Automatic Extractive Text Summarizer (AETS): Using Genetic Algorithm

Alok Rai[1], Yashashree Patil[2], Pooja Sulakhe[3], Gaurav Lal[4],Dr.Rajesh.S.Prasad[5]

[1] *Student, Computer Department, NBNSSOE, Maharashtra, India*
[2] *Student, Computer Department, NBNSSOE, Maharashtra, India*
[3] *Student, Computer Department, NBNSSOE, Maharashtra, India*
[4] *Student, Computer Department, NBNSSOE, Maharashtra, India*
[5] *Principal, NBN Sinhgad School of Engineering , Maharashtra, India*

ABSTRACT

Text summarization is largely summarizing the source text into a simplified short version maintaining its actual information content and also the abstract that it means. Therefore supporting analysis of an already planned model of extractive text summarization, this paper focuses on the development of this extractive text summarization tool which is based on genetic algorithm and is named as AETS (Automatic Extractive Text Summarizer). Traditional summarizers use normally feature extraction and various techniques but they do not provide suitable extracted summary always. In this approach we have used feature extraction, fuzzy logic and genetic algorithm in order to train the machine for automatic summary generation in order to produce better results. The summarizer is also tested using standard datasets. In this way this system can be used to develop an automatic extractive text summary.

Keyword:- *Feature extraction, text summarization, part of speech, automatic text*

1. INTRODUCTION

In practical manner the large volume information is just too huge to store and handle. User won't be able to summarize this information in economical manner. Summarization done by machine is theoretical and report by human is an extractive outline. Text summarization ought to be temporary enough, however it should mean its original information. It plays very important role to extract helpful information. It is also helpful in convalescent data for user.

Hence, the algorithm mentioned below proposes a technique to expeditiously summarize the text so as to get sorted information. In this paper, we tend to contemplate the system of text Summarization as evolving system that learns incrementally through expertise within the surroundings. Text Summarization will be outlined as " extracting brief and correct information from given information, which is able to satisfy the user" .

2. LITERATURE SURVEY

From last few years, problem of text summarization has increased. In order to tackle with this problem various techniques are proposed. Out of the various techniques these are some well known proposed examples and some already existing summarizers.

- A genetic algorithm is proposed with special emphasis on the fitness function which permits to contribute with some conclusions.[1]
- An Evolving connectionist System is adaptive, incremental learning and knowledge representation system that evolves its structure and functionality.[2]
- Microsoft word summarizer just highlights the summary in abstractive way according to the demand.
- Copernic summarizer gives extracted summary which is not always relevant to the topic.
- A new text summarizer based on fuzzy logic.[3]
- Automatic text summarization by sentence with important features based on fuzzy logic.[4]

3. RELATED WORK

The extractive summaries are the ones which can be composed through precise phrases or phrases which are given within the supplied textual content. Then the problem of achieving extractive summaries from the bottom-text is reduced to discover the smallest set of sentences that constitute the entire text appropriately.

In practice, the extractive summaries are limited by means of size; for example, extractive summary needs to be no longer than the 10% of the entire textual content where the period of the summary extraction is calculated by means of the number of words [1]. This implies that for actual troubles, extractive summaries are definitely the first-rate feasible approximation of the base-textual content which fulfils the defined precise-constraints.

4. SUMMARIZATION PROCESS

The summarization process goes through following phases:

- Pre-processing
- Feature Extraction
- Fuzzy logic Scoring
- Genetic Algorithm
- Selection based on overall scoring
- Assembly of summary sentence

4.1 Pre - processing

4.1.1 Parsing

This process divides the sentences in different clauses defining its type, structure etc.For a given sentence as per the parts of speech the sentences are defined. E.g. : Stanford Parser**.**

4.1.1 Filtering Stop words: In this procedure the stop words i.e. articles, conjunctions are filtered.

4.1.2 Stemming: The root words are derived from its morphological variants eliminating the affixes (prefixes and suffixes).This pre-processed data is saved in the database and is used for further processing.

4.2 Feature Extraction

For calculating summary, it is essential to represent sentences within the vector shape with the intention to offer us and attributes so one can constitute the input records.[3] For our need, we are going to use 5 features to extract the exact sentences from input data and so that it will provide output as ‘ 0’ or ‘ 1’ .

