TEXT MINING TECHNIQUE ON BIG DATA USING GENETIC ALGORITHM (A REVIEW)

[1]Deepankar Bharadwaj, [2]Dr. Arvind Shukla
[1]Research Scholar, IFTM University. [2]HOD, [1.2]Department of Computer Applications, IFTM University, Moradabad (UP)

ABSTRACT:

Mining means to extract something from the source data and give the results that were previously unknown. Big data is a term for bulk and massive data sets having complex, a large structure with the difficulties of storing, analyzing further results. Genetic Algorithm is an algorithm which is used to optimize the results. This paper gives an overview of all three Mining, Big Data and Genetic Algorithm concept, samples, scope, methods, advantages and challenges etc.

Keywords: Mining, Genetic Algorithm, Big Data.

[1] INTRODUCTION

Now a day's there is a rapid increase in data from various document resources like plain documents, web pages etc. [17] Therefore it is necessary to enhance the text processing so that the information or the relevant knowledge which was previously unknown can be mined from the text. So for mining these kind of information from the plain or unstructured text we use Text Mining. It is a way to extract some meaningful information from the bulk amount of text data. There are major challenges that have promoted research into effective & efficient discovery of some meaningful information & use of resources in the form of text on the internet. The general idea of text mining is to get the desired information out of bulk amount of text data without reading it manually. With this increase in data regularly size, complexity, size, security issues are also increasing. There are various sources of storing data like logs, tweets, images, videos, blogs etc and this is increasing its size day by day. This tremendous amount of data is termed as Big Data. A Big Data is the data that are increasing day by day. As per the researchers of Big Data, they have described the Big Data with 4 V's as Volume, Velocity, Variety and Veracity. In this paper, we will discuss about some more concepts of Big Data. Mining is the process to extract some meaningful information from the source data. The source data may be something from which the mining techniques can extract the information. Mining is a set of automated techniques that are used to extract previously unknown or buried information from large sets of databases. A Successful Data Mining makes possible to unearth patterns and relationships, and then to use this "new" information for making proactive knowledge-driven business decisions. Text Mining [17] works on the basis of regular expressions and patterns matching that were defined at the time of the mining process and gives the output on the same basis. Generally Mining techniques are used in case of WWW which serves widely spread, bulk, huge global information service center for news, advertisements and many other information services. For mining the information from plain text there are various algorithms currently being used now a days. These algorithms are being discussed later in this paper. Genetic Algorithms are the algorithms used to solve the optimization problems. These algorithms work on search based inputs which leads to generate useful solutions for such kind of problems. They generate solutions to optimization problems.

There are various techniques used in GA to provide optimum solution. These techniques are later being discussed in this paper. The genetic algorithm identifies combinations of terms that optimize an objective function, which is the cornerstone of the process.

[2] TEXT MINING AND DATA MINING

Mining is the technique which extract something useful from the source data (includes web data, databases, plain text, etc.) and give the results that were previously unknown. Generally mining term is used from the Mines of Coal, Crude Oil, etc. [17] Text Mining is nearly as old as the information retrieval technique. Now a days text mining area has become a very important area of research as text data is increasing day by day and it consumes more time to read all the information given in the plain text. Currently it is very difficult to find the relevant information from huge text as per the users expectation after neglecting all kind of irrelevant information without the use the latest technologies & concepts. It can be taken as one of the classes of Information Retrieval strategies which attempt to avoid the unfairness of human queries, treat entire text collections holistically, and objectify the Information Retrieval process with principled algorithms. These strategies, share many research techniques such as statistical clustering, semantic parsing, etc. They’re also the products like coal, fuel, etc. were extracted or mined from the bulk amount of sauce and after the purification process, and they are used for general use. The same concept is for mining in the computer field. In finding of meaningful & relevant information from the large amount of plain text, the best way is using text mining technique. It gives the desired information within a less friction of time after neglecting irrelevant information from the cluster and finds the relevant information. To use its best performance one has to use the latest concepts and technologies of Text Mining. Text Mining may also be defined as the application of methods & algorithms from the fields of statistics & machine learning to reach the goal of finding useful patterns from plain text. For the purpose of Text Mining it is very important to pre-process the texts accordingly. Many authors use NLPs (natural language processing), IE (information extraction) methods or some simple preprocessing steps in order to extract data from texts. To the extracted data from plain text data mining algorithms can be applied for future use. [16] From this concept, new questions about the use data mining methods arise. The main problem is that we now have to deal with problems of unstructured data sets which can be in any format. If we try to define text mining, we can refer to related research areas also. For each of them, there are different definitions of text mining, which is motivated by the specific perspective of the particular area. In this article, we consider text mining mainly as text data mining. Thus, our main focus is on methods that extract useful patterns from plain texts in order to extract useful information which can be used further for future operations. Text Mining and Data Mining, both are the mining techniques, but there is a small but important difference between them. As described earlier that data mining is the process of extraction the useful and unknown text from the databases and text mining is the process to extract the knowledgeable data from natural language or plain texts. Data mining [21] can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data. With text mining, however, the information to be extracted is clearly and explicitly stated in the plain text. The only sense in which it is “previously unknown” is that human resource restrictions make it infeasible for people to read the text themselves.

