International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 01 Issue 01 | 2013

Knowledge Generation in Scientific Database using Text Mining

Ms. Mrunali L. Vaidya, Prof. M. S. Chaudhari
P.B.C.C.O.E, Nagpur, India

Abstract

Knowledge generation from scientific database has received increasing attentions recently since huge repositories used for development of digital database and internet world. In a corpus of scientific database such as a digital library, scientific articles, scientific subject and so on.At present, the stored information is increasing tremendously day by day. The sudden increase in the amount of texts on the web, it was almost impossible for people to keep up-to-date information. Knowledge generation from textual database referred generally to the process of extracting interesting or non-retrieval patterns or knowledge from unstructured text documents. Using the technique such as information extraction, information retrieval, natural language processing, text mining can be easily found from the corpus of documents set. **Knowledge generation in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.** ***The development of proposed work is the acquirization or selection of target data set, integration and checking of data set, data cleaning, preprocessing and development of transformation model and selection of algorithm which gives generated knowledge as result interpretation, visualization, testing, verification and maintenance. Text Mining is the automatic discovery of previously unknown information by extracting information from text.***

***Index Terms*— Knowledge generation, Corpus, Text mining, Information Extraction, Information Retrieval**

I. INTRODUCTION

A vast amount of online information appears in collections of unstructured text, which is the predominant medium for information exchange among people. The volume of available text resources (e.g., web pages and other online resources) requires techniques such as Information Extraction (IE) as a prerequisite for efficient location, retrieval, and management of relevant information. IE is commonly defined as extracting structured data from unstructured data as provided, for instance, in textual documents [1]. Text Mining is the automatic discovery of previously unknown information by extracting information from text [4]. It typically consists of (i) information retrieval (IR), which gathers and filters documents, (ii) IE, and (iii) data mining for discovering unexpected associations between known facts. Text mining builds largely on research on Knowledge Discovery in Databases (KDD), also known as Data Mining [4]. KDD applies statistical and machine learning methods to discover novel relationships in large relational databases. Standard data and text mining methodologies integrate an IE component that locates specific pieces of data in natural-language documents, and extracts structured information from textual resources and stores it in structured databases

Text mining is the discovery, which discovers the previously unknown information by extracting it automatically from different written sources. Text mining is similar to data mining except that data mining tools. Which are designed for handling the structured data but text mining can work with the [8]unstructured or semi structured data sets such as e-mails, full text documents, HTML files etc. Mostly the Information will be stored in natural form, is called text. Text mining is different from what are all familiar with in web mining. When user searches something in web that is already known and which is written by someone else. The main problem in web mining is purchasing all materials, which are not relevant to our search as well as it will not display unknown information but in text mining the main goal is to discover the unknown information something that no one knows that. The basic form of information is data, which is to be managed and mined in order to create knowledge.

With the overwhelming increase in the amount of texts on the web, it is almost impossible for people to keep abreast of up-to-date information. Text mining is a process by which interesting information is derived from text through the discovery of patterns and trends. Text mining algorithms are [7] used to guarantee the quality of extracted knowledge. However, the extracted patterns using text or data mining algorithms or methods lead to noisy patterns and inconsistency. Thus, different challenges arise, such as the question of how to understand these patterns, whether the model that has been used is suitable, and if all the patterns that have been extracted are relevant. Furthermore, the research raises the question of how to give a correct weight to the extracted knowledge.To address these issues, this paper presents a text post-processing and mining includes different steps.

II. RELATED WORK

The objective behind this paper is to overcome the problem of variable terminology by the aid of concept-based information retrieval.[1] The work is done on systematic generation of concept maps with the help of text mining techniques. They have used GetSmart software package to draw conceptual maps, and develops a publicly accessible repository of concept maps to enable sharing of the knowledge.

