International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-1, November 2019

Issues and Considerations for Effective Text Data Retrieval

D.Saravanan

Abstract: ***Text data retrieval is one of the major domain for extracting knowledge from the stored data sets. Within the text information, the text meaningful numerical codes extracted unstructured process information is to make the free text associated with the unstructured nature of data mining in a different stream. Number of procedure is constructed to Performing this operations most effectively. This paper focuses one of the text retrieval process, experimental results verified proposed methods works well with most of the documents.***

Keywords: Text Mining, Clusters, Text Clusters, Grouping of words, Text extraction.

I. INTRODUCTION

The use of computers to handle the day-to-day operations of a wide range of databases and record companies involved from the foundation. Computers and the ability to efficiently store and retrieve data in the field of marketing, finance , operations, sales for costing and many more. From this available stored huge volumes of information it is difficult to analysis the data sets. For reducing the analysis process today everywhere they start using computers. Computers not only used for analysis it also helps to visualize the result in various users need format. Today many tools supports for their analysis and visualize the data sets one of that knowledge extraction or knowledge discovery process. Traditional methods such as data reporting is a necessity for many companies; Etc. In some cases, the information recovery process using these traditional forms that are difficult to recover it. Several clarifications for the long text documents and databases because of the formless structure of the data sets. Text information are collected from various sources so it is difficult to fit those data sets in particular structure or format. User wish to retrieve any content from the stored document based on the content available on the document or based on the presented terms in the structure. Based on the terms user can group the term and formed complete structure. Using this term handler can done the comparison with any relevant information's.

II. OVERLAYING DATA ONTO TEXT CLUSTERS

Analysis this data structure need preprocessing steps. Because of the nature of the data sets. Data sets are collected from multiple place first it need to be arranged in proper format. After this data sets are groped based on the similarities it helps the users to reduce the searching and processing time.

Revised Manuscript Received on November 30, 2019.
***** Correspondence Author
D.Saravanan*, Faculty of Operations & IT ICFAI Business School (IBS), Hyderabad, India.

Grouping this information reduce most of the users operations effectively. In the proposed technique this grouping described in encrusted procedure. A tree based structure helps solve the problem effectively.

In this structure child nodes helps to form a clusters and root node helps to travel the information from the root to the last node in the structure. This travelling mechanism done either top to down or down to top fashion. It helps to transfers the needed information from the lower node to higher node and vice versa. This grouping mechanism helps gets the parental information. This information's gets transferred from parent to the child node effectively. Other than this tree structure this nodes are well distributed in nature instead of hierarchical structure. In this type of distributed separate nodes are done any operation based on his own, Here travelling or transformations are not done. Every node operations are done separately based on the information present on the node. However, with the increase in popularity of such networks, the attacks have also increased proportionally.

III. APPLICATIONS FOR TEXT EXTRACTION

Knowledge extraction or knowledge mining is the process of bringing the unknown fact or rule from the stored data sets. This source of information's it helps the user in various applications. This technology helps the various industrials functions with their available huge traditional data sets. Today this technology helps the industrial users to solve many applications such as predicting the customer behavioral patterns, risk involved in investment during the new startup operations, finding the errors on the operations and more functions through help of rule based functions. All this operations are done through help of the past result or past rule based operation. This analysis tool help the business to find the risk involved in the business. Some of the data mining functions helps the user to drive the business smoothly they are

A. In Survey Research

One of the major application of knowledge discovery in the area of marketing research. In market segmentation this process helps to understand the customer buying behavior, customer interest, market trend and more. It helps the responded to enter their respective thinking or judgment without any pressure on the previous knowledge or any hint.

B. Automatic Text Classification

One more important claim of text mining used in many mail server and creation of word document i.e. spontaneous text organization. This applications used in many mails with help of certain term or key words system automatically sends the mail to the spam folder. It avoids bulk of mails loaded in the user's inbox.

C. Used in Market Research Trends

Today most of the feed backs, customer query, customer complaints, product specification, user manuals all are comes in the form of text. Applications such as in guarantee declare or any patient medical history record it helps to complete certain information based on the previous history.

IV. TEXT MINING PROCESS

Any text mining process starts with initial scanning of the entire document. After followed by finding the number of terms, special characters any special symbols and numerical etc. After calculating this information text mining process construct a table format consists of number of each entry identified in the second step. This process further extended to finding the similarity of various words, common words available in the list.