Text Mining (TM) [17] is the process of mining the Text from an unstructured set of text or from a natural language text. It is the process which deals with the meaning of some knowledgeable information from a text depends on some set of rules. It may also be defined as the process of analyzing and extracting the meaningful or knowledgeable information from the plain text. That information will be useful for any particular process. Difference with the kind of data stored in the databases, text data is unstructured, ambiguous, amorphous, difficult to deal etc. Generally the term Text Mining denotes to any system(s) which analyze the bulk amount of data and find useful pattern in attempt to find useful information that were not known before to anybody. Text mining finds the patterns that are unknown or unseen from the textual data. But these methods are useless until the required result must be useful for the end user to take decisions. We cannot apply data mining techniques to the plain text data for mining because in case of data mining, we assume that the source data file is in database format instead of plain text format. Therefore, new representations of the text data are used in such cases. It is generally used to denote to analyze the bulk amount of natural text and detects usage patterns in an attempt to extract probably useful information. Also in case of text mining the process of Pattern matching is used while mining any needful information from a text without knowing its type of its format. The text file that is generally being used in the mining process is the Natural Language text file.

The relation between text mining and Information Extraction is somewhat same as because in case of both the strategies we are finding the information from a bulk amount of text. The difference between both the strategies is that in the case of Text mining we do not know the format of the data that is present in the text file but in the case of Information Extraction we generally use the databases to search the data or required information. Text mining and data mining differ in the case that in text mining, the source text may be of any format, generally in the natural language text format, but in case of data mining the source data format is in the format of databases in which the data are arranged into the tables. Only we have to apply algorithms to extract the data from it. Text mining strives to bring previously unknown, meaningful information out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary. Another requirement that is common in both data mining and text mining is that the information extracted should be unknown and useful. In the case of data mining, this notion can be expressed in a relatively domain-independent way. Performance can be measured by counting successes and failures for each case. For this statistical techniques can be applied to compare different data mining methods on the same problem, and so on. However, in case of text mining, it is far harder to characterize the independent of the particular domain at hand. This makes it difficult to find fair and objective measures of success. Both are the mining techniques, but there is a small but important difference between them. As described earlier that data mining is the process of extraction the useful and unknown text from the databases and text mining is the process to extract the knowledgeable data from natural language or plain texts. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data. With text mining, however, the information to be extracted is clearly and explicitly stated in the plain text. The only sense in which it is “previously unknown” is that human resource restrictions make it infeasible for people to read the text themselves.