Decision tree is used to find bigrams that occur nearby. They have evaluated his approach using the sense-tagged corpora from the SENSEVAL word sense [2] disambiguation exercise. They showed that bigrams are powerful features for performing word sense disambiguation. They have also proved that an effortless decision tree where each node checks whether or not a particular bigram occurs near the ambiguous word results in accuracy comparable with state-of-the-art methods.

A novel methodology to extract core concepts from text corpus [3], methodology is based on text mining and social network analysis. At the text mining phase the keywords are extracted by tokenizing, removing stop-lists and generating N-grams. Network analysis phase includes co-word occurrence extraction, network representation of linked terms and calculating centrality measure.

Business databases that have accumulated over many [9] years of business records typically contain a wealth of hidden knowledge that could potentially be utilized by management to make better informed business decisions. They described a framework for knowledge discovery from business databases and demonstrate how the discovered knowledge can be put into good use. The proposed framework uses an application that integrates data mining processes with online analytical processing (OLAP).

There is another proposed algorithm to extract knowledge using predictive Apriori [10] and distributed grid based Apriori algorithms for association rule mining. They presented the implementation of an association rules discovery data mining task using Grid technologies. As a result of implementation with a comparison of classic Apriori and distributed Apriori.

Information Extraction (IE) and knowledge discovery in [6] databases (KDD) were both useful approaches for discovering information in textual corpora, but they have some deficiencies. The aim is to provide a new high-quality information extraction methodology and, at the same time, to improve the performance of the underlying extraction system.

Text mining algorithms suggested the overwhelming increase in the amount [4] of texts on the web, it was almost impossible for people to keep abreast of up-to-date information. Text mining algorithms was used to guarantee the quality of extracted knowledge. However, the extracted patterns using text or data mining algorithms or methods lead to noisy patterns and inconsistency.

Text mining is used to extract the relevant information or knowledge or patterns from different sources, that are unstructured form. Text is unstructured form so it is [5] difficult to find text mining is used. The general framework of text mining consists of two different components; text refining that transforms free-form text documents into an intermediate form and knowledge distillation that deduces patterns or knowledge from intermediate form. Intermediate form (IF) can be semi structured such as the conceptual graph representation or structured such as the relational data representation. IF can be document based, here each entity represents the document or concept based, here each entity represents an object or concepts of interests in specific domain.

III. WORKING OF PROPOSED METHODOLOGY

Knowledge generation in scientific Database brings together current research on the exciting problem of generating useful and interesting knowledge in databases. Scientific database are corpus file collection which file having some scientific text or collection of word from which we have to generate the knowledge and find the best suitable file from repository. Knowledge generation in Databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. In very first phase, text preprocessing and mining, the methods include data collection; tokenization, stop list checking and generation of n grams that are still completed having short description are as below.

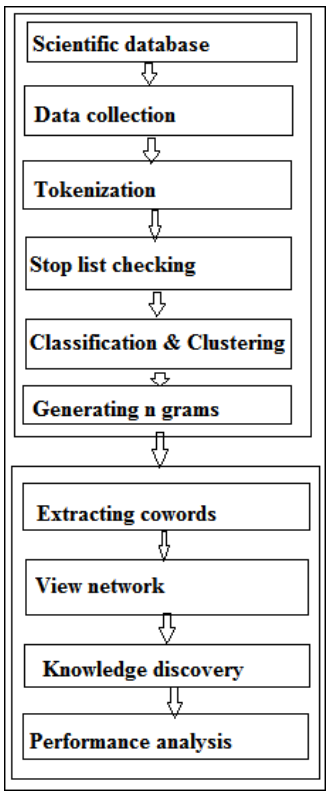

Fig.1. Proposed system for knowledge generation in scientific database

A. Data Collection and Tokenizing

In this step the text corpus is collected and preprocessed. Reuters, the real time data is collected from the website. After collection the tokenizing process is performed. The continuous text should be tokenized to words, phrases, symbols and other elements called tokens.

