KNOWLEDGE DISCOVERY IN TEXT MINING: A REVIEW

Ms. Vaishali Bhujade[1], Prof. N. J. Janwe[2] , Prof S.W.Mohod[3]

[1]*B.D.C.O.E. Sevagram,* [2]*R.G.C.E.R.T. Chandrapur ,*[3]*B.D.C.O.E. Sevagram*

[1]*vaishali.bhujade@rediffmail.com*

***Abstract*- The term 'Data Mining' refers to the finding relevant and useful information from databases. The term knowledge discovery in databases (KDD) and data mining are often used interchangeably another name of this process is discovering useful (hidden) patterns in data. Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. Text mining and Data mining are similar in some sense. One of the Data mining activities which involve extracting meaningful new information from the data is Association rules. The task f association rule is to detect relationship or association between specific values of categorical variables in large data sets. Knowledge discovery in databases (KDD) is the process of finding useful information and pattern in data. The objective of this paper is to provide a deeper insight of the work done till date in data mining and KDD from the text document. Also this paper describes text mining technique for automatically extracting association rules from collection of textual documents, the technique called, Extracting Association Rules from Text (EART). It depends on keyword features for discover work, the EART system ignores the order in which the words occur, but instead focusing on the words and their statistical distribution in documents.**

***Keywords*-Data Mining, Text Mining , Knowledge Data Discovery, Association Rules.**

I. INTRODUCTION

The access to a large amount of textual documents become more & more effective due to the growth of the Web, digital libraries, technical documentation, medical data, …

These textual data constitute resources that it is worth exploiting. In this way knowledge discovery from textual databases, or for short, text mining(TM), is an important and difficult challenge, because of the richness and ambiguity of natural language. Therefore, the problem is the existing of huge amount of textual information available in textual form in databases and online sources. So the question is who is able to read and analyze it? In this context, manual analysis and effective extraction of useful information are not possible. We think the solution is that it is relevant to provide automatic tools for analyzing large textual collections by automatically find relevant information, analyze relevant information and structure relevant information.

Knowledge discovery in databases (KDD) is a process that is defined by several processing steps that have to be applied to a data set of interest in order to extract useful patterns. These steps have to be performed iteratively and several steps usually require interactive feedback from a user. Text mining is an increasingly important research field because of the necessity of obtaining knowledge from the enormous number of text documents available, especially on the web. Text mining and data mining, both included in the field of information mining, are similar in some sense, and thus it may seem that data mining techniques may be adopted in straight forward way to mine text. However, data mining deals with structured data, whereas text presents special characteristics and is unstructured.

II. RELATED WORK

Xin Chen and Yi-Fang WU [6] has presented a text mining technique that discovers novel association rules from documents from a particular user. The system derives user background knowledge from his/her background Documents, and exploits such knowledge to evaluate the novelty of discovered knowledge in the form of association rules by measuring the semantic distance between the antecedent and consequent of a rule in the background knowledge. The experiment result shows that the proposed measure has high novelty prediction, accuracy and usefulness. Rakesh Agrawal and Ramakrishnan Srikant [7] has present two new algorithm for mining

