RESEARCH ARTICLE OPEN ACCESS

Analyzing Different Approaches of Text Mining Techniques and Applications

S.Sheela [1], T.Bharathi [2]
Assistant Professor
Department of Computer Science [1]
Department of Computer Application[2]
Siddhar Sivagnaani Arts & Sscience College
Bommayapalayam, India

ABSTRACT
Text mining is used to extract hidden valuable information from Semi-Structured or Unstructured data. The amount of data is increasingly available in today's world, which is mostly in unstructured form. Hence, it is essential to extract the useful information with the help of Text Mining techniques. Number of techniques are available in Text Mining such as Information extraction, Clustering, Classification, and Summarization. In this paper our focus is to study the basic concept of various Text Mining techniques. In addition to the brief explanation of Text Mining concepts, its applications, benefits and limitations has been presented.

Keywords:— Text Mining, Information Extraction, Unstructred data.

I. INTRODUCTION

The general goal of Information Extraction is to discover structured information from unstructured or semi-structured text. Text Mining (TM) is the extraction of meaningful information from the text. It uses the data mining algorithm. Text mining is also called as Text Data Mining. TM is like a text data mining which is applied on textual data. It is used to read and analyze the textual information. In TM the pattern are extracted from the unstructured data or natural language text. In text mining the input is unstructured text and then the output is structured text. TM contains five key steps to process the text as shown in Figure-1. They are:

1. Collection of text document
2. Pre – processing of text
3. Text mining techniques
4. Analyze the text
5. Knowledge discovery

A collection of text document are in the form of pdf, doc, docx, txt..etc., After collecting the document, apply Pre – processing (compare to NLT – Natural Language Text) of text and then Text Mining techniques. Then, analyze the text document finally discovery the knowledge.Two methods are involved: Filtering and Streaming. Filtering is used to remove the unwanted words or relevant information. Streaming words provide the root for the related words. After applying the streaming method every word is represented by its root node.

Text Mining is an area which is a sudden spurt in adoptions for business applications. The spurt in adoption is triggered by heightened awareness about TM and the reduced price points at which TM tools are available today. Some real-world instances where TM has been successfully applied in three industries:

1. Automotive industry (warranty management)
2. Health care industry
3. Credit card industry

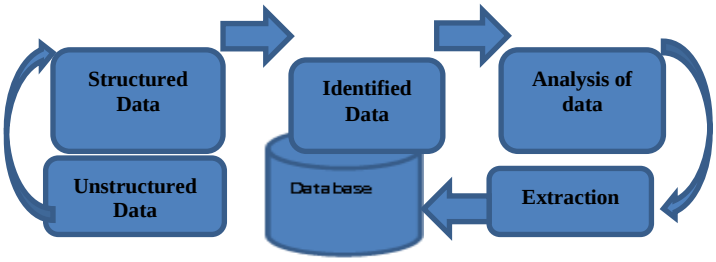


Figure 1: Text Mining process

Manual analysis of unstructured textual information is increasingly impractical, and as a result, text mining techniques are being developed to mechanize the process of analyzing the information.

The fundamental objective of TM is to enable users to extract data from text based assets and manages the operations like Retrieval, Extraction, Summarization, Categorization

(supervised) and Clustering (unsupervised), Segmentation and Association.

This paper explores recent efforts and contributions on TM techniques. Many Data Mining techniques have been planned for mining valuable patterns in text documents. However, how to successfully use and update exposed patterns is still an open research issue, especially in the domain of text mining. Therefore a number of research article and research papers and their contributions are placed in this paper.

II. DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

The difference between TM and Data Mining is based on source of data[1]. In TM, basically input is the unstructured file while for data mining input is of structured data[2].That means patterns are extracted from unstructured text in TM while in Data Mining, structured data is used.

Today the majority of available business data is unstructured information; even though it may also contain numbers, dates and facts in structured fields, unstructured information is typically text (articles, website text, blog posts, etc.)[3][4]. The presence of unstructured information makes it more difficult to effectively perform knowledge management activities using traditional business intelligence tools.

