A Review on the Genetic Optimization of Big Data Sentiment Analysis

Er. Puneet Kaur
Computer Science Department,
Chandigarh university
Mohali, India

Er. Amrita Chaudhary
Computer Science Department,
Chandigarh university
Mohali, India

***Abstract* -Big data is a commonly used data set in many areas. Processing a large dataset takes time, not just because of the large amount of information, but also because the nature and structure of the data can be diverse and complex. The studies on text mining are becoming ever more important recently as the set of digital records from a wide range of sources are increasingly accessible. Few tests specifically on the application of text mining phase of the application of genetic algorithm to the text classification, resuming and knowledge collection method. Due to the nature of genetic algorithms, these studies show improved performance. Genetic algorithm is used to tackle a broad variety of optimization problems. The definition of genetic algorithm and its use in text mining is explored and presented in this paper.**

Keywords–Artificial Intelligence, Big Data, Opinion Mining, Sentiment Analysis, Genetic Algorithm, Fuzzy-C-Means, Text Mining

1.INTRODUCTION

In view of the availability of the increased set of Electronic devices from a combined source bag , the texts mining studies are becoming more relevant late on. The worldwide internet, government electronic registry, news articles and biotic database, talkrooms, computerized bookstores, online forums, e-mail and internet archives compose the properties of unstructured and semi-organized info. Accordingly, an important research field is an effective classification and information discovery from these tools. Text mining is aimed at allowing users for retrieving knowledge from text-focused tools or handle transactions, such as collection, classification (supervised, uncontrolled, semi-supervised) or summary. Since the internet is the primary resource of text records, the number of text data obtainable to us is enhancing about 80% of an organisation's knowledge is processed in unorganized text, in the kind of documents , emails, field of vision and news and so on. This indicates that about 90 % of global information is stored in unorganized forms, and that information considered extremely that move beyond basic database extraction to knowledge discovery. It is quite clear that a huge amount of text data must be used to help human analysis manually in order to retrieve valuable knowledge [1].

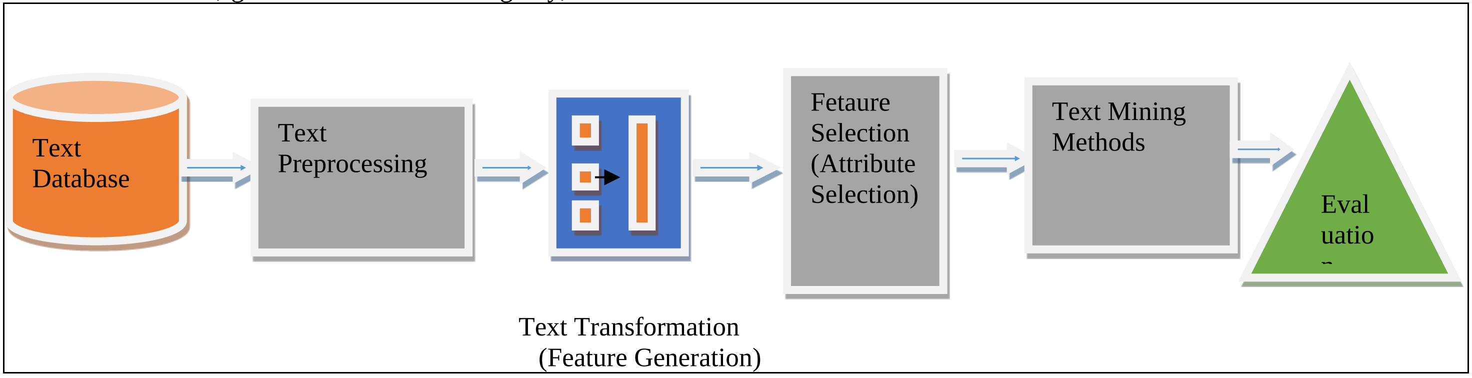
Text Database
Text Preprocessing
Fetaure Selection (Attribute Selection)
Text Mining Methods
Eval uatio
Text Transformation (Feature Generation)

Figure 1.1 Text mining Process[2]

II.GENETIC ALGORITHM

First genetic technique that was developed and identified by J.H. In 1875, the Netherlands fully imitated the natural evolution method [3]. GA works by developing new communities of old cords in an iterative way. The binary encoded, the actual etc, known as the chromosome, are each set. A fitness measurement function associates each string with the fitness of the problem. To automatically retrieve valuable information, it is obvious that a great many text data need to be used to help human understanding[3].Currently, genetic algorithms were developed by nature. The genetic algorithm used to develop was called 'evolution' and "survival of fittest theory." As such, they reflect a smart use of a random search to solve a problem in a particular search area.

