I.J. Modern Education and Computer Science, **2015, 10, 33-39**
Published Online October 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.10.05

Extractive Text Summarization Using Modified Weighing and Sentence Symmetric Feature Methods

Selvani Deepthi Kavila
Assistant Professor, Department of CSE, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, India.
E-mail:selvanideepthi14@gmail.com

Dr.Radhika Y
Associate Professor, Department of CSE, Gitam Institute of Technology, Gitam University, Visakhapatnam, India.
E-mail:yradhikacse@gmail.com

Abstract—Text Summarization is a process that converts the original text into summarized form without changing the meaning of its contents. It finds its usefulness in many areas when the time to go through a large content is limited. This paper presents a comparative evaluation of statistical methods in extractive text summarization. Top score method is taken to be the bench mark for evaluation. Modified weighing method and modified sentence symmetric feature method are implemented with additional characteristic features to achieve a better performance than the benchmark method. Thematic weight and emphasize weights are added to conventional weighing method and the process of weight updation in sentence symmetric method is also modified in this paper. After evaluating these three methods using the standard measures, modified weighing method is identified as the best method with 80% efficiency.

Index Terms—Text summarization, Top Score Method, Weighing method, Sentence symmetric feature Method.

I. Introduction

Text summarization falls under the area of text mining and information retrieval where the main objective is to retrieve valued information from text. In the process of summarization the input could be text documents or multimedia files such as audio, image or video. Text Summarization is used to save time in text mining and information retrieval. Automatic summarization is the process by which computer program creates a shortened version of text. The goal of automatic summarization is reducing the size or volume of source text into a short version that holds the overall meaning and information content.

There are two approaches in automatic summarization systems namely extractive and abstractive. The former approach works by selecting important sentences/phrases/subset of existing words. The selection of important sentences forms the key idea in these methods. Based on a predefined function, each sentence is evaluated and most important ones are extracted from the original text in the original form. On the other hand, abstractive methods construct an internal semantic representation of the text. In these techniques, the intention is to generate a summary which is close to what a human would generate. Unlike in the extractive approaches, the sentences are reformed or regenerated based on the semantic relationships in the original text. This work focuses on automatic summarization of text documents using extractive methods.

In extractive approaches, one of the most important phases in text summarization process is identifying significant words of the text. Significant words play an important role in specifying the best sentences for summary. The top score method[10] extracts significant sentences by giving score to every sentence based on the significant words. A combination of techniques like statistical methods and semantic relationship methods are used to identify significant words.

The rest of paper is organized as follows: Section 2 describes the Literature survey related on Document Summarization. Section 3 presents the architecture of the system and improved methodologies. This section also presents the comparative study of the proposed methods with various summarization techniques. Section 4 describes the Results and Performance Analysis followed by conclusion and future work in section 5.

II. Related Work

A. Back ground work related to document summarization

Text summarization is the process of reducing the text with a computer program to create a summary that keep the most important points of the original document. At first Text summarization was done by Hans Peter Luhn [1] (Father of Information Retrieval) in 1958. His main target is to get summarization of technical literature. It is based on frequency of most significant words and their relevant positions. In this method sentence scoring was done and top scored sentences are extracted.

