Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 4, April-2017

A BRIEF SURVEY ON TEXT MINING AND ITS VARIOUS TEXT MINING TECHNIQUES

Prachi Patel[1], *Prof. Sneha gaywala*[2]
Department of Information Technology (System and Network Security),
Sardar Vallabhbhai Patel Institute of Technology, Vasad, Gujarat, India-388306

Abstract:- *Now a day's most of the information are stored in unstructuredtext. The pattern disclosure from the unstructured data or text document is a well-known issue in data mining. Unstructured text has tremendous sum data which is not effectively utilized by the computer for processing. With the goal that we require certain procedures to fulfill this assignment for extricating required patterns. Text mining is a method to discover important information from the accessible text documents. Text mining assumes an imperative part of removing helpful designs from unstructured content.It is one of the developing advancements for Knowledge Discovery Process. Record association and design revelation turns into the fundamental errand in data mining. In this paper, we are going to discuss about Text mining & its techniques, applications, advantages and disadvantages.*

Keywords: - Data Mining, Text Mining, Text Mining Process, Techniques, Issues.

I. INTRODUCTION

Data mining is the process to extract information from the relevant data and convert it into understandable structure [1]. The various types in data mining are Web Mining, Text Mining, Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mini ng (SDM), Distributed Data Mining (DDM) and Multimedia Mining.Text mining is a type of data mining which is also called text data mining. In Data mining structured data is used to extract the patterns and in text mining unstructured data is used to extract the patterns. Data mining is mainly used for finding the business data and text mining is mainly used for finding the unstructured data that is difficult to manipulate. Pattern discovery is main task in data mining, so text mining is used for discover the pattern. Text mining is extracting the information or a pattern from the huge amount of sources to find a knowledge [2]. Day-by-day this research is more active for the research purpose. Text mining is basically used for finding the relevant information regarding the various databases.

In that computer has not capable to differentiate the unstructured pattern compare to human. But, the computer can highly process the data in larger volume so that text mining is useful for computer to extract the information that is highly processed. This technique is useful for many algorithm to convert the unstructured data into the useful pattern. The main goal [2] of this research is to reduce the time or efforts for obtaining information from a large amount of unstructured data. Text mining is mainly used for analyze the pattern, extract the information from the unstructured data, identify the pattern.

Most of the organizations and institutions store the data in the text format. For example, students roll no, class, and address is store in the database format. So, text mining is more suitable for institution to store the information in the textual database [3]. Section II describe the process of text mining. Section III describe the techniques of text mining. Section IV describe the measures of text mining. Section V depicts the application of text mining.

II. PROCESS OF TEXT MINING

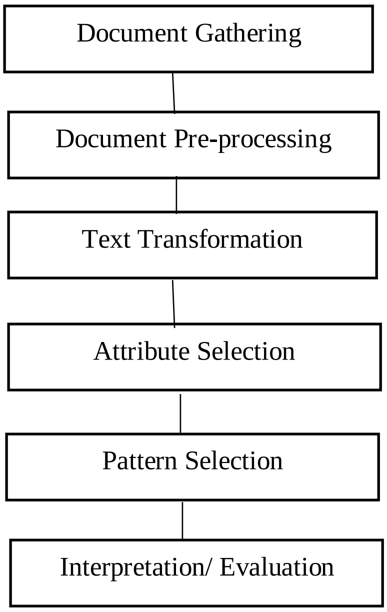

Figure 1. Process of text Mining

Document gathering

In this step, the documents or data are collected from the various resources such as html, css, word, pdf, text, etc [1].

A. Document pre-processing

Text preprocessing is read one text document and processed for Tokenization, Stop Word Removing, Stemming in subtasks [2].

- **Tokenization** the whole sentence is divided into the word by removing comma and spaces.
- **Stop Word Removing** the stop words such as “the”, “a”, “an”, “but”, “of” etc.
- **Stemming** is applied after Stop Word removing that reduce the word into root word. E.g. “working”, “worked” are stemmed to “work”.

B. Text Transformation

Text transformation is convert the text document into the word in large amount of databases.

C. Attribute Selection

Feature selection removes the unrelated features from the data. There are two methods for attribute selection like Filtering and Wrapping methods.

