Text Mining Techniques: A Comprehensive Review

Lalima Choudhary
MSIT, MATS University- Raipur (C.G.)

Abstract: **Rapid progress in digital data acquisition techniques have led to huge volume of data. Now a day's communication and interaction between people, leads to communal learning and sharing of information has been increased through chat, messaging, social networking websites, and search engines. Today 80 percent of data is collection of semi-structured or unstructured data. The innovation of suitable replica and tendency to observe the text documents from enormous amount of data is a big challenge and dilemma. Text mining is a method of extracting interesting and non –trivial patterns from large amount of text documents. There are different methods and tools to extract the text and discover important information for future forecasting and decision making process. Selection of appropriate text mining technique and process helps to increase the speed of extraction and decrease the effort and time required. This paper briefly discuss on text mining techniques and its application in diverse field.**

Keywords: **Retrieval, Extraction, Categorization, Clustering, Summarization**

1.INTRODUCTION:

The most common way of formal information exchange is text and the extracting useful information from text is not easy task. The volume of information has been vastly increasing day by day, so retrieving knowledge and discovering patterns has become a great challenge. Today we need a business intelligent tool for extraction of useful information quickly and in low cost. The most important technique is applications of data mining are text mining and web mining. . In this paper, a discussion over various text mining techniques which are used to solve the problem of text mining [1].

Text mining is new and important research areas that aim to take the dispute and construct the intelligence tool. The tool is a text mining scheme which has the ability to examine huge amount of natural language text and identify lexical and linguistic usage patterns in an effort to extract significant and useful information. The aspire of text mining tools is to be able to answer complicated questions and reach text searches with the constituent of intelligence these information are then stored in text database format which contains structured and few unstructured fields.

Text can be to be found in mails, chats, SMS, newspaper articles, journals, product reviews, and organization records[2].

Text mining is a multidisciplinary field, pertaining to retrieval of information, examination of text, extraction of information, categorization, clustering, visualization, mining of data, and machine learning[2]. Text mining, also known as Intelligent Text Analysis, Knowledge-Discovery in Text (KDT), refers commonly to the process of extracting interesting and non-trivial information and knowledge from unstructured text.

II. TEXT MINING PROCESS

A. Document Gathering

This is the first step in text mining, in this step text documents are gathered which are available in different formats. Such as Word document, html document, cascading style sheet, etc.

B. Pre- Processing of document:

This is the second step in text mining, in this step, the given input document is processed to eliminate redundancies, inconsistencies, separate words, stemming and documents are organized for next step, the phase performed are as follows:

1) Tokenization

The particular document is measured as a string and recognizes single word in document i.e. the selected document string is separated into one unit or token[2].

2) Elimination of end word

The words like a, an, but, and, of, the etc. which are common are eliminated.

3) Stemming:

A stem is a normal group of words with similar meaning. This process explains the foundation of particular word. The two types of method used are called Inflectional and derivational. porter's algorithm is used for stemming[2].

III. TEXT TRANSFORMATION:

A text document is group of words (feature) and their rate. There are two important behavior for demonstration of documents are Vector Space Model and Bag of words[2].

→

A. Attribute Selection:

Outcome in giving low database space, minimal search technique by captivating out unnecessary attribute from input document. Filtering and wrapping of methods are used in attribute selection[3].

B. Data mining/Pattern Selection:

In this stage in the conventional data mining method combines with text mining process. Structured database uses classic data mining technique that resulted from previous stage.

C. Evaluate:

This is the last stage in which result is analyzed and the result can be used further.
The figure below shows the text mining process.

IV. RECENT STUDIES

In this section of paper survey of current efforts and assistance are analyzed. Several data mining technique have been designed for mining information in text documents[2]. But how much are successfully used and updated is still research issue.

Table:1 Types of data available and generated by various sectors

S no	Sector	Video	Image	Audio	Text
1	Insurance	□	□	□	■
2	Banking	▨	▨	□	■
3	Process manufacturing	▨	▨	□	■
4	Discrete Manufacturing	▨	▨	□	■
5	Retail	▨	□	□	■
6	Wholesale	□	□	□	■
7	Health care	▨	■	□	■
8	Transportation	▨	▨	□	■
9	Health care	□	▨	□	■
10	Communication and media	■	▨	■	■
11	Education	■	▨	■	□
12	Government	■	▨	■	■
13	Construction	□	▨	□	□
14	Utilities	▨	▨	□	■
15	Securities and Investment services	□	□	□	■

Penetration high ■ Medium ▨ Low □

A. Text Mining Techniques

1) Information Extraction:

Information extraction is an primary step to examine unstructured text. This process is simplification of text which recognizes phrases and detects the associations between them. This technique is useful for huge texts. Information extraction (IE) is the mission of automatically extracting structured exact information from unstructured or semi-structured natural language. The documents are first changed into the structured databases on which data mining techniques can be applied to mine knowledge or interesting patterns. The task is to identify entities; the result would be a pattern in which all the entities and their relationships with one another can be easily identified and information is entered into the database to apply data mining techniques can be applied in order to find some implicit information[5]. It is simple method for information extraction, but the complexity depends on source text. The extracted data can be given to KDD module for promotion of mined text.

2) Categorization:

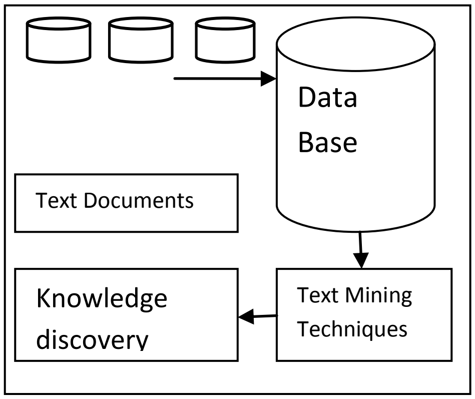

Fig.1 : Process of Information Extraction Technique

It is a organize technique. This technique is based in which the set of input output pattern are basically used to prepare the model being used and arrange to categorize the new documents. Text categorization is the task of natural language documents to evaluate categories according to their content[10].

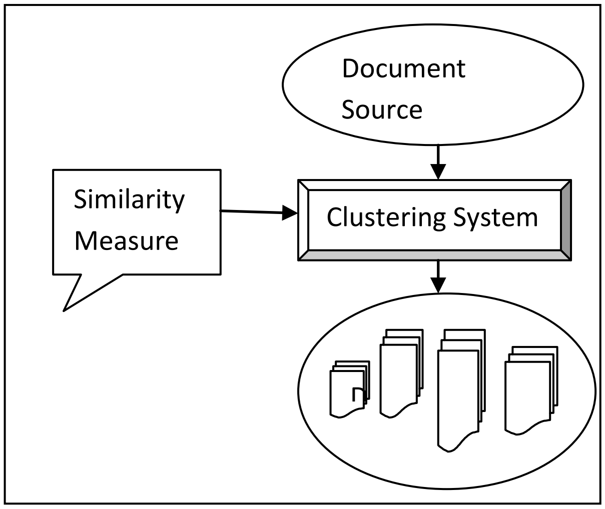

Fig.2 : Process of Information Extraction Technique

This is a process of resulting main matter of document by addition of metadata and analyzing document. This method hit upon calculation of words and from that count choose topic of the document. In this process, text documents are classified into predefined class tag[6]. The variety of classification techniques can be applied to categorize the text. It is used in feedback of customers, filtering emails, etc.

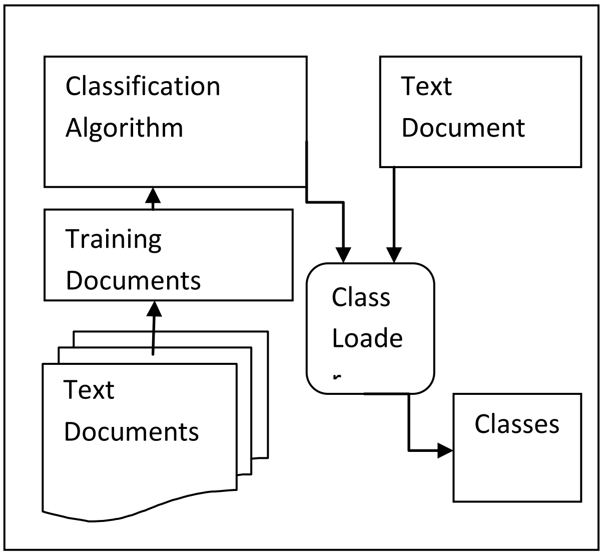

Fig.3 : Process of Information Extraction Technique - 2

3) Text Clustering:

Clustering is a method of forming groups (class) of similar objects from a set of inputs and dissimilar objects in another different class. By this method it creates a group of clusters. It is different from categorization. Words are separated very fast then load are allocated to each word. After calculating relationship and similarity, algorithms for clustering are introduced to create directory of classes. The plan of clustering derived from statistics where it was useful in numerical data.