association rules, In this paper, two algorithm are discussed to solve the problem arises during association rule between item in a large database. This algorithm is Apriori and Aprioritid which are totally different from AIS algorithm and SETM algorithm. The researcher also shows how the best features of the two proposed algorithm can be combined into a hybrid algorithm, called apriorihybrid. The Researcher presented experimental results; showing that the proposed algorithm always outperform mining association rules between sets of items in large databases (AIS) and SET oriented mining of association rules (STEM). Ronen Feldman, Haym Hirsh [8] has described the FACT (Finding Associations in Collection of Texts) system for knowledge discovery in collections of textual documents, which discovers associations amongst the keywords labeling the items in a collection of textual documents. In addition, FACT is able to use background knowledge about the keywords labeling the document in its discovery process. Rather than forcing the user to specify on explicit query repression in some arcane knowledge discovery query language. FACT presents the user with an easy to use graphical interface in which discovery tasks can be specified, with the language providing a well defined semantics for the discovery actions performed by a user through the interface. The author finds association amongst the keywords labeling the documents given background knowledge about the keywords and relationship between them. M. Rajman, R. Besancon [9] has proposed the general framework of knowledge discovery, Data mining techniques. Dedicated to information extraction from unstructured textual data and natural language processing(NLP). They proposed two different text mining tasks, association extraction from a collection of indexed documents design to answer to specific queries expressed by the user and prototypical document extraction find information about classes of repetitive document structure that could be used for automated synthesis of the information content of textual base. The computerized exploration of large amount of data and on discovery of interesting patterns within them Ronen Feldman et.at. [10] has focuses on them. They present an approach to perform text mining at term level. They describe the term extraction module of the document explorer system, and provide experimental evaluation performed at a set of 52000 documents published by return in the year 1995-1996. Term level text mining attempts to benefit from the advantages of these two extremes. On one hand there is no need for human effort in tagging document, and they do not loose most of the information present in the document as in the tagged documents approach. Thus the system has the ability to work on new collections without any preparation, as well as the ability to merge several distinct collections into one. The number of meaningless results and the execution time of the mining algorithms are greatly reduced. Working on the term level also enables the construction of a hierarchical taxonomy which is extremely important to a text mining system. Ronen Feldman and Ido Dagan [11] proposed using a text categorization paradigm to annotate text articles with meaningful concepts that are organized in hierarchical structure. They suggest that this relatively simple annotation is rich enough to provide the basis for a KDD framework, enabling data summarization, exploration of interesting patterns, and trend analysis. Their research combines the KDD and text categorization paradigms and suggests advances to the state of the art in both areas. They have presented a new framework for knowledge discovery in texts. This framework is based on three components: The definition of a concept hierarchy, the categorization of texts by concepts from the hierarchy, and the comparison of concept distributions to find "unexpected" patterns. They conjecture that their uniform and compact model can become useful for KDD in structured databases as well. They conjecture that the concept distributions of

articles marked as interesting by the user can be used for updating the user's personal news profile and for suggesting subscribing to news groups of similar characteristics. The mining process described by Ronen Feldman, Moshe Fresco et. al. [12] they have focused on the computerized exploration of large amount of data and on the discovery of interesting pattern within them. They proposed an intermediate approach, one that call text mining at the term level, in which knowledge discovery takes place on a more focused collection of word and phrases that are extracted from and label each document. These terms plus additional higher–level entities are then organized in a hierarchical taxonomy and are used in the knowledge discovery process. They describe tool that implement text mining at the term level. Experimental result shows that text mining servers as a powerful technique to manage knowledge encapsulated in large document collection. Text mining at the term level thus hits a useful middle ground on the quest for understanding the information present in the large amount of the data that is only available in textual form. Text mining at the term level serves as a powerful technique to manage knowledge encapsulated in large document collection. The survey of basic concept in the area of text data mining and some of the method used in order to elicit useful knowledge from collection of textual data done by Jan Paralic, Peter Bednar [13]. Three different text data mining techniques (clustering/visualization, association rule and classification model) are analyzed and its exploitation possibilities within the Webocracy Project are observed. The WEBOCRAT system will support communication and discussion, publication of documents on the Internet, browsing and navigation, opinion polling on questions of public interest, intelligent retrieval, analytical tool, alternating services, and convenient access to information based on individual needs. Helena Ahonen.ai. [14] They has proposed that general data mining methods are applicable to text analysis tasks such as description phrase extraction. They present a general frame work for text mining. The frame work follows the general knowledge discovery process, thus containing steps from preprocessing to the utilization of the results. The data mining methods apply is based on generalized episodes and episode rules. They presented example application from information retrieval and natural language processing. The applicability of these approach was demonstrated with experiments on real-life data, showing that episodes and episode rules produced discriminate between documents. Both pre- and post processing have essential roles in pruning and weighting the result. How to mine association rule in temporal document collection and describe how to perform the various steps in the temporal text mining process, including data cleaning, text refinement, temporal association rule mining and rule post-processing is given by Kjetil Norvag et. al. [15]. Author also describe the temporal text mining testbench and the experiments are based on a relatively small collection, consisting of 38 days of the front page of the on line version of Financial Times. The size of the dataset is relatively limited and it is therefore not expected that really interesting rules will be found using this dataset. The operator used in the experiment extract terms, extract nouns, remove stop words, and stem words, weight term, filter terms, FITI. In order to determine the quality of the mined rule evaluation criteria have to be defined. In these experiments focus will be on the quality of the rule found. Hany Mahgoud [16] described a system for discovering association rules from collections of unstructured documents. The technique called EART (Extracted Association Rule from Text). The work depends on the analysis of the keywords in the extracted association rules through the co-occurrence of the keyword in one sentence in the original text and existing of the keywords in the sentence without co-occurrence. EART system implements using C# and XML. Author applied