D. Pattern Selection

In pattern selection extracting the pattern from the unstructured data and analyze this pattern. In that useful pattern is discovered from the relevant databases.

E. Evaluation

After discover the pattern this step is to analyze the result based on the pattern and this result is further used for next sequence.

III. TECHNIQUES OF TEXT MINING

Various techniques are available for text mining. The techniques of the text mining is categorized as information extraction, information retrieval, Natural Language Processing (NLP), categorization and clustering. Using this techniques we can mined the text from the large datasets. Fig. 2 shows the different techniques of text mining.

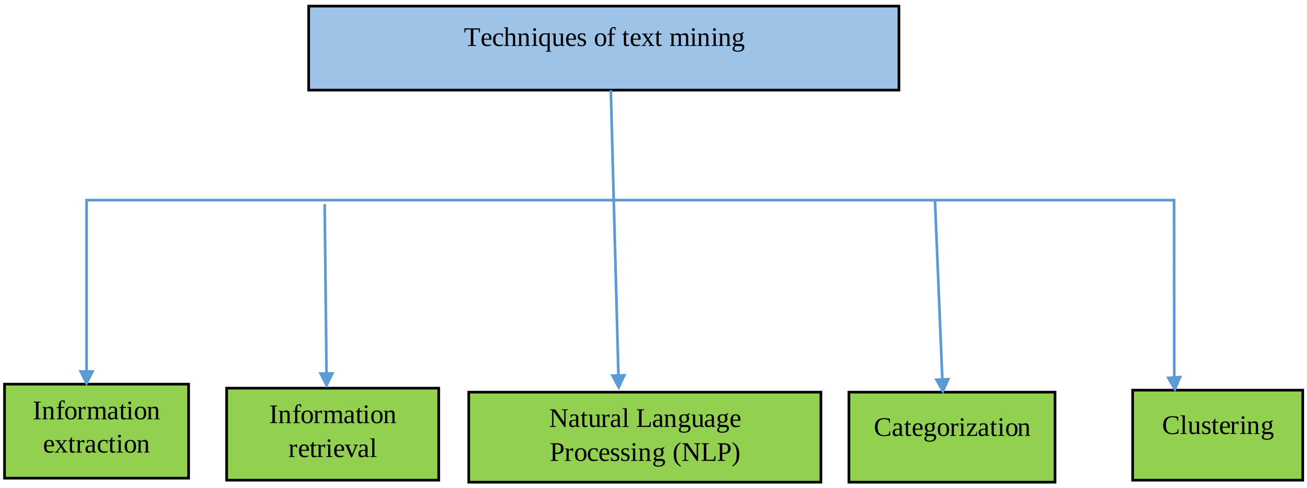

Figure 2. Techniques of text mining

A. Information extraction

It is the process of extracting or access the information regarding the unstructured document. For that it use the Natural Language Text document is used [1]. This technique extract the information from the different sources and structured document is convert into the text. This information is stored in a database like pattern or text. After that various data mining techniques are applied to gain the knowledge.Using this method it identify the term which is useful for finding the information from the document. It identify the organization’s name, place and its id. It also identify the fact from the document. Is used the techniques like domain dictionary, rule selection, lexical analysis, and statistical learning approach to extract the information. In structured data, information is stored in a database in column and rows and it is easy to extract the information like name and id of the students and in unstructured data, the information is difficult to extract [2].

