International Journal of Engineering Research & Management Technology ISSN: 2348-4039

November-2016 Volume 3, Issue-6

Email: editor@ijermt.org www.ijermt.org

MINING OF UNSTRUCTURED DATA WITH CLUSTERING APPROACH

Bhagyashree Pathak
Computer Science & Engineering
Mody University of Science & Technology,
Lakshmangarh, Sikar,
Rajasthan

Niranjan Lal
Computer Science & Engineering
Mody University of Science & Technology,
Lakshmangarh, Sikar,
Rajasthan

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. The numbers of semi-structured and unstructured documents are produced and that are steadily increasing in our daily life. Thus, it will be essential for discovering new knowledge from them. In this paper we have consider the HTML data, implementation is based on extraction of data from text file and web pages by using the popular data mining techniques and final result will be after sentimental analysis of text, and unstructured data extraction of web page with HTML code, there will be an extraction of structure/semantic of code alone and also both structure and content. The sentimental analysis includes the frequency count of number of words extracted from web page, display of main words, and display of counts of these main words and most important it shows the frequency of each word by making WordCloud as a plot diagram. As a result the clustering of the text present in the will be done using two main clustering methods hierarchical and k-means clustering. Execution of this paper is using R is a programming language on Rstudio environment.

KEY WORDS: Data mining, text mining, WordCloud, hierarchical clustering, k-means clustering

I. INTRODUCTION

Internet is an open worldwide added network. It allows global communication linking all the connected computing devices. It is a platform for web services and World Wide Web [2,11]. Web is an admired and interactive medium with powerful amount of data liberally accessible for users to access. It is a collection of documents, text files, audios, videos and other multimedia data [8,12,13]. The different types of data have to be organized in such a way that different users can efficiently Access it. The process of extracting valid, previously unknown, comprehensible, and acted information from large databases and using it to make critical business conclusions is known as Data Mining.

Data mining is concerned with the analysis of data and the use of software systems for finding unknown and unpredicted patterns and relationships in sets of data. The focus of data mining is to find the information that is hidden and unexpected. Data mining can supply gigantic benefits for companies who have made important investment in data warehousing. Although data mining is still a relatively new technology, it is already used in a number of industries. There are many applications of data mining in retail/marketing, banking, insurance, and medicine fields. The storing information in a data warehouse does not propose the benefits a business is seeking. To understand the importance of a data warehouse, it is necessary to extract the knowledge hidden within the warehouse. However, as the amount and intricacy of the data in a data warehouse raises, it happen to ever more tricky, if not impossible, for business analysts to identify trends and relationships in the data using simple query and reporting tools. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses obtainable by data mining progress ahead of the analyses of past events granted by displayed tools typical of decision support systems. Data mining tools can answer business questions that traditionally were also time consuming to determine. They clean databases for unknown patterns, discovering analytical information that experts may miss because it lies outside their expectations.

Data mining is individually the finest way to pull out significant trends and patterns from vast amounts of data. Data mining discovers .information within data warehouse that queries and reports cannot effectively reveal.

The rest of the sections are described as follows. In Section 2 we have discussed the work related to our approach, Section 3 describes the framework for text mining at real example, section 4 implementation and results. Finally, Section 5 concludes this paper.

II. RELATED WORK

Calvillo et.al [5], 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 used up by users are approximately two or more hours looking for papers that produces the possibility to make a search engine to optimize and precision in the results. The initial work of a classification using text mining techniques to search into the documents with natural language contained and get the best words of their content 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. Hiroki Arimura et.al [7], proposed the algorithm and optimizes the performance for the text mining of semi structured data and unstructured data. A basic idea behind their method is to employ a regular set of texts as the organized set worn for revoke the event of frequent and non-informative keywords. The control set will be a set of documents randomly drawn from the whole text collection or the internet. So that we can easily observe that most stop words appear evenly in the target and the control set, while informative keywords appear more frequently in the target set than the control set. as a result, the optimized pattern detection algorithm will find those keywords or phrases that appear more repeatedly in the target set than the control set by reducing a known numerical measure such as the information entropy or the prediction error. Also introduce a class of simple combinatorial patterns over texts such as proximity phrase association patterns and ordered and unordered tree patterns modeling unstructured texts and semi-structured data on the Web. Then, we consider the problem of finding the patterns that modify a given statistical gauge contained by the entire class of patterns in a large collection of unstructured texts. W. Himmel et.al [6], Text mining algorithms have been used in many applications such as summarizing and analyzing web content and managing scientific publications. Text mining generally starts with a text pre-processing step, where unstructured text is transformed into a structured form, which is then used for clustering or classification. B. Liu [8], finally sentiment analysis is the field of study that analyzes people's opinions, sentiments, attitudes, and emotions in text. For example, positive and negative opinions can be mined in customer reviews (text) regarding a specific product. Sentiment analysis is often used to monitor brand reputation and to help businesses understand the perception that customers have about their products or services; this can help improve their marketing and customer relationship management.

