A Short Review on News Articles Clustering

A. V. Rama Krishna Reddy[1], T. Murali Krishna[2], J. Mohan Kumar[3]
[1,2,,3]Asst. Professor, Ashoka Women's Engineering College, Kurnool

Article Info

Page Number: 167 - 172

Publication Issue:

Vol 71 No. 1 (2022)

Article History

Article Received*: 02 February 2022*

Revised*: 10 March 2022*

Accepted*: 25 March 2022*

Publication*: 15 April 2022*

Abstract

Data mining is a notion that has attracted a great deal of interest from a wide range of individuals and organizations as a result of the continuously increasing number of online users and the development of technology to link and exchange huge amounts of data and information. Text mining, a subfield of data mining, has become one of the most important research fields in recent years due to the rapid rise of textual data, which has reached petabytes. A considerable share of the total number of internet users find daily news stories to be of interest. Daily news data is provided by organizations such as Google and Yahoo, among others. In addition, people want to be informed of the most recent news, resulting in a massive amount of textual news data being released every minute by numerous news websites on the internet. In order to make the retrieval of huge data sets easier and more relevant, the administration and categorization of these enormous quantities of data are required. This study provides an overview of the studies on the clustering of news articles.

Index Terms: Text Mining, Document Clustering, News Articles.

1. Introduction

As a result of the evolution of the internet's underlying technology and the simplicity with which it can be accessed by humans, the bulk of our requirements are now satisfied by the world wide web. In addition, the introduction of digitization has accelerated our progress, resulting in the generation of a substantial amount of data. Every day, around 2.5 billion terabytes of data are produced chaology and the simplicity with which it can be accessed by humans, the bulk of our requirements are now satisfied by the world wide web. In addition, the introduction of digitization has accelerated our progress, resulting in the generation of a substantial amount of data. Every day, around 2.5 billion terabytes of data are produced. [1]It is difficult for businesses to organize and handle the massive amounts of data they already possess. Businesses and other organizations must invest a substantial amount of time and effort in order to retrieve this data. A rising number of users, including enterprises, organizations, and non-governmental organizations, are engaging in continuous data processing. These users are looking for methods to use this information to make decisions and solve problems in their lives. The source provides both a structured and an unstructured version of the information.
A recent study indicated that 80% of the data on the planet is unstructured [2]. In contrast, organizations and analysts are unable to derive insights from and exploit the available data in its entirety. This unstructured data can be created in numerous formats, such as blogs, social networking sites, news feeds, emails, papers, log files, etc. Obtaining any sort of knowledge from data presented in an unstructured style is a difficult endeavor. Data derived from media articles is one sort of data that is generated in large quantities by a variety of people who make news stories. Depending on the issues that excite their interest, users anticipate that the information supplied to them will be accurate and current. In this work, we therefore provide a review of numerous

approaches and a study for clustering news articles. Due to the fact that news article data is one of the most reliable sources of textual data, we believed it was essential to incorporate this information. Clustering, a form of unsupervised learning, is one of the text mining techniques that can be implemented. This procedure can be performed alone or it can be incorporated as one of the phases of a bigger algorithm. A collection of things or points will be divided into clusters using this procedure. This is one of the techniques used for text mining, which is the process of extracting information from diverse written sources. Among its many practical, real-world, and real-time uses are fraud detection, medical science, biotechnology, engineering and technology, science, business, database management, software solutions, search engines, recommendation systems, and information retrieval systems.

Clustering methods can be broken down into the following broad categories:

- ✓ Hierarchical and agglomerative clustering
- ✓ Partition clustering
- ✓ Hard clustering
- ✓ Soft clustering
- ✓ Density-based clustering
- ✓ Hybrid clustering

In order to recognize document clustering, either a similarity measure or a distance measure is utilized. The conceptual structure for grouping news stories is presented in Figure 1. There are a number of different approaches to taking similarity measurements. The phases that are most important for clustering textual articles are as follows:

- ✓ Fetch data
- ✓ Preprocessing
- ✓ Feature extraction
- ✓ Similarity measure or distance measure
- ✓ Main clustering algorithm
- ✓ Final clusters formed
- ✓ Performance analysis.

