International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

An Extractive Summarization Technique for Text Documents

Ashwitha Dantis, Roshan Fernandes, Anisha P Rodrigues

Abstract*: In order to read as well as search information quickly, there was a need to reduce the size of the documents without any changes to its content. Therefore, in order to solve this problem, there was a solution to it by introducing a technique called as automatic text summarization which is used to generate summaries from the input document by condensing large sized input documents into smaller documents without losing its meaning as well as relevancy with respect to the original document. Text summarization stands for shortening of text into accurate, meaningful sentences. The paper shows an implementation of summarization of the original document by scoring the sentence based on term frequency and inverse document frequency matrix. The entire record was compressed so that only the relevant sentences in the document were retained. This technique can be applicable in various applications like automating text documents, quicker understanding of documents because of summarization.

Index Terms: Term frequency, Inverse document frequency, Sentence ranking, Sentence score, Pos-tagging

I. INTRODUCTION

There is tremendous amount of textual data generated with the emergence of internet and mass media. It is not possible for a human to handle these large volumes of textual data and there is a need to find certain automatic methods in order to handle it. All these problems fall into the area of Natural Language Processing. We focus on technique called as automatic text summarization where we extract information from the input document and compress it into smaller documents without changing its meaning as well as its relevancy with respect to the original document. There are dual techniques for summarizing of text namely abstractive and extractive technique for text summarization. An approach for extractive text summarization involves extracting important details from the documents and generating the summary in the identical order as that of the initial document. The different techniques of extractive summarization technique are based on term frequency and inverse document frequency, clustering the sentences, graph based method, machine learning approach to classify the sentences as relevant or not, latent semantic analysis by grouping the words that are semantically similar to each other even if they don't share common words, neural networks to summarize the document, query based technique based on vector space model which counts the frequency of terms, fuzzy based logic to summarize the sentence where the output is dependent on the importance of the sentence.

Abstractive approach for text summarization involves understanding the original text document and retelling it in the form of summary. The techniques used in this approach are tree-based method makes use of a dependency tree to represent the document, template to represent the document, ontology-based method to represent the document in the form of knowledge base, rule-based method represented as categories and a list containing aspects, semantic model consisting of multimodal documents, method based on information item, semantic based method for graphs.

II. LITERATURE REVIEW

[1] This research article proposed a technique which was efficient for summarization of text in a document as a service. The proposed system included three modules such as an input module which took the text to be summarized, summarization module which included pre-processor to eliminate the stop words, frequency term generator in order to generate the frequency of words and sentence filtering module so that the sentence is summarized, finally an output module to get the summarized document. It also showed that in order to generate a summary of the document it considered the compression ratio which was set to one-third of the original text. The summary was stored in the database so that the user is able to make some analysis on it. A web service was set up in order to generate the summary by the server when a request to it is made by the client by sending a request message which consisted of the document to be summarized. [2] This article proposed a technique in order to improve the quality of summarization a feature called as pronoun replacement was introduced. The technique included consideration of the word frequency by replacing the pronouns with the proper nouns which are previous. This approach of replacement the pronouns with the nouns and after that considering frequency of words led to generate summaries which had an improved quality. This technique is been applicable to various text applications such as articles of the newspaper, documents of text which are large, internet-based information and many more applications.

Revised Manuscript Received on August 30, 2019.
* Correspondence Author
Ashwitha Dantis*, Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte, India.
Roshan Fernandes, Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte, India.
Anisha P Rodrigues, Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte, India.