this experiment on the 'abstract' part of collection on Medline documents but system can be applied on over all part of the document. So proposed system is flexible to work on all parts or specific parts of documents. Author studied here text categorization using decision tree. Instead of using words, word relation i.e. association rule from these words, is used for building decision tree byMohammad Masud Hasan, Chowdhury Mofiuere Rahman [17]. In this experiment, processes data and then find out association relation amongst these words using Rakesh Agrawal et. al.'s Apriory algorithm applying objective interestingness measures. using the decision tree generator software of Quinlan's C4.5 system. The experimental result obtained made more impressive if increase the number of attributes (i.e. association rule). This can be obtained by decreasing the support level and/or confidence level. But in that case running time would increase. In fact the more association rule use, it does not necessarily mean that the number of input texts will also improve the categorization. Dr. C. A. Dhote and Ms. Prachitee B. Shekhawat [19] introduces a novel approach for text mining and association rules are extracted from unstructured document and taken from fifteen text documents for knowledge discovery process, related to terror attack in India. Author approach seems to perform well as compared to the previous work. The proposed method for text extraction using association rules performs well as it is compared with the earlier approaches using Apriori Algorithm and the EART system. A utilization of modern methods of data mining is describe A. Vesely [20] and especially method based on neural networks theory are pursued. Author discuss advantages and drawbacks of applications of multilayer feed forward neural networks and kohonen's self organizing maps. Author proposed from method based on neural networks, kohonen's self-organizing maps are the most promising.

III. PROPOSED WORK

In Text Mining system generate association rules from unstructured documents. These documents available in various text formats. There is need of converting documents into XML format. Remove all unimportant words that are stored in blockade list. After this process steaming is done, a process that removes a words prefixes and suffixes. Indexing is need to extract keywords, which distinguish the document from remaining documents. The Term frequency, Inverse Document Frequency techniques is used for automated production of indexes associated with documents. Apply algorithm on the document indexes to generate all keywords that have support greater than minimum support. Lastly association rules are generated. The strength of an association rule can be measured in terms of it support and confidence. As shown in fig.

