An approach of Clustering and analysis of Unstructured Data

*Gunisetti Tirupathi Rao, #Dr. Rajendra Gupta

**Research Scholar, Rabindranath Tagore University, Raisen*
#Associate Professor, Rabindranath Tagore University, Raisen

Abstract

Unstructured dataset is a kind of information which is not pre-defined and it is organized in improper manner, dataset contains different data like email, chat, images, video, xml, links etc. It is very complextask to search words in unstructured dataset.To get particular piece of information/search a pair from dataset, there arefour approaches are applied viz. First, Pre-process the dataset, using TPMRFC and assign weight using DTM, Second, Re-calculate the error and update the weights of matrix(DTM), third,Cluster the each term according to its weight using Self Organization Map,lastly, Using Least Frequently Used (LFU) with Dynamic Aging (LFUDA) method by which the pair of words is more frequently used to place in Cache. For testing the proposed scheme PROLOG unstructured dataset is used and results are achieved in terms of accuracy.

Keywords: *Unstructured data, Text Pattern Mining, Cluster, Self-Organized Map*

I. INTRODUCTION

Most of the organizations have two kinds of data; structured data and unstructured textual data. With the development of Internet and social network, huge data is being collected. In today'sscenario, superiority of decisions are made on the basis of structured data because it is easy to analyze the data. If we explore the unstructured data it includes Images, video, audio, education e-mails, emails, tweets, blogs, reviews, status updates, surveys, legal documents, and so much more chats, or other electronic documents or Unstructured Data refer to information which does not have a pre-defined data model or is not organized in a pre-defined manner. Deriving a particular word from unstructured data takes more time. Although unstructured data is usually text heavy and difficult to analyze, many researchers use unstructured data to extract sentiment, construct sentiment index and predict return[1].

Many researchers use text mining and analyze methods for pre-processing the unstructured data. Text mining is also called as knowledge discovery or data mining.It is a intelligent way for discovering useful patterns &knowledge in text documents.Convert the unstructured text into a structured matrix. First transform this unstructured data into a structured dataset and then continue with normal modeling framework [3]. The additional step of transforming an unstructured data into a structured format is simplified by document word matrix. DTM is a matrix that consists of the occurrences of words or terms in documents. In DTM, if the word exists in a particular document, the matrix entry corresponding to that row and column is presented as 1, otherwise it will be recorded as 0. Let say, if the word appears triple in that specific document then it will be recorded as three (3) in that particular matrix entry [4].

II. RELATED WORK

The author in [1] defines structured data and unstructured data. Unstructured data is data comes from machines generated and it is broadly classified into two types i) Non-Textual unstructured data is a multimedia data like still images, videos, and MP3 audio files ii) Textual unstructured data examples are like email messages, collaborative software and instant messagesDr.Goutam Chakra borty [13] et.al in 2014 proposed an outlook at how to analyse textual data for extracting insightful customer intelligence from a large collection of document. Using SAS text miner and SAS sentiment analysis studio artificial neural network regression model is used for variable selection to predict the target variable.

Theauthor M. Siva Lakshmi and MD. Arsha Sultana[2] removed unnecessary character from unstructured dataset by implementing text mining using R programming language. R is used to mine unstructured data which is the most exhaustive statistical analysis package and it incorporates all of the standard statistical tests, models and analyses for managing and manipulating data. By using R only useful information can be gathered by removing unnecessary nonessential characters from unstructured data. So, we can easily predict the status of firm retrieving header frequency from unstructured data".

The author ZurainiZainol, Puteri N.E. Nohuddin, Tengku A.T. Mohd and Omar Zakaria [3] demonstrate different text mining techniques is being discussed. The provides information and brief idea on text mining, its advantages, applications and various text mining techniques that can be used for effective and efficient document analysis that in turn will provide information to build product roadmaps and make better decisions about their activities.

In paper [4], the author provides a state-of-the-art survey of various applications of Text mining to finance. review of Text mining applications in the financial domain These applications are categorized FOREX rate prediction, stock market prediction, customer relationship management (CRM) and cyber security. The author reviewed 89 research papers that appeared during the period 2000–2016, highlighted some of the issues, gaps, key challenges in this area and proposed some future research directions. Finally, this review can be extremely useful to budding researchers in this area, as many open problems are highlighted.

In the paper [5], authorshave proposed a new text mining approach that uses structured knowledge resources to extensively extract a very large amount of information about microorganism habitats and phenotypes from scientific literature in food microbiology.The purpose of this paper is to addresses the lack of available structured information on this subject. The resulting information is structured by relationships and hierarchies that one can efficiently search by using a semantic search engine, AlvisIR Food.

