IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Text Mining Technique on Big Data Using Genetic Algorithm

Deepankar Bharadwaj[1*], Arvind Shukla[2]

[1] School of Computer Engineering & Applications, IFTM University, Moradabad, India
[2] School of Computer Engineering & Applications, IFTM University, Moradabad, India

Available online at: www.ijcseonline.org

Accepted: 16/Jul/2018, Published: 31/July/2018

***Abstract* —** This paper provides the use of three terminologies and gives the best results with the details of methods implemented while applying applications of Genetic Algorithm for the Big Data which is mined from the plain text using Text Mining concept. The focus of this paper is to build up an algorithm that can extract or mine the details from plain text resumes & generates the method to provide the optimum solution. The method which is presented in this paper will help the organizations analyze that weather employee will work with them so long. Our main observation is that the system gives the results on the basis of details mined from the text resumes and based on these results, we can find the result that will helpful for the organizations.

Keywords: Text Mining, Genetic Algorithm, Big Data

I. INTRODUCTION

This section will provide the brief introduction about the paper and the three terminologies used in this paper. First, we will present the [8] Text Mining which is a very important method or process that is used to structure the unstructured data sets. These data sets can be further used for another process to get the best results. Unstructured data is something which is not in any format or we can say it is a plain text written anywhere and with the use of text mining applications and methods, we parse the unstructured data to be processed and gives us some information which is to be used for further processing. As we all know that the Text Mining is an old information retrieval technique and is a very important research area nowadays. As the data in text format is increasing day by day from MBs to GBs to TBs and so on and we do not have enough time to read, analyze all the information given in the plain text we have to adopt some methods that will help us in analyzing the text data and gives us an optimum results. In order to fulfill the user expectations of getting the required and relevant information from bulk amount of plain texts, one has to use the newest concepts and technologies.

Any data which is not easy to store, process, analyze can be considered as big data [2]. Nowadays massive amount 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. [1] Nowadays handling Big Data with old techniques and old algorithms has become a major challenge. With this increase in data regularly size, complexity, size, security issues are also increasing. One of the most used methods for finding the best and optimum solution, Genetic Algorithm [13] is the technique to provide optimum solution after Text Mining process. In this paper we use the concept of Genetic Algorithm to find optimum solution with the use of Text Mining on Big Data. It is the most important part for the mining process because it helps us to produce the optimized results. When adapting the genetic theory to the text categorization problem, the documents represented by a vector of terms become the chromosomes of the population [12]. 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 the 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.

```
1. Generate initial population.
2. Compute fitness of each individual.
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
```

This paper will continue with seven (7) sections that includes problem statement which gives the current problem for the organizations, proposed methodology gives the detail of the methodology we have used, implementation provides the description about the technology used for implementation, rules set gives the set of rules taken for implementation, results after the implementation, a conclusion based on results and the last future directions for others to continue.

II. PROBLEM STATEMENT

It is being noticed that many organizations perform manual efforts on the resumes that were applied to the vacant positions to get some of the required information from the plain text. They perform a manual task and check the information provided in resumes and prepare a summary file only with the required information that will help them to analyze the candidates at recruitment time. Generally for the experience candidates the analysis will be about the information based on their stability in their previous jobs. All of the organizations want to recruit the candidates who will work with them and be bonded with them for long time period. The employees who were not stable in their previous jobs may change this too. So organizations prefer to recruit those candidates who were stable and will stable with them for long periods. This manual process is time consuming and we also require man power to perform this task. We have thought over this situation and planned to develop an automated technique to resolve this problem. We have implemented the concept of Text Mining with the problem and get the information that is useful for the organizations at the time of recruitment.

