International Journal of Advance Research in Science and Engineering
Volume No.06, Issue No. 12, December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

A LITERARY SURVEY ON TEXTMINING APPLICATIONS AND TECHNIQUES

Dr. E. Mary shyla[1], S. Keerthana[2]

[1]*Assistant Professor, Department of Computer Science, Sri Ramakrishna College For Women, Coimbatore*

[2]*Research Scholar, Department of Computer Science, Sri Ramakrishna College For Women, Coimbatore*

ABSTRACT

Text mining plays an important role in research field. It uses the text analysis with supported machine text. The unstructured texts which contains large amount of data information cannot be used for by the computer and knowledge extracted from unstructured text completely uses text mining. The techniques of text mining are retrieval of information, extraction of information with language processing and connect each other with the algorithms and KDD (knowledge discovery data) methods, data mining concepts, machine learning and statistics. Nowadays there are enormous amount of data stored increasing day by day by using the concepts and techniques to extract the useful information. In this paper it is discussed briefly about the text mining applications and the techniques used in the text mining.

Keywords*: *Text mining, Text mining Application, Text mining Techniques.

I.INTRODUCTION

By using the discovering patterns large amount of data are increasing day by day to find the accurate data. Relevant data information is extracted from the huge amount of data by using text mining. To solve this problem various techniques are used for the unstructured text documents to extract the required pattern. Text mining techniques are known as Discovery from Text (KDT).Mining text documents have the some structured and the un structured so that computer is not as much capable to differentiate the patterns as compared to human. Computer can perform the techniques at very high speed and in large volume. Text mining is used to extract the structured data from the unstructured information of data. Function used in text mining summarization, categorization, clustering .In this paper overview of text mining, techniques of text mining, merits, applications and demerits are studied.

II. NEED OF TEXTMINING

Text mining is used to handle the text data. The textual unstructured is very difficult to be manipulated and with the help of the information exchange data mining is used in the business field data. The nontraditional

information retrival which belongs to the text mining .The main goal of the text mining is to obtain the information from the large text documents.

III.DIFFERENCE BETWEEN TEXTMINING AND THE DATAMINING

The text mining and the data mining differs with the source of data. The text mining uses the input as a unstructured file while the data mining uses the input as a structured file. The pattern extracted from the unstructured text is text mining while the data mining uses the structured text. Data mining is that discovery of knowledge with analysis of the data .Text mining is that analysis of data with the discovery of knowledge in data.

IV.APPLICATIONS OF TEXT MINING

Text mining has several applications and the applications are categorized as follows [2]

4.1. Security application

The text mining software packages are used for the security purpose, monitoring the data and the plain text investigation on online references such as internet news, blogs etc... For this application encryption and decryption of the text information is required.

4.2. In software Environment

IBM and Microsoft they study and develop the techniques. Software developed in the data mining techniques is further automated in mining and review processes. The area of exploration and indexing improve the result of the various field. Monitoring the terrorist activities and tracking of the software is concentrated by the public sector.

4.3. In bio-medical field

Bio medical field uses much text mining application. Pub Gene is one of the online text mining applications in biomedical internet service network idea. The navigation tool and exploration in biomedical research analyses is TPX concept.

4.4 .In marketing field

Text mining is used in the marketing field for analysis of the customer relationship management.

4.5. In intellectual

The publishers who have the enormous amount of database for indexing the retrieval of data in the text mining. The scientific methods of information are highly contained in written text. The large number of data is stored and retrieved in the intellectual field with text mining.

4.6 .In Sentiment analysis

The analysis of the movie reviews are estimated by the sentiment analysis. The investigation need the labeled data set for the words .The text is used to detect the effective computing for the detection of emotion. The multiple corpora such as evaluation of students, news stories of the children.

4.7 .In Enterprise Business Intelligence

The business intelligence help the user to store the text for the better decision customer satisfaction and advantage of text mining are essential to gain the aggressive advantages. The data mining gives the deeper approaches in expanding the business intelligence.

V.MERITS OF TEXT MINING

- The data are stored in the emails, memos, feedback. The relevant pattern of the unstructured text is solved by the text mining.
- Storing of the data in the database is not possible due to the size limitation in any organization. Pattern extraction is applied on the text mining.

