ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Review Paper on Extractive Text Summarization

[1] Arpita Sahoo, [2] Dr.Ajit Kumar Nayak
[1]M. Tech. Student, Department of Computer Science and Information Technology, Institute of Technical Education & Research, S'O'A University, Bhubaneswar
[2] H.O.D., Department of Computer Science and Information Technology
Institute of Technical Education and Research, S'O'A University, Bhubaneswar India

Abstract: **- Due to plenty of provided detailed fact, figures and data on information server and "information overload" is becoming an issue for people. It has always been a herculean task to summarize and sort mountains of documents manually and a time consuming job to generate a summary keeping all semantics in consideration. Hence, automatic text summarization can be key solution for this problem. Text summarization provides an apparatus for quick understanding the collection of text documents and has plenty of real life applications. This solution of summarization facility will help users to see at a glance what a collection of text document is about and provides a new way of managing a huge accumulation of information. There are many proficiency of doing text summarization, some are extractive and some are abstractive techniques. But we need that approach which will give significant summary without airing any redundancy or any type of ambiguity even if the summary do not contain any fragment of the original document. This paper is provided with few of these approaches which are preferable to obtain more efficient and accurate summary of the original document.**

Keywords**: Information overload; Multi document summarization; Domain independence; Language independence; Natural Language Processing; Extractive Summary; Abstractive Summary; Text summarization.**

I. INTRODUCTION

With the steady progress in the field of technology, the internet's growth has also increased in a tremendous rate. People can find hoards of information easily in various forms like text document, statistics and data. As it's usually known that the internet provides more than required amount of information. In that way more than one problem were recognized searching for important information through a profuse quantity of documents available and absorbing a large amount of relevant information [1]. Previously storage of large data files were challenging and if could replace these large document files with their summaries then we may overcome this downside. To produce a summary of a large text document we need a reader and an identifier to select between unnecessary and prime words/sentences in the text file cluster to generate summary. A summary which states the gist of the document helps in finding relevant information quickly. Document summarization also provides a way to cluster similar document and present a summary [4]. Automatic document summarization is a primary analysis area in natural language processing (NLP). Natural language processing, comes under the field of science, technology and artificial intelligence and machine learning with the interactions between computers and human language. Generation of summaries without the use of NLP may lack semantic and cohesion.

The mechanism of automatic document summarization is expanding and can impart an explanation to the overburden information problem. A summary can be very purposeful in a characteristic way as an indicator to some fragment of the aboriginal document, or in an elucidative process to encrust most of the similar facts and figures and information of the original text document. A proficient summary generator should be capable of reflecting the assorted fragments from the document while considering the redundancy rate at lowest. A simple example of text summarization is Microsoft Word's AutoSummarize function.

Multiple applications, websites and tools are available for summarization like, text compacter, FreeSummarizer, Sumplify, SummarizeTool, WikiSummarizer, Tools4Noobs are few online summarization tools [6]. Examples of few news article summarizers are Microsoft News2, Columbia Newsblaster, and Google1. Few important biomedical summarization tools are AutoSummarize, MEAD, BaseLine, SumBasic, and SWESUM. Few most searched open source summarization tools are Classifier4J, NClassifier, CNGSummarizer, Open Text summarizer.

II. LITERATURE SURVEY

Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive techniques". In this paper author has explained the over-all idea of extractive

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

summarization and all possible extractive summarization challenges in the paper. He have also explained about the features that are used to generate a summary and features that were earlier used and have described most of the extractive summarization techniques to solve the challenges and to obtain summary using these techniques. According to author "the importance of sentences is decided based on statistical and linguistic features of sentences"[2]. N.Moratanch and S.Chitrakala, "A Survey on Extractive Text Summarization". In this paper, author has described the word level features and sentence level features. In this paper author have categorized all extractive summarization methods into unsupervised and supervised methods and have explained each method and have depicted few evaluation metrics [3].

Rajvardhan Oak, "Extractive Techniques for Automatic Document Summarization: A Survey". In this paper author has explained different extractive summarization features and the summarization process structure basically comprises of two steps pre-processing and processing. Author has described different extractive summarization methods and also describes a comparative study of different extractive summarization methods explaining each method's advantages and disadvantages. He have also explained two summarizer tools MEAD and summarist [4].

