INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

A Survey on Text Mining using Genetic Algorithm

Rekha Dahiya
M. Tech Computer Science Engineering, Galgotias University, Greater Noida, India
Anshima Singh
M. Tech Computer Science Engineering, Galgotias University, Greater Noida, India

Abstract:
Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge discovery, text mining has very high commercial values. Text Mining can be achieved directly by applying Genetic Algorithm to text classification, summarization and information retrieval system in text mining process. The various researches show the performance is improved due to the nature of Genetic Algorithm. Genetic Algorithm has been used to tackle extensive variety of optimization problems. In this paper text mining concept is discussed and how genetic algorithm is applied on it and its usage in text mining.

Key words: *Genetic Algorithm (GA), Text Mining, Classification, Knowledge Discovery*

1. Introduction

The text mining studies are given additional significance as of late due to the accessibility of the increasing number of the electronic records from a mixed bag of sources. The assets of unstructured and semi organized data incorporate the world wide web, government electronic repositories, news articles, biotic databases, talk rooms, computerized libraries, online discussions, electronic mail and website archives. Therefore, proper classification and knowledge discovery from these resources is an important area for research.

The fundamental objective of text mining is to enable users to extract data from text based assets and manages the operations like retrieval, classification (supervised, unsupervised and semi supervised) and summarization.

Today the web is the main source for the text documents, the amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format, in the form of reports, email, views and news etc. Which are shows that approximately 90% of the world's data is held in unstructured formats, so Information intensive business processes demand that we transcend from simple document retrieval to knowledge discovery. The need of automatically retrieval of useful knowledge from the large amount of textual data in order to assist the human analysis is fully apparent [4].

With the fast progression of technology, vast volume of information effectively gathered from regular management of the recent applications, for example retail business, social and health administrations organization and schools. Naturally, this extensive measure of raw stored information holds important hidden knowledge, which could be utilized to enhance the decision-making process of an organization. It is tedious and troublesome to analyze such vast voluminous information and creating relationship between numerous features manually [2].

Quick development of accessible information in digital format expand require for techniques to investigate them. So scrutinize on some topics such as text classification, information retrieval and automatic text summarization turned into a critical field. Scientists in Knowledge Discovery in Databases (KDD) have furnished new tools for analyzing and accessing information in databases. Some of them is dependent upon term recurrence and are utilized within text processing. GA is conveyed to text processing as an optimization problem. GA is utilized within text clustering, Text classification and Text Automatic summarization.

Classification using Genetic Algorithm find the high-level prediction rules by performing a global search in adapt to preferred attribute interaction than the greedy rule induction algorithms where there is no attribute dependency. Genetic algorithm requires less information about the problem. The genetic algorithm uses rule based classifiers in general. One can design an individual to represent prediction (IF-THEN) rules. Classification rules can be considered a particular kind of prediction rules where the rule antecedent ("IF part") contains a combination - typically, a conjunction – of conditions on predicting attribute values, and the rule consequent (“THEN part”) contains a predicted value for the goal attribute [2].

The classification problems have been well studied as a major category of data analysis data analysis in genetic algorithm which generally uses rule based approach. It describes the kind of rule assessment schemes which have been proposed for rule discovery systems. The survey of classification using genetic algorithm rule based approach which includes Michigen versus Pittsburg approach [3]. The genetic algorithms have been associated with greedy techniques. The recent ten years survey of evolutionary algorithms described in [3].

2. Text Mining Process and Technologies

Text mining is a technique which extracts information from unstructured data and find pattern which is novel and obscure prior [5]. Information can be extracted from the summarized words of the documents, so the words can be analyzed and also the similarities between words and documents can be determined.

Text mining is also known as text data mining, which refers the process of deriving high-quality information from text. Text mining includes the process of structuring the input text like parsing and other successive insertion into a database. Text mining derives patterns within the structured data, evaluates them and finally produces the output. Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects [5].

