ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

A Theoretical Review on Text Mining: Tools, Techniques, Applications and Future Challenges

Ashwini R. Kulkarni [1], Dr. S.D. Mundhe [2]

Asst. Professor, Sinhgad Institute of Management and Computer Application (SIMCA), Pune, India[1]

Director -MCA, Sinhgad Institute of Management and Computer Application (SIMCA), Pune, India [2]

ABSTRACT: Text mining is a method of discovering or retrieving information from unstructured data or information. It is also known as knowledge discovery from text, text data mining, text data analytics, etc. It is a multidisciplinary field contains information extraction, clustering, categorization, information retrieval, text analysis, visualization, database technology, machine learning, and data mining. This main objective of this paper is to elaborate on text mining, different techniques of text mining. The paper is also aimed towards the various applications, challenges and tools of text mining. The paper is also conferred on reviews of research work done in this filed by diverse researchers, scholars, organizations etc.

KEYWORDS: Data Mining, Text Mining, Knowledge Discovery, Stemming, Information retrieval, Natural Language Processing (NLP)

I. INTRODUCTION

The data available on internet and worldwide web is very huge and vast, only 10% of data is in structured form and approximate 90% of data is in either unstructured or semi-structured form. Data mining is only feasible for structured data and not for unstructured and semi-structured data. The ample amount of available information or data is stored on web is in the text, images, video, audio formats. To work on such a data or information there is need to focus on text mining. The definition of text mining is as the process of extracting patterns from text documents or text database. Text mining has high commercial potential in all the areas as compared to the data mining.

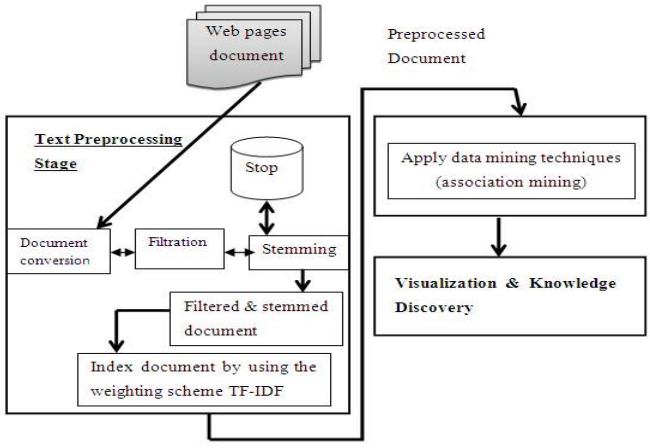


Diagram: Text mining and knowledge discovery process

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

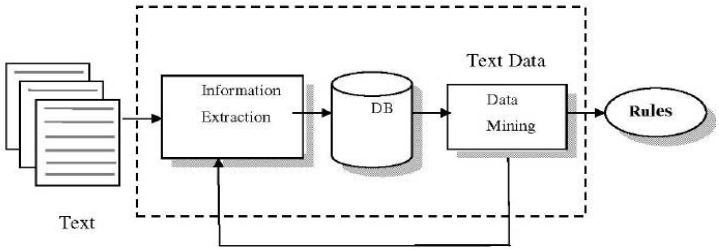
Vol. 4, Issue 11, November 2016

II. TECHNIQUES OF TEXT MINING

There are various text mining methods/techniques such as Summarization, Information Extraction, Topic Tracking, Classification/categorization, Information visualization, Clustering, Concept Linkage, questioning answering and Association Rule Mining. ***Classification*** includes Decision Tree (DT), K-Nearest Neighbor, Support Vector Machine (SVM), Neural Network (NN) and Bayesian Method. ***Clustering*** includes Partition Method, Hierarchical clustering, and Density Based Clustering.

A. INFORMATION EXTRACTION

Information extraction is a process of extracting structured information from unstructured or semi structured information and processing it for pattern matching or pattern discovery. The main problem with information extraction is transforming of corpus document into a structured database for further KDD.