Selvani Deepthi Kavila and Dr.Radhika Y, " Extractive Text Summarization Using Modified Weighing and Sentence Symmetric Feature". In this paper author has mainly laid emphasis on summarization of different research papers of various fields. In this paper, three distinctive algorithms for summarization are shown and results are detected for each algorithm. Author has perceived that sentence score and feature scores used for the summarization process are determined on the basis of the statistical approaches. In this paper author has overcome few challenges like working with huge amount of data to summaries and including unnecessary sentences in the summary while using extractive methodologies by introducing compression ratio that will help to find out importance of each sentence [22].

Deepali K. Gaikwad and C. Namrata Mahender, " A Review Paper on Text Summarization". In this paper author has described both extractive summarization technique and abstractive summarization technique and have described text summarizers and summarization tools for Indian languages and have exhibited the comparison between performance of different methods [6].

Aysa Siddika Asa, Sumya Akter, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal, "A Comprehensive Survey on Extractive Text Summarization Technique". The up-to-date extractive summarization techniques for different languages have been described in this paper. Here in this paper author have mainly laid emphasis to generate a summarization system for Bengali language using different types of features [7].

III. SYSTEM DESCRIPTION

Text summarization method perhaps categorized into two type extractive text summarization and abstractive text summarization. An extractive summarization method abides selecting essential sentences and phases of important sentences etc. from the unchanged first text document and arranges them in sequence.

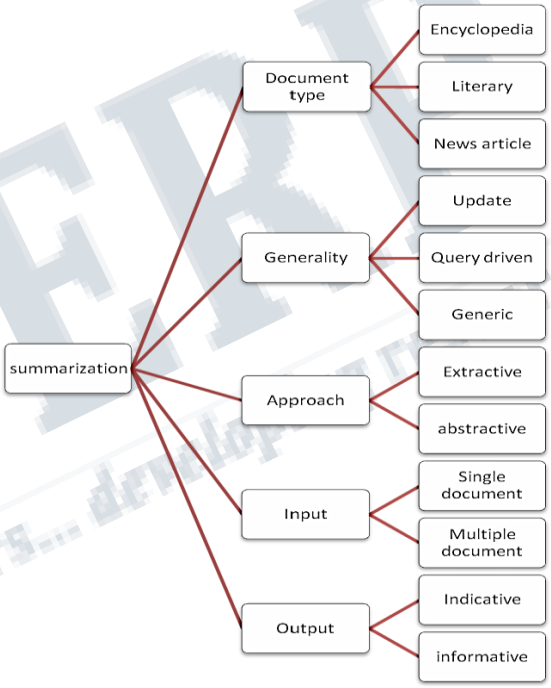

Figure 1. A Summarization Machine

The significance of sentences can be determined through analytical and linguistics attributes of the sentences. Extractive summarization approaches make simpler the difficulty of summarization into the problem of choosing a representative segment of the sentences in the original text document. The extractive summarization is processes based on ability for extracting sentences and extend over the sentences which are more important for the interpretation of the document and focus to cover the phase of sentences that gives the overall understanding of the original document. An abstractive summarization method covers, interpreting the original document along with reciting it using fewer words. It make use of semantic approaches to inspect and explain the original

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

text and to know the up to date concepts and expressions to describe the content of the document by creating a fresh and concise content that emits the maximum infomercial parts from the unchanged raw document. An abstractive text summarization method bids to advance the understanding of the main idea in the document and then communicate those ideas or concepts in a simplified way. In this indicated method, the idea is to create a gist that is similar to what a human being would create/generate. Dissimilar in the extractive ways, the statements are amended based on the linguistic accord in the original text An absolute summarization machine can be designed like the above figure. An indicative summary gives the main concern of the document and contains only a few lines whereas an informative summary is generally lengthy and can be replaced for the original document. The coarseness of the document determines the length of the summary as short, medium, detailed (particular event concerned or an overview) etc.