Text mining and data mining differ in the case that in text mining, the source text may be of any format, generally in the natural language text format, but in case of data mining the source data format is in the format of databases in which the data are arranged into the tables. Only we have to apply algorithms to extract the data from it. Text mining strives to bring previously unknown, meaningful information out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary. Another requirement that is common in both data mining and text mining is that the information extracted should be unknown and useful. In the case of data mining, this notion can be expressed in a relatively domain-independent way. Performance can be measured by counting successes and failures for each case. For this statistical techniques can be applied to compare different data mining methods on the same problem, and so on. However, in case of text mining, it is far harder to characterize the independent of the particular domain at hand. This makes it difficult to find fair and objective measures of success. Various classification of text mining are Keyword extractors, Entity extractors, Entity relation extractors, Document relation extractors. These extractors have their separate classifications and separate functionalities. Keyword Extractors forms the core of any search engine. Another one is entity extractor which tries to classify the terms into basic category. These category may be like organization, person, city, region, money, type, etc. Third extractor is entity relation extractor which not only finds entities mentioned in the document, but also relate with each other. The last one is document relation extractor whose objective is to identify common themes between different documents and to go beyond the limits of a single document. In many Text Mining [22] applications, we can use more structured data representations than just like keywords to perform analysis to uncover unseen patterns from the text. Most Text Mining approaches are implementing the ideas by combining more elaborated information extraction patterns and general lexical resources such as Word Net or specific concept resources such as thesauri. Another approach, relying on Information Extraction patterns, uses linguistic resources such as Word Net to assist the discovery and evaluation of patterns to extract basic information from general documents or plain text.

There are various small and large applications exists that are based on text mining techniques. Few of them are discussed below

1. Analyzing insurance claims, interviews, etc. In this application in most of the business domains mainly the information is collected in textual format. This information is then used to process the data and extract meaningful information.

2. Analyzing open-ended survey responses. In this the idea is to permit respondents to express their opinions or "views" without constraining them to a particular response format or a particular dimension.

3. Investigating competitors by crawling their web sites. Another very useful type of tech mining application is to automatically process the web contents in a particular domain. These can be used afterwards to optimize the result.

4. Automatic processing of emails, messages, etc. Another useful and common application for text mining is to support in the automatic classification of texts such as in the case of emails. For example, in case of mails, it is possible to filter automatically junk mails based on certain terms and conditions or words that are not likely to appear in messages, but instead identify undesirable electronic mail.

[3] TEXT MINING APPROACHES

Following are some text mining approaches used generally.

1. Mining the text as document search.
2. Using well-tested methods of text mining.
3. Black-box approaches to text mining.

[4] TEXT MINING METHODOLOGY

Following is the text mining methodology used generally.

1. Learning the application domain to extract relevant knowledge.
2. Creating a target data set: Data Selection.
3. Data cleaning and Pre-Processing.
4. Data reduction and Transformation.
5. Choosing Text mining approach.
6. Choosing the *mining algorithm*(s).
7. Data mining: search for patterns of interest.
8. Pattern evaluation and knowledge presentation.

9. Use of discovering knowledge.

[5] BIG DATA

As we have heard about the term big data, a question arises in our mind that WHAT IS BIG DATA AND WHY WE NEED BIG DATA? The first term comes in our mind related to big data is the data which is big. In other words the data which is beyond to the storage capacity, beyond the processing power, beyond the analysis that data is considered as Big Data. Any data which is not easy to process, not easy of store and not easy to analyse can be considered as big data. Nowadays massive amounts of data is collecting daily that may be generated from different data generating source or factors. Some of these sources are sensors, CC Cameras, Social Networking Websites, Online Shopping, Airlines, Hospitality Data, etc.

Nowadays handling Big Data with old techniques and old algorithms has become a major challenge. [1] There are very general question still exists.

1. What is BIG DATA?
2. When a data may be called as BIG DATA?
3. How large is BIG DATA?
4. What is the size of BIG DATA?
5. What is the correlation between business intelligence & BIG DATA?
6. What is the optimal solution for storing, editing, retrieving, analyzing, maintaining, and recovering BIG DATA?
7. How can cloud computing help in handling BIG DATA issues?
8. What is the role of cloud architecture in handling BIG DATA?
9. How important is big data in business intelligence?

and many more…

Above questions still exists in everyone's mind. Some users say that a Big Data is anythis when the data is huge, Some says that Big Data may be defined as data that is too complex to capture, process, and analyze, some says that Big Data is something which is difficult to process using current computing infrastructure and many more answers. Recently, big data have attracted a lot of attention from industry, academic as well as from government. Big Data refers to massive volume of data that is not readily handled by the usual practices, common data tools, present unprecedented opportunities for advancing science and informing resource management through data-intensive approaches. For explaining the concept of Big Data there is a theory exists of 4 Vs [2].

