International Journal of Recent Trends in Engineering & Research
ISSN(ONLINE) : 2455-1457
IMPACT FACTOR : 3.344

Text Mining: A Study Of Techniques and Applications

Palwinder Singh[1], Amarbir Singh[2]
[1]*Department of Computer Science, GNDU Amritsar*
[2]*Department of Computer Science, GNDU Amritsar*

Abstract— The Study of text mining is gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. Text Mining is very emerging research area. It is used to find new, previously unknown information by automatically extracting information from different unstructured resources. The unstructured texts contains vast amount of information and therefore is not easy to be used for future processing. Therefore many processing algorithms and techniques are important in order to extract the valuable information which is completed by using text mining. In this paper, we have discussed general idea of text mining, comparison of different text mining techniques and applications of text mining. In addition, to explaining the Text Mining concepts, the use of GA in text mining has been discussed.
Keywords—Text mining, Retrieval, Extraction, Clustering, Genetic Algorithm, Bioinformatics.

I. INTRODUCTION

Text mining is a process for extracting information and finding patterns from unstructured data. Information can be obtained from the summarized words of the documents and also the similarities between words and documents can be determined. Text mining is also known as text data mining, which is the process of deriving high-quality information from text. It is a process that employs a set of algorithms for converting unstructured text into structured data objects. The main purpose of text mining is to minimize the effort required by users to obtain useful information from large computerized text data sources [1]. Fig. 1 shows the Text mining process. From the huge amount of text documents first text preprocessing is done which obtain all words that are used in a given text and then a text document is splitted into a series of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. In second step text transformation means to convert text document into bag of words or vector space document model notation that can be used for further effective analysis [2]. Then next step is feature selection or attribute selection. In this phase the irrelevant features are discarded. It gives an advantage of less computations, smaller dataset size and minimum search space required.

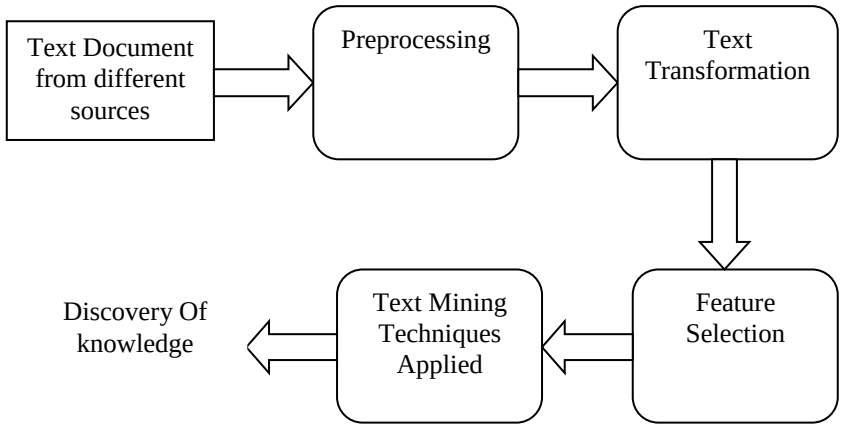

Figure 1 Text Mining Process

@IJRTER-2016, All Rights Reserved

II. DIFFERENCE BETWEEN DATA MINING AND TEXT MINING

The difference between normal data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. One application of text mining is in, bioinformatics where details of experimental results can be automatically extracted from a large corpus of text and then processed computationally. Text-mining techniques have been used in information retrieval systems as a tool to help users narrow their queries and to help them explore other contextually related subjects. Text Mining seems to be an extension of the better known Data Mining [3]. Data Mining is a technique that analyses billions of numbers to extract the statistics and trends emerging from a company's data. This kind of analysis has been successfully applied in business situations as well as for military, social, government needs. But, only about 20% of the data on intranets and on the World Wide Web are numbers - the rest is text. The information contained in the text (about 80% of the data) is invisible to the data mining programs that analyze the information flow in corporations. Text mining tries to apply these same techniques of Data mining to unstructured text databases. To do so, it relies heavily on technology from the sciences of Natural Language Processing (NLP), and Machine Learning to automatically collect statistics and infer structure and meaning in otherwise unstructured text. The usual approach involves identifying and extracting key features from the text that can be used as the data and dimensions for analysis. This process is called feature extraction, is a crucial step in text mining.