The author [6] explored smart grid issues related to the technologies, markets and so on to drive more effective smart grid projects.Using smart grid searching technique through text mining.the paper helps in helps to promote the development of scientific research in future sign analysis through text mining by presenting some challenges in the research methodology.

The author in their paper [7] the author propose a co-clustering with adaptive local structure learning based on nonnegative matrix tri-factorization method The proposed unified learning framework performs intrinsic structure learning and tri-factorization (i.e., 3-factor factorization) simultaneously. The intrinsic structure is adaptively learned from the results of tri-factorization, and the factors are reformulated to preserve the refined local structures of the textual data. In this way, the local structure learning and factorization can be mutually improved. The proposed method solve the optimization problem and efficient iterative updating algorithm is proposed with guaranteed convergence. Experiments on benchmark textual data sets demonstrate the effectiveness of the proposed method.

The Selection criteria for text mining approaches have been demonstrated in [8]. The author proposeddifferent criteria to evaluate the effectiveness of text mining techniques and attempted to facilitate the selection of appropriate technique. The two main criterion used in this paper are: General & Specific categories. In General: General Criteria is based on Usability, Comprehensiveness and Flexibility. Specific: Those criteria divided into different sub-criteria,like Graphic User Interface, goals of research, satisfaction level, and KPI and Support Compliance.

In the paper [9] the author proposed DDKM algorithms built upon the double K-means to address the problem of document-term co-clustering. The proposed algorithms seek a diagonal block structure of the data by minimizing a criterion based on both the variance within the class and the centroid effect. The author evaluate results usingsynthetic data sets, and real data sets commonly used in document clustering.

The Energy Saving in a Wireless Sensor Network by Data Prediction by using Self-Organized Maps is proposed by the author [10]. The author presented predictive analysis method based on a unsupervised type of machine learning algorithm : the kohonen maps. The model exploited in a network of sensors to reduce the number of transmission on the network. The aim of this paper is to present a learning algorithm and the result obtained in the context of smart city.

In [11], the author proposed an alternative scheme which emulates SOM as a visualization tool for large datasets. the proposed scheme complexity for large N is O(N) as compared to SOM which is of order O(N2). Future considerations include sensitivity analysis and addressing the limitation of MDS on higher dimensional data (curse of dimensionality).

TextFlows: A visual programming platform for text mining and natural language processing concept given by Macario O. Cordel II and Arnulfo P.Azcarraga in their research [12]. The paper presents text mining and language processing modules, the author describes pre-composed workflows. First the author compared different document classifiers, part-of-speech, text categorization problem, outlier detection in document. It infeasible and not more efficient for very large dataset.

III. PROPOSED SCHEME

The proposed approaches first collect the web document text. This web text data is having word with space, comma, semi-colon and many more. These include data cleaning, session identification, path completion, and formatting. It is done in order to improve the quality of the results. So first store these word pair in the proper format without space and special character (comma, semi-colon etc.) and convert it in tabular format. Here the synonym of text are also stored and mapped with main text. If text is mapped with two synonyms then priority is also maintain to set to each them based on Radial Based Function using weighted similarity. The flow of the data filtration is demonstration as in following flowchart.

