MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 6, Issue 2, July-December 2021

A Review on Data Mining Techniques, Applications and Methods

Kavitha Reddy
Department of CSE
Vellore Institute of Technology
Corresponding Author's E-mail: *kavithag@rediffmail.com*

Abstract

Data mining is a technique for extracting meaningful information from vast amounts of data. It's utilised to carry out and solve many sorts of research projects. Text mining, web mining, picture mining, sequential pattern mining, spatial mining, medical mining, multimedia mining, structure mining, and graph mining are some of the study fields in data mining. Text mining is also known as knowledge discovery in the text (KDT) or knowledge of intelligent text analysis, and it is often referred to as text of data mining. The technique transforms unstructured data into semi-structured data, resulting in high-quality information. Text mining is the process of mechanically extracting information from a variety of textual sources and extracting new, previously unknown information using a computer. The method of text mining and approaches, tools, applications, and strategies are discussed in this study.

Keywords: *Data mining, Knowledge discovery, Information retrieval, Computer, Text mining*

INTRODUCTION

Data mining is a technique for extracting usable information from a variety of datasets. Data warehouses are great for numerical solutions, but they fall short when it comes to textual data. Text mining is also known as text data mining or knowledge discovery from textual databases because it extracts meaningful information from text data. Text mining is a technique for extracting information from unstructured material and identifying patterns that were previously undiscovered.

It's utilised to get high-quality information from a document that's been buried. It is also known as knowledge discovery from text (KDT), and it is concerned with the

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

machine-assisted analysis of semi-structured or unstructured text documents, such as emails, full-text documents, HTML files, and PDF files. Text mining is the process of analysing vast amounts of natural language text and detecting lexical patterns in order to extract relevant data. Because most information is in text format, text mining is important for organising. Text mining might contain the following steps.

- It transforms unstructured text into organised information.
- Recognize patterns in organised data.
- Using Text Mining tools, analyse the patterns.
- Take the information you need from the text.

Prediction toxicology, Protein interaction, drug development, identification of current product potentialities, detection of correlations between lifestyle and states of health, competitive intelligence, and a variety of other applications are all possible using Text Mining.

The Process of Text Mining

Document Gathering: The documents, which are available in a variety of forms, are gathered. The document might be in Word, HTML, CSS, or PDF format.

Document Pre-Processing: The provided document is processed to remove redundancies, inconsistencies, separate words, and stemming, and the documents are ready for the next phase, using the following stages:

- **Stop word removal:** It removes common words like a, an, and, of, and but.
- **Stemming:** A stem is a naturally occurring set of words that have a very similar meaning. The basis of a word is defined using this way. Inflectional and Derivational stemming methods are the two types of stemming methods.
- **Tokenization:** Recognizing a single word in a document as a string and identifying it. The provided document string is divided into a single token or unit.

Text Transformation: A list of terms and their frequency of recurrence. Bag of words and Vector space model are two methods for representing such papers.

Attribute Selection: By deleting unimportant features from the input document, this strategy results in reduced database space and a minimum search

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

methodology. Filtering and Wrapping are the two strategies for selecting attributes.

Interpretation/Evaluation: This step assesses the outcome, which can be saved or used in the following series of sequences.

TEXT MINING APPLICATION

Text mining is currently being used for a wide range of government, scientific, and corporate purposes. By analytical type or business purpose, applications can be categorised into a variety of groups.

- Enterprise Business Intelligence/Data Mining, Competitive Intelligence are some of the application areas that may be classified using this method.
- Records management and e-discovery
- Intelligence and national security.
- Scientific discovery, particularly in the field of life sciences.
- Listening Platforms, Sentiment Analysis Tools
- Semantic Toolkit or Service for Natural Language
- Automated ad placement and publishing.
- Access to information/search
- Monitoring of social media.

Some uses of text mining are described here:

a **Company Resource Planning:** The mining company's reports and correspondences for operations, so that the company's resource status and problems may be correctly addressed, and future action plans can be established.

b **Application for Security:** Many text mining tools are advertised for security purposes, including the monitoring and analysis of online plain text sources such as web news, blogs, and sites for national security tasks. The research of text encryption and decryption was also included.

c **Biomedical Application:** Text mining is used in the medical field to identify and classify technical phrases in the biological science domain that match to ideas.

d **Market Analysis:** Market analysis is involved with analysing rivals within the market and may also be used to monitor consumer opinions and seek for new prospective customers using a variety of text mining techniques.