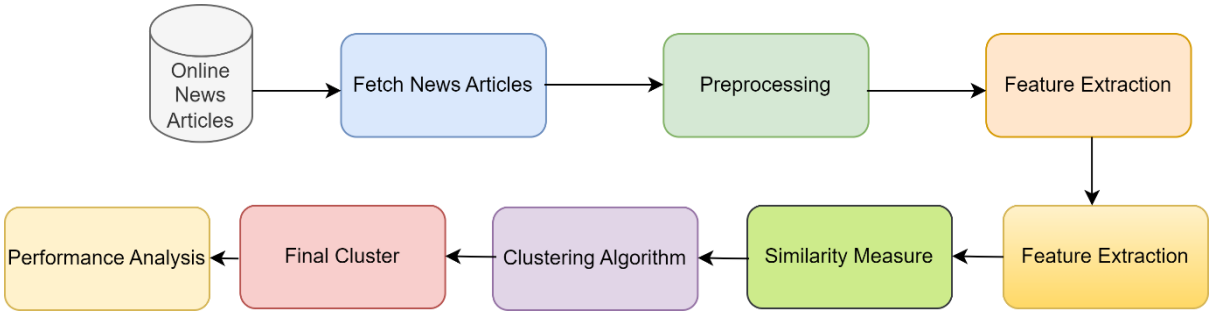

Fig. 1: News Clustering Architecture

2. Literature Survey

Alan F. Smeaton and his co-workers utilized 15,836 news articles, the traditional cosine similarity between the document and the cluster prototype vector, and a bag-of-words representation to assess the degree of similarity between the two groups. For the purpose of retrospective detection, hierarchical clustering defined by group average cluster (GAC) was used. Hierarchical GAC was

proven to perform better than partitioning when compared to this method. The purpose of the comparison was to determine which method was superior. It has been established that non-clustering approaches are preferable in terms of on-line event detection precision [3]. In their research, Yimi Yang et al. employ both agglomerative and hierarchical clustering. Statistics and observations on clustering were compared for three distinct datasets: 34,768 news items from the Irish Times in 1993 (about 100 MB), 48,050 news items from the Irish Times in 1996–7 (approximately 130 MB), and 3050 websites from Dublin City University (nearly 20 MB). They presented a comprehensive method for link clustering that, when implemented, created small, densely populated clusters. They tried to improve the dynamic grouping of news pieces on the internet. The following graphs illustrate the link between the total number of clusters and the average size of each cluster for each of the three collections [4].The authors, Hiroyuki Toda et al., define clustering as the process of labeling a group of items that have been grouped together. They have presented a clustering technique based on the extraction of NE for news articles. They have proposed not just a criterion for label selection but also a mechanism for label placement. During testing, it was established that the suggested approaches were more advantageous than the ones currently in use. Their analysis was limited to Japanese newspaper articles; nevertheless, they determined that their method is language-independent and has the potential to be applied to other languages as well. [5].Gianna M. Del Corso and colleagues suggested a model in which multiple news article excerpts and news sources are connected via a virtual connection. Their experimental settings were based on news data received by comeToMyHead, a search engine that gathered news items from over 2000 news sources over the course of two months (from 7/07/04 to 10/11/04), for a total of over 300,000 news items. Lexical similarity was leveraged [6] to develop a similarity measure for clustering in order to get a ranking for news articles. Marco Aeillo and his colleagues devised three text-processing-based solutions to the newspaper article clustering challenge. These algorithms recognize and group together text chunks that belong to the same article. The conclusions from the clustering were displayed as a link graph. In this comparison, we will examine the three algorithms listed below: A simple agglomerative method has been presented using the bigram indexing algorithm[7].

Using the bigram indexing algorithm and a comparative technique, it was concluded that simple clustering provided the best overall performance due to its simplicity and adaptability [8]. This was due to the fact that the clustering method was shown to have the best overall performance. In order to analyze individuals, regions, and organizations, David Newman et al. researched probabilistic topic modeling. By merging named entity recognizers with topic models, they proved how to study the links between entities (such as persons, organizations, and locations) and themes by analyzing 330,000 news stories from 2000 to 2002. The New York Times and other major and regional newspapers from the United States are included in this collection. Entity recognizers have been included in probabilistic topic modeling. By utilizing topic-model relationships and the bipartite graph to describe the latent structure that exists between entities, a deeper understanding of the latent structure that occurs between things was attained. They realized that statistical language models, such as probabilistic topic models, can aid in the study of enormous text collections[9]. Using the confabulation model, Hiroshi Sekiya and colleagues established a method for automatically generating unit-conceptual fuzzy sets from a collection of 800,000 news articles. Additionally, five statistical association metrics were compared. MI, the Jaccard coefficient, and the chi square were all employed in the process of computing membership values; nonetheless, it was determined that MI was the most effective of the three. [10] According to Maria Soledad Pera and her colleagues, the clustering of RSS news is achieved by utilizing a fuzzy equivalence relation. It

is a method for filtering duplicate news articles from RSS feeds by using word-correlation factors in a fuzzy information retrieval model with max-prod transitivity to separate less-informative articles from non-duplicate ones and clustering the remaining informative articles according to their fuzzy equivalence classes. This is achieved by employing a fuzzy information retrieval model with maximum-probability transitivity. This clustering method can only be used for RSS news articles, although it has the potential to be expanded [11].