[3] This article proposed a technique which is called as query oriented text summarization technique in order to extract the important sentences and then generate a summary. The proposed system consisted of five steps in order to generate a summary. The five steps are preparing data, pre-processing of text, extraction of features, scoring of the sentences, summary generation. Text pre-processing includes tokenization, removal of stop words, stemming operation, pos tagging. There are eleven features which are extracted in this system to generate summaries they are feature of the document, position of sentence, normalized sentence length, numerical data, proper noun, frequency such as topic, topic token, headline, start cluster, skip bi-gram, cluster. This paper shows that large number of features are extracted in order to generate summarizes which improves the quality of summarization. [4] This research article proposed a technique for classification and summarization of text based on information arrangement. This technique includes hierarchically arranging and classifying the sentences. In order to understand the contents of the sentences the Concept Base, the Degree of Association Algorithm, the Time Judgment system and the Place judgment system are proposed by the technique. Concept Base is used to generate semantics from a certain word whereas the Degree of Association Algorithm makes use of the results of expansion of semantics so as to express a relationship in between a word and another one as a numerical value. This paper showed the concept of hierarchically arranging and classifying sentences of an article of a newspaper. [5] This article proposed a technique for summarization of document based on clustering of sentences. This approach consists of generating summaries by considering the similarity measure between the sentences, to estimate the number of clusters in the sentence, sentence clustering, topic sentence extraction. The similarity measure between the sentences are of the following word form similarity, word order similarity, word sematic similarity, sentence similarity. After estimating the number of clusters in the sentence the sentence clustering is performed using k means method of clustering which takes number of clusters and the sentences to be clustered as the input. [6] This article proposed a technique for categorization and summarization of the document on the basis of rule reduction algorithm. The approach included the following steps creation of tokens, identification of feature and categorization and Summarization. This technique derives the structure of the original text making use of a test analyser. The tokens are created by splitting the input text into alphabets, white spaces and punctuation marks. Once the tokens are created the features of the alphabets are identified such as preposition, noun, verb, adjective etc. In the categorization process there is categorization of alphabet token as Noun phrase, Prepositional phrase, Verb phrase and then summarize them in order to formulate a sentence. This approach can be used for a variety of applications such as indexing for retrieval of document, organize and maintain catalogues or web resources, extracting metadata automatically etc. 7] This research article proposed an enhanced summarization approach for Bengali texts. The proposed system included pre-processing, scoring process, synchronized summary generation. The pre-processing included removal of noise, tokenization, stemming, removal of stop words. Noise removal includes removal of headers, footers present in the text document. Separating each word into its lexical form refers to tokenization. In order to calculate the score of the words we use a term frequency and inverse document frequency approach. In order to calculate the score of the entire sentence we use a K means clustering algorithm is used. This approach provides a synchronization of the final summary generated and the original text by extracting top k-sentences from individual cluster containing sentences thereby sorting the above sentences with their appearance in the initial text. [8] This research article proposed a technique for text summarization with the help of a model called as CoRank model. CoRank model is a word sentence- based model. In order to enhance the quality of text summarization it is supplemented with the help of redundancy elimination technique. Currently this approach has been extended to summarization of multiple documents, query oriented or event-driven summarizations. [9] The article proposed an unsupervised text summarization technique which works for any domain. It looks to exploit concept diversity in text summarization. This technique is a diversity-based approach in addition it to an information centric approach has been proposed for evaluation, so as to judge the quality of the summaries not on how they match to the ones created by humans rather to how well the source documents are represented in the categorising as well as retrieval of documents in the text . It takes into account redundancy and diversity as concepts of the text. It includes finding diversity and then reduce redundancy and finally generating summary. [10] This research article proposed an automatic summarization approach for multi-document called as multi-document rhetorical structure. It is a discourse -based approach for text summarization. In this approach hierarchical topic tree was used to calculate the interrelationship between text units, which also included correlation between them. A series of algorithms which includes building of multi-document rhetorical structure, summary generation are proposed. [11] The authors have analysed the twitter data on various products. There is no summarization achieved, but the tweets have been analysed and categorized into positive, negative ad neutral tweets.[12] The data was collected in the mobile devices, as a part of service request. Then this data was analysed using rule based and supervised machine learning techniques. The data was not summarized instead the work was concentrated on retrieving the service name from the English sentence.

III. IMPLEMENTATION

The approach used for the summarization of text is according to term frequency and inverse document frequency. The steps involved in the algorithm are as follows

Algorithm for summarization:

1. Tokenize the sentences of the input document.
2. Compute the frequency of individual word present in the sentence and create a matrix.
3. Compute the term frequency matrix to calculate the occurrence of a particular word in the document. Count the document for each word.
4. Calculate the inverse document frequency matrix to compute the uniqueness or how rare a word is in a document.
5. Generate a matrix based on both term frequency and inverse document frequency.
6. Scoring the sentence based on the computed term frequency and inverse document frequency.
7. Finding threshold is in accordance with average score of the sentences present in the document.
8. Summarize the document based on the score of the sentences and the average score of the sentences in the document.