H.P.Edmundson [2] has proposed a new method for automatic text summarization in 1969. His method is based on four different weighting methods i.e. Cue, Title, Location and Key method. With these four methods he calculated the sentence scores for every sentence and marked the highest scoring sentence as the most important one. The main disadvantage of this method is, irrelevant data and the longer sentences in the document are displayed. A. Das et al [3] proposed a neural net model used to pre-process an input string and match with the user defined string. They extracted featured words from the given text with the user defined words .If there is a match then the value of that word increases. This process repeats until it attains a constant value and total sentence score is then calculated. Therefore, a sentence with higher score will be the first one. Another important characteristic is to integrate a semantic module to refine the search words like detecting association among search words, etc.J Jagadeesh et al [4] proposed Sentence Extraction Based Single Document Summarization. In their research they discussed about the techniques to achieve readable and coherent summaries. Arman Kiani et al [5] proposed Text summarization using Hybrid Fuzzy systems which is based on summarizing a text on the fusion of Genetic System. Saeedeh Gholamrezazadeh et al [6] presented different types of summarization methods and a common summarized system was implemented. They also discussed the most important issues in evaluating a summary and presented common criterion for evaluating a summarized system.Ladda Suanmali et al [7] proposed a Fuzzy logic method for improving text summarization approach. They improvised summary by using general statistic method. Rasim Alguliev et al[8] proposed a sentence based extraction method by using new functions for finding the sentence clustering approach. This is most probably used for document summarization. Vishal Gupta et al [9] presented a survey on different extractive summarization methods. Maryam Kiabod et al [10] proposed a Top Score algorithm where they calculate the local and global scores for the words and also identified the significant words for the given text. Masrah Azrifah Azmi Murad et al [11] proposed a similarity method with topic similarity by using fuzzy sets and probabilities. Based on these scores they extracted the important sentences from the given document. Rafeeq Al-Hashemi [12] proposed the text summarization using extracted keywords. In this work operation is performed in four stages. In the first stage pre-processing was done, key phrases are identified in the second stage, sentences were extracted in the third stage and in fourth stage summary is produced. Shaidah Jusoh et al [13] proposed various techniques used in text summarization like Information retrieval etc. and also proposed about the applications and challenging issues in text summarization approach.

B. Existing Summarization tools

There are some summarization tools to generate summaries .Some of them are:

Free summarizer:

It is a tool that generates the summary based on the number of sentences required in the summary. The disadvantage of this tool is that the summary is not efficient.

Auto summarizer:

It is a tool that also generates the summary on the number of sentences required in the summary. The disadvantage of this tool is that semantic relation is missing in the summary.

Online Summarizer:

It is a tool that generates the summary based on the threshold value. The summary varies according to the threshold value. The disadvantage of online summarizer is when document doesn't contain good summary sentences it summarizes poorly and also when user provides url or text it can’t get the right abstract document.

Open text summarizer:

It is a tool to summarize texts. The program reads texts and conforms which sentences are important and which are not. The Open Text Summarizer is both library and a command line tool. The main disadvantage with this tool is it doesn‘t indicate the important sentences because of the repetition. The main sentences are missing in the summary.

Text compactor:

It is a tool in which there are three steps to be followed namely uploading the document, dragging the required percentage and summarizing the document. Whenever the input text is too long text compactor unable to summarize it.

III. Proposed System

Linguistic roles identification is the first module in this work where linguistic roles are identified to make the task of researcher easy. This is performed using methods of keywords extraction and based is on fonts.

Fig. 1 shows the flow of execution. Initially the documents are uploaded in IEEE format. In the first step a document is selected based on the rhetorical roles from the set of documents which are present in the repository and the text besides the keywords is extracted. The extracted text is fed to Text Processing stage where the whole text is divided into number of sentences and tokens. Later, it goes to the Intermediate stage where it performs all the pre-processing steps i.e. Stop word removal, Stemming etc. After that it calculates sentence scores for respective algorithms based on their formulas. Based on the sentence score the sentence extraction is performed, i.e. the highest ranked sentence will be the first one in the summary. Then final output of system generated summary is given. The comparison ratio is found by comparing the system generated summary with manual summary by using relevance measures. The final result i.e. comparison table is measured.

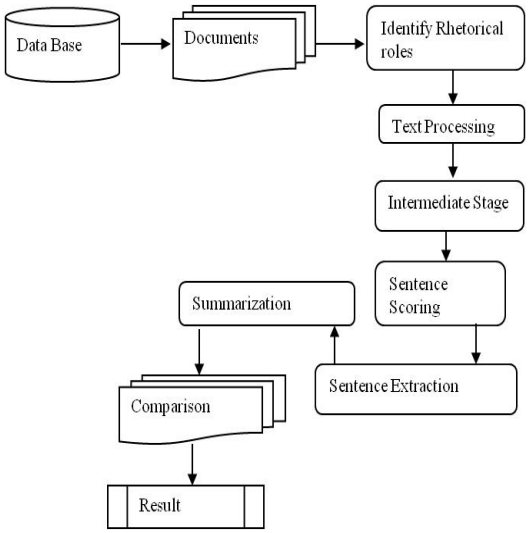

Fig.1. System Architecture

The following important rhetorical roles are used in this paper.

