ENHANCING ACCURACY OF TEXT CLUSTERING USING WEIGHT BASED ALGORITHM

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

MAMTA KUMARI

11505404

Supervisor

Mrs. Kirandeep Kaur

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

MAY 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P173::M.Tech. (Information Technology) [Full Time]

COURSE CODE : INT545 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0237

Supervisor Name : Kirandeep Kaur **UID :** 15901 **Designation :** Assistant Professor

Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Mamta Kumari	11505404	2015	K1520	9814472604

SPECIALIZATION AREA : Software Engineering **Supervisor Signature:** _____

PROPOSED TOPIC : Text mining in document clustring

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	6.67
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.17
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	6.83
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	6.83
5	Social Applicability: Project work intends to solve a practical problem.	6.67
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.17

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): Yes
PAC Member 2 Name: Mandeep Singh	UID: 13742	Recommended (Y/N): Yes
PAC Member 3 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes
PAC Member 5 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
DAA Nominee Name: Kanwar Preet Singh	UID: 15367	Recommended (Y/N): Yes

Final Topic Approved by PAC: **Enhancing accuracy of text clustering using weight base algorithm**

Overall Remarks: Approved (with major changes)

PAC CHAIRPERSON Name: 11011::Rajeev Sobti **Approval Date:** 22 Nov 2016

ABSTRACT

The management and extraction of useful information from large dataset is very big challenge for each organization. In this paper we discuss about data mining, its application text mining using clustering technique. The working on textual documentation is an application of clustering analysis which includes automatic topic extraction, document organization, and information retrieval. In this paper we introduce document clustering, procedure used in document clustering, challenges faces in document clustering. In existing approach the weight based algorithm (LDA) used for calculate the weight of each word to generate final cluster. In weight based algorithm some words are not cluster which reduce the accuracy of document clustering. In proposed work, DMNB algorithm used to increase the accuracy and reduce execution time. The neural network used to apply sort the similar kind of words in dictionary order to improve the accuracy of text clustering.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "ENHANCING ACCURACY OF TEXT CLUSTERING USING WEIGHT BASED ALGORITHM" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mrs. Kirandeep Kaur.. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Mamta Kumari

R.No:- 11505404

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**Enhancing Accuracy of Text Clustering Using Weight Based Algorithm"**, submitted by **Mamta Kumari** at **Lovely Professional University, Phagwara, India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Kirandeep Kaur

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____

2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

I owe a debt of deepest gratitude to my dissertation supervisor, **Miss. Kirandeep Kaur**, Department of computer science and engineering, for her guidance, support, motivation and encouragement throughout the period within which this work is caring out. Her readiness for consultation at all times, her educative comments, her concern and assistance even with practical things have been invaluable. I would also like to express my gratitude to all my friends in the department of computer science for their support and encouragement.

I am grateful to **Mr. Janpreet Singh,** head of the department, computer science and engineering for providing me the necessary opportunities for working on thesis. I also thank the other faculty and staff members of my department for their invaluable help and guidance.

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

Inner first page – Same as cover i

PAC form ii

Abstract iii

Declaration Statement iv

Supervisor's Certificate v

Acknowledgement vi

Table of Contents vii

List of Figures x

CHAPTER1: INTRODUCTION 1

1.1 Data Mining **1**

1.1.1 Knowledge Discoveries Step In Database **1**

1.2 Text Mining **2**

1.3 Document Analysis **3**

1.3.1 Tokenization **3**

1.3.2 Stop Removal Method **4**

1.4 Document Clustering **4**

1.4.1 Preprocessing **5**

1.4.2 Feature Extraction **5**

1.4.3 Document Clustering **6**

1.4.4 Document Representation **6**

1.5 Document Clustering Challenges **6**

1.5.1 Cluster Analysis Requirement **7**

1.6 Technique 0f Text Mining **8**

1.6.1 Natural Language Processing **8**

TABLE OF CONTENTS

CONTENTS	PAGE NO.
1.6.2 Information Extraction	9
1.7 Algorithm Used In Document Clustering	9
1.7.1 K-means Clustering Algorithm	9
1.7.2 Hierarchical Clustering	10
1.7.3 Clustering Based Upon Density	11
1.7.4 Clustering Based Upon Grids	11
1.8 Basic Level To Measure The Text Clustering	12
CHAPTER2: REVIEW OF LITERATURE	15
APTER3: PRESENT WORK	31
3.1 Problem Formulation	31
3.2 Objective of The Study	31
3.3 Research Methodology	34
3.3.1 Implementation of Tool	35
CHPTER4: RESULTS AND DISCUSSION	37
4.1 Experimental Result	37
4.2 Comparison With Existing Technique	41
CHAPTER5: CONCLUSION AND FUTURE SCOPE	46
5.1 Conclusion	46
5.2 Future Scope	46
REFERENCES	47
APPENDIX	50

LIST OF FIGURES

FIGURE NO.	FIGURE DESCRIPTION	PAGE NO.
Figure 1.1	Shows the step include in text data mining	1
Figure 1.2	Shows various operation perform in document clustering	5
Figure 1.3	Shows the document clustering	6
Figure 1.4	Shows the clustering using k-mean	10
Figure 1.5	Represent agglomerative and divisive method	11
Figure 1.6	shows the density based clustering	12
Figure 1.7	shows the grid based clustering	13
Figure 3.1	Flow chart of existing work	33
Figure 3.2	Research Methodology	34
Figure 4.1	Loading of Basic code	37
Figure 4.2	LDC function applied	38
Figure 4.3	LDA function applied	38
Figure 4.4	Accuracy calculation is 57.14	39
Figure 4.5	Accuracy calculation and Clustering of Data file	40
Figure 4.6	Coding of research methodology	41
Figure 4.7	DMNB algorithm implementation	42
Figure 4.8	DMNB algorithm execution complete	42
Figure 4.8	Accuracy calculation	44
Figure 4.9	Final clustering result	44

Checklist for Dissertation-III Supervisor

Name: _____ UID: _____ Domain: _____
Registration No: _____Name of student:_____

Title of Dissertation:

_____

- ☐ Front pages are as per the format.
- ☐ Topic on the PAC form and title page are same.
- ☐ Front page numbers are in roman and for report, it is like 1, 2, 3…….
- ☐ TOC, List of Figures, etc. are matching with the actual page numbers in the report.
- ☐ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.
- ☐ Color prints are used for images and implementation snapshots.
- ☐ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.
- ☐ All the equations used in the report are numbered.
- ☐ Citations are provided for all the references.
- ☐ **Objectives are clearly defined.**
- ☐ Minimum total number of pages of report is 50.
- ☐ Minimum references in report are 30.

Here by, I declare that I had verified the above mentioned points in the final dissertation report.

Signature of Supervisor with UID

CHAPTER 1

INTRODUCTION

1.1 Data Mining

Document clustering is application of cluster analysis which analyzed using data mining technique. Data mining is a process of knowledge discovery includes integration the data, data cleaning, and transformation of data. It also involves data selection, evaluation of pattern, pattern discovery, and knowledge presentation. Knowledge discovery is a process of data summarizing in into useful information and analyzing data from different perspectives. The users can analyze data from different angles or dimensions using data mining. Data mining applied on various kinds of data such as ordered data, graph data streams, or network, text, multimedia, and web data. Data mining has many applications such as web search, finance, digital libraries, bioinformatics, health information, and business intelligence.

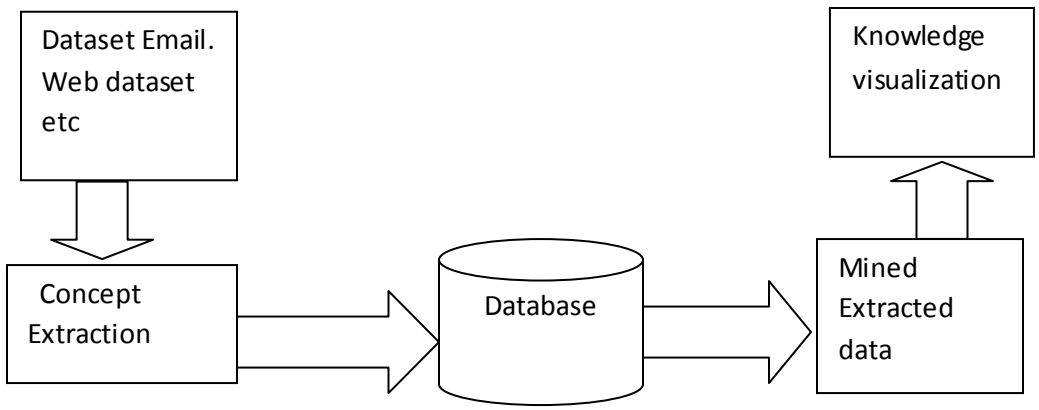

Figure1.1: Shows the step include in text data mining

1.1.1 Knowledge discoveries step in database: - In databases process the knowledge discovery focus on following step to discover the meaningful knowledge.

- Data cleaning: It is a process in which irrelevant data noise data is removed from large database. The noisy data contain stop word stemming word that need to be reduce into text information.

- Data integration: in this process the data similar kind of data is combined that are related to each other.
- Data selection: the relevant data is selected from database for retrieving information.
- Data transformation: the data is transform for mining process by performing aggregation or operation summary in this process.
- Pattern evaluation: In this process, interesting patterns are evaluated and representing knowledge is identified based on given measures.
- Knowledge representation: to represent knowledge to the user, knowledge representation techniques visualization is used.

1.2 Text Mining

The information stored in unstructured textual format in the world. For example large amount of online text document available on internet. It is impossible to organizing manually such rapidly and vast amount of data. We need to extract relevant and useful information from large dataset. So solve this purpose, important to develop efficient text mining techniques or algorithm. We finding interesting pattern from large database. It's also known as intelligence pattern text analysis. Text mining applied with unstructured or semi structure dataset such as full text document, emails and HTML etc. Text Mining is the discovery by unknown information extract automatically and relating information from different written resources. The extraction information from large dataset is very complex task.

The irrelevant words are removed from textual dataset that may reduce the dimension of textual document.

Text mining extracts useful information from a collection of documents that is unstructured. It is field of data mining in which the data sources are semi-structured or documents unstructured. Text Mining is the unknown information that is define by extracting information from different type of written resources. Text mining is mined the data that are unstructured form with dissimilar to each other. The user wants to access data that is known by them. The user wants to access data that is already known. The text mining follows the technique of preprocessing, document extraction, text clustering and classification of data. This also represents the document with different dimension. The data with lower dimension is easy to mine as compare to data with higher dimension.

1.3 Document Analysis

The document analysis is analyzed by various step such as document extraction, preprocessing, document clustering and classify and represent the document into different dimension. The stop word and stemming word are removed during preprocessing. Then we cluster that have the similar kind of object we have cluster them into one cluster and the dissimilar object are cluster into another cluster. The clustering technique applied in document to reduce the dimension of data. The data is available to user in the form that the user wants to search. The weighting term technique is applicable for document clustering by assigning the weight to each term in the document. The document with highest weight is clustered one then another one. The clustering is possible basis on the priority assign to word. The term frequency and entropy used to measurement the similarity between documents. The cluster analysis is applied basis on semantic weight and term analysis. When we reduce the irrelevant word from document or access only structure data then the dimension also reduce of document. It's easy for view the data to user with low dimensionality.

Document pre-processing is the method of introducing a new document to the information retrieval system in which each and every document introduced is highlighted by a index terms set. The goal of document pre-processing is to signify the documents in a manner such that the way that their memory in the system and repossession from the system are very efficient. It has following methods.

1.3.1 Tokenization

Tokenization in text mining is the method of splitting a given stream of text or sequence of characters into symbols, words, phrases or other meaningful rudiments called tokens. These tokens are grouped together as a semantic unit and can be used as input for further processing such as parsing or text mining. Tokenization is very useful in the field of data security and Natural Language Processing. It is used as a body of text segmentation in Natural Language processing and as a unique symbol representation for the sensitive data in the data security without compromising its security importance. As tokenization is occurring at every level of word but meaning of words vary accordingly. White space character, punctuation separates the tokens. The resulting list may or may not contain whitespace and punctuations.

1.3.2 Stop Removal Method: -

Stop word are word that are irrelevant in document that must be need to remove into the document. These word such as a, the, an, with, at, etc. these word are removed during preprocessing.

After document analysis section is divided into two sections:

- Analysis of term
- Analysis of term

Term analysis is calculate the term approximately frequency and document integrated frequency value .on the other hand the semantic analysis is designed to calculate semantic weights. The term cube are generate in the term analysis. The semantic cube is generating in the semantic analysis. The term weight and semantic analysis are used in the clustering process. The term analysis and semantic analysis are evaluated after the document preprocessing

1.4 Document clustering

The useful information extraction in from large amount of data is very big challenge for various organisations. The data is not available in the structure form, so there is need to manage this large amount of data according to user query. The text document that is unstructured, clustering process as the most difficult due to huge volume, complex semantics, high dimensionality etc. to overcome this problem document clustering is used. Document clustering more efficient techniques for organizing document in unsupervised manner and is operation used in organization information, summarizing, information retrieval and automatic topic extractions. We apply text mining technique using clustering method to fastest access of information when we search it. Cluster is a process of grouping document into a cluster; similar documents are grouping in the one cluster and dissimilar documents in another cluster. The complexity in document clustering is increasing when, its high dimension of textual information. We also reduce the dimension by apply efficient technique such as stemming. Dimension reduction is necessary process in data mining and also inrease the performance of clustering technique.

The basic step involved in document clustering such as preprocessing, text document encoding, dimension reduction technique, text document clustering, and optimization. There have following procedure of document clustering.

