"A CONCEPTUAL STUDY OF KNOWLEDGE DISCOVERY USING TEXT MINING AND ITS APPLICATIONS"

Mrs. Ashwini Brahme[*]

Dr. S.D. Mundhe[**]

ABSTRACT

Text mining is defined as the process of discovering hidden, useful, and interesting patterns from unstructured text documents. Text mining is also known as intelligent text analysis, knowledge discovery in text (KDT), text data mining etc. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, computational linguistics and natural language processing etc. This research paper is intended towards text mining methodology and its applications. Also it explores on problems in text mining. This Paper is focused on future challenges and areas of text mining and knowledge discovery.

KEYWORDS

Text mining, Data Mining, Knowledge Discovery, KDT, DSS

[*] Assistant Professor, Sinhgad Institute of Management and Computer Application (SIMCA), Pune, Savitribai Phule Pune University, Maharashtra, India,

[**] Director , Sinhgad Institute of Management and Computer Application (SIMCA), Pune, Savitribai Phule Pune University, Maharashtra, India,

INTRODUCTION

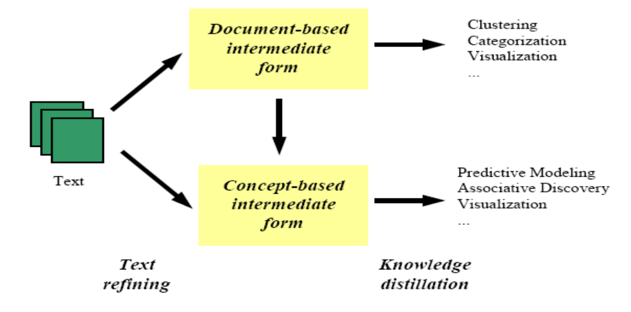
The data stored in computer can be in any one of the form structured, semi structured and unstructured. The data stored in database is structured and semi structured and unstructured data is emails, full text documents, html files, web data, audio, video and many more; approximately 90 % of corporate data is in unstructured format. The information retrieval from unstructured text is very complex requires specific processing methods and algorithm to extract useful pattern, to solve this Text mining technique is used to work on semi structured and unstructured data. Text mining is defined as the process of discovering hidden, useful, and interesting patterns from unstructured text documents. Text mining is also known as intelligent text analysis, knowledge discovery in text, text data mining etc. Text mining is an interdisciplinary filed which incorporates data mining, web mining, information retrieval, computational linguistics and natural language processing etc. Text mining is an inspiring research area which is useful to discover knowledge from unstructured data is important for knowledge management applications such as communities of practices, decision support system (DSS), expertise locations and competency management, business intelligence. There are various applications of text mining like telecommunication, bank, IT, media, insurance, political analysis, pharmaceutical, health care, bioinformatics, business intelligence, national security and many more.

DATA MINING AND TEXT MINING

There are main two objectives of data mining namely Prediction (using some variables in data set in order to predict unknown values to other relevant variables) and Description(involves findings human understandable patterns and trends in the data) which is categorized into classification, regression, clustering, dependency modeling, deviation detection, summarization etc.

Text mining consists of two steps text refining and knowledge distillation. The various steps involved in text mining are

1. Text preprocessing : contains tokenization, stop word removal, stemming
2. Text transformation
3. Feature selection
4. Text mining (data mining) methods such as classification, clustering, association etc.