FICUS is a clustering and filtering technique introduced by Maria Soledad Pera and her coworkers. Utilizing fuzzy set information retrieval techniques, redundant RSS news articles are identified and eliminated. The subsequent stage is the clustering of the remaining non-redundant RSS news articles according to their degrees of similarity. In FICUS, the clusters are organized in the shape of a tree according to their rank. The contents of the various clusters were gathered using keywords that were reflective of the RSS news articles that comprised each cluster. The FICUS system is easy to use since it searches for relevant articles using a set of predefined word correlation criteria. This technique uses a fuzzy approximation to identify phrases that are related in articles [12]. Milos Krstajic and his coworkers proposed a visual analytics strategy that can be utilized to analyze the connections between news articles. Their methodology makes use of a carrot-inspired clustering module, which is an open-source clustering system. This framework incorporates Lingo and Suffix Tree Clustering (STC) as clustering methods. Both of these clustering methods have the benefit of not needing a fixed number of clusters. They are also very efficient in terms of processing time and computing power [13].

Shilpi Malhotra et al., Shilpi Malhotra et al., Shilpi Malhotra I have described a method for summarizing the data contained in news items and, as an intermediate step, have calculated the similarity between documents and sentences by calculating the frequency of beyond and the length of the sentence [14].

Richard Elling Moe and co-workers did clustering studies using a corpus of Norwegian news stories. According to the outcomes of their study[15], the suffix tree clustering approach was chosen since it outperformed a number of competing algorithms.
Flat clustering is the foundation of a new method developed by M. Uma Devi and colleagues for identifying news article data sets that have sentences that are semantically identical. They utilized fuzzy clustering, which proved to be more effective. They employed the EM framework and the page rank method to identify overlapping clusters based on the semantic relatedness of the data in order to conduct entropy and purity performance analysis. The similarity score improved by 39 percent, according to the data [16]. In this study, N. Dangre and colleagues compare a variety of clustering techniques, including the following: The Marathi news applied the K-means, KNN, and SVM clustering methods, with the SVM algorithm proving to be the most effective of the three. Utilizing a graphical user interface, they supplied graded clusters from a variety of sources [17].

Ilya Blokh and his co-workers suggested a method for the clustering of news that might classify the news into semantically related groups. A number of Facebook experiments were conducted on the volume of news from official news media pages. From January 2014 to May 2017, we were able to acquire 4,150,000 texts. Through the use of WordNet-based semantic similarity, an ontology-based similarity estimation was performed. The results of the experiments show that messages can be grouped by subject, and the data show how news clusters are spread out over time. In the study conducted by Tom Nicholls and colleagues, automated methods were demonstrated for finding related news stories within a corpus of 61,864 documents. The identification of textual closeness between pairs of articles is performed using techniques derived from the field of information

retrieval, while the clustering of these articles is performed using approaches derived from the field of network analysis. The BM25F score algorithm is utilized to find which articles are most closely linked. We were able to show data in a network format and find communities by making the process easier and putting the network nodes into groups [18]. Following a survey of the pertinent literature, we identified numerous applications of article clustering, some of which are shown in Figure 2. Text accessibility can be achieved through the use of content analysis. Two independent systems comprise text access: the recommendation system and the search engine system. These technologies can be used for a variety of purposes, including the clustering of news articles.

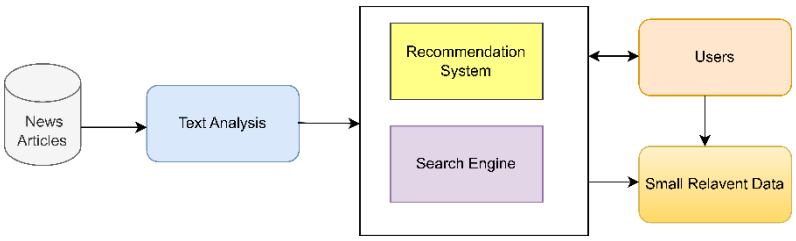

Fig. 2: Text Analysis Architecture.

Text accessibility can be achieved through the use of content analysis. Two independent systems comprise text access: the recommendation system and the search engine system. Using this approach, the clustering of news articles can be used for a variety of purposes.