The basic terms in genetic algorithm are described below.

- Individual: Any solution necessary
- Population: community of all entities
- Space Search: all objective function to the issue
- Chromosome: an individual's blueprint
- Allele-:Possible role configurations
- Locus: The location of the gene on the chromosome
- Genome: Selection of all genomes for each entity

In Addition to determine a full creation of latest strings, standard GA applies such collection, transforming and mutation to an initially random population. The genetic operators have the following functions:

1)Selection: Selection is concerned with the deterministic survival of the fittest, when more fit chromosomes are selected for survival. While strength is a corresponding measure of chromosomes effectiveness.

2) Crossover: The method is done by picking a random gene across the width of the chromosome and then exchanging both genomes. Rates . The range from 0% to 100% for choosing the crossover limit. If it is 0%, So each chromosome of another creation population will be the result of a collaboration of two different chromosomes of the current generation. The rate of fusion indicates that next phase chromosomes are the carbon replicas of current generation chromosomes and 100%.

3)Mutation : Adjust the existing ideas to make effective solutions more stochastic. That is the possibility of a bit being reversed inside a chromosome (0 is 1, 1 is 0).
The rate of mutation also indicates how often genes can mutate in a population of one creation. Here, the scale could also be among 0 and 100 percent. If the mutation rate is 0 percent, then neither genes will be identified. Even if it is 100%, it implies that all genes in a generation population will mutate. As stated above, the mutation is an operator that provides a certain level of population diversity and thus prohibits GA from becoming best collected localized [4] .

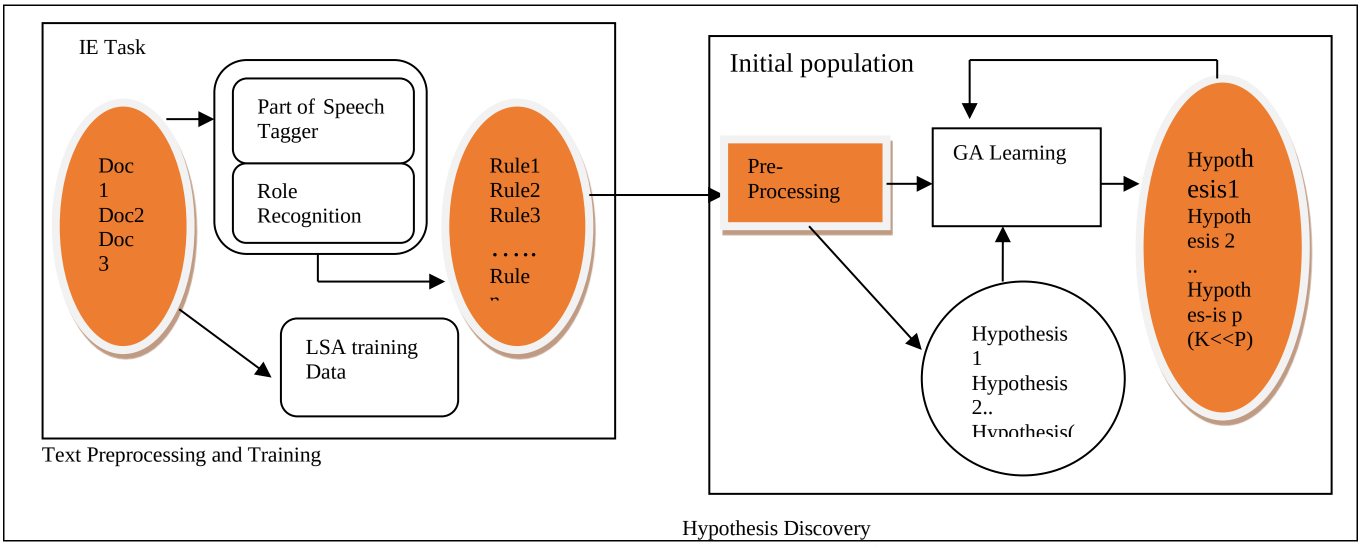
IE Task
Doc 1 Doc2 Doc 3
Part of Speech Tagger
Role Recognition
Rule1 Rule2 Rule3 ….. Rule n
LSA training Data
Text Preprocessing and Training
Initial population
Pre-Processing
GA Learning
Hypothesis1 Hypothesis 2 .. Hypothes-is p (K<<P)
Hypothesis 1 Hypothesis 2.. Hypothesis(
Hypothesis Discovery