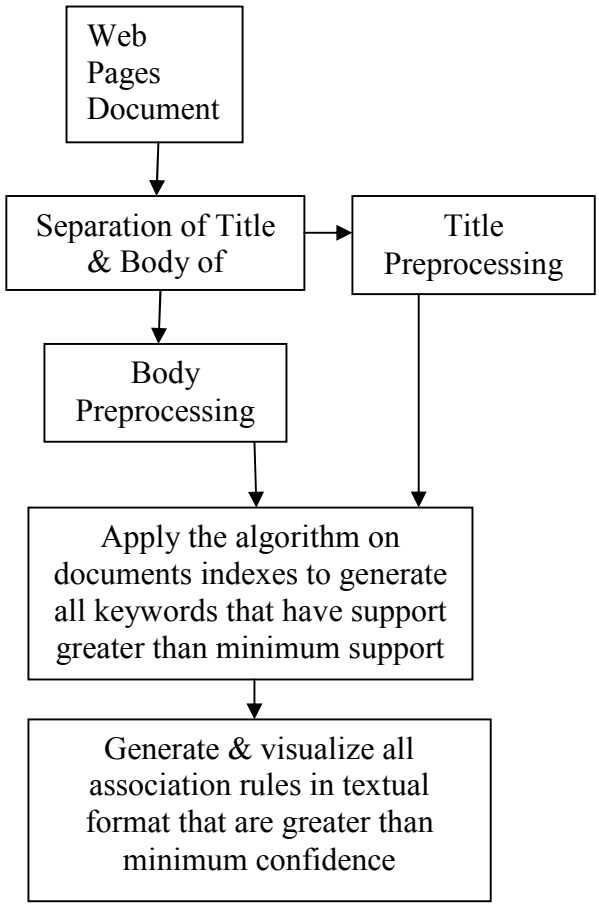

Fig:- Text Mining System Architecture

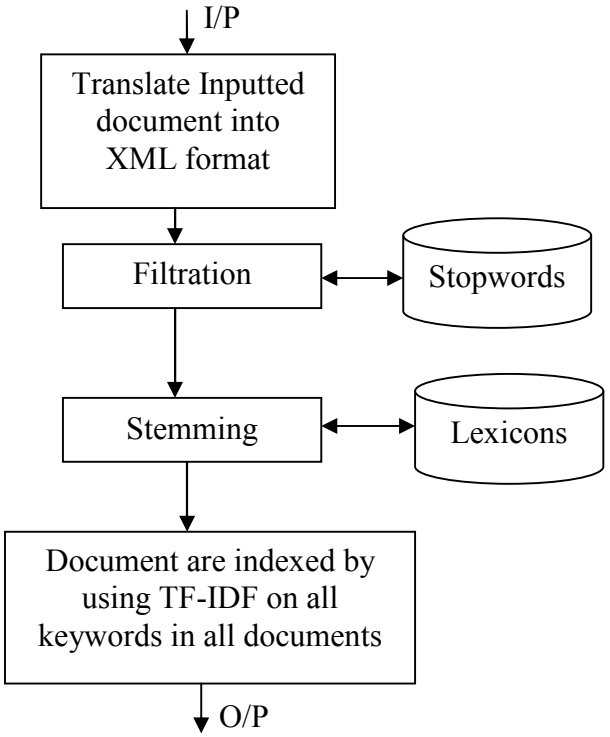

Fig:-Text Preprocessing Phase

The propose text mining system, Extracting Association Rules from Text (EART) is shown in fig 1. It automatically discovers association rules from textual documents. The main contributions of the system are that, it integrates XML technology with an Information Retrieval Scheme (TF-IDF) and with data mining techniques for association rules extraction. The EART system ignores the order in which the word occurs, but instead focusing on the words and their statistical distributions. The system begins with selecting collections of documents from the web or internal file systems. The EART system consists of three phases: Text Preprocessing Phase (Transformation, Filtration, Stemming and Indexing of the documents), Association Rule Mining (ARM) Phase, and Visualization Phase (Visualization of results).

IV. CONCLUSION

The paper is focused on presenting the literature published in the area of data mining in general and KDD in particular. The significance of generating association rules is a vital parameter in deciding other parameters in KDD. Our proposed work in generating rules is based on apriori algorithm using neural network technique.

V. REFERENCES

[1]. J.Han and M.Kamber, *Data Mining: Concepts and Techniques*, 2nd ed., Sanfransisco, CA: Elsevier Publication, ISBN: 978-81-312-0535-8, 2008, pp.234-261.