The summary generated can be categorized into 2 basic types update, query driven and generic. If the generated summary is the output of a query taken a input it can be called as query driven otherwise the generated summary can be called as a generic one. Another type of summary i.e. topic oriented summaries as per the name it specifies the user's interested topic and bring out the information from the document that precise the interested topic of the user. Coming to the next category of summary i.e. generic summary, under this category of summary most of the informative phase from the document is covered keeping the general topical organization of the original text.

Text summarization can be grouped into three basic areas:

- Selection based (tf-idf, ranking, etc.)
- Understanding based (syntactic analysis, semantic analysis)
- Information Extraction / Information Retrieval

The selection method is more preferable than the understanding based as the latter connected to the Natural Language Processing (NLP). The extractive text summarization methods are preferred depending on TF-IDF (Term Frequency-Inverse Document Frequency), cluster based methods, the single value decomposition is preferred by the Latent Semantic Analysis (LSA) and the vector space model is preferred by the concept based summarization. There are many other methods that are deployed based on the graphs, neural networks, fuzzy logic, regression, etc.

The input document provided to the automatic text summarization is categorized into two basic categories:

- Single document summarization
- Multiple document summarization

In single document summarization when a summary is generated it required to predetermine the length of the summary according to the original document (size of the summary as per the size of the document). This is called compression rate. For example, when a document consisting of 10 sentences is compressed by 10%, the summary generated as the result will be a one line summary.

The consisting sentences are scored using the appropriate sentence scoring mechanism then these sentences are arranged in an order on the basis of their scores. Then according to the compression rate the peak sentences are chosen and included in the summary.

The sentence score for a sentence i can determine in the following way:

$$S_i = w_1 * C_i + w_2 * K_i + w_3 * T_i + w_4 * L_i$$

Where,
Si – score of sentence i
Ci – score of sentence i based on cue words
Ki - is the score of the statement i dependent on keywords feature
Ti - is the score of the statement i dependent on title words feature
Li - is the score of the statement dependent on its location or position property
w1, w2, w3 and w4 – are the weights assigned

Or in other word, for summary generation, score of a sentence can be determined as the sum of the frequency of the word in that sentence and their related weightage as per the provided details in the above section.

In multiple documents it is required that the provided documents are related to each other by keeping the main topics as concern. If we are provided with multiple documents belonging to different topics it is possible to generate its summary if few steps are followed. The first step includes text clustering, each sentence of each document avail this approach until clusters of same documents are formed. Once the clusters of the documents are formed, for each cluster respective summaries can be generated.

As the summaries generated from cluster obtained from multiple documents, there are chances to counter few similar sentences are chosen from different document and they are included in the final summary. To create a summary with low inter-sentence similarity the following formula can be used:

$$\text{Cos}(t_i, t_j) = \sum_{h=1}^{k} t_{ih} t_{jh} \Big/ \sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2}$$

Where,

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

i, j – the ith and jth sentences
ti, tj – term frequencies of ith and jth sentences
Based on the similarity measures and compression rate, sentences with higher ranking order and non-overlapping are picked from multiple documents.
But a limitation arises here in this method i.e. the sequence or the order in which these sentences would be displayed. This limitation can be terminated if we note the locations of these selected sentences in its respective document (beginning, center and end) and seek to put each sentence according to their location.

IV. ATTRIBUTES OF EXTRACTIVE SUMMARIZATION

Extractive text summarization features find appropriate sentences and select important sentences from the original document and include these sentences to the summary. Few features are described below that can be applied to select these important sentences:

A. Cue words features:
cue words or clause are the group of words located around vital words like "summary", "reflects", "concludes", "purpose", "because" etc. That specifies the over-all content of the document and can be used as a indicator for the sentence to be included in the summary.

B. Keyword features:
Keywords have the key role in the criteria of selecting important sentences. Sentences that contain most of the keywords are considered to be included in the summary. The keywords can be the verbs, noun, adjectives and adverbs and are determined based on the TF-IDF method. These particular words can also be identified by their acronyms, capitalized or italicized property.

