MULTI-DOCUMENT TEXT SUMMARIZATION A Survey

Avishek Gangopadhyay, Ankush Khandelwal, KalyaniJadhav, Varad Dingankar

Student,Computer Department,KKWIEER, Nashik, Maharashtra, India

***Abstract*—Multi-document text summarization approach summarize the required information from multiple documents and gives output to the user. Summarization helps the user to get an idea about the required keyword in short by saving a lot of time. In this paper, we made survey on offline search engine and multi-document text summarization. In this approach,word graph generation is used for assigning weight to sentences from documents and this weightage is used further for summarization. Research in summarization has been made but we are going to combine search engine and multiple document summarization in one package.Research on summarization of multiple document is bit behind so we have concentrated on the same.**

Keywords—Abstractive Summary, LDA,Multi-document summarization,,Sentence Extraction information retreival,Text mining , Topic identification.

I. INTRODUCTION

Over the past few years, interest in multi-document summarization is increasing. Summarization is method of compressing large amount information into main points which gives an abstract idea about the source information. A computer program has no basis for knowing which sentences are to be included in the final summary so with the help of Word Graph Generation in assigning weightage to the sentences we can easily generate proper summary. Mainly, Multi-document summaries should include relevant information across all documents only once and also the information which is directly related to the user query to have optimal result.

Word Graph Generation is based on weighting function which is used to identify the links between the words which appear significantly across the documents. WGG is also used to generate an informative compression which promotes paths passing through main nodes. Dijkstra's algorithm is also a part of WGG whose application is to find the shortest path.

Current multi-document summarization systems can successfully extract sentences but with many restrictions including document size, inaccurate extraction to important sentences, redundancy, feeble resemblance between selected sentences.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before understanding and implementing the tool it is important to consider the time factor, economy and company strength.

In [1] the summarization process is done by first tokenizing the sentences, sentence segmentation and stop-word detection for identification of the important terms from the paper. It considers the words which are occurring most frequently, proper nouns, sentence length, Cue-phrases such as "in conclusion", "this letter", "this report", "attempts", etc. The summary is generated by considering these points.

In [2] produces extractive single/multi-document generic or query-focused summary. The favorable conditions of the framework that has been effectively utilized as a part of current summarizer, test summarization features, assessment metric, short query machine interpretation framework.

In [3]relying on the combination of statistical and positional similarity feature. The advantage of that it makes available well-established summarization method through the implementation and integration of algorithms and evaluation of resources.The disadvantage is that the study of problem of opinion summarization is needed.

In [4] Extraction based multi-document summarization algorithm consist of choosing sentences from the document using some weighting mechanism and combining them into summarization. It is a mixture of significant and insignificant topic in terms of sentence weight. Meanwhile we use traditional characters such as term frequency, sentence position and sentence length.

In [5] A new Bayesian sentence based topic model for summarization is presented. This model makes use of both term document and term sentence association to help the context understanding and guide the sentence selection in summarization procedure.
In [6] A novel supervised approach taking advantage of both topic model and supervised learning is proposed. This

ISSN: 2278 – 1323

approach can incorporate rich sentence feature into Bayesian topic models.

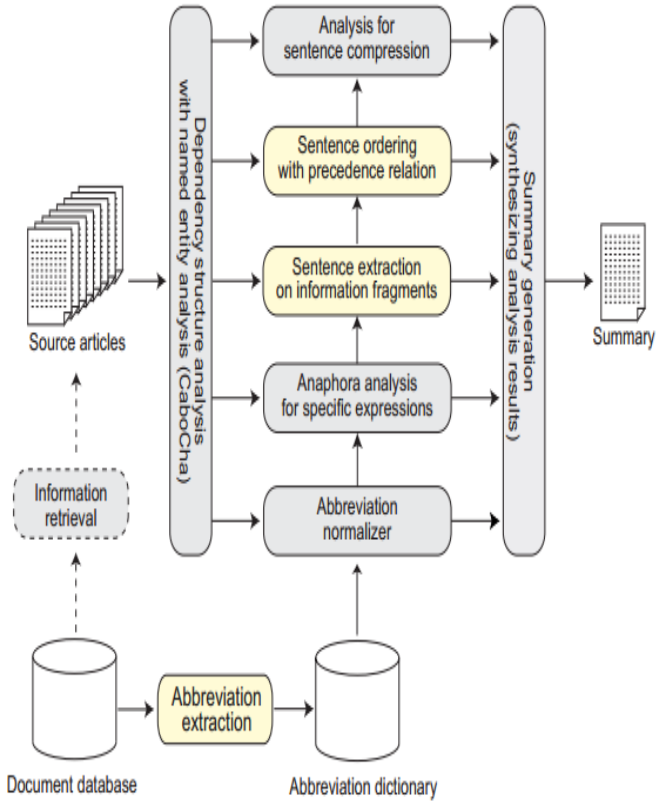

Figure 1: Architecture of Traditional system

The phases of summarizer can be split into three general categories 1) Interpret This is where a representation of document to be summarized id produced. Also known as analysis. 2) Transform This is where the representation of document is turned into one of a summary of the document. 3) Generate Here the summarized text is produced, also known as synthesis.

III. Types of Summary

1. Extractive: This where the summary consists of sentences that have already appeared in text
2. Abstractive: Here summarizer generates some new text, clearly extractive summary is the simpler option of the two because they avoid language generation problem.
3. Indicative: These summaries give the reader an idea about how it would be worthwhile reading the whole document.
4. Informative: In this important factual content of the text is expressed.

