Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2019, 6(2):66-73

Research Article ISSN: 2394 - 658X

Discovering Hidden Themes by Enhancing Document Cluster Interpretability

Akshata Upadhye

Data Scientist

ABSTRACT

In this era of digital transformation, as the digital information continues to grow it presents a formidable challenge in organizing and understanding documents. Most of the clustering algorithms efficiently group similar documents, yet they have limitations such as revealing the hidden themes that define each cluster that hinder effective comprehension. Motivated by this gap that lies between clustering and understanding the clustering, this research addresses this gap by leveraging Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) representation within document clusters. This novel approach enhances interpretability by extracting distinctive themes from the cohesive document clusters thereby providing nuanced insights into key terms shaping each cluster's content. Additionally this research work not only bridges the gap between cluster formation and interpretability but also enabled generating actionable insights in large document collection. This transformative approach advances the potential of document clustering for a more interpretable and insightful analysis of textual content.

Key words: Index Terms—Document clustering, Latent themes, Interpretability, TF-IDF analysis, Text document analysis, Digital information

INTRODUCTION

With the adaptation of digital transformation, the digital repositories continue to expand. Due to growth of digital information the task of understanding and organizing documents has become increasingly challenging. Clustering which is a method of grouping similar documents, offers a solution to this challenge by allowing the identification of inherent structures and hidden themes within a large corpus. While these traditional clustering algorithms produce well-formed clusters, a persistent challenge lies in uncovering the underlying themes that define each cluster. This remains as a significant barrier to the effective understanding of document clusters.

The motivation for this research emerges from the inherent limitations of existing clustering algorithms in providing interpretable insights into the themes present within each cluster. While clustering techniques are efficient in organizing documents based on similarity metrics, the lack of a interpretability of the type of documents within these clusters hampers the utility of the analysis. This critical gap might have practical implications across various domains such as information retrieval, content summarization or knowledge discovery, where a deeper understanding of cluster themes is essential.

To address this challenge, our work focuses on harnessing the power of Term Frequency (TF) and Term FrequencyInverse Document Frequency (TF-IDF) within the context of document clustering. By integrating these methodologies, we aim to extract and articulate the distinctive themes that characterize each of the well-defined document clusters. This novel approach not only enhances the interpretability of document clusters but it also helps in providing a nuanced understanding of the key terms that contribute to the documents within each cluster.

Through an exploration of TF and TF-IDF within the text document collection, this research seeks to bridge the gap between cluster formation and interpretability, offering a comprehensive solution to the challenges posed by traditional clustering algorithms. By focusing on the themes within clusters, our approach aims to unlock new possibilities for extracting actionable insights from large document collections, thereby advancing the state of the art in text document analysis.

RELATED WORK

In the extensive field of text document analysis, clustering algorithms have been instrumental in revealing latent structures within large corpora. However, a persistent challenge acknowledged in the literature is the limited interpretability of themes within the clusters generated by these algorithms. This section reviews existing research, focusing on the interpretability aspect of document clustering methodologies and the motivation for adopting innovative approaches.

In this paper [1] the authors conduct an experimental study comparing common document clustering techniques, with a focus on the comparison between agglomerative hierarchical clustering and K-means. The study includes a comparison of both the standard K-means algorithm and a variant known as “bisecting” K-means. The paper emphasizes that while hierarchical clustering is often considered superior in terms of clustering quality, it has quadratic time complexity which poses limitations when working with larger datasets. On the other hand, K-means and its variants, with linear time complexity, are generally known to produce less satisfactory clusters. However, the results in the study reveal that the bisecting K-means technique outperforms the standard K-means and is comparable or superior to hierarchical clustering approaches across various cluster evaluation metrics. The authors present an insightful explanation for these results, by focusing in on a detailed analysis of the clustering algorithms and the inherent characteristics of document data.

In this research paper [2], the authors present a detailed overview of document clustering for automatic organization of documents to create cohesive clusters. The study emphasizes the importance of document clustering across various domains such as web mining, search engines, and information retrieval. The authors cover a spectrum of clustering methods, ranging from traditional approaches to more recent developments such as fuzzy-based, genetic, co-clustering, and heuristic-oriented methods. The document outlines the document clustering procedure, including feature selection processes, similarity measures, and evaluation criteria for clustering algorithms. Furthermore, the paper dives deep into various improvements, including TF-IDF processes and techniques for dimensionality reduction. This exhaustive survey spans the last fifteen years of research and aims to serve as a valuable resource for researchers in the field of document clustering and provides insights as to how the advancements have contributed in unsupervised document organization and information retrieval.