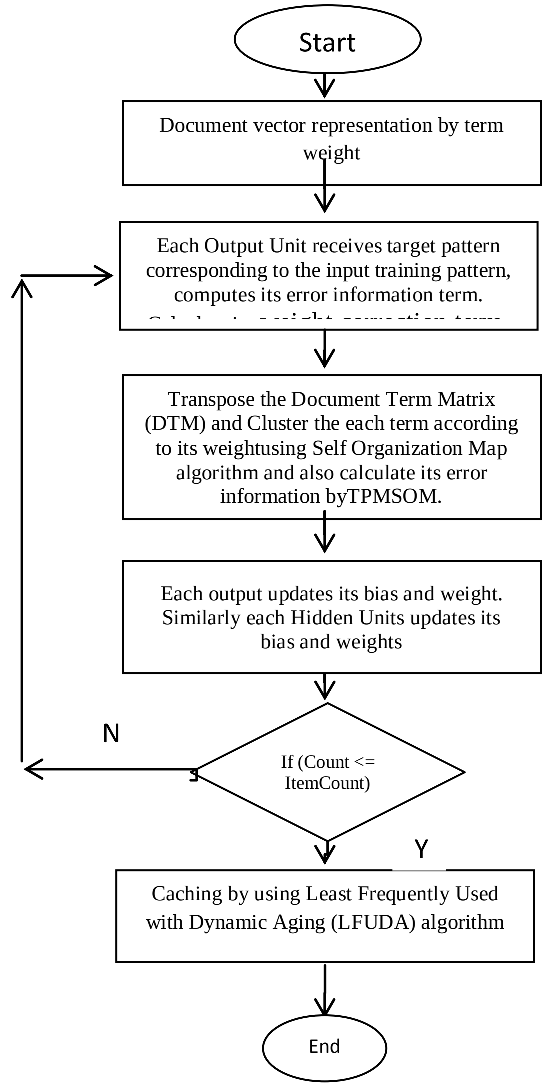

Start
Document vector representation by term weight
Each Output Unit receives target pattern corresponding to the input training pattern, computes its error information term.
Calculate its weight correction term
Transpose the Document Term Matrix (DTM) and Cluster the each term according to its weightusing Self Organization Map algorithm and also calculate its error information byTPMSOM.
Each output updates its bias and weight. Similarly each Hidden Units updates its bias and weights
N
If (Count <= ItemCount)
Y
Caching by using Least Frequently Used with Dynamic Aging (LFUDA) algorithm
End

Figure : Process of filtering Term Documentusing proposed algorithm

A. Mathematical Models

a) A Document Term Matrix

A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms.

The term in the documents can be calculated as follows :

Table 1 : Document Term Matrix using four document files

doc1	Two for tea and tea for two
doc2	Tea for me and tea for you
doc3	You for me and me for you
Doc4	We for us and us for world

Table 2 : Occurrence of words in all the document files

	Two	tea	Me	You
doc1	2	2	0	0
doc2	0	2	1	1
doc3	0	0	2	2
doc4	0	0	0	0

We can transfer the same model to modeling documents as terms. Each term belong to term |T| in the vocabulary which becomes a dimension. The document's position in the dimension of term t is determined by sd,t (score of term t in document d – for now, just the term frequency). Each document becomes a point in |T|-d term space.

The weight of a term's appearance in a document is frequently calculated by combining the Terms Frequency (TF) in the document with its Inverse Document Frequency (IDF). This can be expressed as follows :

$$w_{t,d} = tf_{d,t} * idf_t$$

This term-document score is known as TF*IDF, and is quite widely used.

B. Generating Method

In first phase, the document vector has been generated.The document vector is represented in terms of weight, during this process in which irrelevant data are omitted for calculation purposes. Here vector forms by determining the term weight by using Radial Basis Function. In this process, tokens are created by segmenting the strings by white space and punctuations called Tokenization.

C. SELF ORGANIZED MAP

Self-Organized Map (SOM) is a type of artificial neural network(ANN), that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Best thing about SOMs is visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. A self-organizing map consists of components called nodes or neurons. We Transpose the document term matrix and cluster each term according to weight using self-organization map.

Syntax:
selforgmap(dimensions,initNeighbor,topologyFcn, distanceFcn)

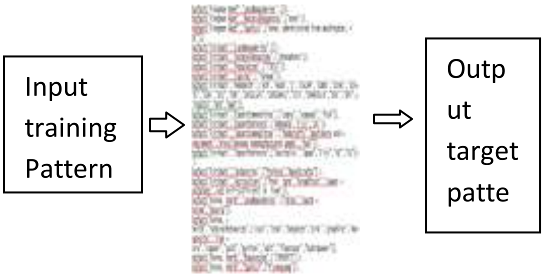
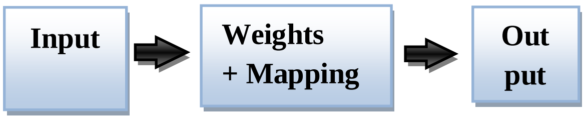


Fig 3:Text preprocessing

SOM Layer

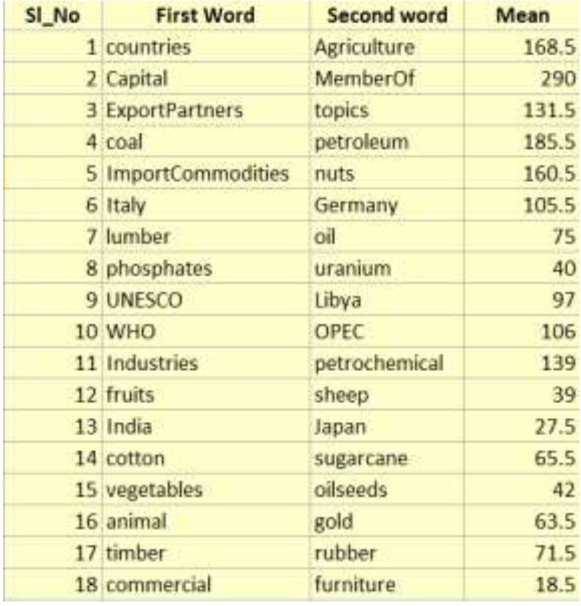


Results

To analysis the data, MATLAB R2016a is used for implementing the algorithm. As per the data given in standard PROLOG dataset, a pair of words Coal and petroleum, Algorithm scan's whole document and calculate the pair of words and Self Organized Map(SOM) is generated. In figure 1,Count of "***Coal***" is 61, it represents *orange* color, Count of second word "***petroleum***" is 280 it represents *brown* color.