V. PROPOSED METHODOLOGY

A. Removing special characters, Numeric, small terminology etc.:

After scanning entire document text mining finds the special character, numerical, repeated terms, similar terms are identified. This process was done before the indexing of the entire file.

B. Include lists, exclude lists:

Creation of key words or stop words are most important step in text mining because this terms are helps to find the relevant document more quickly and more accurately. With help of this stop words user may analysis or retrieve the needed relevant documents easily. Help of this stop words many researchers are creating an indexing this helps to categorize the documents and also helps to retrieve the content very quickly based on the users input query. Most of the stop words i.e. “what”, “how”, “or” normally most of the computer searching process are ignored this words. This words are used for message purpose most effectively, but in the transcribed information this stop words are not play any important role. In text mining user need to identify which of the words to be included or which are all the words are to be removed before user can create any documents. It is one of the important steps to identify which need to include which not to include.

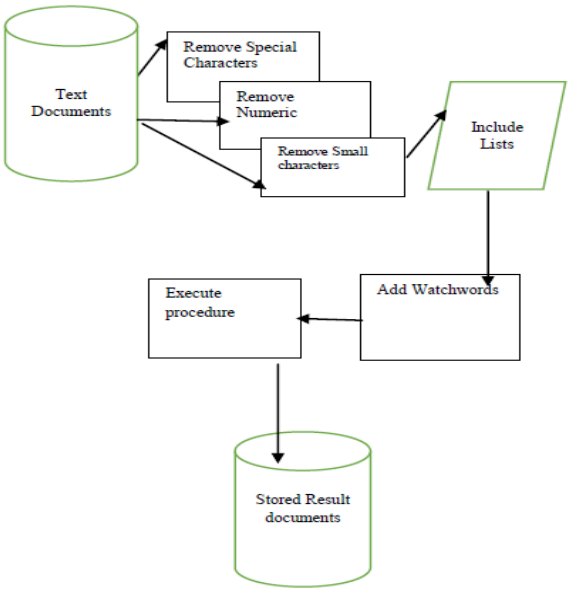

Fig 1. Block diagram- Text Extraction process

C. Substitutes and watchwords:

In text mining process some of the words are given the equal meaning i.e. “beauty” or “loveliness” both words are given equal meaning in the sentence or paragraph. User need to avoid such terms it occupy more space and text mining process it give more time to process the operation.

Pseudocode for Removal for words:

Step 1: Start the word removal process.

Step2: Collect the documents where words need to be removed.

Step3: Create list of stop words

Step4: like “S”, “T” “U” “$” “&”

Stept5: Check in the document list

Step6: int w=0, sum, inspect;

Step7: if inspect=0

Step8: String s= Stop words;

Step9: if(s.equals(str))

Step10: Repeat step 3 and 4 until all documents are done.

Step11:Stop.

D. Stemming algorithms:

After step one in text mining and before creation of indexing stanching can be done. The term “Stemming or stopping “states replacing the bigger information with smaller one. Stemming algorithm used for grouping the words based on the user operations. This is done based on the words used in the text documents. Second this stemming used to retrieve the information from the stored documents based on the user query. In this retrieval process based on the users request information’s are extracted and send back to the user. The process starts when user enter the query to find the document or find the needed content from the stored database or stored documents.

