A Framework for Semantic Clustering of News Articles Based on Fuzzy

Nidhi Dewan[1] and Shagufta Farzana[2]

Research Scholar, Department of Computer Science and Engineering[1]
Assistant Professor, Department of Computer Science and Engineering[2]
Dr. C. V. Raman University, Bilaspur, India

Abstract: *Text mining is a process that uses data mining approaches to extract valuable information held in the hidden form in textual data. In this paper, we are proposing a framework for fuzzy clustering of news articles. These news articles originate on different news portals on the web. The data obtained need to be stored in a central database and then pre-processing reduces the noise. The keyword extraction is used to extract keywords from the text and then word-frequency vector is generated. On these vectors, distance measure or similarity measure function is used to find the similarity between articles. One article may belong to more than one cluster so fuzzy context vector must be generated. Mutual Information can be used to find fuzzy membership values. The threshold values are required for the identification of clusters. The proposed framework shows that fuzzy clustering does not restrict each news article to belong exactly to one cluster. Therefore this framework when applied to informationretrieval systems or other application systems, system gives better performance and relevance to the users.*

Keywords: News article, Clustering, Fuzzy, Text Mining

I. INTRODUCTION

In today's era, with the increased usability and involvement of the internet into our lifestyle, the data on the web is increasing. The ever-increasing share of our communication, interaction, and culture is in recorded digital text form. This availed opportunity for research scholars to use the information encoded in text, which is also a very rich source of commercial values. A recent study revealed that greater than 80 percent of the information a company stored is in digital form[1]. Also, we can read our choice of newspaper on the web, we can be updated minute by minute on the latest news by blogs and newsletters.

Clustering is a technique aimed at grouping a set of elements into clusters or subsets. The objective is to create clusters that areinternally coherent but different from each other, substantially. In other words, the similarity between elements in the same clustershould be as high as possible, whereas the similarity between elements in different clusters should be as low as possible. Clusteringplays an important role in many fields like Information Retrieval, data mining, knowledge discovery, data management etc. Clustering is a form of unsupervised learning. The aim of a good clustering scheme is to minimize intra cluster distance measure(similarity measure) while maximizing inter-cluster distance measure.

Text mining is the more complex task as compared to data mining, since text mining involves working with underlying unorganized and unclear data. Such data is known as fuzzy data and clustering this data is known as fuzzy clustering, which is alsoknown as soft clustering. Unlike soft clustering technique, the hard clustering techniques, which are conventional techniques, restricts each point in the data set to belong to one cluster exactly. In 1965, theory on fuzzy set was proposed by L.A. Zadeh , whichgave an idea about the uncertainty of belonging which was explained by a membership function. The fuzzy sets are used to calculate imprecise or fuzzy membership information[2]. In 1966, the application of this theory in cluster analysis was proposed inthe work of author[3].These papers made the door opened for research on this theory. Now, this clustering technique hasbeen widely researched and applied in various application areas.

II. RELATED WORK

Text clustering is also known as document clustering. We found a lot of related work on text clustering. It has wide applications. Also, clustering is one of the important areas of research since it has applications that affect our daily activities and needs, such as search engines, recommendation systems, information retrieval systems, etc. The applications of text clustering canbe divided into two categories online and offline. One of our need is to be updated on recent and accurate news of our choice. Thismakes news clustering important for research. The different approaches can be broadly categorized into hierarchical, non- hierarchical, fuzzy and crisp. Google news uses many state of the art algorithms for indexing, ranking and to solve problems relatedto clustering.

Alan F. Smeaton *et.al.*, have proposed Group Average Cluster(GAC)-based hierarchical clustering technique was used for clustering 15,836 news stories. Comparison of GAC clustering method was done with partition approach and results show that hierarchical GAC gave better performance[4].

Yimi Yang *et.al.*, presented one clustering algorithm for online news clustering. The Agglomerative & hierarchical clustering isused. they showed a complete link clustering method which generated small, tight clusters as result. The experiment was executedon dynamic clustering of online news articles[5].

Hiroyuki Toda *et.al*, have anticipated clustering for news articles retrieval system in which he used Named Entity extraction method, which was evaluated using news articles from Japanese newspaper articles only. This shows the problem of evaluation being language specific[6].

Gianna M. Del Corso *et.al*, have presented the lexical similarity calculation method was used to find similarity measure valuesto obtain ranking of news articles[7].