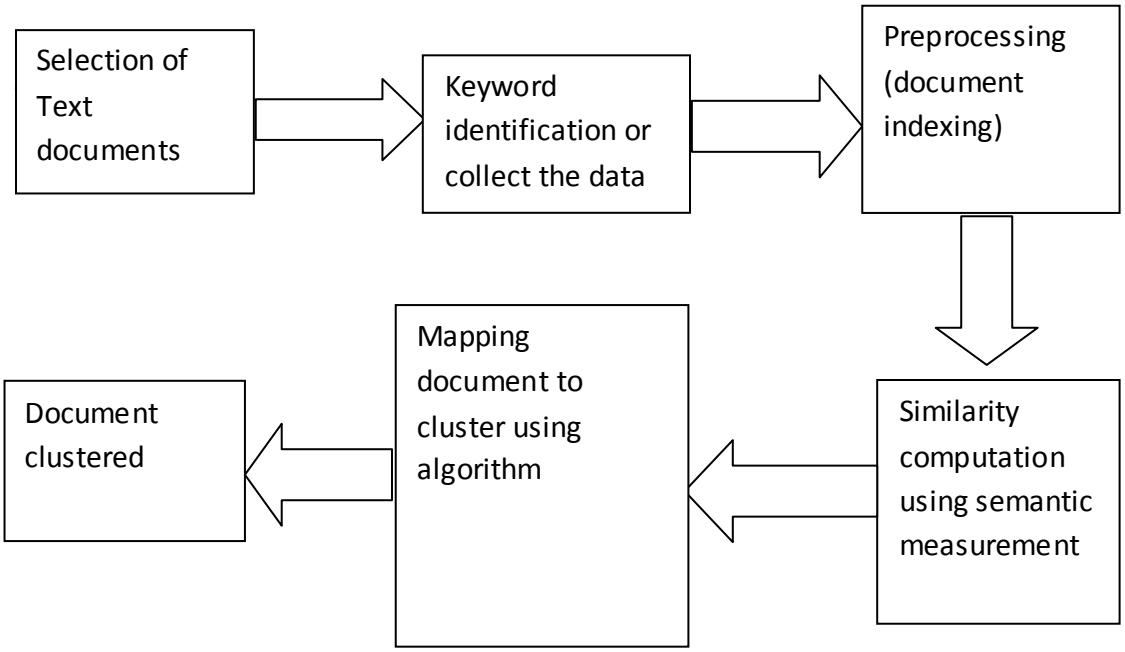

Figure 1.2: - Shows various operation perform in document clustering

1.4.1 Preprocessing

The concept of pre-processing is used to prune all character and term from the document with poor information. Pre-processing operation perform such as filtering, stemming, tokenization, stop word removal, pruning. To removing special characters from input document is the process of filtering that is not enough to hold any discriminative power under vector model. Tokenization splits the words into tokens. The stop word carries common word which is filter before and after processing of data. We need to remove stop word such as 'a' 'the' 'any' etc. the stemming process of reducing word which have similar meaning such as flow and flying are stemmed to fly.

1.4.2 Feature extraction

Feature extraction is generating of new feature from the original feature through some technique such as word clustering. It reduces the computation complexity; obtain batter classification performance by reducing the amount of irrelevant and redundant information in classification problem. The feature extraction method has a limitation is to generate new features may not have a clear understand by its meaning so that clustering result are difficult to optimized.

1.4.3 Document representation

Document representation signifies that finding a document model, a set of feature that can be used to represent a document. The vector space model (VSM) is the most common used model of document representation referred as bag of words, in which document is converted into vector of word. To define the relationship of document can also be represented as concept vector. This allows the similarity between document using geometric measures such as cosine similarity and Euclidean distance.

1.4.4 Document clustering

The target documents are grouped into different clusters on the basis of selection features. For document clustering two techniques have used, hierarchical clustering and partitioning method. Partitioning clustering devided the data point into a set of disjoint group. The type of partitioning clustering is K-means.

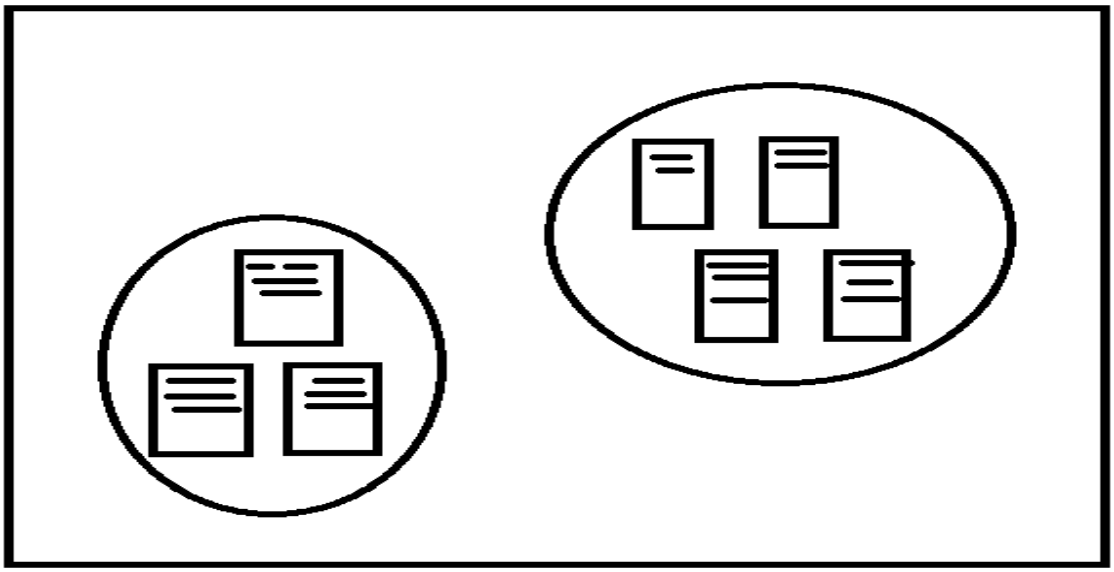

Figure 1.3: -shows the document clustering

1.5 Document Clustering Challenges

Clustering is implementing from many year but still have facing the many problem during clustering. We need to focus some constraint and challenge during clustering.

- Feature selection: The feature selection is major challenge in document clustering because the clustering is an unsupervised technique. It is difficult to select the feature due to absence of class labels.
- Selection of appropriate similarity measure: The similarity measure like Euclidian Manhattan can be used for numerical attributes, but for the categorical attributes, the selection of similarity measurement is difficult.

- Implementation of clustering algorithms: By making optimal use of available memory and CPU resources, implementation of clustering algorithm is big challenge.
- Meaningful cluster label: Meaningful clustering label, in process of browsing the cluster can guide user by providing a meaningful cluster description. Cluster method should provide labels to cluster that must be understandable to non-expert.
- High Accuracy requirement: The outcome of clustering procedure must be accurate and resultant cluster should be high clustering quality. High quality means that it clustering should be done with low similarity of inter-cluster and high similarity of intra-cluster.
- Knowledge of input parameter: Before clustering, many cluster algorithm required certain information provide by user example: - number of clusters. Clustered accuracy result may be sensitive to such input parameters. But before execution, must defined exact value of parameter is difficult. Hence this may degrade cluster quality.

1.5.1 Cluster analysis requirement

In data mining to analysis the cluster, the following some requirement for clustering algorithm:

- **Feature to handle different type of attribute: -** it must be required that by using clustering algorithm, handle nominal, integer, ordinal, image, document etc data type.
- **High scalability: -** clustering algorithm required high scalability to handle large dataset that difficult to manage
- **To determine input parameter requirement of domain knowledge: -** to calculate the desired number of cluster, many clustering algorithm required the domain knowledge about input parameter. The exact value must be known by domain knowledge of input parameter that controls the quality of clustering.
- **Deal with noisy data: -**the clustering algorithm must be able to handle noisy that that contain outlier, unknown data, erroneous data, or irrelevant data from large dataset.

- **Ability to deal with high dimension data: -** the dataset with two and tree dimension is easy to cluster. But handle to large dataset with multiple dimensions its challenging task for handle through clustering algorithm. Because dataset with large dimension consist of thousand of words that difficult to manage
- **Constraint based clustering: -** the user needs to focus that what kind of data need to be cluster. For example if we want to cluster the data by online website for business strategy we must be constraint that the data is cluster that is benefit for business need.

The good clustering with high quality exists if cluster have high intra-class similarity and low inter-class similarity. Basis on similarity measurement and document representation the quality of clustering is depended.

1.6 Document clustering techniques

1.6.1 Term weighting

We count the frequency of terms in document. Term weighting method applied to assign weight to the document. The term weight with high frequency generate good cluster than term weight with low frequency. The TF-IDF technique used to process the weighting term.

1.6.2 Similarity measurement

The similarity measurement method is used to measuring the similarity between the document by Euclidean distance method, cosine and f-measurement.

1.7 Technique 0f Text Mining

In document clustering the main technique used is neural language processing or information extraction

1.7.1 Natural language processing

It is the area which is related to computer and human interaction. Most of the current natural language processing algorithms are based on the machine learning. In NLP tasks different of classes machine learning algorithms are applied.[2]

NLU have the following parts:

- Tokenization: in this case the sentence is break down forming the token list. This token can be any word or any special symbol.

- Morphological: lexical analysis, we can say morphological can be defined as a process Where words are tagged its corresponding components of speech. A single word can have more than one part of speech.
- Analysis of syntactic: it is the phase where a parse tree is assigned to given syntactic structure for a given natural language sentence.
- Analysis of semantic: the process of translation of syntactic structure of the sentence into a semantic representation which can be expressed in semantics is known as semantic analysis.

1.7.2 Information extraction

Information extraction is process of extracting useful information from the text. With help of information extractions we can extract important patterns, particular patterns. The main objective of information is to discover useful information from various structured and unstructured texts, and then this useful information is used in various fields like business analytics. [3]

1.8 Algorithm Used In Document Clustering

Document clustering applied using the various algorithm such as k-mean clustering, hierarchical clustering, density based clustering, grid based clustering etc.

1.8.1 K-means clustering algorithm

We will select the centred point randomly initially. Then we perform number of iteration basis on the mean value. At each iteration, the new cluster is generate the centroid do not change between iteration.

Algorithm: k-means:-The k-means algorithm is one of partitioning algorithm, in which each cluster's centre is represented by the mean value of the objects in the cluster.

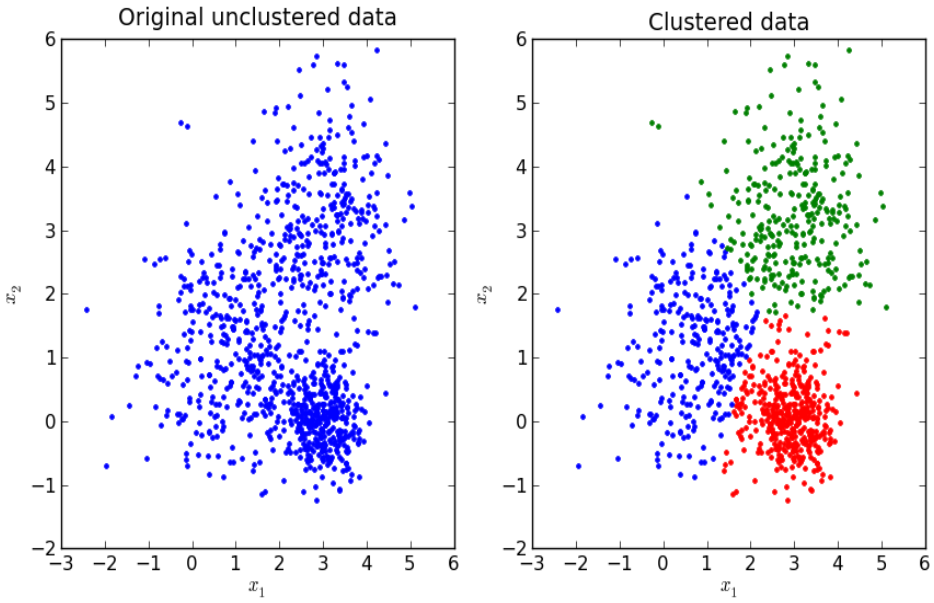

Figure1.4: - shows the clustering using k-mean

Input:

K: represent number of clusters,

D: specify a data set contain n objects.

Output:

A set of k clusters are generated.

Method:

(1) Randomly choose k objects from D as the initial cluster centres;

(2) **Repeat**

(3) (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(5) **Until** no change;

The selection of initial partition can significantly affects the final clusters that result, in terms of inter-cluster and intra-cluster space and consistency.

It is widely used in many applications, but it has following drawbacks:

- Number of clusters that are generated needed to specify in advance. But it is not true and possible in real-world applications.
- It is an iterative technique; the k-means algorithm is mainly sensitive to initial centres choice.

- The k-means algorithm may meet to problem of local minima.
- There is efficiency problem in k-mean clustering algorithm. Like it take large amount of time to perform computation and constructed clusters also does not have good quality.

1.8.2 Hierarchical clustering

The number of cluster is selected basis on hierarchal tree cluster. The hierarchical clustering has two types such as agglomerative hierarchical clustering and divisive method. The agglomerative hierarchical clustering follows bottom to top approach. It merges the common cluster bottom to top. The divisive method clustered the words from top to bottom. It's first splits the data into subpart then cluster the common object. It analysis that the divisive method is more complex than bottom up hierarchical clustering

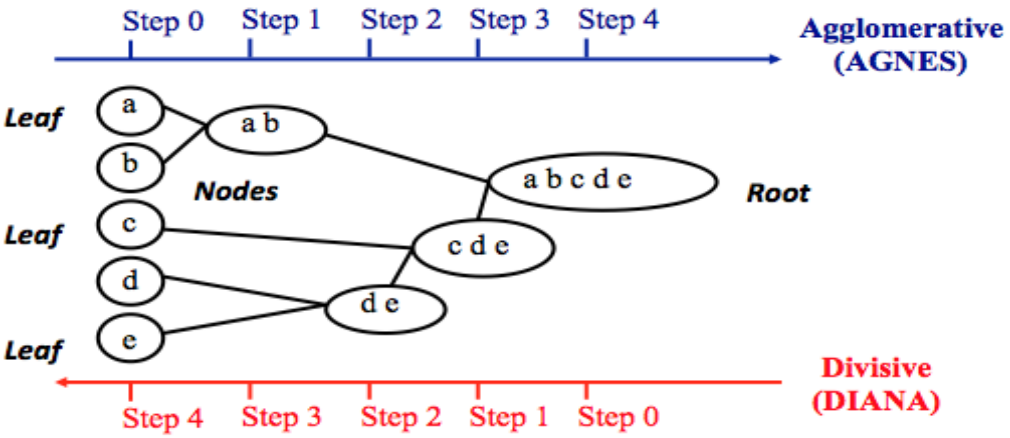

Figure 1.5: - hierarchical clustering that represent agglomerative and divisive method.

