Text Mining Approaches To Extract Interesting Association Rules from Text Documents

Vishwadeepak Singh Baghela [1], Dr. S.P.Tripathi [2]

[1] CSE Department, UTU Dehradun ,India
vdsbaghela@yahoo.com

[2] CSE Department ,IET, , GBTU Lucknow ,India
tripathee_sp@yahoo.co.in

Abstract

A handful of text data mining approaches are available to extract many potential information and association from large amount of text data. The term data mining is used for methods that analyze data with the objective of finding rules and patterns describing the characteristic properties of the data. The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification. In general, data mining deals with structured data (for example relational databases), whereas text presents special characteristics and is unstructured. The unstructured data is totally different from databases, where mining techniques are usually applied and structured data is managed. Text mining can work with unstructured or semi-structured data sets
A brief review of some recent researches related to mining associations from text documents is presented in this paper.

Keywords: *Text Mining, Association Rule Mining, Information Extraction, Natural Language Processing, Knowledge Discovery from Database*

1. Introduction

The emerging field of data mining promises to provide new techniques and intelligent tools to encounter these challenges [4]. The term "Data Mining" also known as Knowledge Discovery in Databases (KDD) is formally defined as: "the non-trivial extraction of implicit, previously unknown, and potentially useful information from large amount of data" [5]. Data mining [6], [7], [8], [9] as a multidisciplinary joint effort from databases, machine learning, and statistics, is championing in turning mountains of data into nuggets. The term data mining is used for methods that analyze data with the objective of finding rules and patterns describing the characteristic properties of the data. The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification. Data mining techniques have increasingly been studied [10], especially in their application in the real-world databases. The ultimate goal of a data mining task in a real-world application might be e.g. to allow a corporation either to improve its marketing, sales, and customer support operations or to identify a fraudulent customer through better understanding of its customers. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, bioinformatics, information retrieval, adaptive hypermedia, electronic commerce and network management [11] [12], [13, 50].

In recent times, the amount of textual information available in electronic form is growing at staggering rate. The best example of this growth is the World Wide Web (WWW), which is estimated to provide access to at least three terabytes of text (that is, three million megabytes). Even in commercial and private hands text collection sizes which were unimaginable a few year ago are common now, and the challenge is to efficiently mine interesting patterns, trends and potential information that are of interest to the user [35]. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text [20]. It is therefore crucial that a good text mining model should retrieve the information that users require with relevant efficiency [38]. In general, data mining deals with structured data (for example relational databases), whereas text presents special characteristics and is unstructured [18]. The unstructured data is totally different from databases, where mining techniques are usually applied and structured data is managed [19]. Text mining can work with unstructured or

semi-structured data sets such as emails, full-text documents and HTML files and more [21]. Some general approaches about text mining and knowledge discovery in texts can be found in [22], [23], [24]. Text mining shares many characteristics with classical data mining, but differs in many ways [25].

- Many knowledge discovery algorithms defined in the context of data mining, are irrelevant or ill suited for the textual application
- Special mining tasks, such as concept relationship analysis, are unique to text mining.
- The unstructured form of the full text necessitates special linguistic pre-processing for extracting the main features of the text

.

Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining [26]. In text mining approaches, initially the unstructured text documents are processed using natural language processing techniques to extract keywords labeling the items in that text documents. Then, classical data mining techniques are applied on the extracted data (keywords) to discover interesting patterns. Starting with a collection of documents, a text mining process would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted [19].

For mining large document collections, it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing than a plain text file [28]. Text preprocessing classically means tokenization and then Part of Speech Tagging [29] or in a bag of words approach word stemming and the application of a stop word list. Tokenization is the process of splitting the text into words or terms. Part of Speech (PoS) Tagging tags words according to the grammatical context of the word in the sentence, hence dividing up the words into nouns, verbs and more [30]. This is important for the exact analysis of relations between words, as it is needed in the extraction of relations between the texts [31]. Most text mining objectives fall under the following categories of operations: Search and Retrieval, categorization (supervised classification), summarization, Trends Analysis, Associations Analysis, Visualization and more [17].

