INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

Document Clustering using improved K-means Algorithm

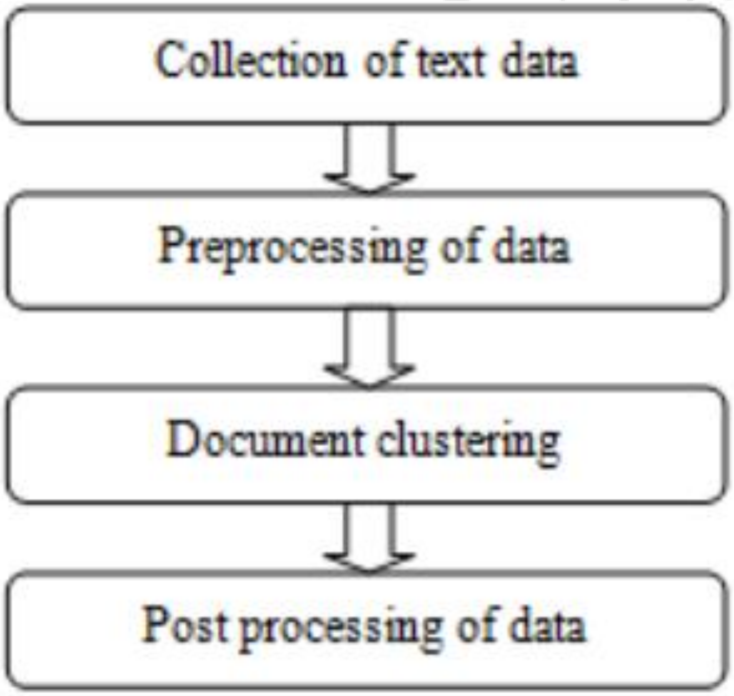
ARYAN BATRA	*21114019*
TUSHAR RAJU CHINCHOLE	*21114032*
MANOJ KUMAR SIAL	*21114059*
SACHIN JANGID	*21114086*

Abstract:

- Clustering is an efficient technique that organizes a large quantity of unordered text documents into a small number of significant and coherent clusters.
- k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters by using centroid.
- But K-means have some limitation
- This paper describe a new algorithm improved k-means clustering which is developed to overcome these limitations .

Stages of Document Clustering:

- Getting relevant data from a collection of documents include following stages:-



Collection of text data

Preprocessing of data

Document clustering

Post processing of data

Collection of text data includes the processes like filtering etc. which are used to collect the documents that need to be clustered.

Pre-Processing of data is done to represent the data in a form that can be used for clustering. There are many ways of representing the documents like, Vector-Model.

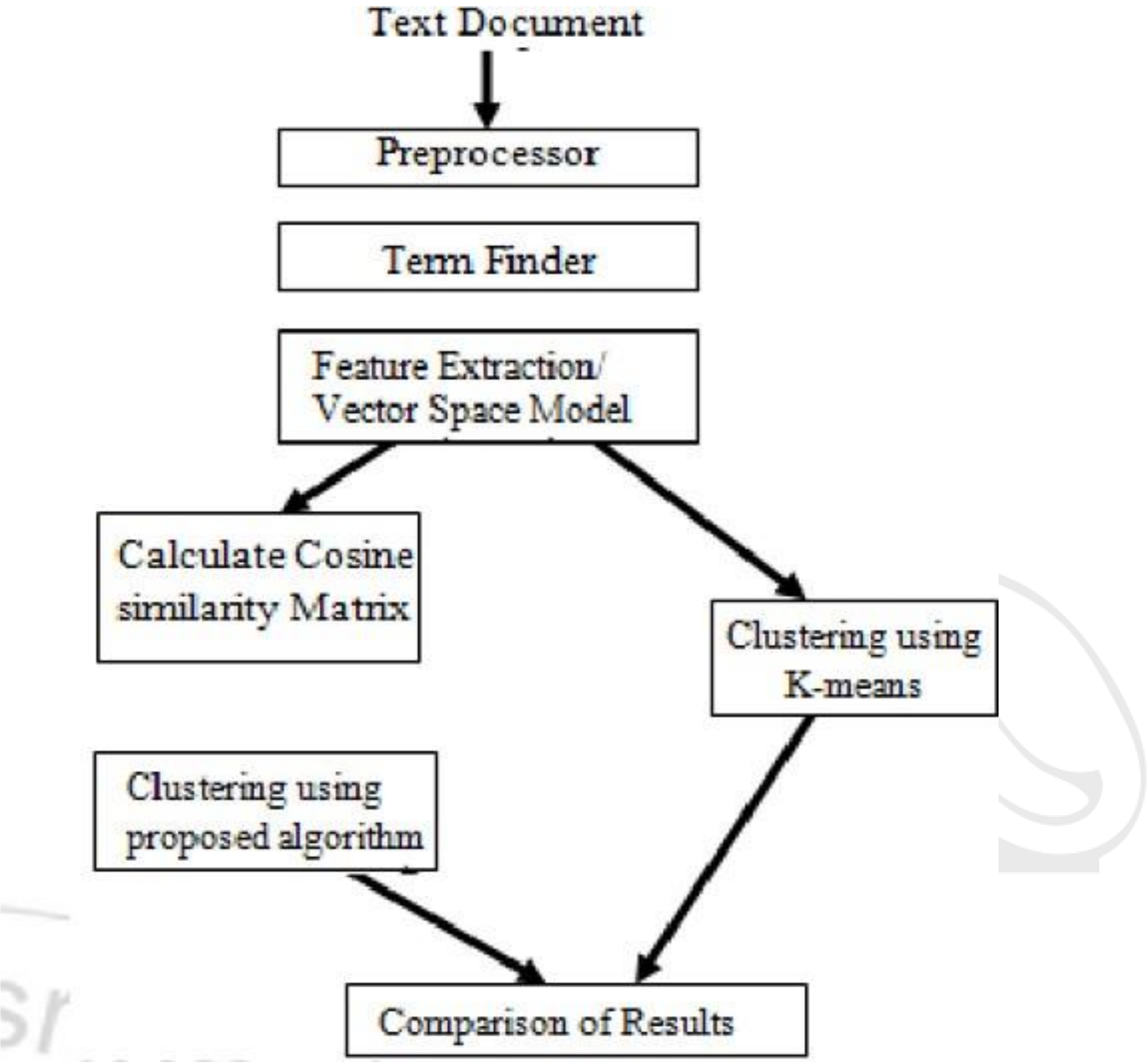
Document clustering is the main focus of this thesis work and will be discussed further.