1.8.3 Clustering based upon Density

Most partitioning methods cluster objects based on distance between objects. Spherical shaped clusters can be discovered by these methods and encounter difficulty in discovering clusters of arbitrary shapes. So for arbitrary shapes new methods are used known as density-based methods which are based on the notion of density. In these methods the cluster is continues to grow as long as the density in the neighbourhood exceeds some threshold. This method is based on the notion of density. The basic idea is to carry on the growing the given cluster as long as the density in the neighborhood exceeds some threshold i.e. for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points [13].

It helps to discover arbitrary shape clusters. It also handles noise in the data. It is one time scan. It requires density parameters also.

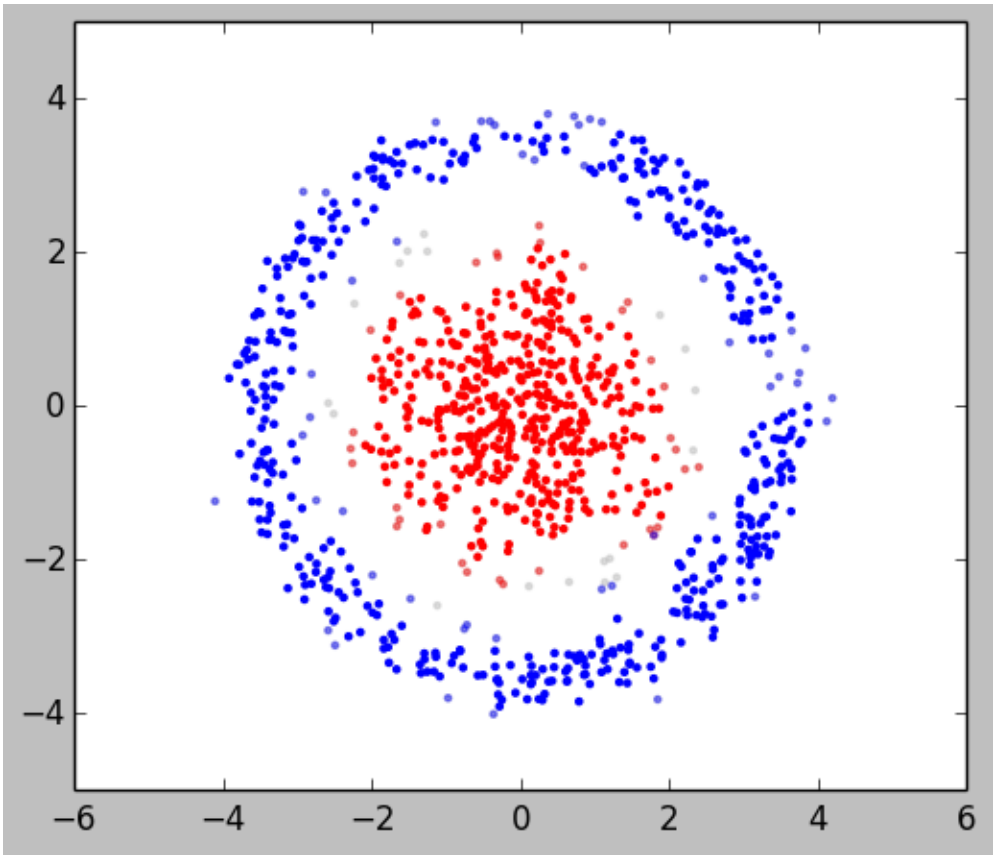

Figure 1.6: - shows the density based clustering

1.8.4 Clustering based upon Grids

Grid based methods quantize the object space into a finite number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space. In this objects are together to form grid. The object space is quantized into finite number of cells that form a grid structure. It assigns to the object grids cells and compute density of each cell. After that eliminate whose density is below threshold t [22]. Now form cluster according to group of dense clusters. In this no distance computations so it is fast process. In this it is also easy to determine which cluster is neighboring.

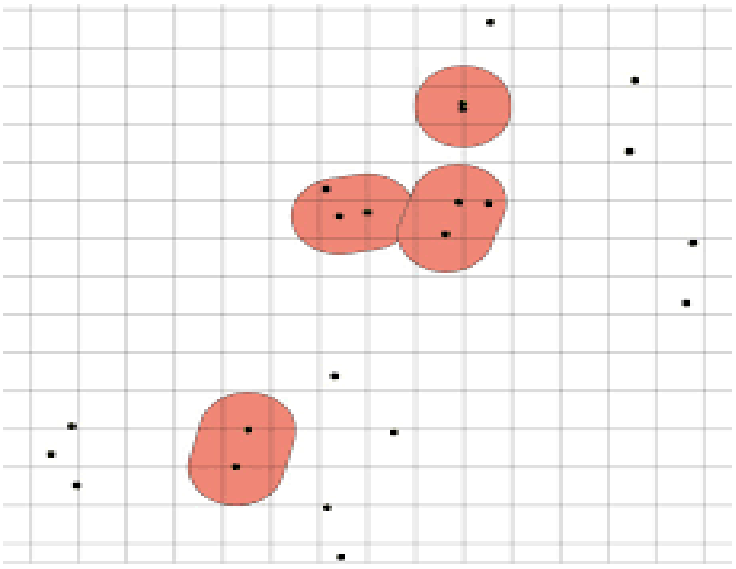

Figure 1.7: - shows the density based clustering

Here shapes are limited to the union. Complexity of the clustering is depends of the grouping of the cells. Grid-based algorithms quantize the space into a finite number of grids and perform all operations on this quantized space. These approaches have the advantage of fast processing time independent of the data set size and are dependent only on the number of segments in each dimension in the quantized space.

1.9 Basic level to measure the text clustering

. The Information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Basic measures for text retrieval are Precision and Recall.

- Precision: This is the percentage of retrieved documents that are in fact relevant to

  the query (i.e., "correct" responses). It is formally defined as:

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{\{\ Retrieved\ \}}$$

- Recall: This is the percentage of documents that are relevant to the query and were, retrieved. It is formally defined as

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{\{\ Relevant\ \}}$$

- An information retrieval system often needs to trade off recall for precision or vice versa. One commonly used trade-off is the F-score, which is defined as the harmonic

Mean of recall and precision:

$$F\_score = \frac{recall * precision}{(recall + precision)/2}$$

Precision, recall, and F-score are the basic measures of a retrieved set of documents.

Text clustering scope and usage

The text clustering applicable in many area to access the text information in sequence order

- The aim of document clustering is finding similar topic based on user query with lesser degree search at minimum time, display the relevant material.
- Based on the user interest, allow the system to identify or extract automatically documents and for a task.
- help the user to drill down and find desired information by guided and interactive search
- The scope of document clustering has in many fields such as
- Online library catalogue
- Online document management system
- Web search system
- In universities, the student record management used to store the large database record in document as clustered form that easily access by user.

CHAPTER 2

REVIEW OF LITERATURE

(Jadhav Bhushan, 2014) Presented Search engine based on fastest reading algorithm which provides best result. When we search or read research paper then its required more time to search. So we apply text mining technique using clustering method to fastest access of information from engine. It's important to access information has grown based on the need of user on search engines. For searching a specific topic, distribution of information take a many hours of search paper. The main aim is to reduce the time inverted in searches. So we will apply the clustering to collect the similar data which is dissimilar to other group. To send the parameter for classification of research papers, clustering will use K-means algorithm. We have easy information categorization by implementing a clustering for fast search and also textual analysis entered by user. The Future work is to implement an automatic learning that increased manipulated text. These techniques allow making a best search engine.

(Oyelade, 2010) This paper tells about how to measure the students' academic skills by using the parameter values for clustering. The clustering analysis is basis on two algorithms comparison such as K-means algorithm and farthest first algorithm. In this paper the data of student of private universities in Indonesia is used. This propose the comparison used between k-means algorithm and farthest first algorithms to measure the performance of both algorithms. We have first for prediction of students' Academic performance, study application of k-means clustering algorithm and the grade point average explained to measure the ability of students. Weka tool apply to study data mining algorithm. It can be observed that the algorithm farthest first and K-Means algorithm can be used for grouping the dataset for contained proximity values. Research using the farthest first takes less time than k-means algorithm and more random access memory needed in the calculation algorithm. But still there in the calculation algorithm farthest first is more number of clusters desired, and for the calculation of the longer the time and RAM required is also greater. In future, to

predict student performance the help of clustering algorithms make it possible to find out the student characteristics.

Shady Shehata et.al (2006) presented paper proposed a new concept-based mining model that are basis on the analysis of the document and the sentences, rather than introduce and analysis of the document dataset only. The mining model analysis of a concept based similarity and terms measurement. The model can efficiently find significant similarity terms basis on words or phrases, according to the semantics of the text of the documents. The sentence semantics analysis term is an-analyzed with sentence and document levels. The similarity between documents can measure basis on a new concept-based similarity measure technique which is used to the matching terms between documents. Using the concept-based term analysis and similarity measure in text clustering are conducted Experiments. The experimental results show that the currently developed concept-based mining model enhances the clustering quality of number of documents.

(K.T.Mathuna, 2015) Presented one of the most popular social media micro blog is Twitter. In this paper the analysis of Twitter data is performed through the text data of millions of users in the word. We extract text data and then apply Pre-processing clustering algorithms are applied on text data. The different clusters formed are compared through various parameters. Visualization techniques are used to represent the results from which inferences like time series and topic flow can be easily made. The observed result shows that the hierarchical clustering algorithm performs batter than other algorithms. The most popular social sites includes you tube Face book, and Twitter. The social media has opened up many research opportunities because of the increased amount of information. The results of this study show the clustering of twitter data. Twitter is fast growing and widely used social networking site on the World Wide Web. Mining of Twitter data has gained importance in the past decades. The better knowledge on information discovery and decision making by integrating both text mining and visualization provides. This work can be further applied on government portals, medical text mining, etc.

(Xiaohui Cui, 2005) Document clustering is used operation such as information organization, summarizing, information retrieval and automatic topic extraction. In this paper we use particle swarm optimization document clustering algorithm to perform search globalized in entire solution space. We take four different text

document dataset and applied PSO, hybrid PSO and k-mean clustering algorithm on them. The similarity matrix Euclidian distance measure and correlation measure are used as in each algorithm. In each experiment for comparison PSO and K-mean approach run 100iteration. In hybrid PSO approach less number of iteration compare to both algorithm. The result shows that hybrid PSO algorithm can generate more effected cluster result than PSO and K-means algorithm because K-means conduct the localized searching but required less number of iteration and PSO algorithm conduct globalize searching but required more iteration. So hybrid PSO algorithm combine both feature of fast access of K-means clustering algorithm and of globalized searching of PSO algorithm. In this way we can avoid the drawback of both algorithms. In future we will also increase the algorithm's efficiency and reduce the complexity of algorithm.

(Xu, 2002) Document clustering is organized the document into group. The problem generate during text document clustering is high dimension. Preprocessing is performing to reduce the dimension of the document vector space. The stemming technique applied for dimension reduction of document vector. Then frequent vector corresponding to each document are represent frequency of term which is applied to find out the similarity words, sentences between every two corresponding document. We used two dataset RTS (Reuters transcribed subset) and newsgroup dataset. Then we applied on it some stage such as preprocessing, frequent term vector generate, then applied the proposed algorithm. The result shows that f-vector produce the high similarity than cosine similarity. There may be many possible improvements that can be implemented on this proposed algorithm. Using other clustering techniques, proposed algorithm can be applied i.e. fuzzy c-means, bisecting k-means etc. This is also being applied on some large dataset.

(Khunteta, 2013) This is presented about document clustering using agglomerative and hierarchical clustering. We also applied document representation, matrix representation, vector space model etc for similarity measurement we have clustering the document by K-means algorithm compare with clustering the document by new approach. For dataset testing we have taken 140 document and 4 classes as mechanical engineering civil engineering, electronics engineering, and computer science. We collected the keyword for each specific branch and we show the 10 keyword for each branch. Then we applied the algorithm by getting document term matrix. The result represent that we have clustering the document by K-means

algorithm compare with clustering the document by new approach. In new approach we get a good cluster than getting the cluster by K-mean. In future work we are planning to make our algorithm more general by performing more improvement over it.

Abdulmohsen Algarnietial (2014) this presented that Term based method efficient computational performance and right way for term weighting. Term based approach to extract many feature from text document but it still include noise. To reduce the noise information from the extract feature, there have much popular text mining technique applied. The stemming words removal also reduces the noise in document.

The text information that contains noise is removed and just consist meaningful information that also reduce the size of document. the neural network applied to training data using clustering algorithm and we also tested document that not all document is use to training only useful document is train the classifier. We extract knowledge from feedback document to describe user need. For information analysis, we use web intelligence and information retrieval (IR) communities. Number of terms is extract from feedback document and phrases have also been used in IR model. Rocchio algorithm is used to filtering information and categorization of text data. BM25 model used for term weights. In future, we will group the training documents using the clustering method. In training process the clusters that contain large amount of noise that can be reduce. It is a big challenge to determine how to cluster is select that contain less noise using neural network.

(Ammar Ismael Kadhim, 2014) Text mining defines the knowledge from unstructured text documents. Text document preprocess have huge effect on extract knowledge. In this paper we implement TF-IDF for term weighting and SDF (singular value decomposition) technique for dimension reduction. We are clustering the document using K-means algorithm. Dimension reduction is important process in data mining and also enhances the performance of clustering technique. Term weighting and dimension reduction step is implement on two dataset, BBC news and BBC sport and. There have multiple stage applied as texts preprocess, term weighting, dimension reduction, document clustering, performance evaluation. The BBC news dataset consist 2225 text document and 9366 words unique. The dataset have 5 classes such as business, entertainment, politics, sport and tech. BBC sport dataset consist 737 text documents and 4613 word unique. The dataset is classified into five classes as athletics, cricket, football, rugby, and tennis. We applied same algorithm (K-means)

on these dataset and compare both dataset. We use the same platform dataset but with varying the sizes of documents. The result shows that the accuracy of these two dataset is approximately same. In future we can apply the different algorithm on different dataset and enhance the performance of clustering.

