*https://doi.org/ 10.33472/AFJBS.6.10.2024.3870-3878*

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

# A Topical Reflection Measure Based Classification Model for Improved Document Clustering

**Dr.Abhijeet Madhukar Haval**, Associate Professor, Faculty of CS & IT, Kalinga University, Naya Raipur, Chhattisgarh, India.
Mail ID:ku.abhijeetmadhukarhaval@kalingauniversity.ac.in
ORCID ID: 0009-0003-1776-4795
**Akanksha Mishra**, Assistant Professor, Faculty of CS & IT, Kalinga University, Naya Raipur, Chhattisgarh, India.
Mail ID:ku.akankshamishra@kalingauniversity.ac.in
ORCID ID: 0009-0003-7723-5529

Article History

Volume 6,Issue 10, 2024

Received:13 Apr 2024

Accepted : 05 May 2024

doi: 10.33472/AFJBS.6.10.2024.3870-3878

**Abstract:**
The problem of document clustering has been approached with different techniques in literature. The popular k means clustering algorithm uses Euclidean distance as the similarity measure in clustering the documents. Similarly, different approaches use various similarity measures like Euclidean distance, term frequency in measuring the similarity among the documents towards document clustering. However, they suffer with poor clustering accuracy and overlap. To solve this, an efficient Topical Reflection Measure based classification model (TRMCM) is presented in this article. The proposed TRMCM model preprocesses the document text to obtain meaningful terms and applies feature selection to identify subset of terms. The selected features are used to measure TRM value at the classification phase. Based on the value of TRM, the method identifies the class of document to perform clustering. The proposed TRMCM model improves clustering accuracy with less overlap.
Index Terms:Document Clustering, Supervised Learning, TRMCM, TRM.

## 1. Introduction:

The organization maintains different information in form of documents. Such documents are huge in volume and need to be grouped in a meaningful way to explore the documents when necessary. When the volume of documents is increasing it is difficult to identify the required and similar documents from the document pool. To solve this problem, document clustering has been used. Clustering is the process of grouping related and similar documents under a specific name and by labeling the document with a common name helps the document retrieval system to identify the related document as result for a search query.

The document clustering is performed in several ways like supervised and unsupervised learning. There exist number of approaches available for document clustering. For example, the K-means clustering algorithm measures the Euclidean distance measure between the document text of the input and documents of the class to perform clustering. Similarly, the support vector machine is used in the same problem, which measures the similarity by measuring the support value according to the terms of the document given. On the other side, Bayesian classification is used for the problem which works according to the rule available. Similarly, you can name number of approaches to solve the problem.

The performance of clustering and classification algorithm is depending on the kind of feature being considered and kind of similarity measure used. There are number of features being considered like topic, terms, semantics and so on. By considering effective features the performance of document clustering can be improved. On the other side, the feature selection plays vital role in the achievement of clustering performance. When the method selects non important and irrelevant features, the performance of clustering has been gets affected. Similarly, when you miss important features, then the accuracy of clustering gets affected. So, the feature selection plays vital role in the accuracy of document clustering.

With the consideration to improve the performance of document clustering an efficient Topical Reflection Measure (TRM) based clustering model is presented in this article. Any document would contain number of terms which is related to different topics. In order to become a class of document, it is necessary that the document should reflect the concern topic in more precise way. By reflecting the topic of the class, the document can be assigned with the class label. By considering this , the proposed TRMCM model cluster the documents according to the TRM measure computed for various class of documents. The working of the TRMCM model is briefed in this section.

## 2. Related Works:

Number of clustering schemes is recommended in literature and this section analyzes set of methods around the problem.

An graph based auto encoder (GAE) based scheme is presented in [1], towards document clustering. The method construct the graph from different documents and measures cosine similarity to identify the document of the class.

An deep convolution auto encoder network (DCAN) based clustering model is presented in [2], which uses a integrated loss function in the softmax layer to perform document

clustering. A tensor based clustering framework is presented in [3], which cluster the hyper spectral data to detect the forged papers in multi-page documents. The method uses the diagonal structure of the documents to measure the similarity towards clustering.

