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 Semantic Indexing (LSI), Single value decomposition for unstructured document which was quickly filtered and viewed, but it is much harder to comprehend 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, a new 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 term that portrays an expansive volume of structured, semi-structured and unstructured data that can be extracted from information 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, or Not Only SQL (NoSQL) systems. The data types are shown in the below diagram.

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 such 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 rejoicing patterns to detect. The process of text mining is shown in below figure.
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 recognize the latent associations.

The most important format for gadget file organization is to insert them into folders and 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 an idea 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 number when their contents can't be renamed 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 Value Decomposition (SVD).

Numerous authors are implementing 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 utilizes a key phrase extraction algorithm[6]
- Stemming is done by using the proposed 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-means clustering algorithm [8] has been utilized.

B) Overview of Apache Hadoop

Hadoop:

It is a purely free source programming 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 kept each and every part 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 is there 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](image3.png)
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 various different 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.

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 of the 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.

![Map Reduce Model](image)

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 proposed NMF 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 idea 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 MapReduce 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 KNMF 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-medoids 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-Medoids.

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. Tajoo 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. Mungunathodevi 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.

Zhao 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 of resemblance 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.
The below figure explains about supervised and unsupervised clustering.

![Figure 6: Types of text mining algorithms](image)

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 marked with 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 previously accessible one. If it is a new one, the renovation is included in the updated 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 comprehension 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 been deleted 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.
The NMF algorithm is as follows: 

1. Initially, the feature matrix \( V \) 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.

\[
V_{ab} \approx W_{ab} \times H_{bc}, b < \min \{i, c\}
\]

- Here, \( b \) is the number of features to be extracted, or the number of clusters required when applying document clustering can be called.
- \( V \) contains columns and rows. The column represents a document/record vector whereas rows represent term vectors. The contemporary element of document vectors asserts the connections between the documents and the 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 \( V \), 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 \{ \( h_1, h_2, ..., h_k \) \} and the equation is written as

\[
V_i \approx W_1 \cdot h_1 + W_2 \cdot h_2 + \ldots + W_k \cdot h_k
\]

NMF is 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 algorithms, 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 are called a multiplicative method.

### C. MM Algorithm

- At first, \( W, H \) are taken as non-negative values.
- Repeat for each \( c, j \) and \( l \) until approximation error over one iteration converges:

\[
(a) H_{cf} \leftarrow H_{cf} \frac{(W^T V)_{cf}}{(W^T W)_{cf} + \epsilon(a)H_{cf}}
\]

\[
(b) W_{lc} \leftarrow W_{lc} \frac{(V H^T)_{lc}}{(W H^T)_{lc} + \epsilon}
\]

In steps (a) and (b), \( \epsilon \), a, 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 \( F = \{f_1,f_2,...,f_n\} \) which are calculated by NMF. Consider term-document matrix documents as \( V = \{d_1,d_2,d_3,...,d_{m}\} \) when the angle between the \( d_i \) and \( f_k \) is minimum then, the document \( d_i \) is supposed to belong to cluster \( f_k \).

**Procedure**
1. Build the document term matrix \( V \) using the tfidf value from 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 \( w_c \) is minimum, allocate \( d_i \) to \( w_c \). This is correspondent to k-means algorithm by a particular turn.

\( k \)-means:

- 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.

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.

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.
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 Fig 11. This process repeats for the remaining two clusters.

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


7. Martin RajmanRomaric Besançon, “Text Mining - Knowledge extraction from unstructured textual data”, Advances in Data Science and Classification pp 473-480


9. Dr. E.Laxmi Lydia, "Document Clustering Based On Text Mining K-Means Algorithm Using Euclidean Distance Similarity”.


11. Dr.E.Laxmi Lydia, Dr.M.BenSwarup, Dr.ChallaNarshimham, A Disparateness– Aware scheduling using K-Centroids clustering and PSO techniques in Hadoop cluster.


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 Karupagam 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.