Text Mining with Hadoop:Enforcement of Document Clustering using Non-Negative Matrix Factorization KNMF

E. Laxmi Lydia, K.Vijaya Kumar, K. Shankar

Abstract*–Big data is recognized as information coming from many sources with an innovative analysis of information.The data in documents are mostly unstructured data such as text processing documents, audio, webpage, log results, etc. Problem Statement: To Order these files manually in folders, it is essential to know the entire contents of the files and the name of the files in order to process files,so that certain files are aligned as a lot. Another characteristic of this information is that it is prone to continuous change, hence clustering is required. Existing approach: uses Latent SemanticIndexing(LSI),Single value decomposition for unstructured document which was quickly filtered and viewed, but it is much harder tocomprehend for computer machines. Proposed approach: A prototype is prepared by deducting redundancy structures to organize the data by similarity, NMF's updated rules along with k-means are proposed in this paper which is used to find the top terms in a respective cluster. For the purposes of exploration, anew data set called Newsgroup20 is considered. To accomplish this,preprocessing steps like Documents indexing, removal of stop words,Stemming.In specific, the words of the text document must be identified for the extraction of key features. The actual work was distributed in parallel with all documents in this project here, Apache Hadoop Map reduce was used for parallel programming.

***Keywords* – Big Data, Hadoop, LSI, Newsgroup20, NMF, , SVD.**

I. INTRODUCTION

. Big data[1] is a termthat portrays an expansive volume of structured, semi-structured and unstructured data that can be extracted forinformation purposes and utilized in Artificial intelligence projects and other progressed analytics.

Big data contains a massive variety of data types, including structured data such as SQL databases and data warehouses, whereas unstructured [2] data, for example, text and archive document in Hadoop clusters,orNotOnly SQL (NoSQL) systems. The data types are shown in the below diagram.

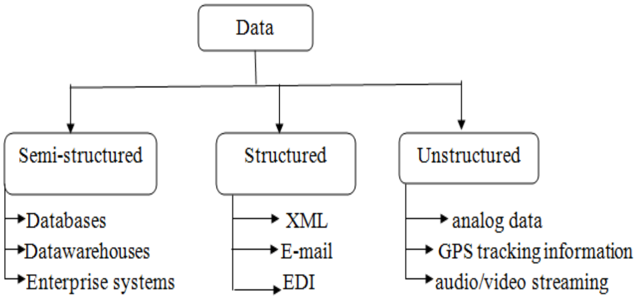


Figure 1Different types of data

Almost all information on the planet is stored in the form of unstructured textual format [21]. In spite of the fact that procedures, for example, Natural Language Processing (NLP) can achieve constrained content investigation, there are as of now no PC programs accessible to investigate and translate content for assorted data extraction needs.

Now a day's technology is growing day by day. The information which we want is being gathered into very large datasets. For example, the internet contains a huge amount of online text documents which are quickly increasing day by day.It is impossible to manually get useful or relevant information from that large datasets. Hence,to extract valuable and appropriate information from such large data sets has guide significant need to develop computationally well-organized text mining algorithms [3].

A. Text Mining

Text and data mining (TDM) is the gadget-read material recovery technique. It works by duplicating an immense amount of material, trying to extract data, and rejoiningpatterns to detect. The process of text mining is shown in below figure.

Revised Manuscript Received on 30 May 2019.
* Correspondence Author

E. Laxmi Lydia, Department of Computer Science Engineering, Vignan's Institute of InformationTechnology(Autonomous), Visakhapatnam, India.

K. Vijaya Kumar, Department of Computer Science Engineering, Vignan's Institute of Engineering for Women, Andhra Pradesh, India.

K. Shankar, School of Computing,Kalasalingam Academy of Research and Education,Krishnankoil, India.

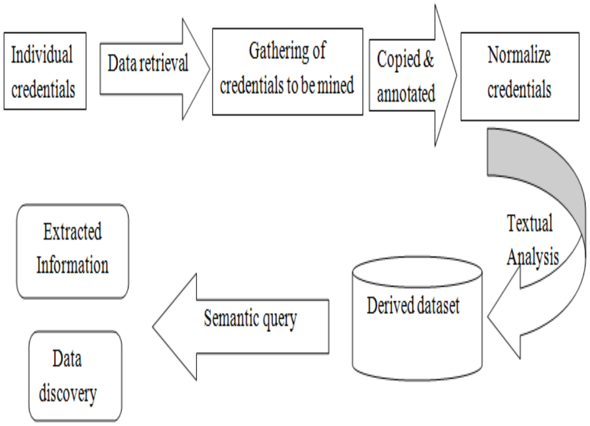


Figure 2 Procedure of text mining

The above figure involves 4 steps

- First, all significant credentials are recognized.
- These individual credentials are then converted into a gadget-readable design so that structured data can be extracted which is important and needful information
- Finally, data are mined to find out novel information, test hypothesis, and recognizethelatentassociations.

The most important format forgadget file organization is to insert them into foldersand place the folders in some folders of the highest level. To hand-place all these data in folders, files data information is required. Normally the name of a record is adequate to give anidea of the substance of the documents as needs be to which the documents can be assembled. There are some situations in which it becomes hard to physically group the files, for example when they are in huge numberwhen their contents can't be renowned from their names. Because of these difficulties, there is a need for computer-aided clustering of the documents.

In our project, clustering of documents takes place by using a combination of k-means and NMF. Lee and Snug's [4, 5] bring up to date set of laws for NMF. Because of these improved rules of NMF, we get better performance than Latent Semantic Indexing (LSI) with Singular ValueDecomposition (SVD).