(Kumar) Document clustering is best technique of finding nearest neighbors of a document and returns a result by search engine that response to a user's query. Document clustering use in different areas of text mining and information retrieval. We compare between K-means andhierarchical clustering. We use a K-means, bisecting K-means and K-means variant. Result shows that bisection k-means technique better than K-means and hierarchical testing approach. Run time required bisecting K-means fast than hierarchical clustering algorithm. In proposed technique first, we run of the K-means algorithm. Then hierarchical clustering technique compared to the single run of K-means. We incremental update of cenroids to improve K-means algorithm. We will also refinement hybrid hierarchical clustering using k-means and hierarchical clustering technique. It produces better result than hierarchical clustering. In future, we focus on to make a better performance of bisecting K-means so that it produces uniformly sized clusters rather than clusters of widely varying size.

(Beil, 2002)Presented the hypertext and text document managed in organization intranets, representing the knowledge of organizations that more important for their success in information society. It is very challenging task to find the content that is relevant for some user due to the huge size of database, high dimensionality, and understandable description of the cluster. We introduce frequent term based approach for text clustering uses frequent item sets. With respect to set of supporting document we measure the frequent term set. We applied two algorithm for frequent term based text clustering, FTC and HFTC. FTC creates flat clustering and HFTC create hierarchical clustering. An experiment evaluate using text document and web document and demonstrate that the purpose algorithm obtain clustering quality more efficient than state text clustering algorithm. We implement all algorithms FTC, bisecting K-means, K-section, K-means. Frequent term set measure using the technique known as apriori algorithm. We use java 1.2 tools because it allows flexible and less time for development. FTC was more efficient than its competitor on all test dataset. HFTC generate hierarchical clustering which are easy to browse. For generate the frequent term set, we use apriori algorithm. We use three datasets such as classic,

Reuters, and WAP. We measure quality of clustering and runtime of each number of clusters. For future research, we need to integration for generate of frequent term sets could significant increase speed of FTC. To solve the frequent term based clustering problem Dynamic programming might also be adoptive using hierarchical clustering are interest for many application but it capture the low quality according to user perspective. Further improvement is required for this method development.

Shady Shehata et.al (2006) in this paper we discuss about the importance of term within the document and the sentence is the statistical analysis of a term frequency (phrase or word) captures. Text mining techniques are used to word or sentences analysis of the text. A concept based model analysis of sentence and the document. The model include of similarity measure and concept-based analysis of term. The model can accurately find matching terms of phrases or words of the document according to the semantics of the text and concept-based similarity measure applied to the matching term to the document. We Experiments using concept-based similarity measure and term analysis in text clustering. The experiment consist three datasets, ACM digital library Reuters dataset and brown corpus dataset. The concept based model is used among document to compute a similarity matrix. For testing the concept-based similarity on clustering, we use three clustering technique k-means nearest neighbor, hierarchical agglomerative clustering and single pass clustering. For measuring the quality of clustering, we use two methods that are F-measure and entropy. Entropy measure that how clusters are homogeneous measure. In the future, number of probabilities for extends this work. Experiment result show that enhances the clustering quality of set of documents by newly developed concept-based mining model. One we improve the accuracy of the similarity calculation of the documents by apply different similarity calculation methods or strategies, secondly we link the present work to web document clustering.

P. Chiranjeevi et.al (2015) Document clustering is the most critical techniques for document organizing in sequence manner. It is the application of cluster analysis to textual documents which include automatic document organization, topic extraction, and fast information retrieval.

Example of document clustering is web document clustering for search engine. A user can locate the document and navigate browse with web search engine. Many document clustering technique cluster the document with low inter similarity and high intra-similarity. Various cluster document technique are effectively globally the

optimal solution can be obtained high speed and high quality clustering algorithms. In this paper we represent the basic step in text clustering such as preprocessing that reduce the noise in text collection by eliminating list of stop words that do not carry semantic meaning. The next process is document term matrix construction, dimension reduction technique text document clustering, and finally optimization of clustering document. We have selected datasets for clustering purpose such as Reuters-21578 which contains 21578 text documents. The Reuters contains 22 files. Each 21 file contains 1000 documents, while the 22 last file contains 578 documents. The document divides into 5 categories: exchange, people, topics, organizations and place. 20 newsgroup datasets contains 20.000 articles from 20 newsgroups on variety of topics. These dataset get from some website on web. We have applied hybrid model of clustering based on GA and PSO which use to solve the clustering problems. To reduce the latent semantic structure, GA used to improve the accuracy and efficiency. It produces much more time for computation due to reduction of dimensions. This paper survey on the research work done on text document, clustering based on extension technique. Our work has to be carried out based on semantic measurement to improve the quality of text document clustering.

(Solka, 2007)Text clustering analysis is the one of the fastest growing research technique and many applications required analyzing large amount of textual data. The text documents have high dimension data, need to reduce noise of such data is very challenging processing. So we use an approach to represent every document using only low dimensions. We have applied the dimension reduction techniques to reduce the dimensions. We experiments the five reduction techniques impact on the accuracy of two supervised classifier on three dataset. We observed that DR can be successful at improving accuracy with low dimensions by compare to using the original word as features. We implement techniques in java, and validated each against existing METLAB. After experiment, the result show that 10 neighbor produce good result. We examined how perform pre-processing with reduce of dimension which can improve text analysis, accuracy to measuring performance using classification. We have evaluated five DR technique. All techniques were able to achieve the improvements rather than compare to not performing DR. We have achieved accuracy using little number of dimensions. In future we should take large dataset and classifier applied DR technique on it.

(Rakesh Chandra Balabantaray, 2013) It introduces the clustering for retrieve the relevant information in a cluster. K-medoids and K-means clustering algorithms used and compare to find which one algorithm is best executed are based on document summarization. Document summarization is executed based on weight assign to sentence to focus on key point of whole documents, which make an easy to people access the information. Dataset f hundred documents were executed for clustering. It has measuring on WEKA tool using k-means algorithm. K-medoids was implemented on java tool. At the end of execution it has experiment that k-means give better result than k-medoids. The cluster form in k-means algorithm is better result than k-medoid. Because using k-means the Manhattan distance of obtain batter than the Euclidean distance. In future, the k-mean algorithm can be used for multi document summarization that help to save the user time by providing the key point of a document that make for user easily retrieve information that they want to access.

(Zhang, 2013) It presents a term weighting technique such as cluster-based term weighting scheme (CBT) for document clustering using term frequency-inverse document frequency (TF-IDF). Our method assigning a term weight using the information obtained from generated cluster. We compare our method using k-means clustering technique with three widely used term weighting scheme such as Norm-TF, TF-IDF, TF-IDF-ICF. We use 20 newsgroup, and Reuter dataset for experiment. Using the stemming algorithm, we removed the stemmed and stop words. We calculate the selective information for the four term weighting schemes, then we computed the average of entropies obtain in each run. We have seen that k-means algorithm with the cluster-base term weighting scheme have generate better results than to other term weighting schemes on each dataset. We observed that some of deviation in the result is due to the k-means clustering algorithm's not better of handling of noise in the data collection.

Sapna Gupta et.al (2014) in this paper we discuss that document clustering is mostly based on the number of occurrences and the existence of keywords. Similarity phrase based clustering technique ignores the semantic behind the words, only capture the order in which the word appear in the sentence. The term frequency based technique take the documents as unstructure words while ignoring the semantic relationship between the words. The concept based clustering technique used to overcome this drawback. In concept based technique, the data set dimensionality is also condensed. It uses medical field heading MeSH ontology for extract the concept and the concept

based calculation is done by its identity and relationship with its synonym. The result is analysis for document clustering using k-means clustering algorithm.

(J.Sthya, 2012) this paper presented the document clustering performing some step such as preprocessing, document clustering and feature extraction. In preprocessing phase term weighting and semantic weight estimate are applied. Feature selection used for each text document in each cluster and semantic relationship is defined by rule mining process on cluster document. We experiment using text document take from IEEE website. We download html pages from website. We remove HTML tag element from web document. The text content maintain separate file. The document analysis is applied to preprocessing and weight estimate process. The stop word eliminates and stemming process is performing on text document. The feature selection is process of extract contents of text document. We experiment on semantic clustering and term clustering operation. The document text content is optimizes of document clustering with semantic analysis.

(Jing 2007) this paper present the clustering the document using K-mean algorithm with reduce dimensional data. We extract the keyword from document and cluster them that the keyword in one cluster may not occur in the document of other cluster. The high dimension data, data sparse problem is generating. In propose work, the K-mean used to calculate the clustering weight for each dimension in each cluster. The weight value identifies dimension subset and different cluster categorized basis on weight. We experiment to decrease the time of clustering process. We just add the step of automatic weighting process in document clustering for dimension reduction purpose.

Shady shehata et.al (2006) this paper presents the document analysis model and sentences analysis using semantic measurement. The term weight and similarity measurement is part of this model. the weight is assign to each word using training set and which word have highest weighting we have clustering them first the second so on. The word is clustered at sentences level and document level. The quality of document clustering can be measure by f measurement and entropy. These weighting term are called concept based model.

(E.V.Prasad, 2015) This paper clustering the document based on relevant meaning of words. The model consist preprocessing of document, calculating the weight based on semantic measurement, similarity measurement of the document, apply the

clustering algorithm. The chameleon algorithm is applied for clustering the document and also hierarchical algorithm is applied to find the possible result. The vector space model used to represent the document. The information is retrieved based upon large database. It is very challenging task to manage the large data. The classification is used to classify the data into common classes. The similar kind of data is collected and divided according to classes. the classification used for two purpose.

- To classify the keyword or index set.
- To classify the text document into classes subject

. (Alfawareh S. J., 2012)In this paper various techniques and challenging issues in text mining are used. In these days text mining is very interesting area of mining, document clustering also known as text mining. The two main techniques for text mining are:

- Natural language processing
- Information extraction

Natural language processing used as processing the text dataset for information extraction and unstructured data in sequence form. Then clustering applied to cluster the irrelevant dataset. Stop word and stemming is removing during preprocessing. After extraction the information it can used in various database for future used, these information can be used with the help of queries fired on the database or the information can be used in commercial or business world for analysis. There is various application of text mining specially the structured information mining. the information used as explore the pattern various model to be construct..

1. The first area where text mining is used is bioinformatics which can be also referred as biomedical text mining.

2. National security is other area where it is used; text mining is used by many government organizations for investigation purpose. Text mining is one of the important technologies for national security defense.

3. The text mining is used in business intelligence for taking decision. For making any decision first we have to analysis the data which is done with the help of text mining techniques for the prediction. The problem faces in natural language complexity is that the one word has multiple meaning so it’s difficult to cluster the these kind of data

(Shaikh, 2010) this paper present that document clustering using similarity measurement between the document using the semantic measurement technique. In the document we assign the value between of 0 to 1 for measurement the similarity function. The 1 is defining that the word is similar to each other and 0 indicate that the word is dissimilar to each other. For compute the document similarity or represent the document, the vector space model is used. The text document is written by common language such as human language that contains word and context that are semantically related. In this paper we represent the similarity measurement based upon the topic map which is the industrial standard with constraint extraction and search. This paper present that new similarity measurement technique is more attractive rather than commonly used approach.

Nicola cinefra et.al (2012) this paper presents the dataset with portioning points into different group. in this clustering the two point are semantically similar and other two point are dissimilar to each other in clustering. Similarity measurement function applied to measure the similarity in dataset. The semantic classification model is used for extract the complex feature. The purpose of this technique is to discover the path to more and wide general knowledge about the meaning related to the component and in semantic unit they association them.

(Adrain Kuhn, 2006) In this paper they explained many of the existing approaches in Software Comprehension focus on program program structure or external documentation. However, by analyzing formal information the informal semantics contained in the vocabulary of source code are over-looked. To understand software as a whole, we need to enrich software analysis with the developer knowledge hidden in the code naming. This paper proposes the useof information retrieval to exploit linguistic information found in source code, such as identifier names and comments. They introduced Semantic Clustering, a technique based on Latent Semantic Indexing and clustering to group source artifacts that use similar vocabulary. They call these groups semantic clusters and they interpret them as linguistic topics that reveal the intention of the code. They compare the topics to each other, identify links between them, provide automatically retrieved labels, and use a visualization to illustrate how they are distributed over the system. Their approach is language independent as it works at the level of identifier names. To validate their approach we applied it on several case studies, two of which they present in this paper [15].

Adrain Kuhn et.al (2006)In this paper they explained many of the existing approaches in Software Comprehension focus on program program structure or external documentation. However, by analyzing formal information the informal semantics contained in the vocabulary of source code are over-looked. To understand software as a whole, we need to enrich software analysis with the developer knowledge hidden in the code naming. This paper proposes the useof information retrieval to exploit linguistic information found in source code, such as identifier names and comments. They introduced Semantic Clustering, a technique based on Latent Semantic Indexing and clustering to group source artifacts that use similar vocabulary. They call these groups semantic clusters and they interpret them as linguistic topics that reveal the intention of the code. They compare the topics to each other, identify links between them, provide automatically retrieved labels, and use a visualization to illustrate how they are distributed over the system. Their approach is language independent as it works at the level of identifier names. To validate their approach we applied it on several case studies, two of which they present in this paper [15]

Anwiti Jain et.al (2012) in this paper, the k-mean algorithm is modified to improve the time of text clustering. The k-mean algorithm is applicable for locally optimal solution. Modified k-mean reduce the probel of local searching and also decrease the time for execution. The result show that the modified k-means algorithm for larger document and smaller document take less time and also increase the quality of clustering. the modified k-mean generate less nmber of erroneous and outlier as compare to k-means.

(Sharma, 2010)In this paper, introduce that the major data mining task is cluster is aim to grouping the object which have similar to each other and dissimilar object is grouping in another cluster. the similarity of one group is maximizing than dissimilarity of other group is minimizing. The k-mean cluster technique applied to perform clustering. The benchmark dataset used to implement the clustering technique. The objective of this work is to make the cluster in a way that tahat increase the quality of text clustering.

