IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 4, April 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104123 |

The Study of Document Clustering Using TF-IDF

Abhishek Mankuskar, Abhishek A. Jambure, Pranal kale, Abhilesha Bhagwat, Dr. Rubeena Khan

Department of Computer, MESCOE, Pune, India

ABSTRACT: The rise of the internet platform and the successful digitization of a number of documents and other text for use in various environments such as corporate environments libraries and news agencies have been increasing exponentially every day. This large-scale increase has been apparent with the growth in the number of documents that are being seen every single day. A huge number of documents and other types of textual information is regularly intercepted by the news agencies that have to sift through large amounts of information to understand the context as well as do accurate and effective reporting. The process of manually analyzing this kind of information is extremely difficult as it can take large amounts of time even by a dedicated team of individuals. Therefore, there is a need for an effective technique that can accurately identify and cluster documents semantically to achieve effective improvement in the entire process. For this purpose, a number of related works have been analyzed and outlined in this survey article to approach our methodology which will be further elaborated in the upcoming editions of this research article.

KEYWORDS: Natural Language Processing, Deep Belief Networks, and Decision Making.

I. INTRODUCTION

In today's world, the technological revolution has provided individuals with a wide range of interconnected digital components that create massive amounts of data on an unprecedented scale. In the following years, it is expected that the data volume would increase dramatically. This massive amount of data comes in a variety of formats, including video, audio, images, and text. Analysts believe, however, that the bulk of data created nowadays is unstructured, primarily text. Organizations and businesses find it challenging to examine and extract useful information due to the vast volume of text data.

The document clustering has garnered considerable interest in recent times. It organizes the text collection into completely distinct clusters. The goal is to separate the text documents into separate categories so that text documents inside one category are similar to one other but different from text documents in those other categories. The more the resemblance or uniformity seen between text documents in a group, as well as the greater the divergence seen between chunks, the finer the clustering outcomes. Using text document clustering algorithms provides us with a number of benefits. Clustering the text documents is advantageous in the enhancement of navigation and discovering procedures. Take, for example, a library with tens of hundreds of books.

Users will be able to find their desired book more quickly if the books are organized by subject, theme, or substance. Individuals are also adept at clustering in two or three dimensions. Text data, on the other hand, is regarded a high-dimensional data set that is nearly difficult for a person to deal with manually, necessitating the use of text document clustering algorithms for automatic categorization. Furthermore, these approaches aid in the reduction of time complexity, with some of these completing tasks in a third of the time.

Though most effort has gone into making the procedure of clustering text documents as painless as possible, there are several challenges that arise when working with a large number of documents. The obstacles and challenges arising when text documents texts are grouped are the major topic of this research. Understanding what to expect in terms of future obstacles in the clustering process will assist us generate the perfect judgments possible when developing successful technologies and strategies.

The prevalence and co-occurrence of terms in a document have been used in conventional document clustering approaches. This means that such algorithms treat texts just as a string of words, ignoring any potential meaningful correlations between them. Because clustering is an unsupervised activity, the outcomes would not be of the highest possible quality leading to a shortage of advice on which papers genuinely correspond to the same group. To solve this challenge, creative techniques to text clustering are thoroughly examined in order to identify our solution. The approach that was developed will be detailed in future editions of this research paper.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104123 |

This literature survey paper segregates the section 2 for the evaluation of the past work in the configuration of a literature survey, and finally, section 3 provides the conclusion and the future work.

II. RELATED WORKS

S. Liu et al. [1] suggested word vectors based on automated news text segmentation. The procedure is based on pre-processing and cleaning each word from the chosen news texts. The selected word, such as nouns, verbs, adjectives, and so on, will be moved to the word tag. This allows them to assign distinct tags to the text expressions of different subjects. The most significant tags in the database are verbs and nouns, which might assist them grasp the subjects. Considering the characteristics of the news material in particular, they include adjectives as an indicator in their scope. As a result, the messages may be separated into distinct portions based on their frequency of recurrence.

S. Alaee et al. created an ontology-based technique for categorising e-Learning content. The proposed approach used term re-weighting and a clustering procedure, specifically a two-step algorithm, on the ICVL papers set [2]. The authors started by developing an ontology using materials from a well-known international e-Learning conference. In the clustering technique, idea weights were computed by taking the TF-IDF of the words and the related weights of ontology ties between each word and the other words into consideration after deleting stop words, stemming and merging synonyms, and unifying complex terms. The word weight vectors of these texts were then clustered using the clustering approach. The results of the experiments showed that reweighting enhances document clustering.