It is primarily concerned with the management of customer messages. CRM

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

entails providing customers with relevant services based on their requests and offering quick responses to their questions.

TEXT MINING METHODS

There have been various ways established in the past to handle the problem of text mining, which is nothing more than retrieving relevant information based on the user's requirements. In terms of information retrieval, there are basically four approaches.

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

1. The Term-Based Approach

In a document, a term is a word with semantic meaning. The text is examined on the basis of terms in the term-based technique, which has the benefits of efficient computing performance and established theories for term weighting. The information retrieval and machine learning communities have developed these approaches during the previous few decades. Polysemy and synonymy are issues with term-based techniques. Polysemy refers to the fact that a word may have several meanings, whereas synonymy refers to the fact that multiple words can have the same meaning. For responding what users desire, the semantic meaning of many identified phrases is unknown. Many term-based strategies were offered by information retrieval to overcome this problem.

2. Phrase-Based Approach

Phrases have more meanings, such as information, and are less perplexing. The material is examined on a phrase-by-phrase basis in the phrase-based technique because phrases are less ambiguous and discriminative than individual words. The poor performance is likely due to the following factors: 1) Phrases have poor statistical features compared to words, 2) They have a low frequency of occurrence, and 3) They contain a large number of redundant and noisy phrases.

3. Concept-Based Approach

On a sentence and document level, the words are examined. The majority of text mining approaches rely on statistical analysis of words or phrases. The phrase frequency encapsulates the significance of a word in the absence of a document. Two terms may appear in the same paper with the same frequency, but the implication is that one phrase contributes more correctly than the other term. A new concept-based mining is presented to give greater weight

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

to phrases that represent the semantics of the text.

1. There were three parts to this model.
2. Examines a sentence's semantic structure.
3. To characterise the semantic structures, creates a conceptual ontological graph (COG).
4. The first two components, which use the conventional vector space paradigm to create feature vectors. A concept-based approach can efficiently distinguish between non-essential phrases and meaningful terms that describe the meaning of a sentence. Natural language processing methods are commonly used.

4. Taxonomy of Patterns

Documents are studied using the pattern taxonomy approach. The is-a connection may be used to organise patterns into taxonomies. Data mining techniques such as association rule mining, frequent item set mining, sequential pattern mining, and closed pattern mining can be used to find patterns. Because certain valuable lengthy patterns with high specificity lack support, using obtained information (patterns) in the field of text mining is difficult and unproductive. Pattern deploying and pattern developing are two methods used in the pattern-based methodology. This method refines the patterns seen in text materials. Pattern-based models outperform not only other pure data mining-based approaches and concept-based models, but also term-based models, according to the results.

TEXT MINING TECHNIQUES

Natural Language Processing (NLP) and Information Extraction are two of the most used text mining approaches (IE). Other text mining techniques, such as knowledge-based, statical, rule-based, and machine learning-based approaches, are being investigated. NLP is concerned with text processing, whereas IE is concerned with extracting information from text.

Natural language processing technologies are used to analyse, comprehend, and generate text. In the text mining process, methods such as information extraction, summarization, classification, clustering, and information visualisation are employed.

We'll go through each of these technologies and their role in text mining in the sections that follow. The sorts of scenarios in which each technology may be beneficial to people are also highlighted.

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

A. Natural Language Processing

Natural Language Processing (NLP) is a branch of computer science that studies how computers can read and alter natural language material. NLP researchers want to learn how people interpret and use language so that appropriate tools and techniques may be developed to help computers understand and manipulate natural languages to execute the jobs they're supposed to. NLP's foundations may be found in a variety of fields, including computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robots, psychology, and so on.

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 are all examples of NLP applications.

B. Information Retrieval

For many years, the notion of information retrieval (IR) has been established in conjunction with database systems. The association and retrieval of information from a vast number of text-based documents is known as information retrieval. Each of the information retrieval and database systems deals with different types of data. Concurrency control, recovery, transaction management, and update are some database system issues that are rarely encountered in information retrieval systems.

In addition, some typical information retrieval issues, such as unstructured documents, approximated search based on keywords, and the concept of relevance, are rarely seen in traditional database systems. Information retrieval has a wide range of applications due to the vast amount of text data available. On-line library catalogue systems, on-line document management systems, and the more recently developed Web search engines are just a few examples of information retrieval systems.