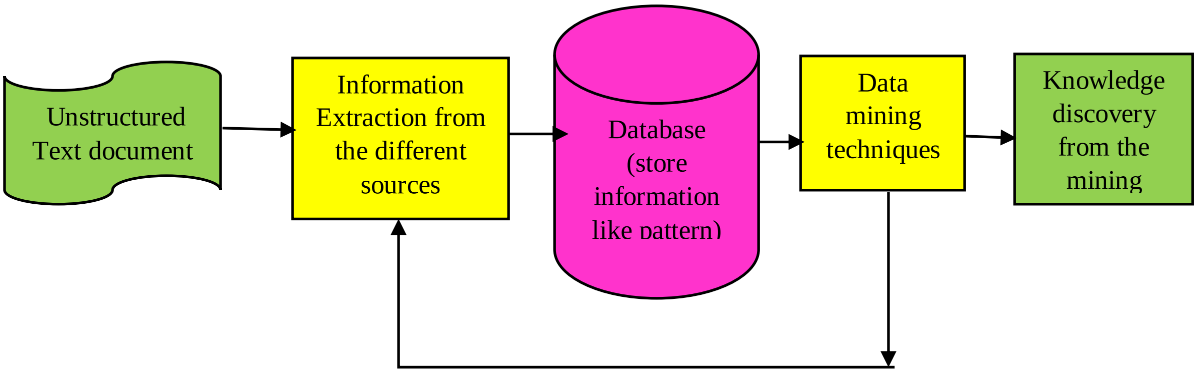
Unstructured Text document
Information Extraction from the different sources
Database (store information like pattern)
Data mining techniques
Knowledge discovery from the mining

Figure 3. Information extraction

B. Information retrieval

This technique is basically retrieve the query and structured data[2]. It is used to identify or retrieve the relevant information from the structured document. In that information retrieval and database is two different things. In that the information of recovery, transaction is not stored in the database. Information retrieval is retrieve the huge query or search the document in a large subset of data. Search Engine Google is the application to find or search the data in very huge manner. It retrieve the document automatically in the document collection. Information retrieval system is used in global library, search engine to search any data. For the effective retrieval various techniques are used such as retrieval algorithm, association rules. Fig. 4. Shows how information retrieve from the document is given.

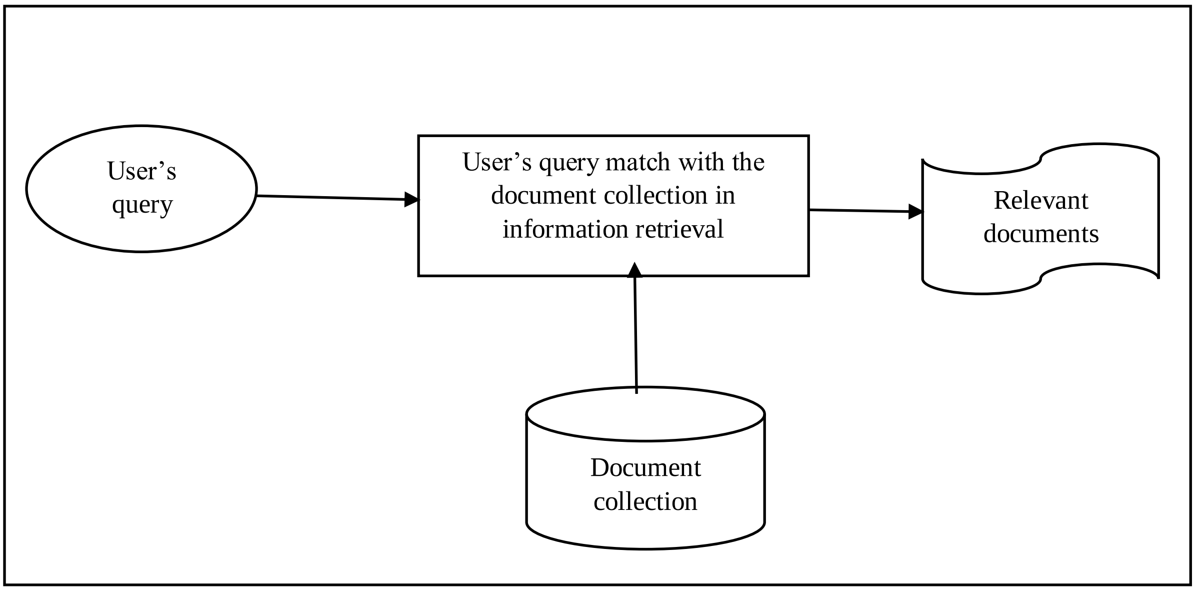
User's query
User's query match with the document collection in information retrieval
Relevant documents
Document collection

Figure 4. Information retrieval

C. Natural Language Processing

NLP mainly used for understanding the human language in computer manner[1]. Computer does not understand the human language because computer only understand the language in terms of bits and to convert the human language into the bits form this technique is used. This technique analyze the human language naturally. NLP is based on machine learning system and it is the major component of artificial intelligence that mainly understand the human in artificially system. Differentiate of word such as noun, verb and adjective this is used. It is also used in probability of the word to convert the natural language. This technique also used for the prediction of the semantic words.

