International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

# A Survey on Text Mining and Its Techniques

**Amrut M. Jadhav[1], Devendra P. Gadekar[2]**

[1]Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India

[2]Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering & Research, Wagholi, Pune, India

**Abstract:** ***Unstructured text has huge amount information which is not easily used by the computer for processing. So that we require certain techniques to accomplish this task for extracting required patterns. Text mining plays an important role of extracting useful patterns from unstructured text. It is one of the emerging technologies for Knowledge Discovery Process. Document organization and pattern discovery becomes the main task in data mining. In this paper, a survey of Text mining & its techniques, applications, merits and demerits of text mining have been presented.***

**Keywords:** Text mining, information extraction, summarization, topic tracking, classification.

## 1. Introduction

Discovering patterns is a great challenge due to huge amount information increases day by day to find accurate knowledge [1] [11]. Text mining is used to perform this task which gives better performance for finding the relevant information. To resolve this problem, various techniques [5] of text mining are discussed in this work. Text mining is nothing but extracting patterns from number of unstructured text documents. This technique is also called as Knowledge Discovery from Text (KDT) [6].Text documents are considered as semi-structured or unstructured format. Computer has not that much capability to easily differentiate linguistic patterns as compare to human. But the computer can process text at high speed and in large volume. So, text mining becomes useful for computer to examine unstructured data. This technique employs numbers of algorithms to for converting unstructured text into useful patterns. Text summarization, text categorization and text clustering these are the functions of text mining [7].This paper provides the general overview of text mining, techniques of text mining, their merits, applications and demerits.

## 2. Difference between Text Mining and Data Mining

The difference [9] between text mining and data mining is based on source of data. In text mining, basically input is the unstructured file while for data mining input is of structured data. That means patterns are extracted from unstructured text in text mining while in data mining, structured data is used.

## 3. Need of Text Mining

Text mining is useful [1] for handling textual data. Textual data is unstructured, difficult to manipulate and unclear, so that text mining becomes most useful method for information exchange whereas data mining is basically applied on business data [12]. Text mining belongs to a nontraditional information retrieval strategy. The main goal of this strategy is to reduce efforts required for obtaining information from large set of textual documents.

## 4. Text Mining Process

Overall process of text mining is depicted in the figure 1[9].Text mining process comprises of text pre-processing, text transformation, feature selection, pattern discovery and evaluation.

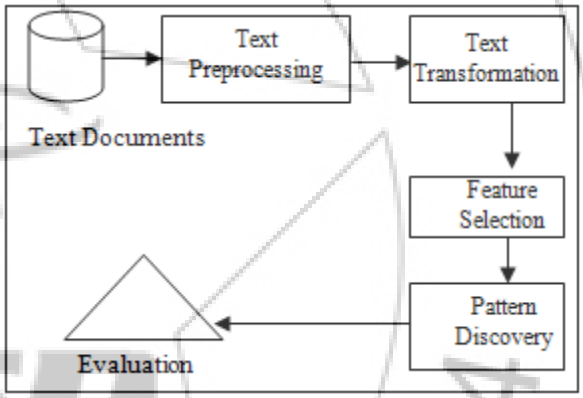


**Figure 1**: Text Mining Process

### 4.1 Text preprocessing

Text preprocessing is the initial step of text mining which reads one text document at time and processes it. This step divides into following main three subtasks-

#### 4.1.1Tokenization

Generally text document contains multiple sentences. So this process divides whole sentence into words by removing comma, spaces, punctuations etc.

#### 4.1.2Stop Word Removing

This process removes stop words such as "the", "are", "a" or any tags like HTML tag etc.

#### 4.1.3Stemming

Stemming is applied after stop word removal by reducing the word to its root word. E.g. "playing", "played" are stemmed to "play".

### 4.2 Text Transformation

Text transformation has the role of conversion of text document into words so that it will useful for further processing.

### 4.3 Feature Selection

It performs removing features that are considered unrelated for mining purpose.

### 4.4 Pattern Discovery

Pattern discovery is one of the important processes that use methods for discovering patterns. Methods include clustering, classification, summarization, information retrieval, topic extraction etc.