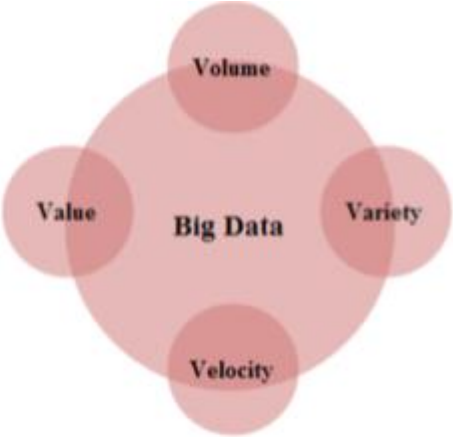

Fig. 1. **Four V's of Big Data**

VOLUME: Data measurement is in terabytes (2^{40}) or even petabytes (2^{50}), and is rapidly heading toward Exabytes (2^{60}).

VARIETY: Data is heterogeneous and can be highly structured, semi-structured, or totally unstructured.
VELOCITY: Data production occurs at very high rates, and, because of this sheer volume, some applications require real-time data processing to determine whether to store a piece of data.
VALUE: Through predictive models that answer what-if queries, analysis of this data can yield counterintuitive insights and actionable intelligence.

[6] TYPES OF BIG DATA

There are two types of big data as we found, one is structured data and another one is unstructured data. [3] Structured data are numbers and words that can be easily categorized and analyzed. These data may be generated by network sensors that are embedded in electronic devices, smart phones, and global positioning system (GPS) devices, other things like transaction data, account balances and sales figures. Unstructured data includes more complex information, which is difficult to process such as customer reviews of commercial websites, photos, multimedia, comments on social networking sites etc. This kind of data can't easily be separated into categories or analyzed numerically or in other words it can't be classified. Clustering of these kind of data will become very difficult and analysis of these unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The massive and explosive growth of the Internet usage now a days means that the variety and amount of big data is continuously in growing stage and many researchers are working on keeping this big data into an efficient manner so that it can be used further.

[7] GENETIC ALGORITHM

Genetic Algorithms (GA) [18] are those algorithms that are used to solve optimization problems. These algorithms are based on the inherit processes of biological organisms. GA simulates those processes in natural populations which are essential to evolution. Evolution inspired these computational model algoriothms. They apply operators like recombination or crossover and encode a potential solution to these structures of data in such a manner that it provides best & optimum result. GAs are often viewed as function optimizers which optimize the results as per the requirements of the users. There are various types of GAs now a days available which on multiple problems provide better results. Genetic Algorithms are applied on the basis of population of chromosomes. Chromosomes are the key of Genetic Algorithm. They are considered as the individual solutions itself. In a wide usage, Genetic Algorithm is any population-based model that uses its operators for optimum solution. Operators of Gas are selection or reproduction, crossover or recombination and mutation. These operators can work on any environment and provide the solution as per the requirements. Many of the genetic algorithm models have been introduced by researchers who are working largely from an experimental perspective. Many of these researchers are application oriented and are typically considered genetic algorithms as optimization tools for their work.

Terminologies or Explanation used in Genetic Algorithm are:

- Chromosome (string, individual) as a Solution (coding).
- Genes (bits) as Part of solution.
- Locus as Position of gene.
- Alleles as Values of gene.
- Phenotype as Decoded solution.
- Genotype Encoded solution.

GAs is those search algorithms that are based on the concepts of natural genetics and natural selections. These algorithms were developed to reproduce some of the processes that are observed in natural evolution, and to optimize the processes that operates on chromosomes. It uses the concept of the objective function or fitness function without any gradient information. The traditional methods use gradient information, whereas the transition scheme method of genetic algorithm is probabilistic. Because of these features genetic algorithms are used as general purpose optimization algorithm to optimize the results. They provide means to search irregular spaces and hence they are applied to a variety of machine learning applications, parameter estimation and function optimization.

