Text Mining Techniques and Its Applications: A Survey

[1]R.Annamalai Saravanan, [2]Dr. M.Rajesh Babu
[1]Research and Development Center, Bharathiar University, Coimbatore, Tamil Nadu, India
[2]Dept. of CSE, Karpagam College of Engineering (Autonomous), Coimbatore, Tamil Nadu, India

Abstract
The massive growth of databases in nearly all area of human activity has created a huge demand for new, the great tools for turning data in helpful knowledge. Text mining is supported for research area. Text Mining is finding with new computer, previous unknown information, by automatically extracting information from various written resources. This survey article is main objective is to provide easy accessibility to the major ideas for non experts about Text Mining Techniques, applications and Difference between text mining and data mining, text mining applications have been presented.

Keywords
Text Mining, Information Extraction, Preprocessing, Classification, Summarization, Clustering, Question Answering etc.

I. Introduction

Nowadays, due to computational automation various different text document sources are available. Extraction of patterns and organizing the text document is a key target of text mining technique development. Text mining is related to data mining, with the exception of that data mining tools are considered to manage structured data, but text mining can work with formless or semi structured data sets [1]. Text mining or knowledge discovery is that sub process of data mining that is widely used to discover hidden patterns and significant information from the huge amount of unstructured written material. This text mining uses techniques of different fields like machine learning, case based reasoning, visualization, database technology statistics, text analysis, knowledge management, natural language processing and information retrieval. Text mining is growing field of computer science simultaneously to big data and artificial intelligence [2]. Text mining is useful for computer to check unstructured data. This technique utilizes amount of algorithms to for transferring unstructured text into useful patterns. Text summarization, text categorization and text clustering these are the functions of text mining. This paper provides the general idea of text mining, techniques of text mining and applications [3].

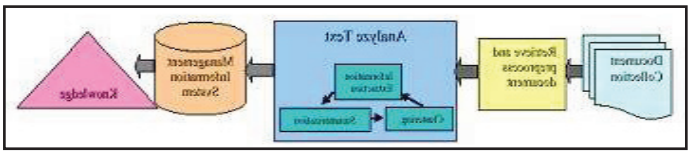

Fig. 1: An Example of Text Mining

II. Data Mining vs Text mining

Text mining is a variation on a field called data mining [3]. Data mining is Discover new knowledge through analysis of data. Text mining is Discover new knowledge through analysis of text. That means patterns are extracted from unstructured text in text mining as in data mining, structured data is applied [4].

III. Need of Text Mining

Text mining is useful 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. Text mining belongs to a nontraditional information retrieval strategy. The main objective of this method is to minimize efforts needed for getting information from huge set of textual documents.

IV. Text Mining Process

The overall processes of the text mining are shows in the fig. 2.

A. Text Preprocessing

The text preprocessing step divided into number of sub steps as given below.

Tokenization: Text document have a set of sentences. This step divide entire statement into words with removing spaces, commas etc.

Stop word Removal: This stop word removal removing XML and HTML tags from web pages. Then process of removal of Stop words like "a", "of" etc. is executed. The last step is using stemming method.

Stemming: This technique is used to get the root or stem of a word. Stemming technique converts words to their stems. For example Flying, Flew word to fly. The algorithm proposed by Port is known as a Port's stemming algorithm is commonly used for the same.

B. Text Transformation or Feature Generation

Text transformation means to convert text document into bag of words that can be used for more efficient analysis task.

C. Feature Selection/Attribute Selection

This phase mostly performs removing features that are considered unrelated for mining purpose. This procedure is to provide benefit for less computations, lesser dataset size and reduce search space..

D. Pattern Discovery

Pattern discovery is main processes that methods used for discovering patterns. This Method includes clustering, summarization, information retrieval, topic extraction etc [5].

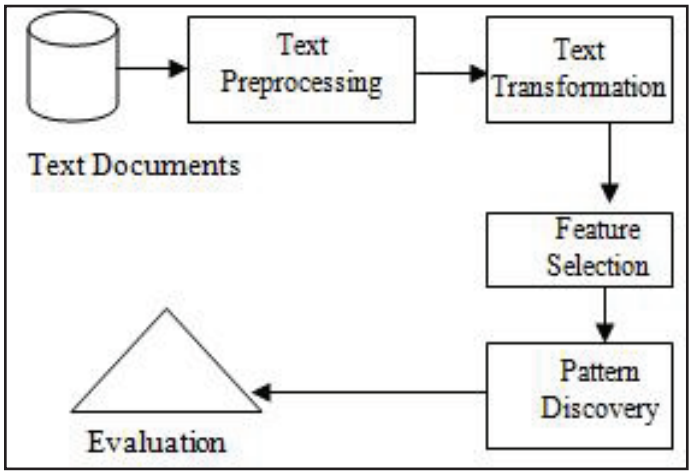

Fig. 2: Text Mining Process

V. Techniques of Text Mining

In this section discuss about text mining techniques so it will be helpful to work further in the interested area. There are given below:

A. Information Extraction

Information extraction method knows key phrases and relationship within a text. For that it is using pattern matching method. Pattern matching method means matching already defined sequences of text along with user text. This technique is more helpful in analyzing bulk text dataset. The extracted information cannot be represented directly into a structured form. Consequently need for post processing [6].

B. Summarization

Summarization is important techniques for text mining. In simple words summarization is a process of creating summary of any document containing huge amount of information while theme or main objective maintained document. It supports user to understand whether a particular document is useful or not. Compression is also interrelated to summarization, but is not in human's readable form [7].