Abstract:

This Keyword is found after the title of the paper and names of the authors. It contains the text which is of around 200 words which gives the essence of the whole document This abstract can be further summarized so that user can get the essence of the paper by reading only a few lines.

Introduction:

This keyword is found after the keywords or index terms. This contains text which is of 3 to 4 paragraphs. It gives information related to domain, existing system, proposed system and the sections that will be further dealt in the paper. Once the "Introduction" keyword is identified based on rhetorical roles, the text beside "Introduction" is extracted and it undergoes all the phases till summarization. The output text of this "Introduction" contains domain of the paper and important points are to be extracted based on scoring factor.

Conclusion:

This keyword is identified by the word "conclusion". The text besides this undergoes all the stages and finally a summarized text will be produced which gives information about the work done in the paper and also the future work.

A. Implementation details

In this paper three summarization algorithms are implemented which mainly focuses on research papers of the given area. The three algorithms are, as follows:

- Top-Score Algorithm
- Modified Sentence symmetric feature Algorithm
- Modified Weighing method Algorithm

Top score algorithm [10] is an existing well defined method. In this work sentence symmetric algorithm is used in a modified way to be compared with the top score method. The modifications are done to include more features like thematic weight and emphasize weight. Weighing method is used in the conventional manner but the way in which weights are given is changed and also a graphical matrix representation is used.

B. Modified Sentence Symmetric Feature Method

In Sentence Symmetric feature algorithm the following attributes are used to calculate the sentence score.

- Cue
- Key
- Title
- Location

To calculate the sentence score the formula S= aC+bK+cT+dL is used.

Where C – Cue weight, K – Key weight, T – Title weight, L – Location weight and a,b,c,d are set of positive integers in the range [0,1] .

The main disadvantage of using this method is irrelevant data is also being displayed. To overcome this disadvantage, a modified version of the above scheme is used in which instead of calculating the key weight, two more features are added i.e.,

1. Thematic weight of the sentence.
2. Emphasize weight of the sentence.

So, the Modified Sentence Symmetric Feature consists of

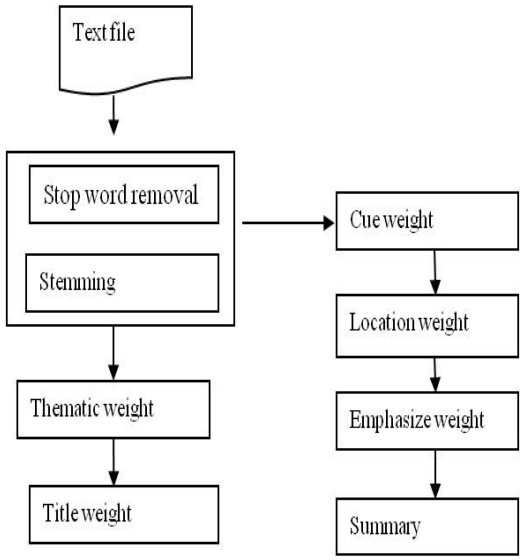

Fig.2. Data flow diagram of Modified Sentence Symmetric feature Method

Cue Weight for sentences:

The Cue Weight for sentences is calculated by adding the cue weight of its constituent words, it is a quantitative description. This depends up on the hypothesis that has significant implications for language acquisition, and is applicable for the specification of a particular sentence

by the its existence or nonexistence of particular cue words in the cue dictionary.

Total number of cue words present in a sentence s is denoted by Cwj (Sj) and total number of cue words in the document is denoted by Cwi.

Thematic Weight for sentences:

Thematic words are defined as most frequent words. The functions of the thematic words frequencies are Sentence scores.

Where indicates Total number of thematic words present in a sentence s is denoted by Thej(Si) and total number of thematic words present in the document is denoted by Thei.