VI. DEMERITS OF TEXTMINING

- The text in the text mining requires the lot of unstructured text in data collection.
- Ambiguities in the natural text require the human support.
- There is no program for handling the text in text mining to analyses the unstructured text.

VII.TECHNIQUES OF TEXT MINING

In this section discuss about text mining techniques so it will be helpful to work further in the interested area. There are given below:

7.1.Information Extraction

According to the view of Raymond J. Mooney and Un Yong Nahm Department of Computer Sciences, University of Texas, Austin information extraction is that the ease of the text mining is that to discover the knowledge of the structured and the unstructured text. The information extraction (IE) and the knowledge discovery in database (KDD) are used to extract the text.IE is a techniques used to analyze and discover the bulk of text in the data set. The information which is extracted cannot be directly presented into the structured form by the side it needs for the post processing [3].

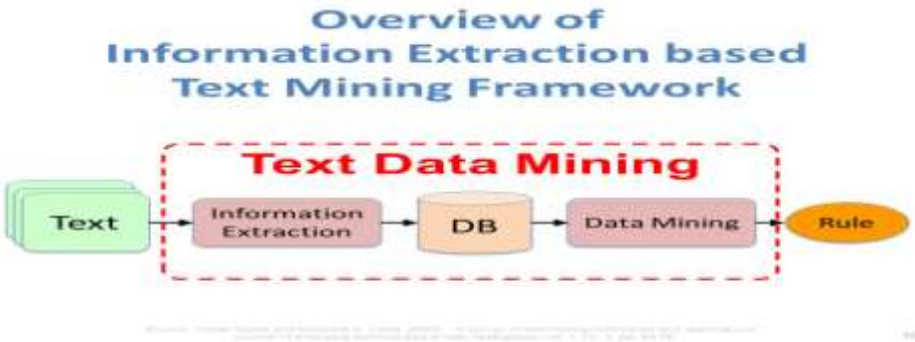

Fig 1 Information extraction frame work

7.2. Summarization

Summarization is also the method of the text mining .Summarization is that creating a summary for the huge information for the objective maintained in the document to be mined. By summarizing the document the user can clearly note whether the document is useful or not. The compression is not in the human readable format.

7.3Topic tracking

Topic tracking helps the user by tracking the topic searched by the user. If the same user search for the documents top .The topic detection deals with the new and the upcoming topics in the time order. This method is used in the news broad cast [4].

7.4. Classification

In this technique the text documents are classified into predefined labels. Classification method has several applications that is customer feedback classification in online, SMS classification in mobile, report classification of a business etc…Topic tracking is therefore used to classify the documents by topic for making the process faster.

7.5. Clustering

In this mining techniques it find out the similarity measures between the dissimilar objects It relates the object in one class and unrelated object in other class. The text clustering algorithms is of different types partitioning algorithms, clustering algorithms.

7.5.1. Hierarchical Clustering algorithms

Hierarchical clustering algorithm is to build the clusters in the hierarchy of clusters .The hierarchy constructed in a top down called divisive and bottom up called agglomerative. In the hierarchical clustering algorithm the documents is split up into clusters and the sub clusters.

7.5.2. K -means clustering algorithm

It is one of the partitioning algorithms in data mining .In this partitioning the n documents with the text data into k clusters .K-means algorithm is that finding the optimal solution for the computationally difficult NP-hard clusters [3].

Disadvantage of the k-means clustering algorithm is that the clustering is very sensitive with number of k. Therefore for using the light weight clustering algorithm agglomerative a brief survey on the classification, clustering and extraction techniques are used .

Algorithm

K-means clustering algorithm

Input: Document set D, similarity measures S, number k in the clusters.

Output: Initialize the k clusters randomly select the k data points as centroid.

Assign the documents to the centroid based on the closest similarities.

Calculate the centroid for all the clusters.

End return k clusters.

7.6. Concept linkage

It is that finding the related documents which share the common concepts .Browsing for the information rather than searching it as an information retrieval. It is used in the biomedical field for concept linkage method to link diseases and treatment.

7.7. Information Visualization

Information visualization in the text mining visual extraction of the patterns .This is also known as the Visual Text mining. It is that data preparation, data analysis data extraction and visualization mapping etc..

