IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A STUDY OF TEXT MINING METHODS, APPLICATIONS,AND TECHNIQUES

R. Rajamani[*1] & S. Saranya[2]

[*1]Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, India

[2]M.Phil Research scholar, PSG College of Arts & Science, Coimbatore, India

DOI: 10.5281/zenodo.829803

ABSTRACT

Data mining is used to extract useful information from the large amount of data. It is used to implement and solve different types of research problems. The research related areas in data mining are text mining, web mining, image mining, sequential pattern mining, spatial mining, medical mining, multimedia mining, structure mining and graph mining. Text mining also referred to text of data mining, it is also called knowledge discovery in text (KDT) or knowledge of intelligent text analysis. The process is driving high-quality information from not-structured to semi-structured data. Text mining is the discovery by automatically extracting information from different written resources and also by computer for extracting new, previously unknown information. This paper discusses about the process of text mining, methods, tools, applications and techniques.

KEYWORDS: Text mining, NLP,IE,IR, Topic tracking, Summarization, clustering, Categorization, Visualization, Association rule.

I. INTRODUCTION

Data mining technology helps to extract useful information from various databases. Data warehouses are good for only numerical solution but unsuccessful when it came to textual information. As text mining is extraction of useful information from text data it is also known as text data mining or knowledge discovery from textual databases. Text mining is a technique which extracts information from unstructured data and find pattern which is novel and unknown earlier. It is used to retrieve hidden high quality information form document. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text documents are in semi-structured or unstructured format datasets such as emails, full-text documents, HTML files pdf. Text Mining is to analyze large quantities of natural language text and it detects lexical patterns to extract useful information. Text Mining is useful for organization because most of the information is in text format. The following steps can be included in text mining.

- It converts the unstructured text into structured data.
- Identify the patterns from structured data.
- Analyze the patterns using Text Mining techniques.
- Extract the useful information from the text.

The applications of Text Mining are protein interaction, drug discovery, predictive toxicology, identification of recent product potentialities, detection of links between lifestyle and states of health, competitive intelligence and lots of additional.

II. TEXT MINING PROCESS

A. **Document Gathering:** The documents are collected that are present in numerous formats. The document could be in form of word, html, css, pdf.

B. **Document Pre Processing:** The given document is processed for eliminating redundancies, inconsistencies, separate words, stemming and documents are ready for next step, and the stages performed as follows:

- **Tokenization**: The recognizing as a string and identifying single word in document. The given document string is distributed into one unit or token.

- **Removal of Stop word**: It removal of common words like a, an, and, of, but.
- **Stemming**: A stem may be a natural group of words with very similar meaning. This method defines the base of the particular word. There are two types of stemming method, Inflectional and Derivational.

C. **Text Transformation:** A collection of words and their occurrences. There are two ways for representation of such documents is Bag of words and Vector space model.

D. **Attribute Selection:**This method leads to giving less database space, minimal search technique by removing irrelevant feature from input document. There are two methods in attribute selection, Filtering and Wrapping.

E. **Interpretation/ Evaluation:**In this stage measures the result, this result can be put away or it will be used for next set of sequence.

III. TEXT MINING METHODS

Traditionally there are so many techniques developed to solve the problem of text mining that is nothing but the relevant information retrieval according to user's requirement. According to the information retrieval basically there are four methods used

1) Term Based Method (TBM).
2) Phrase Based Method (PBM).
3) Concept Based Method (CBM).
4) Pattern Taxonomy Method (PTM).

1. Term Based Method

Term in document is word having semantic meaning. In term based method document is analyzed on the basis of term and has advantages of efficient computational performance as well as mature theories for term weighting. These techniques are emerged over the last couple of decades from the information retrieval and machine learning communities. Term based methods suffer from the problems of polysemy and synonymy. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods to solve this challenge.

2. Phrase Based Method

Phrase carries more semantics like information and is less confusing. In phrase based method document is analyzed on phrase basis as phrases are less ambiguous and more discriminative than individual terms. The likely reasons for the daunting performance include:

1) Phrases have inferior statistical properties to terms,
2) They have low frequency of occurrence, and
3) Large numbers of redundant and noisy phrases are present among them.

3. Concept Based Method

The terms are analyzed on sentence and document level. Text Mining techniques are mostly based on statistical analysis of word or phrase. The term frequency captures the importance of word without document. Two terms can have same frequency in same document, but the meaning is that one term contributes more appropriately than the meaning contributed by the other term. The terms that capture the semantics of the text should be given more importance so, a new concept-based mining is introduced.