Post-Processing of data includes the major applications in which the document clustering is used, for.

Limitations in k-means algorithm:

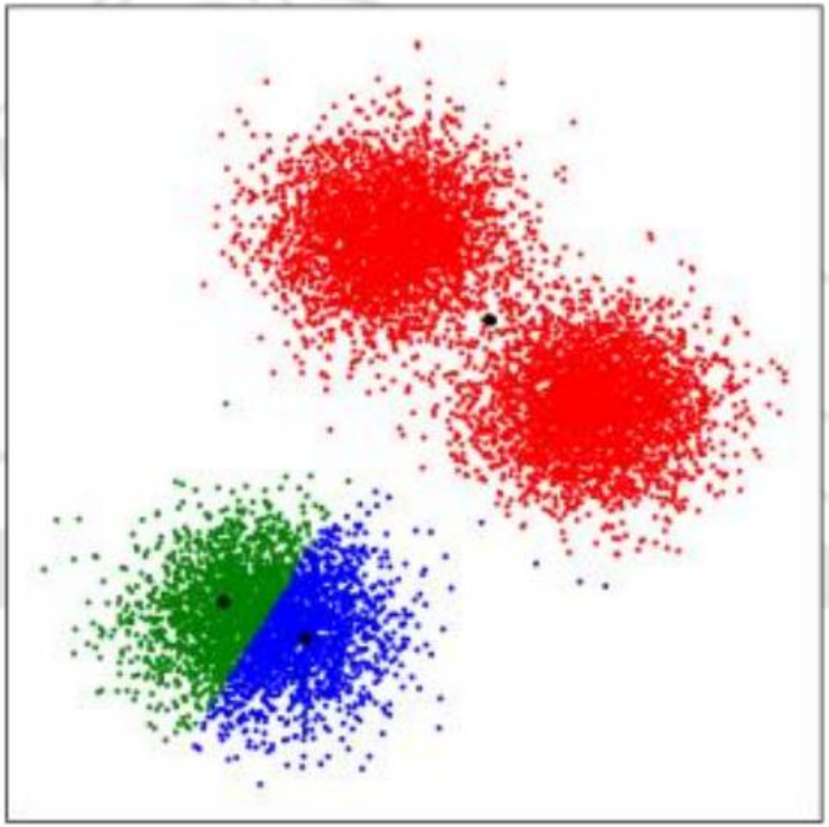
- Sensitive to Initial Centroid Selection: K-means is sensitive to the initial placement of cluster centroids. Different initializations can lead to different final clusters, and it may converge to a local minimum rather than the global minimum.
- It takes nonexclusive words also and do not match the words by semantic basis.
- Requires Euclidean Distance Metric: It relies on the Euclidean distance metric, which may not be suitable for all types of data. For example, categorical or binary data may require a different distance metric.