Dig 1: Text mining framework [1]

Text refining converts unstructured text documents into an intermediate form (IF). IF can be document-based or concept-based. ***Knowledge distillation*** from a document-based IF deduces patterns or knowledge across documents. A document-based IF can be projected onto a concept-based IF by extracting object information relevant to a domain. Knowledge distillation from a concept-based IF deduces patterns or knowledge across objects or concepts. [1]

REVIEW OF LITERATURE

The related literature review carried out for this proposed research is as:

A lot of this information is available over the internet, where firms can use business intelligence tools to take advantage of text mining, which allows business users to see trends and patterns through large amount of informational data. There are two places where text mining can be especially beneficial: on social media platforms and in the medical field. [2]

V.Jayarajand V.Mahalakshmi in the paper entitled "**Text Mining Template Based Algorithm for Text Categorization for Improving Business Intelligence"**, focus on extracting and processing information and deriving knowledge for decision making and improving the scope of business intelligence. The aim of this paper is IRFC (information retrieval based on configuration file) for extracting data from any source of data in the form of configuration file to support all kind of information. This **IRFC** technique is compared with **KNN Text Classification Algorithm** where time optimization is compared and results are interpreted for selecting the proper candidate for job from number of resumes. [3]

K.L.Sumathy and M. Chidambaram in the paper entitled **"Text Mining: Concepts, Applications, Tools and Issues – An Overview** "talk on the information retrieval from unstructured text is very complex requires specific processing methods and algorithm to extract useful patterns. This paper describes about general framework of text mining which contains two main steps text refining and knowledge distillation, text mining process, areas of text mining such as information retrieval, information extraction, data mining and natural language processing, applications of text mining like telecommunication, bank, IT, media, insurance, political analysis, pharmaceutical, health care, bioinformatics, business intelligence, national security and many more. [4]

Mr. Rahul Patel, Mr. Gaurav Sharma reviews the various text mining techniques such as information extraction, summarization, categorization, clustering, concept linkage, information visualization and also briefed about various text mining algorithms such as k nearest neighbor, support vector machine, Bayesian classifier, K-mean clustering in the paper entitled "**A survey on text mining techniques** ". [5]

The information available on web is semi structured or unstructured so it is very difficult to directly carry on data mining on the Web page. The data on Web have to be through necessary data processing for the same text mining is used entitled in paper "**An Information Retrieval(IR) Techniques for text Mining on web for Unstructured data"**, by Santosh Kumar Paul, MadhupAgrawal, Shyam Rajput, Sanjeev Kumar. [6]

Article by Mark A. Anawis , "**Text Mining: The Next Data Frontier"**, 2014 portray text mining methodology , related fields, tools , applications such as spam filters, fraud detection, sentiment analysis , trend and authorship identification. This paper focuses on future challenge of text mining in to a large corpora, domain knowledge, and personalization and muti-lingual capabilities.[7]

Ingo Feinerer, "**Introduction to the tm Package Text Mining in R**", briefs about use of corpus(collection of documents) and removal of stop word , stemming, filtering, term document metrics' and operations on it.[8]

Report of Oracle® Data Mining Concepts 11g Release 1 (11.1), 2008 focus on unstructured data and various text mining algorithms for Such as Naive Bayes, Generalized Linear Models, Support Vector Machine, k-Means, Non-Negative Matrix Factorization, Apriori, Minimum Descriptor Length etc. Report also concludes that Apriori algorithm is used for association mining function which generates various association rules for information retrieval. [9]

Article entitled "**Text Mining: The state of the art and the challenges"** by Ah-Hwee Tan illustrate a general framework of Text mining which contains main two components text refining and knowledge distillation. Text refining converts unstructured text document into intermediate form and knowledge distillation form is deduces pattern or knowledge across the object. The article also talks about text mining challenges such as there is need to capture relationship between the concepts or objects described in the document and it is essential to develop text refining algorithms. [10]

Debora Cheney, in his paper entitled "**Text mining newspapers and news content: new trends and research methodologies**" focus on the news paper documents related to humanities and social sciences text mining including government documents, novel/literature, magazines, newspapers and social media content. Text mining requires a large corpus of documents /contents which is freely accessible and available to the researcher through the newspapers, web pages. [11]

