Proceedings of the Fifth International Conference on Communication and Electronics Systems (ICCES 2020)
IEEE Conference Record # 48766; IEEE Xplore ISBN: 978-1-7281-5371-1

Text Document Clustering Using K-means Algorithm with Dimension Reduction Techniques

Rutuja Kumbhar[1], Snehal Mhamane[2], Harshada Patil[3], Sukruta Patil[4], Shubhangi Kale[5]
[1,2,3,4,5,]School of Computer Engineering & Technology, MIT Academy of Engineering, Pune, India.
rutu9767@gmail.com, smmhamane15@gmail.com, harshada2705@gmail.com, Email: patil28111998@gmail.com, spkale@comp.maepune.ac.in

Abstract - With growing technologies commercial sites, social media, organizations generate lots of data. However, this huge amount of information needs to be organized properly. For this, the text mining process used as extracting relevant features and knowledge of the given unstructured text documents. The document clustering method in text mining is used for retrieving interesting features. In which similar documents organize into different groups called clusters. The large dimensions of data become a barrier in the extraction of useful information, so using dimensionality reduction (DR) technique for reducing the dimensions of the data matrices. Further data divide into groups using the k-means clustering algorithm. This study implements TF-IDF, singular value decomposition (SVD), non-negative matrix factorization (NMF), and k-means clustering. Finally, the results of the comparison of scores of kmeans, SVD with kmeans, and NMF with k-means are shown by graphical representation. The system uses 20 newsgroup datasets for simulating results.

Keywords - TF-IDF, Truncated SVD, NMF, DR K-means clustering.

I. INTRODUCTION

As new technologies coming day by day, there is an increase in textual data on commercial sites, social media, or in the organizations. If there is no proper way to store all this data it will be irritating or frustrating to extract useful data from the bulk of stored data. So, here the clustering comes into the picture. By using clustering one can group the data into different groups. Clustering comes under unsupervised learning techniques in which the data having similar features are placed into one cluster and dissimilar into other clusters. The similarity or dissimilarity between the textual documents can be calculated using different distance metrics like Euclidean distance, Manhattan distance, or Cosine similarity. Here the system is focusing on Euclidean distance. There are also different types in clustering like hierarchical clustering, density-based clustering or data partitioning. Further the more information about clustering the documents using various techniques are provided in Literature survey. Here, TF-IDF is used for vectorization.

Applying clustering on large dimension data such as newsgroup dataset which required lot amount of time. But some applications this situation is not acceptable, to make application more usable, there is need to reduce computational time. Also, to enhance the performance of clustering techniques by using dimension reduction techniques so that text mining procedures process text documents with reduce number of features. Truncated SVD and NMF used as dimension reduction technique to improve the score of clustering the text documents. SVD that is Singular Value Decomposition it is also called latent semantic analysis in case of textual data. This paper combine SVD with K-means clustering to partition the news. NMF that is Negative Matrix Factorization this is a method for approximating dimensions of large dataset. NMF method reduce the vector space by summarizing documents in dataset. This paper combines the dimension reduction technique with K-means algorithm. First case SVD combine with K-means then in second case NMF combine with K-means algorithm.

K-means clustering belongs to the most popular clustering algorithms. Clusters formation done using the Euclidean distance. In .txt documents clustering, the main problem arises is the length of documents. So, to extract only interesting features here two popular dimensional reduction techniques are used like NMF and SVD.

The paper is arranged in a way such that Introduction in section I followed by a Literature survey in section II, section III includes the proposed system, section IV includes Proposed Methodology, Experimental results are discussed in section V, section VI contains application and conclusion is present in section VII. At last, references are provided.

