International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

A Literature Review on Text Mining With Different Techniques

V.Sudheer Goud[1], Prof. P. Premchand[2]

[1]Research Scholar, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, A.P, India

[2]Professor, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, T.S, India.

ABSTRACT:

Text mining referred as intelligent text analysis or text data mining in text and it refers generally to the method of mining or retrieving, knowledgeable and non-trivial information and knowledge from unstructured text. Unstructured text has huge amount information which is not easily used by the computer for processing. So that we require certain techniques to accomplish this task for extracting required patterns. Text mining plays an important role of extracting useful patterns from unstructured text. It is one of the emerging technologies for Knowledge Discovery Process. Document organization and pattern discovery becomes the main task in data mining. In this paper, a review of Text mining and its techniques, applications, merits and demerits of text mining have been presented.

KEYWORDS: Text mining, information extraction, summarization, topic tracking, classification

1. INTRODUCTION

Text Mining is the process of analyzing naturally occurring text for the purpose of discovering and capturing semantic information for insertion and storage in a Knowledge Organization Structure (KOS) with the ultimate goal of enabling knowledge discovery via either textual or visual access for use in a wide range of significant applications. Text Mining is defined as the computational process of extracting useful information from massive amounts of digital data by mapping low-level data into richer, more abstract forms and by detecting meaningful patterns implicitly present in the data. TM is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information, together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text Mining falls into an area called information extraction. Information extraction (IE) software identifies and removes relevant information from texts, pulling information from a variety of sources and aggregates it, to create a single view. IE translates content into a homogeneous form through technologies like extensible Mark-up Language (XML). The goal of IE software is to transform texts composed of everyday language into a structured, database format. In this way, heterogeneous documents are summarized and presented in a uniform manner

2. WHY TEXT MINING

Text mining has engrossed mounting significance and has been dynamically applied in knowledge management. Finding helpful facts or nuggets of knowledge in databases of text is the spirit of text mining, an analysis practice that attempts to uncover hidden patterns in unstructured text data. Text Mining is presently being used in knowledge discovery and business intelligence applications ranging from human resource management to market intelligence to research and development. Its techniques are also being used to extend conventional information retrieval systems with features that create a more interactive and contextually aware search experience. Powerful information technology techniques now exist to identify the relevant S&T literatures and extract the required information. Techniques help to:

i. Substantially enhance the retrieval of useful information from global S&T databases;
ii. Identify the technology infrastructure (authors, journals, organizations) of a technical domain;
iii. Identify experts for innovation-enhancing technical workshops and review panels;
iv. Develop site visitation strategies for assessment of prolific organizations globally;
v. Generate technical taxonomies (classification schemes) with human-based and computer-based clustering methods; 6. Estimate global levels of emphasis in targeted technical areas;
vi. Provide roadmaps for tracking myriad research impacts across time and applications areas.

3. RELATED WORK

Yuefeng Li et al [13]: A Text mining and classification method has been used term-based approaches. The problems of polysemy and synonymy are one of the major issues. There was a hypothesis that pattern-based methods should outperform best compare to the term-based ones in

describing user preferences. A large scale pattern remains a hard problem in text mining. The stateof-the-art term-based methods and the pattern based methods in proposed model which performs efficiently. In this work fclustering algorithm is used. Relevance feature discovery based on both positive and negative feedback for text mining models.

Jian ma et al [4]: The author focused towards the problem by classifying text documents on

axiomatically, for the most part in English. When work with non-English language texts it leads to the forbiddance. Ontology-based text mining approach has been used. Its efficient and effective for clustering research proposals encapsulated with the English and Chinese texts using a SOM algorithm. This method can be expanded to help in searching a better match between proposals and reviewers.

Chien-Liang Liu et al [2]: The paper concluded that the information about the movie-rating is based on the result of sentiment-classification. The feature-based summarizations are used to generate condensed descriptions of movie reviews. The author designed a latent semantic analysis (LSA) to establish product features. It is a way to reduce the size of summary from LSA. They account both accuracy of sentiment classification and response time of a system to design

the system by using a clustering algorithm. OpenNLP2 tool is used for implementation.