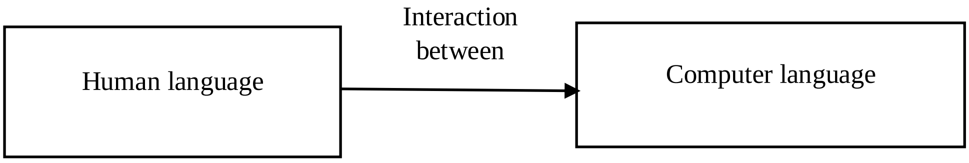

Figure 5. Natural Language Processing

D. Categorization

In categorization all the document is categorized according to subjective and objective manner and classify the object in to a specific group so, that the word or text is easily find from the relevant document object[1]. It recognize and analyze the document of object and mined the data from the databases. It is based on the prediction of the things. It involves the relevant text or word of a document by placing it into correct related topics. It only count the important word in the document so that entire document is covered. It is useful for many applications like in business and organizations to categorize the employees, workers and all. The main goal of this technique is to categorize the whole document into specific topic. To categorize the document it have many technique like KNN (K-Nearest Neighbor), Naïve Bayesian, SVM (Support Vector Machine).

E. Clustering

Clustering is the process of partition the data into set of relevant group[2]. This technique partition the similar type of data in a group and also the functionality of this data is same than it is place into the relevant group. This technique useful for place the data or object in correct place and using that it can find the data very efficiently because in that computer cannot analyze the whole document but only analyze the relevant subgroups based on their functionality. So, the important data is mined easily and very effectively. This technique also used when the data is too large and we cannot find the knowledge so this technique efficiently partition the data in subgroup. So, we can easily find the data from the large document. There are mainly two methods for clustering: Hierarchical Methods and Non-Hierarchical Methods. In hierarchical Methods the cluster is generated in hierarchy in tree format which contain the child and parents cluster. Using child and parents this cluster is partitioned. In Non-hierarchical Methods divide a whole data into number of clusters and it is mutually clustering the whole data and taking into the account.

IV. MEASURES OF TEXT RETRIEVAL

The document is basically denoted by mainly two types: relevant document and retrieved document [6]. A set of relevant data is called as relevant document and set of retrieved data is called as Retrieved document. This is basically how the document is find to extract the information for that this to domain are used. In such conditions, when those document contain relevant as well as retrieved is denoted by {relevant ∩ retrieved}. To retrieve the text there are two measures for that precision and recall. Precision is the percentage of the retrieved data in a particular document and recall is nothing but the percentage of relevant data in a particular document.

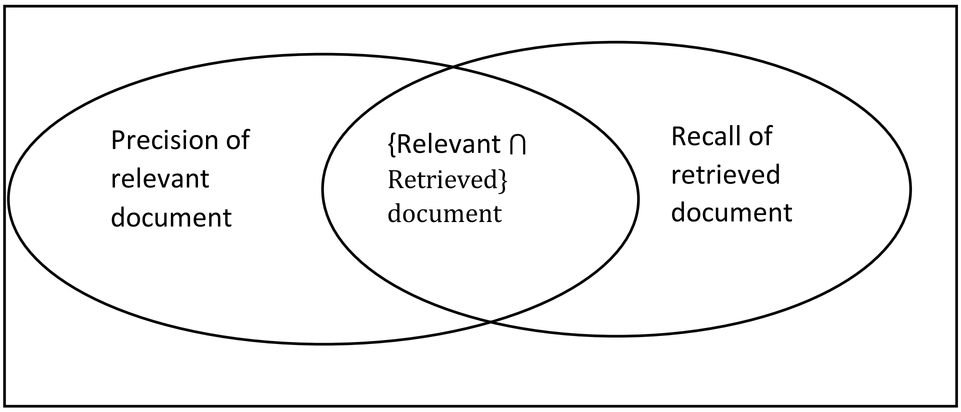
Precision of relevant document
{Relevant ∩ Retrieved} document
Recall of retrieved document