Title Weight for sentences:

Here the sentence weight is calculated by the addition of all the words in the content which are given in the title and sub title of a text.

Total number of title words present in that sentence s is denoted by Tij(Si) and total number of title words in the document is denoted by Tii.

Location Weight for sentences:

The importance of sentence is indicated by its location, sentences tend to occur at the beginning or in the end of documents or paragraphs based on the hypothesis. A greatest correlation is achieved between the human-made exception and automatic exception by adding the three latter methods and the results are shown.

Location of the sentence s is denoted by Lj(Si) and total number of sentences present in the document is denoted by Si.

The proposed algorithm is presented below.

Table 1. Steps for Modified Sentence Symmetric Feature Method

Algorithm
Step 1: Sentence segmentation is performed.
Step 2: for each sentence s do
Step2a:Cue Weight for sentences :
for Cwj in Si do
$C = \Sigma Cwj(Si) / \Sigma Cwi$
Step2b:Thematic Weight for sentences :
for Thej in Si do
$Th = \Sigma Thej(Si) / \Sigma Thei$
Step 2c:Title Weight for sentences :
for Tij in Si do
$T = \Sigma Tij(Si) / \Sigma Tii$
Step 2d:Location Weight for sentences :
for Lj in Si do
$L = \Sigma Lj(Si) / \Sigma Si$
Step2e:Emphasized words Weight for sentences :
for Emj in Si do
$E = \Sigma Emj(Si)$
Step 3.End
Step 4.For each sentence do
Sentence Score :
$Sf = C + Th + T + L + E$
Step 5.End
Step 6.Return sentence score.

C. Modified Weighing Method

a) Pre-processing:

The first step in text summarization involves preparing text document to be analyzed by the text summarization algorithm. First of all we perform sentence segmentation to separate text document into sentences. Then sentence tokenization is applied to separate the input text into individual words. Some words in text document do not play any role in selecting relevant sentences of text for summary, Such as stop words ("a", "an", the"). For this purpose, part of speech tagging is used to recognize types of the text words. Finally, nouns of the text document are separated.

b) Calculating word local score:

Local score of a word is calculated by using term frequency and sentence count Term frequency is defined as frequency of the word normalized by total number of words. Sentence count is the no of sentences containing the word normalized by total no of sentences.

c) Title Weight for sentences:

Here the sentence weight is calculated by the addition of all the words in the content which are given in the title and sub title of a text.

Total number of title words present in that sentence s is indicated by Tij(Si) and total number of title words in the document is indicated by Tii.

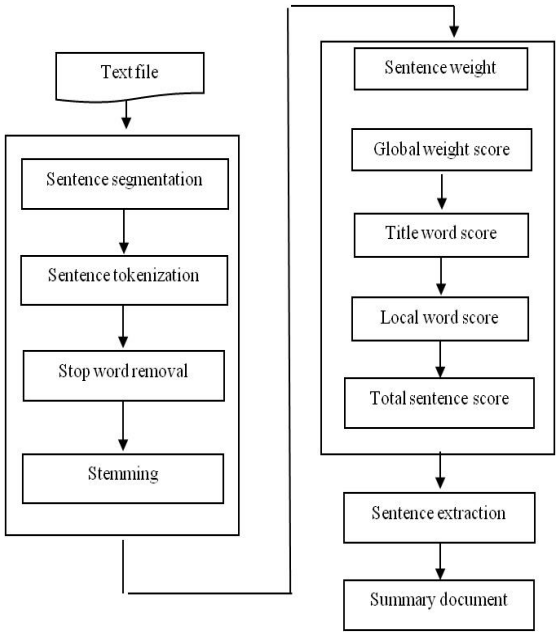

Fig.3. Data Flow Diagram of Modified Weighing Method

d) Sentence-to-Sentence Cohesion:

Calculate similarity between each sentence s and each other sentences of the document and then sum those identical values, acquiring the fibrous value of this feature for s. This process is iterated for all sentences.

$$\text{Sentence weight} = \sum a[i,j] / \sum\sum a[p, q]$$

The proposed algorithm is presented below.