III. PROPOSED METHODOLOGY

There are various methodologies for the information retrieval from the plain text based on the mathematical functions including algebra, probability, statistics etc. As discussed in section regarding the problem of manually checking and analyzing of resumes to see that weather the candidate will stable with them or not we are proposing the method. Text Mining using Genetic Algorithm. We have implemented the concept of Rule Based Text Mining on the data and apply Genetic Algorithm on the mined data to get the optimum results. We have implemented Classification and Prediction methodology for our paper work to provide the best solution. This approach is better because in case of resumes, all the information is classified on the same basis and afterwards we can predict the values for the processing of results. In this methodology, firstly we analyze our data and then classify the type of the information extracted from the text. It is a two step process which includes the building of a classifier that describes the predetermined set of data concepts and gives the accuracy of the text classified to the various classes.

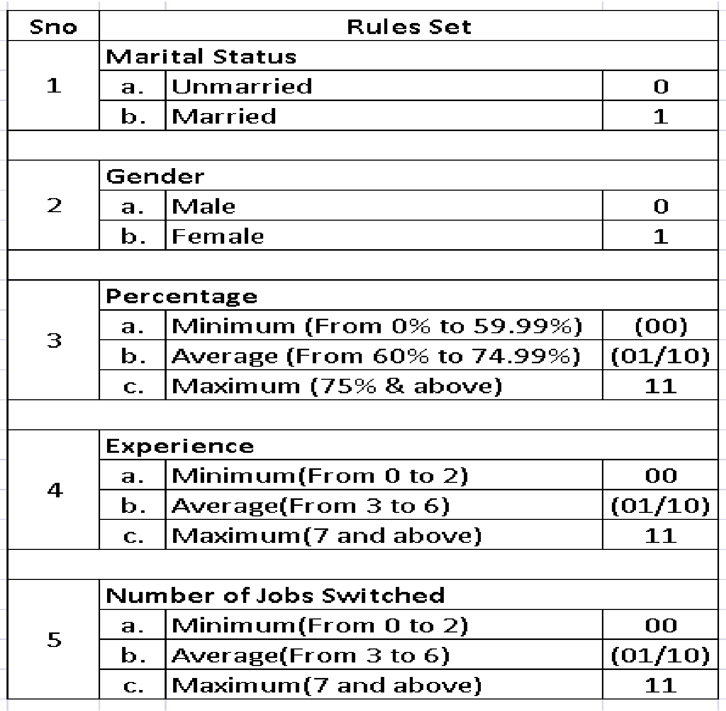
Nowadays lots of Classification methodologies are available, In this paper we have implemented Rule Based Classification method. In this method we can easily define our rule set to implement the process and generate the function. In rule based classification method, IF-THEN rules are used generally to represent the information in the form of bits for the better processing to classification. In this method we have defined some set of Rules for applying Genetic Algorithm to give the desired solution. We have taken some values initially for setting up our rules set. We are processing over the Percentage and Experience for both Male and Female with their Marital Status for the number of jobs changing.

IV. IMPLEMENTATION

In this section we deal with implementation of the proposed methodology and the application has been developed. It can be executed on any platform with some minimum requirement. Here the PHP & MYSQL for application development & storing the data. We use PHP and MYSQL because these are open source technologies.

V. RULES SET

We are considering an eight bit string with the first bit as its marital status, the second bit as its gender classification, the third and forth bit is for the percentage, next two bits are for the experience and the last two bits are for the number of Jobs switched. We are considering the three cases for each of the percentage, experience and the total number of Jobs change



Sno	Rules Set		
1	Marital Status		
	a.	Unmarried	0
	b.	Married	1
2	Gender		
	a.	Male	0
	b.	Female	1
3	Percentage		
	a.	Minimum (From 0% to 59.99%)	(00)
	b.	Average (From 60% to 74.99%)	(01/10)
	c.	Maximum (75% & above)	11
4	Experience		
	a.	Minimum(From 0 to 2)	00
	b.	Average(From 3 to 6)	(01/10)
	c.	Maximum(7 and above)	11
5	Number of Jobs Switched		
	a.	Minimum(From 0 to 2)	00
	b.	Average(From 3 to 6)	(01/10)
	c.	Maximum(7 and above)	11

Fig 1: Rules Set

SET-1

1st ITTERATION CHROMOSMES ARRAY WITH FITNESS TABLE

0	1	2	3	4	5	6	7
00010100	00010000	01000000	11010001	01010101	01110000	11010000	11010101

