Frequent Term Summarization Using Word Probability Based Semantic Similarity

Manoj Kumar

Department of Computer Applications,
Galgotias University Gr Noida(U.P.)

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 mechanism is called text summarization. The whole document is condensed to a smaller version retaining its original meaning. This can be done by two approaches, first is extractive technique and second is abstractive technique of summarization. In extractive summarization technique the important words, phrases, sentences are taken out to make a summary, while abstractive summarization technique is based on the understanding of whole text and then the summary is created by sentence synthesis. There are several methods of extractive summarization but this paper will focus on the Frequent term summarization by considering the semantic similarity of its words as well. The primary purpose of using the combination of two techniques is to remove the limitations of them and to use their best feature to serve the purpose.

Index Terms—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. Therefore, the large amount of data that will be successfully explored for word similarity. Also it cannot forgotten the fact that a large amount of data is also present there but it 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. Two kinds of problems are being faced:

1) Searching a relevant document corresponding to our search.
2) Absorbing maximum amount of information from that bulk data source.

Hence, there is strong need of 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. In this study, the frequency based summarization technique is being explored along with similarity semantics 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 documents:
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 whereas Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. The particular user will speedily aware the information pertaining in big cluster of documents. An example of multi-document summary system is SUMMONS which is designed in Columbia University [9].

B. On the basis of 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. 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], The technical text is the input to proposed system to fabricate the 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 [9] which produces summaries of web documents.

III. ANALYSIS OF VARIOUS EXTRACTIVE SUMMARIZATION TECHNIQUES

On the basis of brief study of various extractive summarization techniques author identified the some limitations about the technique used so far and also the problems that they are found to be suitable for. Here is a brief discussion of those facts about the techniques.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Type</th>
<th>Limitation</th>
<th>Problem Suitable For</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tf-idf</td>
<td>Query Specific</td>
<td>Semantic Similarity not considered.</td>
<td>Decide relevance of document for user’s query</td>
</tr>
<tr>
<td></td>
<td>Generic</td>
<td>0/1 relation is maintained.</td>
<td></td>
</tr>
<tr>
<td>Cluster Based</td>
<td>Query Specific</td>
<td>Summary size may be too small or too large.</td>
<td>Where various topics are discussed in text</td>
</tr>
<tr>
<td>Graph Based</td>
<td>Query Specific</td>
<td>Summary size may be too small or too large.</td>
<td>Mention all topics discussed in text</td>
</tr>
<tr>
<td>Neural Network Based</td>
<td>Generic</td>
<td>No guarantee of same output with same text as input. Improper training may affect results.</td>
<td>Can handle noisy data. Easier to modify the selection criteria to fit to new user’s community.</td>
</tr>
</tbody>
</table>
IV. PROPOSED PROBABILITY BASED SEMANTIC SIMILARITY APPROACH FOR FREQUENT TERM SUMMARIZATION

To remove the above mentioned limitation in term frequency-inverse document frequency technique the flow chart of a proposed method is represented in figure. All the steps to be performed are mentioned in sequential order. The very 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 sentence containing maximum number of words with higher frequency is chosen to give the maximum weight and the same procedure is followed to do the same with rest all 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. The algorithm for proposed method is as follow.

Algorithm: Frequent term summarization with semantic considerations.

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 $S_j$.
   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 and $N$ 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.
   \[ weight(S_j) = \sum_{w_i} p(w_i) \]
Step 7 Pick out the sentences in descending order of their probabilities.
Step 8 If desired summary length is not reached, repeat the above procedure.
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:

V. CONCLUSION AND FUTURE SCOPE

In this paper, the author has introduced an algorithm for single document frequency based text summarizers which consider semantic similarity as well. This technique is found to be more efficient than other existing technique which considers the frequency of text only. The implementation of the proposed approach will be done in future using open source technologies. 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.

REFERENCES


