PalArch's Journal of Archaeology
of Egypt / Egyptology

TEXT MINING AND TEXT ANALYTICS OF RESEARCH ARTICLES

Akshaya Udgave[1], *Prasanna Kulkarni*[2]

Symbiosis Institute of Digital and Telecom Management,

Symbiosis International (Deemed University), Pune, India.

Email: pkulkarni@sidtm.edu.in

Akshaya Udgave, Prasanna Kulkarni: Text Mining and Text Analytics of Research Articles -- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(6). ISSN 1567-214x

Keywords: Text Mining, Text Analytics, Research Papers.

ABSTRACT

There has recently been a tremendous increase in published articles and research papers. Such documents are stored in electronic format, but the data is semi-structured or unstructured. The analysis of the patterns and trends is an enormous task. Text mining is therefore extensively researched today.

Text Mining extracts appropriate knowledge from the text documents. Different text mining techniques convert unstructured data to structured data. Text classification, one of the basic principles of text mining, requires a number of techniques for processing text, the most important being Natural Language Processing (NLP).

Text Mining simplifies data and is useful to researchers, scientists, and academics. Various analytics tools are used for deriving relevant and in-depth information and inferences from the mined text. This paper studies various text mining techniques, and discusses recent advances in the field of design science.

1. Introduction

Data size is rising at daytime exponential levels day by day. All sorts of organizations, enterprises and companies store their data in electronic form. A large amount of data is available in digital format be it in the digital libraries, repositories or blogs, emails, etc. Text mining started with the fields of computing and knowledge management. Text mining is a technique for drawing engaging and meaningful patterns to traverse knowledge from data sources textual in nature. Text mining is a process to explore the data sources for its textual content and gauge a meaningful pattern in them. Text mining is a multifaceted field based on various techniques and technologies like knowledge recovery, data mining, machine learning (ML), and computing languages. This multi-dimensional field touches upon - data science, data mining, retrieval of information, statistics, and also the mathematical languages. The basis of text mining can be put down as - summarizing, classifying, clustering of the identified patterns which can lead to the desired extract of information.

Text awareness typically refers to a mix of interest, novelty, and curiosity. Suppose a student who writes an essay on David Copperfield breaks down sentences and phrases through a text analytics machine, before evaluating something. The first step in almost every NLP functionality is to separate unstructured text data into their component parts including identification of named entities, extraction of themes, and analysis of sentiments. [2].

Text mining is about natural language textual data that is stored in semi-designed and unstructured data format. Text mining processes are used uninterruptedly in business, academics and the web applications, on the internet and many other areas. The text mining techniques have been in constant use in businesses, education, and in the web applications. It finds application in various fields such as search engines, customer relationship management (CRM) system, email filtration, analysis for suggestion of products, detection of fraud, and social media analytics use text analytics and mining for mining of opinions, extraction of features, sentiment, predictive, and trend analysis [3].

2. Objective of the Paper:

Tremendous increase in published articles and research papers in recent years, requires analysis of the patterns and trends in the structured as well as unstructured data. Text mining is useful to researchers, scientists, academics for this purpose.

The objective of this research paper is to analyse the use of text mining techniques, and to explore recent developments in the field of design science.

3. Literature Review

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

a. Text Pre-processing:

The pre-processing step consists of 3 parts - a cleaning up of text, assigning the tokens, and POS i.e. part-of-speech tagging.

- Text Cleanup - A clean-up of text: During this process, Unnecessary and unwanted information is filtered out. The process comprises of filtering out the advertisements from the web pages, normalizing the text which was originally in binary format.
- Tokenization - Assigning the tokens: In this process, the text is simply split into white spaces.

- Part-of-speech (POS) tagging [5]: This step assigns a word class to each token formed. The tokenized text is the input for this process. Taggers have to handle these issues.

b. Generation of Attributes using Text Transformation:

The words it includes and the occurrences thereof form a record of text. A document is represented by using either a bag of words approach or vector space approach

c. Variable Selection of Attributes

This is also called as feature Selection. Here, the subset selected with important features would be used for model creation. Irrelevant or redundant features provide no new information which would be useful.

d. Mining of Data

Structured database using classic data and text mining was produced in previous steps. Various mining techniques are used including frequent item set, closed pattern, sequential pattern, maximum pattern, and association rule.

e. Evaluate

Evaluation of the result is done

.

