Text Summarization Using Semantic Analysis of Frequent Terms

Jyoti Rohilla
Computer Department
Echelon Institute of Technology
Faridabad, India
E-mail: Jyotirohilla689@gmail.com

Usha Yadav
School of Information Technology
CDAC, Noida, India
E-mail: usha.yadav.912@gmail.com

Abstract: *Due to growing amount of data and comparatively less amount of information on web, it becomes necessary to introduce a mechanism that can easily search out relevant information from that bulk of data. This direction approaches to the concept of text summarization where the whole document is condensed to a smaller version retaining its original meaning. There are several methods of extractive and abstractive summarization but this paper will focus on the specialized extractive summarization named frequent term summarization by considering the semantic similarity of its words.The primary purpose of using the combination of these two techniques is to remove the limitations of these two extractive summarization processes and to use their best feature to serve the purpose.*

General Terms *– Your general terms must be any term which can be used for general classification of the submitted material such as Pattern Recognition, Security, Algorithms et. al.*

Keywords *– abstractive summary,extractive summary, Generic summary, Indicative and descriptive Summary, Text Summarization, semantic similarity.*

I. INTRODUCTION

A huge amount of data is available over internet. A large amount of data is uploaded over internet every day, which causes the availability of bulk data here. That means we have a large amount of data that will successfully match our search. Also we cannot forget the fact that a large amount of data is also present there which is not suitable for our search. In that case searching out a relevant data that meet our requirements is a tedious and time consuming task. We can face two kinds of problems

1) Searching a relevant document corresponding to our search.

2) Absorbing maximum amount of information from that bulk data source.

Here we are working for such a technique that will resolve both of these problems i.e. time and information.

This task can be sorted out by summarizing the big document into smaller one, called summary, by extracting out the important data from the document. Here we are working on the frequency based summarization technique along with it we will consider its semantic feature as well for summarizing the document. The main agenda behind summarizing the document is to reduce the time consumption and searching a document corresponding to its relevancy. It must be taken care of that the summary must consists of all the necessary details of the parent document and the length of summary must be less than the original document. In the previous methodology used for this particular task it was felt that few of the important sentences were excluded from the summary just because of the fact that their frequency does not satisfy the threshold value of sentence score because of usage of different phrases used to represent the same fact. The proposed technique will remove this problem up to a certain extent by considering the semantically similar sentences equally important as their original ones are treated as. The idea behind choosing this technique is that it will select the sentences for summary generation on the basis of the frequency of words contained in it

II. TYPE OF SUMMARY

There are different types of summarization approaches depending on what the summarization method focuses on to make the summary of the text, following are few basic criteria's on the basis of which summary can be categorized [9].

A. On the Basis of Input Document

A summarization system can accept one or more documents as input. Single document summary provide the most relevant information contained in single document that helps the user in deciding whether the document is relevant for his search or not. An example of single document summarizer is Summarist, it is an attempt to create a robust automated text summarization system, based on the 'equation': summarization = topic identification + interpretation + generation. Each of these stages contains several independent modules, many of them trained on large corpora of text[11]. whereas Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, to quickly familiarize themselves with information contained in a large cluster of documents. An example of multi-document summary system is SUMMONS which is designed in Columbia University [9].

B. On the basis of Input Approach

Based on the different approaches of analyzing the texts and generation of the summary, text Summarization systems are divided to *extract* and *abstract* systems. An Extractive summarization system consist of selection of important sentence, paragraph , etc. from the source document and concatenate them to form a summary. The importance of sentences is decided on the basis of statistical features of sentences, whereas the abstractive summarization technique is based on the understanding of whole text and re-phrasing it in fewer sentences. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document[1]. One example of a system which use extract summary is Summit applet [9] which is designed by Surrey University.

C. On the basis of Details

This category is based on the details which are important in the desired summary. Two different groups could be considered: *Indicative* and *Informative*.

An indicative summary gives idea about the contents of the article without giving away detail on the article content. Card catalog entries and movie trailers are examples of indicative summaries, whereas Descriptive summary is meant to represent (and often replace) the original document. Therefore it must contain all the information necessary to be conveyed.

One example for detailed based summarization system is SumUM [9], a text summarization system that takes a raw technical text as input and produces an indicative informative summary.

D. On the Basis of Content

Another classification, which is based on the importance of the content in the original text, is *generic* versus *query specific* summaries.

Generic summary do not target to any particular group of readers. It addresses a wide group of readers while Query or topic focused queries are tailor-made for specific needs of an individual or a particular group and represent particular topic. An example of generic summarizer system is SUMMARIST [11] which produces summaries of web documents.

After a brief description of type of summary following table is made for convenience.