III. FRAMEWORK FOR TEXT MINING

The procedure involved in the text mining is described in the Fig. 1 for text mining framework. In this Framework we can clearly see steps to execute text mining and then sentimental analysis of text through WordCloud and clustering.

A. Retrieve text

Texts going to be extracted from website name www.nptel.com with the html code of page using htmlTreeParse() and sapply() in packages XML, xlsx, Rcurl in Rstudio 3.3.1. from this activity we can easily extract the text from website then the required text will be taken and saved automatically in computer's local disk as file name project.csv.

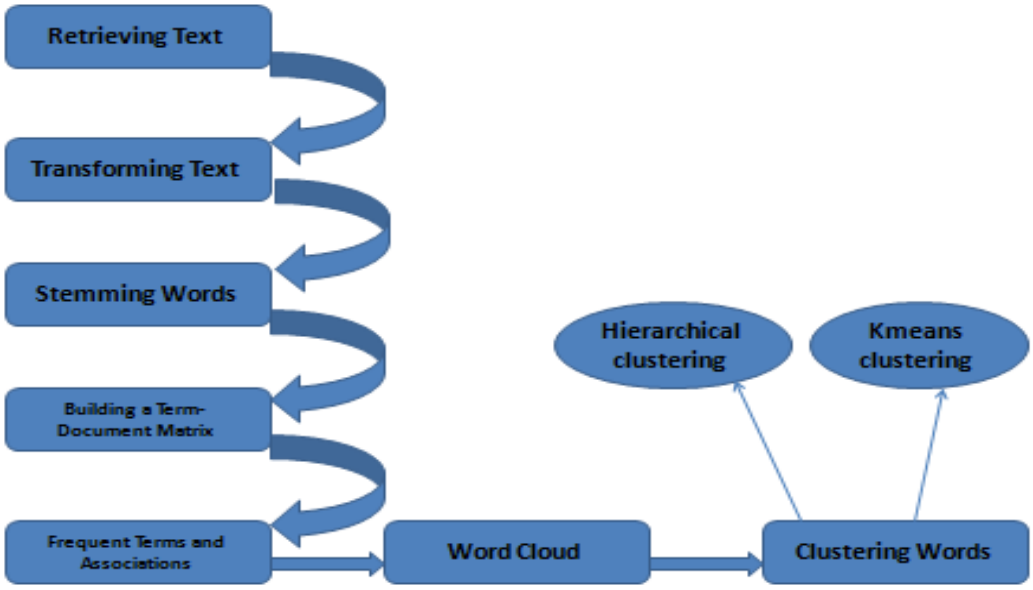

Fig. 1: Steps involve in Text mining

B. Transforming text

The texts are first converted to a data frame and then to a corpus, which is a collection of text document. After that, the corpus can be processed with functions provided in package tm, After that, the corpus needs a couple of transformations, including changing letters to lower case, and removing punctuations, numbers and stop words. The general English stop-word list is modified now via adding up "available" and "via" and removing "r" and "big" (for big data). Hyperlinks are also removed.

C. Stemming words

In many applications, words need to be stemmed to retrieve their radicals, so that various forms derived from a stem would be taken as the same when counting word frequency. For instance, words update, updated and updating should all be stemmed to update. Sometimes stemming is counterproductive, so we have not selected. That's why; we didn't used to do this step in our text mining and sentimental analysis.

D. Building text 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 file, and an access is the number of appearance of the term in the file. On the other hand, one can also build a document-term matrix by swapping row and column. In this section, we build a term-document matrix from the above progression corpus with function TermDocumentMatrix().

