Effective Text Data Extraction using Hierarchical Clustering Technique

D. Saravanan

Abstract— ***In the digital era, text plays an important role it has come forward in a variety of data mining applications. In text mining, information are extracted and clustered from an unstructured data-set. For efficient retrieval many procedures are involved. Text mining is used in a variety of data mining approaches such as market research, survey research, statistical process and more. The objective of this paper is to analyze the relevant data that leads to a novel multidimensional data mining package. The method is based on the use of text mining. Data collection and analysis of data related to the text of a real-world test are also presented***

Index Terms—Text Mining, Clusters, Text Clusters, Segmentation, Text Retrieval, Hierarchical cluster

1. INTRODUCTION

Technology makes the operations of a company more efficient. With the help of technology, mangers are able to perform their day to day operations and decision making more quickly. Today the available data sets are not easy to operate as they are more complex and various different structures are available. From this complex datasets deriving any decision is not an easy task. It requires additional knowledge on the domain and additional efforts are needed . Because of these complex data sets users are diverted to specialized tools such as machine learning, artificial intelligence, data mining, and knowledge extraction is required. Creating and storing of any data types are easy now, the same gives more burden to the user community, because of this unstructured data set user first motivated to data preprocessing steps, because of the impurity in the data sets extract the relevant data sets are not an easy task. Extracting the relevant information required some preprocessing operations, in text retrieval necessary to create stop words, index terms to retrieve the relevant documents. For that many documents today creates a clusters of words to extract any in formations more effectively [3,4]. These words are to help the user to extract similar content or similar words or paragraphs.

Text data mining process is based on a single word or single sentence or a whole document. Extracting such relevant documents in formations is properly indexed and creates index terms for easy access . Before creating the index terms, the user first creates a document matrix that is create a mass of the text document. This task complete with help of collected all related documents it may be any format such as text, doc, pdf, XML any format. Once this collection process was done next immediate steps convert the unstructured information into structured information i.e. different formats are converted into a single format for easy processing. This simple form information’s are stored separately for further operations. Next step to create a document matrix here columns are represents the various available documents and rows are frequent terms. This initial process are required for any text data mining further this operations combined with other tools such as mathematical or algebraic technique used to retrieve the need knowledge.

2. APPLICATIONS OF TEXT MINING

Data mining is the process of extracting the concealed information from the available huge data sets. This tool today is used by most of the researchers and many companies drive a new rule or new procedure for their day to day operations. This tool helps in the day to day performance of the company in terms of doing market analysis, predicting customer behavior, trend analysis, market segmentation and more. This helps the business to forecast upcoming events to run the effective business operations and make effective decisions based on the formations derived. This tool helps to derive any decision based on the past proceedings with that helps to derive new decision making operations. It also helps to solve the business queries quickly which usually is more time consuming, without using this tool. Many of the businesses today run with huge integrated data sets which come in a variety of patterns and structures. Not only business operations apart from academic research, customers’ feedback, medical observations, medical history all are stored in the form of text. These tools are highly support without changing of its nature and perform the operation in terms of additionally without disturbing it functions.

2.1 Investigation study

In investigation study ex customer buying patterns done with help of preparing a variety of open ended question to the particular problem or which need to find solution. It helps the people to find the interesting relationship or finding new ideas or their view based on the available information source. This helps the authority to make conclusions on the particular domain.

2.2 Involuntary Text categorization

The most important operation of text mining was involuntary text categorization. One of the best examples for this application performs the sort our inbox and mechanically pushes certain mail contents in our spam folder based on the users mailing pattern or certain terms or language that are not likely to come into view in rightful communication. In such a

Revised Version Manuscript Received on 12 April, 2019.