Numerousauthors areimplementing with well-organized calculation or computation but the performance is low. In this paper, a novel effective model dependent on Lee and Sung's [5] update rules for NMF is proposed with help utilization of k-means for mechanical archive clustering with an application created for its usage. The Newsgroup20 dataset was used for the testing in this paper.In NMF, there is some pre-processing as follows

- Removal of stop words utilizing keywords which utilizesa key phrase extraction algorithm[6]
- Stemming is done by using theproposed iterated porter algorithm[7]
- The Word count is taken as key value format
- Next, tf-idf(term frequency-inverse document frequency) should be found which takes key-values and gives weights for terms in documents. Tf-idf is also called a "document term matrix".
- Lastly, to know the performance, parallel usage of k-implies of k-means clustering algorithm [8] has been utilized.

B) Overview of Apache Hadoop

Hadoop**:**

It is a purely free sourceprogramming framework based on Java that promotes the dispensation of a huge database in a shared computer environment. For processing and sharing of information, Hadoop makes use of low cost, customary servers in the business. The important characteristics of Hadoop are expense powerful model, adjustability, simultaneous processing of shared information, optimization of local information, mechanical failover organization and subsidies more clusters of nodes.

Working of Hadoop

The structure of Hadoop consists of mainly two important components such as HDFS and MapReduce framework. Hadoop structure separates the data into minor chunks and kepteach and everypart of the data in the specific node under the cluster. Because of this, the time to store the data in the frame is decreased. To provide the data at any situations, Hadoop duplicates each and every part of data on to another gadget that isthere inside the cluster. The number of duplicate copies depends on the replication factor. The benefit of sharing this data throughout the cluster is that it reduces a lot of time during the processing of the data since these data can be processed simultaneously. The Figure below shows the Hadoop functioning model for 4TB of data in 4 nodes of the Hadoop cluster.

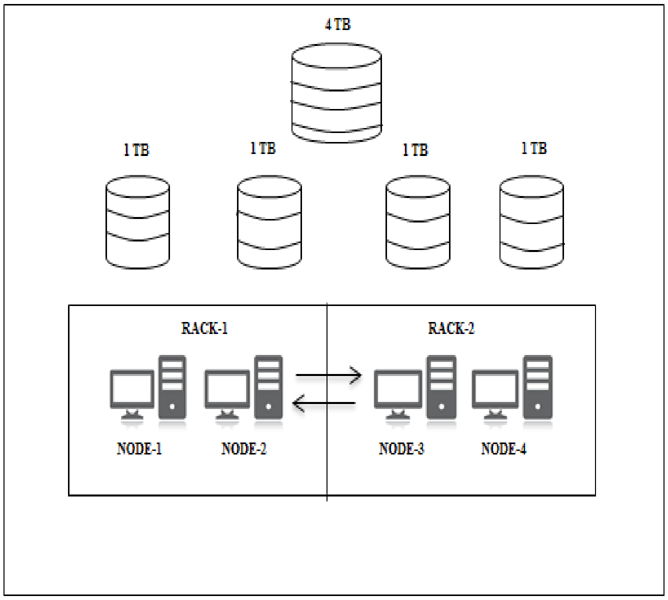


Figure 3 Example of Hadoop working process

Initially, 4 terra bytes of data are dividing into individual terra bytes of size 1 TB. The first two terra bytes are stored in rack 1 with named as node 1,node 2 similarly second two terra bytes are stored in rack 2 with named as node 3 and node 4 as shown above.

Hadoop components are HDFS and map reduce which are explained as follows

HDFS(Hadoop distributed file system)

HDFS is a shared file structure which allows variousdifferent files to be stored and can get at the equal time at an extraordinary speed. It is one of the basic components of the Hadoop framework.

The HDFS model is shown in the figure 5. In this same data is stored in different locations based on the replication factor so that if any crash/deletion of data occurs, data can be retrieved from other stored location. It is extremely trustworthy and fault-tolerant gadget for big data platform Hadoop

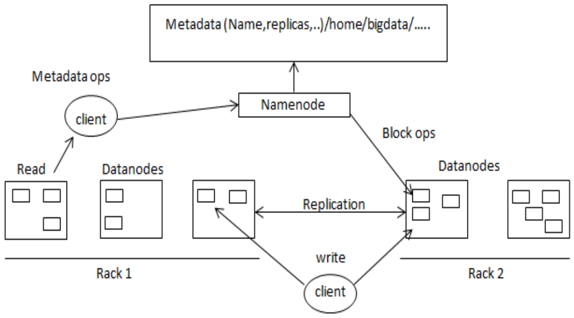


Figure 4 model of HDFS

HDFS features:

- Used to store a very massive amount of data
- Lessinvolvement of the operator
- Simultaneous computing
- Adding/removing more nodes to a system
- Restoring the data
- The data is in an available state at all times

The Hadoop Distributed File System follows the master/slave data model. HDFS file system consists ofthe name node and data node. Name node is the one which acts as a master server used to manage the overall file system and offered correct access to the customers. Data node is used to guide with the task of storage data under the node in which it is executing.

Map Reduce

It is a simultaneous structure that can be simply expanded on a huge amount of material equipment to achieve the augmented necessitate for processing huge amounts of data. This map reduce component can be worked on large amounts of data to get accuracy within a given time. It contains mainly two tasks

- Mapping

  It divides the given input data into an individual pair of (key, value) called tuples to do the task

- Reducing

  It is used to combine the mapper output into small individual elements which are used for searching.

The process of map reduce is shown in the following figure 6.