Dig: Information Extraction

B. TOPIC TRACKING

Topic tracking is a technique in which any user search or view any document it keeps users profile and predict other documents of interest to the user. e.g. www.aletrs.google.com

Topic tracking is used in companies to alert anytime a competitor is in the news to track for products and company. It is also useful for medical industry by doctors and other professionals for new treatment and other new advancements. In education sector topic tracking is also useful for referencing research in the area of interest. In topic tracking keywords identification from huge amount of online news data is helpful to summarize the news articles.

C. SUMMARIZATION

It is used to summarize the lengthy documents, the challenge of summarization technique is to teach the software to analyze the semantics and interpret the meaning. Summarization is used with topic tracking or categorization tools to summarize the particular retrieved document on specific topic. Summarization follows main three steps 1) preprocessing – where the textual document is converted to structured format 2) algorithm transform the text structure to summary structure 3) generation step – final summary is obtained from summary structure. The main goal of preprocessing is to reduce the dimensionality.

This preprocessing contains a) stop word elimination – common words with no semantics and do not aggregate relevant information to the task (e.g. a, an, the, for, etc) are eliminated b) case folding: converting all characters to the same kind either lower case or upper case c) stemming: syntactically similar words such as plurals, verbal variations are considered similar to obtain the stem (root word).

D. CATEGORIZATION OR CLASSIFICATION

It is used to identify the main theme of the document by using predefined set of topics. Categorization counts the words that are available and identify the main topic of the document, ranking of document. The main objective of categorization is to classify a set of documents into a fixed number of predefined categories. It is used in many industries and business to provide a customer support and have question answers on variety of topics from their customers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

E. CLUSTERING

Clustering is used to group the similar documents and it differs from categorization where the documents are grouped as per the predefined topics. It is useful in organization of management information system which contains huge number of documents. The steps in clustering are stop word removal, stemming and filtering.

F. CONCEPT LINKAGE

Concept Linkage connects the related documents by identifying their commonly shared concepts; it promotes browsing of information rather than searching. It is useful for biomedical research to search on which topic the research is done or not. Concept linkage software can identify links between diseases and treatments.

G. INFORMATION VISUALIZATION OR VISUAL TEXT MINING

Visual text mining or information visualization puts large text sources in the visual hierarchy and provides browsing capabilities. DocMiner is tool used for information visualization. It is carried out in three steps as First is Data Preparation: determine and acquire original data of visualization and form original data space, Second is Data analysis and extraction: i.e. analyze and extract visualization data needed from original data and form visualization data space and third is Visualization mapping: i.e. employ certain mapping algorithm to map visualization data space to visualization target.

H. QUESTIONING ANSWERING

Questioning and answering deals with how to find the best answer to a given question. The applications of this technique are in companies where employees can search answer for a questions, education field, medical field etc. It uses many of the text mining techniques for questioning and answering.

I. ASSOCIATION RULE MINING

It is used to discover the relationship among a large set of variables in a dataset. It is widely used in industries, companies for business decision making, supermarkets. Association rule mining is also a part of data mining, also known as knowledge discovery in a database. Association rule for text mining focus on relationships and implications among the topics or descriptive concepts used to characterize a corpus.

III. LITERATURE REVIEW

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 (IF) 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. [2]**

Text Mining: Tools, Techniques, and Applications presentation by Nathan Treloar, describes 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 find the patterns, trends, relationship between the concepts in text.** This presentation can also guide for medical data mining related to diseases, symptoms, drugs, medicines. [3]

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 LawrenceAmpofo, Simon Collister, Ben O'Loughlin, andAndrew Chadwick. [4]

Author Mr. Rahul Patel, Mr. Gaurav Sharma illustrated about the various text mining techniques such as information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, question answering and association rule mining in his research paper entitled**"A survey on text mining techniques".**

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Also author has focused on Association rule mining (ARM) used to discover relationships among a large set of variables in a data set, the relationship between the variables is known as association rule. [5]