(Zhao, 2012)This paper present the work on side information available on social network, web, online site etc. these site information consist document link, text information, from web log user behavior to access etc. so we cluster the text

information using side information using large dataset. The large dataset is an unstructured form as mixture of irrelevant data or relevant data. So we filter that data using the partitioning technique for text clustering. This is challenging task to operate on side information because it may be improve the quality of text document or may be add the noisy data into text document. We combine the probabilistic model and classical portioning algorithm to create the effective approach of clustering. The work can be done using real dataset using different technique. The result shows that the using side information the text clustering quality is more increased.

(Kulkarni, 2013) In this paper, they introduced Explosive and quick growth of the World Wide Web has resulted in intricate Web sites, demanding enhanced user skills and sophisticated tools to help the Web user to find the desired information. Finding desired information on the Web has become a critical ingredient of everyday personal, educational, and business life. Thus, there is a demand for more sophisticated tools to help the user to navigate a Web site and find the desired information. The users must be provided with information and services specific to their needs, rather than an undifferentiated mass of information. For discovering interesting and frequent navigation patterns from Web server logs many Web usage mining techniques have been applied. The recommendation accuracy of solely usage based techniques can be improved by integrating Web site content and site structure in the personalization process. Herein, they propose semantically enriched Web Usage Mining method (SWUM), which combines the fields of Web Usage Mining and Semantic Web. In the proposed method, the undirected graph derived from usage data is enriched with rich semantic information extracted from the Web pages and the Web site structure. The experimental results show that the SWUM generates accurate recommendations with integration of usage, semantic data and Web site structure. The results shows that proposed method is able to achieve 10-20% better accuracy than the solely usage based model, and 5-8% better than an ontology based model

(Twinkle Svadasa, 2014) This paper present that the text clustering apply on online web textual data. Because of growth of text information on, web the information search on a web is also increased day by day. The information that the user want to search on the web need to available and information should be relevant. When the user search for some information, According the user search the maximum output is relevant according user search. Thus the text mining technique is applicable for this purpose.

We cluster the data for extract the meaningful information from web. The ontology and semantic term approach is used to enhance the cluster quality.

(Sharma, 2012) This paper present that the text clustering used to take business related decision. The knowledge based model is developed using the text clustering. The business related decision is taken by collect the information from web or analyzing that information in sequence form using clustering technique. The clustering is used in business to provide the better management of information. This paper discuss about various different technique apply for text clustering. The large text data is difficult to cluster because of high dimensionality of it. So its very challenging task to handle large amount of data.

(Macnaught, 2009)In this paper from two or more different sources they explained that the alignment of definitions. Without changing the meaning of the concept, it is possible to retrieve pairs of words that can be used indistinguishably in the same sentence. As lexicographic work exploits common defining schemes, such as genus and differentia, a concept is similarly defined by different dictionaries. The difference in words used between two lexicographic sources lets us extend the lexical knowledge base, so that clustering is available through merging two or more dictionaries into a single database and then using an appropriate alignment technique. Since alignment starts from the same entry of two dictionaries, clustering is faster than any other technique. The algorithm introduced here is analogy-based, and starts from calculating the Levenshtein distance, which is a variation of the edit distance, and allows us to align the definitions. As a measure of similarity, the concept of longest collocation couple is introduced, which is the basis of clustering similar words. The process iterates, replacing similar pairs of words in the definitions until no new clusters are found [19].

J.L Stricker et.al (2002)In this paper they introduced that the original California Verbal Learning Test (CVLT) employed a semantic clustering index that used the words recalled during a given trial as the baseline for calculating expected values of chance clustering. Although commonly used in cognitive psychology, clustering indices that use recall-based calculations of expectancy are implied by the assumption that organizational processes do not occur until after words are retrieved from memory. This assumption contradicts the generally held assumptions among

neuropsychologists that (1) organization is an antecedent to recall, and (2) increases in the use of organizational strategies will result in better recall performance.

(Priyadarshani, 2012)In this paper they introduce the method of feature selection. This method increases the accuracy of text clustering and quality. The feature selection technique reduces the irrelevant words and irrelevant word from text clustering. The text clustering with feature selection is reducing the unwanted words or stop word from document. The learning process technique applied to learn the training data. The propose system is design to identify the semantic relationship. The ontology used to represent the term. The statistical method is used in text clustering for future selection and represents the text clustering.

Neelima Bhatia et.al (2015) In this paper we present the DLVN (deep-learning vocabulary network). . The existing term frequency-based methods only calculate the number of words, but the relations of words are not considered in feature extraction. The approach constructs vocabulary network to mine the importance of words using related-word set, which contains “co occurrence” relations of words. so the text with same feature are short distance than text with different feature are longer distance among different category Text clustering is an effective approach to collect and organize text documents into meaningful groups for mining valuable information on the Internet. They also used the technique of feature extraction and data dimension reduction. We present a novel approach as deep-learning vocabulary network. The vocabulary network is deployed based on related-word set, which contains the “co-occurrence” relations of words or terms. Future sparse-group deep belief network is proposed to reduce the dimensionality of feature vectors, and we introduce coverage rate for similarity measure in Single-Pass clustering. To verify the effectiveness of our work, we compare the approach to the representative algorithms, and experimental results show that feature vectors in terms of deep-learning vocabulary network have better clustering performance.

Saiyed Saziyabegum et.al (2016) this paper presents the work on information available on internet. Important information can be considered by creating summary from available information. Manual creation of summary is complicated task. Therefore research community is developing new approaches to for automatic text summarization. Automatic text summarization system creates summary. Summary is

shorter text that covers important information from original document. Summary can be created using extractive and abstractive methods. Abstractive methods are requires deep understanding of text. After understanding, it represents text into new simple notions in shorter form. Extractive approach uses linguistic and statistical approach for selection of sentences for summary. This paper presents an survey of recent text summarization extractive approaches developed in last few decades. Summary evaluation is also covered briefly in this paper. Finally this paper ends with conclusion of future research needed. Text summarization is motivating field of research and it has variety of applications. The objective of this paper is to study some important information related to the past of automatic text summarization and current trends. In this paper, more focus is given to Text summarization extractive approaches and they are categorized into different categories.

Ladda Suanmali et.al (2014) this paper present the Automatic text summarization used is a wide research area. There are several ways in which one can characterize different approaches to text summarization: extractive and abstractive from single document or multi document, characteristic of text summarization, level of processing. The most researchers for automatic text summarization have transferred their efforts from single document summarization to multi document summarization but they have to be aware of the issues of redundancy, ordering in sentence, etc. In this paper, we compare techniques that have done for multi document summarization. We describe evaluation method for automatic text summarization to weigh the quality and performance of system. The future work performs on multi document summarization. We will try to develop an algorithm or new model that supports multi document summarization area combine Lexical Chain with cluster-based method.

CHAPTER 1

PRESENT WORK

3.1 PROBLEM FORMULATION

Here, we are going to discuss about the proposed technique that is text clustering. In this technique we are clustering the document in way that the user can easily access the information or read the text information in dictionary order or in index form. In existing technique we have applied the weight base LDA algorithm. That calculates the each word by assigning the weight to every word in document. But this technique is time consuming or not simplify to calculate each word in document. In the present work, we are going to apply text clustering using DMNB algorithm that does not need to assign weight to each word they just collect word randomly and cluster the similar type of word. These methods increase the accuracy of text clustering and also simplify as compare to LDA algorithm.

3.2 OBJECTIVE OF THE STUDY

The following objective applied in document clustering:

- To Study and analyze text clustering algorithm in data mining
- To propose enhancement in the semantic clustering algorithm to increase cluster quality
- To compute the effectiveness and performance of proposed technique on the basis of accuracy and execution time.

3.3.1 Existing Algorithm Methodology

We apply the neural network technique with semantic based analyzer. First, we will read the text file from the database then define the number of neurons for the network that will act as an input. The input data that has been selected, it must be preprocessed that is done in the pre-processing layer and then comes learning layer, in this layer Learning is occurred by changing the connection weights after each word is processed, based on the amount of error(Error = expected value - actual value). After that there will be training network. So with this process one word tries to attach for matching many other words for creating efficient synonyms. This methodology will

reduce the processing time and reduce the algorithm escape time. The following step performs using this technique:

- Fist we load text data that include the number of words that are in a random position .we have cluster data in way that the output will be effective and in serial order.
- All value which we take that is from mat file we have to convert into ascii format then after process is continue.
- Then latent Dirichlet allocation algorithm is applied in which we assign the weight to each word in document. The word which has maximum weight we have clustered them. The iteration is continuously performed until all word is not clustered.
- Under LDA they have to find phi, Perplexity and error rate.
- Then the document is clustered in serial form.

3.3.1.1 Latent Dirichlet Allocation

LDA algorithm is weight based algorithm used to observation number of words in document. The document is mixture of various words. The various steps involved in document clustering through LDA algorithm as below:

- First we computing the number of topic (K) in document.
- Then we compute number of N word in vocabulary (example: 60,000 to 10,00000)
- Now we calculate number of words in all documents by sum of all N (d) value.
- αk =Then we assign the weight to topic k in document
- (α) collect all αk value and view as single vector
- βw now we assign weight of word w in topic usually same for all word.
- β collect all beeta word value and view as a single vector.
- Now we check the number of word occur in document d
- (Z)Then we identify the number of word in document.
- (W) Identify of all word in all document.

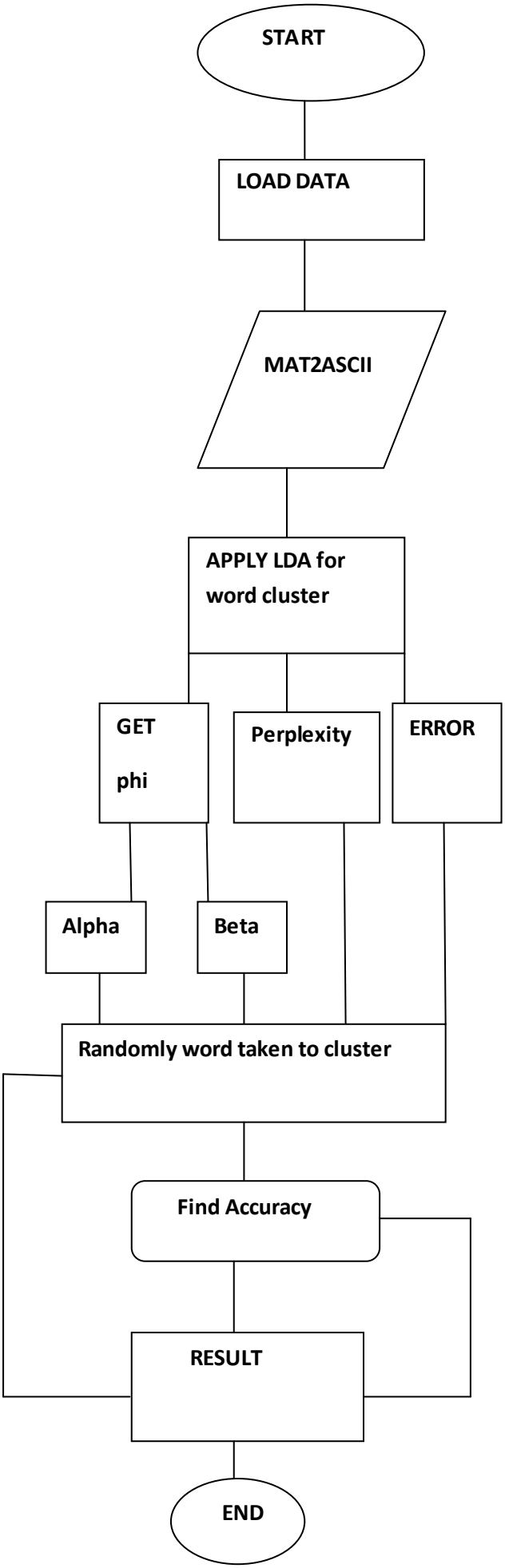

Figure3.1: - Flow chart of existing work

3.3 RESEARCH METHODOLOGY

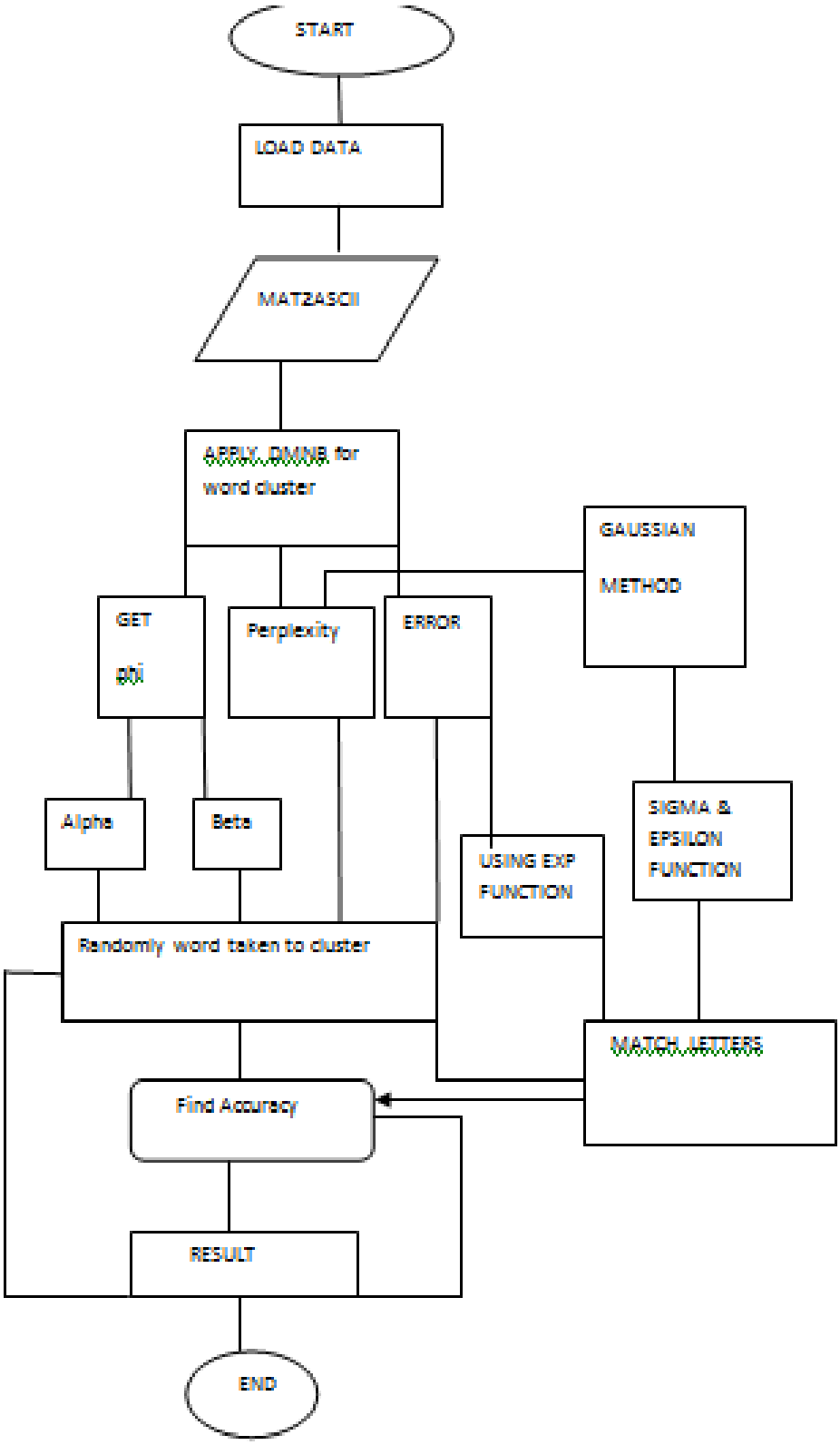

Figure3.2: - Research Methodology

STEP 1: - First we load the data that is in unstructured forms .we have clustered like that the output will be effective and in serial order.