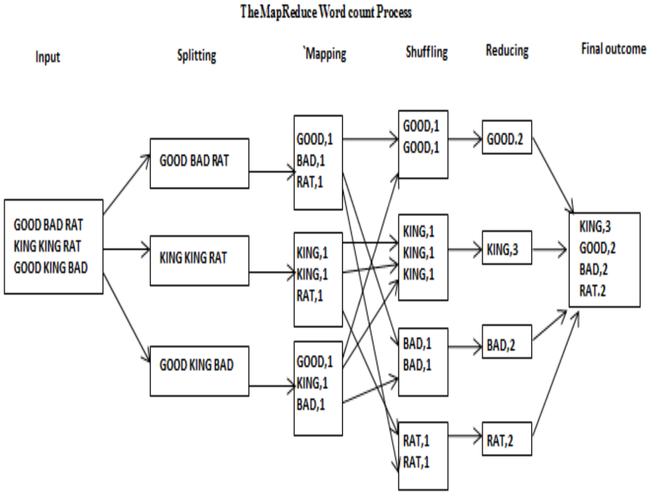


Figure 5 map reduce model

Mapper stage:

This is the primary step of the Map Reduce and it incorporates the approach of interpreting the information from the Hadoop Distributed File System (HDFS). The data should be in the structure of a list of the directory or a file. The given data file which is given by the user is given as input to the mapper function by each line at a time. The mapper then processes the data and decreases it into minute blocks of data.

Reducer stage:

In this stage, initially shuffling takes place which provides mapper output to the reducer for further processing. Then after shuffling, sorting takes place which reduces the data by combining or counting the values. Finally, the output of the reducer stage consists of a key-value pair.

The main aim of our project is to classify the documents in the form of user distinct categories based on the data. This can be skilled through the K-means algorithm. The proposedNMF model with K-means on dataset named newsgroup20 gets more accuracy than the previous system.

II. LITERATURE SURVEY

E. Laxmi Lydia et al(2015) [9] solves the clustering problem by K-means algorithm, which is one of the simplest unsupervised learning algorithms. E. Laxmi Lydia with the continuous work disparateness cluster environment is created along with the property of resource such as resource type, processing speed, and the memory. E.Laxmi Lydia et al: The well-known clustering problem can be solved by K-means, which is one of the simplest unsupervised learning algorithms.

Assume the K number of clusters for classifying a given grid processor in a simple and easy way. K-means clustering does not have a guarantee for optimal solution as the performance is based on initial centroids. Thus, the proposed system uses the partitioning clustering, say, K- centroids clustering [10].

E.Laxmi Lydia et al: With the continuous work of E.Laxmi Lydia et.al Disparateness cluster environment is created along with the properties of resource such as resource type, processing speed, and the memory. In order to avoid the scheduling delay, the system needs to form a cluster using the K-centroids clustering. Depending up on higher priorities, the node will move to the cluster [11].

Document Clustering is extensively used text mining ranging the capability with the growth in possibility of available text data. Text document clustering is applied to certainly to a group of document that associate to the same topic to provide users peruse of improved results by Gao et al. [12]. Andrews et al proposed an experimental information confirm that the prosperity from document clustering. Document clustering is consistently been used as a mechanism to enhance the achievement of improvement and operate spacious data. Currently clustering has been advanced for reading a collection of documents.

Guduru, N (2006): conventional clustering techniques use words to detect similarities between documents. These words are believed to be generally autonomous, which may not be the situation in genuine application. However, conventional VSI uses words to portray documents as a general rule, the ideas semantics highlights points are what the documents are portrayed. The highlights extracted contain the most vital document-related thought idea. Effectively, extraction included was used as part of text mining with unsupervised algorithms such as Principal Components Analysis (PCA), Singular Value Decomposition (SVD), and Non-Negative Matrix Factorization (NMF), including factoring the word matrix document.

In this paper “Text Mining With Lucene And Hadoop: Document Clustering With Updated Rules Of NMF Non-Negative Matrix Factorization” [13] by E. Laxmi Lydia,D.Ramya, Key conclusion is to run huge data-"Big Data "approach has maintained traditional data processing applications and database management tools.A new processing technique, Iterated Lovins Stemmer algorithm, has produced better results compared to the stemmer algorithm of Porter Stemmer and Lovins. And a new furthest - used KNMF algorithm called the “Text Mining Lead”.Helping cluster documents through the K means the clustering of labels with these defined KNMF features.Similarly, parallel implementation with Map Reduce for large - scale documents minimizes time computing and increases the average computing speed.Every process is done in Apache Hadoop, from token generation to clustering.

This paper “Charismatic Document Clustering Through Novel K-Means Non-negative Matrix Factorization (KNMF) Algorithm Using Key Phrase Extraction” [14] by E. Laxmi Lydia,The new system for processing and analyzing this huge data "Big Data" approach provided relief from database management tools and traditional applications for data processing.A comparison is perfected between Iterated Lovins algorithms, Lovins algorithms and Porters algorithms with comparative factors such as ICF, WSF, CSWF and Iterated Lovins algorithms resulting in maximum minimized stem words. Thus a new KNMF algorithm is used and the application is called Progressive Radical Text Mining. With these defined features of KMNF therefore, the documents are clustered, because we consider them to be the ultimate label of Kmeans clusters. And parallel implementation with MapReduce also leads to the minimisation of time computation for large documents.

Balabantaray et al.,[15]correlates the K-means clustering with K-mediods clustering. K-means was performed based on Euclidean distance and Manhattan distance measures in WEKA. Lastly, it was examined that K-means produce improved outcome than K-Mediods.

Greene Derek et al., popularized text clustering with groundwork advanced unsupervised text mining works. Jain et al., deliberates regarding pre-processing of documents, operations of text clustering. and also with their pros and cons of text clustering along with some key approaches that finalizes the algorithms which allow not to perform overlapping of clusters.Jajoo et al., explains how to progress the performance and accuracy of clustering in documents. Techniques like partitioning clustering applied in document clustering to execute more desirable performance results than standard clustering algorithms. This decreases the noise in data and enlights the quality of clusters.Khadhim et al., Implemented TF-IDF and SVD dimensionality reduction techniques and implemented the reduction techniques and given the performance for text clustering in documents.