Figure 1.2 GA Based Knowledge Discovery from Texts [5]

III.FUZZY C MEAN (FCM) ALGORITHM

Let x= xi, I = 1 to N be the micro-array spot pixels whereby N is the pixel numbers in the spot. The pixels must be divided into two BG and FG groups. Let cj j=1,2 be the FG and BG pixels cluster centres. For each cluster, each pixel must have uij membership degrees. A specific cluster is allocated with the pixel based on the member's membership function value. Therefore, the algorithm aims to boost the membership degree function iteratively until the cluster centers are not modified. The sum of the pixel membership values of all clusters should be achieved

$$\sum_{j=1}^{2} uij = 1 \quad \forall\, i = 1,2, \ldots\ldots N \qquad (1.1)$$

The Euclidean distance from its pixel to the cluster center is specified through [18]

$$D_{ij}= \| x_i - c_j \| \qquad (1.2)$$

The objective of this methodology is to lessen the outright estimation of the differentiation between the 2 successive target capacities K^t and K^{t+1} set out in Equation 1.3 and 1.4.

$$K^t=\sum_{i=1}^{N} \sum_{j=1}^{2} u^m{}_{ij} \;\; dij, m€ [1,Infinity] \qquad (1.3)$$

$$\|K^{t+1} - K^t \|< £ \qquad (1.4)$$

Where m is the capacity variable and the mistake that should be decreased. Iteratively in each progression, the refreshed participation u_{ij} and the cluster centers c_j are given by Equations 1.5 and 1.6.

$$u_{ij} = \frac{1}{\sum_{K}^{2} (dij/dij)2(m-1)} \qquad (1.5)$$

$$C_j =\frac{\sum_{i=1}^{N} uijm\ Xi}{\sum_{i=1}^{n} uij\ m} \qquad (1.6)$$

IV. SENTIMENT ANALYSIS USING BIG DATA

The study of sentiment intends to evaluate the user's views, perceptions and moods in relation to some of the judgments or full document contextual isolations. The basic task of evaluating the polarity at various levels such as the document level, the sentence level and the aspect level or class of the individual. Emotions conveyed by feeling or thought are graded as positive, negative and neutral in various classes[6]. Emotional states such as "angry," "happy," "sad," etc. are articulated.

"Big data is a process of examine large data sets that include different types of data, i.e. big data, to discover hidden patterns, ties, consumer dynamics, customer desires and other useful business data[7].

Advantages of Big Data:

The following advantages of big data are[8]:

- Understanding and Targeting Customers
- Understanding and Optimizing Business Process
- Improving Science and Research
- Improving Healthcare and Public Health
- Optimizing Machine and Device Performance
- Improving Security and Law Enforcement

Disadvantages of Big Data:

Following are disadvantages of big Data[9]:

- A bulk of large data is unorganized.
- Big data analysis contradicts privacy laws.
- Big data analysis is not useful in the short term. It needs to be analyzed for a longer period of time to reach its objectives.
- Modern storage can cost more money to hold big data.
- Speedy updates in big data can mismatch real figures.

V. EXAMPLE OF TEXT CLASSIFICATION WITH DIFFERENT TECHNIQUES

Build on enhanced approaches (GAFCM) of fuzzy c-means text mining techniques (FCMs). The effective approach will be tested by a sample in this section, and GAFCM will be compared to fuzzy c-means text mining.

- **Select the data source**

  The content information were browsed the content corpus of the Open Platform for the Analysis of Chinese Natural Languages (www.nlp.org.cn). Pick six sorts, specifically aviation, governmental issues , innovation, sports , instruction and economy, each category has 200 text files, a total of 1200.