A research paper entitled **"Advanced Knowledge Based Systems: Model, Applications & Research"** by Priti Srinivas Sajja, Rajendra Akerkar describes about the various research areas in pure knowledge based system as KBS development, Knowledge management, Information and query based system, user interface , expert system, distributed KBS, knowledge grid, multi agent system, tutoring system, soft computing based KBS. Knowledge discovery is one of the application and current research area of Knowledge based data mining. The researcher also described about the use of knowledge-based systems for interpretation, prediction, diagnosis, design, planning, monitoring, debugging, repair, instruction, control, etc. [6]

"A Survey of Text Mining Techniques and Applications" by Vishal Gupta and Gurpreet S. Lehal elaborates on text mining techniques as Information Extraction, Topic Tracking, Summarization, Categorization or Classification, Clustering, Concept Linkage, Information visualization, questioning answering. And association rule mining. The text mining application are in publishing and media, telecommunication, IT sector, bank, insurance, financial market, politics, healthcare, pharmaceutical and research companies and many more.
The researcher also talks on only 20% of data on intranet and World Wide Web is structured and rest 80% of data is in either unstructured or semi-structured form. Knowledge discovery from text is needed to focus where main challenge is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. [1]

A research paper entitled "**A Cogitate Study on Text Mining**" by Y. Jahnavi and Y. Radhika is enlighten on a cogitate study on preprocessing, term weighting algorithms, concept based term weighting algorithms, categorization, pattern discovery, domain ontology based frame work for text mining and summarization techniques is presented. Also researcher had focused on various applications of text mining. Preprocessing of test documents is one of the important steps in text mining, in pre-processing various things are followed like stop word elimination, stemming and pruning. The researcher had studied on Porters stemming algorithm which is implemented by using suffix list and it is speedy small and easy; but it may not work for all the words. This paper also describes various term weighting algorithms.
Text mining technique is useful in various sectors such as publishing and media, telecommunications, energy, services industries, Information technology, Internet Banking, insurance, financial markets, Political institutions, political analysts, public administration and legal documents, Pharmaceutical and research companies, healthcare, etc. [7]

There are various text mining tools such as Intelligent Text Analysis (ITA), Intelligent Miner for Text (IBM Software), Text Finder (Paracel Inc.), Callable Personal Librarian (CPL), Vantage Point, Fulcrum Search Server, Isaac and Amberfish, ISIS , AE1, WordSmith Tools, Harvest etc are also described in **"A Comprehensive Study of Text Mining Approach"** paper . Researcher also pointed on various applications of text mining like competitive intelligence, detection of junk mails, management of human resources, customer relationship management, multilingual applications of natural language processing, classification of news as a text, classification of scientific documents, sentiment classification. Researcher has suggested that there is need to design and develop efficient technique of text mining. [8]

The researcher elaborates on structured and unstructured data; also it aims on discovery of relationship between disease, symptoms, drugs etc in healthcare. It also point out the challenges of text mining in terms of large dimensions of data, complexity in data, ambiguity and sensitivity of information etc. The researcher has given various tools for text mining namely IBM Intelligent Miner for Text, Semio Map, InXightLinguistX / ThingFinder, LexiQuest, ClearForest, Teragram, SRA NetOwl Extractor, Autonomy etc tools are suggested by Nathan Treloar in his presentation tilted **"Text Mining: Tools, Techniques, and Applications".** [3]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

IV. CHALLENGES OF TEXT MINING

There are some important challenges in text mining as:

- Current research in the area of text mining tackles problems of classification, clustering, text representation, information extraction, pattern discovery from hidden patterns.
- Intermediate form
- Multilingual text refining
- Domain knowledge integration
- Personalized autonomous mining
- Large dimension
- Complexity of natural language
- Ambiguity and context sensitivity in data

V. APPLICATIONS OF TEXT MINING

The text mining applications are useful in different areas like in publishing and media, telecommunication, Information Technology sector, banking, insurance, financial market, politics, healthcare, pharmaceutical, research companies, education, social media monitoring, bioinformatics, business intelligence, national security, and many more.