C. Title word feature:
Words contained in the title part are considered to be important words and sentences that contain these are necessary sentences and these sentences are included in the final summary. The sentences that are contained in the original document that has the stated words of the title of the document are contributed to the final summary as it can demonstrate the subject of the document.

D. Location or position word feature:
The location or position of sentences will determine whether the sentence is feasible to be included in the summary or not. Sentences that are placed in the beginning part of the document represents the introduction part of the original document and the sentences that are pointed towards the end part of the document can represent the conclusion part of the summary giving a proper meaning to the document.

E. Sentence position feature:
As most of the documents are hierarchically arranged containing sentences with more important content in the initial and conclusion part of the paragraphs in a document. Hence, sentences situated in the initial and edge parts are more likely to be contributed in the summary.

F. Proper noun feature:
Proper noun a name used for a particular person/ individual which can be stated starting with Capital letter. For example: Jane, Loan and Oxfam. Including the names of persons, places, concepts and organization in the final summary is very important to generate a shorten form of the document keeping the over-all meaning of the document constant. [1]

G. Sentence length feature:
The length of the sentences is a essential feature in selecting sentences which one is to be included in the summary and which one is not to be. Shorter length text not having much words do not carry vital information so as the long length sentences consisting of many words but do not hold up any information are needed to be push aside.

H. Upper case word feature:
The words that contain only upper case or called abbreviation can be reflect on the summary and sentences holding these abbreviations can be included in the summary.

I. Similarity or cohesion feature:
Similarity feature is necessary to remove redundancies and reorder the segments to obtain a coherent summary. Similarity can be calculated among the sentences.

J. Term frequency:
Frequently occurring words rises the sentence score.TF-IDF is used to calculate frequency of each word and the importance of the sentence increases for more number of times the word is visible in the sentence.

V. MOTIVATION

Text summarization is a emerging sub-topic of NLP as short version of any large text document or huge amount of information will be more helpful than going through the whole document to know the purpose of the document. Few inferences for motivation of text summarization are the followings:

- Precise information can be helpful to understand the document more effectively and efficiently.
- News headlines can be more helpful to keep ones update.
- Movie reviews generated can be helpful to people.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

- Summarization can make searching easier.
- Summaries can help people to take decisions in less time.

VI. LIMITATION OF EXTRACTIVE METHOD

- Computers deficient in human intelligence and sometimes fail to understand the language. This makes the automatic/ spontaneous summarization of text very difficult & challenging and difficult.
- deeper analysis of document is necessary.[8]
- Extraction of crucial information or optimization of the whole document in order to minimize the time required to review the document.
- Problem in selection representative subset of sentences from the original text document.[12]
- Generation of summary with minimum redundancy, maximum relevancy and referring elements of document in the summary.
- Contradicting information can not in include in the summary [2].
- Extracted sentences are used to lengthier than other sentences and unnecessary part of the information is also including increasing the over-all size of the summary.

VII. METHODOLOGIES OF EXTRACTIVE SUMMARIZATION

A. Cluster Based Method:

Usually documents are scripted in a way that each section of the document belongs to different topics in an organized manner and the documents are categorized either implicitly or explicitly to obtain a summary. This approach is called clustering method. Sentences can be grouped based on their sentence score [20] which can be calculated using the formula:

$$S_i = w_1 * C_i + w_2 * K_i + w_3 * T_i + w_4 * L_i$$

Where,
Si – score of sentence i
Ci – score of sentence i based on cue words
Ki - is the score of the statement i dependent on keywords feature
Ti - is the score of the statement i dependent on title words feature
Li - is the score of the statement dependent on its location or position property
w1, w2, w3 and w4 – are the weights assigned

Other factors that can make this sentence selection process more particular are the follows:
- Consider a theme of each cluster and determine sameness of each sentence by comparing them with the cluster.
- Position of each sentence in the text document

B. Machine learning method:

For a set of training documents including its extractive summaries is provided as input in the training stage. Here in machine learning [5] approach the summarization method is basically a classification problem where the sentences are categorized into two, either summary sentence or non-summary sentence focused on the properties which each sentence holds. The sorting method is carried out statistically from the provided set of training document using Bayes' rule:

P (s∈<S | F1, F2,..., FN) = P (F1, F2, ..., FN | s∈S) *P (s ∈S) / P (F1, F2,..., FN)

Where,
s = the sentence from the document collection
F1, F2.. up to.. FN = features or properties used in classification/ sorting.
S = is the summary to be generated
p (s E< S | F1, F2, .. up to .. FN) = is the extent to which an event is likely to occur that statement s shall be selected from the gist given that it possess properties F1, F2, .. up to.. FN

C. Query Based method:

Here in this query based approach, the sentences score is determined using the above formula or using the term frequency feature but sentences that contain the query phrases are considered with greater score as compared to the terms that contains solitary query terms that contains single query word [23]. Statements with the greatest score are included in the gist. Sentences are extracted from the original document using query based approach using the following algorithm:

1. Sentences are arranged as per their score
2. Add sentences from title or topic of the document .
3. First level -1 heading is added to the gist.
4. Whereas (the summary size limit does not exceed)
5. The following greatest scored statement is added.
6. The structural context of the statement is added (if any & further not already added in the statement)
7. The greatest level heading above the extracted text is added (let's take this heading as 'h').
8. The heading before 'h' is added in the same level.
9. The heading after 'h' is added in the same level.
10. Steps 7,8 & 9 are repeated for the next greatest level headings.
11. While loop is ended.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Other way of query based approach [21] is by considering a whole text document as a interconnected text elements and consider the query keywords because searching a keyword has been the most famous data extraction process. This method can be used by following few steps, at first the preprocessing step, it determines the structure of the document, which would seems like a characterized and complete graph called document graph. Then next step is keyword proximity search in the document graph to determine the association of these keywords with the document graph. The summary of each document is the least extending tree of the respective document graph which consolidates entire associated keywords.

D. Graph Theoretic Method:

Graph theoretic process is used to find out the idea of the paragraph. Each statements of the original record are considered as junction of an undirected graph after the preprocessing step is performed in the original document. Sentences that contain a common word then these nodes are interconnected through common edge. The edges with higher cardinality are selected for generation of final summary [7].

E. Fuzzy Logic Method:

In fuzzy logic approach each element possesses a degree that means each element constitute of another element. In this method feature of each sentence like sentence length, similarity to title, similarity to other sentences etc. as input to the system. Then the necessary rules are entered required for summarization. A threshold value is generated for each and every statement as the output/harvest on the basis of statement properties and all possible guidelines. The output value determines the importance of each sentence that is included in the final summary. Components of fuzzy logic system are:

(1) Fuzzifier: input data are converted into linguistic principles which use a membership function to be use for input linguistic variables.

(2) Inference Engine: the interface engine cites to the principles that are being used for the judgment making process.

(3) Knowledge Base: it has a set of principles that are being used for the decision making process.

(3) Defuzzifier: it executes activity adverse to that of the fuzzifier. The linguistic variables from the conjecture are being changed to the concluding crisp values by the defuzzifier by the help of membership function for giving the concluding score.

In this method sentences are considered as per their sentence score and all sentences are arranged in a descending order with respect to their sentence score. Top statements present in the order are selected & also comprised in the final gist on the basis of the compression rate.

VIII. CONCLUSION

As plenty of information is available on World Wide Web and it is not possible to go through each document available to know the purpose of the document and to know if it is a necessary document or not. Hence, a summary of these document will be more helpful to reader to decide if the available document is relevant or not and extraction of gist of each document will be easier. Although extractive text summarization is easier to implement but it also holds few limitations causing ambiguity and miscommunication is the summary. Abstractive summarization can generate more relevant and precise summary but more complex heuristic algorithms is required. The summarization methods should generate summaries with more accuracy in less time with minimum redundancy. Summary evaluation can be done using either extrinsic or intrinsic method. Intrinsic method measure quality of the summary using human evaluation method while an extrinsic method measures the summaries using task based performance measures.

REFERENCES

[1]Reeve Lawrence H., Han Hyoil, Nagori Saya V., Yang Jonathan C., Schwimmer Tamara A., Brooks Ari D., “Concept Frequency Distribution in Biomedical Text Summarization”, ACM 15th Conference on Information and Knowledge Management (CIKM), Arlington, VA, USA,2006.