Index	First 6 Bits Match	Total Match	Fitness	Fitness/Sum	Round(FS)	Cumulative Fitness	Round Off(CF)
0	11	11	100	0.15384615384615	0.2	0.15384615384615	0.2
1	11	11	100	0.15384615384615	0.2	0.30769230769231	0.3
2	11	11	100	0.15384615384615	0.2	0.46153846153846	0.5
3	11	11	100	0.15384615384615	0.2	0.61538461538462	0.6
4	4	3	75	0.11538461538462	0.1	0.73076923076923	0.7
5	2	2	100	0.15384615384615	0.2	0.88461538461538	0.9
6	11	0	0	0	0	0.88461538461538	0.9
7	4	3	75	0.11538461538462	0.1	1	1

Above is the set of 8 chromosomes in which each bit represents the set of rules for the data used for the research work. Each set of chromosomes is a 8 bit string generated by the combination of rules set. In another table first column displays the count of the first 6 bits matches in the data set, the second column gives the count of total bits match in data set, third column displays the fitness value calculated in Ist iteration, fourth column displays the fraction value of fitness and total sum, fifth column shows the round off value for fourth column, the next and sixth column shows the cumulative fitness based on the fractional value of fitness and total sum. The last and seventh column shows the round off of the sixth column for each and every bit.

20th ITTERATION CHROMOSMES ARRAY WITH FITNESS TABLE

0	1	2	3	4	5	6	7
00010100	00010000	01000000	11010001	01010101	01110000	11010000	11010101

Index	First 6 Bits Match	Total Match	Fitness	Fitness/Sum	Round(FS)	Cumulative Fitness	Round Off(CF)
0	95	89	93.684210526316	0.2063617470759	0.2	0.2063617470759	0.2
1	1	0	0	0	0	0.2063617470759	0.2
2	7	0	0	0	0	0.2063617470759	0.2
3	1	1	100	0.22027377496865	0.2	0.42663552204455	0.4
4	25	16	64	0.14097521597994	0.1	0.56761073802449	0.6
5	1	0	0	0	0	0.56761073802449	0.6
6	27	26	96.296296296296	0.21211548700685	0.2	0.77972622503135	0.8
7	1	1	100	0.22027377496865	0.2	1	1

This table represents 20th iteration for the same set of chromosomes. Below is the another table representing the data processed by the another chromosomes set with the same column with different calculated values.

SET-2

1st ITTERATION CHROMOSMES ARRAY WITH FITNESS TABLE

0	1	2	3	4	5	6	7
01011100	00101110	01001000	00110101	00100110	01100000	00100000	01011001

Index	First 6 Bits Match	Total Match	Fitness	Fitness/Sum	Round(FS)	Cumulative Fitness	Round Off(CF)
0	11	11	100	0.14285714285714	0.1	0.14285714285714	0.1
1	2	2	100	0.14285714285714	0.1	0.28571428571429	0.3
2	11	11	100	0.14285714285714	0.1	0.42857142857143	0.4
3	11	11	100	0.14285714285714	0.1	0.57142857142857	0.6
4	2	2	100	0.14285714285714	0.1	0.71428571428571	0.7
5	11	11	100	0.14285714285714	0.1	0.85714285714286	0.9
6	11	11	100	0.14285714285714	0.1	1	1
7	2	0	0	0	0	1	1

Above is the set of another chromosome set which is implemented on the same data set with same number of iterations implemented. In this paper we use two sets of chromosomes to find weather the results can be more optimize or it may provide any other findings.