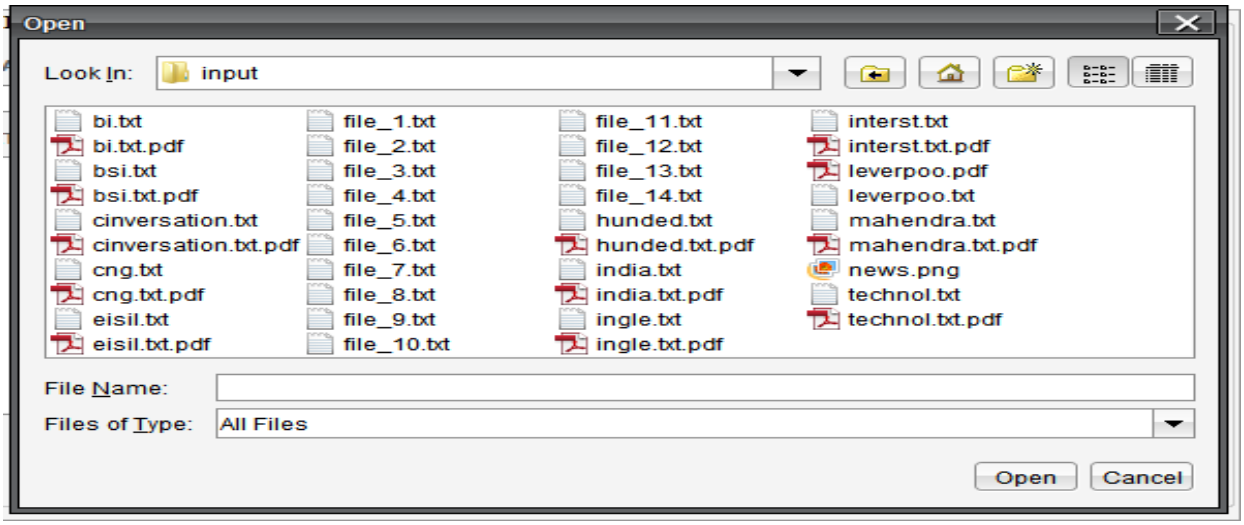

Fig.2. Data collection from text files

Fig.3. Processing the collected data

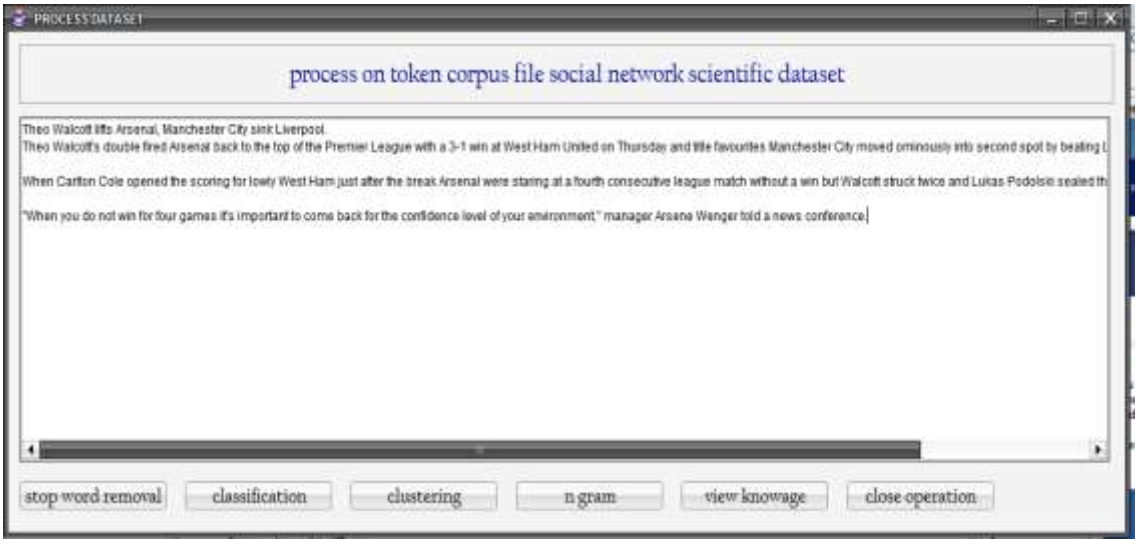

Fig.4. Tokenization of collected data

B. Stop words Removal

In this step the list of stop words are removed from the data set. Stop words are unnecessary words (example a, an, the, is, was). Stop-lists (or 'stop-words') are lists of non information bearing words.