C. Topic Tracking

Topic tracking is used to facilitate user by managing the topic searched. Next time system predict user's other search documents interrelated to previous topic very effective manner. Topic detection studies the problem of detecting new and upcoming topics in time ordered documents. The methods are repeatedly used to detect and monitor news tickers or news broadcasts.

D. Classification

Classification method classifies text documents into predefined class label. Classification has been used in several applications like as online customer feedback classification, Mobile SMS classification, business reports classification etc. Classification can be combined with topic tracking to classify the documents by topic and therefore making the process faster.

E. Clustering

Clustering is a no predefined class labels but using similarity measures between dissimilar objects, it put most related object in one class and unrelated object in another class. Text clustering algorithms are divided into different types such partitioning algorithms, as agglomerative clustering algorithms and standard parametric modeling based methods [7].

F. Concept Linkage

Concept linkage finds interrelated documents who share common concepts. The purpose of concept linkage is providing browsing for information rather than searching for it as in information retrieval. For example in biomedical field this concept linkage method used to link diseases and treatment.

G. Information Visualization

It provides visual representation for text mining rather than simple searching for extracting the patterns. That's why this method is known as Visual Text mining. Information visualization has three main steps for performing text mining namely data preparation, data analysis, data extraction and visualization mapping. User can interact with document by perform more numbers of operations like scaling, zooming, etc.

H. Question Answering

This mechanism is used to finding best answer for a question. There are many websites are available this type of question answer. This receives user question and provides best answer to the user. This type of technique used for extracting the exact answer for user.

I. Association Rule Mining

The main purpose of association rule mining (ARM) is to find relationship between enormous set of variables in a data set. There is given database records and that includes amount of variables with its value. Hence, ARM gets variable value combination within the database records that are frequently occur. Fundamentally ARM find outs relationship between two or more variables. That relationship is known as Association rule. This technique is used to locate items those are generally purchased by customer and place adjacent with another item. So that customer is purchase these items after that increase sales automatically.

J. Natural Language Processing (NLP)

Natural language is simply human language and also it processed with computer language, this entire interaction is known as Natural Language Processing (NLP). The main objective of natural language processing is to design and form such as computer system that will examine, understand and make NLP [9]. It is used for different areas for example robotic system, fiction , etc and translation of one human language text into another human language text.

VI. Measures of Text Retrieval

The set of relevant document is represented by {Relevant} that is relevant to particular query. Similarly set of retrieved document is represented by {Retrieved}. It follows certain conditions, there are also documents those are relevant with retrieved is represented by using notation {Relevant} $\cap$ {Retrieved}. The quality of text retrieval measures using two measurement of precision and recall method.

1. **Precision: -** Precision is calculating percentage of retrieved documents which are relevant to the query.
2. **Recall: -** Recall is calculating percentage of relevant documents which are relevant to the query.

VII. Applications

The main Text Mining applications are most often used in the following sectors [10]:

1. Publishing and media.
2. Telecommunications field.
3. Energy and other services industries.
4. Information technology sector and Internet.
5. Financial markets, Banks and insurances.
6. Political institutions, public administration and legal documents.
7. Research companies, healthcare and Pharmaceutical.

VIII. 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. The main focus on this survey is basically on how text is to be mined. We have also discussed process of text mining, its applications.

References

[1] K.Thilagavathi, V.Shanmuga Priya,"Survey on text mining techniques", International Journal of Research in Computer Applications and Robotics, 2014.

[2] Kaushik, A., Naithani, S.,"A Comprehensive Study of Text Mining Approach", International Journal of Computer Science and Network Security (IJCSNS), 16(2), pp. 69, 2016.

[3] Jadhav, A.M., Gadekar, D.P.,"A Survey on Text Mining and Its Techniques", International Journal of Science and Research (IJSR), 2012.

[4] Falguni N. Patel, Neha R. Soni,"Text mining: A Brief survey", International Journal of Advanced Computer Research, Vol. 2, No. 4, Issue 6, 2012.

[5] Patel, F.N., Soni, N.R.,"Text mining: A Brief survey", International Journal of Advanced Computer Research, 2(4), pp. 243-248, 2012.

[6] N. Kanya, 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, pp. 1111-1118.

[7] Gupta, V., Lehal, G.S.,"A survey of text summarization extractive techniques", Journal of emerging technologies in web intelligence, 2(3), pp. 258-268, 2010.

[8] Rashmi Agrawal, Mridula Batra,"A Detailed Study on Text Mining Techniques", International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue 6, January 2013.

[9] Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang,"Tapping into the Power of Text Mining", Journal of ACM, Blacksburg, 2005.

[10] Vishal Gupta, Gurpreet S. Lehal,"A survey of text mining techniques and applications", Journal of Emerging Technologies in web intelligence, pp. 60-76, 2009.

R. Annamalai Saravanan received his B.S. degree in Computer Science from Kongunadu Arts and Science College, Coimbatore, India in 1998, the MCA., Degree from Madurai Kamarajar University, Madurai, India in 2001 and the M.Phil in Computer Science Degree in Datamining from Manonmaniam Sundaranar University Tirunelveli, India in 2007. He was a lecturer, Assistant Professor and Head of the Department of Computer Application, SNMV College of Arts and Science in 2007. He was heading the Department of Computer Science and Application at Sankara College of Arts and Science, Coimbatore, India in 2012. Presently, working as a Head Department of Computer Application, Nehru Arts and Science College, Coimbatore, India. His research interests include Text mining and Data mining. He is engaged in Text Classification techniques.