This article [3] provides a review of recent research focusing on document retrieval using agglomerative hierarchical clustering methods. The paper begins with an introducing the computation of inter-document similarity measures and moves on to discussing suitable clustering methods for document organization and implementation of these methods on sizable databases are discussed. The validation of document hierarchies is explored using tests based on the theory of random graphs and empirical characteristics of document collections. Further the article presents an evaluation of a range of search strategies for information retrieval using document hierarchies and discusses results from various research projects employing different types of hierarchical clustering methods. Of all the different type of linkage criterion, the complete linkage method is identified as the most effective in retrieval performance, despite the challenges in its efficient implementation with larger data. Finally the article briefly discusses other applications of document clustering techniques. It also highlights experimental evidence suggesting that nearest neighbor clustering approach which uses a network representation of documents also offers an efficient means of incorporating inter-document similarity information into document retrieval systems.

This paper [4] introduces a novel document clustering approach which utilizes the non-negative factorization of the term-document matrix within a given corpus. By using the non-negative matrix factorization (NMF) method the proposed approach transforms the document space into a latent semantic space where each axis captures the foundational topic of a specific document cluster. Documents are then represented as additive combinations of these base topics, facilitating the straightforward determination of cluster membership based on the axis with the highest projection value. Through experimental evaluations, the proposed clustering method outperforms latent semantic indexing and spectral clustering not only in the ease and reliability of deriving clustering results but also in terms of document clustering accuracy. This innovation presents a promising advancement in document clustering techniques.

This study in paper [5] addresses the increase in the volume of text documents due to the recent advancements in computer technology, emphasizing the imperative need to classify these documents by type for efficient decision-making. The paper suggests the classification of related documents is particularly crucial for researchers who engage in interdisciplinary studies, as they have to uncover insights from research documents covering diverse topics. The authors conduct experiments on real and artificial datasets, including NEWS 20, Reuters, emails, and research papers, using the Term Frequency-Inverse Document Frequency in conjunction with fuzzy K-means and hierarchical clustering algorithms. Initial experiments on small datasets involve cluster analysis, and the best-performing algorithm is chosen and applied to an extended dataset. The study presents results in the form of different clusters of related documents, accompanied by the trends in silhouette coefficient, entropy, and F-measure to showcase the performance of each algorithm for every dataset. This research contributes valuable insights into the application and performance of clustering algorithms in document classification scenarios.

BACKGROUND

A. Clustering

1) Document representation: The effectiveness of document clustering largely depends on the quality of document representation. Therefore document representation techniques used to transform raw textual data into a format suitable for clustering plays a vital role in overall document clustering process. Some of the techniques are utilized to transform documents and enhance clustering outcomes are as follows:

- Bag-of-Words (BoW): BoW [6] model is used to represent a text document as unordered collections of words by disregarding the word order. It creates a vector space model based on the frequency of terms, thus providing a basic but widely used representation. Term Frequency (TF) representation is one of the types of BoW model.
- Term Frequency-Inverse Document Frequency (TFIDF): TF-IDF model [7] is a BoW based representation with assignment of weights to terms based on their frequency within a document and their rarity across the entire corpus. This technique helps mitigate the impact of frequency of common words while giving more weightage to the distinctive terms.
- Word Embeddings: Word embeddings, such as Word2Vec [8] and GloVe [9], capture semantic relationships between words by representing them as dense vectors in a continuous vector space. Document vectors are derived by aggregating word vectors, capturing contextual information.
- Doc2Vec (Paragraph Vectors): Doc2Vec [10] is an extension of word embeddings used to represent entire documents as vectors. Each document is assigned a unique vector, capturing its semantic meaning and allowing for similarity comparisons.

It is important to select the most appropriate document representation technique based on the nature of the data and the desired level of semantic understanding since it impacts the clusters formed.

2) Document clustering: Document clustering techniques are often used to organize large document collections into meaningful groups, allowing for efficient information retrieval and analysis. Several methods have been developed to find natural clusters within the textual data such as:

- K-Means Clustering: K-Means [11] clustering partitions the documents into k clusters based on similarity and is one of the most widely used methods. It relies on an iterative process to optimize cluster centroids, making it computationally efficient.
- Hierarchical Clustering: This clustering approach [12] builds a tree-like hierarchy of clusters, either top-down also known as divisive or bottom-up also known as agglomerative. Agglomerative hierarchical clustering is commonly used in document clustering, where clusters are successively merged based on similarity.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN clustering [13] method identifies clusters by examining the density of data points. It is effective in discovering clusters of varying shapes and sizes and is robust to outliers in the data.

