Text Mining Technique using Genetic Algorithm

Deepankar Bharadwaj
CS, Department,
CET, IFTM, Moradabad

Suneet Shukla
IT Department,
CET, IFTM, Moradabad

ABSTRACT
The focus of this work is to develop a technique/algorithm which can mine the details from text resumes and give the optimized solution on the basis of the information extracted from the text. The details extracted are: communication details such as email ids, contact numbers, address, location etc.; academic details such as percentage in Matriculation, Intermediate, Graduation and Post Graduation; some personal details such as gender; and other details such as total experience in years. The details extracted from the text resumes are presented and processed for further mining operations. On these details, we have applied Genetic Algorithm to mine optimally the useful knowledge which can be of potential help to prospective employers.

Keywords
Text Mining, Genetic Algorithm, Knowledge Discovery

1. INTRODUCTION
There is a rapid increase in the text data in various document resources like plain documents, web pages etc. The need is to enhance the processing of the text so that the relevant knowledge which was not previously known can be mined from the text. Text Mining is a way to extract some meaningful information from bulk amount of textual data. [1] Text Mining gives the output on the basis of the patterns and regular expressions that were defined at the time of processing. Generally Text Mining techniques are used in case of World Wide Web which serves a huge, widely distributed, global information service centre for news, advertisements and many other information services. These challenges have promoted research into efficient and effective discovery and 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.

Text Mining is nearly as old as the information retrieval technique. Currently text mining is a very important area of research as the textual data is increasing day by day and we do not have time to read all the information given in the plain text. [4] In order to fulfill the user expectations of finding the relevant information from the bulk amount of texts, after neglecting the irrelevant information, one has to use the latest concepts and technologies. [5] It can be taken as one of the class 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.

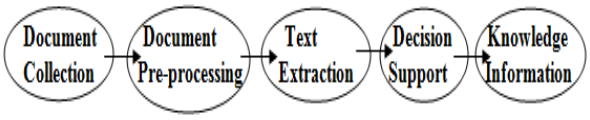

Fig. 1 Text Mining

Genetic Algorithms are the algorithms used to solve optimization problems. These algorithms are search based algorithm used to generate useful solutions for search problems. [3] They generate solutions to optimization problems using techniques such as selection, crossover and mutation. [8] These are the algorithms that encode a potential solution and apply recombination or crossover operators in such a manner so that it provides optimized result. Genetic Algorithms are often viewed as function optimizers which optimize the results as per the requirements of the users. Multiple problems to which genetic algorithms can be applied are quite broad and complex. [7] An implementation of Genetic Algorithm begins with a population of chromosomes. Chromosomes are to be considered as the individual solutions. [8] In a broader usage, a genetic algorithm is any population-based model that uses selection or reproduction, crossover or recombination and mutation operators to generate new results within the same space. This is illustrated below:

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.

   CONTINUE (until desired numbers of offspring are generated)
   - FOR population size DO Select 1 or 2 offspring randomly.
     a. For each selected offspring, select bits randomly to apply Mutation operation.

   CONTINUE (until desired numbers of bits are mutated)
   - Compute fitness of each offspring in new generation.
   - IF acceptable level of change in fitness achieved
     THEN Finished
   - ELSE
     Repeat Step 3

2. MOTIVATION
In finding the relevant and meaningful information from the bulk amount of plain text, Text Mining is the best way to get the desired information. For neglecting the irrelevant information and finding the relevant information, one has to

use the latest concepts and technologies of Text Mining. A resume in the form of plain text document consists of a collection of bulk amount of data. Many organizations perform the manual tasks at their host computers to process the document and to find only the relevant information that are important to them. We have implemented a system that will provide all the relevant information from those bulk resumes and give the details that are usually required by many organizations to compare and judge the applicants. The focus is to find the pattern relation within the data extracted from the resumes. The information may be regarding the details like contact information including Email ID, Address, personal details, the academic percentages in high school, intermediate, graduation, post graduation (if exists), total experience in years and total no of Jobs switched.