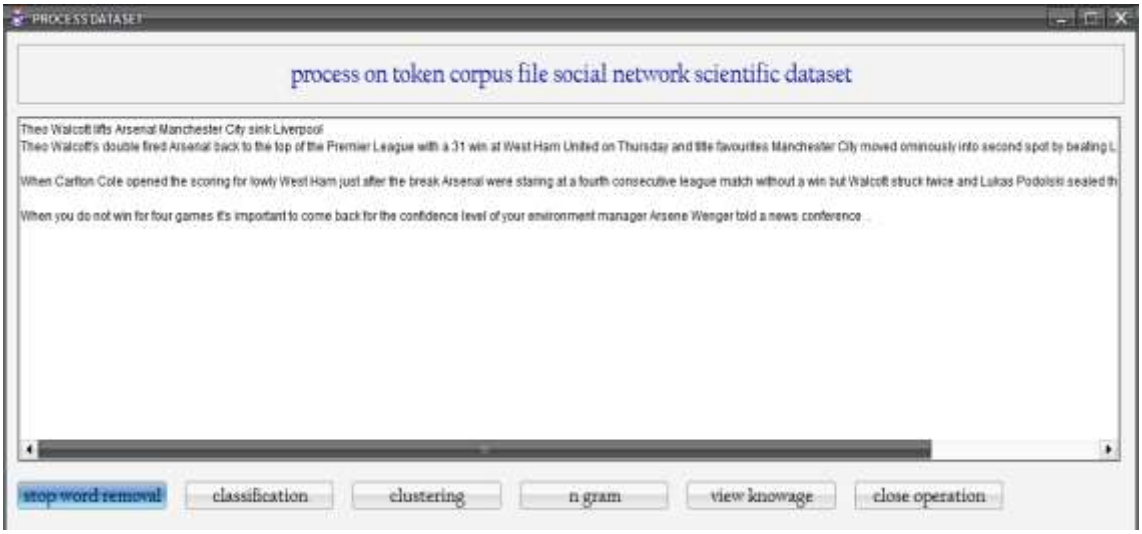

Fig.5. Stop word removal

C. Classification algorithm

The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. It is based on the Bayesian theorem. It is particularly suited when the dimensionality of the inputs is high. Parameter estimation for naive Bayes models uses the method of maximum likelihood. In spite over-simplified assumptions, it often performs better in many complex real world situations. Advantage is to require a small amount of training data to estimate the parameters. The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

