TEXT MINING: AN UNSTRUCTURED DATA MINING CONCEPT, USE, CHALLENGES AND FUTURE DIRECTION

[1]Dr. Vishal H. Bhemwala,[2]Dr. Kirit I. Chokhawala,[3]Dr. Jayesh N. Modi

[1]Assistant Professor,[2]Assistant Professor,[3]Assistant Professor,
[1]Department of Computer Science,
[1]Hemchandracharya North Gujarat University, Patan, Gujarat, India.

Abstract: Text mining is also known as text data mining or knowledge discovery process. There are actually structured and unstructured Data Mining Techniques. Structured Data Mining Techniques use the concept of relational database as input or first point of the knowledge discovery process. There is well defined structure and everything seems to be perfectly planed for the knowledge discovery. But Text mining has not structured source of input. The data arrives from various unstructured sources and there is no universal mechanism or formation, so mining process can be done in efficient way. In this paper, the focus is on the architecture, challenges and use and future direction of Text mining. Though, the data are not properly organized but if we can form architecture very important information can be derived and this will make useful in decision making or discovering the new knowledge. There is huge amount of information is stored in article, books and journals. Only proper algorithm is required to handle this information. The text mining serves the purpose of discovering the knowledge which is previously unknown.

IndexTerms - **Web Mining, Data Mining, Web Content Mining, Web Usage Mining, Web Structured Mining and Text Mining.**

I. INTRODUCTION

Text mining, also referred to as text data mining, similar to text analytics, is the process of deriving high-quality information from text. It involves "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." Written resources may include websites, books, emails, reviews, and articles. High-quality information is typically obtained by devising patterns and trends by means such as statistical pattern learning. We can differ three different perspectives of text mining: information extraction, data mining, and a KDD (Knowledge Discovery in Databases) process. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interest. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP), different types of algorithms and analytical methods. An important phase of this process is the interpretation of the gathered information. A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted. The document is the basic element while starting with text mining. Here, we define a document as a unit of textual data, which normally exists in many types of collections.☑

The term text analytics describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation. The term is roughly synonymous with text mining; also describe as "text analytics". The latter term is now used more frequently in business settings while "text mining" is used in some of the earliest application areas, dating to the 1980s, notably life-sciences research and government intelligence. The term text analytics also describes that application of text analytics to respond to business problems, whether independently or in conjunction with query and analysis of fielded, numerical data. It is a truism that 80 percent of business-relevant information originates in unstructured form, primarily text. These techniques and processes discover and present knowledge – facts, business rules, and relationships – that is otherwise locked in textual form, impenetrable to automated processing. Text mining computer programs are available from many commercial and open source companies and sources. Few names are Carrot2, GATE, SAS, MATLAB, RapidMiner and many more software for text mining these days available.

II. TEXT ANALYSIS PROCESS

THE STEPS FOR TAX ANALYSIS PROCESS ARE EXPLAINED AS BELOW:

1. DIMENSIONALITY REDUCTION IS IMPORTANT TECHNIQUE FOR PRE-PROCESSING DATA. TECHNIQUE IS USED TO IDENTIFY THE ROOT WORD FOR ACTUAL WORDS AND REDUCE THE SIZE OF THE TEXT DATA.
2. Information retrieval or identification of a corpus is a preparatory step: collecting or identifying a set of textual materials, on the Web or held in a file system, database, or content corpus manager, for analysis.
3. Although some text analytics systems apply exclusively advanced statistical methods, many others apply more extensive natural language processing, such as part of speech tagging, syntactic parsing, and other types of linguistic analysis.

