A Study of Text Mining Techniques: Applications and Issues

Dr. Noopur Goel
Assistant Professor
Department of Compute Application V.B.S.P.U. Jaunpur

Abstract

Data Mining is the method of retrieving meaningful information from the ocean of data. The data are in the form of text, audio, video and images. Obtaining information from these data is not an easy task. It requires different techniques to extract information. Text mining is one of them. Text mining is the process of extracting information, pattern or knowledge from different text documents available on different resources. Every day million bytes data are added in exiting data. Most of data stored in text documents which are unstructured data and cannot be used for any processing to extract useful information. So different techniques such as classification, clustering and information extraction are applied for this purpose. There is a number of text categorization techniques are developed. Some of them are based on supervised and some of them unsupervised manner of document arrangement. In this paper focus is based on Text Mining, different text mining techniques and its application.

Keywords—Classification, information retrieval, information extraction, text categorization, Knowledge Discovery, Applications.

1. INTRODUCTION

The data size in the computer world increases by exponential rates day by day. Every day million Megabytes data are added in the exiting data. Almost all types of institutions, organizations and commercial industries store their data in electronically digital form. A standard amount of text circulates on the Internet in the form of digital libraries, repositories and other textual information such as blogs, social networks and e-mails. It is a very difficult task to determining the trends and patterns appropriate to extract appropriate valuable knowledge from this large volume of data. Traditional data mining tools cannot handle text data because it takes time and effort to extract information.

Text Mining is a process to extract meaningful and interesting models for exploring knowledge from textual data sources. Text Mining is a multidisciplinary field that is based on Data Mining, Information Retrieval, Machine Learning, Statistics, and Computational Linguistics.

Several text mining techniques such as summarization, classification, clustering, etc., can be applied to extract knowledge. Text

extraction processes the text in natural language that is stored in semi-structured and unstructured format. The Text Exploring techniques are continuously applied in the industry, the university, the web applications, internet and other fields. Application area like search engines, customer relationship management system, filter emails, analysis of product suggestions, detection of fraud and social media analysis uses text mining for the exploration of opinion, characteristics extraction, feeling, predictive analysis and trend.

Figure1 : General process of Text Mining

Generic process of text mining performs the following steps :

1. Unstructured data collection from different sources, available in different file formats such as plain text, web pages, pdf files, etc.

2. Pre-processing and cleaning operations are performed to detect and remove anomalies. Cleaning process be sure to capture the real essence of the available text and is performed to delete the stop words (process of identifying the root of certain words) and index the data.

3. Processing and controlling operations are applied to audit and further clean the data set by automatic processing.

4. Pattern analysis is implemented by Management Information System (MIS).

5. Information processed in the above steps is used to extract valuable and relevant information for effective and timely decision making and trend analysis.

Extraction of valuable information from a corpus of different document is a tedious and tiresome task. The selection of appropriate technique for mining text reduces the time and effort to find the relevant patterns for analysis and decision making. The objective of this paper is to analyze different text mining techniques which help to perform text analytics effectively and efficiently from large amount of data. Moreover, the issues that arise during text mining process are identified.

2. DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

In general sense data mining and text mining are often used interchangeably to explain how data is processed or extracted to get information. Data mining is more general term in comparison to text mining. It refers all techniques such as Association, classification, clustering etc. Which are applied for extraction of data and text mining is one of these techniques. Technically data mining and text mining are different things. Data mining is the process of extracting information from structured data which are stored in forms of tables, spreadsheets, database files. The data stored in these files are generally in numbers or able to easily transform into numbers. Data mining analyses these billions of numbers and extract statistical data and emerging trends or patterns. This kind of analysis has been successfully applied in business situation as well as for military, social and government needs.

Text mining is the process to extract information/patterns from text files which are unstructured or semi- structured in nature. Approximate 80% digital data are stored in text format. Text mining technique is used as information retrieval system which helps the user to get the result of their queries. Actually text mining is the extension of data mining. Text mining uses Natural language processing and Machine learning to automatically collect statistical and meaningful information. Text Mining is a comprehensive technique. It relates to data mining, computer language, and information searching and knowledge management.

3. TEXT MINING TECHNIQUES

Text mining is interdisciplinary techniques which include a number of techniques such as Information retrieval, Information Extraction, text classification etc.. Different text mining techniques are applied to analyze text patterns and their extraction process.