Table 2. Steps for Modified Weighing Method

Algorithm
Step 1: Sentence segmentation is performed.
Step 2: for each sentence do
Title word score(f1)= Σ Tij(Si) /ΣTii
Global keyword score(f2)=no of global keywords present in a sentence
Local keyword score(f3)= no of local key words present in a sentence
Sentence weight(f4)= Σ a[i,j] /ΣΣ a[p,q]
End.
Step 3:for each sentence do
Sentence score= $\frac{(f2*s)+(f3*s)+(f4*s)}{\text{Total no of words in sentence i}}$ + f1

Where s=1 for title words
S=0.9 for global keywords
S=0.8 for local keywords.
End
Step 4: Return sentence score.

IV. Results and Performance Evaluation

Initially selected document is uploaded and the linguistic roles in it are identified. Later the sentence scores for the given document are calculated. Next, extract the sentences of the document based on their sentence scores.

Once the summarized text for the three algorithms is achieved then the precision and recall values are calculated to find the best method.

The performance of the proposed system is evaluated based on available manual summaries as the dataset using the evaluation measures. For experimentation, the summary is generated for different compression rate and is evaluated on the extractive summary provided in the dataset using the evaluation measures.

By comparing the average of precision, recall and F-measure scores of the three algorithms, the best method among the methods is found to be Modified weighing method.

The table 3 presents the values collected while measuring the performance of all the systems.

A. Performance comparison

To test the summarization process, different research documents have been used as input. The purpose was to test the context understanding by the summarizers developed in this work. The table gives the results of three approaches with their average precision, recall and f-measure. Therefore it is observed that Modified Weighing method is the best method among the other two methods.

Summaries	Precision	Recall	f-measure
Modified Weighing method	0.4832	0.2968	0.3678
Top score	0.4625	0.2806	0.3488
Modified Sentence symmetric	0.4439	0.2418	0.3129

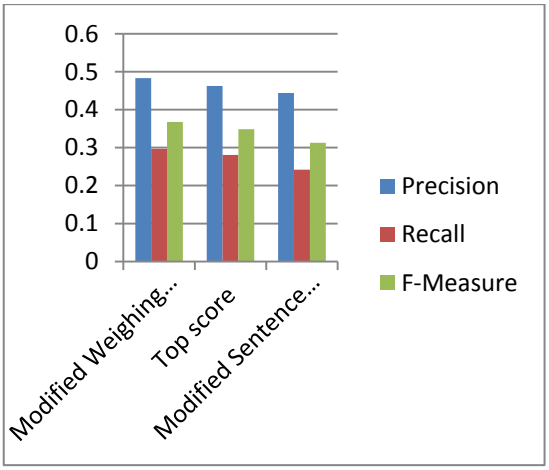

Fig.4. Performance Comparison

Table 3. Measuring the Performance for all Three Methods

SNO	DOC NO	MODIFIED WEIGHING METHOD			TOP SCORE METHOD			MODIFIED SENTENCE SYMMETRIC METHOD		
		PRECISION	RECALL	F-MEASURE	PRECISION	RECALL	F-MEASURE	PRECISION	RECALL	F-MEASURE
1	AS001	0.3444	0.2303	0.2759	0.3333	0.1636	0.3078	0.3358	0.1636	0.342
2	AS002	0.5259	0.2939	0.2937	0.4259	0.2039	0.3492	0.4629	0.2196	0.2809
3	AS003	0.4222	0.3755	0.3973	0.3888	0.3175	0.2142	0.5135	0.2755	0.3275
4	AS004	0.5512	0.3017	0.3878	0.5253	0.2615	0.2652	0.4938	0.2812	0.3125
5	AS005	0.4925	0.3125	0.3765	0.4125	0.3218	0.2256	0.5246	0.2615	0.3185
6	AS006	0.4812	0.2725	0.3598	0.4821	0.3025	0.2025	0.5315	0.2912	0.3001
7	AS007	0.5816	0.2985	0.3927	0.3961	0.2827	0.4014	0.4521	0.2127	0.2812
8	AS008	0.4998	0.2935	0.2861	0.5142	0.2569	0.3252	0.4925	0.2412	0.2912
9	AS009	0.4514	0.3885	0.3411	0.5652	0.3599	0.3851	0.3215	0.2231	0.3215
10	AS010	0.4821	0.3012	0.3712	0.5841	0.3321	0.3951	0.3112	0.2489	0.3101
11	**AVERAGE**	**0.4832**	**0.2968**	**0.3482**	**0.4625**	**0.2806**	**0.3071**	**0.4439**	**0.2418**	**0.3057**

