A REVIEW ON TEXT MINING IN DATA MINING

Yogapreethi.N[1], Maheswari.S[2]

[1]M.E.Scholar, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India
[2]Associate Professor, Department of Computer Science & Engineering, Nandha Engineering College, Erode-638052, Tamil Nadu, India

ABSTRACT

Data mining is the knowledge discovery in databases and the gaol is to extract patterns and knowledge from large amounts of data. The important term in data mining is text mining. Text mining extracts the quality information highly from text. Statistical pattern learning is used to high quality information. High –quality in text mining defines the combinations of relevance, novelty and interestingness. Tasks in text mining are text categorization, text clustering, entity extraction and sentiment analysis. Applications of natural language processing and analytical methods are highly preferred to turn text into data for analysis. This survey is about the various techniques and algorithms used in text mining.

KEYWORDS

Data mining, Text mining, knowledge discovery

1. INTRODUCTION

Text mining is to handle textual data. Textual data is unstructured, unclear and manipulation is difficult. Text mining is best method for information exchange. A non-traditional information retrieval strategy is used in text mining. For obtaining information from large set of textual documents which was done by the text mining. The figure1 is elaborated with the process of text mining.

In recent times, language analysis would be done by the computer is better than the human being. The manual techniques were expensive and time consuming method. To achieve this goal of text mining, there are various technologies are deployed. The technologies are information extraction, summarization, topic tracking, classification and clustering. Knowledge Discovery from Text (KDT) [6] is one of the issues to derive implicit and explicit concepts .Natural Language Processing (NLP) [8, 13] techniques are used to find the semantic relations between concepts. Large amount of text data is accounted by the knowledge discovery. Knowledge Discovery from Text (KDT) is generated from Natural Language Processing (NLP), carry out the methods from knowledge management. Discovery process is deployed for the rest. KDT plays a progressively significant role in trending applications, such as Text Understanding.

DOI:10.5121/ijsc.2016.7301

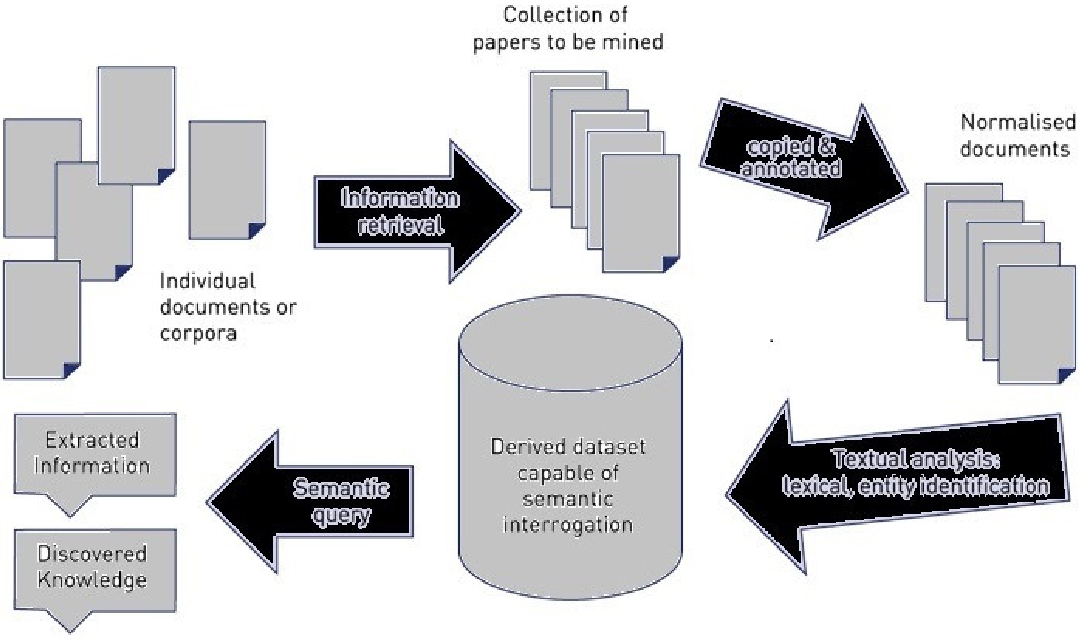

Collection of papers to be mined
Information retrieval
copied & annotated
Normalised documents
Individual documents or corpora
Derived dataset capable of semantic interrogation
Textual analysis: lexical, entity identification
Semantic query
Extracted Information
Discovered Knowledge

Figure 1 overall process of text mining

2. TECHNIQUES OF TEXT MINING

The text mining has numerous techniques to process the text. The main techniques are explained here.

2.1 Information Extraction

Information extraction is an initial step for unstructured text analysing [6]. Simplification of text is the work of information extraction. The main work is to recognize phrases and finds the relation between them. It is suitable for the bulky size of text. It extracts structured information from unstructured information. The figure 2 explains the information extraction.