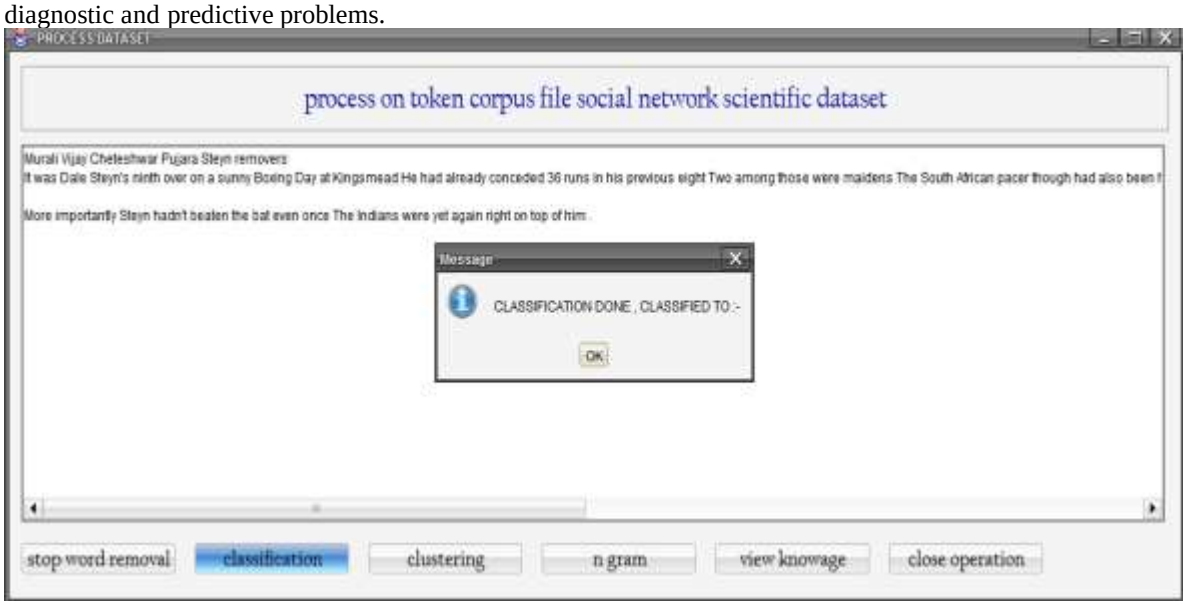

Fig.6. Classification on data

D. Clustering algorithm

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. Mahalanobis distance is a well known statistical distance function. Here, a measure of variability can be incorporated into the distance metric directly. Mahalanobis distance is a distance measure between two points in the space defined by two or more correlated variables. That is to say, Mahalanobis distance takes the correlations within a data set between the variable into consideration.