4.2.1 Title features

It can be calculated by using the ratio of number of titles in words by number of word in title [4].

$$\text{output} = \frac{\text{Number of title words in document}}{\text{Number of words in title}}$$

4.2.2 Sentence length

It can be calculated in order to find out short sentences such as book name, authors, sub meanings [5].It can be calculated as,

$$\text{output} = \frac{\text{number of words in sentence}}{\text{number of words in longest sentence}}$$

4.2.3 Term weight

Weight age of sentence means importance of sentences in the document. It can be calculated by taking occurrences of sentence in document [4]. It can be calculated by,

$$\text{output} = \frac{\text{Sum of TF ISF}}{\text{Max(sum of TF ISF)}}$$

Where TF-ISF is nothing but term Frequency and ISF is Inverse term frequency.

4.2.4 Sentence position

Position of sentence is important for deciding the importance of sentence. So for the first sentence output will be 1 and for other it will be zero [5].

4.3 Fuzzy Logic Scoring

The above mentioned features are applied on the sentences to determine their significance. These extracted text features are used to map into the fuzzy logic to score each sentence of the document. This score is then used to extract the sentences for the summary of the document.

- **Fuzzifier:** It gives linguistic values to the input feature values in set of low, very low, medium, high, very high. These values constitute a fuzzy set FS for the sentences. The supports of these fuzzy sets determine the importance of the sentence.
- **Rule Base:** It defines the fuzzy IF-THEN rule that specifies the occurrence of the sentence in the various paragraphs and phrases of the document.
- **Defuzzifier:** The values defined by the fuzzifier are converted to the crisp values. These values determine the closeness of the sentence to the given linguistic values.

4.4 Genetic algorithm Overview

Genetic algorithms are stochastic search strategies coping with a population of simultaneous seek positions [1] [2] [5]. A traditional genetic algorithm consists of three important elements:

- **The coding of the optimization hassle**: It produces the required discretization of the variable values (for optimization of actual functions) and makes their management easy in a population of search factors viable. Generally the maximum variety of seek points, i.e., the population size, is constant at the start.
- **The mutation operator**: It determines the opportunity with which the facts structures are changed. This may arise spontaneously (as in stochastic search) or a handiest whilst the strings are combined to generate a brand new populace of seekspoints [1]. In binary strings a mutation corresponds to a bit flip [8]. The records trade operators manipulate the recombination of the hunt points which will generate a brand new, better population of factors at each new release step [4].
- **Exchange of operators:** Earlier than recombining, the feature to be optimized must be evaluated for all records structures in the population. The search points are then looked after inside the order of their

function fee, i.e., within the order of their so-called fitness [1]. In a minimization problem the points which might be located at the beginning of the listing are the ones for which the characteristic fee is lowest. The points for which the feature to be minimized has the best characteristic cost are placed on the give up of the listing.

The overall working of AETS system can be explained with following diagram:

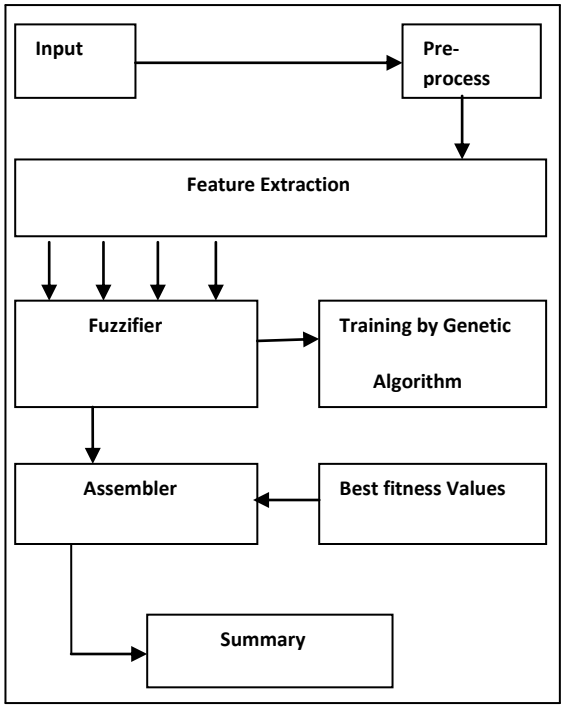

Fig1. System Architecture Overview

4.5 Genetic Algorithm Working

Genetic algorithm is the search method in which principle of natural selection and genetics is used. Solutions for the problem are considered as set of candidate solutions. This candidate solution is known as " chromosomes" . And the alphabets from the string are called as genes and the values of genes are known as alleles. Take an example, in travelling salesman problem. In these problem chromosomes is nothing but the route. And gene is nothing but city. Genetic algorithms have made real impact on all those problems in which there is not enough information to build a differentiable function or where the problem has such a complex structure that the interplay of different parameters in the final cost function cannot be expressed analytically[2] [5].In genetic algorithm, it mainly focuses on the population of candidate solution. This is a user defined parameter which will affect on factors such as scalability and performance of algorithm [2] [5].

Following are the processes of genetic algorithm:

4.5.1 Initialization

The initial population of candidate solutions is usually generated randomly across the search space. However, domain-specific knowledge or other information can be easily incorporated. Though it is simple to implement but

when the text size grows it may be complex to generate candidate solutions (chromosomes). Chromosomes consist of sequence of positive integers that represents the sentence number of original text. Its size is fixed and decided by the compression rate of the text.

4.5.2 Selection

" Survival of the fittest" can be achieved through selection operator. Selection allocates more copies of those solutions with higher fitness values and thus imposes the survival-of-the-fittest mechanism on the candidate solutions [3]. In AETS mechanism we are trying to prefer better solutions to worse ones, and many selection procedures have been proposed to accomplish this idea, including roulette-wheel selection, stochastic universal selection, ranking selection and tournament selection[4][8]. In our selection process the selection of parent string:-Selection with replacement is used, i.e., the whole population is the basis for each individual parent selection. It can occur that the same string is selected twice. We use selection criteria such based on Fuzzy logic as mentioned above.

4.5.3 Crossover

The document is divided into integer valued sizes. Chromosomes are selected randomly based on the selection criteria defined above. These values are called as parent sets. Crossover of these parents gives us new individuals. Now we start comparing the current values to the previous rejected values and try to get relation between them. If the parent the same genes as of its neighbours then it will be discarded for redundancy purpose. Recombination combines parts of two or more parental solutions to create new, possibly better solutions (i.e. offspring). There are many ways of accomplishing this (some of which are discussed in the next section), and competent performance depends on a properly designed recombination mechanism [1].

For the recombination of two strings a cut-off point is selected between the two positions and then a crossover is carried out. The probability that a schema is transmitted to the new string depends on two cases. If both parent strings contain, then they pass on this substring to the new string.

4.5.4 Mutation

Mutation is an operator that produces random changes in various chromosomes and maintains the diversity of the population. A random gene in the chromosomes can be selected and replaced with the values which do not duplicate the sentence. While recombination operates on two or more parental chromosomes, mutation locally but randomly modifies a solution. Again, there are many variations of mutation, but it usually involves changes being made to an individual' s trait or traits. In other words, mutation performs a random walk in the vicinity of a candidate solution. The offspring population created by selection, recombination, and mutation replaces the original parental population. Many replacement techniques such as elitist replacement, generation-wise replacement are also used [4] [8].When two strings are recombined, the information contained in them is copied bit by bit to the child string. A mutation can produce a bit flip with the probability. This means that a schema with fixed bits will be preserved after copying with probability [4] [8].

4.5.5 Stopping Criteria

The implementation of genetic algorithm takes numerous iterations. Thus stopping criteria must be used. They can as follows:

a) When an upper limit of the number of generation is reached or

b) Chance of getting changes in the consecutive generation is extremely low

4.6.6 Assembly

These output sentences are in their descending order of importance. Depending upon the required summary length these sentences are appended in a summary text file which provides as an output for our Summarizer.

5. IMPLEMENTATION

User can use this application to filter out required data. In first stage, user will be provided login credentials to login to application. User will enter username and password. Validation is done. If user enters wrong login credentials then pop window will be shown. It is shown in fig 2(a) and fig 2(b).