II. LITERATURE SURVEY

Data mining is areas in text mining include the retrieval of useful information from a large amount of data. In which classification lies under the complex process of forming groups of the interrelated documents. Document clustering approach is used for grouping of documents with more similarity in same cluster, one of the methods of clustering a document is k-means clustering, cluster the documents by considering centroid similarity and cluster similarity [1]. Paper published by Ammar

Ismael Kadhim, Yu-N Cheah, Nurul Hashimah Ahamed, et.al. [2] have presented document clustering is method of retrieving relevant information form pre-process text. The pre-process text has a huge effect on an extracting knowledge. Number of documents are directly proportional to increase in complexity of finding the document. Therefore, clustering algorithms are not able to handle the large dimensionality of data. For solving this problem technique like DR come into picture. Authors shows reducing difficult using effective process i.e., pre-processing and SVD with k-means clustering. Results are simulated on BBC news and BBC sports dataset; in outcome the accuracy is 95% and 94.6667%. Which shows that proposed method enhance of text document clustering of English texts.

Prafulla Bafna, et.al [3] has proposed the system that shows the importance of TF-IDF in document clustering. Author's has performed experiments for different datasets i.e., 20 Newsgroup, Reuters, emails and different research papers to analysis the performance of TF-IDF using fuzzy kmeans and hierarchal agglomerative clustering algorithms. Among both algorithms, the algorithm having highest entropy and f-measures scores is used to evaluate results of large datasets. In first experiment with research paper dataset Fuzzy K-means give F-measure up to 0.8845. And HAF used on News 20 give entropy and F-measure 0.2562 and 0.8612 respectively. In [4] paper, the comparison of result of K-Means and kmeans with DR technique is done. Where BBC sports dataset is use to analysis the result. DR technique is of three types – Infogain, Best first search and Greedy Stepwise. DR techniques reduces the dimension of TF-ID matrix. K-means with Infogain (IG) DR technique is more effective than K- Means clustering without DR techniques. The K-means clustering with Infogain (IG) DR technique has 97.8% precision, 96.4% recall, 96.7% accuracy and 97% F-measure.

Improved k-means algorithm is used to form groups of text documents. Author's used Euclidean distance formula for similarity measure so that similar types of documents in appropriate clusters. Here the value of F-measure, precision and recall is calculated for each algorithm. F-measures value id greater for new algorithm compared to existing one. Same in case of recall and precision which is 0.54474 and 0.47201 in case of k-means, in case of improved k-means the values are 0.8177 and 0.82 respectively. And also, the existing algorithm take more time than proposed algorithm. The algorithm is tested on 300 documents of mini_newsgroup dataset [5]. Wei Xu, Xin Liu, Yihong Gong et.al [6] has been taken novel dataset. Every axis lay content of the document. Every document represents as multiple basic topics. Cluster of each documents can find basic topics based on largest projection values. Their method is different from SVD with latent semantic indexing (LSI). Latent semantic space is derived by NMF does not require to be orthogonal. Method take non- negative value. The performance ranking is in the order of AA, NMF, NC and NMF-NCW as it contrast the accuracy in each case which is 75%, 83%, 79% and 89% respectively. FCDC (Frequent Concept Document Clustering) is method only focuses on frequent concept and not on frequent item sets. Algorithm uses the approach of semantic relationship among the words. Using WordNet Ontology, low dimensional feature vectors are created which are helpful to create more accurate clustering algorithm. In FCDC, concept is the set of synonyms in which first it finds concepts and then finds the frequent concepts using apriority paradigm. Initial clusters are processed by finding score function, inter-similarity value and disjoint clusters are created. Finally, the results are represented by Tree-Like Structure by applying pruning and sibling merging on disjoint clusters. From the results it is conclude that FCDC performed well in terms of accuracy as compared to another algorithms and it is also less sensitive to the number of datasets [7].