Liu Tao et al., proven that feature selection approaches can increase the efficiency and accuracy of algorithms in clustering. The term condition is favourable than DF and Entropy based.Mugunthadevi et al., worked and studied on various feature selection approaches along with their pro and cons. Lastly ended by stating that the feature selection holds on the field by giving better results facing new challenges in text mining.Tang bin et al., handles a transformation mechanism considerably to diminish the computational cost combined with the finest transformation approaches like Independent component Analysis (ICA) and Latent Semantic Indexing (LSI) defending the clustering accomplishment. .

Zhoa et al.,[16] recommended a new mechanism that initiates the cloud model theory to feature selections in building up clustering documents. The practical results shown in the field with K - means have significantly enhanced the accuracy of text clustering under circumstances.

III. INTERCONNECTED STUDY

A. Document clustering:

Document clustering is defined as “clustering of documents”. Clustering is a procedure for interpretation ofresemblance and difference between the given words and thus, separating them into significant subsections.

The document clustering is divided into two subsections such as flat clustering and hierarchical clustering which are explained as follows

- Flat clustering:

This technique divides the document into disconnected clusters.

Different approaches used in this technique are k-means clustering, probabilityclustering, latent semantic indexing(LSI), NMF. In this paper, we are using a combination of K-means and NMF for document clustering.

- Hierarchical clustering

This method returnsconsecutive clusters of a document from acquired clusters either usingagglomerate or divisive techniques.

Document clustering [17] is also called as text clustering which is used for analysis of documents in clusters for easy and retrieval process of information.

The main aim of text clustering is as follows:

- to allocate documents to various terms
- i.e. when the terms are not well-known in advance
- as opposite to Document categorization when labels are well-known
- document clustering is a Cluster Analysis task which is anUnsupervised Learning that is applied to textual data

B. Text Mining

Text mining is a method to know and mine the important hidden information from a hugequantity of semi-structured or unstructured text data[18]. This processis a combination of both human ability and computational power of the machine. The computational power of machine includes the ability to theprocessing of huge amount of data at a high rate. Some application of text mining is:

- Information mining,
- Topic recognition and tracking,
- Overview/outline,
- Classification,
- Clustering,
- model association,
- Data view etc [22-37].

Generally, the text mining algorithms can be divided into

- Supervised learning and
- Unsupervised learning.

The below figure explains about supervised and unsupervised clustering.

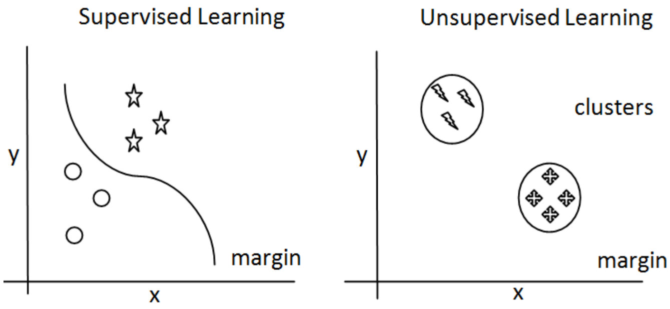


Figure 6 types of text mining algorithms

In supervised learning, the machine efforts to accomplish from the existing data that are specified. On the contrary, in unsupervised learning, the method tries to identify the model directly from the specified data. So if the database is markedwith the name its fall under a supervised problem, if the database is unnamed then it is an unsupervised problem.

V. EXISTING SYSTEM

Firstly, if we want to process any document first step here is indexing which considered as an important role of setting up each and every document in a well-ordered format. The document which is provided initially checks if the document is the new one or previouslyaccessible one. If it is a new one, the renovation is included in the unupdated list of documents.

- Once the document is updated or the provided document is already upgraded one we go further to subsequent steps without any changes of the particular document which are used for indexing and if the document is not customized then we erase the previous document by designing a new document.
- In the next step, we attempt to remove stop words
- Stemming is a procedure which is used for improved apprehend of root to the word in the document is specified. All the gathered information which is collected as of now is represented in a specified index. At last, all the unrelated data files are beendeleted from the folder.

Then tf-idf is calculated which gets the output as a document-term matrix. This document matrix is given as input to the NMF to extract the features. In this method, data mining techniques for feature extraction like PCA (principal component analysis),SVD[19](singular value decomposition), LSI(latent semantic indexing) are used. These techniques considered both positive and negative values for document clustering which degrades performance. To improve the performance of the clustering we propose the combination of K-means and NMF [20] which does not consider any negative values.

VI. PROPOSED METHOD

For the clustering of automatic documents, another updated model.i.e.KNMF compared to Lee and Sung's NMF is used. The Proposed method is implementing using a NewsGroup20 dataset. In the clustering of documents, extracted characteristics play a major role.

In the past, document clustering methods use words as a metric to know connecting documents. It is assumed that these words are mutually independent, which may not be the case in the actual application. Words are used in existing methods to describe the documents, but the concepts / semantic / features / themes actually describe the documents.

A. System architecture

the system framework for the proposed model is shown in the below figure.