The discovery of knowledge sources that contain text or unstructured information is called "Text Mining"[5]. So, the main difference between data mining and text mining is that in text mining data is unstructured.

III. NEED OF TEXT MINING

Text mining is useful for handling textual data. Textual data is unstructured, difficult to manipulate and unclear, so text mining becomes most useful method for information exchange whereas data mining is basically applied on business data[6]. Vast amounts of new information and data are generated everyday through economic, academic and social activities, much with significant potential economic and societal value.

Techniques such as text and data mining and analytics are required to exploit this potential. The main goal of this strategy is to reduce efforts required for obtaining information from large set of textual documents.

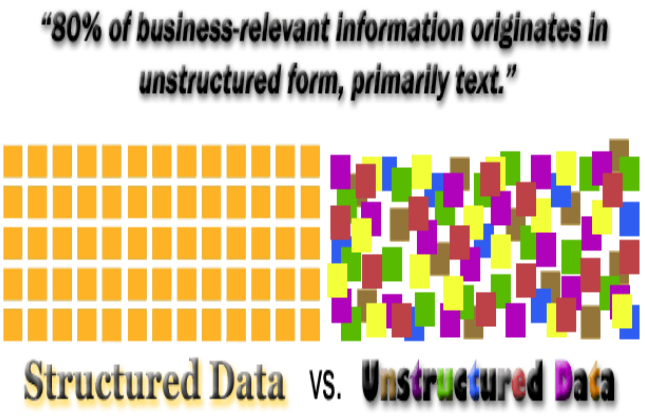


Figure 2: Structured data vs. Unstructured data

Structured Data

It concerns all data which can be stored in database SQL in table with rows and columns. They have relationnal key and can be easily mapped into pre-designed fields. Today, those data are the most processed in development and the simpliest way to manage informations. So let's introduce semi structured data.

Semi structured data

Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. With some process you can store them in relational database (it could be very hard for some kind of semi-structured data), but the semi - structure exist to ease space, clarity or compute.

Examples of semi-structured : CSV but XML and JSON documents are semi - structured documents, NoSQL databases are considered as semi - structured.

Unstructured data

Unstructured data represent around 80% of data. It often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents.

IV. TEXT MINING PROCESS

1) Document Gathering: In the first step, the text documents are collected, which are present in different formats. The document might be in form of pdf, word, html doc, css etc [7].

2) Document Pre- Processing:

In this process, the given input document is processed for removing redundancies, inconsistencies, separate words, stemming and documents are prepared for next step, the stages performed are as follows[8][9].

a) Tokenization:
The given document is considered as a string and identifying single word in document[8] i.e. the given document string is divided into one unit or token.
b) Removal of Stop word:
In this step the removal of usual words like a, an, but, and, of, the, etc.,
c) Stemming:
A stem is a natural group of words with equal (or very similar) meaning. This method describes the base of particular word. Inflectional and derivational stemming are two types of method[7]. One of the popular algorithm for stemming is porter's algorithm. e.g. if a document pertains word like resignation, resigned, resigns then it will be consider as resign after applying stemming method.

3) Text Transformation: A text document is collection of words (feature) and their occurrences[10]. There are two important ways for representations of such documents are Vector Space Model and Bag of words.

4) Feature Selection (attribute selection): This method results in giving low database space, minimal search technique by taking out irrelevant feature from input document[10][11]. There are two methods in feature selection i.e. filtering and wrapping methods.

5) Data mining/Pattern Selection: In this stage the conventional data mining process combines with text mining process[11]. Structured database uses classic data mining technique that resulted from previous stage.

6) Evaluate: This stage Measures the outcome. This resulted outcome can be put away or can be used for next set of sequence.

V. TEXT MINING TECHNIQUES

There are different kinds of techniques available. Some of the essential techniques are discussed in this section.

