UGC JOURNAL NUMBER 44557

International Journal of Allied Practice, Research and Review

Website: www.ijaprr.com (ISSN 2350-1294

A Detailed Approach for Data Mining and Clustering of Unstructured Data using R

Khushbu Singhal and Reshu Grover
Computer Science & Engineering
LIET, Alwar, Rajasthan, India

Abstract - Data mining is a phenomenon of extraction of knowledgeable information from large sets of data. Now a day's data will not found to be structured. However, there are different formats to store data either online or offline. So it added two other categories for types of data excluding structured which is semi structured and unstructured. Semi structured data includes XML etc. and unstructured data includes HTML and email, audio, video and web pages etc.

Now a day the amount of data, complexity of data and format of data is changing day by day and difficult to manage these data, as many companies or organizations are storing these data in data warehouse, so we need a technique to retrieve useful information data warehouse or from any location and for business analyst to identify trends and relationships of the unstructured data and making simple query with reporting tool. Data mining is reported as a tool for unstructured data.

In this paper, the framework for web mining is implemented using data mining tool Rstudio. Most important aspect of this paper is to extract data from website which is obviously unstructured data. It found difficult to extract content from unstructured data source.

Keywords - ***Unstructured Data, Semi-structured Data, Heterogeneous Data, DataMining,Clustering-means.***

I. Introduction

Due to the wide availability of huge amount of multimedia data in various modalities such as image and text documents, having a great amount of similarity among them is inevitable. In this paper, we present an efficient model which correlates the similarity among documents belonging to various modalities to achieve cross-media retrieval. Cross-media retrieval is a content based information retrieval system where heterogeneous data is mined to retrieve results of various modalities, i.e., input object and returned results may be of different modalities. For example, text objects can be retrieved as a result to image input. First, features are extracted from multimedia objects by which the objects are labeled. Using the labels, similar documents are grouped to generate Multimedia Documents. We construct a cross-media correlation

graph with documents as vertices, where positive weight is assigned to every single edge according to the amount of similarity between vertices. The cross-media retrieval system identifies the input document and as a result returns required number of documents with highest weights.

The advent of smartphones, social networks and cloud computing has added to the amount and sparse of data creation in the world, so much so that 90% of the world's total data has been created in the last 5 years and 70% of it by individuals. Studies predict that approximately 4 trillion gigabytes of data will exist on earth. As the world becomes increasingly digital, new techniques are requested, needed to search, analyze, and understand these huge amounts of unstructured data. This requires an automatic processing for unstructured data. This is BIGDATA R&D problematic, more specifically in the field of textual Data researches. Text mining is a set of techniques, which aim to process those huge amounts of data and gain value from it. Introduced by Ronen and Dagan as KDT [2], we find as main branches of text mining: text extraction, summarizing, categorization, etc. In opposite of Data mining, KDT aims to process unstructured texts, complex and over dimensioned data. Generally KDT is based on an automatic process to analyze the entire contents. The paper is organized on three sections: In the first section, we present a text clustering system for KDT. In the second section, a number of classification algorithms are described. The third section presents a comparative study of clustering algorithms in a KDT context. At the last section some conclusions are drawn

With the development of database, the data volume stored in database increases rapidly and in the large amounts of data much important information is hidden. If the information can be extracted from the database they will create a lot of profit for the organization. The question they are asking is how to extract this value. The answer is data mining. Most objects and data in the real world are interconnected, forming complex, heterogeneous but often semi-structured information networks. However, most people consider a database merely as a data repository that supports data storage and retrieval rather than one or a set of heterogeneous information networks that contain rich, inter-related, multi-typed data and information. Most network science researchers only study homogeneous networks, without distinguishing the different types of objects and links in the networks. We systematically introduce the technologies that can effectively and efficiently mine useful knowledge from such information networks. Today wide adoption of Internet has become an integral part of human life in terms of communication, gathering information, conducting business etc. Also, the web has grown exponentially in size and contains a large amount of publicly accessible web document distributed all over the world on thousands of servers. As document collection grows larger, they become more expensive to manage. The different types of data have to be managed and organized properly so that they can be accessed efficiently. However, the retrieval techniques based on the Information Retrieval (IR) and Data Mining research have found their way into major information services and the World Wide Web (www).

Data mining is one of the best ways to extract meaningful trends and patterns from large and unstructured data that discovers appropriate information from data warehouse of queries to display correct reports.

1.1Data Mining

The term "Data mining" was introduced in the 1990s, which is the evolution in the field of data retrieval with a long history, explained in [4]. Data mining roots are traced back along three family lines: classical statistics, artificial intelligence, machine and deep learning:

- Statistics are the foundation of most technologies on which data mining is built, e.g. regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals. All of these are used to study data and data relationships.

- Artificial intelligence is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS). Machine learning combines both statistics and AI.

1.2 Stages in Data Mining

The major components and the Stages of the data mining process are shows in Figure 1 are:

1) Data pre-processing (Heterogeneity resolution, Data cleansing, Data transformation, Data reduction, Discretization and generating concept hierarchies).