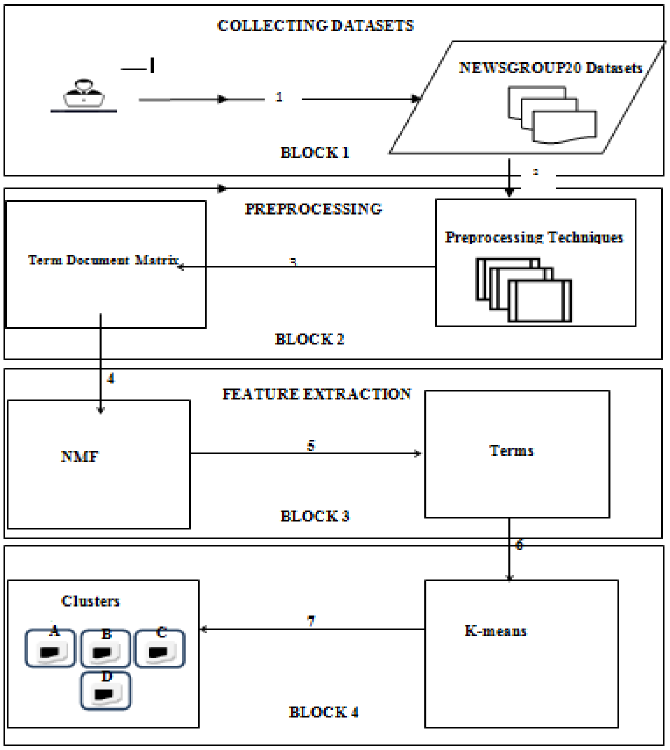


Figure 7 design of the proposed system

Previously for feature extraction unsupervised algorithms like PCA, SVD is used. Using SVD the collection of objects in the vector space also considers negative values, it makes hard for representation. Because of this reason we are using NMF in our proposed with K-means for better results.

B. NMF algorithm

It is a better type of matrix factoring where the non-negative values are not considered as matrices. The steps for the NMF algorithm is as follows

- Initially, the V_{ac} matrix is divided into two lower-class matrices W_{ab} and H_{bc}. The product of both W and H is therefore equal to V shown in the equation below.

$$V_{ab} \approx W_{ab} \times H_{bc,b \ll \min(a,c)}$$

- Here b is the number of features to be extracted orthe number of clusters required when applying document clustering can be called.
- V contains columns and rows. The column represents document/records vectors whereas rows represent term vectors. The contemporary element of document vectors assets the connections between the documents andthe terms.
- W is made up of columns which are represented as feature vectors. These are not in the form of orthogonal (for instance, if there are any overlaps in features) a). H consists of columns with some weights related to each base vector in W.
- Each document vector of the document term matrix can hence be gathered by a straight blend of the essential vectors of W weighted by the relating segments of H.
- Consider V_i any document vector with W column vectors in matrix V as { W_1, W_2, ..., W_k } and the matrix H column related components are represented { h_{i1}, h_{i2}, ..., h_{ik} } and the equation is written as

$$V_i \approx W_1 . h_{i1} + W_2 . h_{i2} + \cdots\cdots + W_k . h_{ik}$$

- NMF uses an iterative method to change W_{ab} and H_{bc}'s starting values so that V_{ac} approaches the result. Upon reaching the specified number of iterations, the NMF method will stop. The NMF division is not a unique division.
- Two simple cost functions studied by Lee and Seung are the square error and the extension of the Kullback - Leibler divergence to positive matrices. Every cost function tends to other NMF algorithm, which frequently minimizes the variance by iterative update rules. According to the Frobenius norm, the minimization problem intended for matrices can be stated as

$$Min(W,H)||V - WH|| \quad _{p^2}$$

- Where W,H are non-negative. The updated rules of NMF according to the Frobenius norm arecalled a multiplicative method.

C. MM Algorithm

- At first W, H are taken as non-negative values.
- Repeat for each c, j and I until approximation error or one iteration converges:

$$\boldsymbol{(a) H_{cj} \leftarrow H_{cj} \frac{(W^T V)_{cj}}{(W^T W H)_{cj} + e_{(a) H_{cj}}}}$$

$$\boldsymbol{(b) W_{ic} \leftarrow W_{ic} \frac{(V H^T)_{ic}}{(W H H^T)_{ic} + e}}$$

In steps 2(a) and (b), e, a minute affirmative bound equal to 10−9, is supplementary to keep away from division by zero. From this MM Algorithm, it is observed that W and H stay non-negative during the updates.

D. KNMF

The clustering of documents is carried out in this KNMF algorithm based on the resemblance among the extracted features and the individual documents. Assume feature extracted vectors as F={f_1,f_2,f_3....f_k} which are calculated by NMF. Consider term-document matrix documents as V = {d_1,d_2,d_3....d_n}. when the angle between the d_i and f_x is minimum then, the document d_i is supposed to belong to cluster f_x.

Procedure

1.Build the document term matrix V using the tfidf value fro m the records of a given input folder
2. The length of columns of V is standardized by using the Euclidean distance
3. NMF is applied on V and calculate the values of W and H by using the below equation

$$V_{ab} \approx W_{ab} \times H_{bc}$$

4. To calculate the distance between the documents d_i and extracted vectors of W K-means algorithm is used. When the angle between d_i and w_x is minimum, allocate d_i to w_x. This is correspondent to k-means algorithm by a particular turn.

k-means:

k-means is the one of the most important unsupervised algorithm for the process of clustering. Clustering is the mechanism of splitting the points into classes based on the resemblance.If the value of K is given as input, then the process of the K-means as follows

- Divide the objects or data into K subsets in which data is not null.
- Recognize the mean point of the clusters for the current split.
- Allocate each point to a particular cluster.
- Calculate the distances from each point and then allocate the points to the cluster based on the minimum distance from the centre.
- After rearranging the points calculate the new mean based on the newly assigned points.