III. CHALLENGING ISSUES

The main challenging issue in the process of text mining is complexity of natural language. The natural language is not free from the ambiguity problem. One word may have multiple meanings and multiple words can have same meaning. The capability of being understood in two or more possible ways means ambiguity [4]. This ambiguity leads to noise in extracted information. Ambiguity cannot be entirely eliminated from the natural language as it gives flexibility and usability. There are various ways to interpret one phrase or sentence thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain. It is challenge to answer what user wants as semantic meanings of many discovered words are uncertain. Many researchers are working on resolving this problem, but problem still exists and proposed approaches are dedicated for specific domain [5]. The main merits of text mining are the names of different entities and relationship between them can easily be found from the corpus of documents set using the technique such as information extraction and the challenging problem of managing great amount of unstructured information for extracting patterns e is solved by text mining. As far as demerits are concerned, the information which is initially needed is no where written and To mine the text for information or knowledge no programs can be made in order to analyze the unstructured text directly.

IV. TEXT MINING TECHNIQUES

Text Mining is a technique of collecting information from unstructured data, converting the information received into structured data, identify the pattern from structured data, analyze the patterns and extract the valuable information and store in the database. The various techniques which are available for text mining are:

A. Information Retrieval

Information retrieval is the process of obtaining relevant information from a collection of various resources. Each user tries to locate documents that can provide information required and satisfy information needs for a particular user [6]. The process of acquiring, identifying and searching the possible documents that may meet this information need is called retrieval process. Hence information retrieval is defined as a set of methods and techniques for formulating information needs of the users in form of queries. For many applications challenging is electronic

@IJRTER-2016, All Rights Reserved

information is in the form of free natural language documents rather than structured databases like relational databases. Information extraction solves this problem of transforming a corpus of textual documents into a more structured database

B. Information Extraction

The Information Extraction is the process of extraction of useful information from text [7]. It identifies the extraction of entities, events and relationships from semi-structured or unstructured text. Most useful information such as name of the person, organization and location are gathered without proper understanding of the text. Then those are stored in database like patterns and are then available for further use.

C. Clustering

Clustering aims at finding intrinsic structures in information and then arranges them into useful subgroups for further study and analysis. Clustering is an unsupervised process in which objects are classified into groups called clusters. It basically organizes a large amount of documents into a number of meaningful clusters, document clustering can be used to browse a collection of documents or organize the results returned by a search engine in response to a user's query. In data mining K-means is frequently used clustering algorithm in text mining field also it obtains good results [8]. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. The organization of management information systems makes use of clustering technology as organizational database contain thousands of documents.

D. Visualization

In text mining visualization methods can improve and simplify the discovery of relevant information. To represent individual documents or groups of documents text flags are used to show document category and to show density colors are used. Visual text mining puts large textual sources in a visual hierarchy. The user can interact with the document by zooming and scaling. Information visualization is applicable to government to identify terrorist networks or to find information about crimes. Following fig.3 shows steps involved in visualization process [9]. The goal of information visualization divided into three steps: Data preparation step includes deciding and obtaining original data of visualization and form original data space, The process of analyzing and extracting visualization data needed from original data and to form visualization data space is known as Data analysis and extraction and Visualization mapping step employ certain mapping algorithm to map visualization data space to visualization target.

