Frequent Concepts-based Document Clustering using Correlation (FCDCC)

Rekha Baghel
Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh

Kanika Singhal
Indraprastha Engineering College, Ghaziabad, Uttar Pradesh

Ruchira Goel, PhD
Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh

ABSTRACT

Document clustering (text clustering) is the way by which meaningful patterns can be extracted from large datasets. It helps to achieve prompt information retrieval, precise topic mining or in short streaming. Document clustering is useful in the extent of filtering out similar documents and further finding the distinct topics and subtopics. Huge amount of information is available as documents on online sources such as Newswire and different Blogs. Document clustering techniques can be used for managing such document datasets.

But high dimensionality is still a big challenge for mostly Document clustering algorithms. In this paper a new and efficient methodology of document clustering using correlation analysis is presented to address the problem of high dimensionality and finding a better solution.

Keywords

Data mining; Document Clustering; Correlation Analysis;

1. INTRODUCTION

Clustering is a process by which data points can be partitioned in to clusters in such a way that clusters have supreme intra-cluster similarity and lessen the inter-cluster similarity. Document clustering has been extensively studied in the literature, with traditional methods focusing on techniques like hierarchical clustering, k-means, and agglomerative clustering. Recent advances in natural language processing have led to the development of topic modeling approaches, such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), which have gained popularity.

Concept-based clustering has also been explored, but the emphasis has primarily been on concept extraction and document representation, rather than utilizing concepts for clustering purposes. In contrast, our approach leverages frequent concepts and correlation measures for improved clustering quality.

Document clustering research area consistently attracting the researchers and is widely used in various field such as specific topic extraction, Topological data Analysis, Web mining etc.

There are few issues that need to be addressed in order to promote the wider adoption and application of Document clustering. These issues includes High dimensionality, scalability, capability to treat heterogeneous attributes, Constraint- based clustering, Detection of clusters with random shape, Incremental clustering and insensitivity to input order etc. [1].

A new method Frequent Concepts-based Document Clustering using Correlation (FCDCC) is proposed in this paper. This paper uses a pure correlation analysis to find dependency among documents. This helps in getting better results.

Recent work related to document clustering is explained in Section II. Section III includes the details of proposed clustering technique. Section IV presents the experimental results. Finally, section V summarizes the article.

2. RELATED WORK

In [2], authors focused on the occurrence of the word present in the document. They proposed a novel approach based on lexical chains.

In [3] the authors focused on the unsupervised document clustering and applied it on the polish language. Their results indicate that the WordNet based similarity measures outperforms the recent embedding approaches such as BERT.

In [4] the authors proposed a new approach that incorporates features of both semantic and statistical based approaches.

In [5], in this paper the authors offered the solution for the effective navigation and browsing for the large datasets. They designed a new algorithm using two approaches Rider Optimization Algorithm (ROA) and Moth Search Algorithm (MSA).

3. PROPOSED ALGORITHM

Figure 1 explain the working of FCDCC algorithm for document clustering. In FCDCC algorithm for document clustering we take a new text document and apply preprocessing in it. After that all the root words are identified. Then we search for those root word in domain specific dictionary where the complete meaning of those words are present and also there synonyms. After all this we extract the concepts of the words. Like if we have ball, bat, pitch types words then the concept of this is cricket. Here clustering occurs which is based on Concepts. The clusters that are already formed is compared with new documents. Most similar concepts have similar cluster.

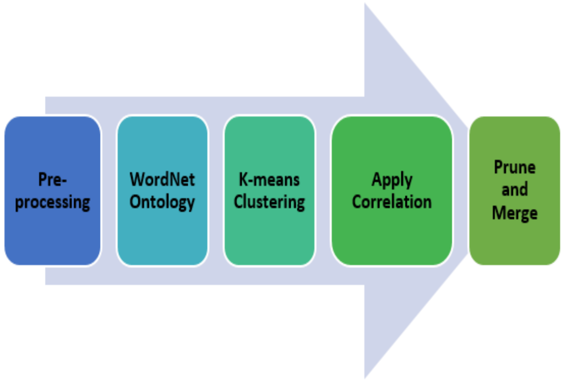

Figure 1: Frequent Concepts based document clustering using Correlation

3.1 Stop words removal

Preprocessing is a vital step that needs to be applied in advance to clustering algorithm. This step can lead to better clustering results.Prepocessing include separation of sentences and then it identify the part of speeches and tag them. After that all stopwords are removed. Stop Words are considered as meaningless words since they generally appear as articles or prepositions instead of index. So these stop words are cleaned out from search queries. Stop words have a long list to go but few of the most commonly occurring words are, a, an, and, am, as, are, is, in, it, the, be, by, from, for, has, etc. Stop words are removed after the preprocessing step.

3.2 WordNet

Mostly popular document clustering algorithms follow the BOW approach. In bag of words approach the semantic relationship between words does not considered. In order to deal with this issue WordNet database is used after preprocessing step. It is a huge English language vocabulary database, a grouping of a lexicon and glossary.

In WordNet[11], synsets are used which represents the synonymous words, collocations and corresponds to a concept. WordNet bifurcate the word into four classes nouns, verbs, adjectives and adverbs. A unique synset number is assigned to each synset. One word can belong to more than one synset for e.g. Apple can be a fruit or a mobile phone. To investigate the current techniques used to tag words into parts-of-speech Look at correlation techniques to determine whether any may be useful for solving the problem. Obtain a variety of samples of written text on which to apply correlation techniques on.

Table 1. WordNet database description

Words	Synsets	Word-sense pairs
155,287	117,659	206,941

3.3 Using k- means clustering to make cluster

K-means clustering is the widely used partitioning method which partitioned the data sets in to n different clusters automatically.

