# NEWS ARTICLES CLUSTERING: A REVIEW

[1]Priyanka, [2]Sanjay Joshi, [3]Dr.H.L.Mandoria, [4]B.K.Pandey
[1]Research Scholar, [2]Assisstant Professor, [3]Head and Professor, [4]Assisstant Professor
[1]Department of Information Technology,
[1]College Of Technology, GBPUA&T, Pantnagar, Uttarakhand, India

***Abstract :*** With the ongoing growth of users and advancement in technology to communicate and to share large volumes of data & information on the web, data mining have been widely researched topic by different groups and organizations. Text mining which is a subclass of data mining is one of the important research areas since textual data is increasing rapidly to petabytes. A large portion of total number of users on internet is interested in daily news articles. So producers of news data, such as Google, Yahoo etc produce news data on daily basis. Also users want to be updated of recent news, so every minute vast amount of textual news data is generated by different news portals on web. Therefore the management and organization of these large amounts of data is important so that their retrieval becomes more easy and relevant. This paper presents a review on the study of clustering of news articles.

***Index Terms* - Clustering, News Articles, Text Mining, Document Clustering, Unstructured data.**

## I. INTRODUCTION

The evolution in internet technology and its easy availability to mankind has brought us to depend on world wide web for most of our requirements. Also digitization has helped us evolve and therefore a lot of data is being created. Nearly 2.5 billion GB data is generated each day. It is a struggle for companies to manage and organize this much large volumes of data[1]. This data is ever increasing so its retrieval is also a very complex and tedious task for companies and organizations. Increasing users such as business companies, organizations, NGOs etc are constantly working on data. These users are searching for ways to utilize this data to make decisions and to get solutions to their problems. The data is present in two forms: structured form and unstructured form. A recent study has estimated that 80% of the total data in this world is in an unstructured form[2]. But companies and analyzers are unable to gain insights and are unable to make utilization of the entire potion of available data. This unstructured data is generated in different forms such as blogs, social media sites, news feeds, emails, documents, log files etc. Because of its unstructured form, it is a very difficult task to gain insights from such data.

News articles data is one such form of data which is produced in a huge amount by different news story producers. Since users want to be up to the date with the accurate and current news on the topic of their interest. So in this paper, we are presenting a review of various techniques and study for news articles' clustering since news articles data is one of the appropriate sources of textual data.

Clustering is an unsupervised learning process which comes under text mining techniques. This process is used either as standalone or as an intermediary step in other algorithms. In this process, a set of entities or points is categorized into clusters. This process is used in text mining to get insights into textual data. It has many useful, real-life and real-time applications such as fraud detection, medical science, biotechnology, engineering and technology, science, business, database management, software solutions, search engines, recommendation systems, information retrieval systems etc.

The different techniques of clustering are categorized broadly into the following :

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

Document clustering is identified on the basis of similarity measure or distance measure. The architecture for clustering news articles is shown in figure 1. There are various approaches to find the similarity measures. The main steps involved in the clustering of textual articles are :

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

- Performace analysis

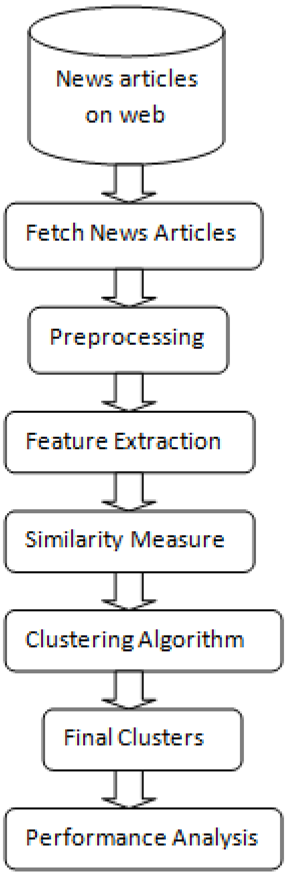

Figure 1: Architecture for clustering of news articles

## II. RELATED WORK

Alan F. Smeaton *et.al.*, used 15,836 news stories, standard cosine similarity between document and cluster prototype vector to measure their similarity and bag-of-words for representation. Group Average Cluster (GAC)-based hierarchical clustering was used for retrospective detection. This method was compared with the partition approach and concluded that hierarchical GAC performed better. For on-line event detection,\ non-clustering techniques have demonstrated better detection accuracy[3].