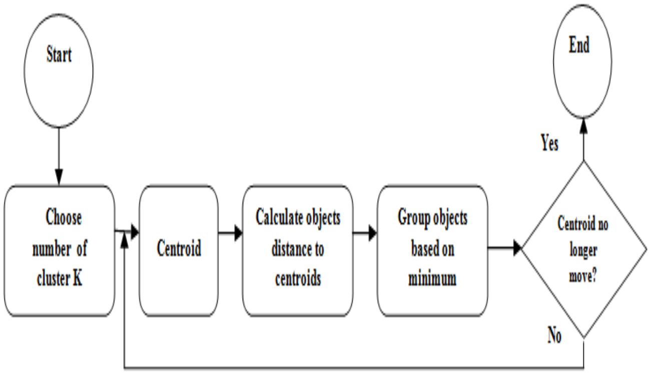

Start
End
Yes
Choose number of cluster K
Centroid
Calculate objects distance to centroids
Group objects based on minimum
Centroid no longer move?
No

Figure 8 K means procedure

5. To run the K-means parallel Hadoop is executed in local and pseudo mode.

The procedures of NMF, MM and KNMF are explained above. The proposed method follows the below procedure.

- Initially, the dataset NEWSGROUP20 is downloaded for the implementation.
- By using pre-processing techniques the term document matrix of a particular document is obtained in a dataset
- To obtain features of a particular document NMF method is used
- For cluster formation of individual document based on similarity k-means algorithm is used.
- Finally, from formed clusters identify top terms in that.

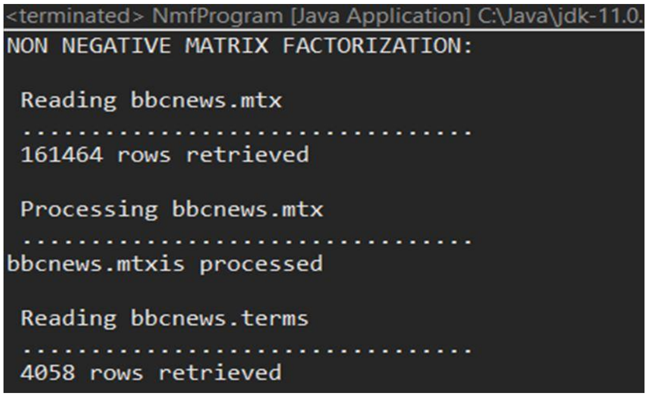
VII. RESULTS

The results of the proposed method are shown below. NEWSGROUP20 dataset is used for clustering of documents and to retrieve top words of a particular cluster. Initially pre-processing techniques are used on dataset to obtain the matrix format. Because the NMF can process through only matrix format documents. Now NMF is used to read the dataset which is in the form of matrix which retrieves the 161464 rows as shown in Fig 9. After reading the matrix that is processed through NMF. To obtain the TF-IDF values the terms of the dataset should also be read and processed. After this 4058 terms are processed as shown in Fig 9.

In this proposed method two algorithms are used

- NMF
- K-means

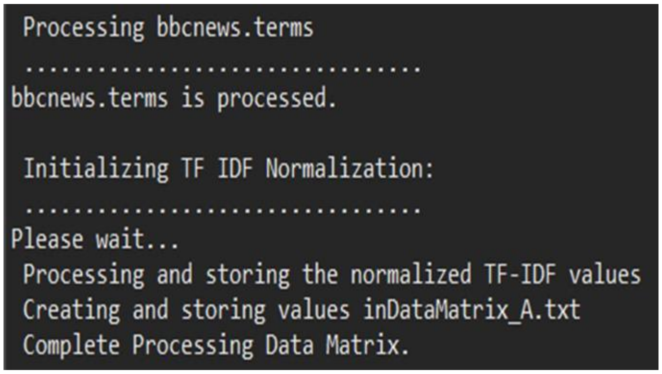
These both are used parallel to implement the clustering of documents and to retrieve the top term of a particular cluster.



```
<terminated> NmfProgram [Java Application] C:\Java\jdk-11.0.
NON NEGATIVE MATRIX FACTORIZATION:

 Reading bbcnews.mtx
 ..................................
 161464 rows retrieved

 Processing bbcnews.mtx
 ..................................
bbcnews.mtxis processed

 Reading bbcnews.terms
 ..................................
 4058 rows retrieved
```

Fig 9 screen for reading dataset

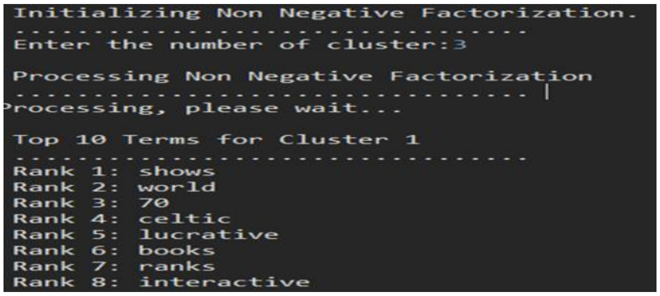
Now the terms are processed using NMF algorithm. Finally we got TF IDF values of the document. Those TF IDF values are stored in inDataMatrix_A.txt which is shown in Fig 10.



```
 Processing bbcnews.terms
 .................................
bbcnews.terms is processed.

 Initializing TF IDF Normalization:
 .................................
Please wait...
 Processing and storing the normalized TF-IDF values
 Creating and storing values inDataMatrix_A.txt
 Complete Processing Data Matrix.
```

Fig 10 Screen for initializing

Now after obtaining TF IDF values clustering is applied on it. Number of clusters here used are 3. After entering the value NMF with K-means is used and gives the top terms in the cluster 1 as shown in below Fig 11. This process repeats for the remaining two clusters.



```
Initializing Non Negative Factorization.
..................................
Enter the number of cluster:3

Processing Non Negative Factorization
.................................. |
Processing, please wait...

Top 10 Terms for Cluster 1
..................................
Rank 1: shows
Rank 2: world
Rank 3: 70
Rank 4: celtic
Rank 5: lucrative
Rank 6: books
Rank 7: ranks
Rank 8: interactive
```

Fig 11 Screen for cluster-1

VIII CONCLUSION

This paper proposes a new model with updated rules called NMF with a combination of K - means for document clustering and application development based on this medium.The developed model is used to organize folders in such a way that the documentation can be divided into subfolders without any knowledge of content. Therefore, the performance in the retrieval of documents in any situation really increases.