Text mining is nothing but "nontraditional information retrieval strategies." The goal of these strategies is to reduce the effort required of users to obtain useful information from large computerized text data sources. Traditional information retrieval strategies simultaneously retrieve both less and much information from the text. The nontraditional strategies represent a useful system that must go beyond simple retrieval [5].

Text mining process is shown in Fig. 1. From the bulk amount of text documents first text preprocessing is done which obtain all words that are used in a given text, a text document is split into a stream 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, which can be used for further effective analysis task [6]. Then next step is feature selection or attribute selection. This phase mainly performs removing features that are considered irrelevant for mining purpose. This procedure give advantage of smaller dataset size, less computations and minimum search space required [6]. There are different text mining methods such as Clustering, Classification, Information retrieval, Topic discovery, Summarization, Topic extraction in Data mining had been proposed. Then the results of this method are evaluated and interpreted in terms of calculating precision and recall, accuracy etc.

2.1. Mining Plain Text

There are different methods of text mining. Text can be mine from plain text and structured text. Mining plain text describes the major ways in which text is mined when the input is plain natural language, rather than partially-structured Web documents [5]. Text summarization, Document retrieval, information retrieval, Assessing document similarity, Text categorization are different methods of mining plain text.

2.2. Mining Structured Text

Much of the text that we have on the Internet contains explicit structural markup and differs from traditional plaintext. Some markup is internal and indicates document structure or format; some is external and gives explicit hypertext links between documents. These information sources give additional benefits for mining Web documents. Wrapper induction, Document clustering with links, determining authority of web documents are the different methods of mining structured text [5].

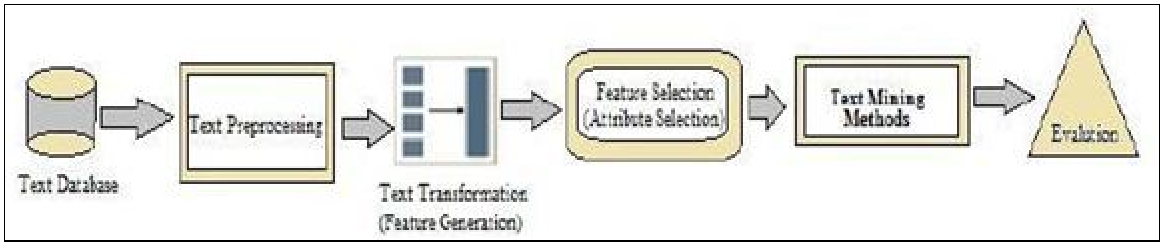

Fig. 1. Text Mining Process [6]

3. Genetic Algorithm

Genetic Algorithms (GAs) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. GAs are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so will work well in any search space [8].

Genetic algorithm was in fact invented by nature. Charles Darwin named it 'evolution' and “survival of fittest” theory used to evolve genetic algorithm. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem. GA works in an iterative manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc which has known as chromosome. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. The basic terms in genetic algorithm are described below.

- **Individual -** Any possible solution
- **Population -** Group of all individuals
- **Search Space -** All possible solutions to the problem
- **Chromosome -** Blueprint for an individual
- **Allele -** Possible settings for a trait
- **Locus -** The position of a *gene* on the chromosome
- **Genome -** Collection of all chromosomes for an individual

Standard GA apply genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings. The functions of genetic operators are as follows:

3.1. Selection

Selection deals with the probabilistic survival of the fittest, in that, more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.

3.2. Crossover

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

3.3. Mutation

Alters the new solutions so as to add stochasticity in the search for better solutions. This is the chance that a bit within a chromosome will be flipped (0 becomes 1, 1 becomes 0).

3.4. 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 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 [9].

3.5. Mutation Rates

Similarly, 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. As indicated earlier, 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 [9].