Table 1: Table Type of Summaries

Criteria	Type of summary	System used to implement
Input documents	Single Document	Summarist
	Multi Document	Summons
Approach	Extractive	SummIt Applet
	Abstractive	Summarist
Details	Indicative	SumUM
	Informative	SumUM
contents	Generic	Summarist
	Query Specific	Mitre'sWebSumm

III. PROPOSED METHOD

The proposed method summarizes the bulk of text data to a smaller version and keeps the original meaning by collaborating the two different approaches i.e. frequency based summarization technique and the

semantic method for summary extraction for taking out the summary. It includes various modules. All the functions to be performed are mentioned in sequential order. The first step of proposed method involve preprocessing, which further includes sentence separation by punctuation marks like comma, semi colon, full stop etc., word separation by blank spaces , and stop word elimination like helping verbs , articles etc. Further stemming of the remaining document is done where all the derivations of words are converted to its root word. As the title says the frequency of each word is calculated by considering semantically similar words as well. Weights are assigned to each word on the basis of probability given to each word on the basis of formula mentioned below in algorithm. Then the sentecnce containing maximum number of words with higher frequency is choosen to give the maximum weight and the same procedure is followed with rest of the sentences. Ranking of each sentence is done using the probability assigned to each word and further the weight of sentence which will be done by following the appropriate algorithm and all the sentences belonging to those top ranked words are considered in final summary and if the length of summary do not match the requirement of user then the above mentioned steps are repeated as per the requirement.

A. Algorithm

The algorithm for proposed method is as follow.

Algorithm: Text Summarization Using Semantic Analysis of Frequent Term

Input - ***I.*** Text Data for which Summary is required.

Output - ***O.*** Summary for the Original Text Data.

Steps:

Step 1 Accept input document.

Step 2 Preprocessing of the document is done in following manner.

a) Separate all the sentences on the basis of punctuation marks and record in array Sj.

b) Separate all the words on the basis of space, punctuations etc.

c) Eliminate all the stop words like is, are , the etc.

Step 3 All the derivations of the words are replaced by its root word, for example all the words like computerized, computerization etc are replaced by its root word Computer, this process is called stemming.

Step 4 Calculate the probability of each word(w_i) occurring in the input document by the following formula

$$p(W_i) = \frac{n}{N}$$

Where n denotes for frequency of particular word

N denotes total number of words

Step 5 Replace semantically similar words by their root words and update frequencies.

Step 6 Assign Weight to each sentence.

Weight assigned to each sentence is equal to the average of the probability of all the words assigned in step 1.

i.e. $weight(S_j) = \sum_{w_i} \frac{p(W_i)}{|\{W_i : W_i \in S_j\}|}$

Step 7 Pick out the sentences in descending order of their probabilities.

If desired summary length is not reached, repeat the above procedure.

B. Architecture

Above mentioned algorithm will remove the limitations of purely frequent term summarization as the frequency of semantically similar words will also be counted in the root word which will increase the frequency of root word and deserving importance will be given to that word. No separate frequency is counted for those words. The flow chart of for the specified method is as follows.

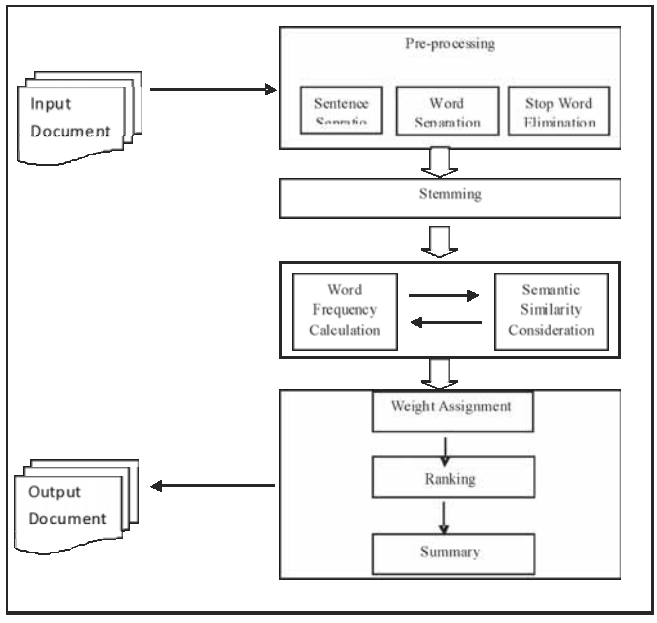

Fig 1: Architecture for Proposed algorithm

IV. COMPARISION WITH EXISTING TECHNIQUES

Here is a comparison of some of the well known online available summarizers. All these summarizers uses the technique of term frequency – inverse document frequency without even thinking about the semantic similarity of the words present in the text which contains the same meaning. The proposed algorithm finds out the summary based on the frequency of terms in the document and also pays attention on the semantic similarity of the words present in the text that contains the same meaning.