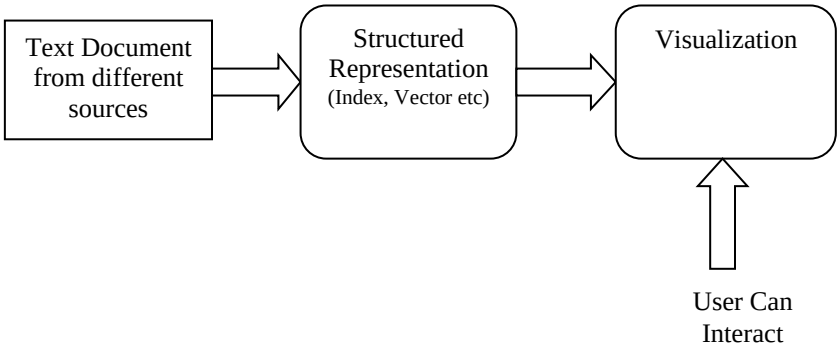

Figure 2 Visualization

E. Summarization

Text summarization is to reduce the length and detail of a document while retaining most important points and general meaning. Text summarization is helpful for to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information hence

summary can replace the set of documents [10]. In the time taken by the user to read the first paragraph text summarization software processes and summarizes the large text document. It is difficult to teach software to analyze semantics and to interpret meaning of text document even though computers are able to identify people, places, and time. Humans first reads entire text section to summarize then try to develop a full understanding, and then finally write a summary, highlighting its main points. Summarization process include following steps:

1. Pre-processing obtain a structured representation of the original text.
2. To transform summary structure from text structure algorithm is applied in next processing step.
3. In the invention step the final summary is obtained from the summary structure.

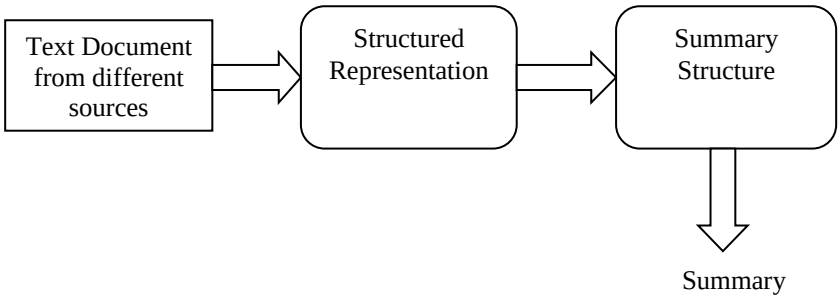

Figure 3 Summarization

V. TEXT MINING USING GENETIC ALGORITHM

Genetic Algorithms (GAs) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. GA is frequently used algorithm and so will work well in any search space [11]. It can be appropriately said that Genetic algorithm is the invention of nature itself. It was named as Survival of fittest theory by Charles Darwin, which is used to evolve the GA. Genetic Algorithm was first named and presented by J.H. Holland in the 1875.Genetic algorithm basically shows an intelligent way to exploit out of a random search from a defined search space in order to solve problem [12]. The basic genetic operators of GA are: selection, crossover and mutation which are applied on an initially random population so as to calculate a new generation. The steps involved in Genetic Algorithms are:

A. Selection

The Selection is the process in which individual genomes are selected from population and are evaluated according to the fitness function defined. More fit will be the chromosome more is it chances to survive or to be selected.

B. Crossover

This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point.

C. Mutation

Make changes in the new solutions in order to find better solutions. In this randomly the bit within a chromosome is selected and is changed accordingly the technique for mutation being used. As in flip mutation, the chosen bit is flipped.

D. Crossover Rates

The range for selection of crossover rate is from 0% to 100%. If the crossover rate is 0% it means that chromosomes in the next generation will be the exact copies of chromosomes in the current

@IJRTER-2016, All Rights Reserved

generation and if it is 100% then every chromosome in the population of next generation will be the result of crossover between any two chromosomes of the current generation [13].

E. Mutation Rates

The mutation rate means how many genes in a population in one generation would get mutated. Here also the range could be from 0% to 100%. If the mutation rate is 0% then it means none of the genes would get selected. But, if it is 100% then it means all the genes in a population of a generation would get mutated. Mutation is an operator that creates a certain level of diversity in a population and hence GA is prevented from getting trapped into local optimum [13].

An initial population is created and it contains randomly generated transactions. String of bits is used to represent each transaction. Then the various genetic operators are used to obtain the final results. To perform text mining using GA two levels of processing are considered. The input is in the form of scientific and technical natural language documents; the output is a small set of the hypotheses that the GA discovered [14]. Automatic text summarization takes an input text and extracts the most important content in the text.