A. Information Extraction: Information Extraction is an initial step of analyzing unstructured text. Information Extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents[5]. In most of the cases this activity concerns processing human language texts

Text mining process

	CLASSIFICATION	CATEGORIZATION	CLUSTERING
Motivation	Automatic classification for the large number of on-line text documents (Web pages, e-mails, corporate intranets etc.)	Pre-given categories and labelled document examples (Categories may form hierarchy)	Automatically group related documents based on their contents No predetermined training sets or Taxonomies Generate a taxonomy at runtime
Processes & Models used	Data pre-processing Definition of training set and test Sets Creation of the classification model using the selected classification algorithm Classification model validation Classification of new/unknown text documents	Automatic: Typically exploiting machine learning techniques Vector space model based Prototype-based (Rocchio) Neural Networks (learn non-linear classifier) Support Vector Machines (SVM) Probabilistic or generative model based	Data pre-processing remove stop words, stem, feature extraction, lexical analysis etc. Hierarchical clustering-compute similarities applying clustering algorithms. Model-Based clustering (Neural Network Approach)- clusters are represented by exemplars (e.g.: SOM)
Algorithm Used	Support Vector Machines K-Nearest Neighbours Naïve Bayes Neural Networks Decision Trees Association rule-based Boosting	Naives Bayes, SVM K-Nearest Neighbour Decision Tree Neural Networking	Sequential algorithms Hierarchical algorithms Agglomerative algorithms Divisive algorithms Fuzzy clustering algorithms
Application	Document classification E-commerence interface (Amizon, ebay) Medical domain mesh Geo-demographic classification ACORN Data mining	Web pages New articles/ events tracked & filtered by topic Journal articles index by subject Categories Patents archived using international patent Classification Email message filtering	Document retrieval and texting Web support Pattern classification Image segmentation/ spatial data Analysis Data mining (economic science, scientific data exploration and tools

by means of natural language processing (NLP).

B. Summarization: This process has main goal of precise text from large number of text documents. Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Automatic data summarization is part of machine learning and data mining.

C. Topic Tracking: Basic idea of topic tracking mechanism is to maintain user profile based on previous searches and guesses other documents very effectively based on user profile[6][4][11]. Text mining is a field that automatically extracts previously unknown and useful information from unstructured textual data. It has strong connections with natural language processing. Topic tracking is one of the technologies that has been developed and can be used in the text mining process.

D. Classification

It is process of finding main theme of document by adding metadata and analyzing document[2][12]. This technique find counts of words and from that count decides topic of the document[14]. In this process, text documents are classified into predefined class label.

E.Categorization

Text categorization (a.k.a. text classification) is the task of assigning predefined categories to free-text documents[13]. It can provide conceptual views of document collections and has important applications in the real world.

F.Clustering Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data[13][14]. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

G. Concept Linkage Text mining uses the technique concept linkage to find related document. This mechanism browses documents instead of search. It offers the facility to link related documents.

H.Natural Language Processing Natural language is nothing but human language and that is processed with computer language, this whole interaction is called as Natural Language Processing (NLP). Main goal of NLP is to design and form such a computer system that will examine, understand and produce NLP.

VI. DIFFERENT RELATIVE TERMINOLOGIES

In this section we will show the main differences between classification, categorization and clustering. These terminologies can be differentiated on the basis of process and model used[13]. Another difference we have shown here is with its own algorithm and it is highlighted in Table-1, the main features associated with each terminology[11]. Each technique can be used in different fields on its need and area.

VII. COMPARISON OF TEXT MINING TECHNIQUES

In this section, main algorithms, models and tools are shown. Text mining uses various numbers of techniques which play an important role[14]. The techniques differ from each other. The Summarization technique is used to summarize the document which reduces length and keeps meaning same .The categorization is supervised process and uses predefined set documents according to their contents.

Responsiveness and flexibility of the post-co-ordinate system effectively prohibit the establishment of meaningful relationships because a category is created by individual not the system[16][17]. While as the clustering is used to find intrinsic structures in information and arrange them into related subgroups for further study and analysis[18].