3. Conclusion

Text mining and document clustering have been research priorities for a considerable amount of time. Numerous authors have undertaken research on and implemented numerous techniques and algorithms. Additionally, each solution is validated with its own unique dataset. As a result, the results produced by different algorithms can vary. In the hierarchical clustering method, the clustering procedure is conducted hierarchically. This method provides a deeper understanding of data patterns than the flat clustering method. The agglomerative clustering approach operates in the opposite way from the hierarchical clustering technique. The partition clustering method may be more practical, but it is entirely dependent on the random generation of cluster centers. K-means clustering is the most straightforward approach for clustering. There are several substantial variants of this method that fall within the hybrid approach category. In comparison to soft clustering, hard clustering has fewer uses. When used in retrieval systems, the soft clustering technique produces more relevant results. The dictionary technique, the WordNet approach, and the co-occurrence relation are all based on the similarity of individual words, phrases, or text documents. Examples of such strategies include: It has both domain ontology and taxonomy as potential applications. There is no single platform capable of grouping the several types of text documents and news article datasets. Moreover, each of these algorithms produces high-quality outcomes. In addition, a variety of clustering systems, including IBM Watson [1], Carrot[13], and many more, are available for use. More research and development is needed in this field, though, to get much faster retrieval, the ability to handle huge amounts of data, scalability, and clustering that happens in real time.

References

[1] “1a23aa8c7651ed7ef8d2e5295bf1e7ef596010be @ www.ibm.com.” .

[2] “1bbe3bfd7f4d284c284c7b21586e1dda475d0193 @ www.netapp.com.” .

[3] A. F. Smeaton, M. Burnett, F. Crimmins, and G. Quinn, "An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts," in *20th Annual BCS-IRSG Colloquium on IR 20*, 1998, pp. 1–9.

[4] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 28–36.

[5] H. Toda and R. Kataoka, "A clustering method for news articles retrieval system," in *Special interest tracks and posters of the 14th international conference on World Wide Web*, 2005, pp. 988–989.

[6] G. M. Del Corso, A. Gulli, and F. Romani, "Ranking a stream of news," in *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 97–106.

[7] T. A. S. Srinivas, Y. R. Mohan, R. Varaprasad, G. Mahalaxmi, Y. Sravanthi, and I. Priyanka, "A Survey of Clustering Methods for Health Care Using Data Mining," vol. 4150, no. 5, pp. 100–104, 2022.

[8] M. Aiello and A. Pegoretti, "Textual article clustering in newspaper pages," *Appl. Artif. Intell.*, vol. 20, no. 9, pp. 767–796, 2006.

[9] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, "Analyzing entities and topics in news articles using statistical topic models," in *International conference on intelligence and security informatics*, 2006, pp. 93–104.

[10] H. Sekiya, T. Kondo, M. Hashimoto, and T. Takagi, "Context representation using word sequences extracted from a news corpus," *Int. J. Approx. Reason.*, vol. 45, no. 3, pp. 424–438, 2007.

[11] M. S. Pera and Y.-K. Ng, "Utilizing phrase-similarity measures for detecting and clustering informative RSS news articles," *Integr. Comput. Aided. Eng.*, vol. 15, no. 4, pp. 331–350, 2008.

[12] M. S. Pera and Y.-K. D. Ng, "Using maximal spanning trees and word similarity to generate hierarchical clusters of non-redundant RSS news articles," *J. Intell. Inf. Syst.*, vol. 39, no. 2, pp. 513–534, 2012.

[13] M. Krstajić, M. Najm-Araghi, F. Mansmann, and D. A. Keim, "Story tracker: Incremental visual text analytics of news story development," *Inf. Vis.*, vol. 12, no. 3–4, pp. 308–323, 2013.

[14] S. Malhotra and A. Dixit, "An effective approach for news article summarization," *Int. J. Comput. Appl.*, vol. 76, no. 16, 2013.

[15] R. E. Moe, "Clustering in a news corpus," in *International Conference on Text, Speech, and Dialogue*, 2014, pp. 301–307.

[16] M. U. Devi and G. M. Gandhi, "An enhanced fuzzy clustering and expectation maximization framework based matching semantically similar sentences," *Procedia Comput. Sci.*, vol. 57, pp. 1149–1159, 2015.

[17] N. Dangre, A. Bodke, A. Date, S. Rungta, and S. S. Pathak, "System for Marathi news clustering," *Procedia Comput. Sci.*, vol. 92, pp. 18–22, 2016.

[18] T. Nicholls and J. Bright, "Understanding news story chains using information retrieval and network clustering techniques," *Commun. Methods Meas.*, vol. 13, no. 1, pp. 43–59, 2019.