System Architecture of Proposed System


Text Document
Preprocessor
Term Finder
Feature Extraction/ Vector Space Model
Calculate Cosine similarity Matrix
Clustering using K-means
Clustering using proposed algorithm
Comparison of Results

K-mean disadvantage

It is sensitive to initial condition. Different initial condition may produce different result of clusters. The algorithm may be trapped in the *local optimum.*



Improved K-means

- Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.
- To compute cosine Similarity matrix, we use term frequency vector of the documents.
- Similarity => $\cos(\theta) = \dfrac{A.B}{||A||\,||B||}$
- These matrices are then used as input to K-means and proposed algorithm and clustering is done. Finally results are compared for different parameters like F-measure, time complexity.

Algorithm for Improved K-means

- Input: Dataset set D = {d1, d2 . .. dn}
- Output: Set of Cluster Numbers C along with documents associated.

We will apply improved K- means algorithm on every partition iteratively till we get the same clusters ie until there is insignificant movement of documents across clusters.

1.) Input the number of clusters from user i.e. K (no of clusters).

2.1) Sort the Vector Space Model (VSM) and generate the K parts.

2.2) Take mean of every column (i.e. mean of every part)

2.3) The mean calculated is center of prediction.

3.) Calculate the similarity of the documents using cosine similarity measure.

4.) Assign the nearest (similar) document to the new clusters.

5.) If the clusters are not matched then go to step 3.

6.) If clusters are matched then stop.

Results and Discussion:

Dataset

- Used "20 Newsgroup" English dataset with 20,000 documents.
- Modified version with duplicates and cross-posts removed.

- Details:

1. Number of unique documents: 18,828
2. Number of categories: 20

Result

- Proposed algorithm is faster and uses exclusive words.
- Existing algorithm is slower and lacks semantic matching.
- **Precision (P):** Ratio of relevant documents to the total documents retrieved for a query.
- **Recall :** Ratio of relevant documents retrieved for a query to the total relevant documents in the collection.
- **F1 Score:** Combines precision and recall, providing a balanced measure of performance.

Table 5.1: Quality Comparison of Existing and Proposed System

Results of Existing System		
	CLASS 1	CLASS 2
Cluster 0	54	63
Cluster 1	46	37
Results of Proposed System		
	CLASS 1	CLASS 2
Cluster 0	53	0
Cluster 1	47	100

Table 5.2: Performance Comparison of Existing and Proposed System

Results of Existing System			
	Precision	Recall	F-Measure
Class 1	0.538462	0.63	0.580645
Class 2	0.554217	0.46	0.504732
Results of Proposed System			
	Precision	Recall	F-Measure
Class 1	1	0.53	0.69281
Class 2	0.6802	1	0.8097