2) **Creating a data model:** applying Data Mining tools to extract knowledge from data.

3) **Testing the model:** the performance of the model (e.g. accuracy, completeness) is tested on independent data (not used to create the model).

4) **Interpretation and evaluation:** the user bias can direct DM tools to areas of interest (Attributes of interest in databases, Goal of discovery, Domain knowledge, Prior knowledge or belief about the domain.

5) **Knowledge Presentation:** In this we organize the data in such a way so that mined knowledge is easily available to users and users use it easily.

1.3 Types of Data

Data are categorized in mainly three formats that is structured, unstructured and semi structured. As we are working in distributed environment (internet) with heterogeneous data so we need to know the types of data exists in heterogeneous environment.

1) Structured Data: Forgeeks and developers structured data is very banal. It concerns all data which can be stored in database SQL in table with rows and columns. They have relational key with reverential key to map the pre-designed fields of the relations. Today, those data are the most processed in development and the simplest way to manage information's. But structured data represent only 5 to 10% of all informatics data.

2) Semi structured data: Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. With some process you can store them in relation database (it could be very hard for some kind of semi structured data), but the semi structure exist to ease space, clarity or compute. Examples of semi-structured: CSV but XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured. But as Structured data, semi structured data represents a few parts of data (5 to 10%) so the last data type is the strong one: unstructured data.

3) Unstructured data: The maximum data, now days are in the form of unstructured and multiple formats, which is around 80% of all data. It often includes text and multimedia content. Examples include e-mail, word documents, videos, photos, and audio files, ppt, webpages, pdf , streamed data , face book data , twitter data, linked in data, instagram, and many other kinds of business documents. These data and files may have an internal structure, and still are 'unstructured' they don't fit neatly in a database. Unstructured data is everywhere. Here are some more examples of machine-generated unstructured data *(Satellite images, scientific data, Photographs and video, Radar or sonar data,Text internal to your company,Social media data,Mobile data, and website content.*

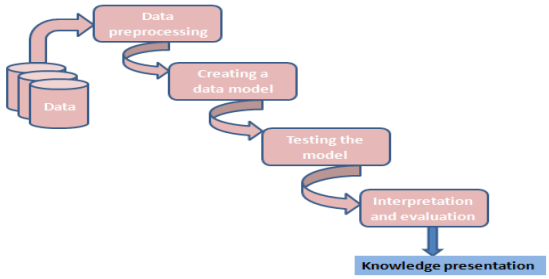

Figure 1. Data processing in Data Mining

1.4 Data Mining Techniques

There are six important data mining tasks are majorly used on heterogeneous data.

1) Association: Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as relation technique. The association technique is used in market basket analysis *to* identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for customer and increase sales.

2) Classification: Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method is based on mathematical techniques such as 1) decision trees, 2) linear programming, 3) neural network, and 4) statistics. In classification groups same type of data. For example, 1) "given all records of employees who left the company; 2) predict who will probably leave the company in a future period." In this case, we make the records of employees into two groups , "leave" and "stay". Then we can ask our data mining software to classify the employees into separate groups.

3) Clustering: This technique is used to dividing data items into groups based on some similarity called clusters, like we have several type of data in one folder, we arranged them in several folder based on similarity like text data saved in text folder, audio in audio folder etc so that user can easily access the data according to their need.

4) Prediction: The prediction, as its name implied, is one of a data mining techniques that discover the relationship between independent variables and relationship between dependent and independent

variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable.

5) Sequential Patterns: For finding the similar type of patters, regular trends and events in transaction, sequential patterns analysis, data mining technique is used over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year.

6) **Decision trees:** A decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. For example, we use the following decision tree to determine whether or not to play tennis: Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the week. And if it is sunny then we should play tennis in case the humidity is normal.

II. Literature Review

Data mining is the process of extracting previously unknown, valid and actionable information from large databases and then using the information to make crucial business decisions.

2.1 Data Mining

Ming-Syan Chen et.al [1], Describes about the basic definition of Data mining. It is defined as a process of extracting or mining useful knowledge from huge amounts of data, or simply knowledge discovery in databases. It has become useful over the past decade in business to gain more information, to have a better understanding of running a business, and to find new ways and ideas to increase the business. Several emerging applications in information providing services, such as data warehousing and online services over the Internet, also call for various data mining techniques to better understand user behavior, to improve the service provided and to increase business opportunities.

S. R. Dhamankaret. al [2], Describes about ontology and it states that more complex the application is, the larger the gap comes into existence between application and users . The data mining applications to illustrate the concepts and selection a better model to match business requirements to data mining categories to connect complex data mining concepts with business problems and assists users to choose the best data mining solution.
Three steps involved are

1). Exploration- Finding different variables on the basis of data base nature, cleaning of data and transform in another form for analysis.

2) Pattern identification- After Exploration, identify, and choose best pattern for the best prediction of data.

3) Deployment Exploration: When patterns are matched, and then deployed these matched patterns for desired outcome.