A Document Vector Extension model is presented in [4], which divide the document in to several sub classes and establishes relationship among them. A black hole based hybrid clustering scheme is presented in [5], which computes external purity and internal silhouette score to perform clustering.

An adaptive Jaro Winkler with Jellyfish search clustering algorithm is presented in [6], which uses reuter data set to group newsletters.

Intention-guided deep semi-supervised document clustering (IGSC) model in [7], which divide the document structure according to user given information. The deep metric learner explores the user's global intention and outputs an intention matrix. The method use the intention matrix to perform clustering.

Fuzzy Local Information C -Means based clustering (FLICM) and Fractional Dwarf Mongoose optimization model is presented in [8], towards document retrieval and uses Bag of words and applies CNN to perform clustering.

A parallelized ontology network based semantic similarity clustering is presented in [9], which preprocess and extract the semantic features to calculate document semantic similarity based on ontology network structure under MapReduce framework. The value of semantic document similarity is used to perform clustering.

An Active Learning with Constrained Document Clustering is presented in [10], which applies SVM to perform initial clustering and generates a distance matrix according to the hyperplane to perform clustering with active learning.

The efficiency of various clustering algorithms are analyzed for their performance in clustering with different document collection [11]. An distance of term frequency-based similarity measure (DTFSM) and presence of common terms-based similarity measure (PCTSM) based document clustering scheme is presented in [12], which computes the DTFSM and PCTSM measures for the document to perform clustering. A Cosine Similarity and K-Main Algorithms based clustering model is presented in [13], which organize the large non sequential text documents into small clusters. An semantic information based document clustering algorithm is presented in [14]. Similarly, Efficient Document Clustering Approach is presented in [15], to obtain semantic clusters from a huge volume of documents.

## 3. Topical Reflection Measure Based Clustering Model (TRMCM):

The proposed TRMCM algorithm reads the document set given and performs preprocessing to remove the noisy features and applies feature selection to get the required features. Preprocessing is performed with term level noise removal algorithm and feature selection is performed with Feature Centric Frequency Analysis Algorithm. Further, the method applies TRM clustering which computes TRM measure for the document towards

various class of documents. Based on the value of TRM, the method identifies the class of document and indexed.