VIII. APPLICATIONS

8.1 Business Intelligence:

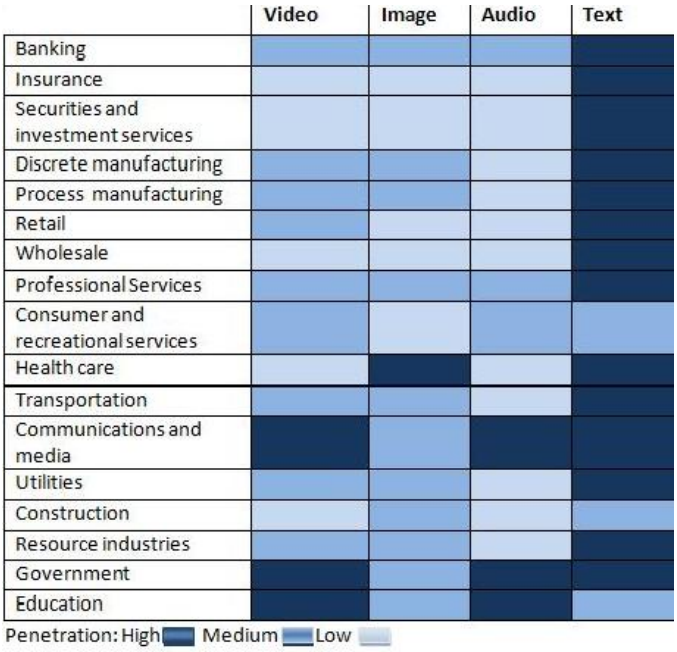
Text mining techniques helps for determining particular topic. Business intelligence, or BI, is an umbrella term that refers to a variety of software applications used to analyze an organization's raw data. BI as a discipline is made up of several related activities, including data mining, online analytical processing, querying and reporting [15].

8.2 Bioinformatics:

Nowadays, Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain[16][18]. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics.Example 1: Text Minning sector

8.3 Security Application:

Application security encompasses measures taken throughout the code's life-cycle to prevent gaps in the security policy of an application or the underlying system (vulnerabilities) through flaws in the design, development, deployment, upgrade, or maintenance or database of the application.



8.4 Human Resource Management:

Human Resource Management (HRM) is the term used to describe formal systems devised for the management of people within an organization. The responsibilities of a human resource manager fall into three major areas: staffing, employee compensation and benefits, and defining/designing work.

8.5 Web Search Enhancement:

In text mining, by using text categorization techniques such as CatS. The presentation of result is by sorting them into (a hierarchy of) clusters which may be displayed to the user in a variety of ways, e.g. as a separate expandable tree (vivisimo.com) or arcs which connect Web pages within graphically rendered "maps" (kartoo.com).

8.6 Text mining is also used in following sectors -

- Publishing and media
- Telecommunications, energy and other services industries.

Table 1: Comparison of Text Mining techniques

Technique	Characteristics	Tools
Retrieval	**Retrievals valuable information From unstructured text**	**Intelligent Miner, Text Analyst**
Extraction	**Extract information from structured database**	**Text Finder, Clear Forest Text**
Summarization	**Reduce length by keeping its main points and overall meaning as it is**	**Tropic Tracking Tool, Sentence Ext Tool**
Categorization	**Document based categorization**	**Intelligent Miner**
Cluster	**Cluster collection of documents, Clustering, classification and analysis of text Document**	**Carrot, Rapid Miner**

- Information technology sector and Internet.
- Banks, insurance and financial markets.
- Pharmaceutical and research companies and healthcare.

IX. PRACTICAL APPLICATION OF TEXT MINING TECHNIQUES

Some of the practical applications of text mining techniques include:

- Spam filtering
- Creating suggestion and recommendations (like amazon)
- Monitoring public opinions (for example in blogs or review sites)
- Customer service, email support
- Automatic labeling of documents in business libraries
- Measuring customer preferences by analyzing qualitative interviews
- Fraud detection by investigating notification of claims
- Fighting cyberbullying or cybercrime in IM and IRC chat

X. ADVANTAGE AND DISADVANTAGE IN TEXT MINING

Advantages of Text Mining

- ✓ As database can store less amount of information, this problem has been solved through Text Mining.
- ✓ Using the technique such as information extraction, the names of different entities, relationship between them can easily be found from the corpus of documents set.
- ✓ Text mining has solved the problem of managing such a great amount of unstructured information for extracting patterns easily; otherwise it would have been a great challenge.