2.2 Unstructured and Semi-structured Data mining Algorithm:

In the previous section, we have regarded Web pages as unstructured texts. However, it is often more natural to consider Web pages as markup texts or semi-structured data. Using such structure information, we

may obtain more interesting pattern that characterize a given dataset well. Thus, it will be an interesting problem to develop an efficient algorithm for finding optimal pattern in such semi-structured data. Hence, we develop efficient pattern discovery algorithms from large collections of semi-structured data on the Web based on the framework of the optimized pattern discovery in [1, 5, 6], We model such semi-structured data and substructure patterns by labeled ordered trees and study the problem of discovering all frequent patterns with frequency above a given threshold on a given collection of semi-structured data. We present an efficient mining algorithm for discovering all frequent patterns from unstructured huge amount data of labeled ordered trees, shown in Figure 2.

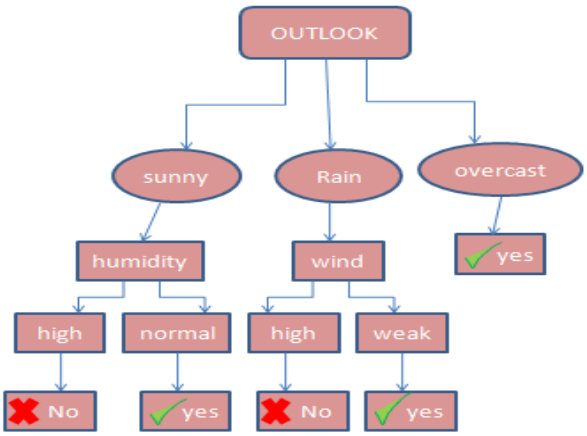

Figure2. Example of decision tree

2.3 Text Mining

Calvilloet.al [10], Describes about the text mining and about usage of data mining technique clustering. Automated text classification is the task of assigning a category to a document. The time spent by users are almost two or more hours looking for papers that generates the possibility to make a search engine to optimize and precision in the results. The classification of text mining is used to search in the documents using natural language to find the best words theircontents to get a database knowledge, that's the first step to get the desired knowledge about documents and use the same engine to make searches classifying the information introduced by the final user and searching in the correct cluster.

2.4 Clustering algorithms k-Means Clustering (O-Cluster)

K-Means algorithms support identifying naturally occurring groupings within the data population. the algorithm divides the data set into k number of clusters according to the location of all members of a particular cluster in the data. Clustering makes use of the Euclidean distance formula to determine the location of data instances and their position in clusters and so requires numerical values that have been properly scaled..

III. Web Mining

Data mining means extraction of data in terms of patterns or rules from huge amount of data [1,4]. The research in the field of web is classified on two aspects: the retrieval and the mining. The retrieval focuses on retrieving relevant Information from large repository whereas mining research focuses on extracting new information already existing data. Web mining is integration of information that is gathered by traditional data

mining techniques with information gathered over World Wide Web. Web mining is decomposed into following subtasks:

I). Resource Discovery: It helps in retrieving services and unfamiliar documents on web.

II). Information selection and preprocessing: This step is used for automatically selection of the specific and required information, and then preprocesses this information from the web sources.

III). Generalization: It uncovers general pattern at individual web sites as well as across multiple sites.

IV). Analysis: It validates and interprets the mined pattern.

V) . Visualization: It presents the result in visual and easy to understand way.

Web mining is divided into three main categories depending on the type of data as web content mining, web structure mining and web usage mining [17].we are mainly focusing on web content mining then applying sentimental analysis through frequency count and WordCloud.

3.1 Web Mining categories

This section divides web mining into three categories depending on the type of data i.e. Web Content Mining, Web Structure Mining, and Web Usage Mining. Let us have a bird's eye-view on each of the above three mining techniques. Later on we will focus on the web content mining, its significance and features in this paper [6, 17, 18] is shown in Figure 3.

3.2 Processes Involved in Web Mining (Shown in Figure 4).

Web mining is defined as the "process of studying and discovering web user behavior from web log data. "Generally the web data collection is done over a long period of time (one day, one month, one year, etc.). Later on, four steps, 1) Web Data Collection, 2) Preprocessing of unstructured web Data, 3) Discovery of Pattern from Web Data and 4) Analysis of Pattern of Web Data that are indexed. Pre-processing of web data is the process of transformation of the raw data into a usable data model. Pattern discovery step uses several data mining algorithms is used to extract the user patterns. Finally, pattern analysis from web data uncovers useful and interesting user patterns and trends. These steps are normally executed after the web log data is collected.