STEP 2:- The text file which we take that is from mat file we have to convert into ascii format.

STEP 3: - Then Discriminative mixed-membership naive Bayes (DMNB) applied. The Neural networks is applied which will traverse the whole dataset and find the number of members in each cluster

STEP 4: -Under LDA we have to find phi, Perplexity, error. We add the Gaussian functionality computation using DMNB.

STEP5:- the random word are taken to cluster the data

STEP 6: - The alpha function compute to cluster the number of topic in document basis on the similarity and beta function compute to cluster the number of words under document basis of the similar type of words. The perplexity defined that sets the number of effective nearest neighbors.

STEP 7: -Error will be occurring less when no of iterations will continue. Under Gaussian method we perform sigma and epsilon value which help us to match letter. It means that we cluster the alphabet e in serial order like if first alphabet is A, after B so on. The output will be in form of A-Z alphabet.

STEP 8: -Then document is clustered in serial alphabetic.

3.3.1 Implementation of tool: - To implement this technique we will use MATLAB tool. MATLAB tool is widely used in all areas of research universities, and also in the industry. It is beneficial for mathematics equations (linear algebra) moreover numerical integration equations are also solved by MATLAB. It also provides a one of the simplest programming languages for writing mathematical programs. It has various types of tool boxes that are very beneficial for optimization and so on. MATLAB is used for machine learning, signal processing, communications, computational finance, control design, robotics, image processing, etc. The MATLAB platform is used to solving scientific and engineering problems. MATLAB helps you take your ideas applied on the desktop. You can run your analyses on larger data sets

and scale up to clusters and clouds. MATLAB code can be integrated with other languages, enabling you to deploy algorithms and applications within web and production systems.

CHAPTER4.

RESULT AND DISSCUSSION

4.1 EXPERIMENTAL RESULTS

The final result that we have obtained by implements the LDA and DMNB algorithm. After implementation, we got that the text clustering using DMNB algorithm simplifies the clustering process and increase the efficiency of document clustering. The result shows that we have got the document with dictionary or index form in sequence order than unstructured document.

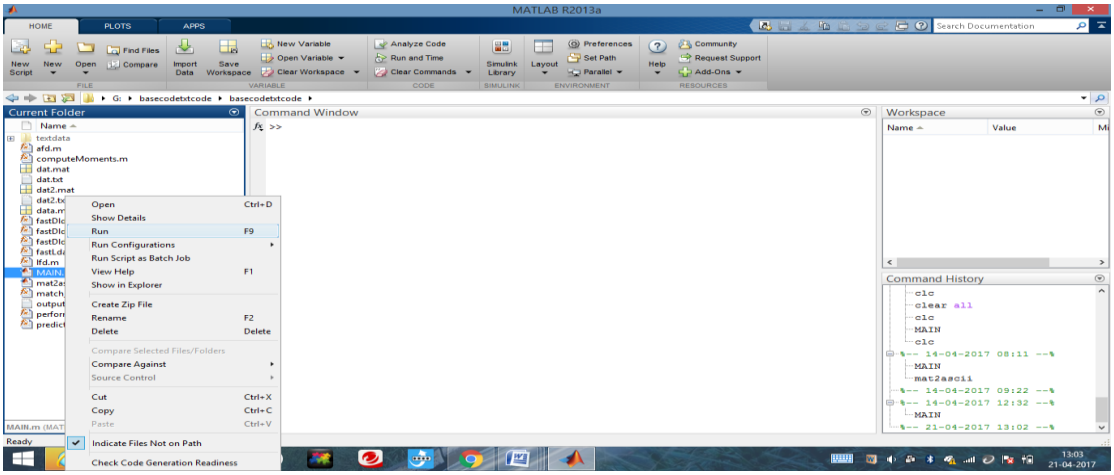

Figure 4.1:- Loading of Basic code

As shown in figure 1, the base code of text clustering has been implemented in which the LDC function is applied which cluster the similar words of the dataset.

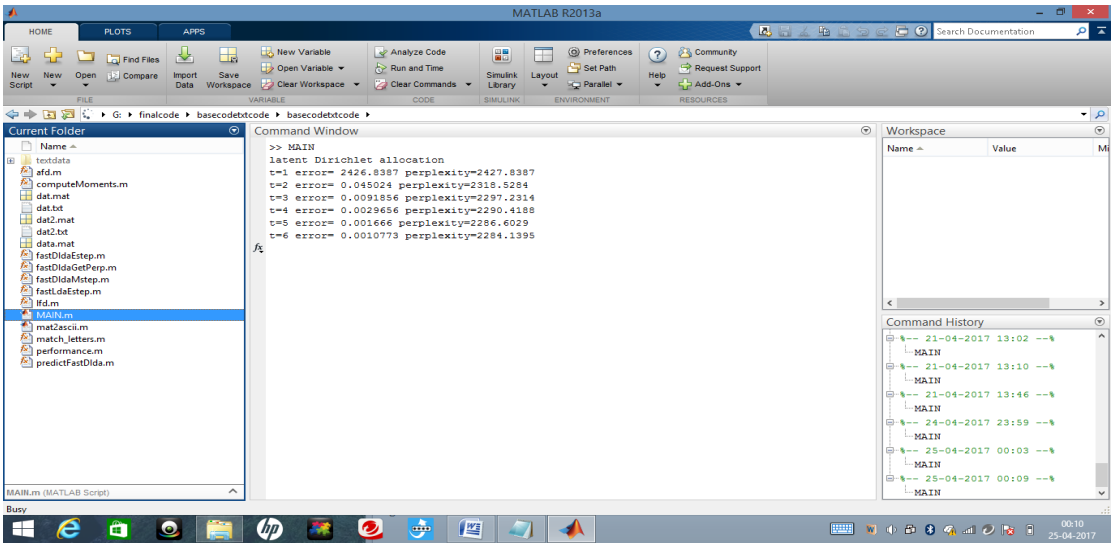

Figure 4.2: - LDC function applied

As shown is figure, Start LDA algorithm in which we calculate their error, perplexity, and accuracy of output result. According to input iteration will be changes.

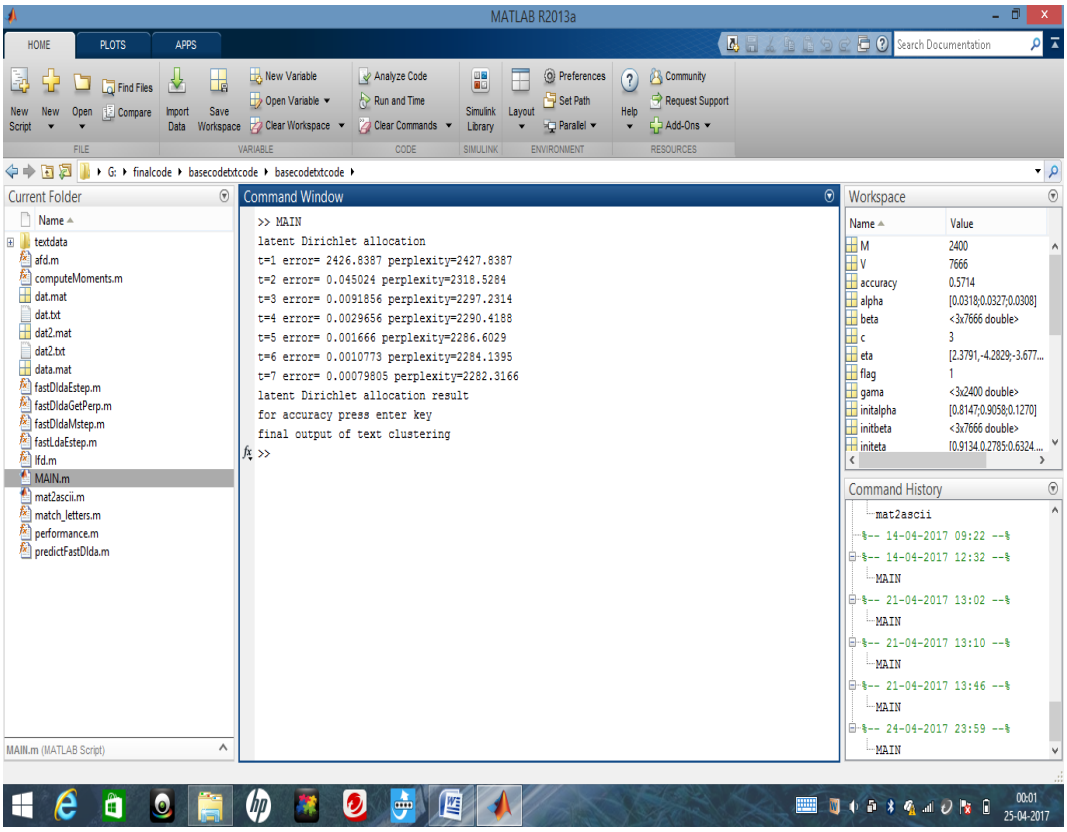

Figure 4.3: -LDA function applied

As shown in figure 3, all iterations done now LDA algorithm complete now we have to press enter for getting result in text file or we calculate their accuracy of word cluster data.

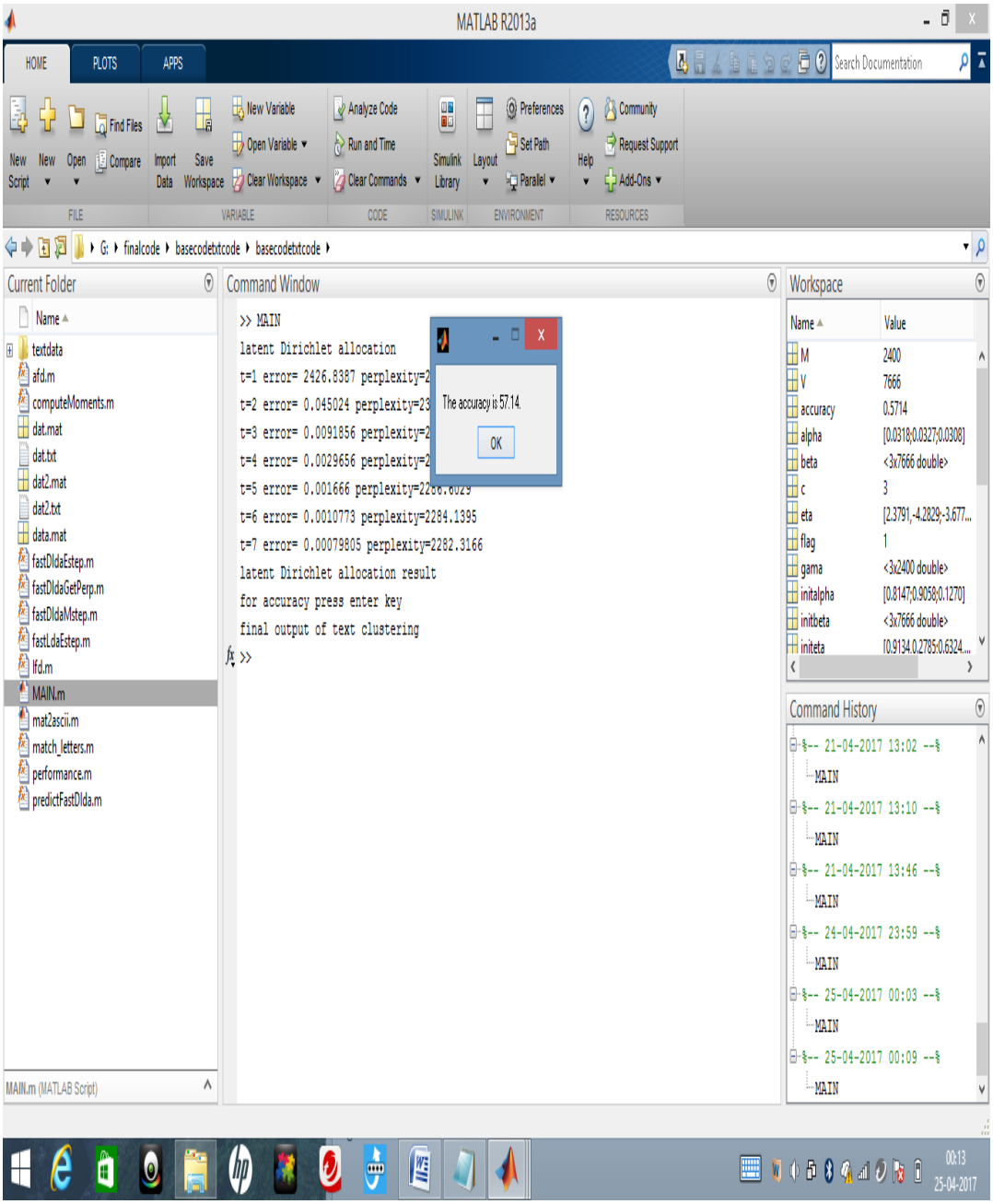

Fig 4.4:-Accuracy calculation is 57.14

The result show that the multiple iteration perform using LDA algorithm and we find out the accuracy of text clustering with 57.14. we further enhance the accuracy by using other algorithm.