Yimi Yang *et.al.*, agglomerative & hierarchical clustering is used. Comparision of statistics and some observations on the clustering for three collections of datasets : 34,768 Irish Times news articles from 1993 (nearly 100Mb), 48,050 Irish Times news articles from 1996-7 (nearly 130Mb), 3050 Web pages from Dublin City University (nearly 20Mb). They have shown a complete link clustering method which produced small, tight clusters. They worked on dynamic clustering of online news articles. The graphs below show the number of clusters against the size of the cluster for each of the three collections[4].

Hiroyuki Toda *et.al.*, the authors regard the clustering as the tasks of marking labels for a list of objects and have focused on news articles and proposed a clustering method that uses NE extraction. They have proposed a label selection criterion and a label organization method. Evaluations indicated that the proposed methods were more useful than the existing methods. Their evaluation used only Japanese newspaper articles but concluded that their method is not language specific and so it could be used to handle other languages[5].

Gianna M. Del Corso *et.al.*, proposed a model in which the exploitation of a virtual linking relationship between different pieces of news stories and news sources. Their experimental settings were based on the news data collected by comeToMyHead, a search engine to gather a collection of news articles in two months (from 8/07/04 to 10/11/04) by more than 2000 news sources, and consist of about 300,000 pieces of news. The lexical similarity was used to find similarity measure for clustering with the objective to rank news articles[6].

Marco Aeillo *et.al.*, proposed three text processing-based algorithms for the problem of article clustering in newspaper pages, that is, the identification of text blocks which belong to the same article. The results were shown using a connection graph for clustering. Three different algorithms are compared: an agglomerative technique with bigram indexing algorithm, simple technique with bigram indexing algorithm and comparative technique with bigram indexing and proved that simple clustering gave overall best performance because of low complexity and more flexibility[7].

David Newman *et.al.*, authors have researched in probabilistic topic modeling to analyze persons, places, and organizations. By the combination of named entity recognizers with topic models, they have illustrated how to analyze the relationships between entities (persons, organizations, places) and topics, using news articles from 2000 through 2002, resulting in 330,000 separate articles. These include articles from the NY Times, regional US and other urban newspapers. The probabilistic topic modeling is combined with entity recognizers. The topic-model relationships were used and representation is done using a bipartite graph which resulted into the better understanding of the latent structure between entities. They have concluded that Statistical language models, such as probabilistic topic models, can play an important role in the analysis of large sets of text documents. Representations based on probabilistic topics go beyond clustering models because they allow the expression of multiple topics per document[8]

Hiroshi Sekiya *et.al.,* presented a method to generate unit conceptual fuzzy sets automatically using confabulation model on a collection of 800,000 news articles. And five statistical measures for relations were compared. MI, Jaccard coefficient and chi square are used to calculate membership values and concluded Mutual information was most effective[9]

Maria Soledad Pera *et.al.,* clustering of RSS news is done using fuzzy equivalence relation. It is the approach in which the use of the word-correlation factors in a fuzzy information retrieval model with max-prod transitivity to filter duplicate news articles from RSS feeds shed less-informative articles from the non-duplicate ones and clusters the remaining informative articles in accordance to the fuzzy equivalence classes on the news articles. This clustering approach applies only to RSS news articles but some extension can be done[10].

Maria Soledad Pera *et.al*, presented a clustering and filtering approach, called FICUS, which starts with identification and elimination of redundant RSS news articles using a fuzzy set information retrieval methods and then clustering the remaining non-redundant RSS news articles according to their degrees of resemblance. FICUS uses a tree hierarchy for the organization of clusters. The contents of the respective clusters were captured by the representative keywords from RSS news articles in the clusters. FICUS is simple, as it uses the pre-defined word correlation factors to find related articles. This method uses a fuzzy approximation to identify related words in articles[11].