Vivek Kumar Singh, Nisha Tiwari, Shekhar Garg et.al [8] applied 3 different algorithms on 3 different datasets and analysed the results by using different types of vectorization and pre-processing techniques on datasets. The dataset like 20-newsgroup, Reuters-21578 and classic 2095 are used by authors. From these datasets some documents are selected for further process. After this they have done pre-processing on these datasets. In pre-processing they removed stop words and also stemming is done on all the dataset. Stemming and stop words removal task remove the term count in the documents, for perfect clustering. After pre-processing the documents converted into vector format using TF, TF-IDF, Boolean techniques. By using this all documents are converted into vectors and used for clustering. After vectorization K-means, Heuristic K-Means and Fuzzy. In [9] paper shows the comparative results of detecting topic of tweets on twitter using k-means clustering with and without dimension reduction technique. The authors form the Topic Detection and Tracking (TDT) system will help end users to detect the topic automatically by providing tweets as text documents. To get exact cluster, tweets documents are pre-processed and TF-IDF scores are provided to Latent Semantic Analysis (LSA). LSA is used as a DR method to reduce the document matrices with the help of Truncated SVD. The dimension of TF-IDF matrices are reduced using truncated SVD and this document are provided for clustering. Accuracy and computation time of detecting tweets using K-means clustering with truncated SVD is more comparable to K-means clustering without truncated SVD. The results show that combination method gives relative accuracy in terms of topic recall, keyword recall and precision of keyword. With FA Cup dataset topic recall, key precision and key recall are 89%, 22%, 52% respectively. Similarly, in case of Super Tuesday dataset topic recall, key precision and key recall are 23%, 40%, 69% respectively.

The paper written by Aysun Güran, Murat Can Ganiz, et.al [10] proposed K-means clustering on two types of Turkish dataset. Here, Text Summarization technique helps in reducing dataset size. Text Summarization process comes under pre-processing part of the dataset. Author used Dimension reduction technique

to convert original vectors space into low dimensional vector space before applying K-means clustering on Turkish datasets.
Among all this paper it is observed that the dimension reduction techniques that are SVD and NMF gives better result compare to K-means algorithm for forming the clusters, in this project, consider SVD and NMF method to reduces the dimension of the matrices obtained by the TF-IDF vectorization. To reduce the noise in the system, pre-processing is done. Therefore, SVD and NMF along with k-means comparison are plotted as the result.
In [11] paper Flow of the process given in this paper is as basic K-means algorithm, Residual Sum of Squares, Termination Condition, Bad choice of initial seed, TF-IDF, TF-IDF stands for term frequency-inverse document frequency. It is a numerical statistic which reflects how important a word is to a document in a collection or corpus. Most common weighting method used to describe documents in the Vector Space Model, particularly on IR problems. Algorithm was run on three different scenarios first is document vectors formed using features(words) then document vectors formed using sub-category of features last one is Documents vectors formed using parent category of features. Mohammadreza Shams, Zeinab Shamaee, Amir mehdi ghazifard et.al [12] used new method for text clustering, by constructing a term correlation graph, and then extracting topic word sets fromit and finally, categorizing each document to its related topic with the help of a classification algorithm like SVM. This method provides a natural and understandable description for clusters by their topic word sets, and it also enables us to decide the cluster of documents only when needed and in a parallel fashion. Hamshahri contains more than 160,000 newspaper articles, labelled with 19 categorizes. INEX: contains more than 12,000 scientific articles, collected from 18 different IEEE journals. According to results proposed algorithm shows better result than traditional K-means. That is less entropy values showing clusters with more purity.

III. PROPOSED SYSTEM

Text document clustering includes Clustering of data. Preprocessing is done to remove the noise for dataset the algorithm for clustering is Kmeans. Finally, after calculating TF-IDF score the two different DR techniques used are SVD and NMF. The model includes comparison of algorithms of getting maximum accuracy.

