Analysis of different techniques for information retrieval from high dimensional Unstructured Text Data

Ankit J. Faldu* & Dr. Kishor H. Atkotiya**
*Author, Assistant Professor-Faculty of Science-MCA, Atmiya University.
**Corresponding Author, Professor-Department of statistic, Saurashtra University.

Received: January 28, 2019 Accepted: March 02, 2019

ABSTRACT: *Text analysis (TA) is atrending research area that used to identifiedhelpful information that could derived from large amount of high dimensional unstructured text data by using difference methodfromvery trending technology that is machine learning, data mining, NLP means natural language processing and information retrieval (IR). Text analysisincludes the preliminary processingmethod to collecting data and earlyaware of the different patterns and attributes that present in the collecting data from different resources. The different Methods like"Information Retrieval (IR)", "Information Extraction (IE)", "Categorization of data", "Clustering and Summarization" of data that are used to analyze these in-between representations like"distribution analysis (DA)", "association rules (DR)" and "visualization". Text analysis helps to find valuable information from customer email, documentation, comments, social network etc.*

Key Words: *Big Data, Unstructured Data, Unstructured Text Data, Algorithms, Information retrieval.*

1. Introduction:

Nowadays a most of business organizations collect and accumulate massive amounts of new information in data warehousesand cloud platforms and this unstructured data remains to raise exponentially by the minute as new information comes driving in from several sources. Asoutcome, it becomes a great issues and challenge for corporate organizations to store, process, and analyze massive amounts of unstructured textual data with currently available traditional tools. This is the main reason that Text analysis comes in the picture.The text analysisobtains the helpful information from different data sourceslike emails, html, comments, documents etc. through the elaboration and identifications of exciting patterns and unique design in data. So Natural processing language is used to translate unstructured data into structured data. Text analysis is used to discovery useful information for this data. Text analysis combines and assimilates the tools of statistics, data mining, information retrieval (IR),machine learning,and computational linguistics, Text analysis also pacts with Natural language processing (NLP)texts that is stored in semi-structured or unstructured formats.Text is an important intermediate for exchange different type of information. The aim of text analysis applications is to trace data, retrieve data and operate data of applicable information proficiently from the large amount of bulk of text which continues to grow expeditiously and exponentially.

2. Text-Analysis Pre-Processing method:

Pre-processing method are dynamic in TA and its different applications. It is also knowsas cleaning the text, which is used to identify bug and eliminate bugs in the given large amount of unstructured text. Different pre-processing methods are available to make the text to be prepared for more analysis process to accomplish notable performance [1]. All Most of thefunctionalities of pre-processing methodswork with Natural language processing (NLP). The succeeding are some key pre-processing techniques: Word Sense Disambiguation (WSD), Parts-of-Speech (POS) Tagging,Tokenization, Grammatical Parsing and Chunking,Stop Word Filteringand Chunking, Text Summarization, Term Frequency and Inverse Document Frequency,Lemmatization and Stemming.

3. Information Retrieval (IR)

IRis a procedure of mining appropriate and related patterns baseon a given large set of different words or expressions. It's very strong relationship in text analysis and information retrieval form large amount of unstructured text data. Information retrieval systems used various algorithms that are helpful to analysis and find pattern the user's comportment and findappropriate data[2]. Most of search engines are more or less using information retrieval (IR)

system regularly to findappropriateinformation and documents according to input text. That's way search engines used query relateddifferent algorithms to find the more relevantoutput. [3]

4. Information Extraction (IE)

Information extraction (IE) is used of repeatedlytaking out structured data from verity of unstructured machine-readable documents. This activity focus on processing human language under stable texts by helpof natural language processing (NLP). Current activities in hypermedia document processing like programmedcomment and content mining out of different unstructured documents could be understood as IE.

The purpose of IE techniques is the to obtaining helpful information from unstructured text. It's used finds the entities, extraction or obtaining of eventsand relationships between unstructured text data.IErecognizes important phrases and relationships available within given unstructured text data. Information extraction (IE) is anxious with obtained of relateddata and information from the unstructured text data. [4].