Yue Hu et al [19]: PPSGen is a new system which was proposed to solicitation of the presentation slides been generated can be used as drafts. It helps them to prepare the formal slides in a faster way for the proprietor. PPSGen system can bring out slides with better quality suggested by the author. The system was developed by the Hierarchical agglomeration algorithm. Tools are a Microsoft Power- Point and OpenOffice. A 200 combo of papers and slides are taken as tests set from the web demonstrate for evaluation process. PPSGen is comparably better than the baseline methods that were evident by the user study.

Xiuzhen Zhang et al [10]: The problem faced by all the reputation system is concentrated by the

author. However the reputation scores are universally high for sellers. It is a situation requiring great effort for promising buyers to select trustworthy sellers. Author proposed CommTrust for trust evaluation by feedback comments through mining. A multidimensional trust model is used for computation job. Data set are collected from ebay, amazon. In this technique used a Lexical-LDA algorithm. CommTrust can effectively address the good reputation problem issue and rank sellers are finally by showing definitely through the extensive experiments on eBay and Amazon data.

Dnyanesh G. Rajpathak et al [9]: The challenging task is In-time augmentation of D-matrix

through the finding of new symptoms and failure modes. Proposed strategy is to construct the fault diagnosis ontology abide with concepts and relationships frequently observed in the fault diagnosis domain. The needed artifacts and their dependencies from the unstructured repair verbatim text were found out by the ontology. Real-life data collected from the automobile domain. Text mining algorithms are used. To establish automatically the D-matrices by the unstructured repair verbatim data that was mined done by the ontology based text mining composed while fault diagnosis. A graph and the graph comparison algorithms have to be generated for each D-matrix.

Jehoshua Eliashberg et al [11]: To forecast the box office performance of a movie at the crenulation point, it's suitable only if it holds the script and production cost. They extract textual features in three levels particularly genre and content, semantics, and bag-of- words from scripts using domain knowledge of screenwriting, input given by human, and natural language processing techniques. A kernel-based approach is to assess box office performance. Data set are collected from 300 movie shooting scripts. The proposed methodology predicts box office income more exactly 29 percent is reduced mean squared error (MSE) compared to benchmark methods.

Donald E. Brown et al [17]: Rail accidents present image of a valuable safety point for the

transportation industry in many countries. The Federal Railroad Administration needs the railroads muddled in accidents to submit reports. The report has to be cuddled with default field entries and narratives. A combination of techniques is to automatically discover accident characteristics that can inform a better understanding of the patron to the accidents. Forest algorithm has been used. Text

mining looks at ways to extract features from text that takes advantage of language characteristics particular to the rail transport industry.

Luís Filipe da Cruz Nassif et al [6]: In forensic analysis that was computerized with millions of files is usually examined. Unstructured text was found in most of the files performing analyzing process is highly challenging revealed by computer examiners. Document clustering algorithms for the analysis of computers on forensic department seized in police an investigation which was suggested by the author. Variety of mixture of parameters that leads to prompt of 16 different algorithms consider for evaluation. K-means, K-medoids, Single, Complete and Average Link, CSPA are the clustering algorithm are used. Clustering algorithms motivate to induce clusters formed by either relevant or irrelevant document which is used to enhance the expert examiner's job.

4. NECESSITIES OF TEXT MINING

For comprehensive access to the global S&T literature, and maximum extraction of useful information from this literature, five primary conditions are required.

1. A large fraction of the S&T conducted globally must be documented - information comprehensiveness.

2. The documentation describing each S&T project must have sufficient information content to satisfy the analysis requirements - information quality.

3. A large fraction of these documents must be retrieved for analysis - information retrieval.

4. Techniques and protocols must be available for extracting useful information from the retrieved documents - information extraction.

5. Technical domain and information technology experts must be closely involved with every step of the information retrieval and extraction processes - technical expertise.

5. CONCLUSION

So in this paper, our focus is basically what text mining is and why it is useful. We have also discussed requirements of text mining, finally we taken review of different author's regards text mining. In next paper we will discuss about text mining process and its applications.