Unstructured data are data that have no fixed data model. Without preprocessing, unstructured data, it cannot be stored in a table, Examples: social media (tweets, blogs, posts, etc.), call center data, email, etc. The dataset available in PROLOG database which contain countries and its related items. The prolog document includes different search log across the country. Using Matlab we analyze the pair of words from dataset. It clearly shows how combining the text data with numerical data gives better accuracy in predicting the target attribute. Based on the weighs we can predict the frequently used words from the huge data set, and place the pair of word into cache.

Table : Dataset contains occurrenceof first and second wordin PROLOG dataset and its mean value



Sl_No	First Word	Second word	Mean
1	countries	Agriculture	168.5
2	Capital	MemberOf	290
3	ExportPartners	topics	131.5
4	coal	petroleum	185.5
5	ImportCommodities	nuts	160.5
6	Italy	Germany	105.5
7	lumber	oil	75
8	phosphates	uranium	40
9	UNESCO	Libya	97
10	WHO	OPEC	106
11	Industries	petrochemical	139
12	fruits	sheep	39
13	India	Japan	27.5
14	cotton	sugarcane	65.5
15	vegetables	oilseeds	42
16	animal	gold	63.5
17	timber	rubber	71.5
18	commercial	furniture	18.5

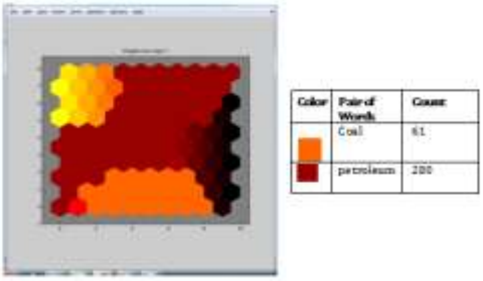


Figure : Simulation of word count for 'Coal' and 'Petroleum' keywords

Table :PROLOG dataset Index value of Pair ofWords

Sl.No	First word	Second word	Mean
1	**Countries**	**Agriculture**	
	264	73	168.5
2	**Capital**	**Member Of**	
	323	257	290
3	**Export Partners**	**Topics**	
	258	5	131.5
4	**Coal**	**petroleum**	
	91	280	185.5
5	**Import Commodities**	**Nuts**	
	258	63	160.5
6	**Italy**	**Germany**	
	78	133	105.5
7	**Lumber**	**Oil**	
	18	132	75
8	**Phosphates**	**uranium**	
	39	41	40
9	**UNESCO**	**Libya**	
	177	17	97
10	**WHO**	**OPEC**	
	198	14	106

For example, a search for the keyword "***Countries***" will yield far more results than a search for the keywords "***Agriculture”***. In the above table the word “***Capital***” and “***MemberOf***” weight is 290 which shows these two key words are used frequently by the user. At the next hit time the weight of same pair of words may varies,that all depends on search engine.The frequently used word which having high-test weight is placed in cache.

Table : Document Term Matrix for keyword 'India' and 'country' for different document files

'File Name'	Key word 'India'	Keyword 'country'
'doc1.txt'	[7]	[0]
'doc2.txt'	[11]	[0]
'doc3.txt'	[0]	[0]
'doc4.txt'	[15]	[0]
'doc5.txt'	[1]	[0]
'doc6.txt'	[4]	[0]
'doc7.txt'	[1]	[0]
'doc8.txt'	[0]	[0]
'doc9.txt'	[4]	[0]
'doc10.txt'	[0]	[0]