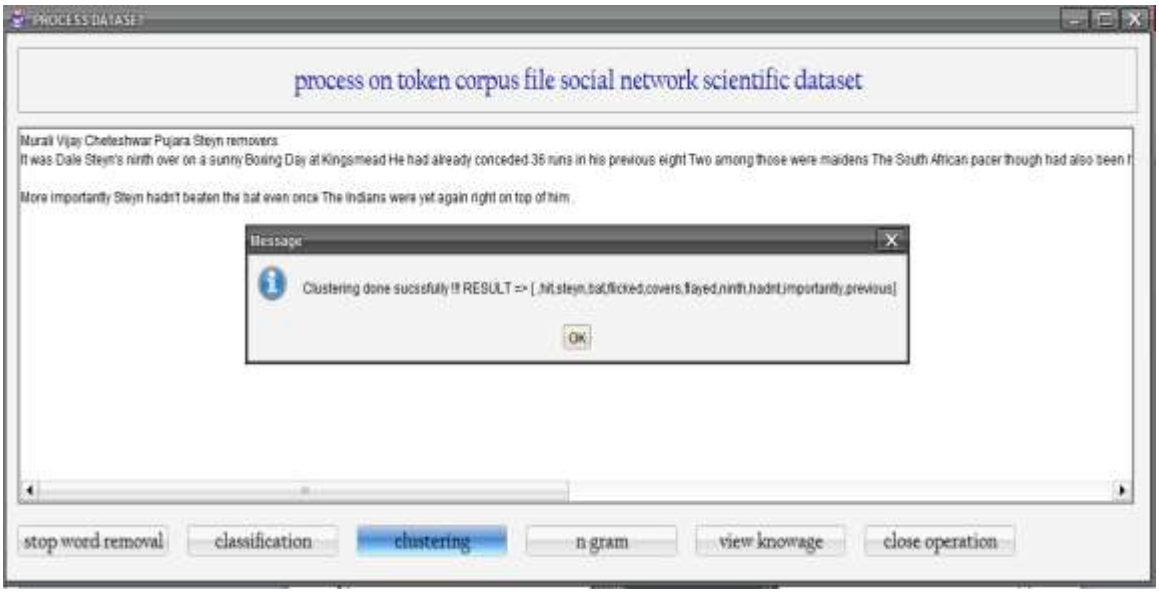

Fig.7. Clustering the data

IV. CONCLUSION

After the investigation of all the techniques for generation of knowledge in scientific database, it can be concluded that text mining and the social network analysis methods are used for generation of useful knowledge from large scientific databases. Overall we tried to achieved text preprocessing and mining phase. The future work is to achieve social network analysis phase. The text mining algorithms are effectively apply in the context of text data is critical for a wide variety of applications. Hence Social networks require text mining algorithms for a wide variety of applications such as keyword search, classification, and clustering. So the combination of above is important for knowledge generation.

REFERENCES

1. Ammar Jalalimanesh, "Knowledge Discovery in Scientific Databases UsingText Mining and Social Network Analysis",IEEE Conf. on Control, Systems and Industrial Informatics (ICCSII),PP.46-49,September 23-26,2012.
2. Zhen Guo, Zhongfei (Mark) Zhang et.al, "A Two-Level Topic Model towards Knowledge Discovery from Citation Networks", IEEE Transaction on knowledge and data engineering, PP.1-30, 2013.
3. K. M. Sam, C. R. Chatwin, "Ontology-Based Text-Mining Model For Social Network Analysis",PP.226-231,IEEE 2012.
4. Mubarak Albathan, Yuefeng Li et.al, "Using Patterns Co-occurrence Matrix for Cleaning Closed Sequential Patterns for Text Mining", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology,PP.201-205,2012.
5. M.Sukanya,S.Biruntha, "Techniques on Text Mining",IEEE International Conf. on Advanced Communication Control and Computing Technologies (ICACCCT), PP.269-271,2012.
6. Christina Feilmayr, "Text Mining-Supported Information Extraction", 22nd International Workshop on Database and Expert Systems Applications, PP.217-221,2011.
7. Xin Guo, Yang Xiang, Qian Chen, "A Vector Space Model Approach to Social Relation Extraction from Text Corpus", Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)PP.1756-1759,2011.
8. Vaishali Bhujade ,N.J.Janwe, "Knowledge discovery in text mining techniques using association rule extraction",international conf.on computational intelligence and communication system,PP.498-502, 2011.
9. A.C.M. Fong, "A Generalized Framework for Knowledge Discovery in Business Environments", Second International Conference on Communication Systems, Networks and Applications, PP.48-51,2010.
10. Mrs. R. Sumithra, Dr (Mrs). Sujni Paul, "Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery", Second International conference on Computing, Communication and Networking Technologies, PP.1-5, 2010.
11. M. Fritsch and M. Kauffeld-Monz, "The impact of network structure on knowledge transfer: an application of social network analysis in the context of regional innovation networks," The Annals of Regional Science,vol. 44, PP. 21-38, 2010.

12. T. Opsahl, et al., "Node centrality in weighted networks: Generalizing degree and shortest paths," Social Networks, vol. 32, PP. 245-251, 2010.
13. N. Santoro, et al., "Time-Varying Graphs and Social Network Analysis: Temporal Indicators and Metrics," 2011.
14. J. Jayabharathy , S. Kanmani , "Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature"
15. ,IEEE,PP.425-429,2011.
16. Michele Coscia, Fosca Giannotti, Ruggero Pensa, "Social Network Analysis as Knowledge Discovery process", IEEE,Advances in Social Network Analysis and Mining, PP. 279-283,2009.
17. Peter A. Gloor,Jonas Krauss et.al, "Web Science 2.0: Identifying Trends through Semantic Social Network Analysis", International Conference on Computational Science and Engineering ,IEEE,PP.215-222,2009.
18. Barahate Sachin R.,Shelake Vijay M, "A Survey and Future Vision of Data mining in Educational Field", Second International Conference on Advanced Computing & Communication Technologies,IEEE,PP.96-100,2012.
19. David Combe,et.al, "Combining relations and text in scientific network clustering", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, PP.1248-1253,2012.