This model included three components.

i. Analyzes the semantic structure of sentences.
ii. Constructs a conceptual ontological graph (COG) to describe the semantic structures.
iii. The first two components to build feature vectors using the standard vector space model. Concept-based model can effectively discriminate between non important terms and meaningful terms which describe a sentence meaning. It usually relies upon natural language processing techniques.

4. Pattern Taxonomy Method

In pattern taxonomy method documents are analyzed on pattern basis. Patterns can be structured into taxonomy by using is-a relation. Patterns can be discovered by data mining techniques like association rule mining, frequent item set mining, sequential pattern mining and closed pattern mining. Use of discovered knowledge (patterns) in the field of text mining is difficult and ineffective, because some useful long patterns with high specificity lack in support. The pattern based technique uses two processes pattern deploying and pattern evolving. This technique refines the discovered patterns in text documents. The experimental results show that pattern based model performs better than not only other pure data mining-based methods and the concept-based model, but also term-based models.

IV. TEXT MINING APPLICATION

The text mining technology is now mostly applied for a wide variety of government, research, and business needs. Application can be sorted into a number of categories by analysis type or by business function. Using this approach to classifying solutions, application categories include:

- Enterprise Business Intelligence /Data Mining, Competitive Intelligence.
- E-Discovery, Records management.
- National Security/ Intelligence.
- Scientific Discovery, especially Life Sciences.
- Sentiment analysis Tools, Listening Platforms.
- Natural Language/ Semantic Toolkit or service.
- Publishing and automated ad Placement.
- Search/ Information Access.
- Social Media Monitoring.

Text Mining, there are some applications explained here:

A. **Security Application:** Many text mining packages are marketed for security applications, particularly observation and analysis of online plain text sources like web news, blogs, and site for national security functions. It also concerned with the study of text encryption and decryption.

B. **Biomedical Application:** Text Mining is used in medical specialty for identification and classification of technical terms within the domain of biological science corresponding to the concepts.

C. **Company Resource Planning:** Mining Company's reports and correspondences for activities, so its resource status and problems will be handled properly and future action planned can be design.

D. **Market Analysis*:*** With the help of numerous text mining techniques, market analysis is concerned to analyze the competitors within the market and can also be used to monitor customer opinions and searing for new potential customers.

It mainly deals with managing the customer messages. CRM consists of providing applicable service to the customer as per their request and providing fast answers to their queries.

V. TEXT MINING TECHNIQUES

There are many techniques used for text mining, but most popular are Natural Language Processing (NLP) and Information Extraction (IE). Research is going on to analyze other techniques for text mining such as knowledge based, statical, rule-based and machine learning-based approaches. NLP focuses on text processing while IE focuses on extracting information from actual text. To analyze, understand and generate text, technologies are produced by natural language processing. The technologies like information extraction, summarization, categorization, clustering and information visualization, are used in the text mining process. In the following sections we will discuss each of these technologies and the role that they play in text mining. The types of situations where each technology may be useful in order to help users are also discussed.

A. Natural Language Processing

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text. NLP researchers aim to collect knowledge on how human beings understand and use language so that fitting tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the preferred tasks. The basics of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems and so on.

B. Information Retrieval

Information retrieval (IR) concept has been developed in relation with database systems for many years. Information retrieval is the association and retrieval of information from a large number of text-based documents. The information retrieval and database systems, each handle various kinds of data some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in conventional database systems, such as unstructured documents, estimated search based on keywords, and the concept of relevance. Due to the huge quantity of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines.

C. Information Extraction

The information extraction method identifies key words and relationships within the text. It does this by looking for predefined sequences in the text, a process called pattern matching. The software infers the relationships between all the identified places, people, and time to give the user with meaningful information. This technology is very useful when dealing with large volumes of text. Traditional data mining assumes that the information being “mined” is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural language documents rather than structured databases.