The four stages involved in the extraction of the most relevant sentences in the text are

1. Pre-processing: The pre-processing of the input document to be summarized involves segmentation of the sentence, tokenization of the sentence, removal of stop words, performing stemming of the words. Segmentation of the sentence is a procedure of breaking down the entire input document into sentences by identifying the features of the sentence like full-stop, question mark, exclamation marks, as well as identifying the total amount of sentences present in the whole document. Tokenization of the sentence is the breaking down of sentences into words by identifying the special symbols which are present between the words. The stop word elimination refers to the removal of the words which do not carry a lot of meaning to the final summary to be generated. Stemming operation of the words refers to getting the root word and the system used here makes use of a stemmer called as Porter stemmer to get the root word of the given word.
2. Scoring of the sentences: The scoring of the sentence in this approach makes use of term frequency and inverse document frequency approach. This approach is based on statistics which gives the relevance of a particular words in the document. Term frequency is the frequency of occurrence of a certain word present in a document, when the frequency of a word is likely to be more relevant whereas inverse document frequency tells how unique or rare a particular word is in consideration with the input document to be summarized.
3. Sentence ranking: Once the sentences are scored, they are ranked based on the threshold which is computed using the sentences average score. There is the ranking of the sentences by the system with the highest score that begins at the starting position and the lowest rank that ends at the last positions.
4. Extraction of the summary: The summary is extracted based on the scoring of the sentences and this system generates the compressed output as that of the initial sequence of the document.

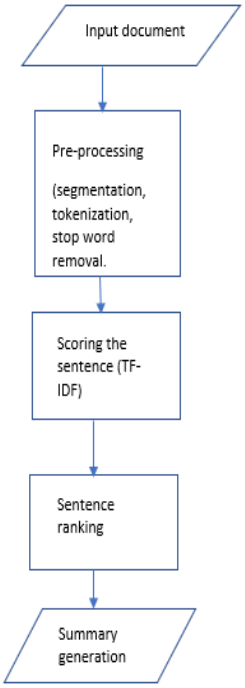

Fig 1: Flowchart of the algorithm

IV. RESULTS

In order to generate the summary, we need to calculate the score of the sentences in the initial document in accordance with certain parameters. The table below shows the scoring of the sentences with a frequency of words in the document containing sentences as well as scoring of the sentences based on the term frequency and inverse document frequency matrix.Table showing sentence score in accordance with the frequency of the words.

Document no	Length of the summarized text (based on the number of sentences)	Length of the original text (based on the number of sentences)
1	32	17028
2	49	25591
3	12	2943
4	162	67132
5	6	5286

Table 1: Frequency of words based on sentence score

Table showing sentence scores in accordance with the term frequency and inverse document frequency

Document no	Length of the summarized text (based on the number of sentences)	Length of the original text (based on the number of sentences)
1	28	17028
2	47	25591
3	11	2943
4	170	67132
5	8	5286

Table 2: Term Frequency and inverse document frequency based on sentence score

The compression ratio can be defined as the ratio of the length of the summarized document to the length of the original document.

CR=LS/LO where LS =length of the summarized document, LO = length of original document, CR= compression ratio

Document no	Compression ratio based on frequency matrix	Compression ratio based on term frequency and inverse document frequency
1	0.001879	0.001644
2	0.001914	0.001836
3	0.004077	0.003737
4	0.002413	0.002532
5	0.001135	0.001513

Table 3: Compression ratio

Document 1 words	TF-IDF values	Document 2 words	TF-IDF values
grate	0.3713	Appli	0.41655
alway	0.3713	Theoret	0.3405
centr	0.3069	Toolkit	0.3405
journey	0.2784	Introduct	0.2167
konkani	0.2629	Highest	0.2167

Table 4: TF-IDF values

Document 3 words	TF-IDF values	Document 4 words	TF-IDF values	Document 5 words	TF-IDF values
Root	0.2187	sound	0.8076	improv	0.386
Get	0.2187	collect	0.2276	concur	0.2205
Present	0.1701	receiv	0.1552	close	0.1544
matrix	0.132	order	0.1306	Reconcile	0.1544
word	0.1181	evid	0.1172	data	0.1290