E. Frequent Terms and Associations

We have a look at the popular words and the association between words. findFreqTerms() finds frequent terms with frequency no less than ten. Note that they are ordered alphabetically, instead of by frequency or popularity. To show the top frequent words visually, we next make a barplot for them. From the termdocumentmatrix, we can derive the frequency of terms with rowSums(). Then we select terms that appears in ten or more documents and shown them with a barplot using package ggplot2. Alternatively, the above plot can also be drawn with barplot() as shown in Fig. 7.

F. Word Cloud

After building a term-document matrix, we can illustrate the significance of terms with a word cloud (moreover identified as a tag cloud), which can be easily produced with package WordCloud. In the code below, we first convert the term-document matrix to a standard matrix, and after that estimate word frequencies. Afterwards, we set gray levels based on word frequency and use WordCloud() to make a plot for it. A colorful cloud can be generated by setting colors with rainbow(7). It shown in the Fig. 9.

G. Clustering Words

We then try to find clusters of words with hierarchical clustering and kmeans clustering. Sparse terms are removed, so that the plot of clustering will not be crowded with words. Then the distances between terms are calculated with dist() after scaling. After that, the terms are clustered with hclust() and the dendrogram is cut into 5 clusters. The agglomeration method is set to ward, which denotes the increase in variance when two clusters are merged. For kmeans clustering we need to use kmeans() and clustplot() functions. Two main packages are used for both clustering i.e library(cluster) and library(fpc).it shows in the Fig. 10,11,12.

IV. IMPLEMENTATION AND RESULTS

A. Extraction of webpage data/ html data

In this paper the 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 here in the following sections.

B. Extracted website data:

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

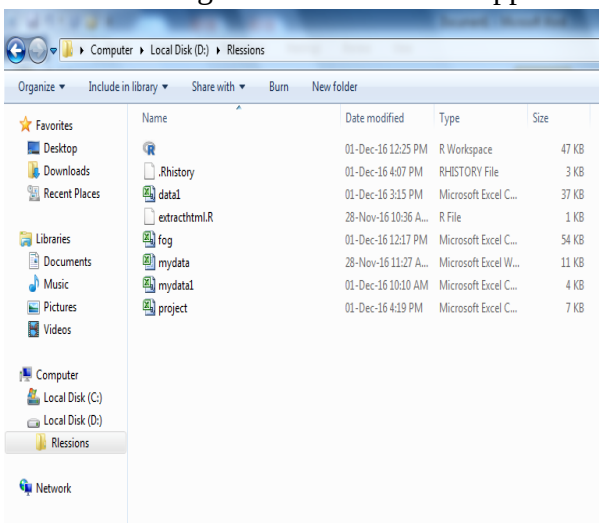

Fig 2: Windows Representing File

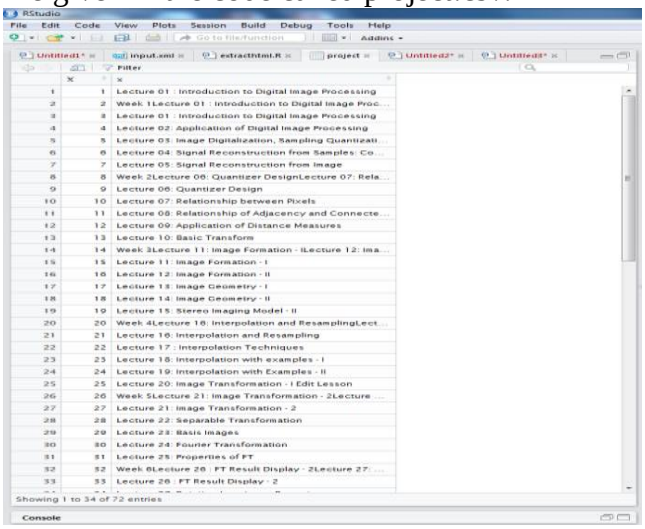

Fig 3: Representation of windows file in Rstudio

C. View the contents

Copying entire view of csv file is copied into R as it appeared in the Fig 3. Using command view(project). Project is a file name stored in the local disk of computer.

B. Retrieving Text from the Project.csv File

See this process of retrieving text from project.csv file in Fig. 4.