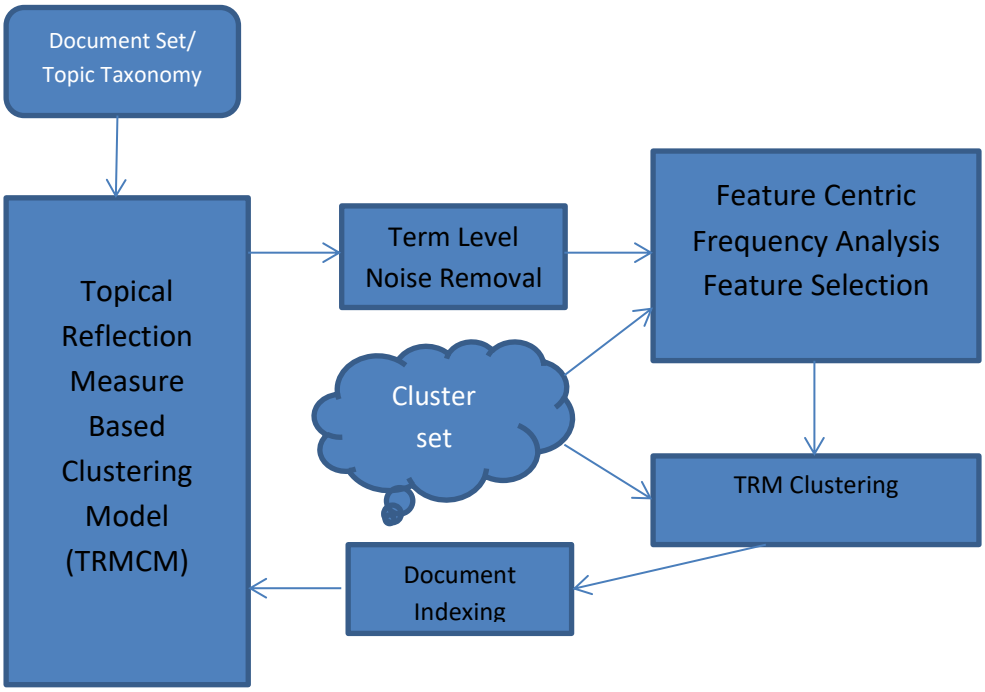


Figure 1: Process Diagram of TRMCM clustering model

The working process of TRMCM clustering model is presented in Figure 1, and the sequence of process is explained in detail in this section.

Term Level Noise Removal Preprocessing:

The term level noise removal preprocessing algorithm removes the noisy features from the document to support document clustering. The document set given has been read and the text features from each document are retrieved. From the text features, the method generates number of terms to produce term set. The terms of the set are mapped with the stop word list maintained by the model to eliminate the terms with no meanings. Further, remaining terms are stemmed to trim the terms and obtain pure terms. At last, the terms are tagged to identify the nouns with the Part of speech tagger. The terms tagged as noun only will be retained for the further process.

Algorithm:

Given: Document D, Stop word list Sl

Obtain: Term set Tset

Start

Read D and sl.

$$sentences(D)$$

Term set Tset = $Sentences\big(D(i)\big).split(".", \text{" space"})$

$$i = 1$$

For each term Tk

If Sl ∈ $Tk$ then

Tset = Tset ∩ $Tk$

Else

Tk = Stemming (Tk, remove (“ing”,”ed”))

Tktag = PosTagger(Tk)

If Tktag!=Noun then

Tset = Tset ∩ $Tk$

End

End

End

Stop

The term level noise removal preprocessing algorithm identifies the pure terms from the document to support document clustering.

Feature centric frequency Analysis Feature Selection:

The feature centric frequency analysis algorithm selects the subset of term features according to the frequency of appearance. To perform this, each term in the set has been measured for its frequency of appearance in the own term set. According to the value of frequency of appearance FoA, the method selects a subset of features or terms are selected features. The selected features are used to perform document clustering.

Algorithm:

Given: Term set tes

Obtain: Feature Set Fes

Start

Read Tes.

For each term T

Compute FoA = $\frac{\sum_{i=1}^{size(Tes)} Count(Tes(i)==T)}{Size(Tes)}$

If FoA > Th then

Fes = $\sum Terms(Fes) \cup T$

End

End

Stop

The feature centric frequency analysis based feature selection algorithm computes the value of Frequency of Appearance for different terms in the set and based on that the method identifies the set of features as selected features. Selected features are used to perform clustering of documents.

**TRM Clustering:**

The topical reflection measure based clustering algorithm reads the document set given and the cluster set available. For each document d, the method applies the term level noise removal algorithm to preprocess the document and obtain the set of terms in that. Further, the method

applies, term level frequency analysis to perform feature selection. With the selected features, the method computes Partial reflection measure (PRM) and completes reflection measure (CRM) values for each term identified. Using the value of PRM and CRM, the method computes the value of TRM. Based on the TRM value, the method identifies the class of the document and performs indexing.

Algorithm:

Given: Document Set Ds, Cluster set Cls, Topical Taxonomy Tt

Obtain : Null

Start

Reads Ds , TT, and Cls.

For each document d

Term set Tes = perform term level noise removal preprocessing (d)

Feature set fes = Apply term level frequency analysis feature selection (tes)

For each class c

Compute Partial Reflection Measure PRM.

$$PRM = \frac{\sum_{j=1}^{size(TT(c))} \sum_{i=1}^{size(Fes)} Count(TT(c(i)) partially\ matches\ with\ Fes(j))}{size(Fes)}$$

Compute Complete Reflection Measure CRM.

$$CRM = \frac{\sum_{j=1}^{size(TT(c))} \sum_{i=1}^{size(Fes)} Count(TT(c(i))\ completely\ matches\ with\ Fes(j))}{size(Fes)}$$

Compute TRM = $\frac{PRM}{CRM} \times size(TT(c))$

End

Class C = Choose the class with maximum TRM.