5. Natural Language Processing (NLP)

NLP focuson automatic analysis and processing of high volumes unstructured text information. NLP doesused various types of data analysis techniques likeNER means "Named Entity Recognition" for recorded and itsalternative wordfindingand to identified the relationships betweenwords [5]. NER recognize the occurrences of definiteentity from differentlarge cluster of various documents. These all the entities and their objectpermit the finds theassociation and other info to achieve their important concept. Main disadvantage of NER technique is to absencesof whole dictionary for completely named objects used for recognition [5][6]. Query relatedto the algorithms essential to be used to achievesuitableoutput. In reality data, a single unit or entity has abundantrelationssuchfreeze and refrigerator. Many times, a collection of successive words have a multiple word names to recognize the boundaries value and solve overlying issues to use classification technique.NER systems have accomplishedsignificancenearto accurate result between 75 % to 85% [7].

To abstractalternative expression and abbreviation from unstructured text data, coreferencing techniques areregularlytaking for used. Natural Languages Processing (NLP) its self-to much of difficulties as anunstructured text mined from different resources do not have similar words or abbreviation. So, it is anessential to distinguish such problems and make guidelines or protocol for equals identification [8].

6. TEXT CATEGORIZATION

Aim of text categorization is to categorizelargeset of verities documents inpersistent number of advance established categories. Everytext file, document or articlearerelating to one or multiple class. Responsibility of categorization is organized a given unstructured data objectto predefine set of categories. TC is one kind of "supervised" learning approach. When categorizing any documentanalgorithm will regularlyconsider as a "Bag of Words". It is a group of unstructured text file, article and documents, the procedure of discovery the perfectresult or result for every document. Automated text categorization techniquesused mixof contexts. Further some important methods of textualanalysis were presented with ordinary text analysis process to expand the significance and correctness of results from unstructured data[9].

To categorizethe various text documents one of method is weighted heuristics which extracts useful information and features by given listed some specific rules. fixed phrase,Sentence length,related word, paragraph. Those methodare implementing and analyzing for text categorization. Text categorizationhas been used on manyfile, documents and article at the same time. Quality of classifiers & types of classifiers is totally depending on the theme and nature of the unstructured text data[10].

7. CLUSTERING

Clustering is most attentive and most important methods in text analysis which is helped to grouping similar type of documents in one group. Its purpose is to identifiedbasic structures available in information, and organizeinformation into relevant sub-groups for additional study of data and analysis of data. It is anon-supervisedprocedure. With these different instances are arrange into different groups knows as acluster. The problemwith clustering is that creationgroup without labelgathering into relevant clusters without any prior information. Any labels relatedtoinstances are found from the given unstructured data. More advantage of clustering is that documents can materialize in multiple topic and sub-topics, socertifying that a valuable document must be presentto search results. A fundamental clustering algorithm makes vector topics for everyvaluable data and procedures the weights of howthe datafix into everycreate cluster.Clustering techniques is useful in many real time applications like pattern

recognition,biology, data analysis,image segmentation, document retrieval, pattern classification, web search, security and BI. Frequently used clustering algorithms are the "K-means", "the EM-based mixture resolving", and the HAC. [11]

8. VISUALIZATION

Visual contentanalysis or data visualization places huge unstructured text data sources in a hierarchy order. The datagiven by form of chart, picture, graph etc. is best, complete and fastprocessing than plain text-based explanation so it is better for analysis the hugeset of documents. Visualization is applicableinbroadvariety of documents and findrelated information. visualization can expand to findingof related useful patterns or related information for text analysis. Information or data that permit a pictorial representation includesontology, result set,relations of keyword are considered the important elements of the search activities. [12]construction of information visualization conducted into following step.

1. Data assembling
2. Data examinationor analysis and extraction
3. Visualization

9. ISSUES RELATEDTO TEXT ANALYSIS

Various issues face in the text analysis process. Difficulties have rise at in-between of text analysis techniques. Pattern analysis is used to translate unstructured text data intostructured but analysis process has generatedits own difficulties[13]. Another mainproblem is a text modificationof group is dependent that make problems. Existing some tools that support multiple languages. Numerous algorithms and different techniques douseful for autonomouslyand support multi language text. Important documents continueexternal the text analysis process. Several tools and technologies don't supportit. [14]. Requirements of the domain, experts'people have been required to work together fromprecise results [5], [6].