Basic steps of simple genetic algorithm are as shown in Fig. 2. GA runs to generate solutions for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence the quality of the solutions in successive generations improves. The process is terminated when an acceptable or optimum solution is found.

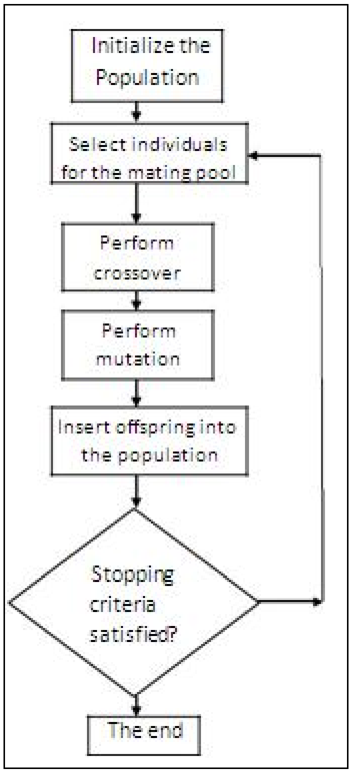
Initialize the Population
Select individuals for the mating pool
Perform crossover
Perform mutation
Insert offspring into the population
Stopping criteria satisfied?
The end

Fig. 2. Flowchart of Genetic Algorithm

4. Text Mining Using Genetic Algorithm

There are many different sources from which information can be extracted. Genetic algorithm is used in Xml mining, opinion mining, web mining, knowledge discovery, feature extraction, classification and different text mining techniques to optimize the solution using the mechanisms of genetic evolution and survival of the fittest in natural selection.

The major advantage of using GA in text mining is that they perform global search and its time complexity is less compared to other algorithms as the genetic algorithm is based on the greedy approach [9].

In [10] Genetic algorithm (GA) is applied on large XML data sets to discover the frequent item sets. First the samples of records are loaded from the transaction database that fits into memory. An initial population is created consisting of randomly generated transactions. Each transaction can be represented by a string of bits. Then by applying genetic operators, correct and appropriate results are gathered.

Figure 3 shows working model divided into two general levels of processing. The input is a corpus of technical and scientific natural language documents; the output is a small set of the hypotheses that the GA discovered [11].

Automatic text summarization takes an input text and extracts the most important content in the text. In [12] two different approaches have been used in the text summarization domain. The first one is using genetic algorithms to learn the patterns in the documents that lead to the summaries. The other one is using lexical chains as a representation of the lexical cohesion that exists throughout the text. The experiments performed on the CAST corpus showed that combining different classes of features and the results showed that features like sentence location, sentence centrality and named entities give better performance than the other features. However, the combination of the features yields better success rates than any individual feature.

In [2] the rule based genetic algorithm classifier is improved by improving the fitness function parameter modification. Also, it compares the results with the probabilistic approach such as Naïve Bayes which is always gives better results and very efficient in case there is no attribute dependency in the problem, which is not true in most of the real world problem. Nursery database which contains 12960 instances of 8 attributes is used for experiment and the result shows the accuracy in percentage which is 49.32 using probability base classifier and 79.65 using GA classifier.

Thus, Genetic algorithm is used in E-mail classification [13], to extract details from text resumes [14], text summarization [12] 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.