Association is a powerful data analysis technique that appears frequently in data mining literature [36], [37]. Since its introduction by Agrawal *et al*., the task of association rule mining has received a great deal of attention [41]. Today the mining of such rules is still one of the most popular pattern-discovery methods in KDD. Association Rule Mining (ARM) is the process of discovering collection of data attributes that are statistically associated in the underlying data. Association rules "aim to extract interesting correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories". An association rule generated is of the structure *A-->B,* where *A* and *B* are disjoint conjunctions of attribute-value pairs. Association rule generation is a two-step process. First, minimum support is applied to find all frequent itemsets in a database. In second step, the frequent itemsets and the minimum confidence constraint are used to form rules. The main advantages of association rules are simplicity, intuitiveness and freedom from model-based assumptions. The important application of association rule mining is market basket analysis which is a famous tool among retail enterprises, for example they inform the user about items most likely to be purchased by a customer during a visit to the retail store. They are widely used in many other areas such as telecommunication networks, market and risk management, inventory control and more [39].

2. Motivations for the Research

Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted great attention with an imminent need for turning such data into useful information and knowledge. The popularity of the Web and the large number of documents available in electronic form has motivated the search for hidden knowledge in text collections. Consequently, there is growing research interest in the general topic of text mining [52]. Usually, the text is a collection of unstructured documents with no special requirements for composing the documents. Massive wealth of knowledge is embedded in these texts and waiting to be discovered and extracted. Thus there is a great need for efficient and effective techniques to process these texts in order to extract knowledge and discoveries significant for the advancement of science and technology [51]. Traditionally, text documents have been mainly analyzed by natural language processing techniques. Generally, classical data mining techniques are typically applied to large databases of highly structured information in order to

discover new knowledge. Therefore, there is a pressing need for developing efficient approaches to handle and mine information from unstructured data. One way of providing shallow understanding on the text corpora is with the use of information extraction techniques.This mining process will be ineffective if the samples are not a good representation of the larger body of the data.Therefore an important part of the process is the verification and validation of patterns on others samles of data.[60]

Information extraction systems can be used to directly extricate abstract knowledge from a text corpus, or to extract concrete data from a set of documents which can then be further analyzed with traditional data-mining techniques to discover more general patterns [27]. In short, information extraction is the task of locating desired pieces of data from a natural language document. Since information extraction addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge. Many text mining methods have been developed in order to achieve the goal of retrieving useful information for users [45, 46, 47, 48, 49,]. Most text mining methods use the keyword-based approaches, whereas others choose the phrase technique to construct a text representation for a set of documents. It is believed that the phrase-based approaches should perform better than the keyword-based ones as it is considered that more information is carried by a phrase than by a single term. Based on this hypothesis, Lewis [56] conducted several experiments using phrasal indexing language on a text categorization task. Ironically, the results showed that the phrase-based indexing language was not superior to the word-based one.

Although phrases carry less ambiguous and more succinct meanings than individual words, the likely reasons for the discouraging performance from the use of phrases are: (1) phrases have inferior statistical properties to words, (2) they have a low frequency of occurrence, and (3) there are a larger number of redundant and noisy phrases among them [38]. In recent times, extracting semantic relationships among entities in text documents has gained enormous popularity. Association rules are interesting patterns that are discovered from a given dataset. The earliest form of association rule mining is the market-basket analysis, which searches for interesting relationships between shoppers and item bought, for example. Mining association rules in transaction databases has been demonstrated to be useful in several application areas. However, its application on text databases is still very challenging because characteristics of text and transaction databases are different. This leads to the motivation of this research, that is to apply association rule mining to text databases to capture the relationships among words (terms) [44]. Association rules have been researched and applied extensively, in diverse domains and applications [1-3, 40, 42]. In text mining, extracted rules can be interpreted as co-occurrences of terms in texts and consequently are able to reflect semantic relations between terms [55]. In general, association rules highlight correlations between features in the texts, e.g. keywords. Moreover, association rules are easy to understand and to interpret for an analyst or may be for a normal user. However, it should be mentioned that the association rule extraction is of exponential growth and a very large number of rules can be produced. The extracted association rules identify the relations between features in the documents collection. The scattering of features in text contribute to the complexity of define features to be extracted from text. These kinds of features relationships can be better described with the association rule mining of text [17]. Several researchers have presented algorithms and approaches for mining associations from text document collections [14, 15, 32, 33].