Thus, Genetic algorithm is used in E-mail classification [14] in order to obtain the details from text summarization, text resumes, and there is a vast area of research in bioinformatics to mine text from large dataset in which manually text mining process is almost impossible.

VI. APPLICATIONS

Text mining application uses unstructured textual information and examines it in attempt to discover structure and implicit meanings hidden within the text [15]. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with. Through text mining, we can uncover hidden patterns, relationships, and trends in text. Reference [16] addressed that text mining enables organizations to explore interesting patterns, models, directions, trends, rules, contained in text in much the same way that data mining explores tabular or “structured” data.

A. Bioinformatics

Research work for IE has grown dramatically in a bioinformatics domain, where biomedical journal articles have become an important application area in the recent years. The motivation for this work comes primarily from biologists, who find themselves faced with an enormous increase in the number of publications in their field since the advent of modern genomics is too many; keeping up with the relevant literature is nearly impossible for many scientists [17]. In the bioinformatics domain, biomedical research literature has been a target for text mining. The first textbook on biomedical text mining with a strong genomics focus appeared in 2005, where it has reported that industry has suggested that 90% of drug targets are derived from the literature. The goal of text mining in this area is to allow biomedical researchers to extract knowledge from the biomedical literature in facilitating new discovery in a more efficient manner. Most of the text mining research in this domain has been done in the context of MEDLINE. MEDLINE records consist of a title, an abstract, a set of manually assigned metadata terms.

B. Business Intelligence

Market Analysis, instead, uses text mining mainly to analyze competitors and/or monitor customers' opinions to identify new potential customers, as well as to determine the companies’ image through the analysis of press reviews and other relevant sources. For many companies tele-marketing and e-mail activity represents one of the main sources for acquiring new customers. The TM instrument makes it possible to present also more complex market scenarios. The major concerns in any business are to minimize the amount of guessing work involved in decision making. The risk of making wrong prediction should be reduced. Most of the data mining techniques are

@IJRTER-2016, All Rights Reserved

created to deal with prediction [18]. The problem with data mining is that it can help only up to a certain point, since most of data are available in texts (reports, memos, emails, planning document, etc). Data mining and text mining techniques can complement each other. For example, data mining techniques may be used to reveal the occurrence of a particular event while text mining techniques may be used to look for an explanation of an event.

C. National Security

The use of text mining tool in national defence security domain is very significant. Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, chats in chat rooms. Email is used in many legitimate activities such as messages and documents exchange. Unfortunately, it can also be misused. Thus automatic text mining tools offer a considerable promise in this area. Although not much work has been conducted in this area so far, text mining technology is becoming an emergence technology for national security defence. The work of M.corney particularly focuses on investigating and determining the gender of the author based on the gender preferential language used by the author. They claimed that men and women use language and converse differently even though they speak the same language. The work of [19] for example, has applied text mining techniques to existing medical literature to identify viruses which can be potentially be used as biological weapon, and where such capability is not yet recognized. Another example of text mining system is COPLINK system [20].

VII. COMPARISON OF TEXT MINING TECHNIQUES

S.No	Information Retrieval	Information Extraction	Visualizing	Categorization	Summarization
1.	It finds text documents, unstructured in nature that satisfies user's information need.	It extracts the already defined features from structured documents or display information.	It is used to organize a large amount of documents into a number of sub-groups or clusters for further study and analysis.	It is supervised learning process and uses predefined set documents according to their contents.	It summarizes the document thus reduces length by retaining its main points and overall meaning.
2.	Document retrieval	Feature retrieval	Related Documents are retrieved	Collection of documents belonging to particular category are retrieved	Compressed Version of document is retrieved
3.	Tools used are Intelligent Miner, Text Analyst	Tools used are Text Finder, Clear Forest Text	Tools used are Carrot, Rapid Miner	Tools used are Intelligent Miner	Tools used are Tropic Tracking Tool, Sentence Ext Tool
4.	Output of an IR system is a subset of documents that are relevant to user's query.	Difficult as it requires more detailed knowledge of a document. It has to establish relationships between features.	It is an unsupervised process. It is a group of similar type of data and their relationship between them.	It is a supervised process, establishes meaningful relationships.	Summary produced is a text that is obtained from one or more texts which contains a significant portion of the information.