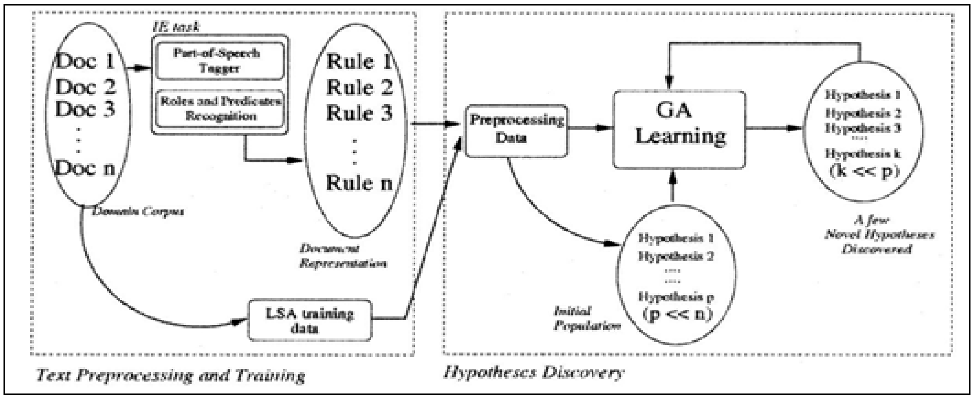

Fig. 3. GA Based Knowledge Discovery from Texts [11]

5. Conclusion and Future Work

In this paper, genetic algorithm is presented in detailed and it has been discussed how GA can be beneficial in text mining. Electronic text based records are broadly accessible because of the development of the Web. Numerous advances are created for the extraction of information from immense collection of text based data using different text mining techniques. This study attempts to provide a thorough understanding of different text mining techniques using genetic algorithm and its application.

The combination of text-mining and Genetic Algorithm technique is a relevant area of research. The survey investigates the recent advancement in the field of text mining using genetic algorithm. This study will definitely provide new ways for researchers to proceed and develop new text mining techniques that will be useful for the analysis of text in large-scale systems.

6. References

1. Vidhya K. A, G. Aghila, “Text mining process, techniques and tools : an overview” International Journal Of Information Technology And Knowledge Management, Volume 2, No. 2, Pp. 613-622, July-December 2010.
2. Keshavamurthy B. N, Asad Mohammed Khan, Durga Toshniwal, “Improved genetic algorithm based classification” Department of Electronics & Computer Engineering, Indian Institute of Technology, Roorkee, Uttarakhand, India.
3. W. B. Langdon, S. M. Gustafson, “Genetic programming and evolvable machines” ten years of reviews, journal of Genet
4. Program Evolvable Machines, pp.321–338, volume 11, 2010.
5. Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, “A review of machine learning algorithms for text-documents classification” Journal Of Advances In Information Technology, Vol. 1, No. 1, February 2010 .
6. Rashmi Agrawal, Mridula Batra, “A detailed study on text mining techniques” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013. A Detailed Study on Text Mining using Genetic Algorithm| ISSN: 2321-9939
7. 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.
8. Shiu Yin Yuen, Member, IEEE, and Chi Kin Chow, “A Genetic Algorithm that adaptively mutates and never revisits” IEEE Transactions on Evolutionary Computation, Vol. 13, No. 2, April 2009.
9. S.N.Sivanandam, S.N.Deepa, “Introduction to genetic algorithms” Springer-Verlag Berlin Heidelberg 2008.
10. M. Srinivas, Lalit M. Patnaik, “Genetic Algorithms: A survey” Motorola Indian Electronics Ltd., Indian Institute of Science, IEEE.
11. Soumadip Ghosh, Amitava Nag, Debasish Biswas, Arindrajit Pal, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, “
12. Xml mining using genetic algorithm” Journal of Global Research in Computer Science, Volume 2, No. 5, April 2011.
13. John Atkinson-Abutridy, Chris Mellish, and Stuart Aitken, “Combining information extraction with genetic algorithms for text mining” Published by the IEEE Computer Society, © 2004 IEEE.
14. Mine Berker, Tunga Güngör, “Using genetic algorithms with lexical chains for automatic text summarization” Boğaziçi University, Computer Engineering Dept., Bebek 34342, Istanbul, Turkey.
15. Jitendra Nath Shrivastava, Maringanti Hima Bindu, “E-mail classification using genetic algorithm with heuristic fitness function” International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013.
16. Deepankar Bharadwaj, Suneet Shukla, “Text mining technique using genetic algorithm” International Conference on Advances in Computer Application (ICACA - 2013) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 –8887)