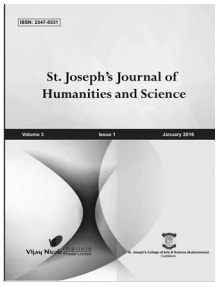
St. Joseph's Journal of Humanities and Science
ISSN: 2347 - 5331
http://sjctnc.edu.in/6107-2/



# Text Mining Techniques in Data Mining – Review

C. Christy[a*]
S. Arivalagan[b]
M.A. Maria Parimala[c]

## ABSTRACT

Discovering Knowledge in Databases and Extract Patterns and Knowledge in erroneous data is Data Mining. The quality of text details is extracted by text mining using Statistical Methods. Relevance, Novelty, Interestingness decides the accuracy of Text Mining. Categorization, clustering, entity extraction and sentiment analysis are used for text mining. Natural language processing, analytical methods related techniques, and algorithms are implemented.

**Keywords:** Data mining, Text mining, knowledge discovery. This survey is about the various techniques and algorithms.

## 1. INTRODUCTION

Text mining handles Textual data. It is Difficult to manipulate unstructured, unclear Textual Data. To Exchange the information a non-traditional retrieval strategy for information is used. For this process, Text mining is applied. Figure 1 shows the overall process of text mining. Nowadays, computers did the process better than human. The manual techniques are very expensive and time-consuming process. To achieve text mining, Extraction of information, conclude the information, tracking the topic, classification, and clustering technologies are used. For finding concepts both Implicit and Explicit, natural language processing (NLP) [8, 13] and semantic relation knowledge discovery text (KDT) is applied. To process knowledge management, natural language process (NLP) employed a great role in generating knowledge from text. Remaining is done by Discovery process. For understanding text, KDT is very important.

## 2. TEXT MINING TECHNIQUES

Various types of techniques involved in process of text mining. The techniques are Information Extraction, Clustering, Classification, Information visualization.

[a]Research Scholar, Department of Computer Science, Annamalai University, Chidambaram, Tamil Nadu, India.
[b]Department of Computer Science and Engineering, FEAT, Annamalai University, Chidambaram, Tamil Nadu, India.
[c]PG & Research Department of Computer Science, St.Joseph's College of Arts and Science (Autonomous), Cuddalore - 1.
[*]E-mail: vincentchristy4@gmail.com, Mobile: +91-9443878396.

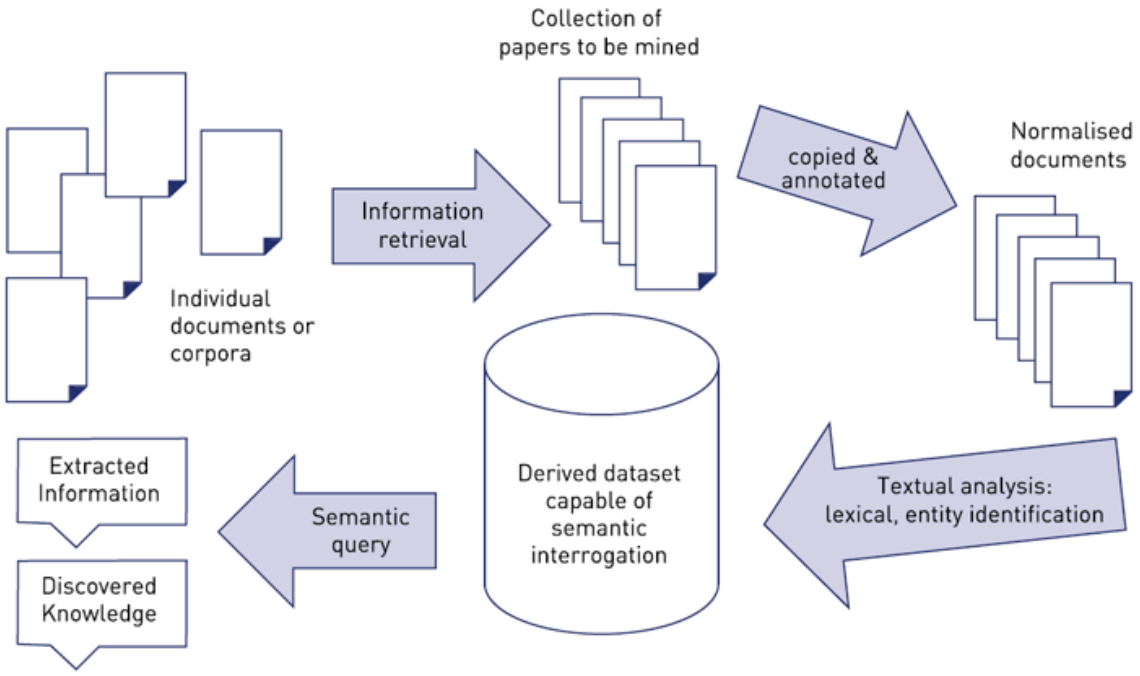



**Figure 1: The Overall text Mining Process**

## 2.1 Information Extraction

It's used for analyzing non formal text [6], and its simplification. The main process is understood the meaning and relation of the group of data. Formal information is derived from unformed information. Figure 2 Explains Information Extraction.