Theme	Text Mining	Text number
Computer	200	20090101-200901200
Education	200	20090101-200901200
Economy	200	20090301-200903200
Art	200	20090401-200904200
Politics	200	20090501-200905200
Sports	200	20090601-200906200

Table 1. List Of Text Data

- **Evaluation criteria**

The clustering models are calculated according to two requirements for the right rate and F-Measure[19] parameters. The accuracy of the measurement equation as shown below :

$$A=\frac{\sum_{i=1}^{k} \max_{0<j<T)}^{(eij}}{N} \qquad (1.7)$$

In which K is the measure of target bunches in the content, N is set of text sets, T is the measure of grouping results, ij e is the arrangement of text wherein I-incorporates the j-group results in the report. Measure utilizing the condition and get the correct rate to come to the condition:0 < A<1. The higher the worth, the further solid the grouping tests. F – measure is an outside evaluation apparatus and is a basic factor for surveying the achievement execution.

$$\text{F-Measure} = \frac{2PR}{P+R} \qquad (1.8)$$

Using F-Measure to test prediction performance, then A(precision)and B(recall) of cluster i and category j are as obeys:

$$A(i,j)=\frac{N\ 1}{N\ 2} \qquad (1.9)$$

$$B(I,j)=\frac{\boldsymbol{N1}}{\boldsymbol{N3}} \qquad (1.10)$$

N1 is the amount of text in which the i cluster refers to group j. N2 is the amount of all text in the i cluster. N3 is the sum of all texts in category j. So the F-Measure of Category J is as follows:

$$F(j)=\frac{2\ AB}{A+B} \qquad (1.11)$$

When the F-measure price is better, the i cluster is a classification mapping. The clustering effects of the F-Measurement may be provided by the weighted average F-Measurement for each class j. The equation is:

$$F = \frac{\sum j\ (F\ (j)\times|j|)}{\sum\ j\ |j|} \qquad (1.12)$$

|j|means the set of all texts in category . F-measuring price is larger, clustering is stronger.

- **Implementation process**

  The working condition of the framework is Windows Vista, and the programming language is C++. The outcome is as per the following:

1) **Select two gatherings text from the information source, three as a gathering.**

   The one and two classes of text information are found in Table II:

Group 1		Group 2	
Category	Total	Category	Total
Computer	205	Art	205
Education	205	Politics	205
Economy	205	Sports	205

Table 2. Texts Of Group 1 And Group 2

2) **Word processing**

   For segmentation Hailiang Chinese Intelligent Participant Program is used. Implementation process is :

a) Load text data to memory and place it in a predetermined string.

b) The participle function of the mass segmentation is named as the string is split and the results are stored in the predefined data structure string array.

c) The feature loads the hailiang participle user dictionary, manages the stop words, stores the character words in the predetermined data structure string list.

3) Features weight calculation and vectors expressions

Firstly, the dimension is decreased by using the document word frequency, and by measuring the document operating frequency of each sub-word variable, also by eliminating the text terms with a lower frequency quality. The specification is as follows:

a) Call the string array that holds the outcomes of the segmentation, calculate the frequency of the document for every word and then eliminate the phrase in which frequency value of the document is too high then returns the result in a string array.

b) measure the amount of each string using the weight formula then return the data in the corresponding numeric array.

c) Use the text convey design and text vector.

4) Cluster Analysis

Execution of the FCM text processing technique:

(a) simply select c texts as the center of the cluster;

(b) Define the initial clustering matrix;

(c) Measure||p(b)-(b+1)||,compare the results with the given value ∈, and stop when the results of the measurement are less than ∈;

d) Measure the set of clusters that each cluster comprises, calculate the number of the right cluster type, and display the data.

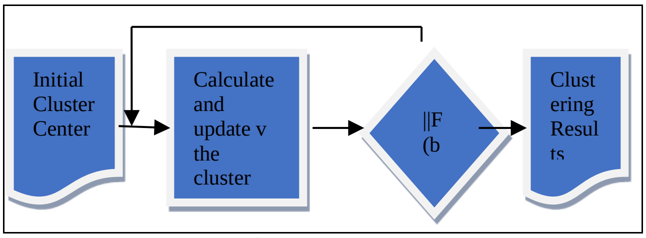

Figure 1.3. Flow chart of FCM text mining method[18]

5) Clustering results

Once the clustering process is complete, the performance result will be in Table III and Table IV:

Category	FCM Clusters	Output
Computer	198	162
Education	209	165
Economy	196	162

Table 3. Results Of Fcm Text Mining Method

Category	GAFCM Clusters	Output
Computer	196	171
Education	213	176
Economy	194	169

Table 4. Results Of Gafcm Text Mining Method

The test results show that the c-fuzzy method of text mining is faster in text mining. Nonetheless, local convergence is simple to achieve, the optimum solution is not obtained, and the result is low stability and less precision. By using the genetic algorithm global search, the improved GAFCM method can achieve the optimum overall and obtain the ideal solution. Introduce simultaneously the class concept vector to improve the method's accuracy[18]. The experimental results indicate that the enhanced text mining model is more effective and consistent than the fuzzy c mean process.