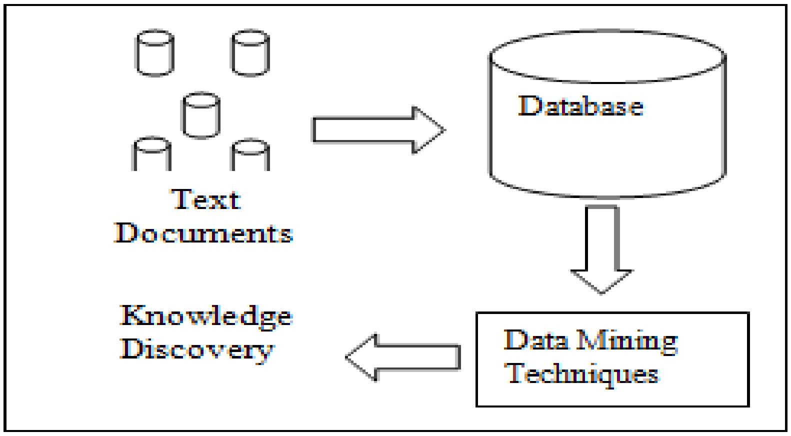


Figure 2 Information Extraction Processes

2.2 Clustering

Clustering focus towards the similarity measures on different objects and places, it has no predefined class labels. It segregate text into one group and in the same way generates cluster of group [4]. Words are isolated quickly and weights are assigned to each word. List of classes are generated by using clustering algorithms after calculating similarities.

2.3 Classification

Classification is to find the main theme of document by adding Meta and analysing document. The count of words and from that count decides the topic of the document which was done by the classification technique. It has predefined class label.

2.4 Information visualization

Instead of searching for extracting the patterns. They provide visual representation for text mining. Text mining used to perform particularly preparation of data, analysis & extraction of data, visualization mapping [19] on Information visualization. Zooming, scaling operations are used for user interaction with the document.

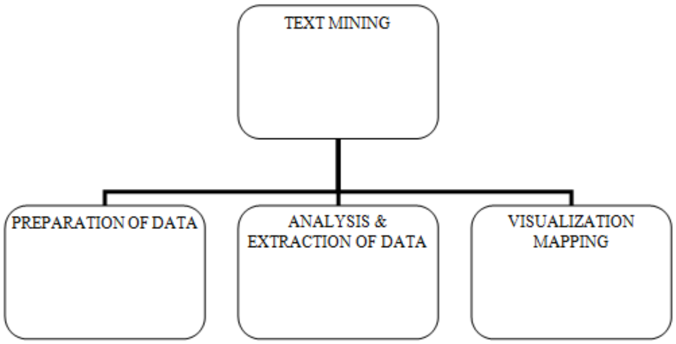


Figure 3 Information Visualization Processes

3. LITERATURE SURVEY

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 state-of-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.

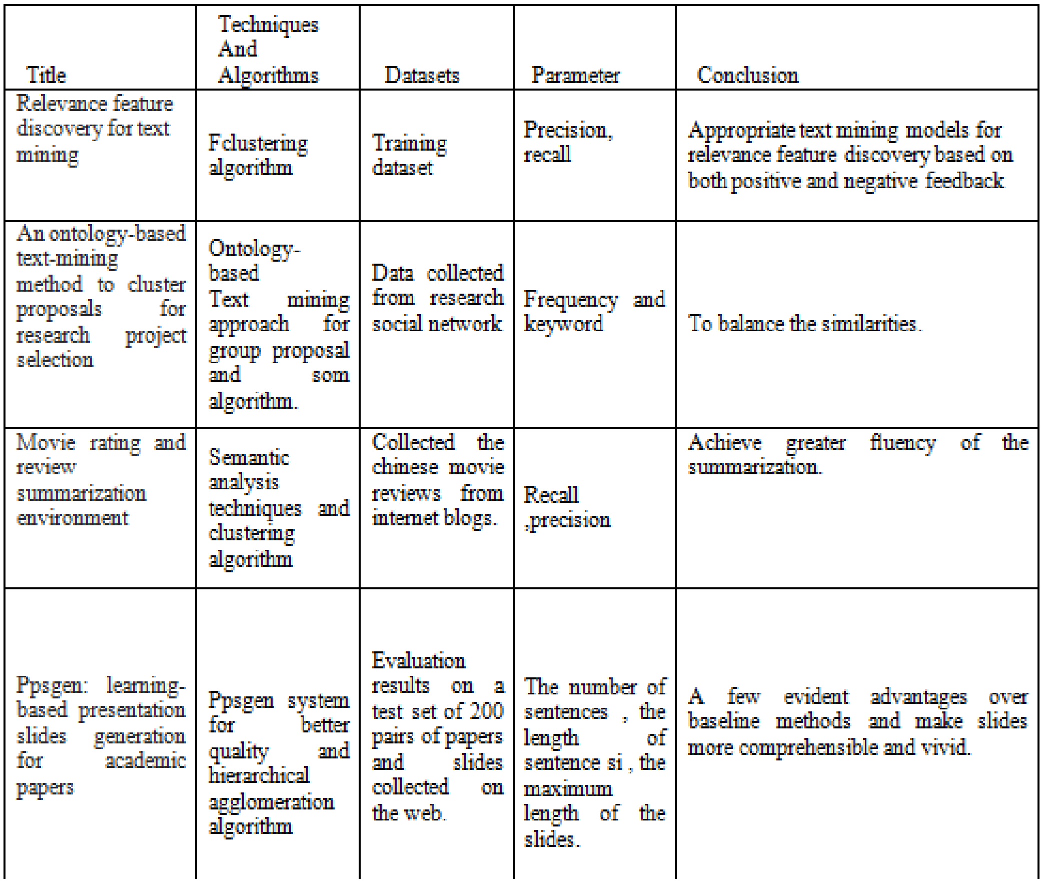
Jehoshua Eliashberg et al [11]: To forecast the box office performance of a movie at the crenulation point, it's suitable only if it holds the script and production cost. They extract textual features in three levels particularly genre and content, semantics, and bag-of- words from scripts using domain 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.