VI. TEXT MINING TOOLS

There are various text mining tools namely IBM Intelligent Miner for Text, Semio Map, InXightLinguistX / ThingFinder, LexiQuest, ClearForest, Teragram, SRA NetOwl Extractor, Autonomy etc.
National center of text mining also suggested various tools as per the use, need and its application such as SureChem for chemical structure/chemist, BrainMap for human brain function and structure, Rely Technology Management Inc. to create information products for pharmaceutical and biotech companies, etc

VII. FUTURE DIRECTIONS AND RESEARCH

To deal with unstructured information, text mining is helpful to user to find accurate information or knowledge from text documents. Text mining provides techniques that discover knowledge from unstructured information. Structured data have particular format for example, structured data is structured query language (SQL) in which rows, columns, contains actual data. Unstructured data usually includes information from web which does not have fixed data model also computer program cannot use it easily. Text mining is an increasing field in the research; the ample amount of information generated by using data and text mining approaches. The main research issue is how to use the information for effective knowledge discovery. Knowledge discovery from text is needed to focus where main challenge is to extract explicit and implicit concepts and semantic relations between the terms using Natural Language Processing (NLP) techniques.

VIII. CONCLUSION

Text mining applications are vide and plays vital role in various sectors like publishing and media, telecommunications, energy, Information technology sector, Internet, Banks, insurance and financial markets, public administration & legal documents, Political institutions, political analysts, Pharmaceutical and research companies, healthcare, etc. The very important point in case of text mining it requires preprocessing and for preprocessing it also has various challenges in terms of stemming, indexing, etc. The proposed research aims to apply text mining techniques with an effective preprocessing in any one of the area and try to advancement in the text mining techniques and to recover the challenges for text mining.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

REFERENCES

1. Gupta V., Lehal G. (2009). A Survey of Text Mining Techniques and Applications**,** *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE,* 1(1)
2. Ah-Hwee Tan, Text Mining: The state of the art and the challenges
3. Treloar N. (2002). Text Mining: Tools, Techniques, and Applications
4. Ampofo L. , Collister S. , Loughlin B., Chadwick A. (2013) Text Mining and Social Media : When Quantitative Meets Qualitative, and Software Meets Humans New Political Communication Unit Working Paper.
5. Patel R. , Sharma G. (2014). A survey on text mining techniques. *International Journal Of Engineering And Computer Science* 3(5) , 5621-5625
6. Sajja P., Akerkar R. (2010). Advanced Knowledge Based Systems: Model, Applications & Research . 1(1) 1 – 11, TMRF e-Book ISBN 978-81-908426-0-0 Retrieved From http://www.tmrfindia.org/eseries/ebookv1.html
7. Y Jahnavi, Y. Radhika (2012). A Cogitate Study on Text Mining. *International Journal of Engineering and Advanced Technology (IJEAT),* 1(6), 189 – 196
8. Kaushik A., Naithani K. (2016). A Comprehensive Study of Text Mining Approach. *IJCSNS International Journal of Computer Science and Network Security,*16(2), 69 -76
9. Online Tutorial: Business Intelligence, Predictive Analytics, and Data Mining Content (Retrieved From http://wps.pearsoned.co.uk/ema_ge_turban_elec_comm_2012/217/55592/14231612.cw/index.html)
10. E-book: Introduction to Data Mining and Knowledge Discovery Third Edition By Two Crows Corporation, ISBN: 1-892095-02-5
11. Jonathan Clark. (2013). Text Mining and Scholarly Publishing. *Publishing Research Consortium* www.publishingreserach.net
12. Brahme A., Mundhe S. (2015).A Conceptual Study of Knowledge Discovery Using Text Mining and Its Applications. *International Journal of Management, IT & Engineering,* 5(9), 174-181
13. Brahme A., Mundhe S. (2016). A Review of Knowledge Based System in Healthcare Using Text Mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(6), 1326-1328

BIOGRAPHY

Mrs. Ashwini Rajendra Kulkarni , working as Assitant Professor at Sinhagd Institute of Managemnet and Computer Application (SIMCA), Narhe , Pune affilated to Savitribai Phule Pune University , Pune Published papers in national/Inetnational conferences and refreed journals.