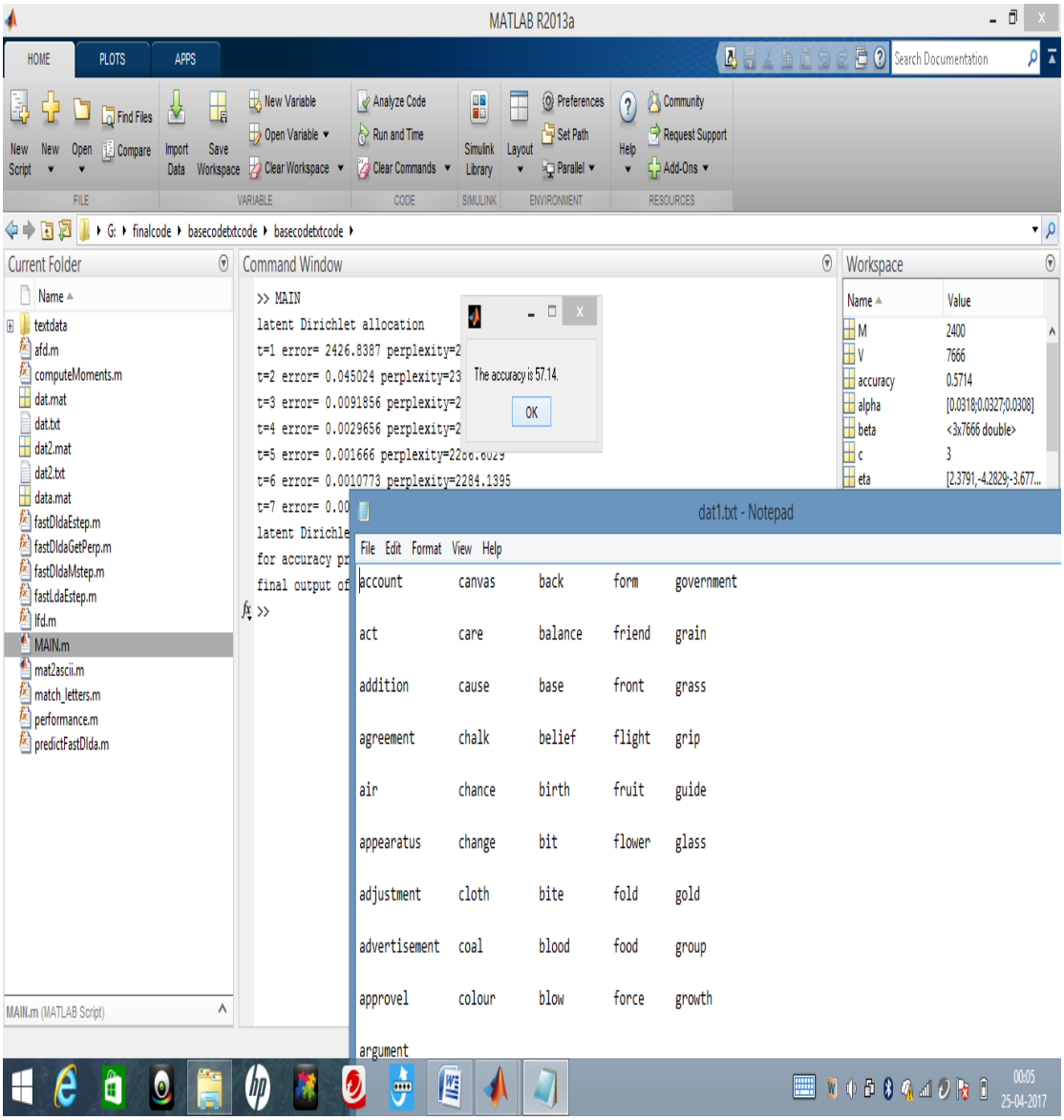

Fig 4.5:-Accuracy calculation and Clustering of Data file

As shown in figure 4, after press enter we getting accuracy result is 57.14 and show text file result in which word will show in column according their alphabet series.

As shown in figure Final output in which we having a cluster data result acc to random series of alphabet.

4.2 Comparison with existing result

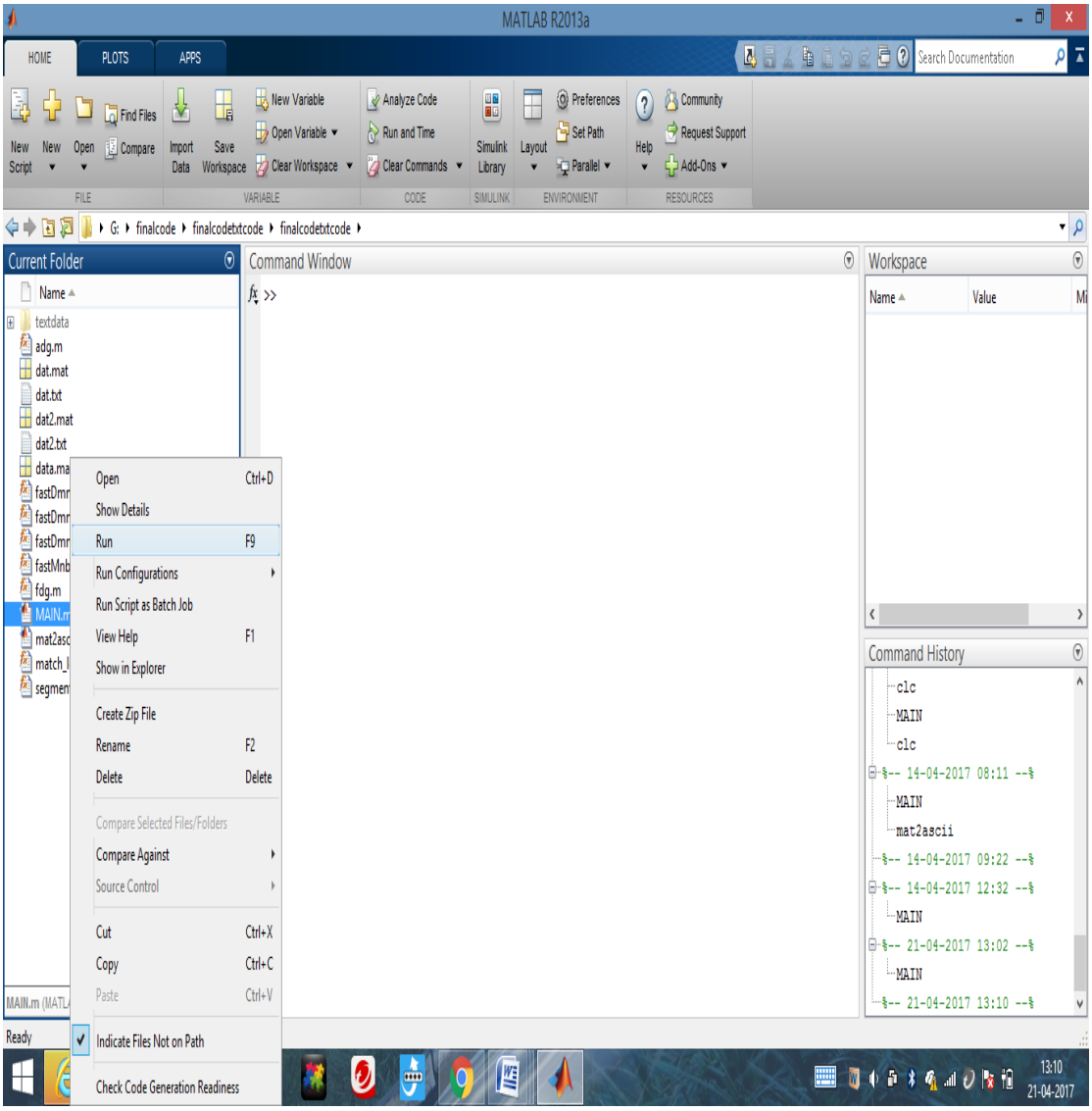

Fig 4.6: - Coding of research methodology

As shown in figure 6, RUN the main file having name MAIN.M in final code by write click on this file. Then after that number of iteration is working using DMNB algorithm.

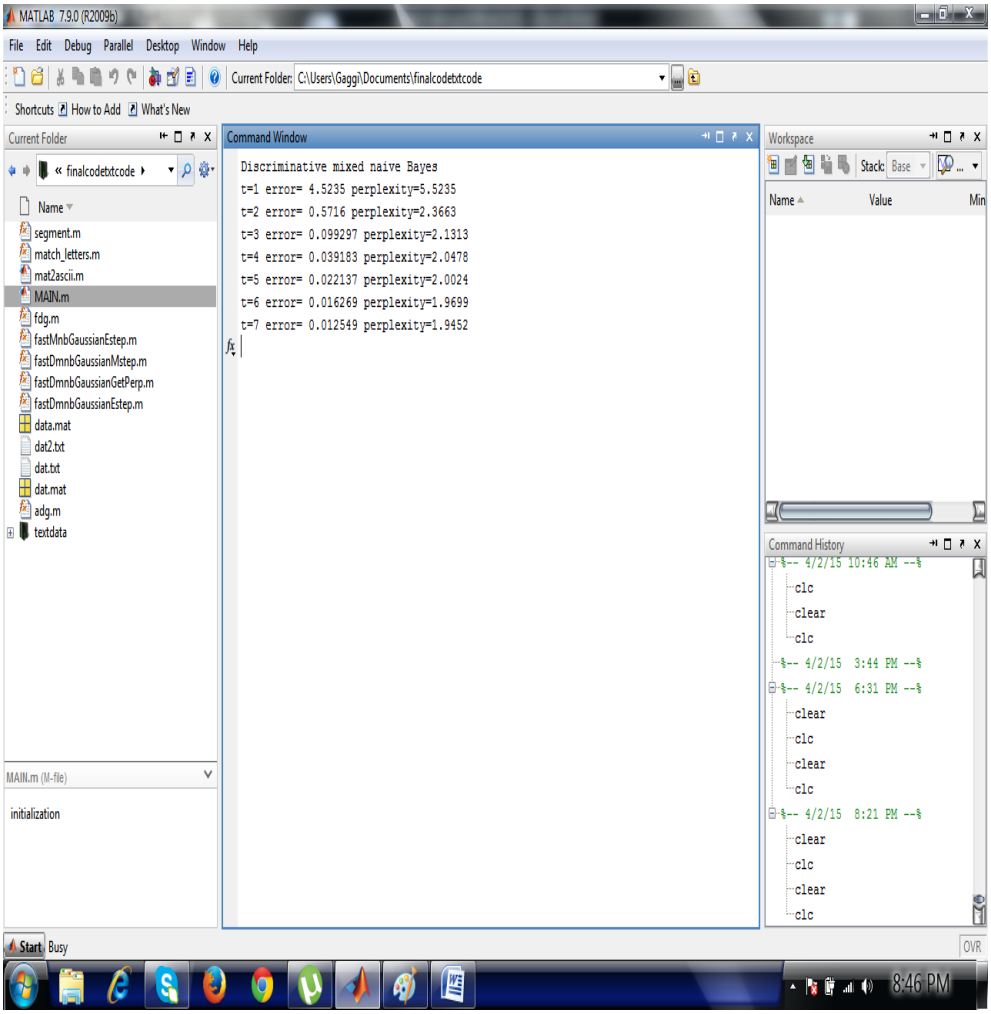

Fig 4.7: - DMNB algorithm implementation

As shown in figure 7, Start DMNB algorithm in which we calculate their error, perplexity, and accuracy of output result. Acc to input iteration will be changes.

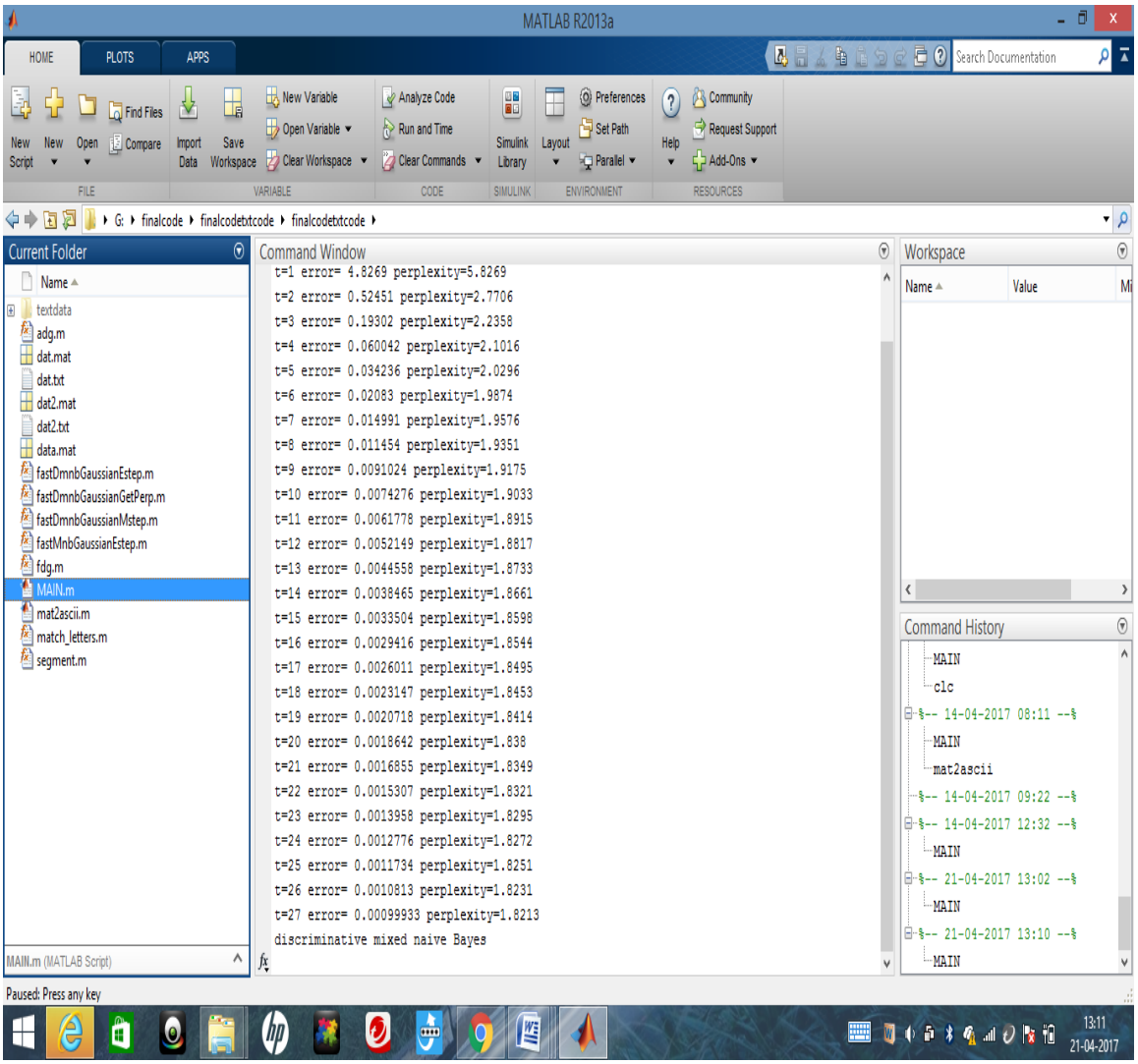

Fig 4.8:- DMNB algorithm execution complete

As shown in figure 8, All iterations done now DMNB algorithm complete now we have to press enter for getting result in text file or we calculate their accuracy of word cluster data.