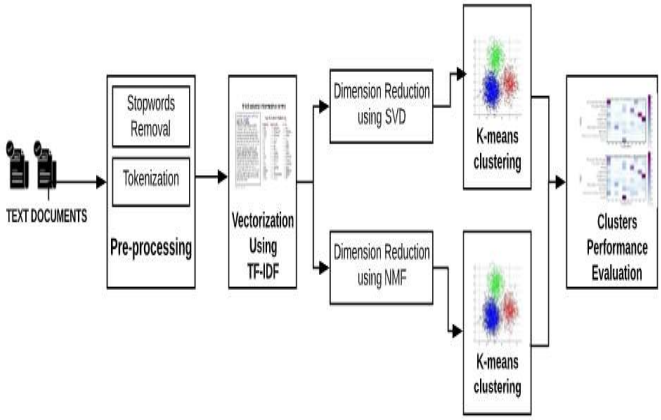

Fig. 1. Proposed System Architecture

IV. PROPOSED METHODOLOGY

A. Input

Text documents are given as the input to the system. 20-Newsgroup dataset is taken from the Kaggle. 20 Newsgroup data set is collection of around 20,000 newsgroup documents, partitioned into 20 different newsgroups. The data is organized into 20 different newsgroups, each of them is corresponds to different topic. Some of them are closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while other are highly unrelated to each other (e.g. misc.forsale / soc.religion.christian). This documents are provided as input to pre-processing phase.

B. Text Pre-Processing

Objective of text pre-processing is to remove the unwanted noise from the collected data. It divides the text document into features called tokens, it represents the document as vector space model with features and their weight. The features which are not important are removed. Still feature spaces has the high dimensionality, therefore; in pre-processing applying specific threshold minimize feature space for each input document based on their count of each terms.

i. Tokenization

Over tokenization raw text documents are get partitioned into tokens the tree operations need to be considered: first is convert document into count of words. The second operation is comprising of cleaning and filtering in which empty sequences get removed. In third operation input document in segmented into features called tokens.

ii. Stopwords Removal

Usually the concurrent words in many of languages, also called as common features. Common features include words like the, is, an etc. need to be remove because this word add no or little value in categorization process. In system, English related stop words are removed natural language processing approach stop words among documents get removed.

C. TF-IDF Weighting

After pre-processing, each word has its own importance, therefore the importance of words in documents and even in dataset is calculate using TF-IDF method.TF-IDF stands for "Term Frequency – Inverse Document Frequency". Term frequency this summarize how often the word appears in document and inverse document frequency downscales words that appear a lot across documents. Finally, it calculates the IF-IDF value for each feature considering each document.

$$TF - IDF(t, f) = tf(t, d) X idf(t) \quad (1)$$

Tf t, d) = count of t in d / number of words in d
Idf(t) = occurrence of t in documents

$$idf(t) = log\left(\frac{n}{1 + df(t))}\right) \quad (2)$$

t - term, d - document

Vector Space Model

After using TF-IDF function, system stores the values of each terms. In this module, considering the document-term matrix, column represents TF-IDF scores of each terms and row represents documents. But the matrices got as output is of higher dimension, which drop the performance to cluster the data. Clustering algorithm can handle data of limited length, because processing of high dimensional data is expensive it requires more time and storing capacity. For reduction of data size, DR approach is used. By using one of the dimension reduction technique reduces the size of vector space model. DR techniques are truncated SVD, Principle Component Analysis (PCA) and independent Component Analysis (ICA).

D. Dimension Reduction (DR)

DR is famous methods to shrink the document dimensions, so that the performance of the clustering technique can be improved. Using SVD and NMF to increase the performance of forming the groups by reducing the dimensions of the TF-IDF matrices.

i. Singular Value Decomposition (SVD)

Truncated SVD is the type of dimension reduction in which the dimension of the TF-IDF matrix is reduced. The input of TF-IDF is given to the truncated SVD, as follow:

$$A_{mXn} = U_{mXr} \cdot \varepsilon_{rXr} \cdot (V_{nXr})^T \quad (3)$$

Where, A is input matrix of order m x n, here m represents documents and n is terms presents in those documents.
U is the left singular vector of m x r order. In which
represent documents and r represent concepts. ∈ is the diagonal matrix of r x r order, where r represents the concepts. The diagonal values of this diagonal matrix are positive and other values are zero. V matrix is the right singular matrix represented by n x r order, where n represent terms and r represents concepts.