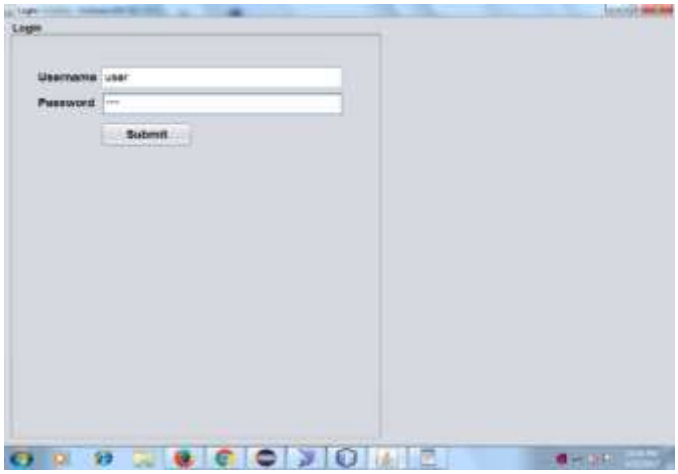

Fig 2(a)

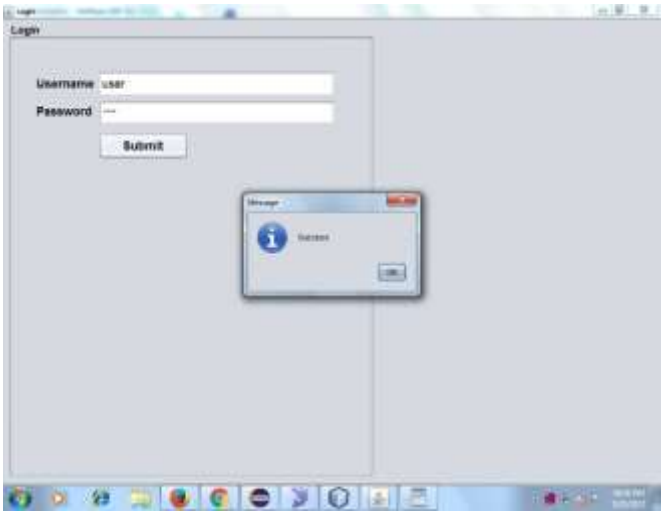

Fig. 2(b)

After successful login, user will select any random text file. As shown in fig. 3, text file is selected from computer. This file is uploaded for further process.

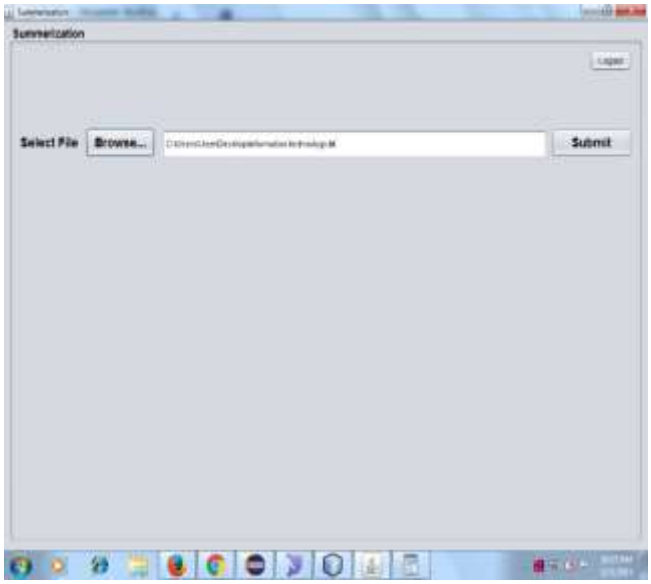

Fig. 3

We apply Stanford LLP parser for used to parse input data written in several languages. After parsing, feature extraction is next stage for process. Features are Title features, Sentence length, Term weight, Sentence position etc. as shown in fig. 4