The remaining paper is structured accordingly. Section II describes the Textual Mining genetic algorithm; Section III gives a small Introduction about Fuzzy-C-Means ;Section IV describes the major figures suggested using sentimental analysis; Section V addresses the previous literature survey. Section VI ends the document.

V.LITERATURE SURVEY

Kumbhar et al.,(2017) Concentrates the use of genetic algorithm (GA) to achieve optimum functionality for unorganized data classification. Here, create a furious rules-depend classification that creates furious classification rules automatically. Two-databases , the 20-Newsgroup and the Reuters-21578 experiments are performed, or the findings show that the GA is above the Principal Component Analytics (PCA)[10].

ZHOU et al.,(2019) The characteristics of these data are discussed in this paper. First, use the TF-IDF analysis to obtain text features and convert them into vectors. The decision tree algorithm is then used to process the objects. In Addition to enhance the accuracy of the Bagging Ensemble category, a random sample learning on a text vector transformed through TF-IDF generates Bagging classification performance since the Bagging classification model is a set of Basic Classifier outcomes with a high recognition performance, Gene has been usedThe effects of this classification are produced by the Ensemble Classifier on the basis of genetic equations. Through experimental assesses in the Railways, the precise safety classification, the recall rate and the f-score values of the Evolutionary Ensemble Classifier framework were significant improvements in safety incident data hidden from the power supply intake catalysts [11].

Mortezanezhad, et al.,(2019) Addressing in this paper a proposal for a new Genetic Algorithm (GA) automated proposed technique in which cluster numbers are not necessary. The proposed method uses a very short genome encoding and recommends the correct crossover and mutation governors that generate an outstanding clustering impact. Our technique requires an unsupervised data point cluster learning model. To display the performance of the suggested algorithm, balanced/unbalanced real-world data with a 13-fold data vector is analyzed and an artificially produced random data set of 1,000,000 specimens is also given. In any case, our software beats the other implementations[12].

Hao et al.,(2017) Described a Twitter Time Series video examination that blends sentiment and stream observation with geological and time-based, interactive visualizations to analyze Twitter data streams in the real world. In order to effectively post -purchase survey information and Twitter amusement park data to detect fascinating patterns in consumer reviews, successfully applied visual sensation techniques to film tweets. The tools used today for visual analysis (e.g. SAS JMP, Vivisimo, Polyanalyst, etc.) include feedback data with yes / no questions, quantitative scores and specific statements [13].

SHI et al.,(2019) This study summarizes the existing results of the sentiment analysis in recent years and reflects on the methods and implementations of transfer learning in sentiment analysis[14].

El Alaoui et al.,(2019) This implies an adaptable method for sentiment analysis, evaluating social media responses and in real time extracting the consumer opinion. The solution proposed consists of first creating a dynamic word polarity dictionary depend on a select group of hashtags relating to particular subject. Then, the tweets are categorized under different groups by adding new technologies that greater refine the polarity of a message. We have coded tweets for the 2016 US election in order to confirm our approach. The prototype test results have been successful enough to identify both positively and negatively groups and their sub-groups [15].

Zhang et al.,(2015) Google proposed not a single algorithm, but two learning models, Continued Bag of Words (CBOW) and Skip-gram, are endorsed or suggested by Google. Word2Vec transmits wordvectors that can be interpreted either as a broad text or even the entire article by adding text data to either one of the learning models. In the first part of our research, we trained the data using the Word2Vec model. Furthermore, we group together similar terms and use the created clusters to fit into a new data dimension to reduce the data dimension [16].