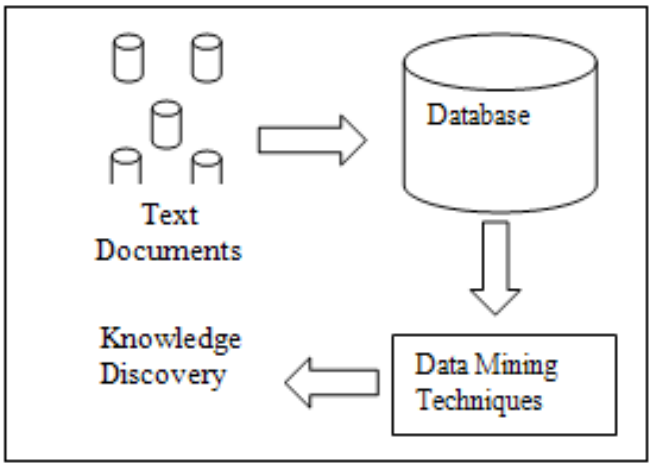



**Figure 2: Process of Information Extraction**

## 2.2 Clustering

Clustering Measures various kinds of information like objects, places etc. But not assigning any predefined Labels and classes. Grouping the text and the cluster of the group [4] generated by segregation. Separating and assign value for each word. Creating classes and similarities are calculated by clustering algorithms.

## 2.3 Classification

A huge collection of information is analyzed by Classification. Calculate the word count and topic of the collective information is decided by classification technique. It predefines the name of the class.

## 2.4 Information Visualization

Text mining is visually represented by this technique. Preparation of data, analysis & extraction of data, visualization mapping [19] on Information visualization is done by Information visualization. User's interaction with the document is based on scaling and zooming methods.

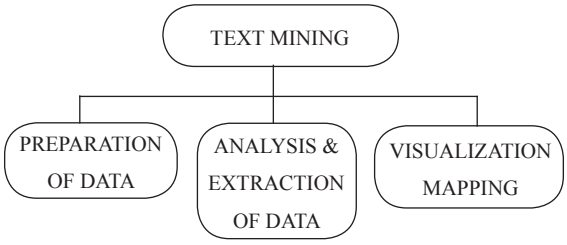



**Figure 3: Process of Information Visualization**

## 3. SURVEY OF LITERATURES

Yuefeng Li et AL [13]: classification method and Term based approaches are used by Text mining. The major issues are the problems of autonomy and synonymy.

The hypothesis is: Users prefers the comparison of pattern-based methods. One of the problems in text mining is the pattern of huge scale. In this work of clustering algorithm is used. Feedbacks for positive and negatives relevance based features are identified through methods of text mining.

Jianma et al [4]: classified automatically the text of English document is analyzed by the author. It's very difficult in other languages. In his work an approach of Ontology based text mining is used. SOM algorithm is efficiently developing Research proposals of both English and Chinese texts. This SOM algorithm helps to find absolute match of proposals and reviewers.

Chien - Liang Liu et al [2]: conclusion of this paper is sentiment-classification about rating of movies. The collaborated details of movie reviews depends feature based conclusions. Latent semantic analysis(LSA) Determines the features of product and minimize the summary size. Clustering algorithm is the best way of analyzing the classification of sentiment accuracy and calculating the response time during system design. implementation is based on NLP2 tool.

Xiuzhen Zhang et al [10]: The problem faced by all the reputation system is concentrated by the author. However, seller's reputation scores are large. Due to this buyers must select believable sellers. Comm. Trust is proposed by author to evaluate the trust. It's mostly based on mining of feedback through mining. Computation process is done by a model of multidimensional trust. Data sets are taken from E-BAY, AMAZON. In this work Lexical-LDA algorithm is used. Rank sellers got a large effective experiment through E-BAY, AMAZON data.

Dynaesh G.Rajpathak et al [9]: Finding new symptoms and failure of models by in time argumentation of D-matrix is the challenging task. The finding solution is based on fault diagnosis domain. In this work construct the concepts, relations abide with ontology of fault diagnosis. Through ontology find the needed facts and their dependencies from un formal verbatim repair data. Automobile domain is used for real life data collection and algorithms related to text mining are used. While composing of fault diagnosis, Ontology based text mining is applied to establish automatically mining the D-matrices of unformed repair verbatim data. For each D-matrices graph and its comparison algorithms are generated.

Johoshua Eliashberg et al [11]: the forecast of box office and crenulations point of movie performance is suitable only if it holds the script and the production cost. Approximately three levels are used for extract the features of text i.e. generate the content, Semantics, and bag of words from scripts used screen writing knowledge domain. Inputs are given by human and NLP techniques applied.