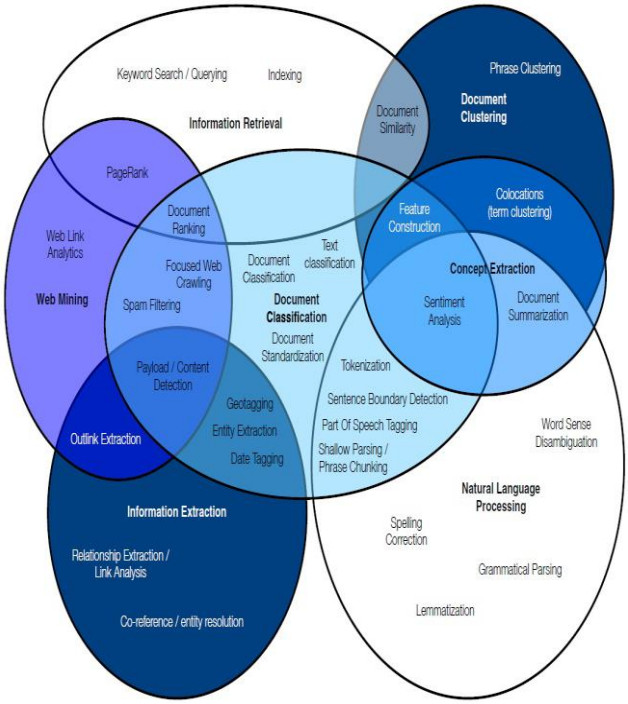

Figure-2: The Venn diagram for interrelation among text mining techniques and their kernel functionality

Classification of documents (classification of text, document standardization), information search (search by keyword / querying and

indexing), grouping of documents (grouping of sentences), natural language processing (spell check, lemmatization, grammatical analysis, and disambiguation of the meaning of the word), information extraction (relation extraction / link analysis), and website exploration (web link analysis) .

A. **Information Extraction:** Information extraction is applied to automatically extract useful information from unstructured and semi structured text documents such as PDFs, web pages and text files. An IE system involves identifying entities such as names of people, companies and location, attributes and relationship between entities.IE can be describe as the creation of a structured representation of selected information drawn from texts.

In IE natural language texts are mapped to be predefine, structured representation, or

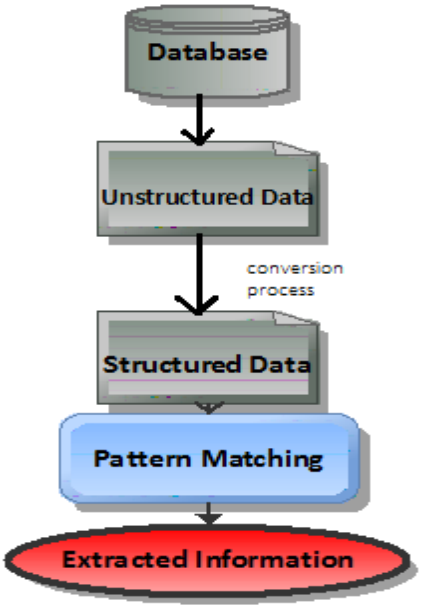

Figure-3 : Steps of information Extraction

templates, which, when it is filled, represent an extract of key information from the original text. The goal is to find specific data or information in natural language texts. Therefore the IE task is defined by its input and its extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web. It does this by pattern recognition. It is the process of searching for predefined textual documents into a more structured database, the database constructed by an IE module can be provided to KDD module for further mining of knowledge. Field experts specify the attributes and relationship according to the domain.

B. **Information Retrieval:** Information retrieval is a process that returns the information that is relevant for a specific query or field of interest. Note that this information could also be in the form of general documents. There is a close relationship in text mining and information retrieval for textual data. In IR systems, different algorithms are used to track the user's behavior and search relevant data accordingly. Google and Yahoo search engines are using information retrieval system more frequently to extract relevant documents according to a phrase on Web. These search engines use query based algorithms to track the trends and attain more significant results. These search engines provide the user more relevant and appropriate information that satisfy them according to their needs.

C. **Natural Language Processing:** Natural Language Processing is a challenging problem in the field of Text mining that uses the concept of artificial intelligence to tackle it. NLP is the study of human language so that computers can understand natural languages similar to that of humans. By NLP a computer system is able to remove text mining ambiguities such as homonymy, polysemy, synonymy and hyponymy. NLP also recognize similar concepts – even if they've been expressed in very different ways. For example, the same word may be spelt differently (hemophilia/haemophilia, tumor/tumour), the same word may be realized differently in different contexts (tumor/tumors, suffers/suffered), the same concept may be expressed by different words entirely (Tylenol / Acetaminophen, heart attack / myocardial infarction).