3. Review of Related Works

A handful of text mining approaches are available in the literature for mining potential information and associations from large collections of text documents. Owing to the exponential increase in the volume of text document collections and the need for analyzing text documents, developing efficient approaches for mining trends, deviations and associations from text documents has received a great deal of attention in research communities. A brief review of some recent researches related to mining associations from text documents is presented here.

Hany Mahgoub [15] has presented a system for discovering association rules from collections of unstructured documents called EART (Extract Association Rules from Text). The EART system has treated texts only not images or figures. EART discovered association rules amongst keywords labeling the collection of textual documents. The main characteristic of EART is that the system has integrated XML technology (to transform unstructured documents into structured documents) with Information Retrieval scheme (TF-IDF) and Data Mining technique for association rules extraction. EART is depended on word feature to

extract association rules. It consisted of four phases: structure phase, index phase, text mining phase and visualization phase. His work depends on the analysis of the keywords in the extracted association rules through the co-occurrence of the keywords in one sentence in the original text and the existing of the keywords in one sentence without co-occurrence. Experiments applied on a collection of scientific documents selected from MEDLINE that are related to the outbreak of H5N1 avian influenza virus.

Hany Mahgoub *et al.* [16] have described that the text mining technique for automatically extracting association rules from collections of textual documents. The technique called, Extracting Association Rules from Text (EART). It depends on keyword features for discover association rules amongst keywords labeling the documents. In their work, the EART system ignores the order in which the words occur, but instead focusing on the words and their statistical distributions in documents. The main contributions of the technique are that it integrates XML technology with Information Retrieval scheme (TF-IDF) (for keyword/feature selection that automatically selects the most discriminative keywords for use in association rules generation) and use Data Mining technique for association rules discovery. Experiments applied on WebPages news documents related to the outbreak of the bird flu disease. The extracted association rules contain important features and describe the informative news included in the documents collection. The performance of the EART system compared with another system that has been used the Apriori algorithm throughout the execution time and evaluating extracted association rules.

Liang-Chih Yu *et al.* [43] have proposed a framework that combines a supervised data mining algorithm and an unsupervised distributional semantic model to discover association language patterns. The data mining algorithm, called association rule mining, was used to generate a set of seed patterns by incrementally associating frequently co-occurring words from a small corpus of sentences labeled with negative life events. The distributional semantic model was then used to discover more patterns similar to the seed patterns from a large, unlabeled web corpus. Suneetha Manne,and S. sameen Fatima [53] have proposed the method of Text Categorization on web documents using text mining and information extraction based on the classical summarization techniques. First web documents were preprocessed to establish an organized data file, by recognizing feature terms like term frequency count and weight percentage of each term. Experimental results were showed, that approach of Text Categorization was more suitable for Informal English language based web content where there was vast amount of data built in informal terms. That method had significantly reduced the query response time, improved the accuracy and degrees of relevancy.

Pablo F. Matos *et al.* [54] have addressed the problem of extracting and processing relevant information from unstructured electronic documents of the biomedical domain. The documents were full scientific papers. That problem imposed several challenges, such as identifying text passages that contain relevant information, collecting the relevant information pieces, populating a database and a data warehouse, and mining these data. For that purpose, that paper have proposed the IEDSS-Bio, an environment for Information Extraction and Decision Support System in Biomedical domain. In a case study, experiments with machine learning for identifying relevant text passages (disease and treatment effects, and patients number information on Sickle Cell Anemia papers) showed that the best results (95.9% accuracy) were obtained with a statistical method and the use of preprocessing techniques to resample the examples and to eliminate noise.