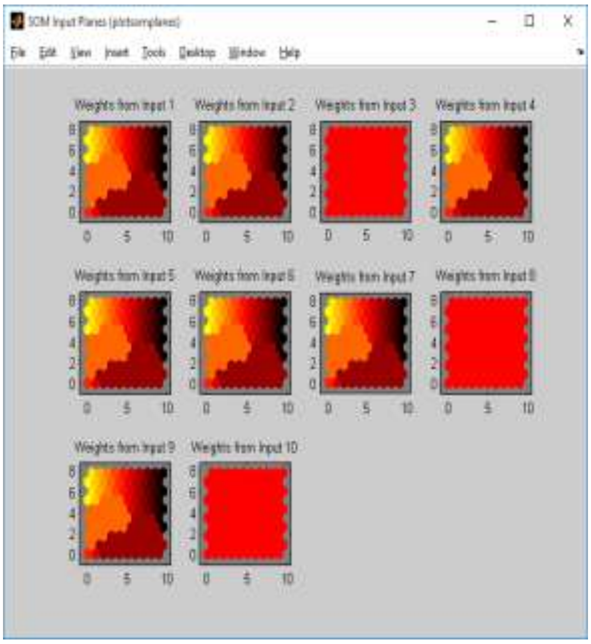


Figure :TPMSOM for pair of words “India”, “Country”

MATLAB produce the above simulation results after applying the proposed algorithms for searching the similar key terms. In the above Figure, the algorithm check each document and calculate appearance of words but inputting the term “India” & “country” as pair of inputs. The weight is assigned to each input in every document which can be seen in 'Weights from Input 3'. The pair of words arenot present in 'Weights from Input 3, 8, 10' so the color is filled with red.

Conclusion

This research work presents an approach of clustering and analysis of unstructured data. The proposed method involve data mining algorithms and various attributes related to unstructured data and filter the data as per the given datasets. The combination of data mining algorithms takes a considerable time to process large data sets and produces better results. The performance of the proposed method is analysedby associating TPMRFC and assign weight using DTM, Weights of matrix (DTM), Clustering the each term according to its weight using Self Organization Map. The proposed method give better outcomes as compared to earlier proposed methods.

REFERENCES

[1] LiGuoa, FengShi, &JunTu, (2016), “Textual analysis and machine leaning: Crack unstructured data in finance and accounting”, The journal of Finance and Data Science, Vol. 2, Issue 3, pp. 153-170.

[2] M. Siva Lakshmi1 and MD. Arsha Sultana,(2016), “Text Mining of Unstructured Data Using R”, International

Journal of Computer Science and Engineering (JCSE), Vol.4,Issue 9, pp.123-130.

[3] ZurainiZainol, Puteri N.E. Nohuddin, Tengku A.T. Mohd and Omar Zakaria, (2017), "Text Analytics of Unstructured Textual Data: AStudy on Military Peacekeeping Document using R Text Mining Package", The 6th International Conference on Computing & Informatics (ICOCI17), At Sepang, Malaysia,pp.1-19

[4] K.V.Kanimozhi1 and Dr.M.Venkatesan,(2015), "Unstructured Data Analysis-A Survey", International Journal of Advanced Research in Computer and Communication Engineering,Vol. 4, Issue 3, pp.223-225.

[5] M. Siva Lakshmi1 and MD. Arsha Sultana, (2016), "Text Mining of Unstructured Data Using R",International Journal of Computer Sciences and Engineering (JCSE), Vol 4, Issue 9, pp.123-130.

[6] B. ShravanKumar&VadlamaniRavia, (2016), "A Survey of the Applications of Text Mining in Financial Domain",Knowledge-Based System, Vol 114, pp 128-147.

[7] ChankookParka&SeunghyunChob,(2017), "Future Sign Detection in Smart Grids Through Text Mining", International Scientific Conference Environmental and Climate Technologies, CONECT, Vol 128, pp. 79-85.

[8] ShudongHuang ZenglinXuand JianchengLv, (2018), "Adaptive local structure learningfor document co-clustering" Knowledge-Based Systems, Vol. 148, pp.74-84.

[9] Hussein Hashimi , Alaaeldin Hafez and HassanMathkour (2015), "Selection Criteria for Text Mining Approaches", Computer in Human Behavior, Vol 51, pp.729-733.

[10] Charlotte LaclauandMohamed Nadif, (2016), "Hard and fuzzy diagonal co-clustering for document-term partitioning", Neurcomputing, Vol 193, pp 133-147.

[11] Adrien Russo , François Verdier and BenoîtMiramond, (2018), "Energy Saving in a Wireless Sensor Network by Data Prediction by using Self-Organized Maps", Procedia computer science, Vol 130, pp 1090-1095.

[12] Macario O. Cordel II and Arnulfo P.Azcarraga(2015), "Fast Emulation of Self-organizing Maps for Large Datasets", Procedia Computer science, Vol. 52, pp 381-388.

[13] Dr. Goutam Chakra borty, Murali Krishna Pagolu (2014) "Text Analytics and Sentiment Analysis of Unstructured Data", International Conference on Analysis of Unstructured Data held at SAS Global Forum, Washington D.C.