The choice of a clustering technique depends on the nature of the data, the desired cluster structure, and computational considerations. Document clustering research has various ongoing efforts to develop methods that can handle the complexities of large and diverse document collections.

B. Cluster Analysis and evaluation

Effective evaluation of the quality of document clusters is crucial for assessing the performance of clustering algorithms. Various cluster analysis and evaluation techniques have been developed to test the efficiency of clustering results such as:

- Silhouette Analysis: Silhouette analysis gauges the cohesion and separation of clusters. The silhouette coefficient measures how well-separated clusters are and it ranges from -1 to 1, with higher values indicating better-defined clusters.
- Rand Index and Adjusted Rand Index: The Rand Index is used to assess the similarity between true and predicted clusters, providing a measure of overall clustering accuracy. The Adjusted Rand Index improves over the Rand Index by adjusting for chance, yielding a normalized metric.
- Normalized Mutual Information (NMI): NMI measures the mutual information between true and predicted clusters, normalized to account for the imbalance due to cluster size. It ranges from 0 to 1, with higher values indicating better agreement.
- Purity: Purity is used to assess the homogeneity of clusters by measuring the proportion of correctly assigned documents within each cluster. It provides a measure of clustering accuracy but may not account for overlaps.

Picking an appropriate metric for evaluation depends on the goals and characteristics of the clustering task. Combining multiple metrics often provides a better understanding of clustering performance.

METHODOLOGY

A. Data Preprocessing

This step involves cleaning and preparing the text documents from the collection by implementing the following steps:

- Cleaning and Tokenization: The 20 Newsgroups dataset is initially cleaned for any special charecters and tokenized, breaking down each document into individual words or tokens.
- Stop Word Removal: The common stop words are removed from the tokenized documents in order to eliminate frequently occurring but less informative words.
- Lemmatization: Each remaining token is lemmatized to reduce words to their root form to ensure consistency in representation.
- N-gram Incorporation: Finally, frequently occurring ngrams are identified and added to the list of tokens to enhance the vocabulary.

B. Document Representation

This step involves training the Doc2Vec model and using it to extract embedding representation for the cleaned texts in the collection.

- Doc2Vec Embedding: A Doc2Vec model is trained to generate a dense 20-dimensional vector representation for each document thereby capturing semantic relationships between words and their context.

C. Clustering

• K-Means Clustering: The Doc2Vec representations are then subjected to K-means clustering to group similar documents together. The number of clusters is determined based on the various evaluation criteria.

D. Evaluation

• Rand Index: The Rand Index is calculated to evaluate the similarity between the true and predicted clusters, providing an overall measure of clustering accuracy.

E. Theme Identification

- Term Frequency (TF) Analysis: For this analysis the sum and mean TF scores are calculated across all documents. The top 20 terms with the highest mean TF scores are identified which are used to offer insights into the most representative terms within the clusters.
- TF-IDF Analysis: For this analysis the sum and mean TF-IDF scores are computed across all documents. The top 20 terms with the highest mean TF-IDF scores are determined to reveal the most distinctive terms that contribute to the cluster separation.

F. Results Visualization

- Results Visualization: The rand index scores for n clusters formed using Doc2vec representation of text and K-means clustering algorithm as presented using line chart.

This proposed methodology combines advanced text processing techniques, embedding methodologies, and clustering algorithms to reveal inherent structures within the 20 Newsgroups dataset. The evaluation metrics and theme identification further enhance the interpretability of the clustering results.

DATASETS

The 20 Newsgroups dataset is a widely used collection of approximately 20,000 newsgroup documents spanning 20 distinct categories, each representing a specific topic such as computers, politics, sports, science, and more. This dataset has originated from Usenet newsgroups in the late 1990s. This dataset serves as a benchmark for text classification and clustering tasks. In our approach for clustering and cluster theme analysis, we leverage the 20 Newsgroups dataset to evaluate the effectiveness of our proposed methodology. By generating advanced Doc2vec text representation techniques combined with K-means clustering, we aim to uncover latent structures within the diverse set of documents. Additionally, our analysis dives into extracting meaningful themes within the clusters by providing insights into the underlying topics by identifying the top words thus enhancing the interpretability of the document clusters.