Chenn-Jung *et al.* [57] have proposed a financial news headline agent to assisting the investors in deciding to buy and to sell stocks in Taiwan market after receiving the essential real-time news headline disseminated by the agent. Weighted association rules and text mining techniques were used to derive the significance degree of each newly arrived news headline on the fluctuation of Taiwan Stock Exchange Financial Price Index on the next trading day. The experimental results revealed that the proposed work indeed achieves significant performance and demonstrate its feasibility in the applications of real-time information dissemination, such as financial news headlines via Internet. Sophia Ananiadou Jung *et al.* [58] have summarized the methods that were currently available, with a specific focus on protein–protein interactions and pathway or network reconstruction. The approaches described will be of considerable value in associating particular pathways and their components with higher-order physiological properties, including disease states.

Yue Dai *et al.* [59] have proposed MinEDec, a decision-support model that combines two well-known and widely-used CI analysis models into a unified model. CI analysis by using this unified model was supported by the use of state-of-the-art

TM technologies. They have also outlined the architecture of a DSS that was based on the MinEDec model and applies various TM technologies. First, they explained that the purpose of our MinEDec model was to transform data into useful knowledge. They then described the functions of SWOT analysis and the FFA framework in a new model for monitoring the business environment. Although there were several CI software tools available, none of them combines TM and several widely accepted CI analysis methods. The proposed model was unique as it analyses the five objectives from the perspective of nine SWOT factors by using TM technologies. Based on this, they have proposed a way of integrating SWOT and FFA models into a unified decision-support model.

Fei Wu and Daniel S. Weld [34] have paper presented WOE, an open IE system which improves dramatically on TextRunner's precision and recall. The key to WOE's performance was a form of self-supervised learning for open extractors - using heuristic matches between Wikipedia infobox attribute values and corresponding sentences to construct training data. Like TextRunner, WOE's extractor eschews lexicalized features and handles an unbounded set of semantic relations. WOE was operated in two modes: when restricted to POS tag features, it runs as quickly as TextRunner, but when set to use dependency-parse features its precision and recall rise even higher.

4. Proposed Methodology

The primary intention of my research is to design and devise approaches for mining potential information and interesting associations among large collections of text documents. The input to the text mining approach will be collection of text documents. Most important approaches to text mining involve the use of Natural Language Processing (NLP) for information extraction. Generally, text documents are a source of unstructured information implausible to be processed by traditional data mining techniques. Hence, Information extraction (IE) is meant to distill structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. Another integral part of the text mining systems is the KDD, which considers the application of statistical and machine-learning methods to discover novel relationships in large relational databases. Both the topics KDD and IE are of significant research interest. There have been so many researches undertaken in these topics aiming to devise innovative techniques for text mining and also to resolve issues faced by the traditional techniques. In recent times, the task of discovering associations, patterns of co-occurrences and significant relationships amongst keywords labeling the items in a collection of textual documents has gained immense importance in text mining.

Thereby, in my research, I am very much interested in extracting associations from unstructured text collections. First, I have intended to perform a detailed study on some of the existing efficient text mining systems that integrate methods from Information Extraction (IE) and Data Mining (Knowledge Discovery from Databases or KDD) for mining potential information and associations. Based on the study, I will devise approaches, either by utilizing some of the effectual IE and KDD techniques available in the literature or by developing innovative techniques, for mining potential information and associations amongst the keywords labeling the items in a text documents collection. Furthermore, the significant measure will be newly proposed to mine the important rules from the text documents instead of support and confidence. Finally, the efficient mining procedure will be devised to mine the important association rules and then, the *information extracted* from the association rules will be analyzed,and the experimentation will be conducted using the text databases and the result illustrate the efficiency of the proposed algorithm. Finally, verification and validation of the proposed algorithm will be tested on other sample of data by using some statistical techniques.

5.Conclusion

There have been so many researches undertaken in these topics aiming to devise innovative techniques for text mining and also to resolve issues faced by the traditional techniques. In recent times, the task of discovering associations, patterns of co-occurrences and significant relationships amongst keywords labeling the items in a collection of textual documents has gained immense importance in text mining. Based on the study, It is possible to devise a new approach, either by utilizing some of the effectual IE and KDD techniques available in the literature or by developing innovative techniques, for mining potential information and associations amongst the keywords labeling the items in a text documents collection. Furthermore, the significant measure of newly proposed techniques is to mine the important rules from the text documents instead of support and confidence.