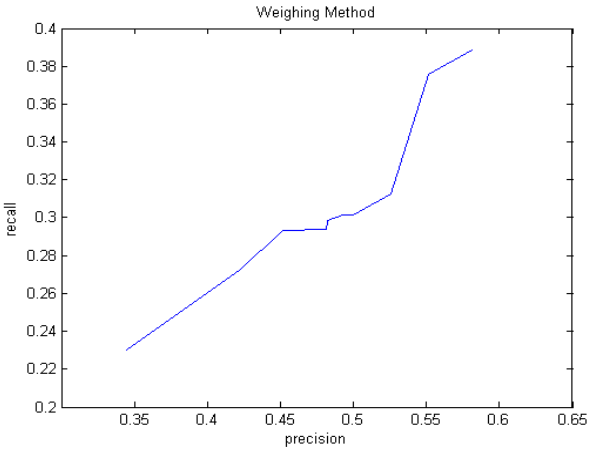

Fig.5. P and R for Modified Weighing Method

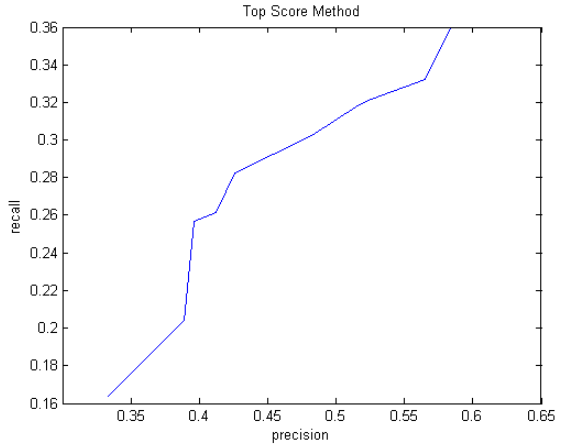

Fig.6. P and R for Top score Method

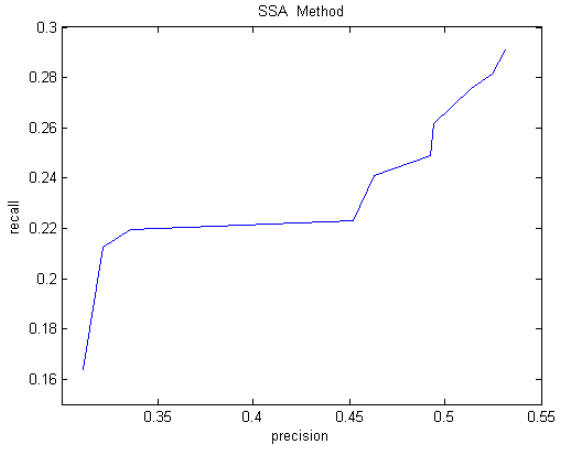

Fig.7. P and R for Modified Sentence Symmetric Method

By comparing Fig. 5 and Fig. 7, Modified Weighing with Modified Sentence Symmetric Method (MSSM) an observation can be made that recall value remained same for an increase in precision for MSSM. Whereas for Modified Weighing Method, the behaviour of recall with precision is linear as should be for a perfect system.

By comparing Figure 6 and Figure 5, Top score method with Modified Weighing method, both are behaving similarly. But for a text summarization system, a system with better precision is preferred. And if both the graphs of Modified Weighing Method and Top score method are observed, for the Top score method the precision dipped for a higher recall but in the Modified Weighing method the increase in precision is consistent.