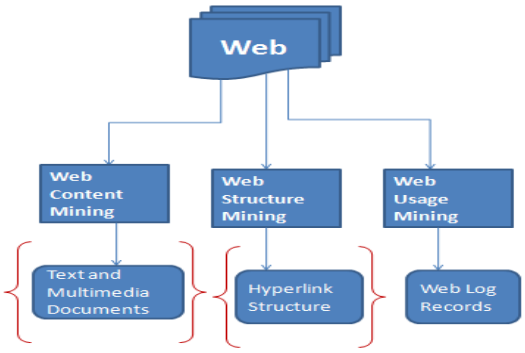

Figure 3. Types of web mining

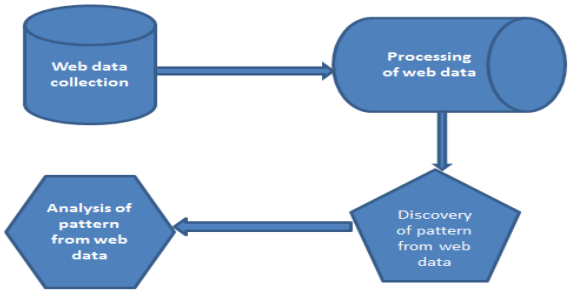

Figure 4. Steps followed in Web mining

IV. Proposed System Design

Here we have proposed a new system to extract the unstructured data, the architecture of proposed approach is shown in Figure 5.

4.1Data mining on Heterogeneous data

Heterogeneous data is the combination of different semantics of data. This means that data stored in different formats such as HTML, XML, audio, video, PDF etc. different types of data is categorized by different authors which is semi-structured and unstructured data, In this paper we are working with data mining of semi structured data (XML) and unstructured data (website content in HTML format) with data analysis in the form of WordCloud and frequency count.

4.2 Methodology: The methodology of our approach is shown in Figure 5.

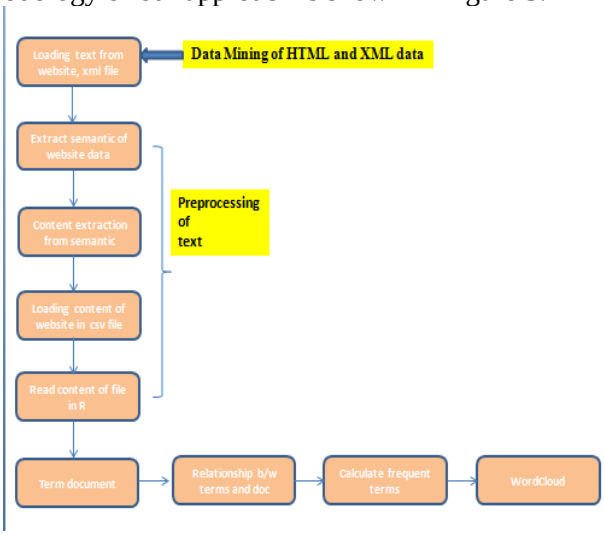

Figure 5.Architecture of paper

4.3 Loading text from website, xml file

It is a first activity of the paper framework, considering website www.nptel.com as the reference for this paper. Extracting the semantic contents from this website(with HTML code) and xml file will be the first step.

```
> library('XML')
> library('methods')
> result<- xmlParse(file = "C://Users//ADMIN//Documents//bhagya m.tech//hmt//rcxml
.xml")
>
```

Figure 6. Loading of semi structured xml file

```
> library(XML)
> library(RCurl)
> library(xlsx)
> getwd()
[1] "D:/Rlessions"
> setwd("D://Rlessions")
> getwd()
[1] "D:/Rlessions"
> ExtractHTML = htmlTreeParse('http://nptel.ac.in/courses/117105135/',useInternalN
odes = TRUE)
>
```

Figure 7.Loading of unstructured data from website (HTML)

4.2 Pre-processing of text

Pre-processing of text data activities are combination of1)Extract semantic of website data and xml data 2) Text extraction from structure 3) Loading content of website in csv file and 4) reading file content in R. this three steps are main steps from which further results will be occurred**.** Extraction of semantic of website means the page source code will extracted as it is as displayed on web in the Rstudio. Now the further step is to extract the relevant contents from that entire HTML programming structure, it is also called cleaning of data by removing html tags and all. Now the third step is to load the data extracted in Rstudio should be stored in some document file so commonly used file formats extensions in Rstudio is CSV file and XLSX file. Fourth step is to read the file content in Rstudio, by this step the data will appear in excel format with contents. Pre-processing step also include removing of numbers, special characters, converting the data into lowercase, remove punctuations etc.

4.6 Term document matrix

A term-document matrix represents the relationship between terms and documents, where each row stands for a term and each column for a document, and an entry is the number of occurrences of the term in

the document. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms and organize the data according to their frequency and also removing the common terms, this will make 10% matrix empty. Term document matrix is shown in Figure 8.

```
> dtm <- DocumentTermMatrix(corpus)
> dtm
<<DocumentTermMatrix (documents: 1, terms: 2)>>
Non-/sparse entries: 2/0
Sparsity           : 0%
Maximal term length: 10
Weighting          : term frequency (tf)
> dtm2 <- as.matrix(dtm)
```

Figure 8. XML file

```
> dtm <- DocumentTermMatrix(corpus)
> dtm
<<DocumentTermMatrix (documents: 1, terms: 119)>>
Non-/sparse entries: 119/0
Sparsity           : 0%
Maximal term length: 21
Weighting          : term frequency (tf)
>
```

Figure 9. HTML content

4.7 Calculate Frequent Word

In this we are going to find out the word that are most frequently occurred in the documents and we also adjust the frequency of word like 3, 4.Suppose we will adjust the minimum frequency 3 26 and then it will plot the word frequency. Note that the frequent terms are ordered alphabetically, instead of by frequency or popularity. To show the top frequent words visually, we make a barplot of them using ggplot2 package of Rstudio.