## 5. Techniques of Text Mining

In present days language analysis would work better using computer as compare to human. So the manual technique was expensive and takes more time. To achieve this goal of text mining, there are different recent technologies are available by which text mining is performed [3].In this section, different text mining technologies are discussed for mining text.

### 5.1 Information extraction

Information extraction is an initial step of analyzing unstructured text. General meaning of this process is simplification of text. It recognizes phrases and finds the relationships between them are the key goal of information extraction [10].So that this technique is useful for bulky size of text. To recognize phrases, pattern matching approach is used in which comparison of user text with predefined sequence of text is done. Its extracts structured information from unstructured information. Figure 2 shows process of information extraction.

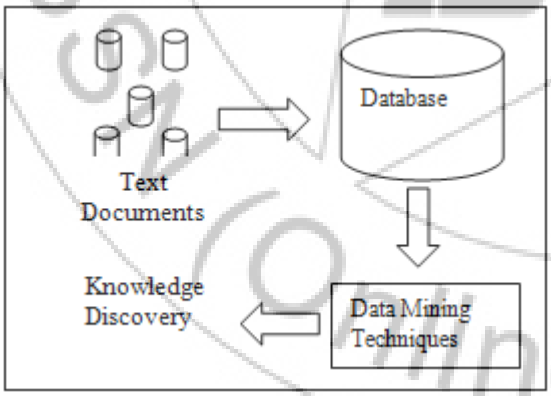


**Figure 2**: Information extraction

### 5.2 Summarization

This process has main goal of precise text from large number of text documents. Manually it is not possible to summarize large documents [9].In many research centers, it is not possible to read all text documents that means researcher has no time to read all this documents. They summarize documents and makes summary of document from main points. Summarization has basically two methods that are extractive and abstractive. In extractive method, important sentences, paragraph etc. are selected from a document and then they are combined to form a short version of text. Importance of sentence or paragraph is decided on the basis of statistical and linguistic features of information. In abstractive method, whole concept of document is to be expressed in natural language by understanding the entire document. It uses linguistic methods to describe the entire document and forms a new text that delivers importance of an original document.

### 5.3 Topic Tracking

Basic idea of topic tracking mechanism is to maintain user profile based on previously searches and guesses other documents very effectively based on user profile [13]. Previously searched records are maintained in user profile. This mechanism is useful for the studies of new and forthcoming news related to search. It has one limitation related to search data because it searches relevant data as well as unnecessary data also. Topic tracking is used in many areas such as radio, news broadcasts etc. In industry, this technology is useful for checking the news as compared with its competitor's products or updates in the market.

### 5.4 Classification

It is process of finding main theme of document by adding metadata and analyzing document [4]. This technique find counts of words and from that count decides topic of the document. In this process, text documents are classified into predefined class label [8]. Classification used in customer feedback, filtering emails etc.

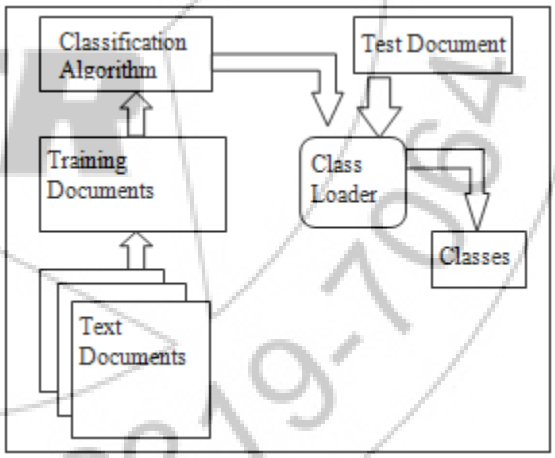


**Figure 3**: Classification

### 5.5 Clustering

Clustering has no predefined class labels, instead of that it uses similarity measures among different objects and places similar object in one class and dissimilar objects in another different class [7]. This technique divides text into one group and in that way creates cluster of group. It is a technique of grouping similar documents but varies from categorization. Words are separated very fast then weights are assigned to each word. After calculating similarity, clustering algorithms [9] are applied to generate list of classes. This process of clustering is depicted in Figure 4.