4. Named entity recognition is the use of gazetteers or statistical techniques to identify named text features: people, organizations, place names, stock ticker symbols, certain abbreviations, and so on.
5. Disambiguation—the use of contextual clues—may be required to decide where, for instance, "Ford" can refer to a former U.S. president, a vehicle manufacturer, a movie star, a river crossing, or some other entity.
6. Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail addresses, and quantities (with units) can be discerned via regular expression or other pattern matches.
7. Document clustering: identification of sets of similar text documents.
8. Co reference: identification of noun phrases and other terms that refer to the same object.
9. Relationship, fact, and event Extraction: identification of associations among entities and other information in text
10. Sentiment analysis involves discerning subjective (as opposed to factual) material and extracting various forms of attitudinal information: sentiment, opinion, mood, and emotion. Text analytics techniques are helpful in analyzing sentiment at the entity, concept, or topic level and in distinguishing opinion holder and opinion object.[13]
11. Quantitative text analysis is a set of techniques stemming from the social sciences where either a human judge or a computer extracts semantic or grammatical relationships between words in order to find out the meaning or stylistic patterns of, usually, a casual personal text for the purpose of psychological profiling etc.

III. APPLICATION (USE)

Text mining technology is now broadly applied to a wide variety of government, research, and business needs. All these groups may use text mining for records management and searching documents relevant to their daily activities. Legal professionals may use text mining for e-discovery, for example. Governments and military groups use text mining for national security and intelligence purposes. Scientific researchers incorporate text mining approaches into efforts to organize large sets of text data (i.e., addressing the problem of unstructured data), to determine ideas communicated through text (e.g., sentiment analysis in social media·) and to support scientific discovery in fields such as the life sciences and bioinformatics. In business, applications are used to support competitive intelligence and automated ad placement, among numerous other activities.

3.1 Security Application

Many text mining software packages are marketed for security applications, especially monitoring and analysis of online plain text sources such as Internet news, blogs, etc. for national security purposes. It is also involved in the study of text encryption/decryption.

3.2 Biomedical Application

A range of text mining applications in the biomedical literature has been described, including computational approaches to assist with studies in protein docking, protein interactions, and protein-disease associations. In addition, with large patient textual datasets in the clinical field, datasets of demographic information in population studies and adverse event reports, text mining can facilitate clinical studies and precision medicine. Text mining algorithms can facilitate the stratification and indexing of specific clinical events in large patient textual datasets of symptoms, side effects, and co morbidities from electronic health records, event reports, and reports from specific diagnostic tests. One online text mining application in the biomedical literature is PubGene, a publicly accessible search engine that combines biomedical text mining with network visualization. GoPubMed is a knowledge-based search engine for biomedical texts. Text mining techniques also enable us to extract unknown knowledge from unstructured documents in the clinical domain.

3.3 Software Application

Text mining methods and software is also being researched and developed by major firms, including IBM and Microsoft, to further automate the mining and analysis processes, and by different firms working in the area of search and indexing in general as a way to improve their results. Within public sector much effort has been concentrated on creating software for tracking and monitoring terrorist activities. For study purposes, Weka software is one of the most popular options in the scientific world, acting as an excellent entry point for beginners. For Python programmers, there is an excellent toolkit called NLTK for more general purposes. For more advanced programmers, there's also the Gensim library, which focuses on word embedding-based text representations.

3.4 Online Media Application

Text mining is being used by large media companies, such as the Tribune Company, to clarify information and to provide readers with greater search experiences, which in turn increases site "stickiness" and revenue. Additionally, on the back end, editors are benefiting by being able to share, associate and package news across properties, significantly increasing opportunities to monetize content.

3.5 Sentimate Analysis

Sentiment analysis may involve analysis of movie reviews for estimating how favorable a review is for a movie. Such an analysis may need a labeled data set or labeling of the affectivity of words. Resources for affectivity of words and concepts have been made for WordNet and ConceptNet, respectively. Text has been used to detect emotions in the related area of affective computing. Text based approaches to affective computing have been used on multiple corpora such as students evaluations, children stories and news stories.