20th ITTERATION CHROMOSMES ARRAY WITH FITNESS TABLE

0	1	2	3	4	5	6	7
01011100	00101110	01001000	00110101	00100110	01100000	00100000	01011001

Index	First 6 Bits Match	Total Match	Fitness	Fitness/Sum	Round(FS)	Cumulative Fitness	Round Off(CF)
0	11	11	100	0.14469453376206	0.1	0.14469453376206	0.1
1	45	43	95.555555555556	0.13826366559486	0.1	0.28295819935691	0.3
2	11	11	100	0.14469453376206	0.1	0.42765273311897	0.4
3	11	11	100	0.14469453376206	0.1	0.57234726688103	0.6
4	45	43	95.555555555556	0.13826366559486	0.1	0.71061093247588	0.7
5	11	11	100	0.14469453376206	0.1	0.85530546623794	0.9
6	0	0	0	0	0	0.85530546623794	0.9
7	11	11	100	0.14469453376206	0.1	1	1

All the above tables represent the data calculated on two sets of chromosomes after 20 iterations. This data will help is finding result set.

VI. RESULTS

TABLE I. Result Set-I

Sno	Remark	Frequency
1	Unmarried Female with Average Percentage Minimum Experience change Minimum number of Jobs.	11
2	Married Female with Average Percentage Minimum Experience change Minimum number of Jobs.	7
3	Unmarried Female with Average Percentage Average Experience change Minimum number of Jobs.	7
4	Unmarried Male with Average Percentage Minimum Experience change Minimum number of Jobs.	7
5	Married Female with Average Percentage Average Experience change Minimum number of Jobs.	5
6	Unmarried Female with Minimum Percentage Minimum Experience change Minimum number of Jobs.	5
7	Unmarried Female with Maximum Percentage Average Experience change Minimum number of Jobs.	5
8	Unmarried Female with Maximum Percentage Minimum Experience change Minimum number of Jobs.	5
9	Married Male with Average Percentage Minimum Experience change Minimum number of Jobs.	3
10	Married Female with Minimum Percentage Minimum Experience change Minimum number of Jobs.	3
11	Unmarried Male with Average Percentage Average Experience change Minimum number of Jobs.	3
12	Unmarried Male with Maximum Percentage Minimum Experience change Minimum number of Jobs.	2
13	Married Female with Maximum Percentage Minimum Experience change Minimum number of Jobs.	2
14	Unmarried Female with Average Percentage Minimum Experience change Average number of Jobs.	1
15	Married Male with Maximum Percentage Minimum Experience change Minimum number of Jobs.	1
16	Unmarried Male with Minimum Percentage Minimum Experience change Minimum number of Jobs.	1
17	Unmarried Male with Minimum Percentage Average Experience change Minimum number of Jobs.	1
18	Unmarried Female with Average Percentage Maximum Experience change Minimum number of Jobs.	1
19	Married Female with Average Percentage Maximum Experience change Minimum number of Jobs.	1
20	Married Male with Average Percentage Average Experience change Minimum number of Jobs.	1
21	Unmarried Male with Maximum Percentage Average Experience change Minimum number of Jobs.	1
22	Married Female with Maximum Percentage Average Experience change Minimum number of Jobs.	1