Figure 6. Measures of text retrieval

V. APPLICATION OF TEXT MINING

This technology is active research area in data mining because text mining is basically used for extracting the information from the unstructured data. Applications of text mining is given below[3]:

- In medical application to analyze the medical field information to gain the meaningful knowledge.
- To monitor the news for the security purposes.
- In organizations also used for extracting the reports and activities to gain the useful information related to the other company or organization.
- Text mining is more useful for commercial environment.
- For market analysis it is more useful for analyzing and monitoring the competitors and also the opinion of the customer it is required.
- It maintain customer relation as per the request of customer and then it solve the problem of customer very effectively.

VI. ADVANTAGES AND DISADVANTAGES OF TEXT MINING

A. Advantages

- Text mining is useful for finding the knowledge or pattern from the large amount of data such as email, blogs etc[2].
- It is extract the data from the unstructured text that computer easily understand.

- In organization there is lots of records or data of its employee and lots of information of organization are there so, this information is not stored in the database because it is too much. So for that purpose text mining is useful for stored the text that we can easily find it.

B. Disadvantages

- Text mining is too complex data mining types because in document one word has lots of meanings in different ways so that the whole sentence because of wrong word is difficult to predict[2].
- It required the human to know about the natural text information.
- There is no one program for analyzing the text document that is unstructured text.
- To handle the unstructured text there is need of data collection.

VII. CONCLUSION

Data Mining is the vital and additionally dynamic research zone separates supportive Patterns from the information. These patterns produced are used for decision making in enterprises. Text mining is a tecnique that handles unstructured or semi structured information.Text mining technique is used for extracting pattern from unstructured data. So in this paper, our focus is basically on how text is to be mined. In addition to that we have outlined the different text mining techniques such as Information Extraction, Information retrieval, Natural Language processing, Categorization and Clustering.And furthermore we have discuss about text mining processing flow, applications of text mining and positive and negative points of text mining.Mining text in diverse dialects might be a noteworthy issue, since text mining tools and techniques should have the capacity to work with a few dialects and multilingual dialects. Integrating a domain knowledge base with text mining engine would increase its efficiency, especially within the information retrieval and information extraction phase.

REFERENCES

[1] R.Janani, Dr. S.Vijayarani, “Text Mining Research: A Survey”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2016.

[2] Amrut M. Jadhav, Devendra P. Gadekar, ” A Survey on Text Mining and Its Techniques”, International Journal of Science and Research (IJSR), Volume 3 Issue 11, November 2014.

[3] C.Uma,S.Krithika,C.Kalaivani, “A Survey Paper on Text Mining Techniques”, International Journal of Engineering Trends and Technology (IJETT) – Volume-40 Number-4 - October 2016.

[4] Falguni N. Patel, Neha R. Soni,”Text mining: A Brief survey”, International Journal of Advanced Computer Research Volume-2 Number-4 Issue-6 December-2012

[5] Andreas Hotho, Andreas N¨urnberger, Gerhard Paaß, “A Brief Survey of Text Mining”

[6] K.Thilagavathi, Mca,2 V.Shanmuga Priya, ” A Survey On Text Mining Techniques”, International Journal Of Research In Computer Applications And Robotics, Vol.2 Issue.10, Pg.: 41-50 October 2014.

[7] https://en.wikipedia.org/wiki/Text_mining

[8] Vijaykumar Ganpatrao Ingawale, Prof. Sunil Damodar Rathod,” A SURVEY PAPER FOR TEXT MINING OF IMPORTANT TERM FROM RELEVANCE DOCUMENT USING PATTERN BASED MODEL”, Multidisciplinary Journal of Research in Engineering and Technology, Volume 2, Issue 4,Pg.788-793.

[9] Falguni N. Patel, ” Large High Dimensional Data Handling Using Data Reduction” , International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.

[10] Dr. G. Rasitha Banu, VK Chitra, “A Survey of Text Mining Concepts”, International journal of innovations in engineering and technology (IjIET), volume , issue2, April 2015.

[11] Vishal Gupta, Gurpreet S. Lehal, ” A Survey of Text Mining Techniques and Applications”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009.