Table 5.3: Time Comparison of Existing and Proposed System

	Existing	Proposed
Time (ms)	109	62

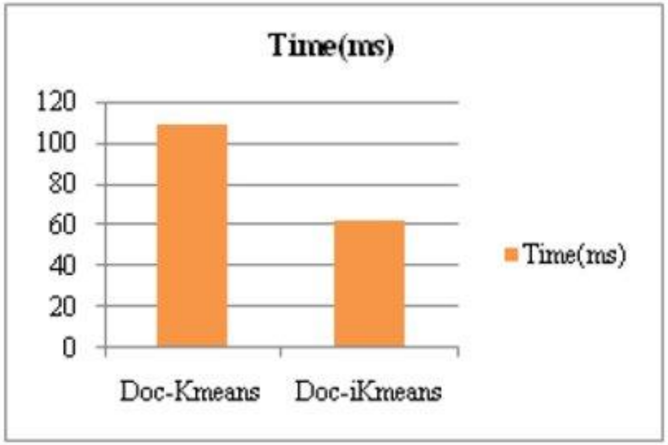


Figure 6: Time comparison between existing k-means and improved k-means

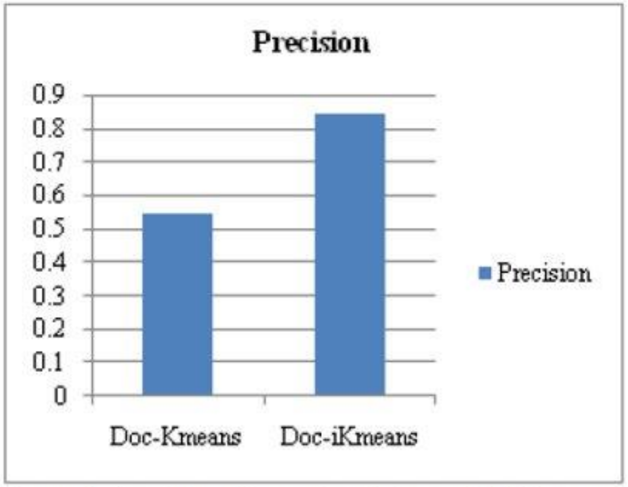


Figure 3: Precision Comparison between K-Means and Improved K-Means

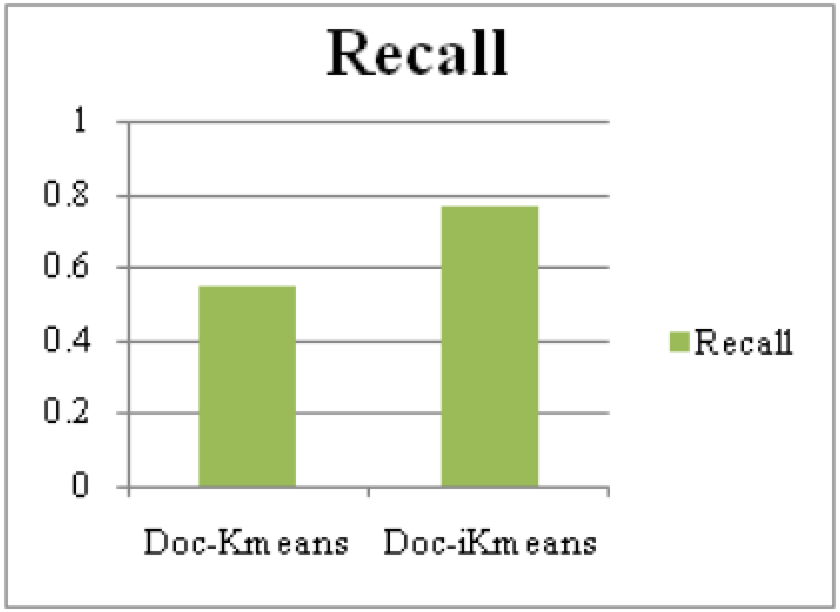


Figure 4: Recall Comparison between Existing K-Means and Improved K-Means

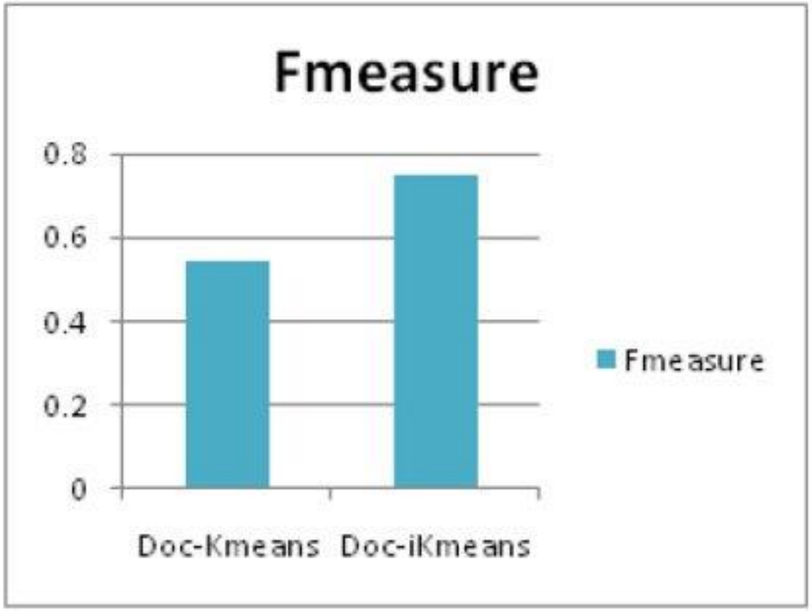


Figure 5: Fmeasure Comparison between Existing K-Means and Improved K-Means

Comparision

Value	K-Means	Improved K-Means
Precision	0.38	0.70
Recall	0.45	0.73
F-measure	0.412	0.714

Conclusions

- Document clustering is crucial for selecting relevant documents from vast collections.
- The proposed clustering process enhances data clusters, providing valuable insights into document content.
- Frequent term-based clustering improves system performance and clustering quality.
- Existing algorithms were investigated, and an improved one was proposed.
- The existing algorithm is slower due to its use of non-exclusive words and lack of semantic document matching.
- The proposed algorithm is faster, using only exclusive words and semantic matching.

Future Scope

- Incorporating keyword searching with document clustering can improve grouping and retrieval efficiency.
- Performance comparisons of similarity measures using various clustering algorithms are potential future research.
- The algorithm presented in this thesis represents an initial improvement; many additional enhancements are possible.

Thanks