6. Khazaei et al. (2021)	FOCT (Fast Overlapping Clustering for Textual Data)	Proposes an overlapping clustering algorithm that allows documents to belong to multiple clusters	Efficiently models the inherent overlap in document topics
7. Hannachi et al. (2023)	Online clustering using infinite extensions of discrete mixture models	Adapts mixture models for clustering short and sparse texts in streaming environments	Effectively handles data sparsity and real-time text streams
8. Inje et al. (2024)	Hybrid Global Search Optimization with Density-Based Clustering	Integrates search optimization with density-based clustering to improve document retrieval	Enhances retrieval quality and convergence rates in large datasets
9. Devi & Siva Kumar (2020)	Hybrid feature extraction with clustering-based classification framework	Combines advanced feature extraction (e.g., GloVe-based) with clustering and classification	Designed for large document sets with high inter-/intra-document variations
10. Lin et al. (2022)	General iterative clustering algorithm	Uses an iterative refinement process on a proximity matrix to improve cluster quality	Offers a flexible, convergence-oriented clustering method

References	Approach / Methodology	Key Contributions	Significance / Remarks
11. Malik et al. (2022)	Hybrid clustering based on the Black Hole algorithm	Leverages a nature-inspired optimization method to escape local minima in document clustering	Yields higher precision and effective global search in clustering
12. Muller et al. (2022)	Automated text clustering using Lingo3G compared with human categorization	Compares machine-generated clusters to manual classifications in systematic reviews	Highlights potential for reducing time and enhancing consistency in reviews
13. Ponnusamy et al. (2022)	Salp Swarm Algorithm (SSA) for text document clustering	Applies swarm intelligence inspired by salp behavior to determine optimal clusters	Outperforms traditional methods across multiple evaluation metrics
14. Subakti et al. (2022)	BERT-based text clustering	Evaluates how deep contextual embeddings improve document grouping over TF-IDF models	Demonstrates significant performance gains using modern language models
15. Urkude & Pandey (2022)	Density-based clustering using domain ontology	Integrates ontology and WordNet to weight features and guide clustering	Enhances semantic relevance and clustering quality in domain-specific corpora
16. Rojas-Thomas & Santos (2021)	New internal clustering validation index	Proposes a validation metric based on density uniformity for arbitrary-shape clusters	Enables automatic determination of optimal cluster partitions
17. Abualigah et al. (2020)	Comprehensive review of nature-inspired optimization algorithms	Surveys various NIOAs (HS, GA, PSO, etc.) as applied to text clustering problems	Serves as a resource for selecting suitable optimization methods in clustering
18. Dayal (2019)	Entropy-k-means clustering	Introduces entropy-regularized terms into K-means to enable feature reduction and optimal k detection	Automatically removes irrelevant features and improves clustering robustness
19. Naeem & Wumaier (2018)	K-means clustering with optimal cluster count techniques	Investigates methods to determine the optimal value of k in text clustering	Provides practical techniques for parameter tuning in document clustering
20. Zamora et al. (2019)	Distributed clustering of text collections	Proposes a distributed algorithm to split and cluster large-scale text collections	Enhances scalability when processing massive textual datasets
21. Jalal & Ali (2021)	Data mining techniques for text document clustering	Employs similarity measures and data mining for grouping web documents	Aids in efficient information retrieval and organized web search results
22. Wang et al. (2018)	Deep representation learning for text clustering	Incorporates deep learning to generate rich text embedding for clustering	Outperforms traditional vector space models in capturing semantic relations
23. Yudi Hidayat et al. (2015)	Automatic text summarization using LDA for clustering	Uses topic modeling in conjunction with summarization to enhance clustering outcomes	Shows that LDA can improve document grouping by capturing underlying topics
24. Madaan & Kumar (2018)	Improved approach for web document clustering	Addresses noise and scalability issues in web documents using enhanced clustering techniques	Results in better cluster quality on vast web datasets
25. Xiaoli Wang et al. (2018)	Improved K-means algorithm for document clustering	Modifies K-means—especially in initialization and centroid updates—to boost performance	Achieves higher accuracy and consistency across text datasets
26. Neepa Shah & Sunitha Mahajan (2017)	Scalability analysis of semantic-based distributed clustering	Analyzes the performance and scalability of distributed semantic clustering algorithms	Provides insights into the trade-offs between semantics and processing efficiency
27. Alaee & Taghiyareh (2016)	Semantic ontology-based document organizer for e-learning	Uses domain ontologies to annotate and cluster e-learning documents with overlapping topics	Ensures clusters are meaningfully labelled and reflect the document semantics
28. Thangaraj & Ponmani (2023)	Enhanced Expectation Maximization document clustering	Proposes an EM-based algorithm optimized for e-content analysis that refines clusters iteratively	Yields improved accuracy and efficiency in clustering educational and digital content

IV. Limitations of the System

Scope and Coverage: Incomplete coverage of relevant studies, reliance on older research, limited access to databases due to this impact on research is may exclude important findings or emerging trends.