Marco Aeillo *et.al.,* have anticipated the results that were generated on the connection graph for evaluation of the proposed clustering algorithm. In this paper, three algorithms were compared: agglomerative clustering with bigram indexing, simple clustering with bigram indexing and comparative clustering with bigram indexing and analysis resulted that simple clustering generated overall best performance because of decreased complexity and increased flexibility[8].

David Newman *et.al.*, proposed a method in which probabilistic topic modeling is combined with entity recognizers. The topic-model-relationships were generated and used while representations were done using bipartite graph which gave a better understanding of the latent structure between entities[9].

Hiroshi Sekiya *et.al.,* have presented the evaluation and experimentation, that was performed on 8,00,000 news stories. The MI, Jaccard coefficient and chi-square algorithms were used to calculate membership values between news articles[10].

Maria Soledad Pera *et.al.,* have proposed an approach in which InFRSS algorithm was introduced for clustering of textual data.This approach consists of two models is a correlation based phrased matching combined with fuzzy compatibility clustering and both of these generate better quality cluster than well-known techniques of clustering[11].

Maria Soledad Pera *et.al*, have proposed a method in which RSS news articles were taken for clustering and fuzzy similarity algorithm was used which was based on fuzzy equivalence relation[12].

Christos Bouras *et.al.,* have presented a new algorithm towards an approach to enhance K-means technique using knowledge from database. This database is external knowledge database. Hence enhancement of K-means algorithm is done in a twofold manner. Therefore, results measured 10 times improve over standard K- means[13].

Srinivas Vadrevu *et.al.,* have introduced a clustering system which consisted of these components offline incremental and real time clustering of the news article. This system is scalable for large collection and resulted in good accuracy in the clustering of thenews articles. In this system similarity is calculated using local sensitive hashing (LHS) to get pairwise similarity values, then similarity graph is constructed and correlation clustering method is applied on this graph to get generate final clusters[14].

Maria Soledad Pera *et.al*, have presented the FICUS method and is used to cluster news articles. This method uses fuzzy approximation to identify related words in articles[15].

Twinkle Swadas *et.al.,* have conducted a study on clustering of news articles. The method proposed was a clustering algorithmin which multi-step terms or feature section process is included. This process is semi-supervised approach. After preprocessing of the news articles data filter methods are applied to evaluation features for the main clustering C-means algorithm is used[16].

Milos Krstajic *et.al.,* the framework for clustering module is based Carrot, which is an open source framework for clustering[17].

Christos Bouras *et.al.,* have proposed a novel approach in which standard K- means algorithm is extended using an external source of knowledge that is WordNet that was used before the clustering process. The corpus used consisted of

DOI: 10.48175/IJARSCT-5415

8000 articles whichbelonged to different fields of : science and technology, business, sports, education, entertainment, politics etc. The proposed algorithm can be clarified into the hybrid technique of clustering[18].

Anestasia N. Soloshenko *et.al.,* have reported a structure for the news articles text. This structure is based on the principle of 'inverted pyramid'. If the structure of all the news articles structures made according to one defined structures, then significantly simplified solutions could be obtained for the clustering problem. The corpus was taken from Russian news articles. After implementations of online news aggregation system and evaluation of its effectiveness authors have concluded that the most appropriate way for clustering news articles is FCM algorithm or neural network[19].

N. Dangre *et.al.,* have presented the comparison between various clustering algorithms was done and determination of best performing algorithm was done among them for Marathi news clustering. They presented ranked clusters from multiple sources[20].Jonathan A. Marshal *et.al.,* have introduced an algorithm MIMOSA for clustering task the results were generated on 10,000,000news articles, which shows that time algorithm achieves liner time complicity for clustering pf these articles, MIMOSA ie., Mark-In, Match-Out similarity analysis algorithm is signature based in which every data item is assigned a signature of limited size[21]

Tom Nicholls *et.al.,* have proposed BM25F algorithm. This algorithm was used clustering of articles which works on scores tofind related articles[22].

In this way, other research methods used different approaches to solve different problems related to clustering of news articles.In this paper, we solve the problem of fuzzyness found in news text, which human easily understand by reading and analyzing the context. Also our proposed method uses soft clustering since one news article may belong to more than one cluster.

III. MATERIAL AND METHOD

The architecture for clustering algorithm implementation is shown in figure 1. The architecture shows that we need a data source to be determined. Some news portals provide API to provide data in a structured format. Another way to get data is by webcrawling or web scraping.