After getting V, the transpose of V is obtained (V^T). U and V^T are the orthogonal matrices.
Latent semantics analysis (LSA) is used for finding the relation between the terms and the documents. LSA takes the three matrices (U, ∈,V^T) obtained from SVD, where it does:

$$A = \cup \cdot \in \cdot V^T \quad (4)$$

$$S = \cup \cdot \in \quad (5)$$

$$A = S \cdot V^T \quad (6)$$

The vector is scaled into S by multiplying orthogonal matrix and the diagonal matrix. Thus, LSA use approach of SVD to fallen off the matrix's dimensions.

ii. Non-Negative Matrix Factorization

NMF is DR reduction techniques. Algorithm used to reduce the dimension of the matrices obtained from TF-IDF vectors. The positive values are representing in NMF, where negative values get ignored. Considering provided document with cluster k=2. Goal factorizing X in 2 non-negative matrices such as U & V that minimizes objective function as:

$$J = \frac{1}{2} \| X - UV^T \| \quad (7)$$

After deriving or solving this equation, the formulas are updated as:

$$u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \quad (8)$$

$$v_{ij} \leftarrow v_{ij} \frac{(X^TU)_{ij}}{(VU^TU)_{ij}} \quad (9)$$

Here unique solution for minimizing value of J is not obtained. If U and V are the solution to J then UD, VD^{-1}. Following Euclidean distance for column vector in matrix U to make unique solution that shown below:

$$u_{ij} \leftarrow u_{ij} \sqrt{\sum_i u_{ij}^2} \quad (10)$$

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \quad (11)$$

Each element i.e., matrix U having U_{ij} represent degree to term f_i ∈ W belonging to cluster j, V matrix having V_{ij} represents degree document i that belongs to cluster j. and if the documents I may belongs to cluster x, then v_{ix} will take large value while
rests of values are close to zero.

i. Create term document matrix X, where weighted term frequency is represented by column I for given document.
ii. Perform NMF on X for obtaining U and V are non-negative matrices using updating formulas SA.
iii. Normalize U and V using above equations.
iv. Use each data of matrix V for determining the labels of the clusters. It examines each row i of matrix V

v. Assign document d_i to cluster x if i'th row vector has maximum value.

E. K-means Algorithm

K-means is one of famous clustering algorithm. It groups the information into number of clusters, which recursively update to get accurate partitions. Kmeans algorithm use two steps i.e., assignment and updating. Euclidean similarity is used to clump data and it is known as assignment step. Partition obtained by previous state help to update the clump and this is Updating step of kmeans. Also, to cluster all related data there is need to find the centroid. Therefore, using kmeans centroid is obtained by various method, here consider the mean of the vector. The algorithm begins with predefine set of centroids. There is continuous update on centroid is happening by using assignment and updates steps.

K-means clustering implement as mentioned below:
Step 1: Setting the value for k. Where k is the count of clusters need to be considered.
Step 2: After that defining the centroid value, this is done using randomization method included in algorithm.

Scores	*K-means*	*SVD k-means*	*NMF, k-means*
Homogeneity_score	0.248	0.771	0.786
Completeness_score	0.331	0.772	0.818
Adjusted_rand_score	0.174	0.855	0.876
Adjusted_mutual_info_score	0.248	0.771	0.782

Step 3: Find relation among category centroids and remaining elements.
Step 3.1: Compute Euclidean Similarity between centroids and elements.
Step 3.2: Updating way is used to update the centroids. Since, the centroid is modified using the minimum distance between element and centroid.
Step 4: All feasible assignments are to be taken care while updating centroids.
Step 5: After that computing distance for all the cases.
Step 6: Obtain best assignment (By analysing the partitions).
Step 7: Return step 3, hence stop when no new assignment.