Milos Krstajic *et.al.,* presented a visual analytics system for exploration of news stories and their relationships. Their framework uses a clustering module based on Carrot, which is an open source framework for clustering. This framework provides two clustering algorithms: Lingo and Suffix Tree Clustering (STC) algorithm whose important advantage is that they don’t require a predefined number of clusters and are very efficient in terms of processing time and computing power[12]

Shilpi Malhotra *et. al.,* Have reported a method for the summarization of news articles data and have calculated the similarity between documents and sentences as an intermediary step by calculating by the frequency of beyond and length of the sentences[13].

Richard Elling Moe *et. al.,* Have used a Norwegian news article corpus for the clustering investigation. They chose to apply suffix tree clustering algorithm because their review shows it outperforms a number of other algorithms[14].

M. Uma Devi *et.al.,* have presented a new method to find semantically similar sentence via flat clustering on news article data sets. They have used an enhanced algorithm of fuzzy clustering. They have used page Rank algorithm EM framework to determine the overlapping clusters on semantically relatedness for performance analysis is done in terms of entropy and purity. Results show 39% higher performance in similarity scoring[15].

N. Dangre *et.al.,* in this paper comparison of various clustering methods: K-means, KNN and SVM is done and presented the best among them for Marathi news clustering. They presented ranked clusters from multiple sources to the user interface[16].

Ilya Blokh *et.al.,* have proposed an algorithm for news clustering that could be able to group news into semantically close sets. Experiments were made on news volumes from several news mass media official pages in Facebook. Retrieved 4,15,000 messages for the period from January 2014 up to May 2017. An ontology-based similarity estimation was performed using semantic similarity based on WordNet. Experimental results show that messages can be grouped into thematic clusters and news cluster distribution over time was obtained[17].

Tom Nicholls *et.al.,* in this work authors presented automated techniques for identification of linked news stories from a corpus of 61,864 articles. This method uses techniques drawn from the field of information retrieval for identification of textual closeness of pairs of articles and then clustering techniques are taken from the domain of network analysis to group these articles. For similarity measure, the BM25F scoring algorithm is used to find related articles. The network-based representation of data and community detection was done by simplifying and grouping the network nodes into groups[18].

After going through some literature we found some applications of news articles clustering as shown in figure 2. The text access can be done on basis of text content analysis. The text access can be categorized into two systems: recommendation and search engine. Using these systems, news articles clustering can be used for various purposes.

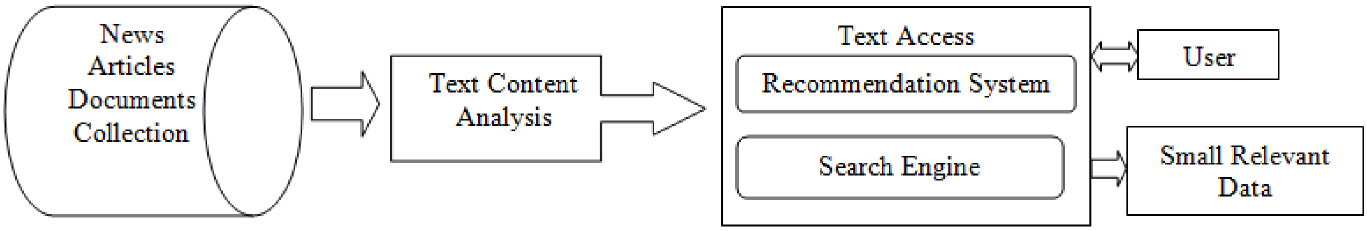

Figure 2: An architecture for Text analysis and its access.

The text access can be done on basis of text content analysis. The text access can be categorized into two systems: recommendation and search engine. Using this system, news articles clustering can be used for various purposes.

