International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Review on Text Mining

[1]Anupam Lakhanpal
[1]Department Of Computer Science and Engineering, Galgotias University, Yamuna Expressway Greater Noida, Uttar Pradesh
[1]anupam.lakhanpal@Galgotiasuniversity.edu.in

Abstract: Rapid developments in automated data processing methods have contributed to tremendous data volume. More than half of the information currently is made up of unorganized or semi organized information. A major problem is the retrieval of common trends and patterns to display textual information from large amounts of data.Text mining is a method of collecting unique and computationally expensive trends from vast quantities of text data. There are many strategies and resources to explore the text documents and explore the detail in policy-making for the potential and operation.Selecting the right and correct text mining methodology allows regain the pace and slows down the time and energy needed to obtain useful data.Through technological innovation, greater and greater information is being distributed in electronic form.Among many, the bulk of the information is in an unorganized structured type.Therefore, it is becoming crucial to develop stronger methods and methodologies in order to obtain interesting and relevant information from a significant amount of conceptual information. Therefore, the area of data mining and text mining have turned out to be the popular analysis fields, in order to obtain unique and necessary information.

Keywords: Information Extraction, Knowledge, Patterns, Text Mining

INTRODUCTION

Text Mining:

Information size is growing step by step at unprecedented rates.Nearly all types of agencies, organizations, and economic sectors store their information digitally.In the form of online databases, archives, as well as other digital material like articles, social networks and e-mails, a great deal of information flow over the web[1].Determining suitable behavioural patterns to derive useful knowledge from such vast quantities of data is a difficult task. Current data gathering techniques are unable to manage text information, as time and energy are needed to obtain information.Text mining is characterized as the isolation of secret and possibly important data from structured information. Text mining is a new field that seeks to obtain useful data from text content that is normal. It can be built as the movement of text analysis to isolate the information needed for a particular purpose.Text mining" is being used to explain how data mining methods are applied to automatically uncover interesting or important information from unorganized data.Many strategies for text mining have been suggested, like logical framework, "association rule mining, episode rule mining, decision trees, and methods of rule induction". Text mining is a method of extracting relevant and important trends from textual sources to investigate information[2]. Text mining is an interdisciplinary field focused on the extraction of data, data mining, artificial intelligence, analytics, and digital linguistics.Text mining works with human language text that is processed in a semi-organized format[3]. Text mining strategies are implemented constantly in business, education, web apps, the internet and other industries.Technology fields such as browsers, client relationship information systems, scan emails, product recommendation analytics, fraud prevention, and social networking evaluation use "text mining for opinion mining", feature retrieval, feeling, analytical, and pattern assessment.Nevertheless, the text is by far the most competitive form for the structured sharing of information in today's society. Text mining works with messages whose purpose is to convey actual information or views.Text mining is close to information extraction, with the exception that data mining techniques are built to use organized information from servers, text mining may also work in fields of unorganized or non-structured databases such as emails, text files and HTML documents etc. Text mining thus has a much better alternative.Most specifically, the term ―text mining is used to describe any program that analyses large quantities of real language content and identifies strategies of syntactic or textual use in an attempt to obtain the required information[4].The Text Mining process starts with the collection of documents via additional resources.A specific report would be obtained by means of a text mining device and its layout and character sets checked; this device would pre-process it.The report will then go via a stage of evaluation of the text.Various methods may be used depending on the goal of the company. In certain cases, methods of text analysis are replicated till the details are collected. The findings can be processed in a "management information system", which presents the operator of that scheme with a large quantity of essential information.Text mining aims to identify the previously unrecognized details by systematically retrieving it from different text-based databases.Text mining is gaining significant significance in analysis due to the growing requirements to acquire understanding from a significant number of textual documents available on the web. The Overview of Text Mining is shown below in Fig. 1 Overview of Text Mining

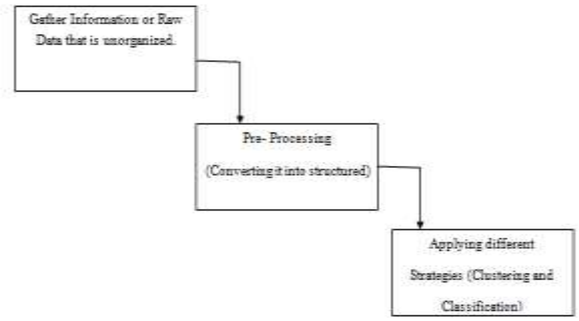
Gather Information or Raw Data that is unorganized.
Pre- Processing
(Converting it into structured)
Applying different Strategies (Clustering and Classification)