Index the document to the selected class.

End

Stop

The TRM clustering algorithm computes the value of CRM and PRM for the document towards various class to compute TRM measure. Based on the TRM value, a optimal class is identified and the document is indexed to the selected class.

## 4. Results and discussion:

The proposed TRMCM model has been implemented using Advanced java and has been evaluated for its performance with Reuters data set. The results obtained have been compared with the results of other approaches.

| Key | Value |
|---|---|
| Tool Used | Advanced Java |
| Data set used | Reuters |
| No of classes | 20 |
| Total documents | 18750 |

Table 1: Experimental data

The experimental details used for performance evaluation of proposed algorithm are presented in Table 1.

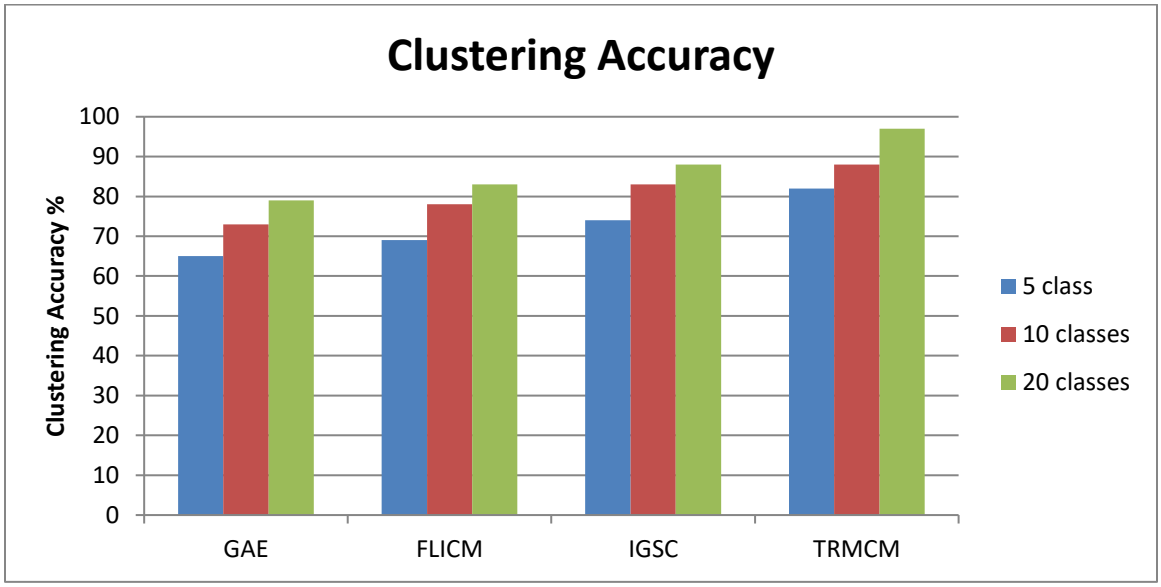


Figure 2: Analysis on clustering accuracy

The accuracy of clustering produced by various methods are measured and plotted in Figure 2, where TRMCM algorithm produces higher clustering accuracy than other methods.