V. EXPERIMENTAL RESULTS

i. Dataset

The first collect 20-newsgroups dataset from Kaggle corresponding to news in six different areas. The dataset consists of 20 folders and each folder contain nearly 1000 text documents. Data set divided into 6 section named computer and electronic, sports, science medicine, religion, politics, automobile and e-commerce

ii. Performance Evaluation

System uses same platform dataset but with different sizes of text documents. System plots confusion matrix based on number of clusters or classes. System calculates homogeneity score, completeness score, adjusted rand score, and adjusted mutual information score for k- means, Singular vector decomposition and Nonnegative matrix Factorization. The value of homogeneity score is always between 0.0 and 1.0. Perfectly homogeneity labelling is represented by 1.0. Homogeneity score calculated as follows,

$$h = 1 - H(Y_{true} \mid Y_{pred}) / H(Y_{true}) \quad (1)$$

A clustering said to be satisfied completeness if the object of same class belongs to the same cluster. Formula to calculated completeness score is,

$$c = 1 - H(Y_{pred} \mid Y_{true}) / H(Y_{pred}) \quad (1)$$

Where,
Y_{pred} = Predicted label
Y_{true} = Actual label

Adjusted rand score it measures similarity between two clustering's by considering all pairs of samples and counting pairs assigned in same or different clusters in predicted and true clustering. Given formula used to find adjusted rand score.

$$ARI = (RI - Expected\_RI) / (max(RI) - Expected\_RI) \quad (1)$$

Where,
RI = Rand Index
Adjusted mutual information score is an adjustment in mutual information score.

Table I: Clustering Performance Measures, K = 2

Table II: Clustering Performance Measures, K = 4

Scores	*K-means*	*SVD k-means*	*NMF, k-means*
Homogeneity_score	0.163	0.225	0.236
Completeness_score	0.171	0.218	0.228
Adjusted_rand_score	0.095	0.229	0.256
Adjusted_mutual_info_score	0.163	0.218	0.222

Table III: Clustering Performance Measures, K = 6

Scores	*K-means*	*SVD k-means*	*NMF, k-means*
Homogeneity_score	0125	0.213	0.209
Completeness_score	0.163	0.231	0.206
Adjusted_rand_score	0.037	0.119	0.152
Adjusted_mutual_info_score	0.125	0.213	0.2

Based on above score, system plot a histogram to show the comparison between between us k-means, Singular vector decomposition and Nonnegative matrix Factorization.