C. Information Extraction

Within the text, the information extraction approach discovers key words and relationships. It accomplishes this by searching the text for predetermined sequences, a process known as pattern matching. To provide the user with useful information, the software infers correlations between all of the listed places, people, and time. When working with vast amounts of text, this technology comes

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

extremely handy. The information being "mined" in traditional data mining is assumed to already be in the form of a relational database. Unfortunately, electronic information is only available in the form of free natural language papers, not organised databases, for many applications.

D. Topic Tracking

Yahoo (www.alerts.yahoo.com) has a topic monitoring service that tells users when news linked to a keyword becomes accessible. However, this technology has limitations. For example, if we put up an alert for "data mining," we will receive various news and stories on mineral mining rather than te E. .xt mining. In industries where firms may produce alerts to locate competitors in the news, topic tracking is often used. It keeps track of news about their own product and business.

The main procedure in topic tracking is keyword extraction. It is a set of words in an article that provides readers with a high-level explanation of the material. It's quite valuable to be able to identify keywords from a large volume of web news data. Keyword extraction by hand is complex and time-consuming. As a result, a rapid and automated keyword extraction procedure is required.

E. Summarization

The goal of text summary is to condense a document's length and content while keeping the most significant elements and overall meaning. Text summarising is useful for determining whether a lengthy document satisfies the user's needs and is worth reading for further information, therefore a summary can be used to replace a group of documents. The text summarising programme processes and summaries the enormous text document in the time it takes the user to read the first paragraph. Even though computers can recognise individuals, places, and times, teaching software to study semantics and interpret meaning of text documents is tough. Humans summarise the entire text section first, then try to grasp the full meaning, and finally provide a summary stressing the essential elements.

F. Categorization

Categorization is the process of recognising a document's key themes by placing it into a pre-defined set of topics. A computer programme will frequently treat a document as a "bag of words" when categorising it. It does not attempt to process the data in the same way that data extraction does. Rather, the categorization just counts the words that appear and, based on the counts, determines the

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

document's primary topics. Categorization frequently relies on a dictionary with established subjects, and relationships are discovered through searching for broad phrases, narrower terms, synonyms, and related terms.

G. Clustering

To locate groups of texts with similar content, the clustering method can be employed. Clustering usually produces a division termed clusters P, with each cluster containing a number of documents d. When the contents of documents inside one cluster are more similar and the contents of documents between clusters are more dissimilar, the clustering quality is deemed better. Even though clustering is a technique for grouping similar articles, it differs from categorization in that documents are clustered on the fly rather than using predefined topics in clustering. Clustering ensures that a helpful document is not removed from search results because documents can appear in several subtopics.

H. Visualization

Visualization approaches can help improve and simplify the discovery of useful information in text mining. Text flags are used to indicate individual papers or groups of documents, while density colours are used to show document categorization. Visual text mining organises enormous amounts of text into a visual hierarchy. Zooming and scaling are two ways for the user to interact with the document. The authorities can use information visualisation to identify terrorist networks or find information about crimes.

CONCLUSION

The field of text mining, often known as data mining, seeks out intriguing patterns in vast databases. The purpose of this study is to provide a brief overview of the text mining process and methods. Some real-time applications are also listed in this paper. Natural language processing, information retrieval summarization, classification, clustering, information extraction, and visualisation are some of the techniques available for extracting usable information, which fall under the area of text mining. These text mining concepts produce the best results that can be utilised in the future.

REFERENCES

1. Fred Popowich, “Using Text Mining and Natural Language Processing for Health Care Claims Processing”
2. Seth Grimes (2005),“The developing text mining market”,

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved

white paper, Text Mining Summit05 Alta Plana Corporation,,Boston, 1-12.

3. U. Y. Nahm and R. J. Mooney. “Text mining with information extraction”.In AAAI 2002Spring Symposium on Mining Answers fromTexts and Knowledge Bases, 2002.
4. G. Ercan and I. Cicekli, 2007, “Using lexical chains for keyword extraction”, published in International Journal of Information Processing and Management,.
5. Vidya k A,G Aghila, ”Text Mining Process,Techniques and Tools: an overview”, International journal of information technology and kno

Page 93-101 © MANTECH PUBLICATIONS 2021. All Rights Reserved