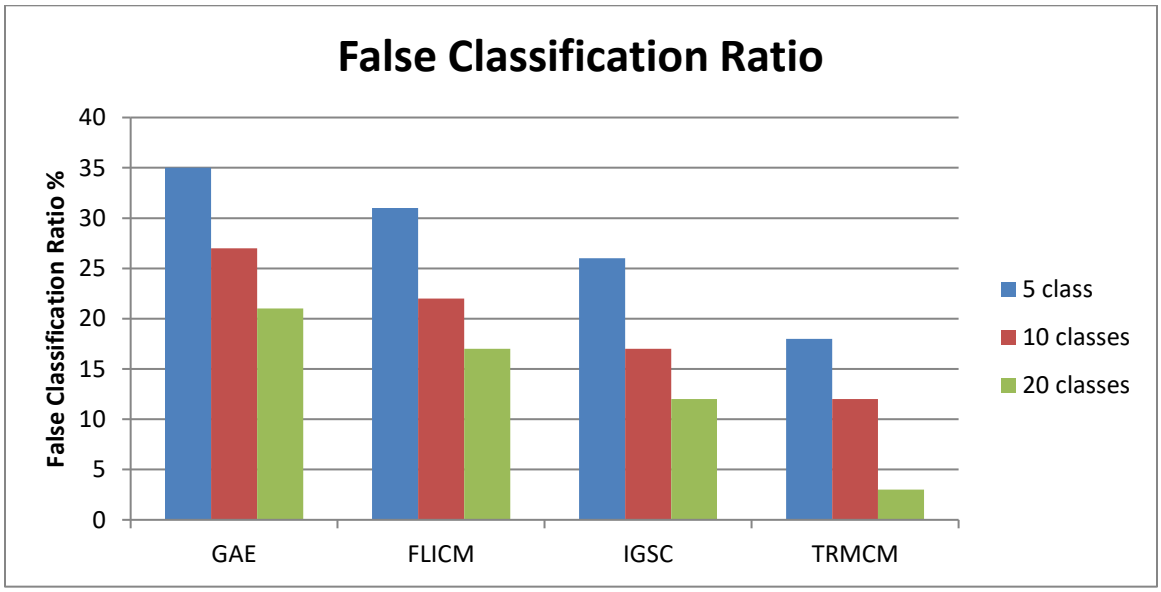


Figure 3: False Classification Ratio

The ratio of false classification produced by various methods are measured and presented in Figure 3, where TRMCM algorithm produces less false ratio than other methods.

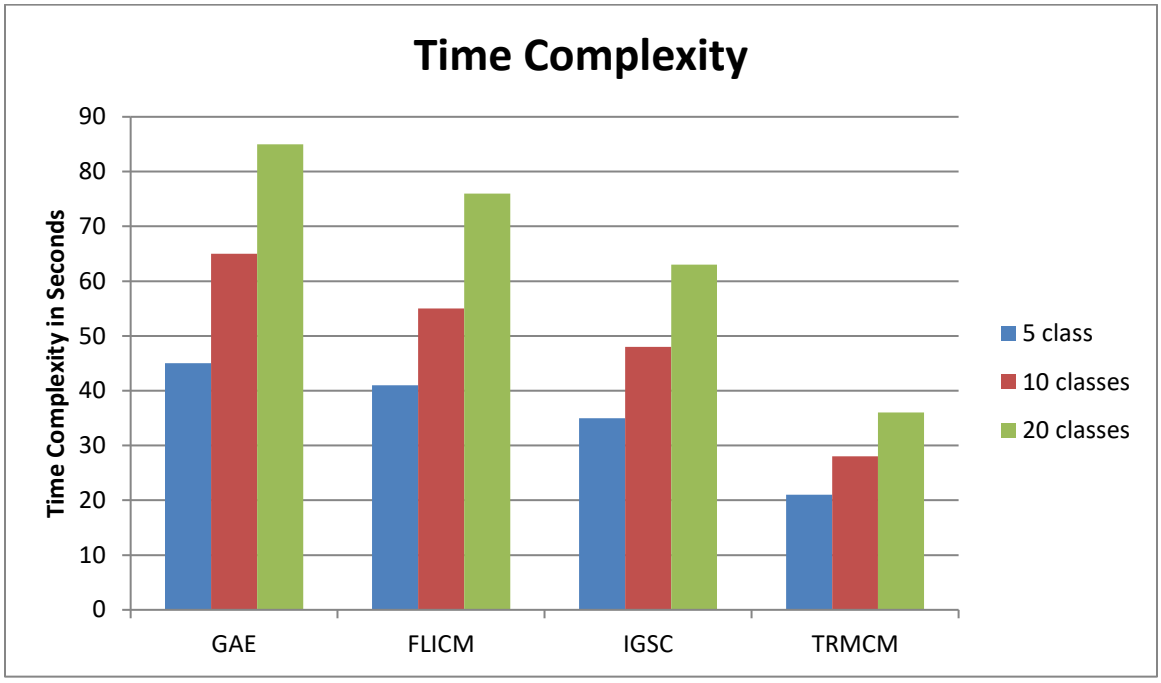


Figure 4: Time complexity

The time complexity on clustering has been measured and presented in figure 4, where TRMCM algorithm introduces less time complexity compare to other methods.