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 are as shown in Fig. 5.

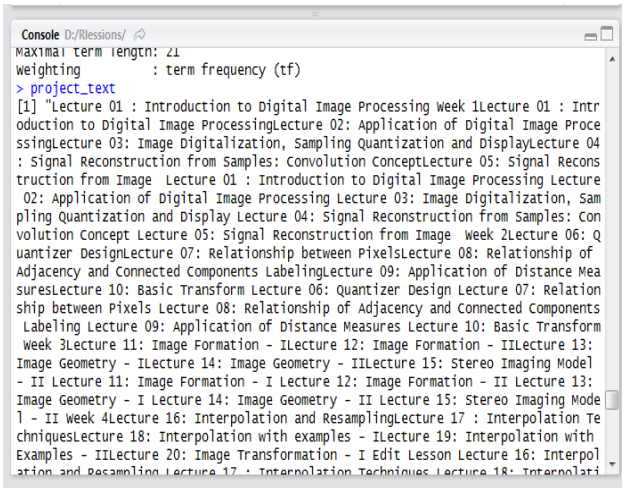

Fig 4: Reading content of webpage in Rstudio

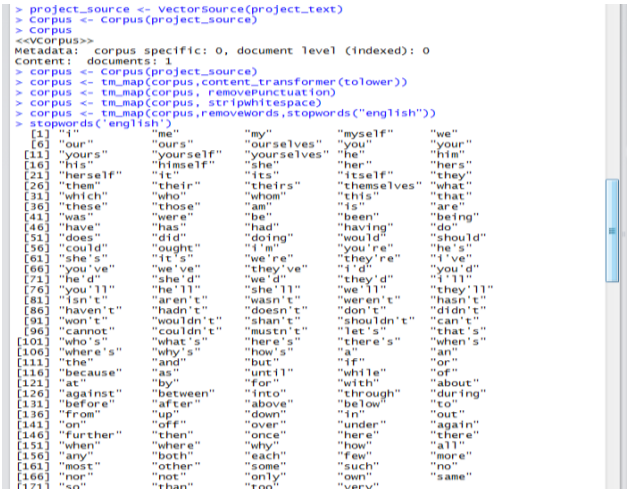

Fig 5: Loading and cleaning process of text

E. Building a term-document matrix

The term document matrix process is shown in Fig. 6.

Fig 6: Term document matrix

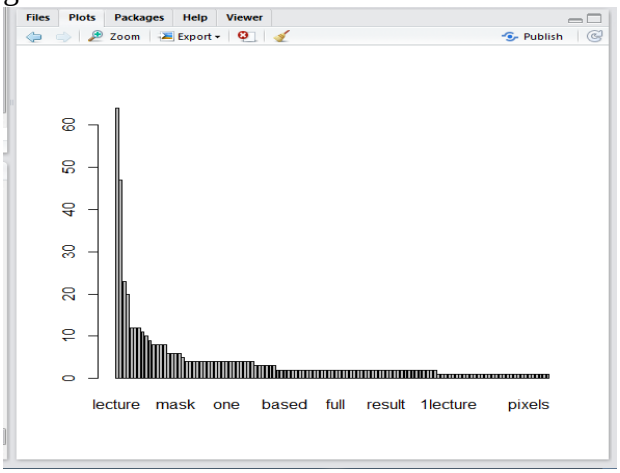

Fig 7: Barplot of words of file

F. Barplot of frequent terms

Plotting of the words according to the appearance in the file project.csv, this is shown in Fig. 7.

G. Calculating head terms

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

H. WordCloud

Worldclout of resultant head terms of text data is shown in Fig. 9. According to the word frequency the WordCloud is plotted in Rstudio using r programming. It is very useful for the data analysis purpose of large amount of data.