Text's synonyms and antonyms availablein the given data generate issuesfor the process of text analysis tools. It's tough to classify the documents whendocumentsarehuge. Abbreviations result gives different meaning in different conditionis largeproblem [15]. So,it's the necessity to define rules depending to the domainthat rules will be used as a standard rule in the required field and shall be integrated in text analysisas a requirement[14], [15]. Natural languages processing (NLP) have manyproblems in text enhancementtechniques and the kind of relationship between entity. Many times, word'sspelling is same but meaning different, for example, bank&bank. TA tools considered both words same. Grammatical regulation is still time an open-endedresearch issue intext analysis [16].

10. CONCLUSION

Present of large textdata required to be analysis to find helpful information. Text analysis are used to analysisof required information from very hugevolumes of unstructured text information. In this research paper we explaintext analysis techniques that is helpful to upgrade and expand the text analysis mechanism. Choosing accurate techniques depending to the domain knowledge support to text analysis to easy, understand efficient. expertiseof various domain and itsintegration, multiple text enhancement, varying concepts granularity and uncertaintyin NLP are importantproblems and challenges that stand up during text analysis process. Many effortsare doing on the document analysis using text analysis methods. But text analysis is tillthe challengingopen-endedissue on present world situation. In future, we will try to design an algorithm and that algorithms arehelp to solvethe different issues mention in this work.

References:

1. Saira, GillaniAndleeb, "From text mining to knowledge mining: An integrated framework of concept extraction and categorization for domain ontology", PhD Dissertation, Budapesti Corvinus Egyetem, 2015.
2. R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," Language Resources and Evaluation, vol. 46, no. 2, pp. 155–176, 2012.
3. Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß, "A Brief Survey of Text Mining. In Ldv Forum", Vol. 20.19–62. 2005.
4. D. Jasmine GunaSundari, D. Sundar, "A Study of Various Text Mining Techniques",vol 08 issue 02 pg.no 82-85 (2017)
5. B. Laxman and D. Sujatha, "Improved method for pattern discovery in text mining," International Journal of Research in Engineering and Technology, vol. 2, no. 1, pp. 2321–2328, 2013.
6. A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravi˘cius, and M. Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," Journal of biomedical semantics, vol. 5, no. 1, p. 1, 2014.
7. A. M. Cohen and W. R. Hersh, "A survey of current work in biomedical text mining," Briefings in bioinformatics, vol. 6, no. 1, pp. 57–71, 2005.
8. E. A. Calvillo, A. Padilla, J. Mu˜noz, J. Ponce, and

J. T. Fernandez, "Searching research papers using clustering and text mining," in Electronics, Communications and Computing (CONIELECOMP), 2013 International Conference on. IEEE, 2013, pp. 78–81.

9. C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," Information Sciences, vol. 275, pp. 314–347, 2014.
10. R. Al-Hashemi, "Text summarization extraction system (tses) using extracted keywords." Int. Arab J. e-Technol., vol. 1, no. 4, pp. 164– 168, 2010.
11. B. L. Narayana and S. P. Kumar, "A new clustering technique on text in sentence for text mining," IJSEAT, vol. 3, no. 3, pp. 69–71, 2015.
12. Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", IJSCE, ISSN: 2231-2307, Vol. 2, Issue-6, January 2013.
13. A. Henriksson, J. Zhao, H. Dalianis, and H. Bostr¨om, "Ensembles ofrandomized trees using diverse distributed representations of clinical events," BMC Medical Informatics and Decision Making, vol. 16, no. 2, p. 69, 2016.
14. A. Kumaran, R. Makin, V. Pattisapu, and S. E. Sharif, "Automatic extraction of synonymy information:-extended abstract," OTT06, vol. 1, p. 55, 2007.
15. A. Kaklauskas, M. Seniut, D. Amaratunga, I. Lill, A. Safonov, N. Vatin, J. Cerkauskas, I. Jackute, A. Kuzminske, and L. Peciure, "Text analytics for android project," Procedia Economics and Finance, vol. 18, pp. 610–617, 2014.
16. N. Samsudin, M. Puteh, A. R. Hamdan, and M. Z. A. Nazri, "Immune based feature selection for opinion mining," in Proceedings of the World Congress on Engineering, vol. 3, 2013, pp. 3–5.