Table 5: TF-IDF values

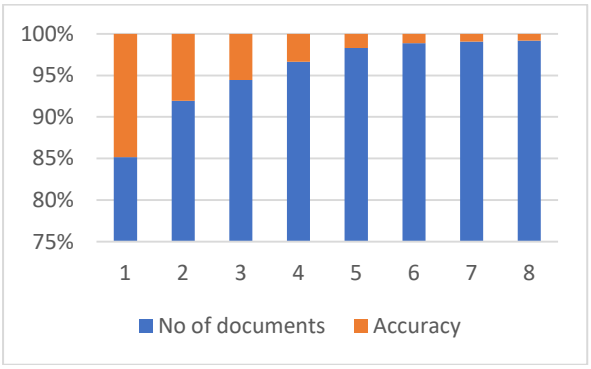

Fig 2: Performance

Fig 3: Original text

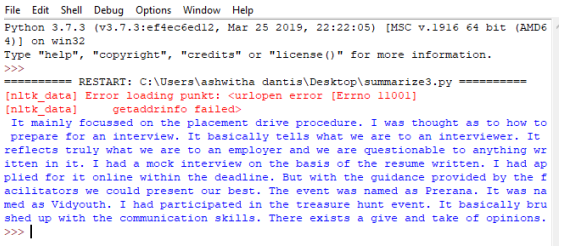

Fig 4: Summarized text

V. CONCLUSION AND FUTURE WORK

This paper has been able to implement the summarization of the text using an extractive approach of summarization by considering the term frequency and inverse document approach. The summary generated is in the identical sequence as that of the initial document. The compression ratio and tf-idf values of the documents are calculated. The approach has been able to ease the task of handling large sized text documents into shorter one by considering the frequency of words approach. The technique emphasizes on considering term frequency and inverse document frequency approach to give a summarized document without losing its relevancy and meaning. This approach has enabled to ease the task of reading large documents by just considering the relevant sentences in the text and summarize a large sized document to a small sized document. The proposed work can be further extended by applying machine learning in order to classify sentence is relevant or not which is considered as a classification problem and generate the summaries of the document.

REFERENCES

1. Bagalkotkar, Anusha, et al. "A novel technique for efficient text document summarization as a service." 2013 third international conference on advances in computing and communications. IEEE, 2013.
2. Urolagin, Siddhaling, and Likitha Satish. "Improving the quality of text summarization using pronoun replacement technique." *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2017.
3. Afsharizadeh, Mahsa, Hossein Ebrahimpour-Komleh, and Ayoub Bagheri. "Query-oriented text summarization using sentence extraction technique." *2018 4th International Conference on Web Research (ICWR)*. IEEE, 2018.
4. Tsuchiya, Seiji, Eriko Yoshimura, and Hirokazu Watabe. "An information arrangement technique for a text classification and summarization based on a summarization frame." *2009 International Conference on Natural Language Processing and Knowledge Engineering*. IEEE, 2009.
5. Zhang, Pei-ying, and Cun-he Li. "Automatic text summarization based on sentences clustering and extraction." *2009 2nd IEEE international conference on computer science and information technology*. IEEE, 2009.
6. Devasena, C. Lakshmi, and M. Hemalatha. "Automatic text categorization and summarization using rule reduction." *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM-2012)*. IEEE, 2012.
7. Tumpa, P. B., et al. "An Improved Extractive Summarization Technique for Bengali Text (s)." *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. IEEE, 2018.
8. Fang, Changjian, et al. "Word-sentence co-ranking for automatic extractive text summarization." *Expert Systems with Applications* 72 (2017): 189-195.
9. Nomoto, Tadashi, and Yuji Matsumoto. "The diversity-based approach to open-domain text summarization." *Information processing & management* 39.3 (2003): 363-389.
10. Xu, Yong-Dong, et al. "MRS for multi-document summarization by sentence extraction." *Telecommunication Systems* 53.1 (2013): 91-98.
11. Fernandes, Roshan, and Rio D'Souza. "Analysis of product Twitter data though opinion mining." *2016 IEEE Annual India Conference (INDICON)*. IEEE, 2016.
12. Fernandes, Roshan, and GL Rio D'Souza. "Semantic analysis of reviews provided by mobile web services using rule based and supervised machine learning techniques." *International Journal of Applied Engineering Research* 12.22 (2017): 12637-12644.

AUTHORS PROFILE

Ashwitha Dantis is currently studying MTech in Computer Science and Engineering at NMAMIT Institute of Technology, Nitte, India. Her area of interests includes Natural Language Processing, Machine Learning.

Roshan Fernandes is currently working as the Associate Professor in the department of Computer Science and Engineering at NMAM Institute of Technology, Nitte. His area of interests include Machine Learning, Mobile Web Services and Semantic Analysis. He is a member of ISTE.

Anisha P Rodrigues Working as Assistant Professor in the Department of Computer Science and Engineering at NMAM Institute of Technology, Nitte, India. Her research interests include Natural Language Processing, Machine Learning, and Big Data Analytics. She is a member of ISTE.

.