Charu C. Aggarwal et al [5]: Author focused on the Use of Side Information for Mining Text Data. an effective clustering approach was done by the classical partitioning algorithm with probabilistic models which was designed by the author. Dataset used is CORA, DBLP-four-area data set and IMDB. Running time and number of clusters are used as a parameter for analyzing purpose. The results can evident that the usage of side-information can improve the quality of text clustering and classification to sustain a high level of efficiency.

4. COMPARISONS ON DIFFERENT TEXT MINING TECHNIQUES



Title	Techniques And Algorithms	Datasets	Parameter	Conclusion
Relevance feature discovery for text mining	Fclustering algorithm	Training dataset	Precision, recall	Appropriate text mining models for relevance feature discovery based on both positive and negative feedback
An ontology-based text-mining method to cluster proposals for research project selection	Ontology-based Text mining approach for group proposal and som algorithm.	Data collected from research social network	Frequency and keyword	To balance the similarities.
Movie rating and review summarization environment	Semantic analysis techniques and clustering algorithm	Collected the chinese movie reviews from internet blogs.	Recall .precision	Achieve greater fluency of the summarization.
Ppsgen: learning-based presentation slides generation for academic papers	Ppsgen system for better quality and hierarchical agglomeration algorithm	Evaluation results on a test set of 200 pairs of papers and slides collected on the web.	The number of sentences , the length of sentence si , the maximum length of the slides.	A few evident advantages over baseline methods and make slides more comprehensible and vivid.

Commtrust: computing multi-dimensional trust by mining e-commerce feedback comments	Commtrust for trust evaluation by mining feedback techniques and lexical-lda algorithm.	ebay, amazon	N-value, trust score	Effectively address the "all good reputation" issue and rank sellers effectively
An ontology-based text mining method to develop d-matrix from unstructured text	The fault diagnosis ontology consisting of concepts and relationships commonly observed in the fault diagnosis domain. Next, we employ the text mining algorithms.	Real-life data collected from the automobile domain.	Fuel tank,hoses fuel	To construct the d-matrices by automatically mining the unstructured repair verbatim data collected during fault diagnosis. Each d-matrix as a graph and develop graph comparison algorithms such that the common patterns emerging from the heterogeneous d-matrices can be amalgamated to construct a single, comprehensive d-matrix
Assessing box office performance using movie scripts: a kernel-based approach	Kernel-based approach	300movie shooting scripts.	Portfolio roi(return of investment) ; number of movies in portfolio	The proposed methodology predicts box office revenues more accurately (29 percent lower mean squared error (mse)) compared to benchmark methods.
Text mining the contributors to rail accidents	A combination of techniques and forest algorithm	Total accident damage from 2001–2011	Accident cast,count	Advances in the use of text mining for train safety engineering.
Document clustering for forensic analysis: an approach for improving computer inspection	Document clustering algorithms to forensic analysis of computers seized in police investigations	Real-world investigation cases conducted by the brazilian federal police department.	Attributes, distance, Initialization, K_ estimate	That clustering algorithms indeed tend to induce clusters formed by either relevant or irrelevant documents, thus contributing to enhance the expert examiner's job.
On the use of side information for mining text data	Classical partitioning algorithms with probabilistic models.	Cora, dblp-four-area data set and imdb	Running time and number of clusters	The results show that the use of side-information can greatly enhance the quality of text clustering and classification,while maintaining a high level of efficiency.

5. CONCLUSION

Text mining technique is mainly used for extracting pattern from unstructured data. Knowledge discovery is mainly focused in this survey. The techniques are clustering, classification, and information extraction and information visualisation was overviewed. The process of text mining and the algorithms are also reviewed. In this paper various problems are surveyed and their solutions are discussed.

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.

[18] Silvana V. Aciar, Gabriela I. Aciar, Cesar Alberto Collazos, And Carina Soledad González," User Recommender System Based On Knowledge, Availability, And Reputation From Interactions In Forums", IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, Vol. 11, No. 1, February 2016.

[19] Yue Hu And Xiaojun Wan," Ppsgen: Learning-Based Presentation Slides Generation For Academic Papers", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 4, April 2015.

[20] Ning Zhong, Yuefeng Li, And Sheng-Tang Wu," Effective Pattern Discovery For Text Mining", IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.