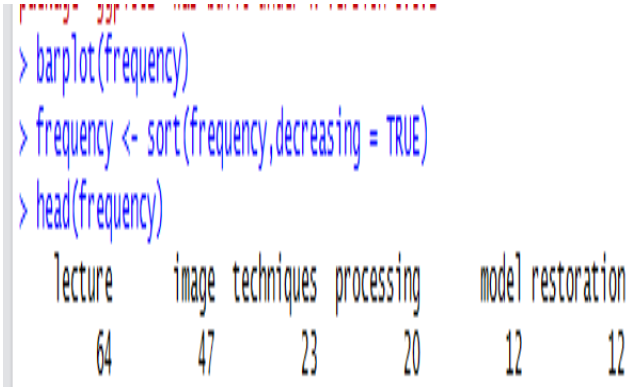

Fig 8: Head terms in file

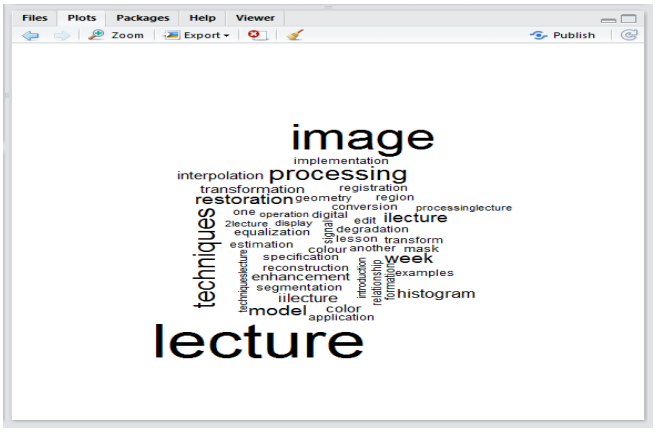

Fig 9: WordCloud

I. Hierarchical clustering

First estimating the distance between words and then cluster them according to similarity (as shown in Fig. 10). 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, we have arbitrarily chosen to look at five clusters, as indicated by the red boxes. It would be easy 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 Fig. 11.

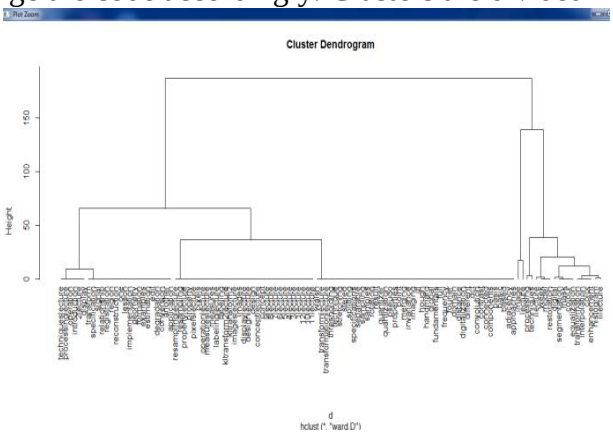

Fig 10: Dendogram of terms

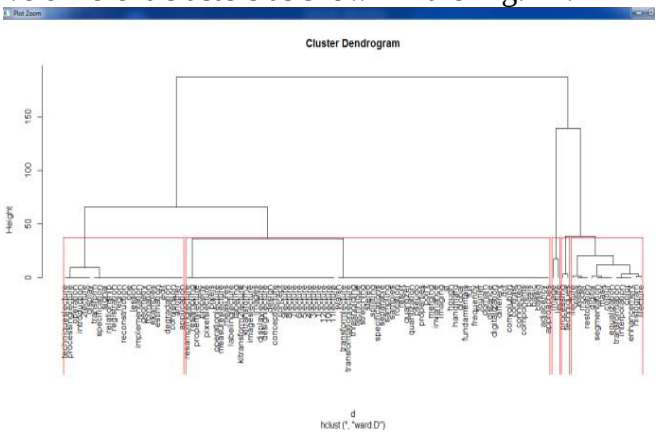

Fig 11: Clusters in red boundary

11. K-means clustering

The k-means clustering method will attempt to cluster words into a specified number of groups is shown in Fig. 12, 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 k-means.