Figure 4 Comparison of Text Mining Techniques

@IJRTER-2016, All Rights Reserved

VIII. CONCLUSION

Text Mining is basically process of obtaining valuable information from unstructured texts. It is observed that most of the information nearly more than 80%, is stored as texts, and text mining is therefore believed to have a high importance. In this paper various text mining techniques and their comparisons have been discussed which can further be enhanced. Also it is discussed how GA can be beneficial in text mining. This combination of Text Mining and Genetic Algorithm is an emerging area of research.The future scope of this can be how Genetic Algorithm can be used on various attributes so that the results can be optimized.

REFERENCES

1. Baumer, E. P. S., Sinclair, J. & Tomlinson, B. 2010. America is like metamucil: Fostering critical and creative thinking about metaphor in political blogs. In Proceedings of 28th International Conference on Human Factor in Computing Systems (CHI 2010) ACM, Atlanta, GA, USA, 34-45
2. Liu, F. & Lu, X. 2011. Survey on text clustering algorithm. In Proceedings of 2nd International IEEE Conference on Software Engineering and Services Science (ICSESS), China, 901-904.S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
3. Xu, X., Zhang, F. &Niu, Z. 2008. An ontology-based query system for digital libraries. In Proceedings of IEEE, Pacific-Asia Workshop on Computational Intelligence and Industrial Application, Wuhan, 222-226.
4. G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management:An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.
5. W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
6. H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.
7. S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006.
8. Falguni N. Patel, Neha R. Soni, "Text mining: A brief survey" International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.
9. John Atkinson-Abutridy, Chris Mellish, University of Edinburg, IEEE 2004, "*Combining Information Extraction with Genetic Algorithms and Text Mining*".
10. R. Rao, "From unstructured data to actionable intelligence," in *Proceedings of the IEEE Computer Society*, 2003.
11. S.N.Sivanandam, S.N.Deepa, "Introduction to genetic algorithms" Springer-Verlag Berlin Heidelberg 2008.
12. G. Desjardins, R. Godin, University of Quebec, Vol 35, 2005 WIT Press, ISSN-1743-3517, "*A Genetic Algorithm for Text Mining*".
13. Indarjit Mukherjee, (ICCTD, 2010), IInd Edition, "Content Analysis based on Text Mining using Genetic Algorithm".
14. Khalessizadeh, S. M., Zaefarian, R., Nasseri, S. H. &Ardil, E. 2006. Genetic mining: Using genetic algorithm for topic based on concept distribution. Journal of Word Academy of Science, Engineering and Technology, 13(2), 144-147.
15. Navathe, Shamkant B., and Elmasri Ramez, (2000), "*Data Warehousing And Data Mining*", in "*Fundamentals of Database Systems*", Pearson Education pvt Inc, Singapore, 841-872.
16. Liritano S. and Ruffolo M., (2001), " *Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining*", IEEE, 454-458 , Italy.
17. JitendraNathShrivastava, MaringantiHimaBindu, "E-mail classification using genetic algorithm with heuristic fitness function" International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013.
18. P. Srinivasan, "Meshmap: A text mining tool for medline,"in *Proceeding the American Medical Informatic Annual Symposium*, 2001, pp. 642–646.
19. M. Corney, O. devel, A. Andersonand, and G. Mohay,"Genderpreferential text mining of e-mail discourse," in *Proceedings of the 18th Annual Computer Security Applications Conference*. Washington: IEEE Computer Society, December09-13 2002, pp. 51–63.
20. P. Hu, C. Lin, and H. Chen, "User acceptance of intelligence and security informatics technology: A study of coplink," *Journal of The American Society for Information Science and Technology (JASIST)*, vol. 56, no. 3, pp. 235–244, 2005.

@IJRTER-2016, All Rights Reserved