D. Saravanan, Faculty of Operations & IT, ICFAI Business School (IBS), Hyderabad, The ICFAI Foundation for Higher Education (IFHE), Deemed to be university u/s 3 of the UGC Act 1956), Hyderabad, India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

fashion this unwanted communications are deleted. In the same way this applications also used segregate the messages based on the usage and direct towards the particular department or particular functions without human knowledge

2.3 Used in Market examine development

In the digital era most of our inputs are collected in text form. For example customer feedback are recorded and analyzed based on the reviews, comments given by the customer. After these samples are collected and clustered based on the feedback received it helps the business improve their efficiency. In same in many of the business today send our problems through text formats. This information's are collected and analyzed with various text mining procedures. This functions today used in most of the places like medical, education, automobile and more.

3. EXPERIMATAL SETUP

In the information age the rapid growth of amount of in formations and facts are collected and stored for various data analysis. Majority of this text data's are unstructured based on the research outcome 90% of the data are stored in site either in web or company domains are un structured. Text mining or text retrieval is the partly programmed technique for extracting the useful information from the available data set. No single technique or method suitable for every application it varies based on the content. Technique used here to extract the meaningful information is:

3.1 FINDING a Key words: Input data's collected from multiple source and most of this in formations are un structured, searching the entire document create time consuming function for that each documents identify the key terms or search terms based on the content. Most of the search engine today works well based on the key phrase searching. This process is done with the help of creating a table denoting the occurrence of a particular word in the document.

3.2 ELIMINATE STOP words: Listing the input document are really challenging job for many researchers. Documents consist of unwanted information such as special symbols, numbers, figures, flow symbols are eliminated from the input document. It reduces the searching time also bring the effective results.

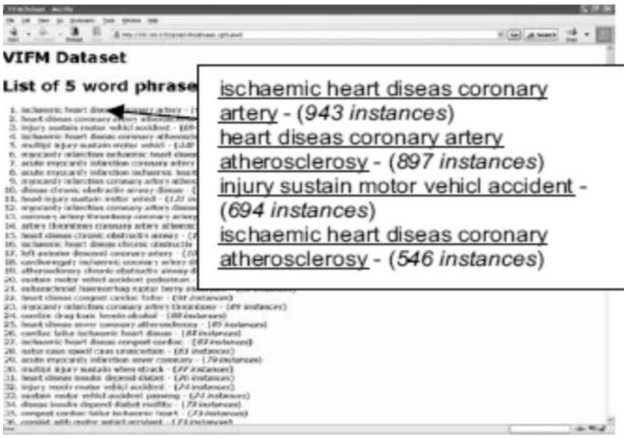
3.4 Synonyms and phrases. Synonyms, such as "sick" or "ill", or words that are used in particular phrases where they denote unique meaning can be combined for indexing. For example, "Microsoft Windows" might be such a phrase, which is a specific reference to the computer operating system, but has nothing to do with the common use of the term "Windows" as it might, for example, be used in descriptions of home improvement projects.

3.5 Stemming algorithms. An important pre-processing step before indexing of input documents begins is the stemming of words. The term "stemming" refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of verbs are identified and indexed (counted) as the same word.

3.6 SUPPORTS for different languages. Stemming, synonyms, the letters that are permitted in words, etc. are highly language dependent operations. Therefore, support for different languages is important. "Reliability" they also include the term "defects" (e.g., make reference to "no defects"). However, there is no consistent pattern regarding the use of the terms "economy" and "reliability," i.e., some documents include either one or both. The idea of latent semantic indexing is to identify such underlying dimensions (of "meaning"), into which the words and documents can be mapped. As a result, we may identify the underlying (latent) themes described or discussed in the input documents, and also identify the documents that mostly deal with economy, reliability, or both. Hence, we want to map the extracted words or terms and input documents into a common latent semantic space.

4. EXPERIMENTAL RESULTS

A search tool was developed. This allowed the user to give several words as input and find the nodes that had the most dominant occurrences of those words. Although searching from words enhanced the navigation of the map, it was a bit difficult to decide on which words to put in the search. It was evident that some guidance in the selection of the words was useful. An additional set of links based on the most occurring 2, 3, 4 and 5 word phrases were listed for the user as illustrated in Figure 3.