Pattern Discovery:

Different algorithms which are used are Pattern Taxonomy Model, Pattern Taxonomy, Pattern Deploying Method, D-Pattern Mining Algorithm and Inner

Pattern Evolution.

Here, the entire document would be split into paragraphs and maintaining important terms in each of the paragraphs. Say the bag of words – which is a set of important terms in a paragraph. Hence, many bags of words i.e. sets would be made from the entire document.

Then, a subset or relation would be formed between the terms in different paragraphs using Pattern Taxonomy. This is the part of Knowledge Discovery in Text (KDT).

In the pattern deploying method, closed patterns in text mining could be formed using knowledge in the taxonomy of the pattern, by evaluating the turn weights. This is nothing but the relation or subset formed using weightage of all bag of words is used for discovering the pattern.

Sequential Pattern (SP) Mining could be used to search all closed sequential patterns by reducing the searching space by using the a priori property. Here, the subsets with the best support threshold and confidence are useful to search all the closed sequential patterns as this shows the most closeness or relations between the text.

Inner Pattern Evolution - Due of the low frequency issue, the technique turns out to be beneficial in reducing the undesired impact of noisy patterns.

Text mining techniques:

These are involved in extracting the document and taking the extract from it. In general, these techniques utilize various tools and applications for their implementation. A few famous text mining methods are: Information Extraction (IE), Information Retrieval (IR), Categorization, Clustering, Summarization [6].

Information Extraction

This is the well-known method of text mining. Information extrication points to the mechanism by which relevant information is extracted from large bites of text. This text mining method mainly focuses upon the detection of the extraction from partially-structured or unstructured textual data of entities, attributes and their relations. Whatever information is retrieved, is thereafter stored, if needed, for accessing and retrieving from the database in the future. The validity and significance of the results is tested and evaluated by using processes of precision and summoning back.

Retrieval of Information

Information retrieval (IR) by which similar patterns are retrieved using a given word sentence. Here, IR uses various algorithms for tracking and controlling the user activities and thus discovering the related data. Google and Yahoo are the two well-known IR services user search engines.

Categorization

This is a "supervised" technique in which text is assigned topics based on content. Natural Language Processing (NLP) is therefore a method of collecting, processing evaluating text data. Co-referencing is commonly used to extract synonyms and acronyms from the text. Today, NLP is used in customized market distribution, filtering of spam and categorization of web pages for many other purposes.

Clustering

It classifies and organizes text structures for further study of subgroups or 'clusters'. The next task is to form coherent groups of text while having no preconditions. This helps in allocating data, and also in running other algorithms on pre-determined groups.

Summarization

Summarization is nothing but a method of automatic generation of a specific text in a compressed format which comprises consumer information. The goal of this text mining method is to search over various sources of text to create text summaries containing a significant and succinct data, essentially retaining same substance and purpose as the initial text. Data summarization incorporates and blends the different techniques of categorizing data, such as regression models, neural networks, swarm intelligence and decision trees. [6]

4. Research Methodology

This is a conceptual research paper exploring how text mining can be used to simplify data and for deriving relevant information and inferences from mined text. It is based on study of various research journals, scholarly articles, professional forums, and other recognised research resources including those on the internet.

5. Why to text mine research papers?

Research papers reflect human intelligence in its most complete form. The idea is not new but provides access to vast quantities of study. Until recently the papers were owned by a handful of firms by conventions with publishers. Recently, the Open Access movement has increasingly led to reducing the barriers to academic papers for text-mining. The availability of software's, advances in machine learning, reduction in costs of memory capacity i.e. storage cost along with computing power has reduced some technical and financial hurdles.

5.1 Opportunities for text mining research papers

5.1.1 Literature based discovery

Literature based discovery (LBD) [7] creates new knowledge by analyzing and merging from the knowledge which is already present in the available literature. Consider that there are several articles and research papers written on any biomedical topic across the world. LBD could be used by scientists in that field to work on different genes, illnesses, vaccines and drugs by finding new connection between them. Swanson [8] pioneered literature-based discovery, hypothesizing that the fusion of two independently reported "A triggers B" and "B causes C" premises refers to the relation between A and C. He found seafood as Raynaud's syndrome diagnosis focused on the mutual relations of the blood viscosity which was 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 for establishing relationship between A and C.