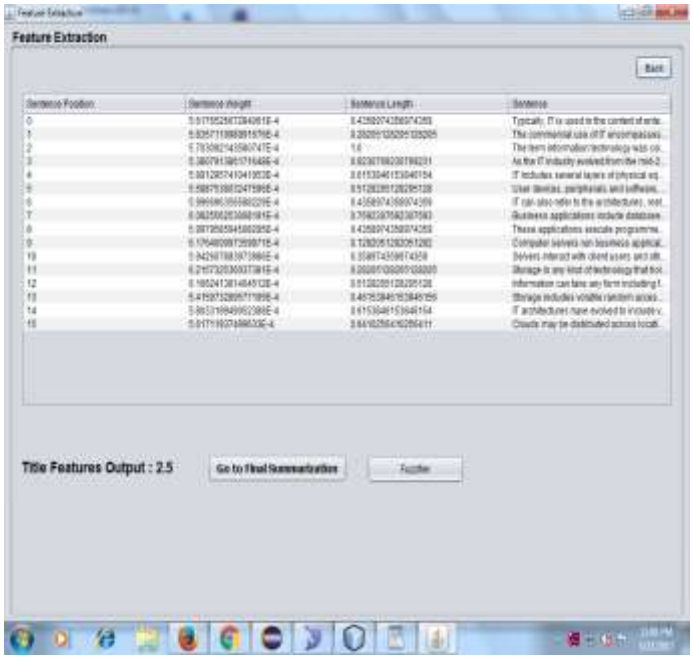

Fig.4

In next stage, we would apply fuzzy logic. User will get fuzzy values.