4.8 Relationship between Terms

- **Term Correlations:** If you have a term in mind that you have found to be particularly meaningful to your analysis, then you may find it helpful to identify the words that most highly correlate with that term.

4.9WordCloud

We can show the importance of words pictorially with a WordCloud. In the code below, we first convert the term-document matrix to a normal matrix, and then calculate word frequencies. After that we use WordCloud to make a pictorial. Humans are generally strong at visual analytics. That is part of the reason that these have become so popular. What follows are a variety of alternatives for constructing word clouds with your text. But first you will need to load the package that makes word clouds in R. We also form a word cloud of words which are frequently occurs in the documents. Similarly we can form a colored word cloud of words.

V. Implementation And Results

5.1 Extraction of xml data

In this implementation, how to do data mining of semi structured data will be shown. Extracting the data from file name **rcxml.xml**. The data in this file is in the form of XML coding. Figure 10 shows the structure of the data in file which in the form of XML coding and Figure 11 shows the result in which the content of the file is extracted using r programming in Rstudio.

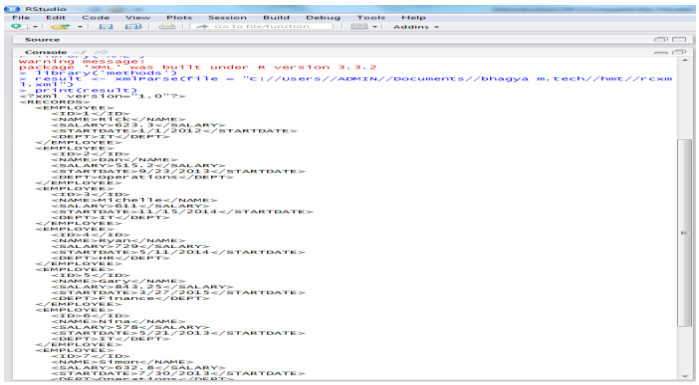

Figure 10. Semantic of Xml data in file

```
attr(,"class")
[1] "XMLInternalNodeList" "XMLNodeList"
> xmlToDataFrame <- xmlToDataFrame("C://Users//ADMIN//Documents//bhagya m.tech//hm
t//rcxml.xml")
> print(xmlToDataFrame)
  ID     NAME SALARY  STARTDATE       DEPT
1  1     Rick  623.3   1/1/2012         IT
2  2      Dan  515.2  9/23/2013 Operations
3  3 Michelle    611 11/15/2014         IT
4  4     Ryan    729  5/11/2014         HR
5  5     Gary 843.25  3/27/2015    Finance
6  6     Nina    578  5/21/2013         IT
7  7    Simon  632.8  7/30/2013 Operations
8  8     Guru  722.5  6/17/2014    Finance
>
```

Figure 11. Content of Xml data in file

5.2 Extraction of webpage data/ html data

In this paper extraction of data from the website 'http://nptel.ac.in/courses/117105135/' which is demonstrated Step by step in the following sections. Extraction procedure will be displayed in the following sections here:

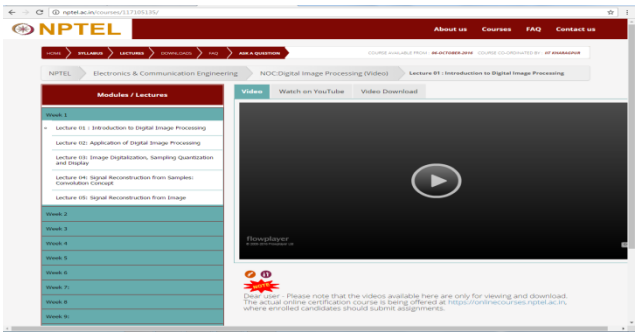

Figure 12.Website page view

A). Extracted website data:

After applying programming for the extraction of website data then the content will be automatically stored in given path to local disk by user. Here the representing the appearance of the automatic storage of file in disk. File appeared as the name given in the code called paper.csv:

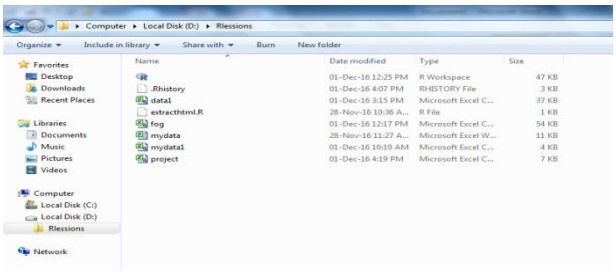

Figure 13.Windows Representing File

B). View the contents

Copying entire view of csv file is copied into R as it appeared in the figure. Using command view (paper). Paper is a file name stored in the local disk of computer.