Bias in Research Selection: Preference for studies with positive results, exclusion of contradictory findings due to this impact on research is leads to an unbalanced perspective.

Quality and Reliability: Differences in methodologies, lack of reproducibility due to this impact on research makes comparison difficult and reduces credibility.

Generalizability Issues: Limited sample sizes, context-specific results due to this impact on research findings may not be widely applicable.

Conceptual and Theoretical Constraints: Variability in definitions, lack of consensus due to this impact on research creates difficulties in synthesizing studies.

Practical Challenges: Time and resource constraints, restricted access to research due to this impact on research limits depth and breadth of the review.

Difficulty in Data Synthesis: Managing large volumes of information, risk of misinterpretation due to this impact on research can lead to oversimplified conclusions.

V. CONCLUSION

Document clustering and tagging systems play a crucial role in organizing, categorizing, and retrieving large volumes of textual data. By utilizing various optimization techniques, machine learning algorithms and semantic analysis, these systems enhance information accessibility and improve efficiency in data management. Recent advancements in nature-inspired optimization algorithms, deep learning models, and ontology-based clustering have significantly improved the accuracy and scalability of these systems. While traditional clustering methods often struggle with high-dimensional data and sparse text representations, hybrid approaches integrate multiple techniques to refine cluster coherence and document classification.

Despite their effectiveness, these systems face challenges such **as** computational complexity, scalability limitations, and the need for accurate semantic interpretation. Addressing these concerns through adaptive clustering methods, real-time processing, and cross-domain integration will further enhance their applicability across various fields, including research, business analytics, and automated information retrieval. In summary, document clustering continues to evolve with the integration of advanced algorithms and AI-driven techniques**,** paving the way for more efficient knowledge organization and improved document management solutions.

REFERENCES

[1] Abualigah, L., Almotairi, K.H. "Dynamic evolutionary data and text document clustering approach using improved Aquila optimizer based arithmetic optimization algorithm and differential evolution. Neural Computing & Applications" vol. 34, pp. 20939–20971 August 2022. (*References*)

[2] Bezdan, T., Stoean, C., Naamany, A. al, Bacanin, N., Rashid, T. A., Zivkovic, M., & Venkatachalam, K. "Hybrid fruit-fly optimization algorithm with k-means for text document clustering. Mathematics", August 2021.

[3] Cozzolino, I., & Ferraro, M. B. "Document clustering. In Wiley Interdisciplinary Reviews: Computational Statistics" vol. 14, Issue 6, John Wiley and Sons Inc. June 2022.

[4] Dubey, Shivkishan "Clustering for Clarity: Improving Word Sense Disambiguation through Multilevel Analysis" vol. 25 No.2 July 2023.

[5] Duo, J., Zhang, P., & Hao, L"A K-means Text Clustering Algorithm Based on Subject Feature Vector. Journal of Web Engineering" vol. 20 Issue October 2021.

[6] Khazaei, A., Khaleghzadeh, H., & Ghasemzadeh, M. "FOCT: Fast Overlapping Clustering for Textual Data". IEEE Access, vol. 9, pp. 157670 - 157680 November 2021.

[7] Hannachi, S., Najar, F., Ennajari, H., & Bouguila, N. "Online short text clustering using infinite extensions of discrete mixture models" Computational Intelligence, vol. 39 Issue 5, pp. 759– 782 October 2023.

[8] Inje, B., Nagwanshi, K. K., & Rambola, R. K. "An efficient document information retrieval using hybrid global search optimization algorithm with density based clustering technique". Cluster Computing, vol. 27 Issue 1, pp. 689–705 February 2023.

[9] Devi, S. A., & Siva Kumar, S. "A Hybrid Document Features Extraction with Clustering based Classification Framework on Large Document Sets" International Journal of Advanced Computer Science and Applications vol. 11, Issue 7 2020.