## 4. COMPARISONS OF VARIOUS TEXT MINING TECHNIQUES

| Title | Techniques and Algorithm | Datasets | Parameter | Conclusion |
|---|---|---|---|---|
| Relevance Feature Discovery for Text Mining | F clustering 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 methods 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 | Semantics analysis techniques and clustering algorithm | Collected the movie reviews from internet blogs | Recall, Precision | Achieve greater fluency of the summarization |

| | | | | |
|---|---|---|---|---|
| Learning based presentation slides generation for academic papers | System for better quality and hierarchical Agglomeration algorithm | Evaluation results on a test set of 100 pairs of papers and slides collected on the web | The number of sentences, the length of sentences, Maximum length of slides | A few evident advantages over base line methods and make slides more comprehensible. |
| Computing multi dimensional trust by mining E-Commerce feedback comments | Evaluation by mining feedback techniques and lexical – Ida Algorithm | E-bay, Amazon | N-value, trust score | Efficiently address the all good reputation" issue and rank sellers effectively |
| An Ontology Based Text mining method to develop d-matrix from un structured text | Concepts and relationships observed from fault diagnoses domain and apply text mining algorithm | Real life data collected from automobile domain | Fuel tank, hoses fuel | Mining the un structured repair verbatim data collected during fault diagnosis. Each D- Matrix is a graph and develop graph comparison algorithms, so common patterns emerging from the heterogeneous d-matrix can be construct a single comprehensive d-matrices |
| Assessing box office performance using movie scripts | Kernel based approach | 100 movie shooting scripts | Portfolio (Return of investment) number of movies in Portfolio | The proposed methodology predicts box office revenues more accurately compared to benchmark methods |
| Text mining contribution to rail accidents | A combination of techniques and combination and forest algorithms | Total accident damage from 2011 - 2012 | Accident cast, count | Advances in the use of text mining for train safety engineering |
| Document analysis 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 - Estimation | The clustering algorithm to induce clusters formed by either relevant or irrelevant documents contributing to enhance expert examiners job |

## 5. CONCLUSION

This paper provides a general idea of text mining techniques in various fields. To discover the knowledge from active data is the main objectives of data mining techniques. These applications use Clustering, Classification, information Extraction, and information visualization and so on. Review various classifications and clustering algorithm and its significance's is the future work.

## REFERENCES

[1] Luis Tari, Phan Huy Tu, Jorg Hakenberg, Yi Chen, Tran Cao Son, Graceiela Gonzalez, Information Extraction using Relational data, IEEE Transactions on Knowledge and Data Engineering, Vol, 24,No 1, January 2012.

[2] Chien-Liang Liu, Wen – Hoar Hsaio, Chia – Hong Lee, Gen - Chi Lu and Emery Jou, Movie Rating and Review Summarization in Mobile Environment, IEEE Transactions on Systems, Man, Cybernetics - Part C: Applications and Reviews. Vol.42, No.3, May 2012.

[3] Fuzhen Zhuang, Ping Luo, Zhiyoung Shen, Qing he, 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, January 2012.

[4] Jian Ma, Wei Xu, Yong – Hong Sun, EFRAIM TURBAN Shouyang Wang and Ou Liu, An OntologyBased Text Mining Method to Cluster Proposals for Research Project Selection , IEEE Transactions on Systems, Man, Cybernetics - Part A: Systems And Humans, Vol.42, No.3, May 2012.

[5] Charu C, Agarwal, Yuchen Zhao, abd 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] Luis Filipie Da Cruz Nassif and Eduardo Raul Hruschka, Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection, IEEE Transactions on Information Forensic and Security. Vol. 8, No 1, June 2013.

[7] Tsang, and Tak - Lam Wong, Discovering Low Rank Shared Concept S pace for Adapting Text Mining Models, IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 35, No 6, June 2013.

[8] Franciso Moraes Oliveira – Neto, LEE D. Han, and Myong Jeon, Online Self Learning Algorithm for License Plate Matching, IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No 4, June 2013.

[9] Dyyanesh G. Rajpathak and Satnam Sing., An Ontology Based Text Mining Method to Develop D-Matrix From Unstructured Text, IEEE Transactions on Systems, Man and Cybernetics Vol. 44, No.7, May 2014.

[10] Xiuzhen Zhang, Lishan Cui and Yan Wang, Multi-Dimensional Trust Mining E-Commerce Feedback Comments, Vol. 26, No.7, May 2014.

[11] Jehoshua Eliashberg, Sam K. Hui, and Z.John Zhang, A Kernal Based Approach on Accessing Box Office Performance Using Moviescripts, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No 11, November 2014.

[12] Riccardo scandariato, James Walden, Aram Hovsepyan and Joosen, Predictong Vulnerable Software Components Via Text Mining, IEEE Transactions on Software Engineering, Vol. 40, No 10, october 2014.

[13] Yuefeng Li, Abdulmohensen 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. IEEE Journal of Biomedical and Health Informatics, Extracting various classes of data from biological text using the concept of Existence Dependency ,Vol. 19, No 6, November.2015.

[15] Beichen Wang, Xiaodo0ng Chen, Hiroshi Mamitsuka and Zhu, Mining Medline for finding Experts in Biomedical Domains based on Language Model, IEEE / ACM Transactions on computational Biology and Bioinformatics. Vol. 12, No 6, November / December 2015.