Given below is a comparison table of data taken out by analysis of various well known techniques and the proposed method gives the result which is best suitable, neither so much nor so less.

Table 2: Comparison of Various Techniques

Summarizer	No of words in test data	No of words in summary	Percentage
Free	169	46	28%
Auto	169	28	17%
Tools4knob	169	156	92%
Text compactor	169	138	81%
Proposed Method	169	69	41%

The result of free available online tool's have been compared with the proposed algorithm and the result is found to be more efficient then the previously used method which is based on frequency of the terms only. The comparison graph of the free online summarizer and the proposed method is as follows:

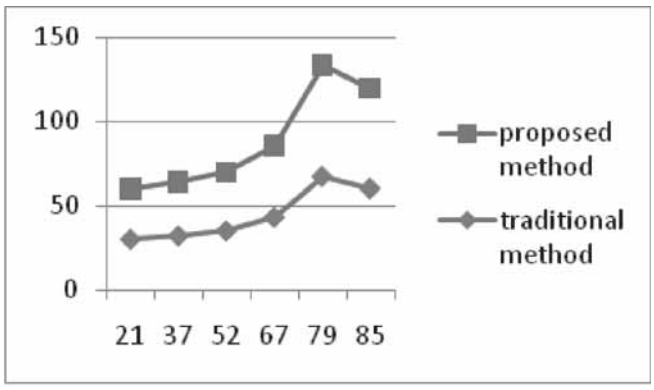

Fig. 2 : Comparison Graph of Different Methods

V. CONCLUSION AND FUTURE SCOPE

In this paper an algorithm has been introduced for single document frequency based text summarizers which consider semantic similarity as well. The result shown by this technique is found to be more efficient than the previously used technique which considers the frequency of text only. The implementation of above mentioned technique is done on open source platform. Looking forward to make the technique more user specified by introducing the concept of query in it as well, where the query keywords would play an important role in selection of sentences for summary. The sentences having the same keywords as that of query would be treated as more important and have more chances of being there in summary.Also the summarizer can be proved to be a helping hand in making the online available data structured. The words chosen by summarizer for summary can be used for ranking the page and avail the page as per the rank.

REFERENCES

[1] Vishal Gupta ,Gurpreet Singh Lehal, August 2010 A Survey of Text Summarization Extractive Techniques. Journal Of merging Technologies In Web Intelligence, Vol. 2, No. 3,.

[2] ArchanaAB,Sunitha 2013 An overview on document summarization technique. International Journal on Advanced Computer Theory and Engineering (IJACTE), Volume-1, Issue-2.

[3] Xinghuo Ye, Hai Wei 2008 Query-Based Summarization for Search Lists. 0-7695-3090-7/08 $25.00 © IEEE DOI 10.1109/WKDD.2008.14

[4] Madhavi K. Ganapathiraju 2002 Overview of summarization methods. 11-742: Self-paced lab in Information Retrieval.

[5] Naresh Kumar Nagwani, ShirishVerma 2011 A Frequent Term and Semantic Similarity based Single Document Text Summarization Algorithm.International Journal of Computer Applications (0975 – 8887) Volume 17– No.2.

[6] A.P. Siva Kumar, Dr. P. Premchand, Dr. A Govardhan 2011 Query-Based Summarizer Based on Similarity of Sentences and Word Frequency. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3.

[7] Vishal Gupta, G. S. Lehal 2009 A survey of Text mining techniques and application. Journal of emerging technologies in web intelligence, VOL.1 NO. 1, 60-76.

[8] A. Kogilavani, Dr. P.Balasubramani 2010 clustering and feature specific sentence Extraction based summarization of multiple documents. International journal of computer science & information Technology (IJCSIT) Vol.2, No.4.

[9] SaeedehGholamrezazadeh, Mohsen AminiSalehi, BaharehGholamzadeh 2009A Comprehensive Survey on Text Summarization Systems. 978-1-4244-4946-0/09/ IEEE.

[10] Ramakrishna Varadarajan, VagelisHristidis 2006A System for Query-Specific Document Summarization. *CIKM'06*, Arlington, Virginia, USA, Copyright 2006 ACM 1-59593-433-2/06/0011.

[11] Hovy, E., & Lin, C. 1999 Automated text summarization in SUMMARIST. *In I. Mani & M. T. Maybury (Eds.)*, Advances in Automatic Text Summarization (pp. 81-94). Cambridge, MA: MIT Press.

[12] Verma, R., Chen, P., and Lu, W. 2007 'A Semantic Free-Text Summarization Systems Using Ontology Knowledge. Document Understanding Conference DUC 2007, pp. 1-5".

[13] Rafeeq Al-Hashemi,, 2010, "Text Summarization ExtractionvSystem (TSES) Using Extracted Keywords", InternationalvArab Journal of e-Technology, Vol. 1, No. 4, June, pp. 164-v168.

❑