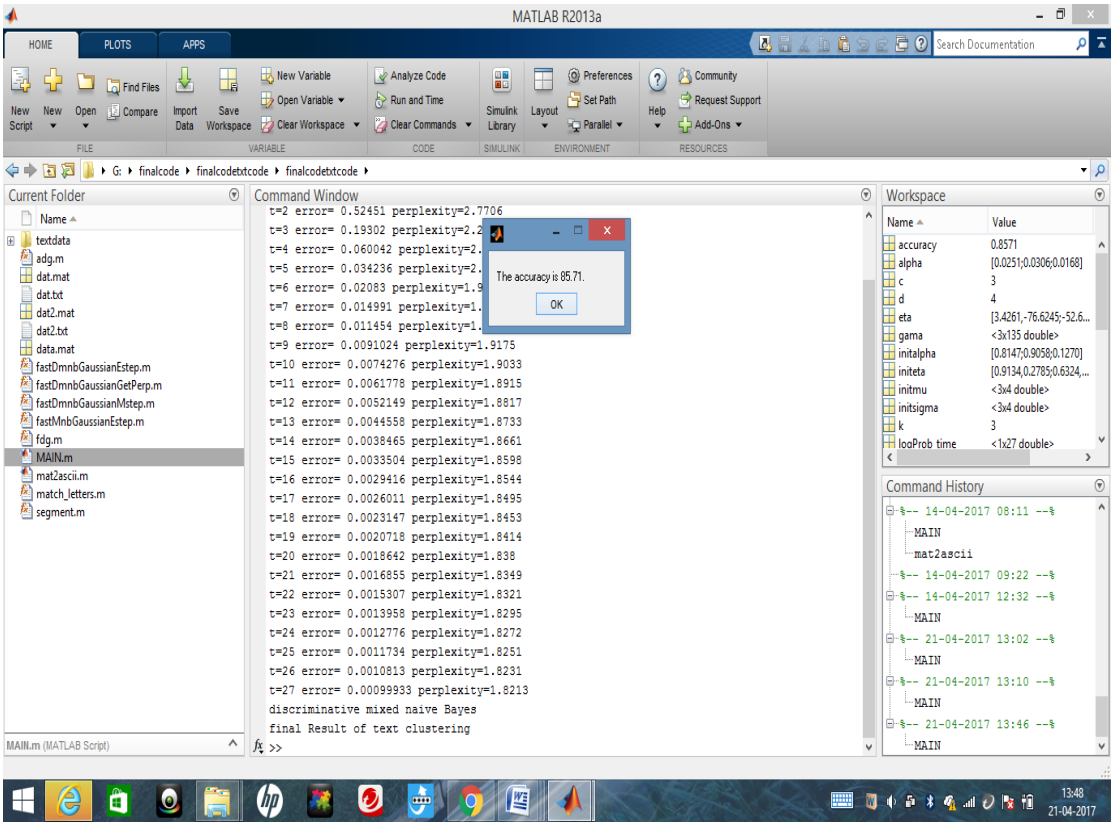

Figure 4.8:- Accuracy calculation

The figure shows the accuracy calculate using DMNB algorithm is 85.71 that give better result than existing technique.

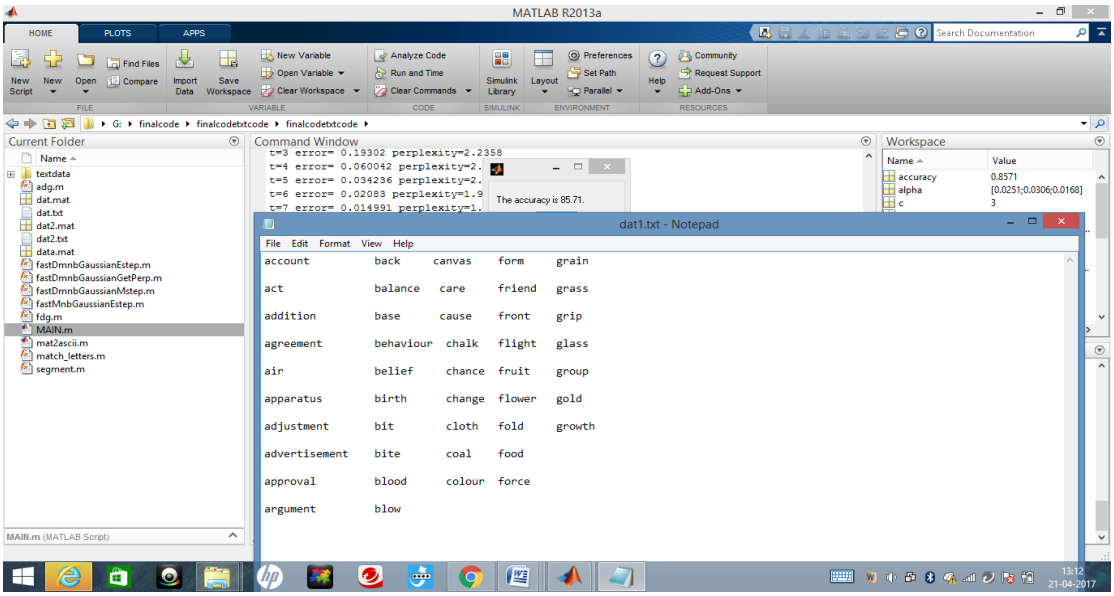

Fig 4.9:- Final clustering result

As shown in figure 10, Final output in which we having a cluster data result acc to Z-A series or will be A-Z series.

As shown in figure 8, after press enter we getting accuracy result is 85.71and show text file result in which word will show in column according their alphabet series.

- In present work we have show that the text clustering using DMNB gives more accurate result compare to existing technique. In the existing technique using LDA algorithm the accuracy is 57.14. The text clustering increase the accuracy using DMNB algorithm as 85.71.
- The present work is simplified the clustering rather than using existing technique.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The overall aim of this thesis work was to evaluate the document clustering that improves the easily information retrieval in document by user who wants to access information. Most specifically focus on improvement in semantic measurement technique to achieve the good clustering of document. We improved the execution time and accuracy using DMNB algorithm. We also simplify the text clustering using DMNB algorithm.

5.2 Future scope

There are number of ways to extend the work. First link it with the web documents. The LDA algorithm limited to support the heterogeneous feature collect. To improve the cluster qualities we can use hybridize the DBMS and LDA algorithm or increase the quality of text document.

REFERENCES

[1] Alfawareh, S. J. (2012). , "Techniques Applications and Challenging Issue in Text Mining uses, Applications",. *,IJCSI International Journal of Computer Science Issues , 9* (6).

[2] Jadhav Bhushan G1, W. P. (2014)," Searching Research Papers Using Clustering and Text Mining", *International Journal of Emerging Technology and Advanced Engineering , 4 (4).*

[3] Sung-min Kim, Y.-g. H. (2016). ,"Automated Discovery of Small Business Domain Knowledge Using Web Crawling and Data Mining", . *Konkuk University, Department of Computer Science and Engineering*

[4] Oyelade, O. J. (2010).," Application of k-Means Clustering algorithm for prediction of Students' Academic Performance.", *(IJCSIS) International Journal of Computer Science and Information Security , 7.*

[5] Ilyas2, M. H. (2015)," A Clustering Based Study of Classification Algorithms", University of Sargodha , 1Dept of Computer Science and Information Technology, Sargodha, Punjab, Pakistan .

[6] K.T.Mathuna, I. S. (2015)," Applying Clustering Techniques for Efficient Text Mining in STwitter Data", *International Journal of Data Mining Techniques and Applications , 04* (02 December 2015), 25-28.

[7] Xiaohui Cui, T. E. (2005),"Document Clustering using Particle Swarm Optimization", *IEEE* .

[8] Khunteta, C. J. (2013). A New Approach of Document Clustering. International Journal of Advanced Research in Computer Science and Software Engineering , 3 (04)

[9] Kumar, M. S. (2014)," *A Comparison of Document Clustering Techniques",* University of , Department of Computer Science and Egineering

[10] Beil, F. (2002)," Frequent Term-Based Text Clustering", Institute for Computer Science Ludwig- Maximilians-Universitae

[11] Rakesh Chandra Balabantaray, C. S. (2013)," Document Clustering using K-Means and K-Medoids", *International Journal of Knowledge Based Computer System , 1* (1).

[12] Zhang, A. K. (2013)," A New Term Weighting Scheme For Document Clustering*", University of Kentucky, 1Department of Computer Science, Lexington*.

[13] Shehata, S. (2006)," *Enhancing Text Clustering using Concept-based Mining Mode",* Proceedings of the Sixth International Conference on Data Mining (ICDM'06) .

[14] Lehal, V. G. (2009). , "A Survey of Text Mining Techniques and Applications",. *Journal of Emerging Technologies in Web Intelligence, VOL. 1, NO. 1, August,* .

[15] Ammar Ismael Kadhim, (2014). Text Document Preprocessing andDimensionReduction Techniquesfor Text Document Clustering. *Universiti Sains Malaysia, 1School of Computer Sciences. Malaysia : International Conference on Artificial Intelligence with Applications in Engineering and Technology.*

[16] Beil, F. (2002)." Frequent Term-Based Text Clustering", *Institute for Computer Science Ludwig- Maximilians-University*

[17] Jusoh Shaidah and Alfawareh Hejab M., "Techniques Applications and Challenging Issue in Text Mining uses, Applications", *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814*

[18] Shehata Shady, "Enhancing Text Clustering using Concept-based Mining Model", Proceedings of the Sixth International Conference on Data Mining (ICDM'06) 0-7695-2701-9/06, 2006

[19] Khare Akhil, Jadhav Amol N., "An Efficient Concept-Based Mining Model For Enhancing Text Clustering**",** *IJAET, Vol.II, Issue IV/October-December, 2011*

[20] Shehata Shady, "A Word Net-based Semantic Model for Enhancing Text Clustering", *IEEE International Conference on Data Mining Works shops, IEEE, 2009*

[21] babak, f. (2013)." A Novel Document clustering Algorithm Based On Ant Colony Optimization Algorithm", *journal of mathematics and computer science ,vol.7, pp.171-180, 2013*

[22] Drakshayani B. and Prasad E.V., "Semantic Based Model for Text Document Clustering with Idioms", *International Journal of Data Engineering (IJDE),Volume(4):Issue(1):2013*

[23] Muhammad Rafi and Mohammad Shahid Shaikh, "An improved semantic similarity measure for document clustering based on topic maps" , *2010*

[24] Walaa K. Gad, Mohamed S. Kamel, "Incremental Clustering Algorithm Based on Phrase-Semantic Similarity Histogram", 2010

[25] Adrain Kuhn, Stephanne Ducasse, Tudor Girba, "Semantic Clustering: Identifying Topics in Source Code", *Elsevier, 2006*

[26] Anwiti Jain, Anand Rajavat, Rupali Bhartiya, "Design, analysis and implementation of modified k-mean algorithm for large data-sets to increase scalability and efficiency", *International Conference of Information Technology, IEEE,* Indonesia, 2012

[27] Nicola Cinefra, "Semantic Clustering for Complex Data Items", 2012

[28] Dharmendra K Roy and Lokesh K Sharma, "Genetic K-mean clustering algorithm for mixed numeric and categorical data sets", *International Journal of Artificial Intelligence and Applications (IJAIA),* Vol.1, No.2 April 2010

[29] Suresh Shirgave and Prakash Kulkarni, "Semantically enriched web usage mining for predicting user future movements", *IJCSI International Journal of Computer Science Issues, Vol. 4, No. 2, 2009, ISSN.1694-0784*

[30] Clustering Technique in Data Mining for Text Documents",. *International Journal of Computer Science and Information Technologies, , 3,* 2943-2947.

[31] Fabrizio Sebastiani "Machine Learning in Automated Text Categorization"*ACM Computing Surveys,* Vol. 34, No. 1, March 2002.

Websites

[32] https://en.wikipedia.org/wiki/Document clustering

[33] https://www.csee.umbc.edu/~nicholas/clustering/

[34] http://brandonrose.org/clustering

APPENDIX

The document clustering is used to access the information in sequence order in the way that the user wants to access. The data is available in unstructured form then we convert the unstructured data into structure data in sequence form as the dictionary order.

The aim of document clustering is finding similar topic based on user query with lesser degree search at minimum time, display the relevant material.

LDA (latent dirichlet allocation) algorithm used in document clustering to implement the work by assigning the weight to each word in document. The LDA basis on term weight has cluster the textual data.

DMNB (discriminative mixed naive bayes) algorithm is used to