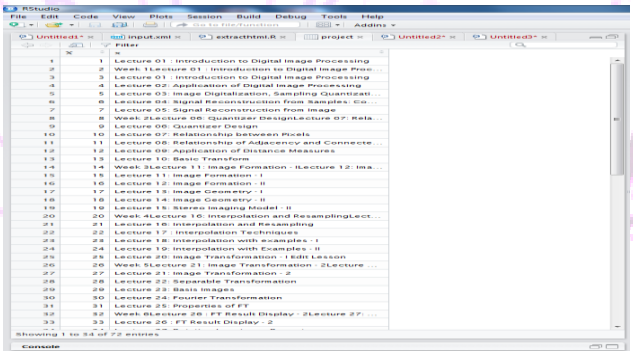

Figure 14 windows representing file

C). Retrieving Text from the Paper.csv File

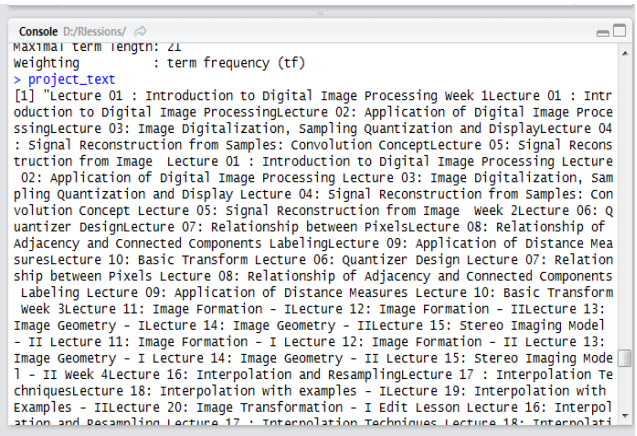

Figure 15.Reading content of webpage

D). Loading and cleaning the corpus

Now we are in a position to load the transcripts directly from our hard drive and perform corpus cleaning using the tm package. Now we use regular expressions to remove at-tags and urls from the remaining documents.

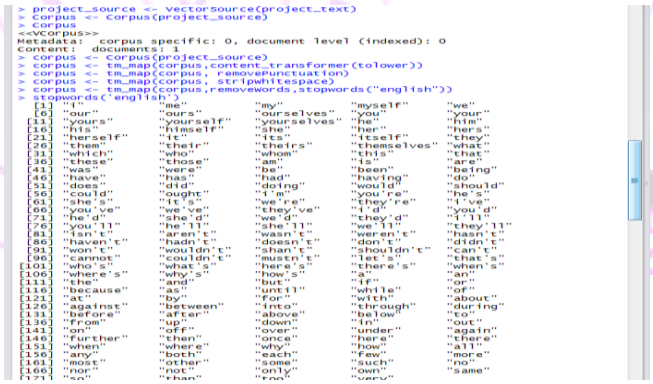

Figure 16. Loading and cleaning process of data

E). building a term-document matrix

A term-document matrix represents the relationship between terms and documents, where each row stands for a term and each column for a document, and an entry is the number of occurrences of the term in the document.

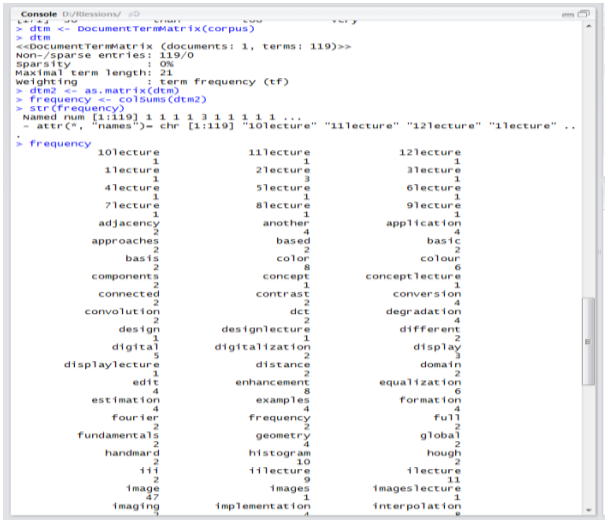

Figure 17. View of term-document matrix

F). Barplot of frequent terms

Plotting of the words according to the appearance in the paper.csv file.

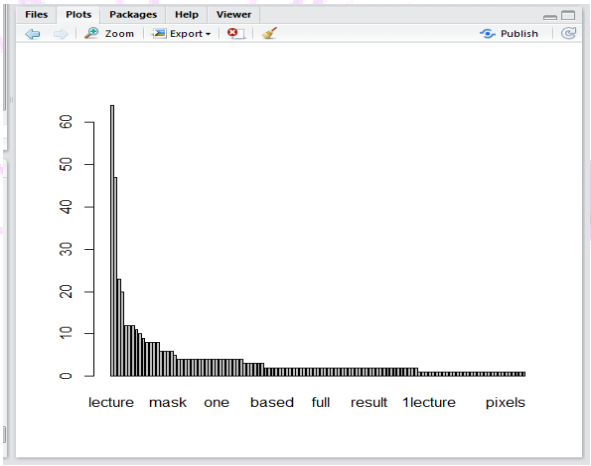

Figure 18.Barplot of words of file

G). Calculating head terms

Calculation of words which are appeared most in the file are counted and shown through program.

```
> barplot(frequency)
> frequency <- sort(frequency,decreasing = TRUE)
> head(frequency)
    lecture       image techniques  processing       model restoration
         64          47          23          20          12          12
```

Figure 19. View of head word

H). WordCloud

Final results of this paper are shown in the following Figures 20 and Figure 21. According to the word frequency the WordCloud will be plotted in the Rstudio using r programming. it is very useful for the data analysis purpose of large amount of data.