Clustering is divided into two categories hierarchical and non hierarchical. The non- hierarchical methods divide a dataset of N objects into M clusters, with or without overlie. These practice are divided into partitioning method, in which the classes are mutually exclusive, and the less common clumping methods, in which overlap is allowed[4].

Every object is a member of the cluster with which it is most similar; however the threshold of match has to be defined. The hierarchical methods creates a set of nested clusters in which every pair of objects or clusters is gradually nested in a superior cluster until only one cluster remains. The hierarchical methods can be further divided into agglomerative or divisive methods.

4) Topic Tracking:

A topic tracking system is a method for supervision of user profile, which checks the contents viewed by the user and studies their profiles. According to the users view it predicts for other document related to users interest. In topic tracking applied by Yahoo, user can give a keyword and anything related to keyword pops up then it will be informed to user.

This can also be applied for unstructured data. For example- if we select competitors name then if anytime their name will come up in the news then this information will be passed to company[7]. Topic tracking method has its own restrictions, however. For example, if a user locates an observant for "web mining", he or she will be given numerous reports and information on web mining. Some of the enhanced text mining tools allows users to choose let users select specific group of importance or the software regularly can even infer the user's importance on his or her searching history[9].

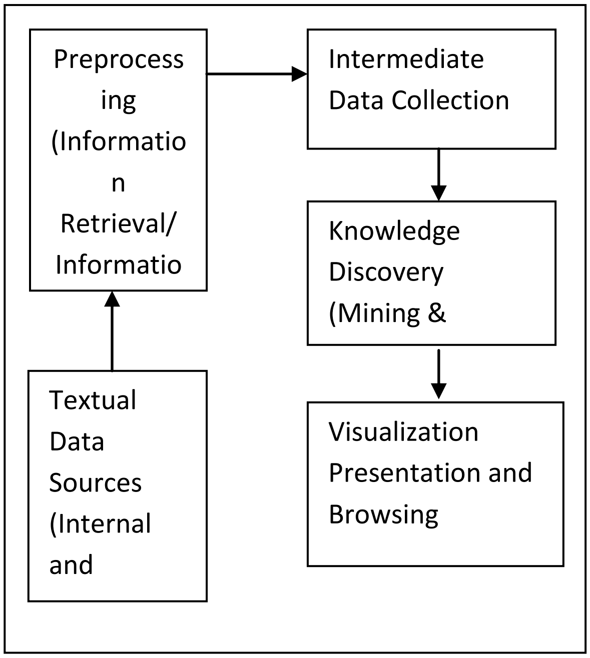
Preprocessing (Information Retrieval/ Informatio
Intermediate Data Collection
Knowledge Discovery (Mining &
Textual Data Sources (Internal and
Visualization Presentation and Browsing

Fig.2: Process of Topic Tracking Technique

5) Visual Text Mining: This process inserts huge textual sources into visual ladder and supply browsing facility with easy searching[8]. This is useful when we have Information visualization is useful when a user needs huge range of documents and survey related topics.

V.CONCLUSION: Text mining is a process of extracting new patterns and knowledge from unstructured, structured and semi- structured documents. In this paper we have described important text mining techniques which are the needs of today's various fields. In near future the proposed technique is implemented using JAVA technology and the comparative results are provided.

REFERENCES

[1] P. Monali , K. Sandip, "A Concise Survey on Text Data Mining" in proceeding of the International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 9, September 2014, pp 8040- 8043

[2] S. Jusoh& H. Alfawareh," Techniques, Applications and Challenging Issue in Text Mining", International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012 ISSN (Online): 1694-0814

[3] Ramanathan V, Meyyappan T, (2013), Survey of Text Mining, International Conference on Technology and Business Management.

[4] Divya Nasa, “Text Mining Techniques- A Survey ”, International Journal of Advanced Research in Computer Science and Software Engineering , ISSN: 2277 128X Volume 2, Issue 4, April 2012 pp 51-540, in IJARCSSE

[5] Lokesh Kumar and ParulKalra Bhatia, “ Text Mining: Concept, Process, Applications” Journal of Global Research in Computer Science Volume 4, No. 3, March 2013.

[6]]Falguni N. Patel, Neha R. Soni,” Text mining: A Brief survey”, International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012

[7] V. Gupta, G.S. Lehal ― A Survey of Text Mining Techniques and applications ―, Journal of Emerging Technologies in Web Intelligence,2009.

[8]http://www.planetdata.eu/sites/default/files/presentations/Big_Data _Tutorial_part4.pdf