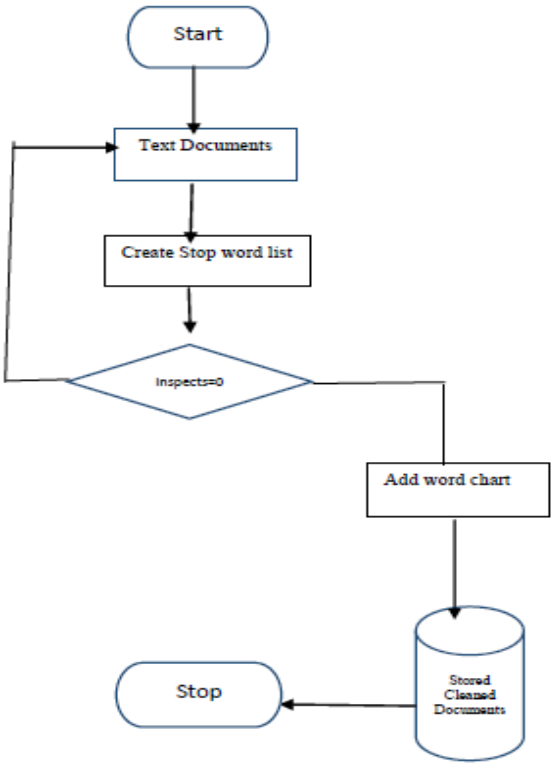

Fig 2. Flow chart for stop words removal

VI. RESULT ANALYSIS

A. Searching the map

In the experimental process a examine implement was constructed. Using this procedure user can define the word to be identified in the documents. Before entering this word before that user need to identify which word or term need to specify in the procedure. Because many documents user created an indexing terminologies i.e. most occurring terms or word in the documents are created by the user. This words are helps to identified or retrieve the document most effectively. This developed tool also helps which word or term most leading incidence in the document. For that most of the text searching technique provide some procedures to the user which word or term need to put on the search bar. User need to follow this procedure carefully to retrieve the need content. This process shown in the fig 1.

Fig 3 Entering permitted word or term in text search bar.

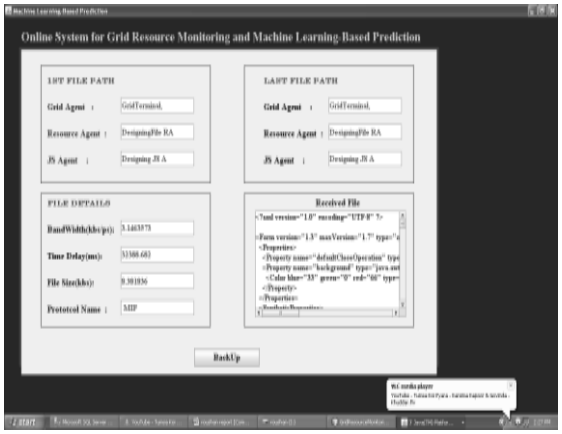

Fig 4. File Uploading

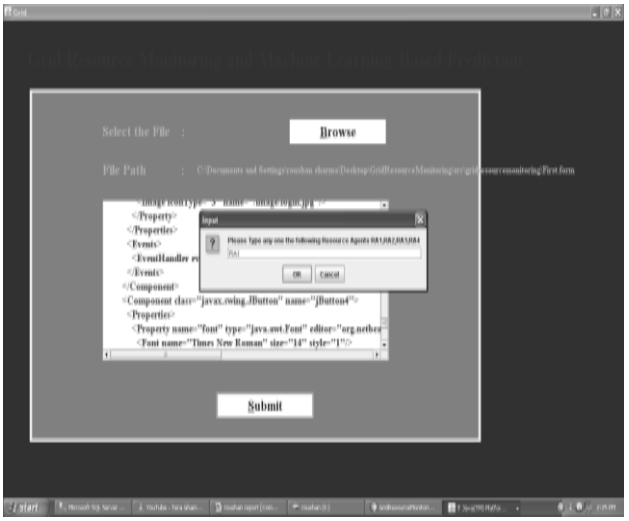

Fig 5. Data Entry (Text Entering)