3. PROBLEM IDENTIFICATION

As we have seen that the text data is increasing day by day especially in organizations. The text data can be some times their information, records or it can be the resumes received at the time of recruitment process. Most organizations manually process the resumes to compile and/or tabulate the information relevant to the job. They manually check the data in resumes and prepare a single file only with the required information that will help them to analyze the candidates at recruitment time. The analysis can be of the information given in resumes based on their stability in previous jobs. In many cases the organizations may prefer candidates who will stay with them for a long period of time. The candidates who have frequently switched jobs in the past are not the likely candidates for consideration under the above mentioned preference of the employers. This manual process is time consuming and we also require man power to perform this task. We have planned to develop an automated technique to resolve this problem. We have implemented the concept of Text Mining using Genetic Algorithms to get the information that is useful for the organizations at the time of recruitment.

4. PROBLEM STATEMENT

The main focus of our study is to find the relevant information from the resumes that are in the form of plain text and apply Genetic Algorithm to optimize the results and give the knowledge to guide the recruitment process from the information extracted by this technique. This paper analyzes the resumes that are in plain text format and extracts the relevant information that is useful for the organization in the process of recruitment. The information is likely to aid the organizations in the recruitment process.

5. PROPOSED METHODOLOGY

There are various methodologies for the information retrieval from the plain text based on the mathematical functions from algebra, probability, statistics etc. As discussed in section regarding the problem of manually checking and analyzing of resumes we have designed and implemented a method which adopts Text Mining using Genetic Algorithm. [6] We have implemented the concept of Text Mining on the data and apply Genetic Algorithm on the mined data to get the knowledge from the data. We have implemented Classification and Prediction methodology for our thesis work to provide the better solution. 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 sets of data concepts and gives the accuracy of the text classified to the various classes. [2] We have number of Classification methodologies present nowadays. For our thesis work 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 considering the Percentage and Experience for both Male and Female candidates with their Marital Status and the number of jobs switched to predict their future stability with the organization.

6. IMPLEMENTATION

To implement the proposed methodology, an application has been developed, which can be executed in any environment with some minimum system requirements. For performing various methods, separate functions have been designed to perform the specified tasks. This methodology is implemented in PHP with MYSQL database for storing the data. We have used PHP and MYSQL because they are open source technologies. 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.

Unmarried	0	Married	1
Male	0	Female	1
Percentage From 0% to 59%			0
Percentage From 60% to 74%			(01/10)
Percentage From 75% & above			11
Experience From 0 to 2 Years			0
Experience From 3 to 6 Years			(01/10)
Experience From 7 and above Years			11
No. Of Jobs Switched From 0 to 2			0
No. Of Jobs Switched From 3 to 4			(01/10)
No. Of Jobs Switched 5 and above			11

Fig. 2 Rules Set

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

Fig. 3 Initial Chromosomes

Fitness	Fitness/Sum	Round(FS)	Cumulative Fitness
66.66666666667	0.10958904109589	0.1	0.10958904109589
66.66666666667	0.10958904109589	0.1	0.21917808219178
100	0.16438356164384	0.2	0.38356164383562
33.33333333333	0.054794520547945	0.1	0.43835616438356
75	0.12328767123288	0.1	0.56164383561644
100	0.16438356164384	0.2	0.72602739726027
66.66666666667	0.10958904109589	0.1	0.83561643835616
100	0.16438356164384	0.2	1

Fig. 4 Fitness Value for each Chromosome

First Random Number=0.05
VALUE=0.21917808219178, LOCATION=1
NEW CHROMOSOMES 00010000,

Second Random Number=0.28
VALUE=0.38356164383562, LOCATION=2
NEW CHROMOSOMES 00010000, 01000000,

Third Random Number=0.18
VALUE=0.21917808219178, LOCATION=1
NEW CHROMOSOMES 00010000, 01000000, 00010000,

Forth Random Number=0.18
VALUE=0.21917808219178, LOCATION=1
NEW CHROMOSOMES 00010000, 01000000, 00010000, 00010000,

Fifth Random Number=0.15
VALUE=0.21917808219178, LOCATION=1
NEW CHROMOSOMES 00010000, 01000000, 00010000, 00010000, 00010000,

Sixth Random Number=0.84
VALUE=1, LOCATION=7
NEW CHROMOSOMES 00010000, 01000000, 00010000, 00010000, 00010000, 11010101,