Natural Language Processing is concerned with Natural Language Generation (NLG) and Natural Language Understanding (NLU). NLG makes sure that generated text is grammatically correct and fluent. Most NLG systems include a syntactic realizer to ensure that grammatical rules such as subject verb agreement are obeyed and a text planner to decide how to arrange sentences, paragraph and other parts coherently. The best Known NLG application is machine translation. The system analyzes texts from a source language into grammatical or conceptual representations and then generates corresponding texts in the target language.

NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. NLU consists of at least of one the following components; a.tokenizer, lexical analyzer, syntax analyzer and semantic analyzer.

- In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark.
- Morphological or lexical analysis is a process where each word is tagged with its part of speech. The complexity arises in this process when it is possible to tag a word with more than one part of speech.
- Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence. It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used.
- Semantic analysis is a process of translating a syntactic structure of a sentence into a semantic representation that is precise and unambiguous representation of the meaning expressed by the sentence. A semantic representation allows a system to perform an appropriate task in its application domain. The semantic representation is in a formally specified language.

.

D. **Document Clustering :** Document clustering is an unsupervised process for classifying text documents in groups by applying

different clustering algorithms. In a cluster, similar terms or patterns extracted from various text documents are grouped together. Grouping is done downward and upward. Different clustering techniques are hierarchical, distribution, density, centroid and k-mean.

E. **Text Summarization:** Text summarization is a mining tool which helps to check whether a document fulfills the user's need and is important to get more relevant information. Text summarization software processes the documents and summarizes it by reducing the length and detail of a document and maintains its key points and overall meaning. To summarize a text, a human first read the entire content and develop a full understanding and then he write a summary highlighting its key points. Since computer does not have the human's language capability, so it uses alternative methods.

One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization. For example, summarization tools may extract the sentences which follow the key phrase "in conclusion", after which typically lie the main points of the document. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document.

F. **Document classification:** Document/Text classification is one of the important and typical tasks in *supervised* machine learning (ML) which place places the document according to content. Text classification of documents such as web pages, library book, media articles gallery etc. has many application like spam filtering, email routing, word sense disambiguation, sentiment analysis etc.. It assign one or more classes to a document according to their content. Classes are selected a hierarchy of categories or classes. For automatic classification, the major text classification processes are extraction of text, tokenization, removal of stop words and lemmatization.

G. **Web mining:** Estimating the size of World Wide Web is very difficult today. In the last decade, its estimated size was about 350 million pages, which was consistently increasing with the rate of one million pages per day. These web pages are repository of data but it is very difficult to extract information from them, because these web pages are actually text documents. Web mining is the mining of data related to the World Wide Web. Web mining is the activity of identifying patterns implied in large document collection. Web mining is an integrated technology in which several research fields are involved, such as data mining, computational linguistics, statistics, and informatics and so on. Web mining examines the content of web pages as well as results of web searching. These contents includes text and graphical data.

Web mining is derived from Data Mining but it has some unique features in comparison to Data Mining .First its source is web data which includes web pages, intrapage structure includes HTML or XML code, interpage structure – the link between Web pages and user profiles. Web mining is to find interesting information and potential patterns from the contents of web page, the information of accessing the web page linkages and resources of e-commerce by using techniques of data mining, which can help people extract knowledge, improve web sites design, and develop ecommerce better.

4. APPLICATION OF TEXT MINING

Different organizations generate a large volume of unstructured data through their day by day transactions. These unstructured or semi structured data contains valuable information that increase their future aspect. But they are unable to get information directly from these text documents which are in form of web pages, text files, e-mails etc. Different Text mining techniques are applicable here to get this valuable information and convert them in structured data. The major areas where text mining applications are used are banks, IT sector, research, energy, media, political analysis, healthcare etc. Text mining applications are:

- **Detection of junk Emails**: Unwanted or unsolicited materials which are sending as email by an organization for advertising or promotional purpose are called junk E-mails. Text Mining techniques are applied to detect unwanted junk e-mails automatically using classification techniques.

- **Management of Human Resources**: Text mining can also be used to manage human resources. For example, analyzing employee opinions is the best use of office mining. In addition to storing new CVs, monitoring company progress and monitoring satisfaction levels of employees are also an important task that can be accomplished by mining technology.

- **Competitive Intelligence**: - Competitive Intelligence is a process of collecting all possible information about market trends and other competitors so that by analyzing this data specific patterns and current requirements can be developed which will further contribute in the company's strategies.

- **Customer Relationship Management**: - It's the duty of CRM to provide quick response to any client's query or message. By using text analysis these messages/queries are diverted to the appropriate person or service for further process.

- **Classification of NEWS as Text**: - Text mining techniques are applied to search a specific news headlines (web pages, PDfs) based

on some terms. Manually doing this job is hectic and time consuming process. Google News service is an example of it.