3.6 Scientific literature mining and academic Application

The issue of text mining is of importance to publishers who hold large databases of information needing indexing for retrieval. This is especially true in scientific disciplines, in which highly specific information is often contained within the written text. Therefore, initiatives have been taken such as Nature's proposal for an Open Text Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD) that would provide semantic cues to machines to answer specific queries contained within the text without removing publisher barriers to public access. Academic institutions have also become involved in the text mining initiative: The National Centre for Text Mining (NaCTeM), is the first publicly funded text mining centre in the world. NaCTeM is operated by the University of Manchester in close collaboration with the Tsujii Lab, University of Tokyo. NaCTeM provides customised tools, research facilities and offers advice to the academic community. They are funded by the Joint Information Systems Committee (JISC) and two of the UK research councils (EPSRC & BBSRC). With an initial focus on text mining in the biological and biomedical sciences, research has since expanded into the areas of social sciences.In the United States, the School of Information at University of California, Berkeley is developing a program called BioText to assist biology researchers in text mining and analysis.The Text Analysis Portal for Research (TAPoR), currently housed at the University of Alberta, is a scholarly project to catalogue text analysis applications and create a gateway for researchers new to the practice.Computational methods have been developed to assist with information retrieval from scientific literature. Published approaches include methods for searching, determining novelty, and clarifying homonyms among technical reports.

IV. CHALLENGES

There are many applications (use) of Text mining but it has also following issues. The challenges of Text mining is explained as below.

4.1 Intermediate Form

Intermediate forms with varying degrees of complexity are suitable for different mining purposes. For a fine-grain domain-specific knowledge discovery task, it is necessary to perform semantic analysis to derive a sufficiently rich representation to capture the relationship between the objects or concepts described in the documents. However, semantic analysis methods are computationally expensive and often operate in the order of a few words per second. It remains a challenge to see how semantic analysis can be made much more efficient and scalable for very large text corporation.

4.2 Multilingual text refining

Whereas data mining is largely language independent, text mining involves a significant language component. It is essential to develop text refining algorithms that process multilingual text documents and produce language-independent intermediate forms. While most text mining tools focus on processing English documents, mining from documents in other languages allows access to previously untapped information and offers a new host of opportunities.

4.3 Domain Knowledge Integration

Domain knowledge, not catered for by any current text mining tools, could play an important role in text mining. Specifically, domain knowledge can be used as early as in the text refining stage. It is interesting to explore how one can take advantage of domain information to improve parsing efficiency and derive a more compact intermediate form. Domain knowledge could also play a part in knowledge distillation. In a classification or predictive modeling task, domain knowledge helps to improve learning/mining efficiency as well as the quality of the learned model (or mined knowledge). It is also interesting to explore how a user’s knowledge can be used to initialize a knowledge structure and make the discovered knowledge more interpretable.

4.4 Personalize autonomous mining

Current text mining products and applications are still tools designed for trained knowledge specialists. Future text mining tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. There have been some efforts in developing systems that interpret natural language queries and automatically perform the appropriate mining operations. Text mining tools could also appear in the form of intelligent personal assistants. Under the agent paradigm, a personal miner would learn a user’s profile, conduct text mining operations automatically, and forward information without requiring an explicit request from the user.

V. Future Scope

Increasing interest is being paid to multilingual data mining: the ability to gain information across languages and cluster similar items from different linguistic sources according to their meaning. The challenge of exploiting the large proportion of enterprise information that originates in "unstructured" form has been recognized for decades. It is recognized in the earliest definition of business intelligence (BI), A Business Intelligence System, which describes a system that will: "...utilize data-processing machines for auto-abstracting and auto-encoding of documents and for creating interest profiles for each of the 'action points' in an organization. Both incoming and internally generated documents are automatically abstracted, characterized by a word pattern, and sent automatically to appropriate action points." Yet as management information systems developed starting in the 1960s, and as BI emerged in the '80s and '90s as a software category and field of practice, the emphasis was on numerical data stored in relational databases. This is not surprising: text in "unstructured" documents is hard to process. The emergence of text analytics in its current form stems from a refocusing of research in the late 1990s from algorithm development to application, as described by Prof. Marti A. Hearst in the paper Untangling Text Data Mining:“For almost a decade the computational linguistics community has viewed large text collections as a resource to be tapped in order to produce better text analysis algorithms. In this paper, I have attempted to suggest a new emphasis: the use of large online text collections to discover new facts and trends about the world itself. I suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results.”