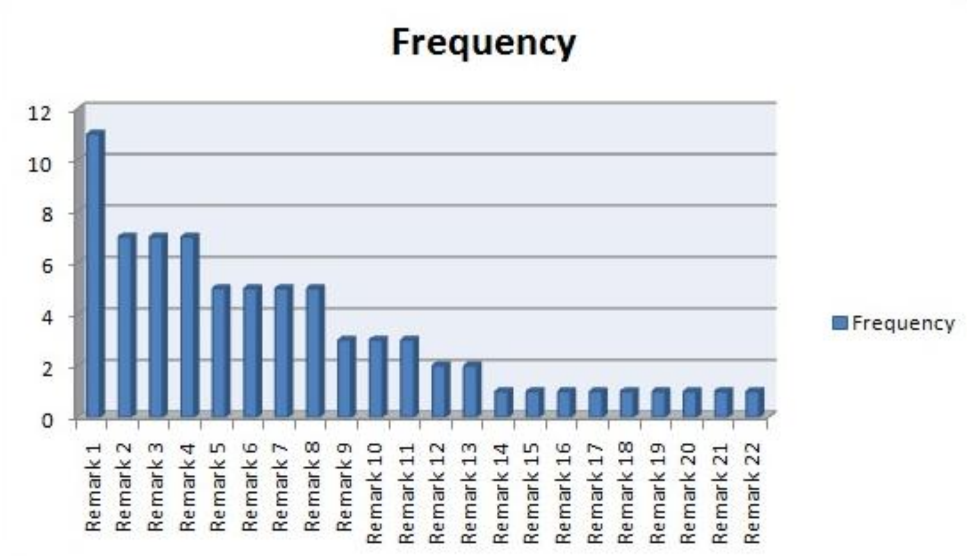


Fig 2: Chart representation for Ist result set

The above chart represents the pictorial representation of the results found after implementation of Genetic Algorithm operators to the first set of chromosomes. In Fig: 2, the first remark **Unmarried Female with Average Percentage Minimum Experience change Minimum number of Jobs** has maximum found with maximum number of frequency after successful iterations. Following is the tabular representation of results found after implementing same algorithm operators on another set of chromosomes.

TABLE II. Result Set-II

Sno	**Remark**	**Frequency**
1	Unmarried Female with Average Percentage Minimum Experience change Minimum number of Jobs.	7
2	Unmarried Male with Average Percentage Minimum Experience change Minimum number of Jobs.	6
3	Unmarried Female with Average Percentage Average Experience change Minimum number of Jobs.	5
4	Unmarried Male with Average Percentage Average Experience change Minimum number of Jobs.	5
5	Unmarried Female with Average Percentage Maximum Experience change Minimum number of Jobs.	3
6	Unmarried Female with Maximum Percentage Minimum Experience change Minimum number of Jobs.	3
7	Unmarried Female with Maximum Percentage Average Experience change Minimum number of Jobs.	3
8	Unmarried Male with Maximum Percentage Average Experience change Minimum number of Jobs.	3
9	Unmarried Female with Minimum Percentage Minimum Experience change Minimum number of Jobs.	2
10	Married Female with Average Percentage Minimum Experience change Minimum number of Jobs.	2
11	Married Female with Average Percentage Average Experience change Minimum number of Jobs.	2
12	Married Male with Minimum Percentage Minimum Experience change Minimum number of Jobs.	2
13	Unmarried Male with Minimum Percentage Minimum Experience change Minimum number of Jobs.	2
14	Unmarried Male with Average Percentage Maximum Experience change Minimum number of Jobs.	2
15	Unmarried Male with Minimum Percentage Average Experience change Minimum number of Jobs.	2
16	Married Female with Average Percentage Maximum Experience change Minimum number of Jobs.	1
17	Married Male with Average Percentage Average Experience change Minimum number of Jobs.	1
18	Married Male with Average Percentage Maximum Experience change Minimum number of Jobs.	1
19	Unmarried Male with Average Percentage Average Experience change Average number of Jobs.	1
20	Married Male with Average Percentage Minimum Experience change Minimum number of Jobs.	1
21	Married Male with Minimum Percentage Average Experience change Minimum number of Jobs.	1
22	Married Female with Maximum Percentage Minimum Experience change Minimum number of Jobs.	1
23	Unmarried Male with Minimum Percentage Maximum Experience change Minimum number of Jobs.	1
24	Married Male with Maximum Percentage Average Experience change Minimum number of Jobs.	1
25	Unmarried Female with Maximum Percentage Maximum Experience change Minimum number of Jobs.	1