Mythili et al.,(2016) This paper is the GA focused technique Fuzzy C Mean (GAFCM) for use in the section spots of complex micro-array (c-DNA) DNA images to determine gene way . The efficiency of the algorithm was assessed by generating simulated microarray slides whose actual mean values were known to be used for research. For the separations between foreground spot (FG) and background (BG), The synthetic images were contrasted and the performance obtained using K-means, Fuzzy C Means (FCM) and the suggested GAFCM method. The method intensity was utilized to evaluate the segmentation similarity factor, the coefficient of determination, the concordance correlation and the gene expression principles. Findings demonstrate that GAFCM segmentation performance is improved opposed to FCM architectures [18].

VI.CONCLUSION

"Big data" is small but larger, equivalent to records. The term 'massive' not only means the amount of data alone in large data. This also refers to the rate of origination of the data, its difficult layout and its roots from various sources. Here, show the comparsion of genetic algorithm with fuzzy-c-means that Improved GAFCM shows better Accuracy of text classification as compared to FCM .Many research works on genetic algorithms for text mining with Big Data are reviewed in this article. The proposed technology should apply a genetic algorithm to predict the exact text documents with sentiment analysis.In contrast with current methods, this will have better performance.

REFERENCES

[1] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, “A review of machine learning algorithms for textdocuments classification” Journal Of Advances In Information Technology, Vol. 1, No. 1, February 2017.

[2] Falguni N. Patel, Neha R. Soni, “Text mining: A brief survey” International Journal of Advanced Computer Research, Volume-2 Number-4 Issue-6 December-2016.

[3] S.N.Sivanandam, S.N.Deepa, “Introduction to genetic algorithms” Springer-Verlag Berlin Heidelberg 2018.

[4] M. Srinivas, Lalit M. Patnaik, “Genetic Algorithms: A survey” Motorola Indian Electronics Ltd., Indian Institute of Science, IEEE,2016.

[5] John Atkinson-Abutridy, Chris Mellish, and Stuart Aitken, “Combining information extraction with genetic algorithms for text mining” Published by the IEEE ,2015.

[6] M. Edison , A. Aloysius , “Concepts and Methods of Sentiment Analysis on Big Data”, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 9, September 2016.

[7] Min Chen, Shiwen Mao and Yunhao Liu “Big Data: A Survey”, Springer Science Business Media , pp: 171-209,2016.

[8] Lenka, V. , Satyanarayana, “A Survey on Challenges and Advantages in Big Data”, IJCST Vol. 6, Issue 2, April - June 2015.

[9] Vincent J. Amoruccio, MS, MA, “The Advantages & Disadvantages of Big Data”,2017.

[10] Kumbhar, P., Mali, M., & Atique, M. , “A Genetic-Fuzzy Approach for Automatic Text Categorization”, IEEE 7th International Advance Computing Conference (IACC),2017.

[11] LI, X., SHI, T., LI, P., & ZHOU, W. , “Application of Bagging Ensemble Classifier based on Genetic Algorithm in the Text Classification of Railway Fault Hazards”, 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD),2019.

[12] Mortezanezhad, A., & Daneshifar, E. , “Big-Data Clustering with Genetic Algorithm”, 5th Conference on Knowledge Based Engineering and Innovation (KBEI),2019.

[13] Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L.-E., & Hsu, M.-C. , “Visual sentiment analysis on twitter data streams”, IEEE Conference on Visual Analytics Science and Technology (VAST),2017.

[14] LIU, R., SHI, Y., JI, C., & JIA, M. , “A Survey of Sentiment Analysis Based on Transfer Learning. IEEE Access”, 2019.

[15] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A, “A novel adaptable approach for sentiment analysis on big social data”,Journal of Big Data”, Volume 5Issue 1, 2019.

[16] Ma, L., & Zhang, Y. , “Using Word2Vec to process big text data”, IEEE International Conference on Big Data (Big Data),2015.

[17] Atkinson-Abutridy, J., Mellish, C., & Aitken, S. , “Combining information extraction with genetic algorithms for text mining”,IEEE Intelligent Systems, Volume 19,issue3, pp 22–30,2017.

[18] Biju V , Mythili P, “ A Genetic Algorithm based Fuzzy C Mean Clustering Model for Segmenting Microarray Images”,International Journal of Computer Applications (0975 – 8887) Volume 52– No.11, August 2016.

[19] S.Wu and H.Yan,, “Processing Based on Clustering and Morphological” ,Pasific Bioinformatics Conference, Adelaide, Australia, pp. 111-118,2017.