7.8. Question Answering

This technique is used to find the best answer for a question. For question answering techniques there are many websites available .This is to provide answers the users question extracting the exact answer for the user.

7.9. Association Rule Mining

The purpose of the association rule mining is that to find the relationship between the data set .In the database records there will be the many number of data records and the value in it.ARM search for the values which are frequently used.ARM check the relationship between two or more variables this process is known as the association rule mining. In this process locate the items purchased by the customer and place the releated item and after the customer purchase the sale automatically increases[5].

7.9.1Application of association rule mining

Different application areas of association rule mining are described below

i. Market Basket Analysis

One of the broadly used application of the association rule mining is market basket analysis In this there will be a large number of record and all items will be brought by the customer at a single purchase or multiple .The managers should know the group of items purchased by the customer . Market basket analysis is the special promotions or sales for combinations of products.

ii.Medical diagnosis

Association rules is for appointing the physicians for curing the patients .Diagnosis is not an easy process for the unreliable tests and noisy data here prediction of correctness is unreliable for the critical medical application .Here association rule mining finds the probability of the illness in diseases and defining the symptoms.

iii.Census data

Census may consists of the huge data with the statistical information .The information relates the population and the economic census which is forecasted on education ,health ,transport and fund .In the public sector business such as shopping mall ,banks ,new factories application of the data mining techniques is used.

7.10. Natural Language Processing (NLP)

The interaction between the human language and the computer language is known as the Natural Language Processing (NLP).It is the translation between the one human language in to the other human language text .It is used in the area of robotics[2] .

7.10.1. Approaches used in NLP

i.Distributional

It is the machine learning and the deep learning process. It includes the large-scale statistical tactics of machine learning and deep learning. In the speech tagging process to find noun or verb .It checks the part of the sentences modify another part. Comparing the words with the other words and result of the different outcome**.**

ii.Frame — Based

The frame data is represented a stereotyped situation explains Marvin Minsky in his seminal 1974 paper called "A Framework for Representing Knowledge."The commercial frame transaction in a frame seller and buyer goods being exchanged, and an exchange price .Example for the frame are" Cynthia visited the bike shop yesterday" and " She bought the bike in cheap" the sentence is incomplete it cannot be adequately analyzed[1].

iii.Model-Theoretical Approach

The category of semantic analysis is under the model-theoretical approach. In this approaches the linguistics concepts are introduced "model theory" and "compositionality". For example to determine the query "What is the largest city in the Europe by population" first you have to find the city and population in the Europe.

iv.Interactive Learning

Paul Grice, a British philosopher of language, a language game between the speaker and the listener .Teach the breadth and depth of the language in the interactive environments where human teach the computer in this approaches the need for the language inform the development**.**

VIII.CONCLUSION

In this article we have given the brief introduction for the text mining field. We have provided the overview of the applications and techniques of the text mining. Overview of the current progresses in the field of text mining is provided. Thus the processing and mining of the large amount of text is of great interest of the researchers.

REFERENCE

[1] Tembhekar Samta, Kanojiya Monika*," Ideas and Innovations in Technology*" *International Journal of Advance Research2014.*

[2] Samta Tembhekar Assistant Professor Department of Computer Technology, Kavikulguru Institute of Technology and Science, *"The Survey Paper on Approaches of Natural Language Processing (NLP)",2014.*

[3] 1 Sathees Kumar B ,2 Karthika R Asst. Professor , M.Phil. Scholar , Department of Computer Science, Bishop Heber College (Autonomous)," *SURVEY ON TEXT MINING PROCESS AND TECHNIQUES" Volume 3 Issue 7, July 2014 2279 ISSN: 2278 – 1323* .

[4] N. Venkata Sailaja Assistant Professor Dept. of C.S.E VNR VignanaJyothi Institute of Engg and Technology, Hyderabad, India*," Survey of Text Mining Techniques, Challenges and their Applications", International Journal of Computer Applications (0975 – 8887) Volume 146 – No.11, July 2016 30.*

[5] 1. K.Thilagavathi 2 . V.shanmuga priya" *A survey paper on text mining techniques" International journal of computer application and robotics ISSN 2320-7345.*