References

[1] Dong, Jianning, Perrizo, William, Ding, Qin and Zhou, "The Application of Association Rule Mining to Remotely Sensed Data", In proceedings of the ACM Symposium on Applied computing, Vol.1, pp. 340–345, 2000.

[2] Brijs, Tom, Swinnen, Gilbert, Vanhoof, Koen and Wets, "Using Association Rules for Product Assortment Decisions: A Case Study", In proceedings of Knowledge Discovery and Data Mining, pp. 254–260, 1999.

[3] Satou, Shibayama, Ono, Yamamura, Furuichi, Kuhara and Takagi, "Finding Association Rules on Heterogeneous Genome Data", In Proceedings of the Second Pacific Symposium on Biocomputing, pp. 397–408, 1997.

[4] Frawley, W., Piatetsky-Shapiro, G., Matheus, C., "Knowledge Discovery in Databases: An Overview", AI Magazine, fall 1992, pp. 213-228, 1992.

[5] Paolo Palmerini, "On performance of data mining: from algorithms to management systems for data exploration", Technical Report, Universit`a Ca' Foscari di Venezia, 2004.

[6] R. Agrawal, T. Imielinski, and A. Swami, "Database Mining: A Performance Perspective," IEEE Transaction Knowledge and Data Engineering, Vol. 5, No. 6, pp. 914-925, 1993.

[7] U.M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, 1996.

[8] J. Han and Y. Fu, "Attribute-Oriented Induction in Data Mining," Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, pp. 399-421, 1996.

[9] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman publishers, 2001, ISBN: 1-55860489-8.

[10] Chen and Liu, "Data mining from 1994 to 2004: an application-oriented review", International Journal of Business Intelligence and Data Mining, Vol.1, No.1, pp.4-11, 2005.

[11] Klaus Julisch, "Data Mining for Intrusion Detection -A Critical Review", Application of Data Mining In Computer Security, Kluwer Academic Publisher, Boston, 2002.

[12] Hewen Tang, Wei Fang and Yongsheng Cao, "A simple method of classification with VCL components", In Proceedings of the 21st international CODATA Conference, 2008.

[13] Qiankun Zhao and Sourav S. Bhowmick, "Sequential Pattern Mining: A Survey" Technical Report Technical Report Center for Advanced 7 Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore, 2003.

[14] Xiaowei Xu, Mutlu Mete and Nurcan Yuruk, "Mining concept associations for knowledge discovery in large textual databases", In Proceedings of the ACM symposium on Applied computing, pp.549 - 550, 2005.

[15] Hany Mahgoub, "Mining Association rules from unstructured documents", World Academy of Science, Engineering and Technology, Vol.20, No.1, pp.1-6, 2006.

[16] Pegah Falinouss, "Stock Trend Prediction using News Articles", Technical Report, Lulea. University of Technology, 2007.

[17] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey, "A Text Mining Technique Using Association Rules Extraction", International Journal Of Computational Intelligence, Vol.4, No.1, pp.21-28, 2008.

[18] Miguel Delgado, Maria J. Martin-Bautista, Daniel Sanchez, J. M. Serrano and Maria Amparo Vila Miranda, ""Association Rule Extraction for Text Mining", Lecture Notes in Computer Science (LNCS), Vol.2522, pp.154-162, 2002.

[19] Vishal Gupta and Gurpreet S.Lehal, "A Survey of Text Mining Techniques and Applications", Journal Of Emerging Technologies in Web Intelligence, Vol.1, No.1, pp.60-76, August 2009.

[20] Navathe, Shamkant and Elmasri Ramez, "Fundamentals of Database Systems", Pearson Education, Singapore, 2000.

[21] Delgado, Martin-Bautista, Sanchez and Vila, "Mining text data: special features and patterns", In proceedings of EPS Exploratory workshop on pattern Detection and Discovery in Data mining, London, UK, September 2002.

[22] Feldman, Fresko, Kinar, Lindell, Liphstat, Rajman, Schler and Zamir, "Text mining at the term level", In proceedings of 2nd European Symposium of Principles of Data mining and Knowledge Discovery, pp.65-73, 1998.