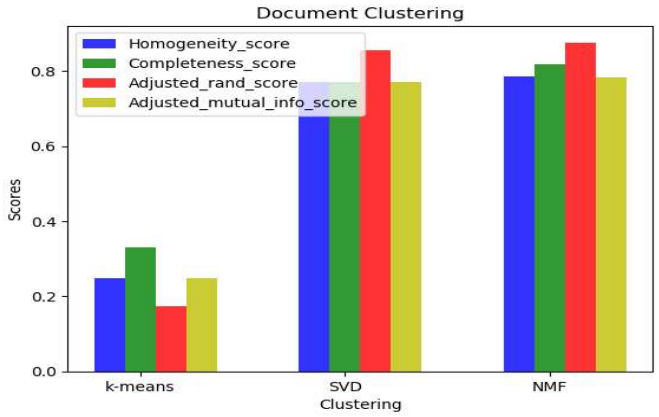

Fig. 2. Histogram when k = 2

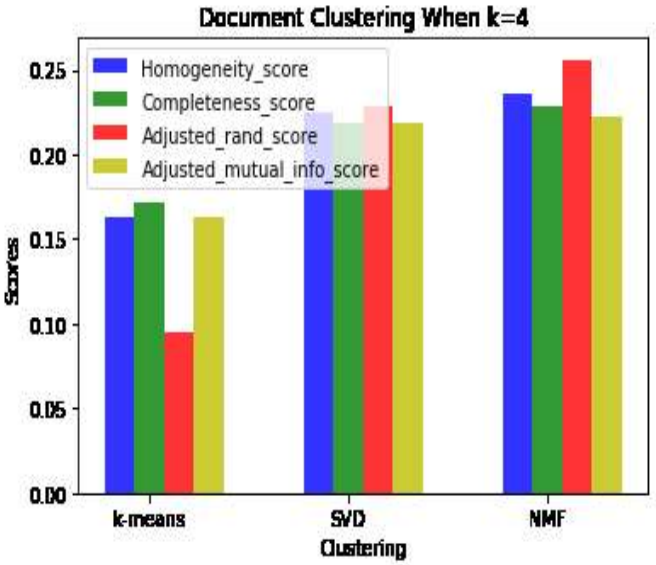

Fig. 3. Histogram when k = 4

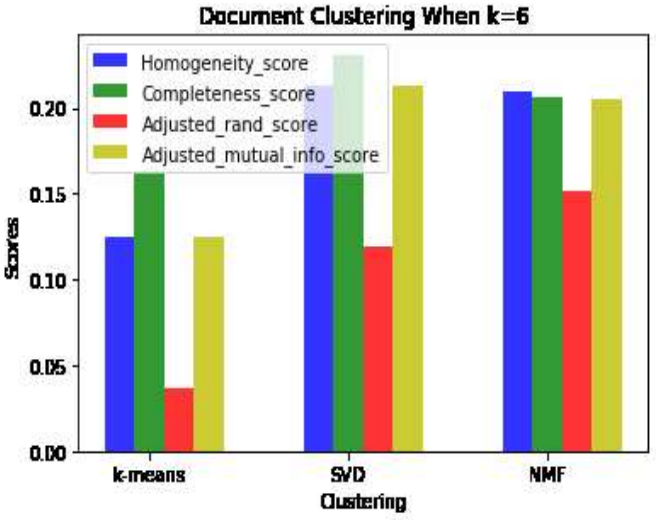

Fig. 4. Histogram when k = 6

It is Observed that k-means and SVD having less scores comparing NMF as shown in fig-2, fig-3 and fig-4 respectively. NMF has higher accuracy to get the proper clusters.

VI. APPLICATION

The most common application of clustering is to gather the data and find all possible meaning to data from huge collection of data. Text Document Clustering System can be used in Duplicate content detection, Customer relationship management system, Personal assistant agent, Recommendation system, Search optimization, organizing large document collections, finding similar documents.

VII. CONCLUSION

As technologies increases, commercial sites, social media, organization or school had large amount of information stores randomly, to collect specific information from large data that require text document clustering. System done text mining on 20 newsgroup datasets for clustering. To improve the result, used Non-negative Matrix Factorization and Singular Vector Decomposition. After that system compared results of NMF-k-means and SVD k-means get more accuracy. It takes information in large scale and provide the results in understandable form to the user.

REFERENCES

[1] Laxmi Lydia, P.Govindaswamy, SK.Lakshmanaprabu, D.Ramya , "Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity ", Jour of Adv Research in Dynamical & Control Systems, Volume 10, 02-Special Issue, 2018
[2] Ammar Ismael Kadhim, Yu-N Cheah, Nurul Hashimah Ahamed, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering", 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, 2014
[3] Prafulla Bafna, Dhanya Pramod and Anagha Vaidya, "Document Clustering: TF- IDF approach", International Conferences on Electrical, Electronics, and Optimization Techniques,2016
[4] Dr. A. Sudha Ramkumar, R.Nethravathy,"Text document clustering using k-means algorithm," International Research Journal of Engineering and Technology Volume: 06 Issue: 06,June 2019.
[5] Shreyatakhatri,Dr. Kanwal Garg, "Document Clustering Using Improved K-Means Algorithm", International Journal of Engineering Research and General Science Volume4, Issue: 3, May-June, 2016.
[6] Xu, Xin Liu, Yihong Gong, "Document Clustering Based on Non-negative Matrix Factorization", NEC Laboratories America, Inc. 10080 North Wolfe Road, SW3-350 Cupertino, CA 95014, U.S.A.
[7] Rekha Baghel, Renu Dhir, "Text Document Clustering Based on Frequent Concepts", IEEE-978- 1-4244-7674-9/10, 2010
[8] Vivek Kumar Singh, Nisha Tiwari, Shekhar Garg, "Document Clustering Using K-Means, Heuristic K-Means, Fuzzy C-Means", International Conference on Computational Intelligence and Communication Systems (ICCICS) ,2011
[9] Khumaisa Nuraini, Ibtisami Najahaty, Lina Hidayati, Hendri Murfi and Siti Nurrohmah, "Combination of Singular Value Decomposition and K-means