In general, many searching and optimizing algorithms exists, but Genetic Algorithms look promising in the case of Text Mining. Compared with classical search-and-optimization algorithms, Genetic Algorithms are much less susceptible to getting stuck in local suboptimal regions of the search space. This algorithm performs global searches by exploring multiple solutions in parallel. Being robust, Genetic Algorithms can cope with noisy and missing data. However, to use Genetic Algorithms effectively, we must tackle several problems easily. After a number of new generations built with the help of the described mechanisms one obtains a solution that cannot be improved any further. This solution is taken as a final one.

Now a days large amount of data have been collected daily for day to day business management, business administration, banking sector, health services organizations, social services, security and many more. Such data is being used for accounting and management of the customer base. Typically, management data sets are very huge, complex, constantly growing. These data sets reflects properties of the managed subjects & relations, and are thus potentially of some use to their owner, they often have relatively low information density. Anyone requires simple, robust, computationally efficient tools to extract information from such big data sets. The development and understanding of such tools are the core business of data mining. Mining of useful information that was previously unknown & helpful knowledge from these large databases has thus evolved into an important research area.

A Genetic Algorithm [10] is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as mutation, selection, and crossover operations. Genetic algorithms give a potential solution to a specific problem. GAs are a very effective way of quickly finding a reasonable and optimized solution to a complex problem. GAs are most effective in a search space for which little is known. They produce solutions that solve the problem in ways you may never have even considered. Then again, they can also produce solutions that only work within the test environment. Genetic Algorithm is the technique to optimize the solution after the process of Text Mining. It is the most important part for the mining process because it helps us to produce the optimized results. Applying genetic algorithms for text mining is not new in the search for better document descriptions. When adapting the genetic theory to the text categorization problem, the documents represented by a vector of terms become the chromosomes of the population. Each term into a vector becomes a gene. The categorization problem turns into finding the best set of terms to represent each document of the collection, with respect to a specific goal, which might be, for example, maximizing the distances between the categories. The goal is modeled as an objective function to optimize, which is termed as the fitness function in the genetic domain. The fitness function plays a very important role of the natural selection. New individuals are generated by exchanging the genes at random between the most fitted sets of terms according to the fitness function.

Genetic Algorithm [19] helps in optimizing the Text Mining results for the problems regarding Documents indexing and retrieval, learning of matching functions and queries, Clustering of Documents and Terms. Genetic Algorithms have been shown to be an effective tool to use in data mining and pattern recognition. An important aspect of Genetic Algorithms in a learning context is their use in pattern recognition. GAs are global search algorithms that work by using the principles of evolution. Traditionally, GAs have used binary strings to encode the features that compose an individual in the population; the binary segments of an individual that represents a specific feature are known as chromosomes. Binary strings are convenient to use because they can be easily manipulated by GA operators like crossover and mutation. Given a problem and a population of individuals, a GA will evaluate each individual as a potential solution according to a predefined evaluation function. The evaluation function assigns a value of goodness to each individual based on how well the individual solves a given problem. This metric is then used by a fitness function to determine which individuals will breed to produce the next generation. In addition, a mutation factor is present which will randomly modify existing solutions. Mutation operator helps the GA to break out of local minima, for a globally optimum solution. While there is no guarantee that GAs will find an optimal solution, their method of selection and breeding candidate solutions means a pool of "good" solutions can be developed given enough generations.

Fitness function F (x) is the backbone for a Genetic Algorithm to work. The main focus of Fitness Function is to give the successive results after applying GA. It is firstly derived from the objective function and the used in successive genetic operations like crossover, mutation. Fitness means quality value which is the measure of the reproductive efficiency of individual string (chromosomes). A fitness functions gives a score to an individual chromosome. Firstly the adjacency index numbers translates into a graph of regular expressions. This graph always has a root node. From this root node the function sets the starting point. The objective function tries to validate the root node of the graph on a line of the function. Secondly it will check if the node has some child edges. If it exists, it tries to fit each child node on the manifest. When a node fits on a part of a line of the manifest, the fitness score is saved. Then the function will look for next child nodes of the current node that fits on the manifest. This process repeats recursively until it reaches to the end of the document or when the root node fits once again. In genetic algorithm, Fitness Function "*F(x)*" is used to allocate reproductive characters to the individuals (chromosomes) in population and acts as some measure of goodness so that it is to be maximized. This means that chromosomes with higher F(x) will have the maximum probability of being selected as candidate for further observations. It means that higher the fitness value, higher the probability of being selected.