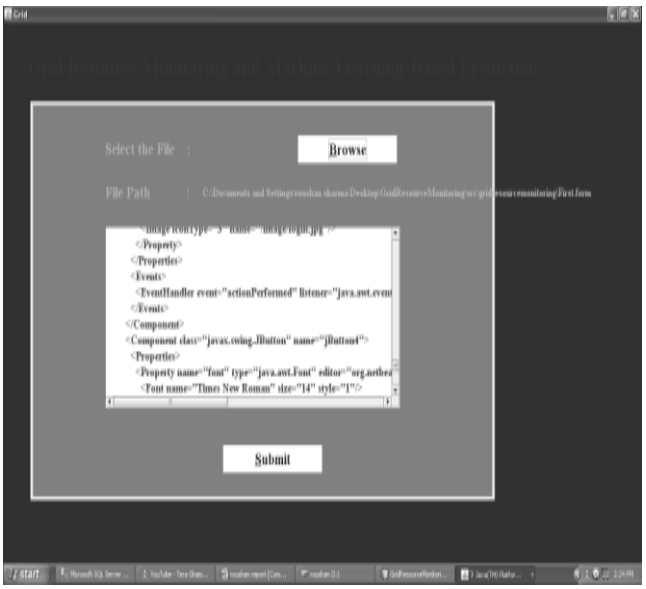

Fig 6. Searching the text information

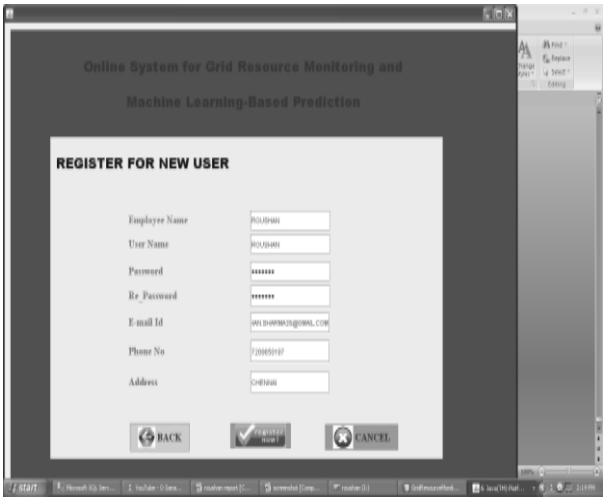

Fig 5. User Registration Screen

VII.CONCLUSION

Traditional text mining and processing methods, mainly from large collections of documents have been used in the detection of similar forms. In the proposed technique the Classical, knowledge extraction processed are using traditional query languages can be difficult to implement, it is provided that can be used on stored content. Corresponding text fields. This paper provides conclusion proposed technique works well compare to the existing technique.

REFERENCES

1. D.Saravanan, ”Clustering of video in formations using BRICH Methodology” Pak journal of biotechnology, Vol 14(Special Issue-II-2017), Pages 377-380, Dec- 2017
2. G. Slaton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing & Management, vol. 24, 1988.pp 513-523.
3. D.Saravanan, Dr.S.Srinivasan, (2013). , Matrix Based Indexing Technique for video data, Journal of computer science, 9(5), 2013, 534-542.
4. D.saravanan, Dr.S.Srinivasan (2012). Video image retrieval using data mining Techniques, Journal of computer applications (JCA), Vol V,Issue 01, 2012. 39-42.
5. 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
6. R. Amarasiri, D. Alahakoon, M. Premaratne, and K.Smith, "HDGSOMr: A High Dimensional Growing Self Organizing Map Using Randomness for Efficient Web and Text Mining", presented at IEEE/ACM/WIC Conference on Web Intelligence (WI) 2005, Paris, France, 2005.pp
7. D.Saravanan, A.Ramesh Kumar, “ContentBased Image Retrieval using Color Histogram”, International journal of computer science and information technology (IJCSIT), Volume 4(2), 2013, Pages 242-245, ISSN: 0975-9646.
8. D.Saravanan “ Effective text data extraction using Hierarchical clustering technique” Int. Journal of Engg., and Advanced Technology(IJEAT),Feb 2019, Vol. 8, Issue 3-S, Pages 741-743(ISSN 2249 8658)June 2019
9. P. Srinivasan and A. K. Sehgal, "Mining MEDLINE for Similar Genes and Similar Drugs", Department of Computer Science, The University of Iowa TR# 03- 02, July 2003.
10. M. A. Hearst, "Untangling Text Data Mining", presented at 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, USA, 1999.pp
11. .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.
12. D.Saravanan, ”Various Assorted Cluster Performance Examination using Vide Data Mining Technique” , Global journal of pure and applied Mathematics , Volume 11, No.6(2015), Nov 2015,ISSN 0973-1768, Pages 4457-4467
13. D D.Saravanan, “Text Information Retrieval using Data mining clustering Technique” International journal of Applied Engineering Research , ISSN:0973-4562, Volume10, No3(2015) ,May2015,Pages 7865-7873
14. 14.D.Saravanan,V.Somasundaram “MATRIX BASED SEQUENTIAL INDEXING TECHNIQUE FOR VIDEO DATA MINING Journal of Theoretical and Applied Information Technology 30th September 2014. Vol. 67 No.3
15. 15. 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

AUTHORS PROFILE

D. Saravanan did his M.E in Computer Science and Engg, and completed his Doctor of Philosophy in the same area. He had 20.05 years of teaching experience. His area of interest is Data mining, Data Base Management systems & Information Retrieval.