[23] Kodratoff, "Knowledge discovery in texts: A Definition and Applications", Proc. of the 11th International Symposium on Foundations of Intelligent Systems, pp.16-29, 1999.

[24] Landeau, Aumann, Feldman, Fresko, Lindell, Lipshat and Zamir, "TextVis: An integrated Visual Environment for Text mining", In Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Nantes, September 1998.

[25] Ah-hwee Tan, "Text Mining: The state of the art and the challenges", In Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, pp. 65-70,1999.

[26] Nasukawa and Nagano, "Text Analysis and Knowledge Mining System", IBM Systems Journal, Vol.40, No.4, pp.967-984, October 2001.

[27] Raymond J. Mooney and Razvan Bunescu, "Mining knowledge from text using information extraction", ACM SIGKDD Explorations Newsletter, Vol.7,No.1, pp.3-10, 2005.

[28] Hotho, Nurnberger and Paass, "A Brief Survey of Text Mining Export", LDV Forum, Vol.20, No.2, pp.19-62, 2005.

[29] Manning and Schütze, "Foundations of statistical

natural language processing", MIT Press 1999.

[30] Shatkay and Feldman, "Mining the Biomedical Literature in the Genomic Era: An Overview", Journal of Computational Biology, Vol.10, No.6, pp.821-855, 2003.

[31] Daraselia, Yuryev, Egorov, Novichkova, Nikitin and Mazo, "Extracting human protein interactions from MEDLINE using a full sentence parser", Bioinformatics, Vol.20, No.5, 2004.

[32] Xin Chen and Yi-Fang Wu, "Personalized Knowledge Discovery: Mining Novel Association Rules from Text", In proceedings of 6th SIAM International Conference on Data Mining, Bethesda, MD, USA, April 2006.

[33] Tien Dung Do, Siu Cheung Hui and Alvis C.M. Fong, "Associative Feature Selection for Text Mining", International Journal of Information Technology, Vol. 12, No.4, pp.59-68, 2006.

[34] Fei Wu, Daniel S. Weld, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,pp.118-127,2010.

[35] Ricardo Baeza-Yates, Alistair Moffat and Gonzalo Navarro, "Searching large text collections", Handbook of massive data sets, pp.195-243, ISBN:1-4020-0489-3, 2002.

[36] Pak Chung Wong, Paul Whitney and Jim Thomas, "Visualizing Association Rules for Text Mining", in proceedings of IEEE symposium on Information Visualization", pp.120, 1999.

[37] Fatudimu, Musa, Ayo and Sofoluwe, "Knowledge Discovery in Online Repositories: A Text Mining Approach", European Journal of Scientific Research, Vol.22 No.2, pp.241-250, 2008.

[38] Sheng-Tang Wu, Yuefeng Li and Yue Xu, "Deploying Approaches for Pattern Refinement in Text Mining", in proceedings of the Sixth IEEE International Conference on Data Mining, pp. 1157-1161, 2006.

[39] Qiankun Zhao, Sourav S. Bhowmick "Association Rule Mining: A Survey ", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003.

[40] Michail and Amir, "Data Mining Library Reuse Patterns Using Generalized Association Rules", In proceedings of International Conference on Software Engineering, pp.167–176, 2000.

[41] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Jorge B. Bocca, Matthias Jarke, Carlo Zaniolo, eds., in proceedings of 20th International Conference on Very Large Data Bases, pp. 487-499, Santiago, Chile, 1994.

[42] Bench-Capon, Frans, Coenen, and Leng, "An Experiment in Discovering Association Rules in the Legal Domain", In Proceedings of the Workshop on Legal Information Systems and Applications, pp. 1056–1060, 2000.

[43] Liang-Chih Yu, Chien-Lung Chan, Chao-Cheng Lin, I-Chun Lin," Mining association language patterns using a distributional semantic model for negative life event classification," Journal of Biomedical Informatics,Vol.44,no. 4, August, 2011.

[44] Alisa Kongthon, "A Text Mining Framework for Discovering Technological Intelligence to Support Science and Technology Management", Technical Report, Georgia Institute of Technology, April 2004.