[8] STEPS USED IN GENETIC ALGORITHM

- Create a population of random chromosomes.
- Test each chromosome for how well it fits the problem.
- Assign each chromosome a fitness score.
- Select the chromosomes with the highest fitness scores and allow them to survive to a next generation.
- Create a new chromosome by using genes from two parent chromosomes (crossover).
- Mutate some genes in the new chromosome.

[9] GA ALGORITHM

```
1. Generate initial population (chromosomes).
2. Compute fitness of each individual (say F(x)).
3. WHILE NOT finished DO
   •    FOR population size DO
        a. Select two individuals randomly from old generation.
        b. Apply Crossover to give two offspring.
        c. Select an individual randomly to apply Mutation operator.
        d. Insert offspring in new generation.
        e. Compute fitness of each offspring.
   •    IF population size converged THEN
```

```
        Finished: = TRUE
4  END
```

[10] ADVANTAGES AND DISADVANTAGES OF GA

The major advantage to the use of genetic algorithms is that they are easily parallelized. There are, however, many disadvantages to their use:

1) Genetic algorithms are difficult to understand and to explain to end users.
2) The abstraction of the problem and method to represent individuals is quite difficult.
3) Determining the best fitness function is difficult.
4) Determining how to do crossover and mutation is difficult.

[11] DRAWBACKS OF GA

The drawbacks of applying genetic algorithms to data mining include:-

1) The large over-production of individuals and
2) The random character of the searching process.

[12] APPLICATIONS OF GA

Following is the list of various applications of Genetic Algorithm:

1. Used in Text Mining.
2. Used in Data Mining.
3. Artificial Intelligence.
4. Bio Informatics.
5. Clustering.
6. Distributed Systems.
7. Neural Network.
8. Natural Language Processing.
9. Wireless Networks.
10. Ad-Hoc Networks and many more…

[13] RELATED WORK

Some related work in context of Big Data rising on Cloud Computing [11], Mr. Ibrahim Abaker and his team worked on research on big data in the cloud computing. In this section they found some key research challenges like scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy & legal issues. In paper Role of cloud computing architecture in Big Data [1], Mr. Mukesh Singhal and team collected massive amount of data from people, sensors, actions, algos and the web. They found some questions which still exist regarding when data may be called big data. How large is big data? What is the correlation between big data and business intelligence? What is the optimal solution for storing, editing, retrieving, analyzing, maintaining, and recovering big data? How can cloud computing help in handling big data issues? What is the role of cloud architecture in handling big data? How important is big data in business intelligence?

Firstly, they review what is big data, what are the important challenges of storing, maintaining analyzing, recovering and retrieving big data, then the role of Cloud Computing Architecture in resolving these issues of big data. After analyzing the role of cloud architecture in big data, the role of business intelligence with big data, the role of major cloud service layers in big data they conclude that any organization who is working with big data can be transformed to a smart organization with the use of business intelligence. Related work in context of Text Mining is that now days, text mining is mainly used in medical field and in case of Big Data, where the quantity and quality of data depend the information extracted from using text mining. [12] In this paper the author introduces a new data science solution as BigAnt for mining Big Data for frequent patterns satisfying commonly used, user specified anti monotonic constraints. As per the authors, this data science solution can be considered as a non-trivial integration of Big Data analytics and mining, frequent pattern mining, uncertain data mining, and constrained pattern mining. BigAnt first reads high volumes of uncertain Big Data. As each item in the uncertain Big Data is associated with an existential probability value, BigAnt computes the expected support of all domain items and returns all and only those patterns that are interesting to the users. Attendance Management System [13] plays a very important role in management of college students. Currently there are various methods in the today's world which provides automated systems of attendance management. These systems use the database architecture for storing the data of attendance. The entire concept behind these systems is to store the data in databases or other sources and to use the data with different technologies. Attendance Management Systems are developed for daily student attendance in schools, colleges and institutes. It facilitates to access the attendance information of a particular student in a particular class. [14] These systems will also help in evaluating attendance eligibility criteria of a student. By just a click on the mouse, the system will be able to produce the student's attendance report thus reducing the need for manual labor which is prone to human errors and time consuming. These applications are built for automating the processing of attendance. It also enhances the speed of performing attendance task easily. The Student Attendance will be based on the department and section. According to the department wise and section wise the attendance will be marked for the students. It includes present, absent and leave column for each student so that they would mark the attendance like period wise. The student can view the attendance record on weekly, monthly, and whole semester basis.