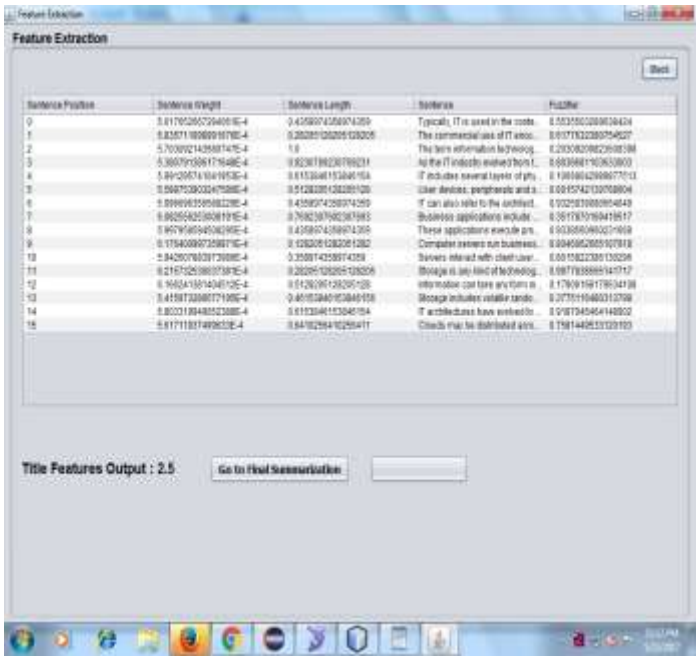

Fig. 5

We are applying genetic algorithm for best fitness value. It is shown in fig. 6

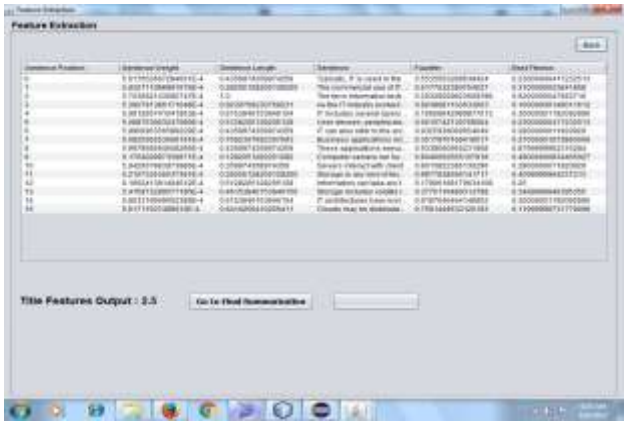

Fig.6

At final step, we will get final text summarization. It is shown in fig. 6

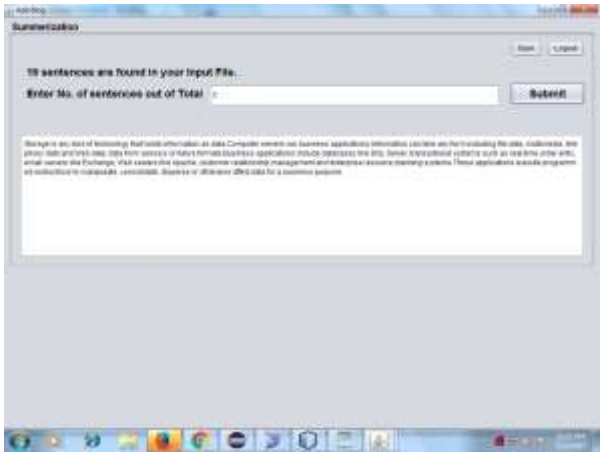

Fig. 7

6. RESULT

The system has implemented all the steps as defined above in the summarization procedure. According to the input file and compression rate given by the user respective number of relevant sentences which form a meaningful summary are extracted and given as output.

As visible from the results genetic algorithm is a sentence choice-based technique. Text summarization is the method of reducing a text file with the help of genetic algorithm that retains the most essential textual content of any authentic file. For preferred domain names is sentence extraction and highlight most effective important text which we need. Using genetic algorithm we can put into effect language modelling, multilingual summaries, summarization of email, spoken document summarization also. Using text summarization, spotting portions of the summary that is in shape of the input files is also easy. We will exactly relate to the existing summarizer. Genetic algorithm helps to build a summarizer from scratch and separates it in to a possibly short form. Thus we get a short summary of the document saving our time trying to understand it.

7. CONCLUSIONS

As visible from the results genetic algorithm is a sentence choice-based technique. Text summarization is the method of reducing a text file with the help of genetic algorithm that retains the most essential textual content of any authentic file. For preferred domain names is sentence extraction and highlight most effective important text which we need. Using genetic algorithm we can put into effect language modelling, multilingual summaries, summarization of email, spoken document summarization also. Using text summarization, spotting portions of the summary that is in shape of the input files is also easy. We will exactly relate to the existing summarizer. Genetic algorithm helps to build a summarizer from scratch and separates it in to a possibly short form. Thus we get a short summary of the document saving our time trying to understand it.

8. REFERENCES

[1]. Sebastian Suarez Benjumea , Elizabeth Leon Guzman, " Genetic Clustering Algorithm for Extractive TextSummarization." 2015, IEEE Symposium Series on Computational Intelligence.

[2] René ArnulfoGarcía-Hernández and YuliaLedeneva" Single Extractive Text SummarizationBased on a Genetic Algorithm" MCPR 2013, LNCS 7914, pp. 374– 383, 2013.

[3] Rajesh S.Prasad, U. V. Kulkarni, and Jayashree.R.Prasad " Connectionist Approach toGeneric Text Summarization" scholar.waset.org/1999.4/1999 , 2009.

[4] Rajesh Shardanand Prasad and UdayKulkarni " Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization." Journal of Computer Science,ISSN 1549-3636, 2010.

[5] Ladda Suanmali, Naomie Salim, Mohammed Salem Binwahlan " Automatic text summarization using fuzzy extraction" Bil 2 , December 2008.

[6] Rajesh S.Prasad, Dr.U.V.Kulkarni " A Novel Evolutionary Connectionist TextSummarizer" 978-1-4244-3882- 2/09/$25.00©2009 IEEE.

[7] Uplavikar Nitish Milind,Wakhare Sanket Shantilalsa, Prof.Dr.R.S.Prasad " Feature Based Text Summarization." International Journal of Advances in computing and Information Researches Vol.1,No. 2,2012.

[8] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Chingiz A.Mehdiyev. " Sentence selection for generic document summarizationusing an adaptive differential evolution algorithm." *Swarmand Evolutionary Computation*, 1(4):213 – 222, 2011.

[9] Rajesh S.Prasad, U. V. Kulkarni, " Two Approaches to Automatic Text Summarization: Extractive Methods and Evaluation." International Journal of Computer Engineering and Computer Applications, Vol.01 Issue no.1,January 2010- March 2010.

[10] Rajesh S.Prasad, U. V. Kulkarni, " An Automated Approach to Text Summarization using Fuzzy Logic." International Journal of Computer Engineering& Information Technology, IJCEIT ISSN 0974-2034, Vol 23, Issue no 01, March 2010-May2010.