TEXT MINING AND TEXT ANALYTICS OF RESEARCH ARTICLES

O'rinov Nodirbek Toxirjonovich,
Teacher, Department of Information Technology,
Andijan State University
E-mail:nodirbekurinov1@gmail.com
Fozilov Murodjon Marifjon o'g'li
Master's student of computer science and programming technologies of
Andijan State University
E-mail:murodjonfozilov@gmail.com

Abstract: Recently, the number of published articles and scientific works has increased sharply. Such documents are stored in electronic format, but the data is semi-structured or unstructured. Analyzing patterns and trends is a huge task. Therefore, text mining is widely researched today.

Text Mining extracts relevant knowledge from text documents. Various text mining techniques transform unstructured data into structured data. Text classification, one of the basic principles of text analysis, requires a number of text processing techniques, the most important of which is natural language processing (NLP).

Text Mining simplifies data and is useful for researchers, scientists and scientists. Various analytical tools are used to obtain relevant and in-depth information and conclusions from the mined text. This article explores various text mining techniques and discusses recent advances in design science.

Keywords: text analysis, text analysis, scientific articles.

I. Introduction

The volume of data is growing at a daily exponential rate day by day. All kinds of organizations, enterprises and companies store their data electronically. A large amount of data is available in digital format, be it in digital libraries, repositories, blogs, email, etc. Text mining began in the fields of computing and knowledge management. Text mining is a technique of drawing interesting and meaningful patterns to extract knowledge from data sources that are textual in nature. Text mining is the process of examining data sources for their textual content and identifying meaningful patterns within them. Text mining is a multi-faceted field based on various methods and technologies such as knowledge recovery, data mining, machine learning (ML) and computer languages. This multi-dimensional field involves - data science, data mining, information retrieval, statistics and mathematical languages. The basis of text mining can be the generalization, classification and clustering of identified patterns that can lead to the desired extraction of information.

Awareness of a text usually involves a combination of interest, novelty, and curiosity. Suppose a student writing an essay on David Copperfield breaks down sentences and phrases using a text analysis machine before evaluating anything. The first step in almost all NLP functions is to separate unstructured text data into its component parts, including named entity identification, topic extraction, and sentiment analysis. [2].

Text mining is natural language text data stored in a semi-processed and unstructured data format. Text mining processes are continuously used in business, science, web applications, the Internet, and many other fields. Text mining techniques are constantly used in business, education, and web applications. It finds application in various fields such as search engines, customer relationship management (CRM), email filtering, product offering analysis, fraud detection, and social network analysis. sentiment analysis, forecasting and trend analysis [3].

II. Purpose of the article:

The huge increase in the number of published articles and research papers in recent years requires the analysis of patterns and trends in both structured and unstructured data. For this purpose, text mining is useful to researchers, scientists and scholars.

The purpose of this research work is to analyze the use of text mining techniques and explore the latest developments in the field of design science.

III. Literature review

The text mining process involves a number of steps. [4] These are the following:

a) Text pre-processing:

The pre-processing stage consists of three parts - text cleaning, token assignment and POS , that is, part-of-speech tagging.

• Cleaning up text. Cleaning up text. During this process, unnecessary and unwanted information is filtered out. This process involves filtering advertisements from web pages and normalizing text that was originally in binary format.

• Tokenization is the purpose of tokens. In this process, the text is simply broken into spaces.

• POS) tagging [5]: In this step, each generated token is assigned a word class. The tokenized text is the input to this process. Tagers have to deal with these problems.

b) Generating attributes using text transformation:

The words included in it and their appearance form a record of the text. The document is represented using either a bag of words approach or a vector space approach.

c) Variable Attribute Selection

This is also called feature selection. Here, a selected subset with important features will be used to create the model. Irrelevant or redundant features do not provide new information that would be useful.

d) Data Mining

A structured database using classical data and text mining was created in the previous steps. Various mining techniques are used, including frequent pattern set, closed pattern, sequential pattern, maximal pattern, and association rule.

e) Evaluate

The result is evaluated

Pattern detection:

D -pattern analysis algorithm and internal algorithm.

Evolution of templates.

Here the entire document will be divided into paragraphs, keeping the important terms in each paragraph. Say “bag of words,” a collection of important terms in a paragraph. Consequently, many bags of words will be created from the entire document, i.e. sets.