Here, $c_1;c_2;...;c_n$ are called data points .The algorithm works in following iterative manner:

Step 1.Initially select the n number of clusters, and initialize the data points $\{c_j\}^n$,j=1.

Step 2.For the input d_t, calculate $I(j|d_t)$

Step 3.Only update the winning data point c_w, i.e.,

$I(w|x_t) =1$, by $c^{new}{}_w = c^{old} + r(x_t-c^{old}{}_w)$ (1)

Here r is a minor positive learning rate. Steps 2 and step 3 are continually applied for each input until all data points converge.

We choose k means algorithm since while dealing with huge datasets it requires minimum computation [7]

3.4 Apply Correlation using q most item-set of each cluster

Correlation-based clustering is employed to group documents based on the relationships between their concept vectors. We compute correlation coefficients, such as Pearson correlation or Jaccard similarity, to measure the similarity between document pairs. Hierarchical clustering or density-based clustering algorithms can then be applied to create meaningful clusters

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})+(x_i-\bar{x})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})+(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2}} \qquad (2)$$

Correlation indicates how strong the relationship between two variables. If two variables are highly correlated then relationship between them must be strong vice versa poor correlation means variables are having the weak relationship. In our approach we have used correlation analysis to find out the dependency between two documents. If the documents under comparison show high correlation then they are highly related and can form one cluster. Equation (2) will be used to find out the correlation between the documents

3.5 Prune and Merge

Pruning is the process of cutting down or removing the unwanted nodes. We can prune the whole sub-tree if not required in the process. This helps in decreasing the computation cost of the nodes which are not required. After pruning the unwanted nodes the algorithm will merge the remaining nodes and the computation will be done on that remaining nodes.

4. EXPERIMENTAL EVALUATIONS

For assessing the efficiency of proposed algorithm, results were compared with the several standard document clustering algorithms such as FCDC [8], k-means and FIHC [9] etc. Comprehensive evaluation has done on large News20 dataset which includes 20.csv files of newsgroup documents. To evaluate the performance of the proposed algorithm FCDCC Evaluation measure F-Measure is used. The performance of various algorithms are evaluated for the 5, 15, 40 and 75 cluster size. FCDCC outperformed other algorithms in each cluster size. Comparative Analysis of results is shown in the table 2.

This technique will help in reducing time and will bring more efficiency to the algorithm by not focusing on the unwanted nodes and thereby saving the time and enhancing the efficiency of the algorithm.

Table 2. Comparison of different algorithms using F-score

Dataset	**No.of Clusters**	**Overall F-Measure**			
		FCDCC	**FCDC**	**FIHC**	**K-means**
News20	**5**	0.58	0.53	0.63	0.62
	15	0.57	0.54	0.53	0.45
	40	0.61	0.58	0.56	0.43
	75	0.62	0.61	0.51	0.35
	Average	0.59	0.56	0.55	0.46

5. CONCLUSION

The main aim of this work was to improve scalability, reduce high dimensionality by replacing the data representation method of vector space model by singular matrix method. Proposed work not only worked on synonyms but it will also include other similarity criteria's for example hyponymy, holonymy, and meronymy for enhancement of accuracy and efficiency of document clustering and making the way for users much easier when they search for a query on a search engine. Future research directions include the exploration of advanced concept extraction techniques, the development of more sophisticated correlation measures, and the application of our approach to other domains and languages. Additionally, investigating the scalability of our method to handle even larger document collections is of interest. In the future, the methodology proposed for document clustering using correlation analysis can be applied to more diverse and large-scale datasets, such as real-time streaming data, academic research articles, and social media content, to validate its scalability and effectiveness in various domains. Complex and unstructured text data can be clustered more accurately with advanced deep learning models like transformers or autoencoders.Developing real-time or near-real-time document clustering would enable its use in streaming data contexts like news feeds and social media monitoring.

6. REFERENCES

[1] Benjamin C.M. Fung Ke Wang Martin Ester "Hierarchical Document Clustering Using Frequent Itemsets"Simon Fraser University, BC, Canada, ester@cs.sfu.ca

[2] Green, S. J. 1999. Building hypertext links by computing semantic similarity. T KDE, 11(5), pp. 50–57.

[3] Gniewkowski, Mateusz , Walkowiak, Tomasz and dkowski, Marcin" "Text Document Clustering:

[4] {W}ordnet vs. {TF}-{IDF} vs. Word Embeddings","Proceedings of the 11th Global Wordnet Conference","2021"

[5] D. R. Cutting, J. O. Pedersen, D. R. Karger, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In Proceedings of the ACM SIGIR, 1992

[6] Vivek Mehta, Seema Bawa, Jasmeet Singh,Stamantic clustering: Combining statistical and semantic features for clustering of large text datasets,Expert Systems with Applications,Volume 174,2021,114710,ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2021.114710.

[7] Madhulika Yarlagadda, K. Gangadhara Rao, A. Srikrishna,Frequent itemset-based feature selection and Rider Moth Search Algorithm for document clustering,Journal of King Saud University - Computer and Information Sciences,2019,,ISSN 1319-1578

[8] Jain, A.K, Murty, M.N., and Flynn P.J., "Data clustering: a review", ACM Computing Surveys, pp. 31, 3, 264-323,1999.

[9] Rekha Baghel and Dr. Renu Dhir, "A Frequent Concepts Based Document Clustering Algorithm", International Journal of Computer Applications, vol.4, no.5, pp.6-12, 2010.

[10] B.C.M.Fung, K.Wan, M.Ester. 2003. Hierarchical Document Clustering Using Frequent Itemsets", SDM‟03.

[11] Miller G. 1995. Wordnet: A lexical database for English.CACM, 38(11), pp. 39–41.