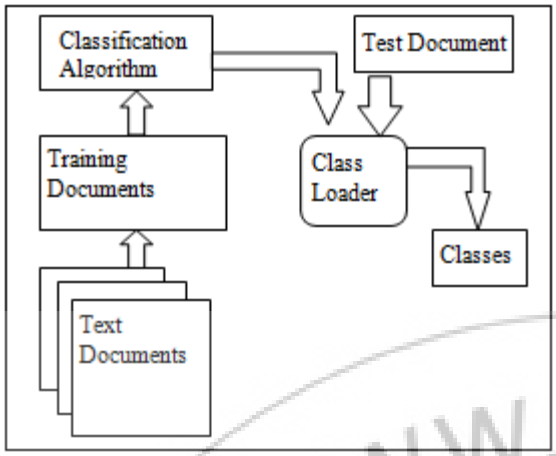


**Figure 4**: Clustering

### 5.6 Concept Linkage

Text mining uses the technique concept linkage to find related document [8]. This mechanism browses documents instead of search. It offers the facility to link related documents. This technique is useful in many areas such as medical field to find documents related to diseases and treatment so that it helps to doctor very efficiently. Government also uses concept linkage for criminal records with previous records for getting idea about criminal and its relationship.

### 5.7 Information visualization

It provides visual representation for text mining instead of simple searching for extracting the patterns. That's why this technique is also called as Visual Text mining [10]. Information visualization has three important steps for performing text mining namely data preparation, data analysis & extraction, visualization mapping .User can interact with document by doing numbers of operations like zooming, scaling etc. Government can use this approach for checking terrorist network and crimes etc. Information visualization process is depicted in figure 5.

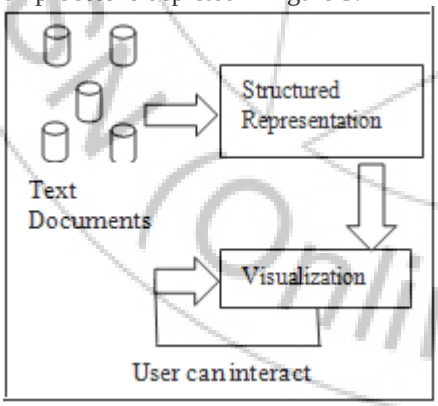


**Figure 5**: Information visualization

### 5.8 Question Answering

It is mechanism of finding best answer for a question. So many websites are available which is of type Question answer. This takes user question and gives best answer to the user [8]. This type of technique uses text mining for extracting the correct answer for user. In many industries, they can use this technique in which employee can ask the question and get appropriate answer.

### 5.9 Association Rule Mining

Association rule mining (ARM) [14] has the main goal of to find relationship between massive set of variables in a data set. That means there is given database records and that contains number of variables with its value. So ARM finds variable-value combination within the database records that are repeatedly occur. Basically ARM identifies relationship between two or more variables. That relationship is called as Association rule. This technique is used to discover items those are commonly purchase by customer and place adjacent with another item. So that customer can purchase these items then automatically increases the sell.

### 5.10 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) [13][2][3]. Main goal of NLP is to design and form such a computer system that will examine, understand and produce NLP. It is used in various areas like robotic system, fiction etc. and does the translation of one human language text into another.

## 6. Measures of Text Retrieval

The set of relevant document is denoted by {Relevant} which is relevant to given query. Similarly set of retrieved document is denoted by {Retrieved}.In certain conditions, there are also documents those are relevant as well as retrieved is denoted by using notation {Relevant}∩{Retrieved}.Figure 6 shows all these notations with concept. Precision and Recall [1] [2] are the two basic measures used for reading the quality of text retrieval.

a) **Precision**: - Precision is nothing but percentage of retrieved documents those are relevant to the query.
b) **Recall**: - Recall is nothing but percentage of relevant documents those are relevant to the query.