A subset or relationship will then be formed between terms in different paragraphs using a template taxonomy. It is part of the Knowledge Discovery in Text (KDT) program.

In the pattern expansion method, closed patterns in text mining can be formed using knowledge of the pattern's taxonomy by estimating the rotation weights. It is nothing but a relation or subset formed using the weight of the entire set of words used for pattern discovery.

Sequential pattern mining (SP) can be used to find all closed sequential patterns by reducing the search space by using a priori property. Here, the subsets with the best support threshold and confidence are useful for finding all closed sequential patterns, as this shows the greatest proximity or relationship between the text.

Internal evolution of patterns. Due to problems with low frequencies, this method proves useful in reducing the unwanted effects of noisy patterns.

Text mining methods:

They are involved in retrieving the document and obtaining an extract from it. Typically, various tools and applications are used to implement these methods. Here are some well-known text mining methods: information extraction (IE), information retrieval (IR), categorization, clustering, summarization [6].

Information extraction

It is a well-known text mining technique. Information extraction refers to the mechanism by which relevant information is extracted from large chunks of text. This text mining method mainly focuses on discovering the extraction of entities, attributes

and their relationships from semi-structured or unstructured text data. Whatever information is obtained is subsequently stored, if necessary, for future access and retrieval from the database. The validity and significance of the results are checked and assessed using precision and callback processes.

Receiving the information

Information retrieval (IR), in which similar patterns are retrieved using a given word sentence. Here, IR uses various algorithms to track and monitor user activities and thus discover related data. Google and Yahoo are two well-known search engines for IR service users.

Categorization

This is a "controlled" method in which topics are assigned to text based on content. Thus, Natural Language Processing (NLP) is a technique for collecting and processing textual data. Co-referencing is commonly used to extract synonyms and abbreviations from text. Today, NLP is used for customized market distribution, spam filtering, and web page categorization for many other purposes.

Clustering

It classifies and organizes text structures into subgroups or "clusters" for further study. The next task is to form coherent groups of text without any preconditions. This helps distribute data as well as run other algorithms in predefined groups.

Summarizing

Summarizing is nothing more than a method of automatically generating specific text in a compressed format containing consumer information. The purpose of this text mining technique is to search through various text sources to create text summaries that contain important and concise data, essentially maintaining the same essence and purpose as the original text. Data summarization includes and combines various data categorization techniques such as regression models, neural networks, swarm intelligence, and decision trees. [6]

IV. Research methodology

This is a conceptual study on how text mining can be used to simplify data and derive relevant information and conclusions from mined text. It is based on examination of various research journals, scientific articles, professional forums and other recognized research resources, including the Internet.

V. Why send text messages about my research?

Scientific works reflect human intelligence in its most complete form. The idea is not new, but provides access to a wealth of research. Until recently, newspapers were owned by a handful of firms under agreements with publishers. Recently, the open access movement has increasingly led to lowering the barriers to scientific articles for text mining. Availability of software, advances in machine learning, reduction in

memory costs, i.e. storage costs as well as computing power have reduced some technical and financial barriers.

5.1 Research Opportunities in Text Mining

5.1.1 Discovery based on literature

Literature-based discovery (LBD) [7] creates new knowledge by analyzing and combining knowledge that is already present in the available literature. Please note that there are several articles and research papers written around the world on any biomedical topic. LBD can be used by scientists in this field to work on various genes, diseases, vaccines and drugs by finding new connections between them. Swanson [8] was the first to make a discovery based on the literature by hypothesizing that the confluence of two independently reported premises, "A triggers B" and "B triggers C," refers to the connection between A and C. He found that seafood is a diagnosis of the syndrome Raynaud, focused on mutual connection. blood viscosity ratios that have been found in the literature. By examining how A is related to B and how B is related to C , LBD is typically used to prove or at least establish a hypothesis to establish a connection between A and C.

5.1.2 Other uses

Supporting research access to research literature

Currently, scientists are faced with an ever-increasing amount of written literature. Although these publications provide rich and useful information, it is difficult for researchers to process and efficiently search for relevant information due to the size of the data sets. Many modern collaborative visual analytics applications are available, such as Literature Explorer , which provide access through mining and collaborative visualization to relevant scientific literature. Some thematic subjects have a clear semantic connection with science topics that are widely used in scientific fields by researchers and hence can be interpreted by humans. They also help in successfully retrieving records. Available Visual packages Analytics Suite [9] are a set of visual components that are carefully linked to the identification of the main thematic object to enable interactive document recovery. All this helps you stay up to date with the latest research. It is also useful in analyzing, comparing and contrasting research results.