5.1.2 Other use cases

Supporting exploratory access to research literature

Nowadays, scientists face an increasingly increased quantity of written literature. Although these publications provide rich and useful information, it is arduous for researchers to handle and search effectively for required information by the size of the datasets. Many modern collaborative visual analytics applications like Literature Explorer that enable the access through mining and collaborative visualization to desired scientific literature are available. Some thematic subjects have a clear semantic connection with the themes of science that are widely used in scientific fields by human researchers, and are thus humanly interpretable. These also aid in the successful retrieval of records. The available Visual Analytics Suites [9] are a collection of visual components that are meticulously linked to the identification of the underlying thematic subject to allow interactive recovery of documents. All this helps in staying up-to-date with research. Also, it is helpful in analyzing, comparing and contrasting research findings.

Summarization of research findings

Text summarization reduces information in an effort to allow users to quicker and more easily recognize and understand related source texts. Substantial work has been carried out in recent years to establish and test various strategies for the various domains. Text summary or rather automated text summary refers to the process by which a computer produces a compact version of the original text (or a group of texts) but still retaining the majority of the information present in the original text. Although this process may incur some information loss, it is highly useful in compression of data. For example, in biomedical domain, various research materials are available and summarizing this data is highly useful to researchers.

Systematic literature review automation

Systematic assessments are performed through a reliable but slow mechanism with a high resource strength. As a consequence, conducting a systematic review could require a substantial amount of money and may take years. Text mining and text analytics prove to be very helpful in this domain as it takes lesser amount of time and money for this. Snowballing technique is highly useful for this.

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

Database tomography is a method of retrieval and interpretation of information that works on textual databases. Its key use to date has been to define omnipresent technological thrusts and themes, and the interrelationships between these themes and sub-themes, which are intrinsic to broad textual databases. This method could be used to work on large databases across the world to draw conclusions over impact of articles, individuals, institutions, countries.

Monitoring research trends

Text mining and analysis of natural language technologies have been used to classify what researchers are searching for and evaluate current research works. A combination of the trend analysis and clustering of queries would lead to forming different priorities. an approach based on systematic concept analysis (FCA) to build a dynamic patent lattice capable of evaluating 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 collection of documents and the terminologies they contain in natural language processing (NLP), especially distributional semantics, where a set of concepts is produced relating to the documents and terms. LSA suggests that identical pieces of text (the distributional hypothesis) would include terms that are close in context. As reuse and plagiarism is becoming a growing problem in academia. Text mining and text analytics form the basis of the plagiarism detection tools which are often used by academics.

6. Conclusion:

With the dramatic rise in world digitization, the number of documents has been on increase as though an explosion. Text Classification is therefore required to classify the documents based on their content according to the predefined classes. This research paper is the analysis on the use of text mining techniques to keep track of recent developments in the field of design science. The result also indicates that the methods developed are universal and could be extended to handle the knowledge of different fields of study. In addition, the text mining techniques used in this study may help researchers gain a thorough understanding of the expertise of a specific field concealed in a vast amount of scientific literature. Choosing and using the right domain-specific techniques and tools helps make the process of extracting text easy and efficient. Integration of domain information, varying granularity principles, refinement of text in multilingual type and ambiguity in the handling of the natural language are major problems and challenges that emerge throughout the text extraction or mining phase. In the future, different design algorithms would be helpful in resolving various issues in the text mining field.

References

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

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

R. Talib, "Text Mining: Techniques, Applications and Issues," International Journal of Advanced Computer Science and Applications (IJACSA), 2016.

DataflairTeam, "Data Flair," 21 September 2018. [Online]. Available: https://data-flair.training/blogs/text-mining/.

V. B. Kobayashi, "Text Mining in Organizational," SAGE, 2018.

A. Rai, "Upgrad," 1 June 2019. [Online]. Available: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/.

M. Yetisgen-Yildiz, "A new evaluation methodology for literature-based discovery systems," Journal of Biomedical Informatics, 2009.

A. Korhonen, "Improving Literature-Based Discovery," Springer International Publishing Switzerland, 2015.

S. Wu, "Literature Explorer: effective retrieval of scientific documents through nonparametric thematic topic detection," Springer, 2019.

H. Yalcin, "Exploring Technology and Engineering Management Research Landscape," IEEE, 2019.