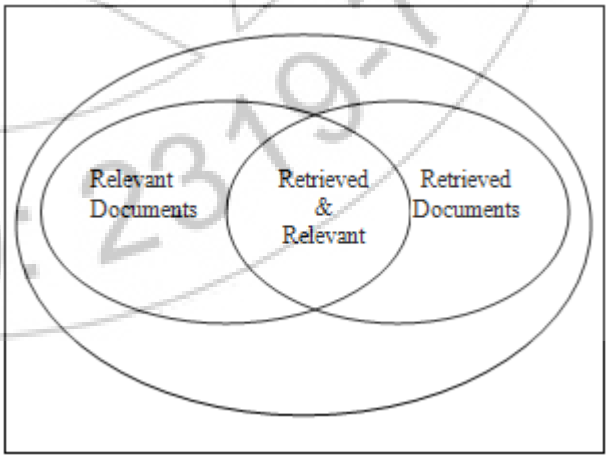


**Figure 6**: Venn diagram of relationship between relevant and retrieved documents

## 7. Applications

Text mining is an emerging technology used for extracting pattern from unstructured data. It has following applications [7] [8]-

- Security applications: Used for monitoring and analyzing plain text on internet such as blogs, news etc.
- Biomedical Application: Identifying technical terms in biomedical field.
- Company Resource Planning: For mining reports, activities in company
- Customer Relationship Management (CRM):Text mining is also useful in Customer Relationship Management (CRM) for supplying immediate answers to frequently asked questions
- Text mining is also used in following sectors [6]-
  a. Publishing and media.
  b. Telecommunications, energy and other services industries.
  c. Information technology sector and Internet.
  d. Banks, insurance and financial markets.
  e. Pharmaceutical and research companies and healthcare.

## 8. Merits

- Most of data in industries are stores in the form of text like emails, memos, feedback etc. So text mining helps to solve the problem of finding relevant pattern from unstructured text [1].
- As data is growing rapidly in any organization, so it not possible to store that data in database due to its size limitation [2]. So most of organization stores data in the form of text. Text mining is applied on that data for pattern extraction.

## 9. Demerits

- Data collection requires handling a lot of unstructured text in in text-mining [1].
- Ambiguities enclosed into natural language texts so that it needs human interference.
- To analyze unstructured text, there is no any program available that handle this text for text mining.

## 10. Conclusion

Text mining technique is basically used for extracting pattern from unstructured data. Various techniques for efficiently performing text mining are discussed in this paper. So in this paper, our focus is basically on how text is to be mined. We have also discussed process of text mining, its applications, merits and demerits.

## References

[1] 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.
[2] Nitin Jindal and Bing Liu, "Identifying Comparative Sentences in Text Documents", University of Illinois at Chicago
[3] Mrs.K. Mythili, and Mrs. K. Yasodha, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining", International Journal of Science and Applied Information Technology, Volume 1, No.3, July – August 2012
[4] Deepshikha Patel, Monika Bhatnagar, "Mobile SMS Classification", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307 (Online), Volume-I, Issue-I, March 2011.
[5] Ranveer Kaur, Shruti Aggarwal, "Techniques for Mining Text Documents", International Journal of Computer Applications (0975 – 8887) Volume 66– No.18, March 2013.
[6] 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
[7] Rashmi Agrawal, 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.
[8] Vishal gupta and Gurpreet S. Lehal , "A survey of text mining techniques and applications", journal of emerging technologies in web intelligence, 2009,pp.60-76.
[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] N. Kanya and S. Geetha ,"Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Comm. Technology in Electrical Sciences, IEEE(2007), Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111- 1118.
[11] http://people.ischool.berkeley.edu/~hearst/text-mining.html
[12] http://www.cis.upenn.edu/~ungar/KDD/text-mining.html
[13] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg.
[14] Dion H. Goh and Rebecca P. Ang," An introduction to association rule mining: An application in counselling and help seeking behaviour of adolescents", Journal of Behaviour Research Methods39 (2), Singapore, 259-266,2007.