[45]K. Aas and L. Eikvil. Text categorisaton: A survey. Technical report, Norwegian Computing Center, Raport NR 941, 1999.

[46]Edda and Jorg, "Text categorization with support vector machines. How represent texts in input space?", Machine Learning, Vol.46, pp.423-444, 2002.

[47]Lam, Ruiz and Srinivasan, "Automatic text categorization and its application to text retrieval", IEEE Transactions on Knowledge and Data Engineering, Vol.11, No.6, pp.865-879, 1999.

[48] Sebastiani, "Machine Learning in automated text categorization", ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002.

[49] Shapire and Singer, "Boostexter: a boosting-based system for text categorization", Machine Learning, Vol.39, pp.135-168, 2000.

[50] Lee, Stolfo and Mok, "A Data Mining Framework for Building Intrusion Detection Models", In proceedings of IEEE Symposium on Security and Privacy, pp. 120–132, 1999.

[51]Hisham Al-Mubaid and Rajit K Singh, "A New Text Mining Approach for Finding Proteinto-Disease Associations", American Journal of Biochemistry and Biotechnology, Vol.1, No.3, pp.145-152, 2005.

[52]Un Yong Nahm and Raymond J. Mooney, "Text mining with information extraction",PreQuestUMI,pp.218,ISBN:0-496-01283-5 2004.

[53]Suneetha Manne,Dr. S. sameen Fatima,"A Novel Approach for Text Categorization of Unorganized data based with Information Extraction,"International Journal on Computer Science and Engineering (IJCSE),Vol. 3 No. 7,July 2011.

[54] Pablo F. Matos , Leonardo O. Lombardi, Thiago A. S. Pardo, Cristina D. A. Ciferri, Marina T. P. Vieira, and Ricardo R. Ciferri," An Environment for Data Analysis in Biomedical Domain: Information Extraction for Decision Support Systems," Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part I,pp.306-316,2010.

[55] Valentina Ceausu and Sylvie Despres, "Text Mining Supported Terminology Construction", In proceedings of the 5th International Conference on Knowledge Management, Graz, Austria, 2005.

[56] D.D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task", In proceedings of SIGIR, pp: 37-50, 1992.

[57] Chenn-Jung,Huang,Jia-Jian,Liao,Dian-Xiu,Yang,Tun-Yu,Chang,Yun-Cheng Luo," Realization of a news dissemination agent based on weighted association rules and text mining techniques, "Journal Expert Systems with Applications,Vol.37,no.9, September, 2010.

[58]Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, Douglas B. Kel,"Event extraction for systems biology by text mining the literature,"Trends in

Biotechnology,vol.28,no.7,pp.381-390,2010.

[59] Yue Dai, Tuomo Kakkonen, Erkki Sutinen,"MinEDec: a Decision-Support Model That Combines Text-Mining Technologies with Two Competitive Intelligence Analysis Methods," International Journal of Computer Information Systems and Industrial Management Applications,vol.3pp.165-173,2011.

[60] VDS Baghela, Dr. S.P. Tripathi,"A Survey on Association Rules in Case of Multimedia Data Mining," International Journal of Computer Sience and Technology,vol.3, Issue 1 , pp.649-652, March 2012.

Mr. VDS Baghela received his graduation degree in statistics from BHU, Varanasi in 1999 and completed MCA Degree from AAIDU, Allahabad in 2004. He obtained M.Tech (CSE) degree from UPTU, Lucknow in 2010. Now he is Persuing Ph.D (CSE) under supervision of Dr. S.P. Tripathi. Currently he is working as an Associate Professor in IT Department at IIMT College of Engg, Greater Noida, UP, India. His research area is Data Mining

Dr. SP. Tripathi has completed MSc degree from Allahabad university in 1979 and M.Tech(CSE) degree from IIT-Delhi in 1985 .He has obtained Ph.D degree in CSE in 2005 from Lucknow university.He is an associate professor in CSE department at Institute of Engineering and Technology Lucknow, a constituent college of Gautam Buddh technical university. He has more than 25 years experience in teaching and research. His area of research is data mining, data base and operating system.