Disadvantages of Text Mining

- ✓ Programs cannot be in order to analyze the unstructured text directly to mine the text for information or knowledge.
- ✓ The initial needed information is not given in the text documents.

XI. CONCLUSION

Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. Here in this work quite big research field ―Text Mining is discussed with its various techniques. In this study, the idea of TM techniques have been introduced and presented. Due to its novelty, there are many potential research areas in the field of Text Mining, which includes finding better intermediate forms for representing the outputs of information extraction or retrieval. In this paper, various TM techniques are discussed with application. In addition to this we have compared the TM technique on the basis of characteristic, algorithm used, models and tools. Applications in the field such as identifying business intelligence, bioinformatics, security application, web search enhancement, CRM and other sectors are also presented.

REFERENCES:

[1] Shilpa Dang, Peerzada Hamid Ahmad, "Text Mining: Techniques and its Application", IJETI International Journal of Engineering & Technology Innovations, Vol. 1 Issue 4, November 2014

[2] 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.

[3] K. Thilagavathi, V. Shanmuga "A Survey on Text Mining Techniques", International Journal of Advanced Research in Computer Science and Robotics , ISSN: 2320 7345 Volume 2, Issue 10, Octl 2014 pp 41-50.

[4] R. Patel, G. Sharma," A survey on Text mining techniques", International Journal of engineering Computer Science ISSN 2319-7242 Volume-2 Issue-5 May-2014.

[5] R. Balamurugan, Dr. S. Pushpa," A REVIEW ON VARIOUS TEXT MINING TECHNIQUES AND ALGORITHMS", 2nd International Conference on Recent Innovation in Science, Engineering and Management, JNU Convention Center, Jawaharlal Nehru University, New Delhi, Nov 2015.

[6] Andreas Hotho, Andreas Nurnberger and Gerhard PaaB, "A Brief Survey of Text Mining", May 13, 2005.

[7] S. M. Inzalkar and Jai Sharma, "A Survey on Text Mining- techniques and application", International Journal of Research In Science & Engineering e-ISSN: 2394-8299 Special Issue: Techno-Xtreme 16 p-ISSN: 2394-8280.

[8] Dr. Shilpa Dang and Peerzada Hamid Ahmad," A Review of Text Mining Techniques Associated with Various Application Areas", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

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

[10] Lokesh Kumar and Parul Kalra Bhatia,"Text Mining:Concept,Process,Applications," Journal of Global Research in Computer Science Volume 4, No. 3, March 2013 .

[11] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, " Effective Pattern Discovery for Text Mining," ", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

[12] Wimalasuriya, D. and Dou, D.. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science,Vol. 36(3), pp. 306-323 article, http://jis.sagepub.com/content/36/3/306.

[13] Dr. G. Rasitha Banu and VK Chitra," A Survey of Text Mining Concepts", International Journal of Innovations in Engineering and Technology [IJIET].

[14] Sathees Kumar B and Karthika R," A SURVEY ON TEXT MINING PROCESS AND TECHNIQUES", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 7, July 2014.

[15] Ananthi Sheshasaayee and R.Jayanthi, "Exploring the potential of Social Media Data using Text Mining to augment Business Intelligence", COMPUSOFT, An International Journal of Advanced Computer Technology, Volume-3, Issue-4, April 2014.

[16] Ruchika R. Tated and Mangesh M. Ghonge," A Survey on Text Mining- techniques and application", International Journal of Research in Advent Technology (E-ISSN: 2321-9637) Special Issue 1st International Conference on Advent Trends in Engineering, Science and Technology "ICATEST 2015", 08 March 2015.

[17] Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009.

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