[14] FUTUTRE DIRECTIONS

This huge data of attendance of any student with all respective entries will be considered as big data. To manage this big data we are going to introduce the new method for storing data which will be beneficial for future use and applying various other techniques. This paper provides the reader a review concept of all the terminologies of the topic. In this paper the reader can find the concept of Text Mining, Big Data and Genetic Algorithm concept, samples, scope, methods, advantages, challenges etc. with their related work and directions.

REFERENCES

[1] Mehdi Bahrami and Mukesh Singhal, "The Role of Cloud Computing Architecture in Big Data", Information Granularity, Big Data, and Computational Intelligence, Vol. 8, pp. 275-295, Chapter 13, Pedrycz and S.-M. Chen (eds.), Springer, 2015.

[2] McAfee, Andrew, and Erik Brynjolfsson. "Big data: the management revolution." Harvard business review 90.10 (2012): 60-66.

[3] Bharti Thakur, Manish Mann, "Data Mining for Big Data: A Review". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014 ISSN: 2277 128X.

[4] Nan Li and Anthony Escalona, A Scalable Big Data Test Framework, 2015.

[5] Mining the Big Data: The Critical Feature Dimension Problem, 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IEEE.

[6] "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, no. 1, January 2014.

[7] A. Kogilavani, "Clustering based optimal summary generation using Genetic Algorithm", 2010.

[8] Combining Information Extraction with Genetic Algorithms and Text Mining

[9] Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. Performance Evaluation Review, 41(4), 70-73.doi: 10.1145/2627534.2627557

[10] Indarjit Mukherjee, "Content Analysis based on Text Mining using Genetic Algorithm" (ICCTD, 2010).

[11] The Rise of Big Data on Cloud Computing (A Review), Ibrahim Abaker Targio Hashem, Faculty of Computer Science, University of Malaya, 2014.

[12] A Data Science Solution for Mining Interesting Patterns from Uncertain Big Data, 2014 IEEE Fourth International Conference on Big Data and Cloud Computing.

[13] Development of a Student Attendance Management System Using RFID and Face Recognition: A Review, Volume 2, Issue 8, August 2014, IJARCSMS.

[14] http://www.studymode.com/essays/Attendance-Management-System-Problem-Statement-47222382.html

[15] Significance & Challenges of Big Data Research, 2015 Science Direct.

[16] Mining Big Data: Current Status, and Forecast to the Future, Volume 14, Issue 2, SIGKDD Explorations.

[17] “Text Mining Technique using Genetic Algorithm”, International Journal of Computer Applications (0975 – 8887) Volume #. 63, February 2013

[18] S.M. Khalessizadeh, R.Zaefarian, World Academy of Science, Engineering and Technology, 2006, “Genetic Mining: Using Genetic Algorithm for Topic based on Concept Distribution”.

[19] Tom V. Mathew, IIT Bombay, “Genetic Algorithm”.

[20] http://viewer.opencalais.com/

[21] Jiawei Han and Micheline Kamber, “Data Mining Concepts & Techniques”, Second Edition, *Morgan Kaufmann* Publishers, Pg 318, 319, 351

[22] http://www3.cs.stonybrook.edu/~cse634/ presentations/ TextMining.pdf