## 5. Conclusion:

This article presented a novel Topical Impact Measure Based clustering model (TRMCM) which preprocess the document for noise removal and feature extraction. Further, the method applies term centric frequency analysis algorithm for feature selection. Third the method applies TRM clustering by computing PRM and CRM values to compute TRM measure. Based on the value of TRM, the method identifies the class of document and indexes them. The proposed TRMCM model improves the clustering accuracy up to 97% with reduced time complexity.

**REFERENCES:**

1. S. Jung and S. Ka, "GAE-Based Document Embedding Method for Clustering," in IEEE Access, Volume. 10, pp. 130089-130096, 2022.
2. Y. Li, W. Wang, M. Liu, Z. Jiang and Q. He, "Speaker Clustering by Co-Optimizing Deep Representation Learning and Cluster Estimation," IEEE (TM), Volume. 23, pp. 3377-3387, 2021.
3. J. Francis, B. Madathil, S. N. George and S. George, "A Comprehensive Tensor Framework for the Clustering of Hyperspectral Paper Data With an Application to Forensic Document Analysis," in IEEE Access, Volume 10, pp. 6194-6207, 2022.
4. S. Guo and N. Yao, "Document Vector Extension for Documents Classification," IEEE (TK&DE), Volume. 33, Number 8, pp. 3062-3074, 2021.

5. F. Malik, S. Khan, A. Rizwan, G. Atteia and N. A. Samee, "A Novel Hybrid Clustering Approach Based on Black Hole Algorithm for Document Clustering," in IEEE Access, Volume. 10, pp. 97310-97326, 2022.
6. Perumal Pitchandi, “Document clustering analysis with aid of adaptive Jaro Winkler with Jellyfish search clustering algorithm”, ELSEVIER (AES), Volume 175, Number 103322, 2023.
7. Li Jngnan, “Intention-guided deep semi-supervised document clustering via metric learning”, Science Direct (JKSU – C&IS), Volume 35, Issue 1, PP 416-425, 2023.
8. Gunjan Chandwai, “Fuzzy Local Information C -Means based clustering and Fractional Dwarf Mongoose optimization enabled deep learning for relevant document retrieval”, ELSEVIER (EAAI), Volume 126, Part B, Number 106954, 2023.
9. Meijing Li, “An Efficient Parallelized Ontology Network-Based Semantic Similarity Measure for Big Biomedical Document Clustering”, HINDAWI (C&MMM), Volume 2021, 2021, doi.org/10.1155/2021/7937573.
10. M. A. Balafar, “Active Learning for Constrained Document Clustering with Uncertainty Region”, HINDAWI ©, Volume 2020, 2020, doi.org/10.1155/2020/3207306.
11. Meng Yuan, “Measurement of clustering effectiveness for document collections”, Springer Link (IRJ), Volume 25, PP 239-268, 2022.
12. R. Laskhmi, S. Baskar, “Efficient text document clustering with new similarity measures”, ACM (IJBI&DM), Volume 18, Issue 1, PP 49-72, 2021.
13. Bambang Krismono Triwijoyo, “Analysis of Document Clustering based on Cosine Similarity and K-Main Algorithms”, ISI (A), Volume 1, Number 2, 2019.
14. Saad Hikmat Haji, “Document Clustering in the Age of Big Data: Incorporating Semantic Information for Improved Results”, JASTT, Volume 4, Number 1, 2023.
15. E.K. Jasila, N. Saleena, “An Efficient Document Clustering Approach for Devising Semantic Clusters”, T&FO (C&S), 2023, doi.org/10.1080/01969722.2023.2175135.