Summary of research results

Text summarization condenses information so that users can more quickly and easily recognize and understand related source text. In recent years, significant work has been done to develop and test different strategies for different areas. Text summary, or rather automated text summary, refers to the process by which a computer creates a compact version of a source text (or group of texts) but still retains most of the information present in the source text. Although this process may result in some loss of information, it is very useful in data compression. For example, in the biomedical

field, various research materials are available and the synthesis of these data is very useful for researchers.

Automation of a systematic literature review

Systematic assessments are carried out using a robust but slow, resource-intensive mechanism. As a consequence, conducting a systematic review may require significant amounts of money and may take years to complete. Text mining and text mining prove to be very useful in this field as it requires less time and money. The snowball technique is very useful for this.

Understanding the research impact of articles, individuals, institutions, countries

Database tomography is a method of information retrieval and interpretation that works with text databases. Its primary application to date has been to identify ubiquitous technology trends and topics, and the relationships between these topics and subtopics, that are inherent in large text databases. This method can be used to work with large databases around the world to make inferences about the impact of articles, individuals, institutions and countries.

Monitoring Research Trends

Text mining and natural language technology analysis have been used to categorize what researchers are looking for and evaluate current research efforts. The combination of trend analysis and query clustering will result in different priorities. a systematic concept analysis (FCA) approach to create a dynamic patent framework capable of assessing complex patent relationships and tracking patterns of research trends.

Evidence of reuse and plagiarism detection

Latent semantic analysis (LSA) [10] is a technique used to analyze the relationships between a set of documents and the terminology they contain in natural language processing (NLP), especially in distributional semantics, where a set of concepts related to documents is generated. and conditions. LSA assumes that identical pieces of text (distribution hypothesis) will include terms that are close in context. As reuse and plagiarism are becoming a growing problem in academia. Text mining and text mining form the basis of plagiarism detection tools that are often used by academics.

VI. Conclusion:

With the sharp increase in global digitalization, the number of documents has grown like an explosion. Therefore, text classification is necessary to classify documents based on their content according to predefined classes. This research article presents an analysis of the use of text mining techniques to track recent developments in the field of design science. The result also indicates that the developed methods are universal and can be extended to handle knowledge in various fields of learning.

Additionally, the text mining techniques used in this study can help researchers gain deep insight into the domain-specific expertise hidden in the vast body of scientific literature. Selecting and using the right domain-specific methods and tools helps make the text extraction process simple and efficient. Integration of domain information, different granularity principles, multilingual text refinement and ambiguity in natural language processing are the main problems and challenges that arise during the text extraction or mining phase. In the future, various design algorithms will be useful in solving various problems in the field of text mining.

Recommendations:

[1] D. Milward, "Lingvamatics", 2020. [Online]. Available: https :// www . linguamatics. com / what - text - mining - text - analytics - and - natural - language - processing.

[2] T. Mohler, "Lexalitics", September 9, 2019 [Online]. Available: https :// www. lexalytics. com / lexablog / text - analytics - functions - explained.

[3] R. Talib, "Text Mining: Methods, Applications and Challenges", International Journal of Advanced Computer Science and Applications (IJACSA), 2016.

[4] Dataflair, “Data Flair,” September 21, 2018. [Online]. Available: https://data-flair.training/blogs/text-mining/.

[5] V. B. Kobayashi, “Text Mining in Organizations,” SAGE, 2018. A. Rai, “Upgrad,” June 1, 2019 [Online]. Available: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/.

a. M. Yetisgen-Yildiz, “A new methodology for evaluating literature-based discovery systems,” Journal of Biomedical Informatics, 2009. Korhonen, “Improving Literature-Based Discovery,” Springer International Publishing, Switzerland, 2015.

[6] S. Wu, “Literary Explorer: Efficient Scientific Document Retrieval through Nonparametric Topic Discovery,” Springer, 2019.

[7] H. Yalcin, “Exploring the Technology and Engineering Management Research Landscape,” IEEE, 2019.

[8] Akshaya Udgave, Prasanna Kulkarni: Text Mining and Text Analytics of Research Articles - Palarha Journal of Egyptian Archeology/Egyptology 17(6). ISSN 1567-214x