RESULTS AND INTERPRETATION

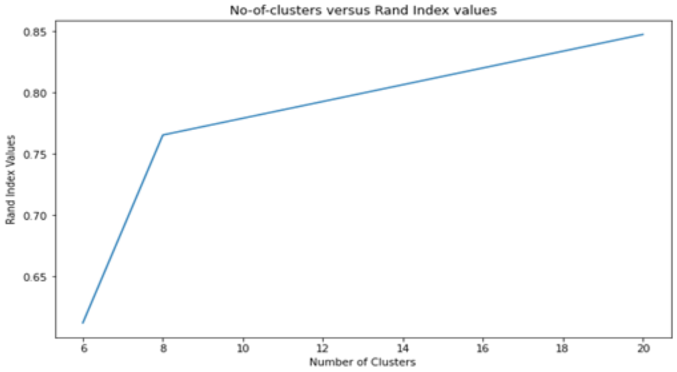

Figure 1: Rand Index line chart for n clusters

A. Clustering Performance

The K-means clustering algorithm is able to effectively grouped documents within the 20 Newsgroups dataset based on their semantic similarities due to a quality text representation generated using the Doc2Vec embeddings. The comparison of Rand Index metric as shown in figure 1 demonstrates that 20 clusters have the best accuracy result of 0.84 for the clustering when compared to 6 and 8 clusters. Therefore a few among the 20 clusters will be further analyzed to determine themes and topics within the cluster to demonstrate the effectiveness of our approach.

B. Key Themes within Clusters

Table 1: Itop 20 Tf and Tf-Idf Terms In Cluster 1 (Sum And Mean)

Top 20 TF Terms (Sum)	Top 20 IDF(Sum)	TF Terms	Top 20 TF Terms (Mean)	Top20 TF-IDFTerms (Mean)
would	team		would	team
go	would		go	would
think	player		think	player
get	game		get	game
say	go		say	go
good	think		good	think
team	get		team	get
make	say		make	say

write	good	write	good
game	make	game	make
see	well	see	well
well	see	well	see
subject	car	subject	car
know	play	know	play
player	people	player	people
people	know	people	know
time	write	time	write
may	year	may	year
take	time	take	time
play	may	play	may

Table 2: Top 20 Tf and Tf-Idf Terms in Cluster 3 (Sum and Mean)

Top 20 TF Terms (Sum)	**Top 20TF IDF Terms (Sum)**	**Top 20 TF Terms (Mean)**	**Top 20TF-IDF Terms (Mean)**
file	use	file	use
use	file	use	file
user	key	user	key
internet	ripem	internet	ripem
key	user	key	user
information	message	information	message
message	anonymous	message	anonymous
may	internet	may	internet
anonymous	privacy	anonymous	privacy
privacy	mail	privacy	mail
mail	information	mail	information
network	rsa	network	rsa
service	may	service	may
system	cipher	system	cipher
site	standard	site	standard
email	email	email	email
see	service	see	service
people	encrypt	people	encrypt
server	anonymity	server	anonymity
post	network	post	network

Top Terms via TF Analysis: Utilizing Term Frequency (TF) analysis, the most representative terms within each cluster were identified. By observing trends in Tables 1-3, the top 20 terms with the highest sum of TF represented in column 1 and mean TF represented in column 3 across all documents provided insights into prevalent themes within clusters 1, 3and 4.

Table 3: Top 20 Tf and Tf-Idf Terms In Cluster 4 (Sum And Mean)

Top 20 TF Terms (Sum)	**Top 20 TF IDF Terms (Sum)**	**Top20 TF Terms (Mean)**	**Top 20 TF IDF Terms (Mean)**
people	armenian	people	armenian
would	people	would	people
say	militia	say	militia
state	would	state	would
armenian	say	armenian	say

right	turkish	right	turkish
make	state	make	state
also	government	also	government
use	right	use	right
take	kill	take	kill
government	village	government	village
turkish	attack	turkish	attack
see	use	see	use
even	gun	even	gun
time	see	time	see
may	take	may	take
think	make	think	make
kill	think	kill	think
many	may	many	may
give	also	give	also

Top Terms via TF-IDF Analysis: Employing TF-IDF analysis enhanced the understanding of distinctive terms contributing to cluster separation. By observing trends in Tables 1-3, the top 20 terms with the highest sum of TF-IDF values represented in column 2 and mean TF-IDF values represented in column 4 across all documents provided insights into prevalent themes within clusters 1, 3 and 4.

C. Observations

By looking at table 1, the significant terms contributing to the cluster 1 have been obtained via TF (sum and mean of TF) in columns 1 and 3 respectively. By looking at this terms the we can say that the cluster contains documents belonging to the sports category. Similarly, the significant terms contributing to the cluster 1 have been obtained via TF-IDF (sum and mean of TF-IDF) in columns 2 and 4 respectively. Similar to the trends depicted by TF, by looking at the top TF-IDF terms the it can be inferred that the cluster contains documents belonging to the sports category.