[10] Lin, Z., Laska, E., & Siegel, C. "A general iterative clustering algorithm. Statistical Analysis and Data Mining" vol. 15 Issue 4, pp. 433–446 August 2022.

[11] Malik, F., Khan, S., Rizwan, A., Atteia, G., & Samee, N. A. "A Novel Hybrid Clustering Approach Based on Black Hole Algorithm for Document Clustering". IEEE Access, vol.10, pp. 97310–97326 August 2022.

[12] Muller, A. E., Ames, H. M. R., Jardim, P. S. J., & Rose, C. J. "Machine learning in systematic reviews: Comparing automated text clustering with Lingo3G and human researcher categorization in a rapid review. Research Synthesis Methods, vol. 13 Issue 2, pp. 229–241 March 2022.

[13] Ponnusamy, M., Bedi, P., Suresh, T., Alagarsamy, A., Manikandan, R., & Yuvaraj, N. "Design and analysis of text document clustering using salp swarm algorithm". Journal of Supercomputing, vol. 78 Issue 14, pp. 16197–16213 May 2022.

[14] Subakti, A., Murfi, H., & Hariadi, N. "The performance of BERT as data representation of text clustering. Journal of Big Data", vol. 9 Issue 1 February 2022.

[15] Urkude, G., & Pandey, M. "Design and Development of Density-Based Effective Document Clustering Method Using Ontology". Multimedia Tools and Applications, vol. 81 Issue 23, pp. 32995–33015 April 2022.

[16] Rojas-Thomas, J. C., & Santos, M. "New internal clustering validation measure for contiguous arbitrary-shape clusters". International Journal of Intelligent Systems, vol. 36 Issue 10, pp. 5506–5529.June 2021.

[17] Abualigah, L., Gandomi, A. H., Elaziz, M. A., Hussien, A. G., Khasawneh, A. M., Alshinwan, M., & Houssein, E. H. "Nature-inspired optimization algorithms for text document clustering—a comprehensive analysis" vol. 13 Issue 12 December 2020.

[18] Dayal, Priyanka "Entropy Reduction using K-Mean Clustering Algorithm". International Journal for Research in Applied Science and Engineering Technology, vol. 7 Issue 4, pp. 1598–1601 April 2019.
[19] Naeem, S., & Wumaier, A. "Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K". In International Journal of Computer Applications vol. 182, Issue 31, 2018.
[20] Zamora, J., Allende-Cid, H., & Mendoza, M. "Distributed Clustering of Text Collections". IEEE Access, vol. 7, pp. 155671–155685 October 2019.
[21] Ahmed Adeeb Jalal, Basheer Husham Ali "Text documents clustering using data mining techniques". International Journal of Electrical and Computer Engineering, vol. 11 Issue 1, pp. 664–670, 2021.
[22] Wang, B., Liu, W., Lin, Z., Hu, X., Wei, J., & Liu, C. "Text clustering algorithm based on deep representation learning". The Journal of Engineering, vol. 2018 Issue 16, pp. 1407–1414, November 2018.
[23] Yudi Hidayat, E., Firdausillah, F., Hastuti, K., & Novita Dewi, I. "Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering". International Journal of Advances in Intelligent Informatics, vol. 1 Issue 3, pp. 132–139, 2015.
[24] Vaishali Madaan, Rakesh Kumar "An Improved Approach for Web Document Clustering". International Conference on Advanced Computing, Communication Control and Networking, 2019.
[25] Xiaoli Wang, Ying Li, Meihong Wang, ZiXiang Yang and Huailin Dong. "An Improved K_means Algorithm for Document". International Congress on Image and Signal Processing, Biomedical Engineering and Informatics, 2019.
[26] Neepa Shah, Sunitha Mahajan "Scalability Analysis of Semantic Based Distributed Document Clustering Algorithm". International Conference on Intelligent Computing, Instrumentation and Control Technologies, 2017.
[27] Sara Alaee and Fattaneh Taghiyareh "A Semantic Ontology Based Document Clustering Organizer to Cluster eLearning Document". International Conference on Web Research, 2019.
[28] Thangaraj M and Ponmani K "An Enhanced Expectation Maximization Text Document Clustering Algorithm for E-Content Analysis". International Journal on Recent and Innovation Trends in Computing and Communication, 2023.