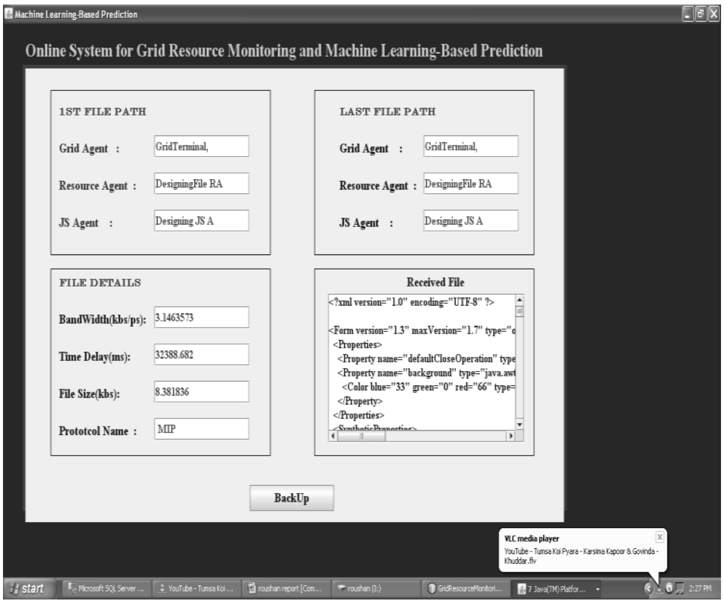
VIFM Dataset

List of 5 word phrase

ischaemic heart diseas coronary artery - (943 instances)
heart diseas coronary artery atherosclerosy - (897 instances)
injury sustain motor vehicl accident - (694 instances)
ischaemic heart diseas coronary atherosclerosy - (546 instances)

Fig 1. List of 5 word phrases in the dataset

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication



Machine Learning-Based Prediction

Online System for Grid Resource Monitoring and Machine Learning-Based Prediction

1ST FILE PATH

Grid Agent : GridTerminal,

Resource Agent : DesigningFile RA

JS Agent : Designing JS A

LAST FILE PATH

Grid Agent : GridTerminal,

Resource Agent : DesigningFile RA

JS Agent : Designing JS A

FILE DETAILS

BandWidth(kbs/ps): 3.1463573

Time Delay(ms): 32388.682

File Size(kbs): 8.381836

Prototcol Name : MIP

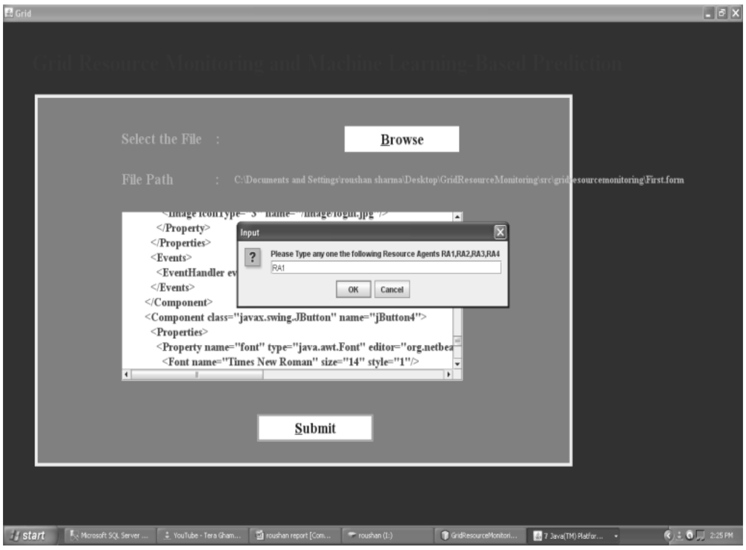
Received File

```
<?xml version="1.0" encoding="UTF-8" ?>

<Form version="1.3" maxVersion="1.7" type="o
 <Properties>
  <Property name="defaultCloseOperation" type
  <Property name="background" type="java.awt
   <Color blue="33" green="0" red="66" type=
  </Property>
 </Properties>
```