Hence, an observation can be made that Modified Weighing Method is a better and consistent method. And as the average Precision (P) and Recall(R) numbers are suggesting, the Modified Weighing Method is suitable.

V. Conclusion and Future Work

This paper mainly focused on summarization of research papers. Three different algorithms for summarization are implemented and the performance is observed. Keywords are used for identifying the rhetorical roles in the document. For the calculation of sentence scores and their feature scores for summarizing the text all these three methods are used based on statistical approaches. The work with text data is difficult at times due to vast amount of data to be summarized. While using extractive methodologies sometimes the sentences that are not important to be included in the summary also get included. In the proposed work this limitation was overcome, by using compression ratio to find out the important sentences.

The scope of the paper is maintained to Extractive summarization approaches only. In future, the scope of this work can be extended to abstractive summarization approaches, so that the system can be more efficiently used by all the researchers by giving semantic meanings to the sentences. Also hybrid approaches of extractive and abstractive methods can also be tried.

Acknowledgement

The authors would like to thank the anonymous reviewers for their careful reading of this paper and for their helpful comments.

References

[1] H.P.Luhn “The Automatic Creation of Literature Abstracts”. IBM Journal of Research and Development, 2(92):159 - 165, 1958.

[2] H. P. EDMUNDSON “New Methods in Automatic Extracting”, Journal of the Association for Computing Machinery, Vol. 16, No. 2, April 19691 pp. 264~285.

[3] A.Das, M.Marko, A.Probst, M.A.Portal, C.Gersheson “Neural Net Model For Featured Word Extraction”, 2002.

[4] Jagadeesh J, Prasad Pingali, Vasudeva Varma, “Sentence Extraction Based Single Document Summarization” Workshop on Document Summarization, 19th and 20th March, 2005, IIIT Allahabad.

[5] Arman Kiani B, M. R. Akbarzadeh “Automatic Text Summarization Using: Hybrid Fuzzy GA-GP”, IEEE International Conference on Fuzzy Systems.Juky 16-21, 2006.

[6] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh, A Comprehensive Survey on Text Summarization Systems, IEEE 2009.

[7] Ladda Suanmali, Naomie Salim, Mohammed Salem Binwahlan “Fuzzy Logic Based Method for Improving

Text Summarization", (IJCSIS) International Journal of Computer Science and Information Security, Vol2 No1 2009.

[8] Rasim ALGULIEV, Ramiz ALIGULIYEV "Evolutionary Algorithm for Extractive Text Summarization" Intelligent Information Management 2009, Science Research.

[9] Vishal Gupta, Gurpreet Singh Lehal "A Survey of Text Summarization Extractive Techniques" Journal Of Emerging Technologies In Web Intelligence, VOL. 2, NO. 3, August 2010.

[10] Maryam Kiabod, Mohammad Naderi Dehkordi and Sayed Mehran Sharafi "A Novel Method of Significant Words Identification in Text Summarization", Journal Of Emerging Technologies In Web Intelligence, VOL. 4, NO. 3, August 2012

[11] Masrah Azrifah Azmi Murad, Trevor Martin "Similarity-Based Estimation for Document Summarization using Fuzzy Sets"**,** International Journal of Computer Science and Security, volume 1 issue 4 2006.

[12] Rafeed Al-Hashemi "Text Summarization Extraction System(TSES) Using Extracted Keywords", International Arab Journal e-Technology, vol 1 No 4, June 2010.

[13] Shaidah Jusoh, Hejab M. Alfawareh, Techniques "Applications and Challenging Issue in Text Mining", IJCSI International Journal of Computer Science Issues,vol 9, issue 6,November 2012.

Authors' Profiles

K.Selvani Deepthi is currently Pursuing PhD (CSE) in Gitam Institute of Technology from Gitam University and She is working as Assistant Professor in Anil Neerukonda Institute of Technology and Sciences at Visakhaptnam. Her area of Interest is Natural Language Processing, Text Mining and Data Mining.

Y.Radhika is doctorate in computer science and engineering. She is working as Associate Professor in Gitam Institute of Technology, Gitam University at Visakhapatnam. Her area of Interest is Data Mining.