D. Topic Tracking

Topic tracking tool is offered by yahoo (www.alerts.yahoo.com) which notifies users when news related to chosen keyword becomes available. But this technology has its own limitations, taking example if we setup alert for “data mining” then we will receive several news and stories related to mining the minerals rather than text mining. Topic tracking is significantly applied in industries where companies can generate alert to find competitors in news. It track news of their own product and company. Keyword extraction is main process in topic tracking.It is important words in an article which gives high level description of its content to readers. Identifying keywords from huge amount of online news data is very valuable. Manual extraction of keywords is very difficult and time consuming. So fast extraction of keywords automated process is needed

E. Summarization

Text summarization is to reduce the length and detail of a document while retaining most important points and general meaning. Text summarization is helpful for to figure out whether or not a lengthy document meets the user’s needs and is worth reading for further information hence summary can replace the set of documents. In the time taken by the user to read the first paragraph text summarization software processes and summarizes the large text document. It is difficult to teach software to analyze semantics and to interpret meaning of text document even though computers are able to identify people, places, and time. Humans first reads entire text section to summarize then try to develop a full understanding, and then finally write a summary, highlighting its main points.

F. Categorization

Categorization involves identifying the main themes of a document by inserting the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a “bag of words.” It does not try to process the actual information as information extraction does. Rather, the categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a glossary for which topics are predefined, and relationships are identified by looking for large terms, narrower terms, synonyms, and related terms.

G. Clustering

Clustering method can be used in order to find groups of documents with similar content. The outcome of clustering is typically a partition called clusters P and each cluster consists of a number of documents d. The contents of the documents within one cluster are more similar and between the clusters more dissimilar then the quality of clustering is considered better. Even though clustering technique used to group similar documents it differs from categorization because in clustering documents are clustered on the fly instead of use of predefined topics. As documents can appear in multiple subtopics clustering ensures that a useful document will not be omitted from search results.

H. Concept linkage

Connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps wouldn't have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research.

I. Visualization

In text mining visualization methods can improve and simplify the discovery of relevant information. To represent individual documents or groups of documents text flags are used to show document category and to show density colors are used. Visual text mining puts large textual sources in a visual hierarchy. The user can interact with the document by zooming and scaling. Information visualization is applicable to government to identify terrorist networks or to find information about crimes.

J. Association Rule Mining

This technique is used to find relationships among large set of variables in data set. It has huge advantage in field of industry, it is discovering relationship among large set of variables, while database of records is present each containing two or more variables and their corresponding values. It checks frequently occurring combination of variable-value. In ARM a relationship can contain two or more variables. It is mostly used to find out which item customers buys together. In text mining ARM is used to study relationships. Association rules are widely used in various areas such as telecommunication networks, market and risk management, and inventory control.

VI. CONCLUSION

Text mining field also known as data mining tries to find interesting patterns from large databases. The aim of this paper is to analyses briefly about the text mining process and methods of text mining. In this paper some real time applications are also listed.To extract useful information, techniques such as natural language processing, information retrieval summarization, classification, clustering, information extraction and visualization are available for the same which comes under the category of text mining. These concepts used in text mining yield best results which can be implemented in the future.

VII. REFERENCES

[1] Susanneviestan, "Three methods for keyword extraction", MSc. Department of linguistics, Uppsds University, 2000.

[2] U. Y. Nahm and R. J. Mooney. "Text mining with information extraction".*In AAAI 2002Spring Symposium on Mining Answers fromTexts and Knowledge Bases*, 2002.

[3] .G. Ercan and I. Cicekli, 2007, "Using lexical chains for keyword extraction", published in International Journal of Information Processing and Management,.

[4] Mr. Rahul Patel, Mr. Gaurav Sharma," A survey on text mining techniques", International Journal Of Engineering And Computer Science ISSN:2319-7242Volume 3 Issue 5 May, 2014

[5] FredPopowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing",

[6] N. Kanya and S. Geetha (2007), "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Communication Technology in Electrical Sciences, IEEE,.

[7] Seth Grimes (2005),"The developing text mining market", white paper, Text Mining Summit05 Alta Plana Corporation,,Boston, 1-12.

[8] Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship to information access. Presentation notes for UW/MS workshop on data mining, July 1997.

[9] VishalGupta,Gupreet S Lehal. " A survey of Text Mining Techniques and Applications".Journal of Emerging Technologies in web inteliignce, No.1, August 2009.

[10] Vidya k A,G Aghila, ”Text Mining Process,Techniques and Tools: an overview”,International journal of information technology and knowledge management.

[11] .Lokeshkumar,ParulkalraBhatia,”Text Mining :concepts,process and applications “ , Journal of global research in computer science ,pp.36-39,march 2013.

[12] http//: www.ai.sri.com/~appelt/ie-tutorial.

CITE AN ARTICLE

Rajamani, R., and S. Saranya. "A STUDY OF TEXT MINING METHODS, APPLICATIONS,AND TECHNIQUES." *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY* 6.7 (2017): 623-28. Web. 15 July 2017