Figure 1: Overview of Text Mining

PROCESS OF TEXT MINING

The process of Text Mining involves the following steps-

Text Pre- Processing:

Text pre-processing is an important part of any NLP process since the words, phrases, and phrases defined at this point are the basic components transferred via implementations like knowledge recovery and machine translation frameworks to all other processing stages, including evaluation and labelling elements.It is a set of operations where pre-processing of Text Documents takes place[5]. Because text information sometimes includes several unique styles like number genres, date formats as well as the most prevalent phrases that may not assist Text mining such as verbs, reports, and pro-nouns. Text Pre- Processing involves the following steps-

Tokenization:

Tokenizing is done clearly by breaking the text into blank spaces and at punctuation not belonging to the acronyms found in the previous step.Tokenization is the method of cracking a textual flow into sentences, words, signs, or any other useful components termed tokens. The purpose of tokenization is to explore the terms in a phrase. For more analysis like sorting or text mining, the collection of tokens represents data[6].All information retrieval procedures involve the terms of the data set. The prerequisite for a translator is, therefore, report tokenization. As the document is already processed in computer-readable forms, it may sound strange.

Stop Word Elimination:

Stop words are used to delete insignificant terms, enabling users to concentrate instead on keywords. And using a specified collection of stop words is fairly simple, in many situations, using these stop words is entirely inadequate for some implementations.Stop words are generally accepted as not adding to the meaning or quality of textual records. Their involvement in text mining creates a challenge in interpreting the quality of the reports because of the high rate of appearance.The operation also eliminates text information and improves the accuracy of the device. Text document works with those terms that are not needed for text mining purposes.

Stemming:

Stemming is the method by which the alternative types of a word are merged into a specific representation, the core. For instance, the terms: "presentation," "presentation," "presentation" can all be condensed to a "present" popular interpretation.It is a commonly utilized method in information retrieval (IR) text extraction based on the premise that uploading a question with the promoting term means a value in reports comprising the "presenting and presenting "terms.Over-stemming is when you branch two terms with separate branches to the same root. This is often called a “false positive”. Under-stemming is when there are no two terms to that same root which should be trimmed in. This is often called a “false negative”.

Text Transformation (Attribute Generation):

The words (functions) it includes and the instances constitute a text document. Word bag and Vector Space are two key methods to text presentation.

Feature Selection (Attribute Selection):

The method of determining a selection of important characteristics to be used in model development is often recognized as variable selection.Using a filtering strategy for functionality, the key presumption is that the information includes several duplicate or unnecessary characteristics[7].Obsolete apps are the one that does not provide any additional details.In no background does insignificant functionality provide any important or useful information. The function selection strategy is a branch of the more general area of selection of data.

Data Mining:

The text mining method is followed by the conventional data mining method in this view. Traditional Data Mining strategies are used in the organized database that gives the previous phases results.Text mining requires an additional step, thus retaining the same analytical objective as data mining.Text mining works with complex data so that the unorganized data must be structured and configured in a manner that allows data processing and analysis to happen before data processing or information processing method can be implemented.

Evaluate:

Test the outcome for the accuracy, the result may be excluded after the modifications or the outcome produced can be used as a reference for the next collection of series.Determine the result, the outcome may be rejected after analysis or the outcome produced may be used as a reference for the next collection of series. The process of Text Mining is shown below in Fig. 2 Process of Text Mining

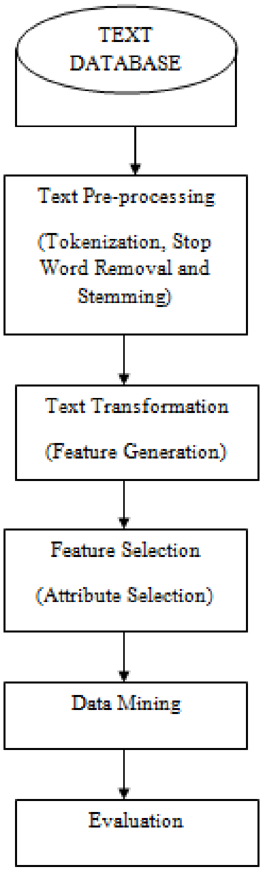

Figure 2: Process of Text Mining

FIELDS OF TEXT MINING

Text analysis requires data recovery, retrieval of data, data mining strategies including analysis of connections and links, simulation and predictive evaluation.The objective is basically to transform the content (unorganized information) into information (organized layout) for examination, using the techniques of natural language processing. The various field of Text Mining are-

Data Mining- Data mining can be characterized broadly as seeking trends in the data.It can be defined more completely as extracting information from concealed, previously unseen, and valuable information.Data mining techniques can forecast patterns and trends of the potential, enabling companies to make optimistic, information-based decisions.Data mining software may respond to industry problems which have historically taken too long to address.There is a need to look for secret and hidden trends in repositories, discovering important information that specialists may overlook as it lies beyond the standards[8].The overall purpose of the data mining method is to obtain information from a database and turn it into a comprehensible framework for further utilization.