[2]Vishal Gupta and Gurpreet Singh Lehal, “A Survey of Text Summarization Extractive Techniques”, Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010.

[3] N.Moratanch and S.Chitrakala, "A Survey on Extractive Text Summarization", IEEE International Conference on Computer, Communication, and Signal Processing, 2017.

[4] Rajvardhan Oak, “Extractive Techniques for Automatic Document Summarization: A Survey”. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 3, March 2016.

[5] J. L. Neto, A. A. Freitas, and C. A. Kaestner, "Automatic text summarization using a machine learning

approach," in Advances in Artificial Intelligence. Springer, 2002, pp. 205-215.

[6] Saranyamol C S, Sindhu L, "A Survey on Automatic Text Summarization", International Journal of Computer Science and Information Technologies, 2014,Vol. 5 Issue 6.

[7] Deepali K. Gaikwad and C. Namrata Mahender, " A Review Paper on Text Summarization", International Journal of Advanced Research in Computer and Communication Engineering(IJARCCE) Vol. 5, Issue 3, March 2016.

[8] Aysa Siddika Asa, Sumya Akter, Md. Palash Uddin, Md. Delowar Hossain, Shikhor Kumer Roy, Masud Ibn Afjal, "A Comprehensive Survey on Extractive Text Summarization Technique", American Journal of Engineering Research (AJER), Volume-6, Issue-1, 2017.

[9] Atif Khan, Naomie Salim, "A Review On Abstractive Summarization Methods", Journal of Theoretical and Applied Information Technology, Vol. 59, No.1,10th January 2014.

[10] Deshpande Anjali R., Lobo L. M. R. J., "Text Summarization using Clustering Technique", International Journal of Engineering Trends and Technology (IJETT) , 2013, Vol. 4 Issue8.

[11] Karmakar Surajit, Lad Tanvi, Chothani Hiten, "A Review Paper on Extractive Techniques of Text Summarization", International Research Journal of Computer Science (IRJCS), Issue 1, 2015, Vol. 2.

[12] S.Mohamed Saleem, R.Krithiga, S.K.Rani, S.Celin Sindhya, "study on text summarization using extractive methods", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 5, May 2015.

[13] Karel Jezek and Josef Steinberger, "Automatic Text summarization", Vaclav Snasel (Ed.): Znalosti 2008, pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[14] A. Kogilavani, P. Balasubramani, "Clustering and Feature specific sentence extraction based summarization of multi-documents", International Journal of Computer Science & Information Technology (IJCSIT), 2(4), 2010.

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

[16] Richa Sharma, Prachi Sharma, "A Survey on Extractive Text Summarization", International Journal of Advanced Research in Computer Science and Software Engineering(IJARCSS), Volume 6, Issue 4, April 2016.

[17] Vishal Gupta and G.S Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, pp. 60-76, August 2009.

[18] N. R. Kasture, Neha Yargal, Nityanand Singh, Neha Kulkarni, Vijay Mathur, "A Survey Methods of Abstractive Text Summarization", International Journal for Research in Emerging Science and Technology, Vol. 1, Issue 6, November 2014.

[19] M. Haque et al. "Literature Review of Automatic Multiple Documents Text Summarization", International Journal of Innovation and Applied Studies, Vol. 3, pp. 121-129, 2013.

[20] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", TECHNIA – International Journal of Computing Science and Communication Technologies, vol. 2, no. 1, Jul. 2009.

[21] Ramakrishna Varadarajan and Vagelis Hristidis, "Structure-Based Query-Specific Document Summarization", in proceedings of CIKM'05, ACM, Bremen, Germany, 2005.

[22] Selvani Deepthi Kavila and Dr.Radhika Y, " Extractive Text Summarization Using Modified Weighing and Sentence Symmetric Feature", I.J. Modern Education and Computer Science, October 2015.

[23] Mohamed Ahmed A., Rajasekaran Sanguthevar, "Query-Based Summarization Based on Document Graphs", Document Understanding Conferences, NIST, 2006.