VLC media player
YouTube - Tumsa Koi Pyara - Karsima Kapoor & Govinda - Khuddar.flv

Fig 2. File Uploading



Grid

Grid Resource Monitoring and Machine Learning-Based Prediction

Select the File :

File Path : C:\Documents and Settings\roushan sharma\Desktop\GridResourceMonitoring\src\grid esourcemonitoring\First.form

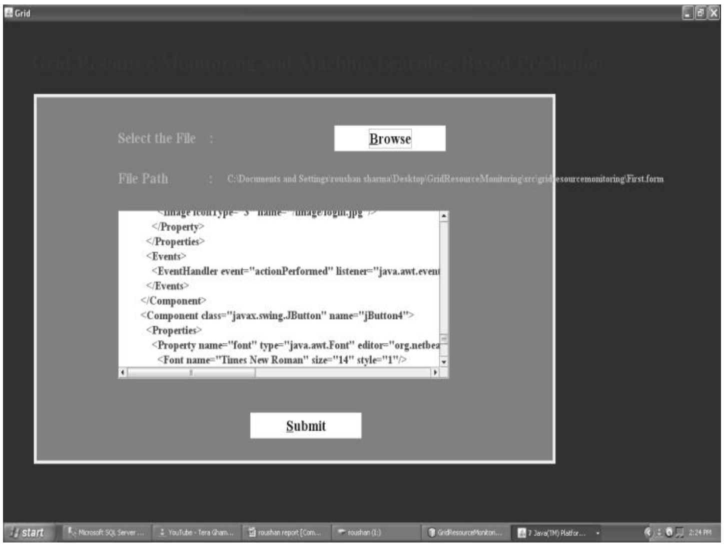
```
     <Image iconType="3" name="/image/login.jpg"/>
    </Property>
   </Properties>
   <Events>
    <EventHandler ev
   </Events>
  </Component>
  <Component class="javax.swing.JButton" name="jButton4">
   <Properties>
    <Property name="font" type="java.awt.Font" editor="org.netbea
     <Font name="Times New Roman" size="14" style="1"/>
```

Input

Please Type any one the following Resource Agents RA1,RA2,RA3,RA4

RA1

Fig 3. Data Entry (Text Entering)



Grid

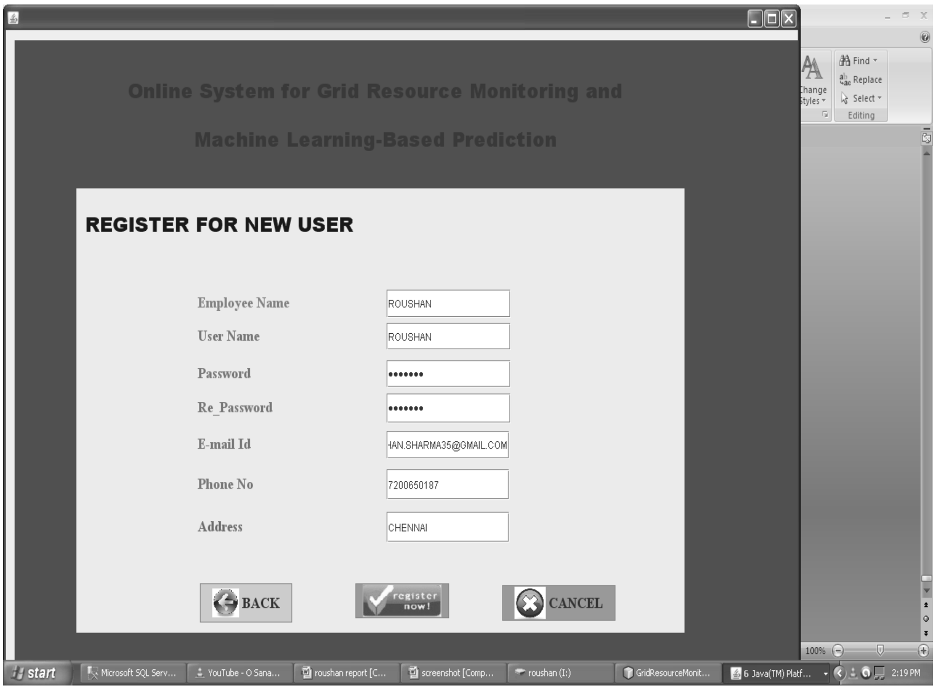
Grid Resource Monitoring and Machine Learning-Based Prediction