J. Mei et al. presented a comprehensive research on fuzzy clustering for massive document classification. New scale-up HFCM and scale-up FCoC clustering algorithms with high scalability and effectiveness are provided. Extensive experimental research using tagged and unlabeled benchmark document datasets represented that these suggested techniques work well [3]. The authors also presented three novel fuzzy coclustering (FCoC) ways to enhance the scalability of FCoC with the three scale-up clustering strategies by utilizing the equivalence between word membership in FCoC and centroids in FCM. Their experimental results reveal that FCoC's greater efficacy in dealing with high-dimensional data is maintained throughout all three scaled versions.

Diaz Harizky Firdaus present a tweet clustering algorithm which used to determine the themes that are covered in a large number of documents. Economic, social, political, and governmental issues are among the subjects covered. The usage of short-term memory is recognised to be particularly significant in boosting the performance of the clustering process, as seen by the shape of the clusters that develop, as observed in the preset test scenarios. The usage of kd and kp has a significant impact on the results of cluster formation. Another finding was that the employment of the kp had a significant impact on the time it took the system to complete the clustering process [4]. With the kd and kp parameter values of 0.1 and 0.005, respectively, the average cluster quality created by the system is 0.3455s.

For document clustering, a Red black Tree-based approach with a Word-Net ontology preparation strategy was developed by E. K. Jasila et al.. This hybrid technique improves text clustering performance in terms of dimensionality reduction, shorter execution time, and increased accuracy. The authors proposed a unique clustering strategy for increasing clustering accuracy while decreasing execution time. Experiment results on the NewsGroup dataset demonstrate that their suggested strategy improves dimensionality reduction, execution speed, and accuracy [5].

X. Pei et al. presented the CF with adaptive neighbours model which is an unique regularised CF model for improving document clustering performance (CFANs) [6]. It enhances the basic CF model by including an ANs regularisation requirement. The approach enhances the encoding matrix's neighborhood preserving property. Therefore, CFAN is utilized to obtain a representation space that maintains neighbourhood features from the original data set. Unsupervised CF approaches are utilized in the CFAN and current CF algorithms. The purpose of the CFAN technique is to adaptively create the neighbourhood graph weights matrix by including the ANs regularisation constraint into the CF model.

On the basis of rapid search clustering and the Density Peak Clustering (DPC) approaches, Vinay Kumar Kotte et al. propose an improved K-means clustering strategy for document clustering. Because text-depend data is frequently high dimensional and inept, they introduce a deep random prediction dimensionality reduction framework based on Stacked-Random Projection (SRP), a greedy layer-wise architecture. To begin, they employ the dimensionality compression approach to reduce the dimensions of the high-dimensional text feature. Following that, they apply the nave Bayes classification method to determine the number of dimensions before grouping the data with k-mean clustering [7]. The

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104123 |

findings show that using the Naive Bayes classification and improved K-Means clustering (KMC) algorithms on the chosen character set, significant changes in the values are obtained without the requirement for the original dataset.

M Thangarasu introduce a MPRK clustering algorithm that efficiently clusters the big text document dataset and provides a straightforward clustering method with fewer computing steps. The authors compare the MPRK algorithm to the PKM and D-PPSOK algorithms in terms of execution time. To test current algorithms, they employ a huge dataset in a single computer. It becomes infeasible for huge data sets, and parallelization is required to test the suggested work on a large dataset[8]. Experiments on a large number of real-time datasets show that the proposed approach is successful and outperforms the PKM and D-PPSOK algorithms

Chunxia Jin et al. offers a microblog short text clustering technique based on a frequent closed word set. The method finds the Top-K most common closed word sets with a length more than or equal to min length. It is possible to considerably reduce the size of vectors by using these often closed word sets to represent feature vectors of micro-blog short content in a clustering technique. Second, microblog documents that use the same common word sets are grouped together. These overlapping papers from different groups are then sorted by word frequency [9]. Finally, these overlapping texts are re-divided using a similarity computation based on the highest frequent word set. The approach is useful for grouping papers with the same topic into the same category and boosting the micro-blog short text clustering algorithm's accuracy.