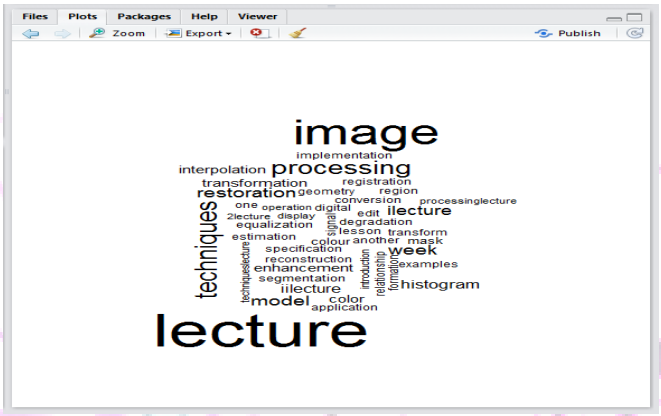

Figure 20.WordCloud

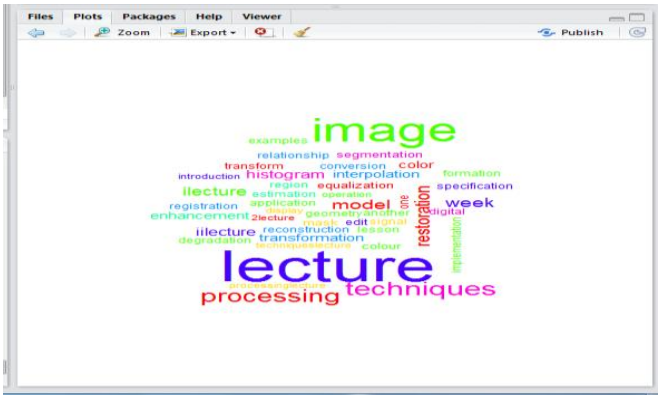

Figure 21. Colored word cloud

G). Hierarchical clustering

First estimating the distance between words and then cluster them according to similarity.

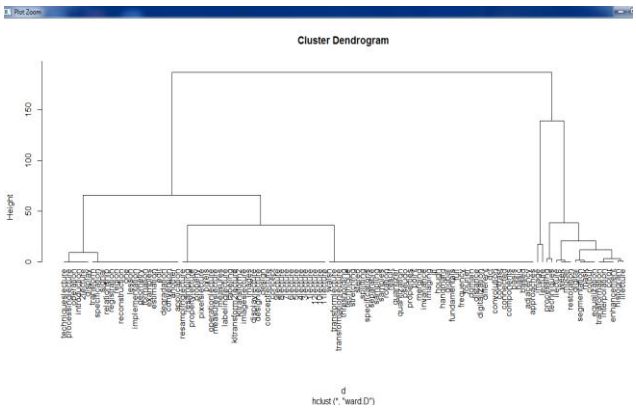

Figure 22. Dendogram Of The Text

Helping to Read a Dendrogram:To get a better idea of where the groups are in the dendrogram, you can also ask R to help identify the clusters. Here, I have arbitrarily chosen to look at five clusters, as indicated by the red boxes. If you would like to highlight a different number of groups, then feel free to change the code accordingly. Clusters are divided in five different clusters as shown in the Figure 23.

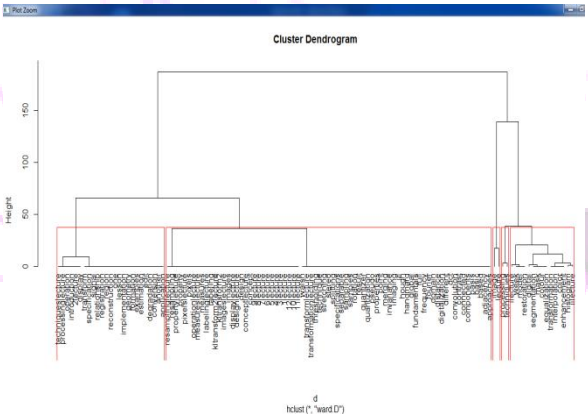

Figure 23 Specified clusters in red

F). K-means clustering

The k-means clustering method will attempt to cluster words into a specified number of groups (shown in Figure 24), such that the sum of squared distances between individual words and one of the group centers. You can change the number of groups you seek by changing the number specified within the kmeans() command.