Clustering Methods for Topic Detection on Twitter", IEEE -978-1-5090-0363-1/15, 2015
[10] Aysun Güran, Murat Can Ganiz, Hamit Selahattin Naiboğlu, Halil Oğuz Kaptıkaçt, "NMF based Dimension Reduction Methods for Turkish Text Clustering", Doğuş University Acıbadem, Kadıköy, 34722, Istanbul, Turkey
[11] Sanjivani Tushar Deokar., "Text Documents clustering using K Means Algorithm", International Journal of Technology and Engineering Science [IJTES] TM Volume 1 (4), pp 282 – 286, July 2013 ISSN
[12] Mohammadreza Shams, Zeinab Shamaee, Amir mehdi ghazifard, "Topic Word Set-Based Text Clustering", 7th international conference, University Iran

AUTHORS PROFILE

Rutuja Kumbhar is a Final Year B. Tech Student. She is pursuing her B. Tech degree for MIT Academy of Engineering in the field of Computer Engineering, Pune. She done project in Database Management System like Vehicle services Management system. She also has done the projects in Machine Learning Fields like Movie Rating Prediction, Face Express Recognition and Handwritten Character Recognition. She has been worked as project intern in PHP institution, where she completed training for Development of Mobile Application Training in Android. She also done Internship in Arin Software Solution and Networking in Java Field. She has interest in Field of Artificial Intelligence and Android Application Development.

Snehal Mhamane is a student in School of Computer Engineering and Technology of MIT Academy of Engineering, Alandi (D). She is in Final year of B. Tech Computer Science and Engineering in MIT. Her interest area is Machine Learning, Deep Learning. She had done the projects in Machine Learning, the titles of projects are Handwritten Character Recognition System, Movie Rating Prediction System. She had worked as intern in Arin Software and Networking Solutions which is start up in Second year and Third year for 2 months. She has also done some project in IOT field, title of the project is Smart Dustbin.

Harshada Patil student of School of Computer Engineering and technology form MIT academy of engineering Alandi(D). She is in final year BTech. Her area of interest are Web development, Artificial Intelligence, Computer Security . She had done project in web development like Food Delivery System, Social Networking Platform for Kids. She also had projects in Artificial Intelligence, the titles of projects are Handwritten Character Recognition System, Personality Prediction System, trafic sign prediction system, emotions recognition system. She has also done some project in IOT field, title of the project is Smart Dustbin, Automatic Water Supply for plants. In Second year she worked as intern in embedded technology and in Third year She had worked as intern in Arin Software and Networking Solutions which is startup for 2 months.

Sukruta Nivruti Patil is a last year Btech computer Engineering student from MIT Academy of Engineering, Alandi (D.). Her area of interest is Machine Learning, AINN. She had done the projects based on AINN and Machine Learning such as flat rent prediction system, Personality Prediction system, Emotion recognition system, Devnagri handwritten recognition system. She has done 2 internships in Embedded Technosolutions, alumini, mumbai based on robotics and Arin Software and Networking Solutions worked on one java project as online billing system. Based on the knowledge of robotics internship, she has done IOT project as Smart Dustbin Monitoring System.

Shubhangi Kale is working as Assistant Professor in School of Computer Engineering and Technology of MIT Academy of Engineering, Alandi (D.). She has completed M.E. in Computer Science and Engineering from MIT, BAMU, Maharashtra, India in 2014 and B.E. in Computer Engineering from NDMVP, University of Pune, Maharashtra, India in 2005. Her research of interest area is Database and Data Mining and Machine Learning. She taught various subjects such as Database Management Systems, Advanced Databases, Business Intelligence and Data Mining, Design and Analysis of Algorithms, Descriptive Analytics and Predictive Analytics. She has guided various UG projects related to data mining and machine learning.