Analysis and Implementation of Text Mining for Different Documents

[1]K. Maheswari, [2]P. Packia Amutha Priya
[1]Department of Computer Applications, Kalasalingam University, Krishnankoil, Tamil Nadu, India
maheswarisnr@gmail.com[1]
[2]Department of Computer Applications, Kalasalingam University, Krishnankoil, Tamil Nadu, India
p.packiaamuthapriya@klu.ac.in[2]

ABSTRACT

The process of making structured data from unstructured and semi structured text is called text mining. Text mining is defined as bag of words. The environment is set up with various documents in a database. The preprocessing of removing unwanted numeric values, uppercase, lower case, frequent words, punctuation is considered. In this work, the frequency of words occurred at least fifty times in a document is identified. The experimental results of the word frequency in a document occurred twenty times, twenty five times, fifty times and hundred times was analyzed and represented visually.

Keywords: Text Mining, Data Mining, frequency of words and text file

I. INTRODUCTION

Text mining and text data mining [1] is a growing field of text analytics. The process of extracting the quality information from text database is known as text analytics. The quality information is extracted through analysis process. The Process of structuring input text is done by the following process,

- Parsing
- Deriving features
- The removal of punctuations numbers and will be updated into a database.

There are many databases used for implementing. They are

- Multimedia databases
- Time-series and sequence database
- Text databases
- World-Wide Web databases
- Spatial databases

A variety of information collected for processing in digital form and stores it in databases for future. They are,

- Business transactions:
- Scientific data:
- Medical and personal data
- Surveillance video and pictures:
- Satellite sensing
- Games
- Digital media
- CAD and Software engineering data
- Virtual Worlds
- Text reports and memos (e-mail messages
- The World Wide Web repositories

Today maintaining large volume of data in a database is a challenging task. The issues and challenges are developing a model of Data Mining

- Scaling Up for High Dimensional Data and High Speed Data Streams
- Mining Sequence Data and Time Series Data

- Mining Complex Knowledge from Complex Data
- Data Mining in a Network Setting
- Distributed Data Mining and Mining Multi-agent Data
- Data Mining for Biological and Environmental Problems
- Data-Mining-Process Related Problems
- Security, Privacy and Data Integrity
- Dealing with Non-static, Unbalanced and Cost-sensitive data.

A. ISSUES AND CHALLENGES IN TEXT MINING

There are large number of issues occur when text mining process is carried out. These issues will affect the performance of decision making. Before applying text mining process there is a need of converting unstructured data in to structured one. The issues and challenges are,

1) Challenges in Text Mining

- Bulky datasets
- Noisy unstructured data
- Ambiguous words
  Example: apple - company or apple - fruit
- Framework understanding
  Example: automobile -car -vehicle
  – Two wheeler – Four wheeler
- Composite and slight relationship Example: “X develops software” “Software is developed by X”
- Multilingual

B. Text Mining Process

- Extracting Documents from the databases
- Text Transformation for processing
- Feature Extraction for analysis
- Reduce Dimensions
- Apply standard Data Mining for performance
- Interpretation / Evaluation for comparison

II. BACKGROUND STUDY

During text mining, there are a lot of issues and challenges which will definitely affect the process. The performance, efficiency effectiveness and accuracy of decision making are achieved by applying best algorithm. Before applying algorithm, pre-processing is performed. In this process, the rules and regulations are imposed to make the text process effective and efficient. This is simply defined as converting unstructured data into structured data. Variety of algorithms is used to perform text mining.

Ah-Hwee Tan [1] described a text mining structure with knowledge distillation and text refining. The author highlighted the challenges and issues of text mining. Ingo *et al.,*[2] the author surveyed the text mining facilities in R using statistical and machine learning methods. The framework was presented for text mining. The grammatical rules and context was not considered in this work.

Mustafa *et al.*, [3] described the study of original and winning pattern-based method such as pattern evolving and pattern deploying to discover the hidden pattern in the text documents. Abhishek Kaushik, and Sudhanshu Naithani [4], presented the review of various text mining tools and techniques. Zhang et al [5], introduced the research position of text mining, text classification, text clustering, association rule mining.

Abhilasha [6], analysed retrieval of text data to select right method for text mining is an important task. The author also focused on automatic text mining to find effective and easy to use method. Michele *et al.,* [7] the authors developed the text mining tool for linguistic, analysed the pros and cons of the text mining tool and compared with conventional pattern classification. Zhou *et al.,* [8], suggested a new improved KNN algorithm for text classification to avoid the complexity. The results are shown with greater accuracy. Songbo Tan [9] dealt uneven text data and NWKNN algorithm was proposed. The experimental result achieves performance improvement.

III. TEXT MINING

The purpose of text classification is to increase the identification of information. The text classification needs the following,

- It obtains documents in the form of text files, pdf files, html files and other file formats.
- It contains tree structured hierarchy which specifies the important information for the

institution, company, organization and association.