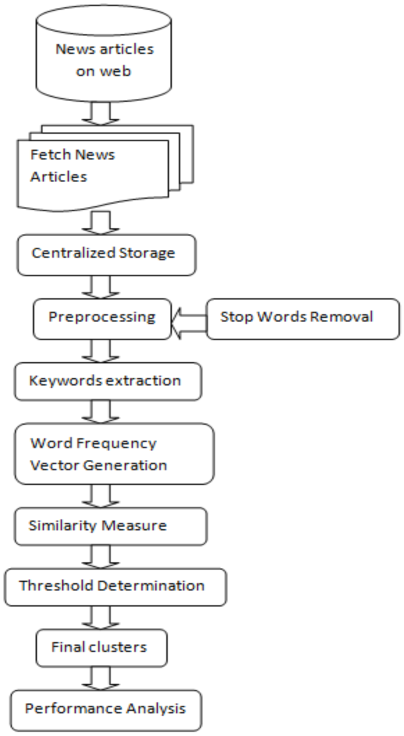

Figure 1: An architecture for the proposed algorithm

DOI: 10.48175/IJARSCT-5415

The steps involved in the proposed algorithm for clustering of news articles are listed below:

- Step 1: There are many news portals available on the web. So first step is to get the news articles data from the internet. Somenews corpora are available on Kaggle, GitHub etc. Other methods include API, Web scraping. Some news portals provide data via their API keys, which is a more convenient and batter method then web scraping. Prior to web scraping, HTML knowledge isrequired.
- Step 2: The data when fetched, must be stored in one place from where retrieval and execution becomes easy.
- Step 3: The pre-processing phase involves several steps. When data is fetched from the web it may come with a lot of unwanteddata such as advertisements, images, GIFs, audios etc., which must be removed as our clustering system work only with textual data. Removal of such unwanted data leads to high efficiency in clustering. Tokenization is used to convert a stream of text data into smaller units called tokens. Stop words removal is a process in which stop words such as 'in', 'the', 'a', 'is', etc. are removedas these word does not contribute in discrimination process to make clusters. Other unwanted data such as symbols, numerical data, and other language data are removed as they contribute as noise in the required dataset. Removal of these data helps to achieve accuracy. POS tagging is a useful process for finding semantic similarity. Stemming and lemmatization are used to find root wordsfrom a given word. Dimensionality reduction is a process which is used to reduce the dimension, since the news articles data consists of a very large vocabulary. All these steps of pre-processing lead to achieving high accuracy and efficiency.
- Step 4: Keywords extraction is a process of feature extraction in textual data analysis. Frequency can be used as a parameter forsuch large textual documents. The frequency count of each word is calculated on all the text documents. The most frequent words are considered as the keywords.
- Step 5: Word frequency vector is generated in which the frequency of keywords from all the documents are enumerated. Thesevectors are then given as input for similarity measure computation.
- Step 6: Fuzzy membership values are calculated to find similarity between news articles. The co-occurence is used to find semantic similarity. We have to get the statistical information among tokens by scanning the documents to get co-occurence probabilities. Mutual information, I(x,y) between two entities or points can be calculated using their co-occurence joint probabilities P(x, y) when two points (keywords), x and y, having probabilities P(x) and P(y). equation (1) gives mutual information. Mutual Information is a method for computation of the dependencies between two entities. Equation (2) gives mutualinformation between two keywords[23].

$$I(x, y) =- \log_2 (P(x, y) /P(x)\, P(y)) \qquad (1)$$

$$MI(k_i, k_j) = \log_2 \left(\frac{\mathbf{Pr(k_i\, k_i)}}{\mathbf{Pr(k_i) * Pr(k_j)}} \right) \qquad (2)$$

- Step 7: Determination of threshold value is done after calculation of membership values as on the basis of threshold values final clusters are built from the data.
- Step 8: Performance evaluation phase is required to check the performance metrics of the resulted clusters of the articles. F- measure, precision and recall are the matrices that are used to evaluate clustering results obtained after the proposed algorithm is applied.

IV. RESULT AND DISCUSSION

The news articles data are unstructured text data that contains a lot of fuzzy data. Also, clustering of news articles is a complex task. As related work shows a lot of work has been done on this topic, many algorithms are developed and applied to different news articles data sets. The clustering of news is important as users want accurate and relevant news. One of the applications of clustering is information retrieval. Partition algorithms produce good results but in this case, clusters depend on initialization. Initialization of cluster centers is done randomly[12].