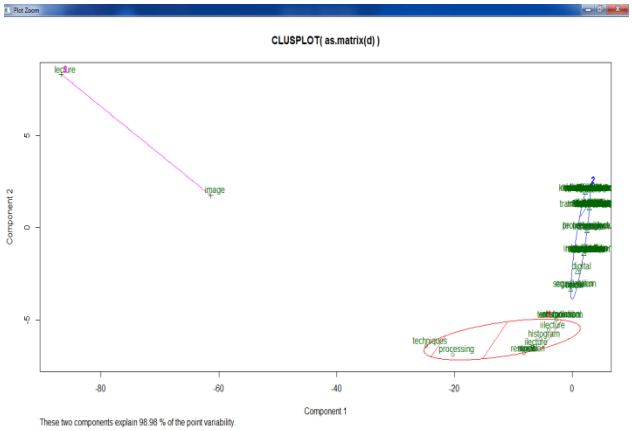

Figure 24. Graphplot of k-means clustering

VI. Conclusion

In this paper, the framework for web mining is implemented using data mining tool Rstudio. Most important aspect of this paper is to extract data from website which is obviously unstructured data. It found difficult to extract content from unstructured data source. Other aspects of this framework is to identify the documents and the data they contained and evaluate the feasibility to apply text mining which may achieve good performance with high efficiency when dealing with thousands of documents, by separating the data contained by documents into bag of words. From our experiment we analyze, pre-processing does play an important role. Frequent words and associations are found from the matrix. A word cloud is used to present frequently occurring words in documents. Two main types of clustering techniques used(Hierarchical and k-means)applied on data set from that we can analyze the data.

The work presented in paper can be enhanced further by applying it to heterogeneous datasets, like Image, Audio, Video, Social Networking etc. we can also apply different tasks data mining such as classification, association, regression analysis and so on, also compare the work of these different tasks on the same data. Due to computer speed and memory limitations, data set was relatively small in this work. One of the future directions for this work is to perform a more detailed statistical analysis of heterogeneous data.

VII. References

[1] Ming-Syan Chen, Jiawei Han, and Philip S.Yu, "Data Mining – An Overview from Database Perspective",Knowledge and Data Engineering, IEEE Transactions on ,Volume 8 , No.6 , pp 866-883,Dec 1996.

[2] S. R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos, "iMAP: Discovering Complex Semantic Matches between Database Schemas", International Conference on Management of Data,ACM SIGMOD, pp 383-394,2004.

[3] E. Rahm, P.A. Bernstein. "A Survey of Approaches to Automatic Schema Matching". VLDB Journal, Volume 10, No. 4, pp 334-350,2001.

[4] Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age" ,IEEE Intelligent Systems, Volume 6, pp 32-34,2000.

[5] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The survey of Data Mining Applications and Future Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[6] Nicholas J Belkin and W Bruce Croft,"Retrieval techniques", Annual Review of Information Science and Technology,Volume 22,pp 109-45,Information Today,1987.

[7] Romero, Cristobal, Sebastián Ventura, and Paul De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors." Volume 14,Issue 5, pp 425-464, 2004.

[8] Ansari S., Kohavi R., Meason L., and Zheng Z., "Integrating E-Commerce and Data Mining: Architecture and Challenges",IEEE International Conference on Data Mining,pp 27-34,2001.

[9] Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features READIT, pp.54-59,2007

[10] Calvillo, E. Alan, Alexandra Padilla, Jaime Munoz, Julio Ponce, and Jesualdo T. Fernandez, "Searching research papers using clustering and text mining." International conference on Electronics, Communications and Computing ,pp. 78-81, IEEE, 2013.

[11] Franklin, Michael, Alon Halevy, and David Maier. "From databases to dataspaces: a new abstraction for information management" ACM Sigmod Record,Volume 34,Issue 4,pp 27-33,2005

[12] Niranjan Lal, Samimul Qamar, "Comparison of Ranking Algorithm with Dataspace", International Conference On Advances in Computer Engineering and Application(ICACEA),pp 565-572, March 2015.

[13] Mark Hall, Eibe Frank, G. Holmes, B. Pfahringer, and P. Reutemann, " The WEKA data mining software: An update",ACM SIGKDD explorations newsletter, volume 11, Issue 1,pp 10-18, june 2009.

[14] Sunita B Aher, Mr. LOBO L.M.R.J., "Data Mining in Educational System using WEKA", International Conference on Emerging Technology Trends (ICETT),Volume 3,pp 20-25,2011.

[15] V. S. Jagadheeswaran, V. N. Saranya, "A Survey on Data Mining Application & Tools", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue5, pp. 477-480, May 2015.

[16] Vikas Gupta, Prof. Devanand, "A survey on Data Mining: Tools, Techniques,Applications, Trends and Issues",International Journal of Scientific & Engineering Research,Volume 4, Issue3, pp 20-33, March 2013.

[17] Bharati m. ramageri," Data mining techniques and applications", Indian journal of computer science and engineering, vol. 1 no. 4 301-305

[18] PrakashR.Andhale1 , S.M.Rokade2, "A Decisive Mining for Heterogeneous Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12,pp. 43-437, December 2015.