Information Retrieval- Recovery of information is assigned as a complete form for document recovery in which the records are recovered and stored in order to compress or obtain the user's specific information.A text summary phase, which concentrates on the person's request presented, or a data extraction phase utilizing strategies, can then be accompanied by data processing.Information Retrieval programs aid in narrowing down the collection of documents related to a specific issue. As text mining continues to introduce very complex techniques to large collections of texts, Information retrieval can greatly accelerate the research by reducing the number of papers for examination.

Natural Language Processing-Natural Language Processing is among the latest and most difficult machine learning issues in the region. It is natural language analysis so that machines can comprehend natural languages as human beings do[9]. NLP analysis seeks a broad question as to how we view a phrase or a document's context.

Information Extraction -Information Extraction is the activity of systematically retrieving organized data from "machine-readable unstructured and/or semi-structured" files. In most situations such task involves decoding documents of the human speech through natural language processing (NLP).Recent developments in digital information processing such as automated editing and image/audio/video extraction data could be interpreted as knowledge mining and Search Engine is the strongest functional and active illustration of Information Extraction.

CHALLENGES

During the text mining process, several problems happen and impact decision-making operational efficiency.At the intermediate level of text mining, difficulties can emerge.Different guidelines are described in the pre-processing level to optimize the message which makes the extraction process productive.There is a need to turn unorganized information into intermediary type before implementing pattern recognition to the text, however, at this point; the extraction process will have its own difficulties. Real theme or data often misleads its value due to the change in the order of the text.Another key issue is refining reliance on the multilingual code that causes issues.There are a few methods available that allow multiple translations.Different methods and strategies are employed to help multiple languages text separately. Such problems create a lot of challenges in the phase of information creation and decision making. Infect real advantage is hard to achieve by using the current text mining tools and methods because its multiple language texts are rarely assisted[10].Domain information incorporation is a key area, as it executes different activities on a given domain and achieves expected results.In these cases, knowledge of the context from which report repository to be derived needs to be combined with the computational capabilities from which knowledge must be obtained.The use of adjectives, synonyms in the records creates problems for the text mining equipment that bring the same meaning into both. When information collection is wide and produced from various fields sharing the very same scope it is difficult to classify the records.Multiple computational complexity definitions change the meaning of the text as per the awareness of the situation and field. There's a need to define rules by field to be used in the region as a norm and can be incorporated as a "plug-in in text mining tools."

APPLICATIONS

Web Mining- Nowadays the internet includes a lot of knowledge regarding things like people, businesses goods, etc. that can be of great concern.Cloud Mining is an essential use of cloud-based machine learning strategies to identify secret and unidentified trends.Web mining is an essential activity that involves the identification of words suggested in a broad set of documents.The first approach to any web-based text processing project would be to accumulate a large number of online pages with topic observation. Instead, the problem becomes not only of finding all the innovations in the subject but also of separating those with the desired context.

Clustering- Clustering is an unmonitored method by using various clustering techniques to identify the text documents into groups? In a group, the same explanations or prototypes from different texts are clustered together. Clustering is achieved in action from “top-down and from bottom up”. In Natural Language Processing, the focus on unorganized content uses several types of extraction tools and methods. Broad clustering approaches include "distribution, density, and centroid, hierarchical and k-mean".

Social Media-For the analysis of social media apps, text mining application features are available to track and interpret the digital plain text from online news, forums, email, etc. Text mining applications help to determine the amount of Social Media messages, comments, and supporters. This kind of observation demonstrates the reactions of people on various comments, media and how well it gets passed about.

Business Intelligence-Text mining performs a key role in business intelligence, helping companies and corporations evaluate their clients and rivals to make good choices.This offers a greater insight into the market and provides data about how to boost customer experience and achieve economic advantages.It also assists in projects in the telecommunications industry, market and trade, and client quality management network.

Life Science-Companies of life science and medicine are producing large amounts of digital and quantitative data concerning patient records, illnesses, medications, disease signs and therapies, and much more.Filtering a suitable and appropriate message to take a decision from a large genetic database is a big challenge.The health records involve differing in design, using complicated, long and complex language which makes the process of information exploration very difficult.