The central challenge of text mining is that the accurate analysis of both structured and unstructured data in order to extract meaningful associations, trends and patterns in large corpuses of text. The various text mining tools are also described such as SysomosMAP, Netbase, Crimson Hexagon Foresight, Discover Text, LinguamaticsI2E etc described in **"Text Mining and Social Media: When Quantitative Meets Qualitative, and Software Meets Humans"** by Lawrence Ampofo, Simon Collister, Ben O'Loughlin, and Andrew Chadwick. [12]

FUTUER RESEARCH & DIRECTIONS

Data mining algorithms act on numerical and categorical data stored in relational databases or spreadsheets. Today there is need to mine the data that is not numerical or categorical such as web page data , images, audio video files, presentations, documents libraries, reports and many more. This type of data is known as unstructured or semi-structured data. Extracting meaning full information and generating knowledge is critical and need of business for decision making and text mining is best solution for the same.

Only 20% of web data contains useful information and remaining 80% data is not useful for mining from this 20% of data several frequent patterns generated using pattern discovery technique like association rule mining which is useful to retrieve particular mining from collection of data, for web prediction and website modification.

Text mining is an inspiring research area which is useful to discover knowledge from unstructured data is important for knowledge management applications ; there are many applications of text mining such as telecommunication, bank, IT, media, insurance, political analysis, pharmaceutical, health care, bioinformatics, national security, etc.

CONCLUSION

Manual analysis and generating effective knowledge discovery from useful information is not possible because of huge availability of information on website. There are various challenges of text mining like large or huge data; documents are in unstructured format complex relationship between the text, ambiguity and context sensitivity in the text. This can be achieved by text mining by finding the patterns, trends, relationship between the concepts in text.

BIBLIOGRAPHY AND REFERENCES

[1]. *Ah-Hwee Tan,* “Text Mining : The state of the art and the challenges”

[2]. http://www.activereportsserver.com/business-intelligence-news/text-mining-a-bi-tool-that-can-help-interpret-mass-amounts-of-information

[3]. V.Jayaraj, V.Mahalakshmi,“Text Mining Template Based Algorithm for Text Categorization for Improving Business Intelligence”,International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS),8(4),2014, ISSN (Print): 2279-0047 ISSN (Online): 2279-0055, Page no. 345 –350,www.iasir.net

[4].K.L. Sumathy ,M. Chidambaram, “Text Mining: Concepts, Applications, Tools and Issues – An Overview” International Journal of Computer Applications (0975 – 8887) ,Volume 80 – No.4, October 2013

[5]. Mr. Rahul Patel, Mr. Gaurav Sharma, “A survey on text mining techniques”, International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 3 Issue 5 May, 2014 Page No. 5621-5625 , www.ijecs.in

[6]. Santosh Kumar Paul, MadhupAgrawal, Shyam Rajput, Sanjeev Kumar , “An Information Retrieval(IR) Techniques for text Mining on web for Unstructured data”, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 4, Issue 2, February 2014, www.ijarcsse.com

[7]. Mark A. Anawis , “Text Mining: The Next Data Frontier”, 2014

[8]. Ingo Feinerer, “Introduction to the tm Package Text Mining in R”, June 10, 2014

[9]. http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/text.htm#BCEDHEDD

[10]. Ah-Hwee Tan, “Text Mining: The state of the art and the challenges”

[11]. Debora Cheney, “Text mining newspapers and news content: new trends and research methodologies”, Library Services to the World Campus/Penn State Online, July 24, 2013,http://creativecommons.org/licenses/by/3.0/

[12].Lawrence Ampofo, Simon Collister, Ben O’Loughlin, Andrew Chadwick, “Text Mining and Social Media : When Quantitative Meets Qualitative, and Software Meets Humans” New Political Communication Unit Working Paper, October 2013

[13].Online Tutorial: Business Intelligence, Predictive Analytics, and Data Mining Content (http://wps.pearsoned.co.uk/ema_ge_turban_elec_comm_2012/217/55592/14231612.cw/index.html)

[14]. E-book: Introduction to Data Mining and Knowledge Discovery Third Edition By Two Crows Corporation, ISBN: 1-892095-02-5