When users search their query, they get a lot of web pages during the search, but browsing through all those documents is time- consuming, so relevant information retrieval is important. After going through a lot of literature, we found that soft clustering is more efficient for news articles clustering. So our proposed clustering algorithm can generate clusters such that one article may belong to more than one clusters.

DOI: 10.48175/IJARSCT-5415

V. CONCLUSION

In unstructured text data, fuzzyness is created by imprecise and unclear data. This fuzzy data creates low precision for clustering applications. This is one of the major challenges seen in text data. Using the proposed fuzzy clustering technique improved accuracy and efficiency is achievable. Traditional algorithms generate clusters such that one article belongs exactly to one cluster only. K-means is the most preferred and convenient clustering algorithm. This is because of its simplicity and good results. But the limitation with this approach is the number of clusters needs to be specified before clustering which is unlike fuzzy clustering. So fuzzy clustering would show good performance when used as the intermediate step in other algorithms.

VI. FUTURE WORK

For clustering of unstructured news articles text data, fuzzy clustering gives better user experience and accuracy. So implementation and experiment results will be analyzed. In this paper, framework is presented. This work requires to be researched for real-time news data. In future aspects, come the scalability, big data and incremental data. Also, some of these algorithms have become language dependent. There is no single algorithm or platform for clustering of news stories of any language. This aspect of text mining and cluster analysis will be taken up in future implementation.

ACKNOWLEDGMENT

I am very much thankful and grateful for the important advice and consistent guidance of Miss. Shagufta Farzana, Assistant Professor, Dr. C. V. Raman University. I am also blessed to the head of the department, Dr. Rohit Miri for his support and guidance. At the end I am very much thankful to committee members who have supported me directly or indirectly support in my research work.

REFERENCES

[1]. HTTPS://EN.WIKIPEDIA.ORG/WIKI/UNSTRUCTURED_DATA

[2]. Zadeh, L. A. 1965. Fuzzy Set. Information And Control,8(3):338-353.

[3]. Bellman, R., Kalaba, R. and Zadeh L. 1966. Abstraction and Pattern Classification. Journal of Mathematical and Applications,13(1):1-7.

[4]. 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.

[5]. 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.

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

[7]. 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.

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

[9]. 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.

[10]. 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.

[11]. 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.

[12]. Pera, M. S. and Ng, Y. K. 2009. Synthesizing Correlated RSS News Articles Based on a Fuzzy Equivalence Relation. International Journal of Web Information Systems (IJWIS), 5(1);77-109.

[13]. Bouras, C. and Tsogkas, V. 2010. W-kmeans: Clustering News Articles Using WordNet, Knowledge-Based and intelligent Information and Engineering Systems, 4278:379-388

[14]. Vadrevu, S., Teo, C. H., Rajan, S., Punera, K., Dom, B., Smola, A., Chang, Y. and Zheng, Z. 2011. Scalable

DOI: 10.48175/IJARSCT-5415

Clustering of News Search Results. Proceedings of the fourth ACM international conference on Web search and data mining, 675-684.

[15]. 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.

[16]. Svadas, T. and Jha, J. 2012. Document Cluster Mining on Text Documents. International Journal of Computer Science and Mobile Computing, 4(6):778-782.

[17]. 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.

[18]. Bouras, C. and Tsogkas, V. 2014. Improving news articles recommendations via user clustering. ,8(1):223-237.

[19]. Soloshenko, A.N., Orlova, Y.A., Rosaliey, V.L. and Zotova, A.V.Z. 2014. Thematic Clustering Methods Applied to News Texts Analysis. Knowledge-Based Software Engineering, 466:294-310.

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

[21]. Marshall, J.A. and Rafsky, L.C. 2017. Exact clustering in linear time. arXiv:1702.05425[cs.DS].

[22]. 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].

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

[24]. Kon B. .2022. Research on the Fusion of Hybrid Fuzzy Clustering Algorithm and Computer Automatic Test Paper Composition Algorithm, School of Information Engineering, Volume, Article ID-4264144, pp.12.

[25]. Jan M. et al. 2022. Interest-Based Content Clustering for Enhancing Searching and Recommendations on Smart TV , Wireless Communications and Mobile Computing , Volume , Article-ID 3896840, pp. 14.

DOI: 10.48175/IJARSCT-5415