Seventh Random Number=1
VALUE=1, LOCATION=7
NEW CHROMOSOMES 00010000, 01000000, 00010000, 00010000, 00010000, 11010101, 11010101,

Eight Random Number=0.91
VALUE=1, LOCATION=7
NEW CHROMOSOMES 00010000, 01000000, 00010000, 00010000, 00010000, 11010101, 11010101, 11010101,

Fig. 5 Randomly generated Offspring

0	1	2	3	4	5	6	7
00000000	01010000	00010000	00010000	00010100	11010001	11010101	11010101

Fig. 6 Two Point Crossover

0	1	2	3	4	5	6	7
00000000	**00010000**	00010000	00010000	00010100	11010001	11010101	11010101

Fig. 7 Mutation I

0	1	2	3	4	5	6	7
00000000	00010000	00010000	00010000	**00110100**	11010001	11010101	11010101

Fig. 8 Mutation II

0	1	2	3	4	5	6	7
11110001	11110001	00010100	00010000	00010000	11110001	11110001	11110001

Fig. 9 Final Chromosomes after Ten Iterations

7. RESULTS

1. Married Female with Maximum Percentage, Minimum Experience change Average number of Jobs.
2. Unmarried Male with Average Percentage, Average Experience change Minimum number of Jobs.

8. RESULT AFTER MULTIPLE EXECUTIONS

Following is the set of results with their frequencies after running the system 25 times successfully.

SNo	Statement	Frequency
1	Unmarried Female with Maximum Percentage, Minimum Experience change Minimum number of Jobs.	11
2	Unmarried Female with Average Percentage, Minimum Experience change Minimum number of Jobs.	8
3	Unmarried Female with Maximum Percentage, Average Experience change Minimum number of Jobs.	7
4	Unmarried Male with Average Percentage, Average Experience change Minimum number of Jobs.	6
5	Unmarried Male with Average Percentage, Minimum Experience change Minimum number of Jobs.	5
6	Married Female with Average Percentage, Average Experience change Average number of Jobs.	3
7	Unmarried Male with Minimum Percentage, Average Experience change Minimum number of Jobs.	2
8	Married Female with Average Percentage, Minimum Experience change Minimum number of Jobs.	2
9	Married Female with Maximum Percentage, Minimum Experience change Average number of Jobs.	1
10	Unmarried Male with Average Percentage, Minimum Experience change Average number of Jobs.	1
11	Married Male with Average Percentage, Minimum Experience change Minimum number of Jobs.	1
12	Married Female with Average Percentage, Minimum Experience change Average number of Jobs.	1
13	Unmarried Female with Average Percentage, Average Experience change Average number of Jobs.	1
	Total Frequency of Statements	**49**

Fig. 10 Results after 25 Complete Iterations

9. CONCLUSION

Through this research work, we have described multiple aspects of Text Mining using Genetic Algorithm and introduce a new concept which illustrates its broad capabilities and functionalities. With the above set of results we can draw conclusion that which set of candidates having specified percentage, experience etc are likely to stay longer with the organization.

10. FUTURE DIRECTIONS

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.

11. REFERENCES

[1] Jiawei Han and Micheline Kamber, Second Edition, *Morgan Kaufmann* Publishers, Pg 318, 319, 351, "Data Mining Concepts & Techniques".

[2] S. Rajasekaran, G.A. Vijaylakshmi Pai, ISBN: 978-81-203-2186-1, PHI Learning, "*Neural Networks, Fuzzy Logic and Genetic Algorithms, Synthesis and Application"*.

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

[4] John Atkinson-Abutridy, Chris Mellish, University of Edinburg, IEEE 2004, "*Combining Information Extraction with Genetic Algorithms and Text Mining*".

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

[6] G. Desjardins, R. Godin, University of Quebec, Vol 35, 2005 WIT Press, ISSN-1743-3517, "*A Genetic Algorithm for Text Mining*".

[7] Tom V. Mathew, IIT Bombay, "*Genetic Algorithm*".

[8] Darrel Whitley, (1994) 4, 65-85, "*A Genetic Algorithm Tutorial*", Statics and Computing.