The accuracy of the proposed model was tested. With the 2 clusters of documents the accuracy of 85 percent was obtained and with the 3 clusters of documents the accuracy of 80 percent is obtained. The average accuracy of all clusters from 2 to 8 is 70 %.NMF is a good metric for document clustering that is more accurate when used with parallel K - means algorithm. The proposed method uses mapreduce implementation of k-means from apache hadoop project.

REFERENCES

1. E. Laxmi Lydia, P.KrishnaKumar, K.Shankar, S.K. Lakshmanaprabu, R. M. Vidhyavathi, AndinoMaseleno, “ Charismatic Document Clustering Through Novel K-means” International Journal of Parallel Programming, Springer 2018

2. Rahul Bekta, “Big Data And Hadoop: A Review Paper”, e-ISSN: 1694-2329 | p-ISSN: 1694-2345, Volume 2, Spl. Issue 2,RIEECE -2015

3. Subramaniyaswamy V, Vijayakumar V, Logesh R, and Indragandhi V, “Unstructured Data Analysis on Big Data using Map Reduce”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15)
4. PravinShinde&SharvariGovilkar, “A Systematic study of Text Mining Techniques”, International Journal on Natural Language Computing (IJNLC) Vol. 4, No.4, August 2015
5. D. Lee, H. S. Seung, “Learning the parts of objects by non-negative matrix factorization, https://www.Researchgate.net/publicatio n/12752 9371999
6. Lee, D &Seung, H (2001). Algorithms for non-negatvie matrix factorization. In T.G. Dietterich and V. Tresp, editors, Advances in Neural Information ProcessingSystems, volume 13. Proceedings of the 2000 Conference: 556-562,The MIT Press.
7. Martin RajmanRomaric Besançon, “Text Mining - Knowledge extraction from unstructured textual data”, Advances in Data Science and Classification pp 473-480
8. Peter Willet, “The Porter stemming algorithm: then and now”, Program: electronic library and information systems, 40 (3). pp. 219-223
9. *Dr.E.Laxmi Lydia,”*Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity”.
10. CH V T E V Laxmi, Dr.K.Somasundaram, 2HARS: Heterogeneity-Aware Resource Scheduling in Grid Environment using K-Centroids clustering and PSO techniques, IJAER Journal, ISSN 0973-4562 Volume 10, No. 7, 2015, Page No.:18047-18062.
11. Dr.E.Laxmi Lydia, Dr.M.BenSwarup, Dr.ChallaNarsimham, A Disparateness– Aware scheduling using K-Centroids clustering and PSO techniques in Hadoop cluster.
12. NeelimaGuduru, Text Mining with support vector machines and non-negative matrix factorization algorithm, Master’s thesis, University of Rhode Island, CS Dept., submitted in 2006.
13. E.Laxmi Lydia, “Text Mining With Lucene And Hadoop: Document Clustering With Updated Rules Of NMF Non- Negative Matrix Factorization”, International Journal of Pure and Applied Mathematics, Volume 118 No. 7 2018, 191-198
14. E.Laxmi Lydia, “Charismatic Document Clustering Through Novel K-MeansNon-negative Matrix Factorization (KNMF) Algorithm UsingKey Phrase Extraction”, International Journal of Parallel Programming, Springer, 2018.
15. Balabantaray ,Rakesh Chandra, ChandraliSarma, and Monica Jha. “ Document Clustering using K-Means and K-Medoids.” arXiv preprint arXiv:1502.07938 (2015)
16. Zhao, Junmin, Kai Zhang, and Jian Wan. “Research of feature selection for text clustering based on cloud model. “ Journal of Software 8.12 (2013): 3246-3252.
17. T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, A.Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation”, IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 24, Issue: 7, Jul 2002)
18. Sunghae Jun, Sang-Sung Park b, Dong-Sik Jang, “Document clustering method using dimension reduction and support vector clustering to overcome sparseness” Elsevier journal, 0957-4174/ 2013
19. Gao, Jing, and Jun Zhang. “Clustered SVD strategies in latent semantic indexing.”Information processing & management 41.5 (2005): 1051-1063.
20. Shintaro Sato, Akihiro Kayahara, Sjin-ichiimai, "Unstructured data treatment for big data solutions" IEEE International symposium on semiconductor Manufacturing.
21. Lakshmanaprabu, S. K., Shankar, K., Ilayaraja, M., Nasir, A. W., Vijayakumar, V., & Chilamkurti, N. (2019). Random forest for big data classification in the internet of things using optimal features. International Journal of Machine Learning and Cybernetics, 1-10. https://doi.org/10.1007/s13042-018-00916-z
22. Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2018). Financial crisis prediction model using ant colony optimization. International Journal of Information Management. https://doi.org/10.1016/j.ijinfomgt.2018.12.001
23. Uthayakumar, J., Metawa, N., Shankar, K., & Lakshmanaprabu, S. K. (2018). Intelligent hybrid model for financial crisis prediction using machine learning techniques. Information Systems and e-Business Management, 1-29. https://doi.org/10.1007/s10257-018-0388-9
24. Lakshmanaprabu, S. K., Mohanty, S. N., Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, 374-382.
25. Lakshmanaprabu, S. K., Shankar, K., Gupta, D., Khanna, A., Rodrigues, J. J., Pinheiro, P. R., & de Albuquerque, V. H. C. (2018). Ranking analysis for online customer reviews of products using opinion mining with clustering. Complexity, 2018.
26. Karthikeyan, K., Sunder, R., Shankar, K., Lakshmanaprabu, S. K., Vijayakumar, V., Elhoseny, M., & Manogaran, G. (2018). Energy consumption analysis of Virtual Machine migration in cloud using hybrid swarm optimization (ABC–BA). The Journal of Supercomputing, 1-17.
27. Lydia, E. L., Kumar, P. K., Shankar, K., Lakshmanaprabu, S. K., Vidhyavathi, R. M., & Maseleno, A. (2018). Charismatic Document Clustering Through Novel K-Means Non-negative Matrix Factorization (KNMF) Algorithm Using Key Phrase Extraction. International Journal of Parallel Programming, 1-19.