CONCLUSION

To obtain useful information, consideration needs to be given to the accessibility of large volume of information-based data. Text mining strategies are used to quickly and effectively analyze the important and relevant data from massive amounts of data. The paper offers a brief summary of the extraction strategies of data that greatly improve the extraction process of the message. In order to gain and obtain valuable information, unique trends and combinations are implemented and used by overlooking irrelevant specifics for statistical analysis. The collection and use of appropriate domain-specific methods and techniques can make the text mining process easier and safer. Integration dependent on domain expertise, granularity dependent definitions, multiple languages text of sophistication, and the use of uncertainty in the production of natural language is significant problems and threats that occur throughout text mining strategies. In addition, the use of accurate text mining tools in the healthcare field helps to measure the efficacy of medical procedures showing positive efficacy by evaluating illnesses, effects. Text mining is of huge advantage in life research and medical care, more than almost any other sector.

REFERENCES

[1] C. C. Aggarwal and C. X. Zhai, “An introduction to text mining,” *Mining Text Data*, vol. 9781461432234. pp. 1–10, 2013, doi: 10.1007/978-1-4614-3223-4_1.

[2] C. C. Aggarwal and C. X. Zhai, *Mining text data*. 2013.

[3] R. Jetley, “Text mining,” *ABB Review*. 2018, doi: 10.4018/978-1-59904-857-4.ch054.

[4] C. Chen, M. Song, C. Chen, and M. Song, “Text Mining with Unstructured Text,” in *Representing Scientific Knowledge*, 2017, pp. 223–261.

[5] A. S. Nayak and A. P. Kanive, “Survey on Pre-Processing Techniques for Text Mining,” *Int. J. Eng. Comput. Sci.*, 2016, doi: 10.18535/ijecs/v5i6.25.

[6] V. S and J. R, “Text Mining: open Source Tokenization Tools – An Analysis,” *Adv. Comput. Intell. An Int. J.*, 2016, doi: 10.5121/acii.2016.3104.

[7] H. Hashimi, A. Hafez, and H. Mathkour, “Selection criteria for text mining approaches,” *Comput. Human Behav.*, 2015, doi: 10.1016/j.chb.2014.10.062.

[8] C. C. Aggarwal, *Data Mining: The Textbook*. 2015.

[9] I. D. Dinov and I. D. Dinov, “Natural Language Processing/Text Mining,” in *Data Science and Predictive Analytics*, 2018, pp. 659–695.

[10] A. Akilan, “Text mining: Challenges and future directions,” 2015, doi: 10.1109/ECS.2015.7124872.

[11] S Balamurugan, RP Shermy, Gokul Kruba Shanker, VS Kumar, VM Prabhakaran, “An Object Oriented Perspective of Context–Aware Monitoring Strategies for Cloud based Healthcare Systems”,Asian Journal of Research in Social Sciences and Humanities, Volume : 6, Issue : 8, 2016

[12] S Balamurugan, P Anushree, S Adhiyaman, Gokul Kruba Shanker, VS Kumar, “RAIN Computing: Reliable and Adaptable Iot Network (RAIN) Computing”, Asian Journal of Research in Social Sciences and Humanities, Volume : 6, Issue : 8, 2016

[13] V.M. Prabhakaran, Prof S.Balamurgan ,A.Brindha ,S.Gayathri ,Dr.GokulKrubaShanker,Duruvakkumar V.S, “NGCC: Certain Investigations on Next Generation 2020 Cloud Computing-Issues, Challenges and Open Problems,” Australian Journal of Basic and Applied Sciences (2015)

[14] Usha Yadav, Gagandeep Singh Narula, Neelam Duhan, Vishal Jain, “Ontology Engineering and Development Aspects: A Survey”, International Journal of Education and Management Engineering (IJEME), Hongkong, Vol. 6, No. 3, May 2016, page no. 9 – 19 having ISSN No. 2305-3623.

[15] Vishal Assija, Anupam Baliyan and Vishal Jain, “Effective & Efficient Digital Advertisement Algorithms”, CSI-2015; 50th Golden Jubilee Annual Convention on “Digital Life”, held on 02nd to 05th December, 2015 at New Delhi, published by the Springer under ICT Based Innovations, Advances in Intelligent Systems and Computing having ISBN 978-981-10-6602-3 from page no. 83 to 91.

[16] Vishal Jain and Dr. S. V. A. V. Prasad, “Analysis of RDBMS and Semantic Web Search in University System”, International Journal of Engineering Sciences & Emerging Technologies (IJESET), Volume 7, Issue 2, October 2014, page no. 604-621 having ISSN No. 2231-6604.