## III. CONCLUSION

Text mining and document clustering have been a topic of research for a long time. There are many different approaches and algorithms researched and applied by different authors. And all the solutions are applied to different datasets. So different algorithms show different results. In hierarchical clustering approach, clustering is done in the hierarchical way which makes understanding of data patterns better than flat clustering. Agglomerative clustering is similar to hierarchical clustering but it works in a reverse way. Partition clustering is more convenient methods but these methods depend entirely on random initiation of cluster centers. K-means clustering is the simplest clustering technique. There are various extended versions of this algorithm that comes under the hybrid approach. Hard clustering has limited applications as compared to soft clustering. Soft clustering approach shows better relevant results in retrieval systems. Semantic similarity approaches are based on the similarity between words, sentence or text documents can be determined using the dictionary approach, WordNet approach, co-occurrence relation. It has application in the domain ontology and taxonomy[19].

There is no common platform for clustering all the types of text documents and news articles datasets. And all these algorithms generate good results. Also, different systems are available for clustering such as IBM Watson[1], Carrot[12] etc. But still, this field needs more research and development to achieve much faster retrieval, big data handling, scalable and real-time clustering.

## REFERENCES

**[1]** https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

**[2]** https://en.wikipedia.org/wiki/Unstructured_data

**[3]** Smeaton, A. F., Burnett, M., Crimmins, F. and Quinn, G. 1997. An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Text. Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, 31(SI):74-81.

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

**[5]** Toda, H. and Kataoka, R. 2005. A Clustering Method for News Articles Retrieval. ACM, 988-989.

**[6]** Corso, G.M.D., Gullf, A. and Romani, F. 2005. Ranking a Stream of News. International World Wide Web Conference Committee (IW3C2) ACM, 97-106.

**[7]** Aiello, M. and Pegorett**,** A. 2006. Textual Article Clustering in Newspaper Pages. Applied Artificial Intelligence, 20(9):767-796.

**[8]** Newman, D. Chemudugunta, C. Smyth, P. and Steyvers, M. 2006. Analyzing Entities and Topics in News Articles Using Statistical Topic Models. Intelligence and Security Informatics LNCS. 3975:93-104.

**[9]** Sekiya, H., Kondo, T., Hashimoto, M. and Takagi, T. 2007. Context representation using word sequences extracted from a news corpus. International Journal of Approximate Reasoning, 45(3):424-438.

**[10]** Pera, M.S. and Ng, Y.K. 2008. Utilizing Phrase-Similarity Measures for Detecting and Clustering. Informative RSS News Clustering. Journal integrated Computer-Aided Engineering, 15(4):331-350.

**[11]** Pera, M.S. and Ng, Y.K.D. 2012. Using maximal spanning trees and word similarity to generate hierarchical clusters of non-redundant RSS news article. Journal of Intelligent Information Systems, 39(2):513-534.

**[12]** Krstajic, M., Araghi, M.N., Mansmann, F. and Keim, D.A. 2013. Story Tracker: Incremental visual text analytics of news story development. Information Visualization, 12:308-323.

**[13]** Malhotra, S. and Dixit, A. 2013. An Effective Approach for News Article Summarization. International Journal of Computer Applications, 76(16):5-10.

**[14]** Moe R.E. 2014. Clustering in a News Corpus. Lecture Notes in Computer Science, 8655:301-307.

**[15]** Devi, M.U. and Gandhi, M. 2015. An Enhanced Fuzzy Clustering and Expectation Maximization Framework based Matching Semantically Similar Sentences. Procedia Computer Science, 57:1149-1159.

**[16]** Dangre, N., Bodke, A., Date, A., Rungta, S. and Pathak, S.S. 2016. System for Marathi News Clustering. Procedia Computer Science, 92:18-22.

**[17]** Blokh, I. and Alexendrov. V. 2017. News clustering based on similarity analysis. Procedia Computer Science , 122, 1715-1719.

**[18]** Nicholls,T. and Bright, J. 2018. Understanding news story chains using information retrieval and network clustering techniques. Social and Information Network, Information Retrieval. arXiv:1801.07988[cs.SI].

**[19]** Ahmed, R. and Ahmad, T. 2018. Fuzzy Concept Map Generation from Academic Data Sources. Internationl Conference on Signals, Machines and Automation NSIT.