IV. Proposed System

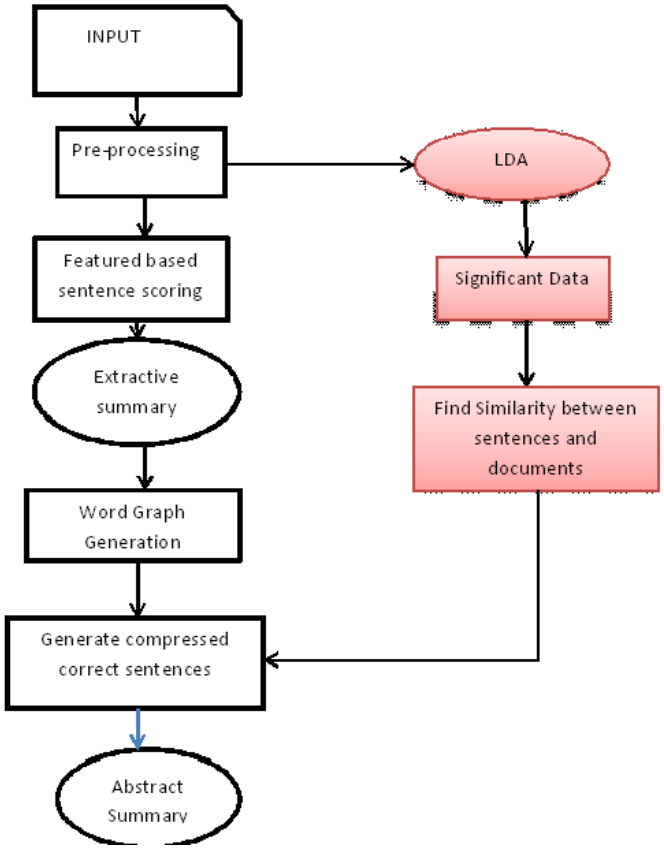

Figure 2: Block Diagram

The traditional system uses the concept of word weightage. It is done by using the following properties word frequency, word positions, proper noun, stop words etc. considering these properties the summaries are made irrespective to the user requirements. The proposed system will ask for a particular key word(s) which will always be referred during the whole calculations. LDA is an approach, which uses similar concept; LDA mainly focuses on the sentence/word score and relation between the weighted entity and the document. Which in return differentiates the significant and insignificant data.

Word Sentence Score (WSS): The word scoring is calculated for summarization of the terms weight, term are nothing but words. The terms weight is calculated by number of words in each sentence divided by number of total sentence in the document.

Term Sentence Frequency (TSF):
The term frequency is very important feature. TF (term frequency) represents how much time the term appears in the

ISSN: 2278 – 1323

document to calculate the term frequency. Thus term sentence frequency is calculated by total number of term frequency divided by total number of sentence in the document. The term Identifying sentence boundaries in a document is based on punctuation such as (. , (, “, [, {, etc.) and split into sentences. These sentences are nothing but tokens.

The Similarity to First Sentence:
This feature is to score the sentence based on its similarity to the first sentence in the document. The first sentence in the document is very important sentence in the document. The sentence has board coverage of the sentence set (document) will get the high score.

Stop word filtering:
In any document there will be many words that appear regularly but provide little or no extra meaning to the document. Words such as 'the', 'and', 'is' and 'on' are very frequent in the English language and most documents will contain many instances of them. These words are generally not very useful when searching; they are not normally, what users are searching for when entering queries. Because of this, it can be beneficial to remove these words from the index. This has the benefit of reducing the size of the index, as well as improving the performance of retrieving documents from the index.

Sentence Length:
In summarization, too long or too short sentences are not expected.

Sentence Position:
Sentence position always gives the importance of the sentences. Research has shown that, the proportion that the first sentence as summarization is 85% and the last sentence as summarization is 7%.

Word Frequency:
The word frequency occurrences within a document have often been used for calculating the importance of sentence. The more feature word the sentence contained, the more information the sentence contained.

The proposed system is mixture of all these concepts and algorithms which in return gives a more accurate and relevant result.

According to the block diagram,input is taken from the user, and that input is sent for the next step i.e. preprocessing, this step is carried to prepare the sentence for further processing. There are six features, which are used to capture the values of sentences, which are term feature, proper noun, position feature, sentence length sentence centrality, cue phrase.

V. CONCLUSION

Multi-document summarization algorithm consists of choosing sentences from the documents using word graph generation and combining them into a summary. In this paper, we have proposed the system, which will generate abstractive summary. This paper has introduced combination of LDA and Improve COPMENDIUMfor multi-document summarization. The proposedsystem divides estimated topic into significance topic and insignificance topic. In term of sentence weight, we use similarity between sentence topic and significance topic. Meanwhile, we use traditional characters such as term frequency, sentence position and sentence length. In the future, we will consider how to determine the number of topic by significance topic automatically.

VI.REFERENCES

[1] Ms.Jagtap Jayanti,Prof.Patel H.H,”Generating abstract of Research paper Using Text Summarization System”
[2] D.Radev,T.Allison,J.Blitzer,A.Celebi,W.Lam,D.Liu,H.Qi,M.Topper,”3 MEAD a platform for multidocument multi lingual text summarization”
[3] H.Saggion,”SUMMA : A robust and adaptible summarization tool”.
[4] Liu Na,Tang Xiao-jun,Lu Ying,Li Ming-xia,”Topic sensative multidocument summarization algorithm.
[5] Dingding Wang, Tao Li, “Multi—Document summarization using sentence-based topic model
[6] Li Jiwei, Li Sujian,”A novel feature based Bayesian Model for Query focused multi-document summarization”

Avishek Gangopadhyay BE Computer K.K.Wagh College of Engineering Nashik
Ankush Khandelwal BE Computer K.K.Wagh College of Engineering Nashik
Kalyani Jadhav BE Computer K.K.Wagh College of Engineering Nashik
Varad Dingankar BE Computer K.K.Wagh College of Engineering Nashik

ISSN: 2278 – 1323