[2]. Arun K. Pujari, *Data mining : Techniques,* Universities Press (India) ISBN 81-7371-380-4, 2006

[3]. Pieter Adriaans and Dolf Zantinge, *Data Mining,* Pearson Education, ISBN: 81-7808-425-2, 1996,pp.37-78.

[4]. Agrawal, R., Imielinski, T., and Swami, A. "Mining Association Rules between sets of Items in Large Databases". Proc. of the 1993 ACM SIGMOD Conf. on *Management of Data*, 1993

[5]. Agrawal, R., Srikant, R. "Fast Algorithms for Mining Association Rules". Proc. of the 20th Int'l Conf. on *Very Large Data Bases*, 1994, pp. 487-499.

[6]. Xin Chen, Yi-Fang WU. "Personalized Knowledge Discovery: Mining Novel Association Rules from Text" available at ww.siam.org/meetings/sdm06/proceeding/ 067chenx.pdf.

[7]. Rakesh Agrawal and Ramakrishnan Srikant. "Fast Algorithm for mining Association Rules" Proc. of the 20th VLDB conference Santiago, Chile, 1994.

[8]. Ronen Feldman, Haym Hirsh. "Mining Association in Text in the Presence of Background Knowledge" Proc. 2nd International Conference on Knowledge Discovery from Databases, 1996.

[9]. M. Rajman, R. Besancon. "Text Mining: Natural language techniques and Text mining Applicationn" Proc. 7th working Conf. on database semantic (DS-7), Chapan & Hall IFIP Proc. Series. Leysin, Switzerland Oct. 1997, pp7-10.

[10]. Ronen Feldman, Haym Hirsh. "Knowledge Management: A Text Mining Approach" Proc. of the 2nd Int'l Conf. on *Practical Aspects of Knowledge Management*, Basel, Switzerland, 1998.

[11]. Ronen Feldman and Ido Dagan. "Knowledge Discovery in Textual Databases (KDT)" Proc. of the 1st Int. Conf. on Knowledge Discovery and Data Mining, 1995.

[12]. Ronen Feldman, Moshe Fresco. "Text Mining at The Term Level" Proc. Of the 2nd European symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Vol.1510, Nantes pp 65-73.

[13]. Jan Paralic, Peter Bednar. "Text Mining for Documentes Annotation and ontology Support" (A book chapter in : "Intelligent system at service of Mankind," ISBN 3-935798-25- 3, Books, Germany, 2003).

[14]. Helena Ahonen, Oskari Heinonen, Mika Klemettinen, A Ikeri Verkamo. "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collection".

[15]. Kjetil Norvag, Trond Oivind Eriksen, and Kjell-innge Skogstad."Mining Association rules in temporal document

collections," Available at http://www.idi.ntnu.no/~noervaag/papers/ISMIS2006.pdf

[16]. Hany Mahgoud. "Mining Association Rules from Unstructured Document" Internation Journal of Applied mathematics and Computer Science Vol. 4, No 4. Paper Submittred 22-06-2006.

[17]. Mohammad Masud Hasan, Chowdhury Mofiuere Rahman. "Text Categorization Using Association Rule Based Decision Tree". Proc. of the 6th Int. Conf. on Computer and Information Technology (ICCIT), Bangladesh, 2003, pp. 453-456.

[18]. Dr. C. A. Dhote and Ms. Prachitee B. Shekhawat. "Mining Association Rules from Unstructured Documents" Internation Journal of Computer Applications in Engineering, Technology and Science (IJ-CA-ES), Oct. 2009, pp. 551-556.

[19]. H. Mahgoub, D. Roesner, Nabil Ismail and Fawzy Torkey. "A Text Mining Technique Using Association Rules Extraction". International Journal of Computational Intelligence vol 4 Number 1 2007 ISSN 1304-2386

[20]. A. Vesely "Neural networks in data mining" AGRIC. ECON. – CZECH, 49, 2003 (9): 427-431.Pages Information Extraction Based on Web", Journal of Luoyang Technology College,3(2005) 30-31 (in Chinese