- Apply software to process document using text classification algorithms.
- To obtain data from the data bases, to process it and assign to the proper classification, text mining algorithm are used.

The frame work of data flow in this work is shown in figure1. Documents are collected and preprocessed for processing. The feature extraction of documents was performed.

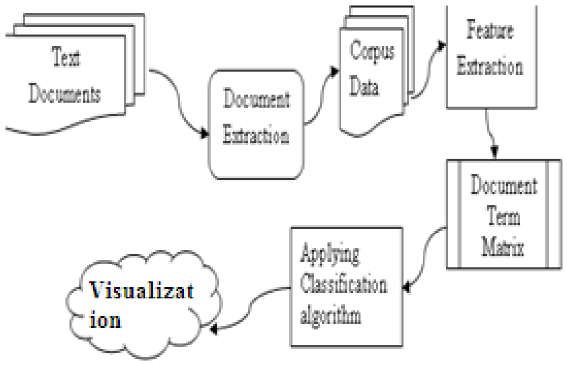

Figure 1: Frame work of proposed work

There are two approaches that are used for classification,

A. **Manual Approach**

This is the simplest approach used in text classification. It collects the key words which qualifies the type of information in a document. If more keywords are present, the key words are assigned to the topic of the document. Among the keyword the frequency of each keyword is calculated. The most frequently occurred word is assigned as topic.

B. **Statistical Approach**

This approach is based on the detection of a "training set" of the documents using data mining algorithms to find out the similarity. To deduce the key elements of the document various text classification algorithm such as Bayesian, LSA are used. In order to categorize the content, it makes use of frequency and key terms to build rules implicitly. If the classification is incorrect, there is no accuracy in result.

- **Searching of information in a database**: Data Storage and retrieval of text documents using keyword search.
- **Retrieval**: If data is present, it is extracted or retrieved from the data base.
- **Text clustering**: Text documents are grouped using data mining clustering methods.
- **Text classification**: the documents are classified using data mining classification methods based on trained dataset..
- **Web mining**: Data and text mining on the Internet is performed for finding interconnections of the web.
- **Information extraction**: The relevant data is Identified, retrieved and extracted from unstructured text;
- **Natural language processing**: The language processing is performed
- **Concept extraction**: Based on the words and phrases from the document, forming similar groups.

IV. EXPERIMENTAL RESULTS

The text files stored in the folder with the size of 1.27 MB was used in this research work. The text file with the size of 1.47 KB was considered for text mining. In this figure 2, the frequency of words in the document was found out and it is represented.

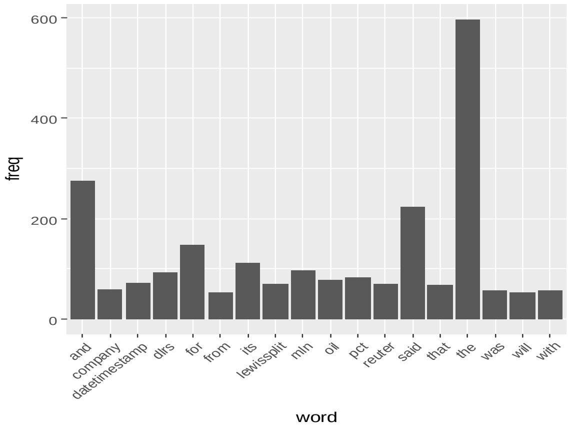

Figure 2: Word frequency of document

The most frequently used word in the document is identified and it is shown in the figure 3.

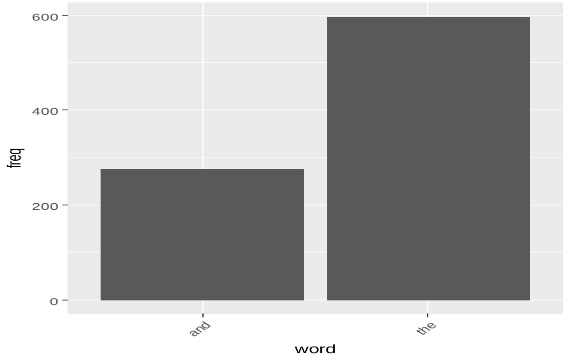

Figure 3: Most Frequent words

The identified frequent words are visually represented in word clouds and it is shown in figure 4.

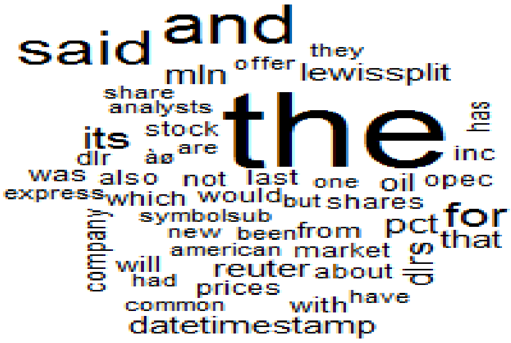

Figure 4: Cloud of Frequency word in a document

The words that occur at least twenty five times are represented visually in the figure 5.