28. Shankar, K., Lakshmanaprabu, S. K., Gupta, D., Maseleno, A., & de Albuquerque, V. H. C. (2018). Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The Journal of Supercomputing, 1-16.
29. Lakshmanaprabu, S. K., Shankar, K., Khanna, A., Gupta, D., Rodrigues, J. J., Pinheiro, P. R., & De Albuquerque, V. H. C. (2018). Effective Features to Classify Big Data Using Social Internet of Things. IEEE Access, 6, 24196-24204.
30. Maseleno, A., Tang, A. Y., Mahmoud, M. A., Othman, M., Negoro, S. Y., Boukri, S., ... & Muslihudin, M. The Application of Decision Support System by Using Fuzzy Saw Method in Determining the Feasibility of Electrical Installations in Customer's House. International Journal of Pure and Applied Mathematics, 119(16).
31. Maseleno, A., Tang, A. Y., Mahmoud, M. A., Othman, M., & Shankar, K. (2018). Big Data and E-Learning in Education 4.0. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 18(5), 171-174.
32. Muslihudin, M., Wanti, R., Hardono, N., Shankar, K., Ilayaraja, M., Maseleno, A., ... & Mukodimah, S. (2018). Prediction of Layer Chicken Disease using Fuzzy Analytical Hierarcy Process. International Journal of Engineering & Technology, 7(2.26), 90-94.
33. Amin, M. M., Maseleno, A., Shankar, K., Perumal, E., Vidhyavathi, R. M., & Lakshmanaprabu, S. K. (2018). Active Database System Approach and Rule Based in the Development of Academic Information System. International Journal of Engineering & Technology, 7(2.26), 95-101.
34. Aminudin, N., Sundari, E., Shankar, K., Deepalakshmi, P., Fauzi, R. I., & Maseleno, A. (2018). Weighted Product and Its Application to Measure Employee Performance. International Journal of Engineering & Technology, 7(2.26), 102-108.
35. Putra, D. A., Jasmi, K. A., Basiron, B., Huda, M., Maseleno, A., Shankar, K., & Aminudin, N. (2018). Tactical Steps for E-Government Development. International Journal of Pure and Applied Mathematics, 119(15), 2251-2258.
36. Sugiyarti, E., Jasmi, K. A., Basiron, B., Huda, M., Shankar, K., & Maseleno, A. (2018). Decision support system of scholarship grantee selection using data mining. International Journal of Pure and Applied Mathematics, 119(15), 2239-2249.
37. Susilowati, T., Jasmi, K. A., Basiron, B., Huda, M., Shankar, K., Maseleno, A., & Julia, A. (2018). Determination of Scholarship Recipients Using Simple Additive Weighting Method. International Journal of Pure and Applied Mathematics, 119(15), 2231-2238.

AUTHORS PROFILE

Dr E.Laxmi Lydia is Associate Professor of Computer Science Engineering at Vignan's Institute of Information Technology(A). She is a Big Data Analytics Online trainer for the international training organization and she has presented various webinars on Big Data Analytics. She is certified by MICROSOFT CERTIFIED SOLUTION DEVELOPER (MCSD). She Published 50 research papers in International Journals in the area Big Data Analytics and Data Sciences and she Published 10 research papers in International conference proceedings. She has been a key note speaker on Big Data Analytics and on Data Sciences. She is an author for the Big Data Analytics Book, currently she is working on Government DST Funded Project and she holds a patent entitled "ACCOMPLISHMENT OF "PAY AS YOU GO" APPROACH ON ENTITY RESOLUTION FOR DATA INTEGRATION AND DATA PRIVACY"

Dr. Kollati Vijaya Kumar completed B. Tech in CSE from Andhra University. He obtained M Tech in Computer Science and Engineering from Acharya Nagarjuna University and completed Ph. D in Computers Science and Engineering at Karpagam University, Coimbatore. He is having 15 years of teaching experience. He is currently working as Assoc. Professor & Head in Department of Computer Science and Engineering at Vignan's Institute of Engineering for Women, Duvvada, Visakhapatnam, Andhra Pradesh, India. He published 10 papers in International journals. His area of interests are Wireless Networking, Network Security, Grid computing, Information security, Big Data.

K. Shankar. received the M.C.A., M.Phil., and Ph.D. degrees in computer science from Alagappa University, Karaikudi, India. He is currently an Assistant Professor with the School of Computing, Kalasalingam Academy of Research and Education, Krishnankoil, India. He has several years of experience in research, academia, and teaching. He has been a part of various seminars, paper presentations, research paper reviews, and conferences as a convener and a session chair, a guest editor in journals. He has authored or co-authored many research papers in reputed journals and conferences. He also has papers in SCI-indexed and IEEE journals. He was a Reviewer in some SCI indexed Journals like IETE Research, Springer and the IEEE conferences. He displayed vast success in continuously acquiring new knowledge and applying innovative pedagogies and has always aimed to be an effective educator and have a global outlook. His current research interests include cryptography, data mining, big data, and Internet of Things.