REFERENCES

1. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledgediscovery: An Overview. In Advances in Knowledge Discovery and Data Mining, U.Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., MIT Press,Cambridge, Mass., 1-36.
2. Feldman, R. & Dagan, I. (1995) Knowledge discovery in textual databases (KDT). Inproceedings of the First International Conference on Knowledge Discovery and DataMining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117.

3. Hearst, M. A. (1997) Text data mining: Issues, techniques, and the relationship toinformation access. Presentation notes for UW/MS workshop on data mining, July 1997.
4. Simoudis, E. (1996). Reality check for data mining. IEEE Expert, 11(5).
5. Tan, A.-H. (1997). Cascade ARTMAP: Integrating neural computation and symbolicknowledge processing. IEEE Transactions on Neural Networks, 8(2), 237-250.
6. Tan, A.-H. & Teo, C. (1998). Learning user profiles for personalized informationdissemination. In proceedings, International Joint Conference on Neural Networks(IJCNN'98), Alaska, 183-188.
7. Shen, Jiaming; Xiao, Jinfeng; He, Xinwei; Shang, Jingbo; Sinha, Saurabh; Han, Jiawei (2018-06-27). Entity Set Search of Scientific Literature: An Unsupervised Ranking Approach. ACM. pp. 565–574. doi:10.1145/3209978.3210055. ISBN 9781450356572.
8. Walter, Lothar; Radauer, Alfred; Moehrle, Martin G. (2017-02-06). "The beauty of brimstone butterfly: novelty of patents identified by near environment analysis based on text mining". Scientometrics. 111 (1): 103–115. doi:10.1007/s11192-017-2267-4. ISSN 0138-9130.
9. Roll, Uri; Correia, Ricardo A.; Berger-Tal, Oded (2018-03-10). "Using machine learning to disentangle homonyms in large text corpora". Conservation Biology. 32 (3): 716–724. doi:10.1111/cobi.13044. ISSN 0888-8892. PMID 29086438.
10. Automated analysis of the US presidential elections using Big Data and network analysis; S Sudhahar, GA Veltri, N Cristianini; Big Data & Society 2 (1), 1-28, 2015
11. Network analysis of narrative content in large corpora; S Sudhahar, G De Fazio, R Franzosi, N Cristianini; Natural Language Engineering, 1-32, 2013
12. Quantitative Narrative Analysis; Roberto Franzosi; Emory University © 2010
13. Lansdall-Welfare, Thomas; Sudhahar, Saatviga; Thompson, James; Lewis, Justin; Team, FindMyPast Newspaper; Cristianini, Nello (2017-01-09). "Content analysis of 150 years of British periodicals". Proceedings of the National Academy of Sciences. 114 (4): E457–E465. doi:10.1073/pnas.1606380114. ISSN 0027-8424. PMC 5278459. PMID 28069962.
14. Flaounas, M. Turchi, O. Ali, N. Fyson, T. De Bie, N. Mosdell, J. Lewis, N. Cristianini, The Structure of EU Mediasphere, PLoS ONE, Vol. 5(12), pp. e14243, 2010.
15. Nowcasting Events from the Social Web with Statistical Learning V Lampos, N Cristianini; ACM Transactions on Intelligent Systems and Technology (TIST) 3 (4), 72
16. NOAM: news outlets analysis and monitoring system; I Flaounas, O Ali, M Turchi, T Snowsill, F Nicart, T De Bie, N Cristianini Proc. of the 2011 ACM SIGMOD international conference on Management of data
17. Automatic discovery of patterns in media content, N Cristianini, Combinatorial Pattern Matching, 2-13, 2011
18. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, N. Cristianini, RESEARCH METHODS IN THE AGE OF DIGITAL JOURNALISM, Digital Journalism, Routledge, 2012
19. Circadian Mood Variations in Twitter Content; Fabon Dzogang, Stafford Lightman, Nello Cristianini. Brain and Neuroscience Advances, 1, 2398212817744501.
20. Effects of the Recession on Public Mood in the UK; T Lansdall-Welfare, V Lampos, N Cristianini; Mining Social Network Dynamics (MSND) session on Social Media Applications
21. Researchers given data mining right under new UK copyright laws Archived June 9, 2014, at the Wayback Machine.