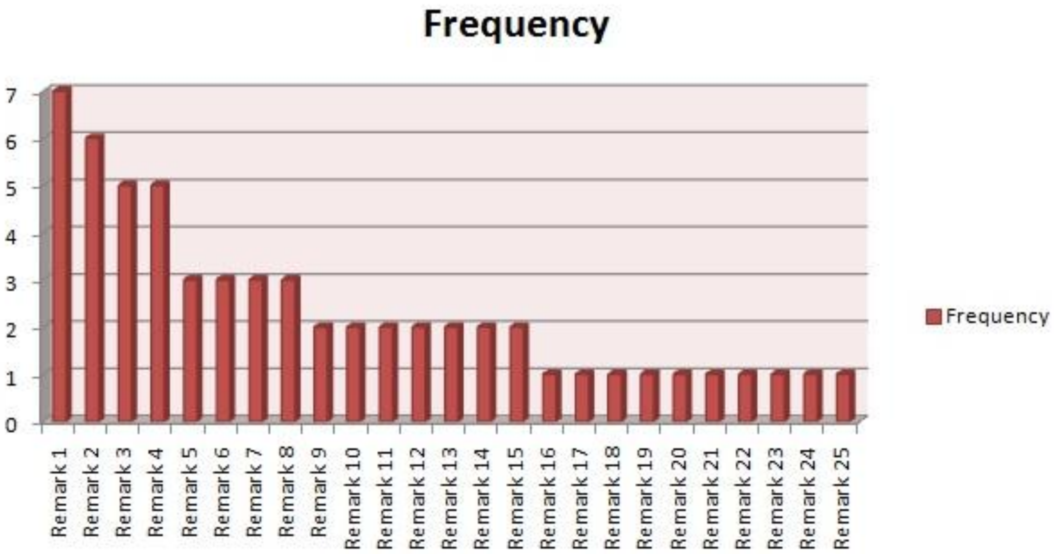


Fig 3: Chart representation for IInd result set

The above chart represents the pictorial representation of the results found after implementation of Genetic Algorithm operators to the first set of chromosomes. In Fig: 3, the first remark **Unmarried Female with Average Percentage Minimum Experience change Minimum number of Jobs** has maximum found with maximum number of frequency after successful iterations. With both the charts in Fig: 2 & Fig: 3 we have successfully implemented the proposed method and gives the optimum solution.

VII. CONCLUSION & FUTURE SCOPE

We have deal with a challenging task for mining problem of finding data set using our proposed Genetic Algorithm based method. This is successfully tested for number of data set such as **Unmarried Female candidates with Maximum Percentage, Minimum Experience changes Minimum number of Jobs**. The method defined here is simple and efficient one and helpful organizations to analyze weather an employee will work with them so long or not. The combination of text-mining and Genetic Algorithm technique is a relevant area of research. The future directions in this topic may be as follows:

- More information can be extracted with using the same methodology.
- Genetic Algorithm can be applied on other attributes also to optimize the results.
- Details can be extracted from the other file formats.
- Experiment could be tried in a real environment.
-

VIII. REFERENCES

[1] M. Bahrami and M. 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: pp 60-66, 2012.

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

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

[6] A. Kogilavani, “Clustering based optimal summary generation using Genetic Algorithm”, 2010.

[7] Suthaharan, S. 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, 2014.

[8] Indarjit Mukherjee, “Content Analysis based on Text Mining using Genetic Algorithm” (ICCTD, 2010).

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

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

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

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

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

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

Authors Profile

Mr. Deepankar Bharadwaj is pursuing Ph.D. from IFTM University, Moradabad (U.P.). He has completed his masters from College of Engineering & Technology, Moradabad (U.P.) in 2012 and completed his Engineering from College of Engineering & Technology; Moradabad (U.P.) in 2009 affiliated to Dr. APJ Abdul Kalam Technical University former Gautam Buddh Technical University. He is currently working as Assistant Professor in School of Computer Engineering & Applications and having 8 years of experience. He has published 5 papers in his area of research in reputed journals.

Dr. Arvind Kumar Shukla is an Assistant Professor, HOD in School of Computer Engineering & Applications, IFTM University, Moradabad. He has taught Data Mining and Mobile Computing to Post Graduate students of Department of Computer Engineering & Applications. He has active research interest in MANET and Mobility Model and has authored papers in the area. Dr. Shukla is an Alumini of Banasthali University, Rajsthan, India. He is currently working as Assistant Professor in School of Computer Engineering & Applications and having more than 13 years of experience. He has published more than 24 papers in his area of research in reputed journals.