❖ **Multilingual Applications of Natural Language Processing**: - Utilization of text mining techniques to analyze different web pages in a different language is a classic example of multilingualism. There are other applications also like speech recognition system.

❖ **Classification of Scientific Documents**: - Text mining techniques such as Classification, clustering, regression analysis are applied by scientist to find articles related to their interest. Replacement of keyword queries with the structured automated process by topic scoring engine reduced research time and improved results quality. By using regression analysis, classification rules are generated to cope up with this problem.

❖ **Sentiment Classification**: - Sentiment Analysis is a case of natural language processing which could mark the mood of the people about any specific product by analysis and classifying it as positive, negative or neutral. Sentiment Analysis is a process of automatic extraction of features by mode of notions of others about specific product, services or experience. The Sentiment Analysis tool is to function on a series of expressions for a given item based on the quality and features. Sentiment analysis is also called Opinion mining due to the significant volume of opinion.

5. ISSUES IN TEXT MINING

Text mining is applied on text document and most text documents are in natural language which is a big issue for extraction. The natural language is not free from the ambiguity problem. Ambiguity means the situation when a word or group of words or sentence has two or more possible senses or ways. In natural language, ambiguity provides flexibility and usability and so it cannot be removed totally. Multiple meanings can be derived from a single words or a phrase or sentence can be interpreted in different ways.

To resolve ambiguity problem from natural language, a number of researches have been conducted but it provide very less success. These researches have been dedicated for a specific domain. On the other hand, most of the IE systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities.

6. CONCLUSION

Today a huge volume of digital data is available in computer world and most of them in textual form. To extract information from this unstructured document text mining techniques are applied. This paper presents a brief overview about Text Mining, data mining and its related terms. Text mining and data mining both are

applied for information extraction using a number of techniques. The major difference between these two processes is based on source of data on which mining techniques are to be applied. Data mining is applied on structured data while text mining is performed on unstructured or semi- structured data. Different text mining techniques are Information Extraction, information retrieval, document classification and clustering etc. Natural language processing and machine learning algorithm are also applied to mine text documents.

Information plays a vital role for an organization's success and generated from extraction of data. The major text mining application area are fraudulent detection, detection of spam, Customer relationship management, Research and development etc. In Text Mining , there are some issues such as ambiguities present in text . A number of researched has been made but still text mining is immature. Processing of natural language text is very difficult. A lots of research opportunities are available in this area.

7. REFERENCES

- *Shaidah Jusoh and Hejab M. Alfawareh, "Techniques, Applications and Challenging Issue in Text Mining", IJCSI International Journal of Computer Science Issues (ISSN (Online):1694-0814) , Vol. 9, Issue 6, No 2, November 2012 ,, pp. 431- 36.*

- *Vishal Gupta and Gurpreet S. Lehal , "A Survey of Text Mining Techniques and Applications " , JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009, pp. 60-76.*

- *Abhishek Kaushik and Sudhanshu Naithani, "A Comprehensive Study of Text Mining Approach ", International Journal of Computer Science and Network Security (IJCSNS), VOL.16 No.2, February 2016, pp 69-75.*

- *Ramzan Talib, Muhammad Kashif Hanif, Shaeela Ayesha, and Fakeeha Fatima,"Text Mining: Techniques, Application and Issues", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 7 No. 11, 2016, pp. 414- 418.*

- *Abhishek Kaushik and Sudhanshu Naithani, "A Study on Sentiment Analysis: Methods and Tools" International Journal of Science and Research (ISSN (Online): 2319-7064, Volume 4 Issue 12, December 2015).*

- *Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction" Multilingualism and*

Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.

- *Patricia Cerrito and John C. Cerrito, "Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs".*

- *Bill Hollingsworth, Ian Lewin and Dan Tidhar, "Retrieving Hierarchical Text Structure from Typeset Scientific Articles – a Prerequisite for E-Science Text Mining".*

- *Marti Hearst, "Text Mining Tools: Instruments for Scientific Discovery" IMA Text Mining Workshop April 17, 2000.*

- *Jadhav Bhushan G, Warke Pushkar U, Kuchekar Shivaji P and Kadam Nikhil V, "Searching Research Papers Using Clustering and Text Mining" International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014).*

- *Jae-Hong Eom and Byoung-Tak Zhang, "PubMiner: Machine Learning-Based Text Mining System for Biomedical Information Mining" AIMSA 2004, LNAI 3192, pp. 216–225, 2004.*

- *Divya NASA, "Text Mining Techniques-A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.*

- *Rashmi Agrawal and Mridula Batra, "A Detailed Study on Text Mining Techniques" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.*