6. REFERENCES

[1] Luis Tari, Phan Huy Tu,Jorg Hakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez, And Chitta Baral,"Incremental Information Extraction Using Relational Databases", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.

[2] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, And Emery Jou," Movie Rating And Review Summarizationin Mobile Environment", IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 3, May 2012.

[3] Fuzhen Zhuang, Ping Luo, Zhiyong Shen, Qing He, Yuhong Xiong, Zhongzhi Shi, And Hui Xiong, "Mining Distinction And Commonality Across Multiple Domains Using Generative Model For Text Classification" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 11, November 2012.

[4] Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu," An Ontology- Based Text-Mining Method To Cluster Proposals For Research Project

Selection", IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012.

[5] Charu C. Aggarwal, Yuchen Zhao, And Philip S. Yu, "On The Use Of Side Information For Mining Text Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014.

[6] Luís Filipe Da Cruz Nassif And Eduardo Raul Hruschka," Document Clustering For Forensic

Analysis: An Approach For Improving Computer Inspection" IEEE Transactions On Information

Forensics And Security, Vol. 8, No. 1, January 2013.

[7] Bo Chen, Wai Lam, Ivor W. Tsang, And Tak-Lam Wong, "Discovering Low-Rank Shared Concept Space For Adapting Text Mining Models" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 35, No. 6, June 2013.

[8] Francisco Moraes Oliveira-Neto, Lee D. Han, And Myong Kee Jeong." An Online Self-Learning Algorithm for License Plate Matching", IEEE Transactions On Intelligent Transportation Systems, Vol. 14, No. 4, December 2013.

[9] Dnyanesh G. Rajpathak And Satnam Singh," An Ontology-Based Text Mining Method To Develop D-Matrix From Unstructured Text", IEEE Transactions On Systems, Man, And Cybernetics: Systems, Vol. 44, No. 7, July 2014.

[10] Xiuzhen Zhang, Lishan Cui, And Yan Wang, "Commtrust: Computing Multi-Dimensional Trust By Mining E-Commerce Feedback Comments", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.

[11] Jehoshua Eliashberg, Sam K. Hui, And Z. John Zhang," Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 11, November 2014.

[12] Riccardo Scandariato, James Walden, Aram Hovsepyan, And Wouter Joosen," Predicting Vulnerable Software Components Via Text Mining", IEEE Transactions On Software Engineering, Vol. 40, No. 10, October 2014.

[13] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, And Moch Arif Bijaksana," Relevance Feature Discovery For Text Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 6, June 2015.

[14] Kamal Taha," Extracting Various Classes Of Data From Biological Text Using The Concept Of Existence Dependency", IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 6, November 2015.

[15] Shuhui Jiang, Xueming Qian, Jialie Shen, Yun Fu And Tao Mei, "Author Topic Model-Based Collaborative Filtering For Personalized Poi Recommendations", IEEE Transactions On Multimedia, Vol. 17, No. 6, June 2015.

[16] Beichen Wang, Xiaodong Chen, Hiroshi Mamitsuka, And Shanfeng Zhu," Bmexpert: Mining Medline For Finding Experts In Biomedical Domains Based On Language Model", I IEEE /Acm Transactions On Computational Biology And Bioinformatics, Vol. 12, No. 6, November/December 2015.

[17] Donald E. Brown," Text Mining The Contributors To Rail Accidents", IEEE Transactions On Intelligent Transportation Systems, Vol. 17, No. 2, February 2016.

BIBLOGRAPHY

V.Sudheer Goud, Post Graduated in Master of Computer Application (**MCA**) from OU,1994 , Post Graduated in Master of Business Administration (**MBA**) from OU,2006, Post Graduated in Master of Computer Science &

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July 2016

Engineering (**M.Tech**) from IETE , Hyderabad in 2013 and Pursuing Phd in Computer Science in ANU. He is currently working as an Associate Professor, Department of Computer Science in **Holy Mary Institute of Technology and Science** (HITS), (V) Bogaram, (M) Keesara, Medchal .Dist, Telangana, India. He has 23 years of Teaching Experience. His research interests include, Data Mining, Cloud Computing and Information Security.

Prof. P. Premchand, He is currently working as an Professor, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, Telangana State, India.