Select the File :

File Path : C:\Documents and Settings\roushan sharma\Desktop\GridResourceMonitoring\src\grid esourcemonitoring\First.form

```
     <Image iconType="3" name="/image/login.jpg"/>
    </Property>
   </Properties>
   <Events>
    <EventHandler event="actionPerformed" listener="java.awt.event
   </Events>
  </Component>
  <Component class="javax.swing.JButton" name="jButton4">
   <Properties>
    <Property name="font" type="java.awt.Font" editor="org.netbea
     <Font name="Times New Roman" size="14" style="1"/>
```

Fig 4. Searching the text information



Online System for Grid Resource Monitoring and

Machine Learning-Based Prediction

REGISTER FOR NEW USER

Employee Name: ROUSHAN
User Name: ROUSHAN
Password: •••••••
Re_Password: •••••••
E-mail Id: HAN.SHARMA35@GMAIL.COM
Phone No: 7200650187
Address: CHENNAI

Fig 2. User Registration Screen

5. CONCULSION

Traditional text mining and processing methods, mainly from large collections of documents have been used in the detection of similar documents. In this paper we Classical, text mining using traditional query languages can be difficult to implement, it is provided that can be used on databases corresponding text fields. This paper architecture, text, and other dimensional data much more efficiently than the traditional radical-based feature the largest collections of maps of the clusters that are capable of self-organizing map

REFERENCES

1. G. Slaton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing & Management, vol. 24, 1988.pp 513-523.
2. D.Saravanan, Dr.S.Srinivasan, (2013). , Matrix Based Indexing Technique for video data, Journal of computer science, 9(5), 2013, 534-542.
3. R. Amarasiri, D. Alahakoon, M. Premaratne, and K.Smith, "Enhancing Clustering Performance of Feature Maps Using Randomness", presented at Workshop on Self Organizing Maps (WSOM) 2005, France, 2005.pp 463-470
4. D.Saravanan, A.Ronald Tony " Text Taxonomy using Data Mining clustering System", Asian Journal of Information Technology, Volume 14(3),March2015
5. M. A. Hearst, "Untangling Text Data Mining", presented at 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, USA, 1999.pp
6. D.Saravanan" Segment Based Indexing Technique for Data file ", Procedia of computer Science,87(2016), Pages 12-17,ISSN: 1877-0509, June 2016
7. A. Rajan, GP Ramesh, J yuvaraj,"Glaucomatous image classification using wavelet transform" ,In the proc.of Advanced Communication Control and Computing Technologies (ICACCCT), 2014 ,Pages 1398-1402.
8. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", presented at 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C.,1993.pp 207-216.
9. D.Saravanan, Dr.S.Srinivasan (2011). A proposed new Algorithm for analysis for analysis of Hierarchical clustering in video Data mining, international journal of Data mining and knowledge engineering , vol 3, no 9.
10. R. Amarasiri, D. Alahakoon, and K. Smith, "HDGSOM: A Modified Growing Self-Organizing Map for High Dimensional Data Clustering", presented at Hybrid Intelligent Systems 2004, Japan, 2004.pp 216 – 221.
11. D.Saravanan," Image frame mining Using indexing technique" Data Engineering and Intelligent Computing , SPRINGER Book series, Chapter 12 , Pages 127-137, ISBN:978-981-10-3223-3,July 2017.
12. Ramesh.G.P A. Rajan," Comparative Study of Glaucomatous Image Classification Using Optical Coherence Tomography", International Journal of Pharmaceutical Sciences Review and Research, Volume 36, Issue1,Pages 277-280