Next, by looking at table 2, the significant terms contributing to the cluster 3 have been obtained via TF (sum and mean of TF) in columns 1 and 3 respectively. By looking at this terms the we can say that the cluster contains documents belonging to the computers, security, networks and information systems category. Similarly the significant terms contributing to the cluster 3 have been obtained via TF-IDF (sum and mean of TF-IDF) in columns 2 and 4 respectively. Similar to the trends depicted by TF, by looking at the top TF-IDF terms it can be inferred that the cluster contains documents belonging to the the computers, security, networks and information systems category. An additional information to note here is that while this approach is useful to identify themes within clusters, it an also be used for evaluation. For instance, looking at the insights for cluster 3, it can be further be broken down into individual clusters for separating the documents within the computers, security, networks and information systems categories.

Finally, by looking at table 3, the significant terms contributing to the cluster 4 have been obtained via TF (sum and mean of TF) in columns 1 and 3 respectively. By looking at these terms the we can say that the cluster contains documents belonging to the politics category. Similarly, the significant terms contributing to the cluster 1 have been obtained via TF-IDF (sum and mean of TF-IDF) in columns 2 and 4 respectively. Similar to the trends depicted by TF, by looking at the top TF-IDF terms the it can be inferred that the cluster contains documents belonging to the politics category.

Through the evaluation and analysis of results the combination of advanced text processing, embedding, and clustering techniques can help in providing an in-depth understanding of the latent structures within the 20 Newsgroups dataset. The revealed themes within clusters showcased the effectiveness of the proposed methodology in uncovering meaningful content relationships. Despite the good results it is essential to acknowledge potential limitations, such as the sensitivity of clustering results to the choice of parameters in the K-means algorithm. Additionally, the interpretability of themes depends on the quality of the underlying data and the effectiveness of lemmatization and stop words removal.

CONCLUSION

In conclusion, this study has leveraged Doc2Vec embeddings and K-means clustering to form clusters within the 20-newsgroup dataset. The clustering performance, as measured by the Rand Index, demonstrates the high accuracy of the proposed approach in grouping documents based on semantic similarities. Further the clusters were analyzed by using the top terms obtained by using both Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). This proposed approach for understanding the themes or topics within the clusters have provided valuable insights into the content of each cluster. The use of these techniques has allowed for the extraction of meaningful and representative terms which contributed to a nuanced understanding of the latent structures within the 20 Newsgroups dataset. While recognizing the success of the current approach, it is essential to acknowledge certain limitations such as the sensitivity of clustering results to algorithm parameters and the interpretability of themes based on data quality are important aspects to address in future research. In essence, the outcomes of this study underscore the efficacy of the proposed methodology in uncovering and interpreting themes within document clusters thus providing a solid foundation for further advancements in the field of text analysis.

REFERENCES

[1] Steinbach, Michael, George Karypis, and Vipin Kumar. "A comparison of document clustering techniques." (2000).

[2] Shah, Neepa, and Sunita Mahajan. "Document clustering: a detailed review." International Journal of Applied Information Systems 4, no. 5 (2012): 30-38.

[3] Willett, Peter. "Recent trends in hierarchic document clustering: a critical review." Information processing and management 24, no. 5 (1988): 577597.

[4] Xu, Wei, Xin Liu, and Yihong Gong. "Document clustering based on non-negative matrix factorization." In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 267-273. 2003.

[5] Bafna, Prafulla, Dhanya Pramod, and Anagha Vaidya. "Document clustering: TF-IDF approach." In 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 6166. IEEE, 2016.

[6] Harris, Zellig S., and Zellig S. Harris. Distributional structure. Springer Netherlands, 1970.

[7] Rajaraman, Jure Leskovec Anand, and Jeffrey D. Ullman. "Data mining (recommended subject) textbook."

[8] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).

[9] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[10] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." In International conference on machine learning, pp. 1188-1196. PMLR, 2014.

[11] MacQueen, James. "Some methods for classification and analysis of multivariate observations." In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297. 1967.

[12] Bridges Jr, Cecil C. "Hierarchical cluster analysis." Psychological reports 18, no. 3 (1966): 851-854.

[13] Ester, Martin, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu.¨ "A density-based algorithm for discovering clusters in large spatial databases with noise." In kdd, vol. 96, no. 34, pp. 226-231. 1996.