A. Sheri In recent times, document clustering algorithms that use discriminating information from documents for K clusters have exhibited effective high performance [10]. These clustering methods repeatedly project articles into a K-dimensional discriminating information space and allocate them to the cluster with the highest value along the axis. The discriminatory intelligence space is defined by the term discrimination information at every cycle, with the term discrimination information approximated from the labelled document collection generated in the previous iterative process. The optimal solution for this clustering approach is any of the known analytical and knowledge metrics, and the task is framed as an optimal control problem.

Maedeh Afzali provides a quick explanation to clustering and the most often used approaches in text document clustering is presented in this work [11]. The article goes over a variety of issues that may arise when clustering a large number of text documents and goes over each one in depth. It is critical to have prior knowledge of the potential issues that others might face when selecting to use cluster analysis to a large quantity of text data; this information assists them in creating the correct framework and methodology selections depending on their needs.

III. CONCLUSION AND FUTURE SCOPE

The process of document clustering based on semantic rules has been one of the most complex and complicated mechanism. This type of clustering is extremely difficult and would take a large amount of time to be performed by an individual manually. This large amount of time and the very fast pace of the news agencies leads to a lot of disparity and time being wasted due to manual operation of this process. There is a need for an effective and useful mechanism for the purpose of clustering the documents received semantically according to the content contained in them. Therefore, a number of related works have been identified and elaborated in this survey article to determine our methodology for document clustering in a semantic manner. This approach will be further elaborated in the upcoming editions of this research article.

REFERENCES

[1] S. Liu, X. Fan and J. Chai, "Clustering analysis of feature words in news text based on co-occurrence matrix," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017, pp. 1-5, DOI: 10.1109/CISP-BMEI.2017.8302144.
[2] S. Alaee and F. Taghiyareh, "A semantic ontology-based document organizer to cluster elearning documents," 2016 Second International Conference on Web Research (ICWR), 2016, pp. 1-7, DOI: 10.1109/ICWR.2016.7498438.
[3] J. Mei, Y. Wang, L. Chen and C. Miao, "Large Scale Document Categorization With Fuzzy Clustering," in IEEE Transactions on Fuzzy Systems, vol. 25, no. 5, pp. 1239-1251, Oct. 2017, DOI: 10.1109/TFUZZ.2016.2604009.
[4] Diaz Harizky Firdaus and Suyanto Suyanto, " Topic-Based Tweet Clustering for Public Figures," 2020 3rd International Seminar on Research of Information Technology and Intelligent System (ISRITI), DOI: 476-481. 10.1109/ISRITI51436.2020. 9315449.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104123 |

[5] E. K. Jasila, N. Saleena and K. A. A. Nazeer, "Ontology Based Document Clustering - An Efficient Hybrid Approach," 2019 IEEE 9th International Conference on Advanced Computing (IACC), 2019, pp. 153-157, DOI: 10.1109/IACC48062.2019.8971594.

[6] X. Pei, C. Chen and W. Gong, "Concept Factorization With Adaptive Neighbors for Document Clustering," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 2, pp. 343-352, Feb. 2018, DOI: 10.1109/TNNLS.2016.2626311.

[7] Kotte, Vinay & Vuppu, Shankar & Thadishetti, Rachana, "High Dimensional Text Document Clustering and Classification using Machine Learning Methods", International Conference on Intelligent Computing and Control Systems (ICICCS 2021), IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2.

[8] M. Thangarasu and H. H. Inbarani, "MPRK algorithm for clustering the large text datasets," 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 224-229, doi: 10.1109/ICACA.2016.7887955.

[9] C. Jin and Q. Bai, "Short Text Clustering Algorithm Based on Frequent Closed Word Sets," 2019 12th International Symposium on Computational Intelligence and Design (ISCID), 2019, pp. 267-270, doi: 10.1109/ISCID.2019.10144.

[10] A. M. Sheri, M. A. Rafique, M. T. Hassan, K. N. Junejo and M. Jeon, "Boosting Discrimination Information Based Document Clustering Using Consensus and Classification," in IEEE Access, vol. 7, pp. 78954-78962, 2019, doi: 10.1109/ACCESS.2019.2923462.

[11] M. Afzali and S. Kumar, "Text Document Clustering: Issues and Challenges," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 263-268, doi: 10.1109/COMITCon.2019.8862247.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details