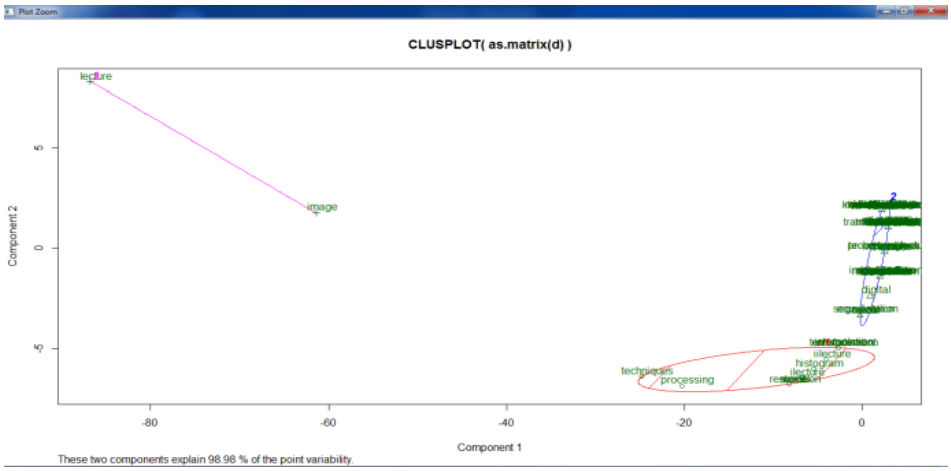

Fig. 12: Plot of Kmeans clustering

V. 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.

VI. References

1. Singh, Brijendra, and Hemant Kumar Singh, “Webdata mining research: A survey." Computational Intelligence and Computing Research (ICCIC), 2010 IEEE international Conference on.IEEE, 2010.
2. Ming-Syan Chen, Jiawei Han, and Philip S.Yu, “Data Mining – An Overview from Database Perspective”, Knowledge, Volume 8, No.6, pp 866-883, Dec 1996.
3. Deepti Sharda and Sonal Chawla. "Web Content Mining Techniques: A Study." International Journal of Innovative Research in Technology & Science.
4. Johnson, Faustina, and Santosh Kumar Gupta. "Web Content Mining Techniques: A Survey.”International Journal of Computer Applications (0975–888) Volume (2012).
5. 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.
6. W. Himmel, U. Reincke, and H. Michelmann, “Text mining and Natural language Processing Approaches for automatic categorization of lay requests to web-based expert forums”, Journal of Medical Internet Research, vol. 11, no. 3, pp. 25, 2009.

7. Asai, T., Arimura, H., Abe, K., Kawasoe, S., S. Arikawa, Online Algorithms for Mining Semi-structured Data Stream, In Proc. IEEE ICDM'02, 27–34, 2002
8. B. Liu, Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, 2015.
9. Dr. Muhammad Shahbaz, Dr. Aziz Guergachi, Rana Tanzeel ur Rehman, "Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text", 27th IEEE Canadian Conference Electrical and Computer Engineering (CCECE), pp 1-7,4-7 May 2013.
10. Amir Ahmad, Lipika De, "A k-mean clustering algorithm for mixed numeric and categorical data" Data & Knowledge Engineering Elsevier,pp. 503-527,2007.
11. 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.
12. Niranjan Lal, Samimul Qamar, Savita Shiwani "Search Ranking for Heterogeneous Data over Dataspace" published in Indian Journal of Science and Technology(0974-6846), SCOPUS index Journal, , Volume 9, Issue 36, September 2016 (pp.1-9)..
13. Niranjan Lal, Samimul Qamar, Savita Shiwani "Information Retrieval System and challenges with Dataspace" published in International Journal of Computer Applications(0975 – 8887), Foundation of Computer Science (FCS), NY, USA , Volume 147 - Number 8 , August 2016(pp.23-28)..

AUTHOR'S PROFILE

Ms. Bhagayshree Pathak received B.Tech Degree in Computer Science and Engineering from Mody University and science and technology, Laxshmangarh, Sikar, Rajasthan, India, in 2015 and Pursuing M.Tech. Degree in Computer Science and Engineering from Mody University, Lakshmangarh, Sikar, Rajasthan, India .Her Research areas Database, Data Mining, Information Retrieval and Data Mining of Heterogeneous Data.

Mr. Niranjan Lal received B.E. Degree in Information Technology from Rajasthan University, India, in 2005, and M.Tech Degree in Information Technology from Guru Gobind Singh Indraprastha University Delhi, India in 2007. He is currently Assistant Professor in Dept. of Computer Science & Engineering at Mody University of Science & Technology Lakshmangarh, Sikar, Rajasthan, INDIA. His research areas are Database, Data Mining, Dataspaces, Computer Networks, Network Security, and Wireless Sensor networks, Cloud Computing, Mobile Computing, and Android Application Development.