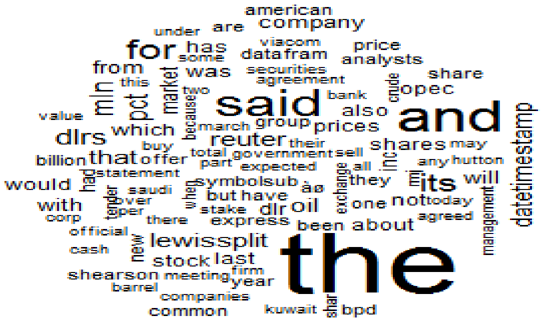

Figure 5: Cloud of Frequency word at least twenty five times in a document

The color cloud was created with the word frequent occurrence of at least twenty times is shown in figure 6.

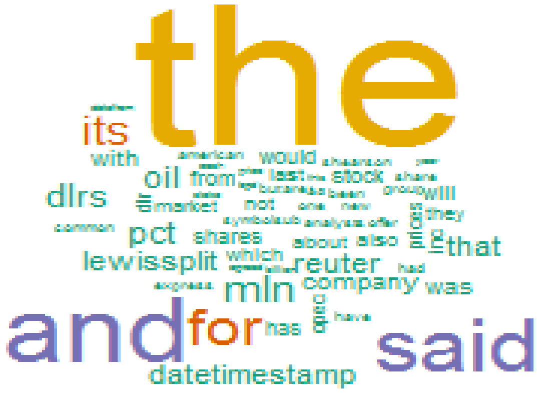

Figure 6: Color Cloud of Frequency word at least twenty times in a document

The color cloud with the word occurrence of at least hundred times in the document is shown in figure 7.

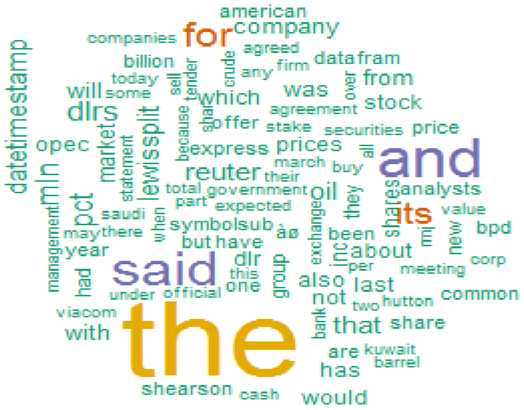

Figure 7: Color Cloud of Frequency word at least hundred times in a document

V. CONCLUSION

Text mining process is data mining is an important task in today’s big data maintenance. When maintaining a huge collection of files, there are many issues and challenges are faced by day to day users. This work provides a solution for finding the frequency of words with different occurrences. The experimental results were represented visually. The future work focused for applying advanced data mining algorithms for huge volume of data.

VI. REFERENCES

[1] Ah-Hwee Tan, “Text Mining:The state of the art and the challenges”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012

[2] Ingo Feinerer, Kurt Hornik , David Meyer "Text Mining Infrastructure in R", Journal of Statistical Software March 2008, Volume 25, Issue 5.

[3] Mustafa M. Shaikh, Ashwini A. Pawar, Vibha B. Lahane, Pattern Discovery Text Mining for Document Classification, International Journal of Computer Applications, Volume 117 ,No. 1,May 2015,PP:6-12.

[4] Abhishek Kaushik, and Sudhanshu Naithani, "A Comprehensive Study of Text Mining Approach", IJCSNS, VOL.16 No. 2, February 2016, PP: 69 – 76.

[5] Yu Zhang, Mengdong Chen, and Lianzhong Liu, "A review on text mining", published in IEEE Xplore digital library, Software Engineering and Service Science (ICSESS), 2015 6th IEEE International Conference on 23-25 Sept. 2015.

[6] Abhilasha Singh Rathor and Dr. Pankaj Garg, "Analysis on Text Mining Techniques", IJARCSSE , Volume 6, Issue 2,February 2016, ISSN: 2277 128X, pp: 132- 137.

[7] Michele Fattoria, Giorgio Pedrazzib, and Roberta Turrab, "Text mining applied to patent mapping: a practical business case" World Patent Information, published in Elsevier, Volume 25, Issue 4, December 2003, Pages 335–342.

[8] Zhou Yong, Li Youwen and Xia Shixiong, "An Improved KNN Text Classification Algorithm Based on Clustering", JOURNAL OF COMPUTERS, VOL. 4, NO. 3, MARCH 2009, pp: 230- 237.

[9] Songbo Tan,"Neighbor-weighted K-nearest neighbor for unbalanced text corpus", Expert Systems with Applications,Volume 28, Issue 4, May 2005, Pages 667–671