Multi-Document Summarization using Fuzzy and Hierarchical Approach

Prof. Shilpa M. Satre
Assistant Professor, Department of Information Technology
Bharati Vidyapeeth College of Engineering
CBD Belapur, Navi Mumbai
shilpa.m.shelar@gmail.com

Mugdha Patil
Department of Information Technology
Bharati Vidyapeeth College of Engineering
CBD Belapur, Navi Mumbai
mugdha1505@gmail.com

Shruti Raju
Department of Information Technology
Bharati Vidyapeeth College of Engineering
CBD Belapur, Navi Mumbai
shrutiraju18@gmail.com

***Abstract*— Multi-Document Summarization is the task of extracting important information from the original text document. With the increasing amount of online information, it becomes extremely difficult for users to find relevant information. Information retrieval systems usually return a large number of documents listed in the order of estimated relevance. It is time consuming for the users to read each document in order to find useful ones. Multi-Document summarization system helps by providing a quick summary of the information contained in the document. This summarization system selects the most important sentences from the input document. It supports selection of documents from multiple formats. A perfect summary does not contain repeated information and includes distinct and precise information from multiple documents on that topic.**

Keywords— Document, Similarity Measure, Page Rank, Expectation Maximization, Clustering, Summary.

I. INTRODUCTION

Data mining is the domain in which rapid changes are evolved in the recent years due to the enormous advances in the software and hardware technology. The advancement has led to the availability of various kinds of data, which is especially suitable for the instance of text data. The World Wide Web has the capability of providing enormous amount of online information globally. Due to this fact, the users submit numerous search queries on the search engine on the Internet every moment. For retrieving information, people widely use internet search engines such as Google, Yahoo, Bing and so on. To understand quickly any kind of information, the users desire to have the maximum coverage of information in a short text representation. The search engine as a response provides information in variety of web pages. This abundant information makes the reading complexity to the users. There is need of efficient multi-document summarization as the internet provides the access to a very large amount of data in a particular language & it also helps them in making decision in a lesser duration.

Document summarization essentially improves the effective retrieval process from the multiple documents. Due to overloading of information which is available on the web, summarization is popular research area. Nowadays, there is a need for strong and powerful document summarizer. NLP is a process of developing a system that can process and produce language as good as human can produce. Document summarization is one of the applications of Natural language processing. Multi-document summarization is the summary of n number of source documents of text in shorter version, that retain the main feature of the content and help the user to quickly understand large volume of information..

II. LITERATURE REVIEW

Muhammad Azhari and Yogan Jaya Kumar, 2017 proposed a English text summarization model based on classification using neuro fuzzy approach. The model is trained to filter high quality summary sentences. Neuro fuzzy approach consists of benefits of both fuzzy and neural networks. This model uses five features for calculation of sentence score. This model combines the explicit knowledge reasoning of fuzzy logic system which can explain input output relationship and implicit knowledge of neural networks which can be learnt. But the quality of summary of sentences was not evaluated in this model. [1]

Taeho Jo, 2017 proposed a specific version of KNN (K Nearest Neighbor) where the similarity between feature vectors is calculated including the similarity among attributes, features and values. This method generates more compact representation of data items and the better performance. It uses modified version as approach to text summarization. Binary classification is done in this technique which judges whether each sentence is important or not based on the parameters. This method can be applied and validated in specialized domains: engineering,

medicine, science and law and it should be customized to the suitable version. [2]

Kalal Gayatri Pradip and D.R.Patil, 2016 used Hierarchical and Fuzzy Relational Eigenvector Centrality-based Clustering Algorithm for the clustering of sentences to overcome the problems of traditional clustering methods. Calculation of similarity measure, page rank and Expectation Maximization were some of the steps used to extract important sentences. In this paper both divisive and agglomerative hierarchical clustering were used to form clusters.[3]

Saif alZahir, Qandeel Fatima and Martin Cenek, 2015 present a new multigraph-based text summarizer method. This method is distinct as it produces a multi-edge-irregular-graph which represents frequency of words in the sentences of the target text. This graph is then transformed into a symmetric matrix from which we can derive the ranking of sentences and hence obtain the summarized text using a threshold. This method is fast and can be implemented for real time summarization. Here the number of edges in the graph between two sentences is equal to the number of same words in both sentences. The total number of edges is stored in a symmetric matrix that represents the text being summarized. Then the values of the rows of the matrix are added to generate the sum vector which is used for ranking the sentences. This method can be used for better comparison between the documents. Sentence fusion(combination) and compression (removal of unimportant sentences) can also be done using this technique. [4]

Sebastian Suarez Benjumea and Elizabeth Leon Guzman., 2015 proposed a novel approach for automatic extractive text summarization called SENCLUS. Using Genetic clustering algorithm, SENCLUS clusters the sentences as close representation of the text topics using a fitness function based on redundancy and coverage , and applies a scoring function to select the most relevant sentences of each topic to be part of the extractive summary. This technique is good to deal with information overload. The drawback of this method is that it can be only used to summarize single document. [5]

III. METHODOLOGY

Text summarization can be done by two ways. They are Sentence Abstraction and Sentence Extraction. Sentence Abstraction includes changing the voice of the sentence or changing the form of sentence. It includes reconstruction of sentences based on analysis of deep natural language. Sentence Extraction includes extracting the sentence as it is from the document. The extraction technique of text summarization consists of selecting significant sentences from source documents and arranges them in the destination documents. Our main focus is on extraction technique for text summarization of multiple documents in English.

The proposed method uses statistical approach to find most relevant sentence. This method uses sentence importance as an important feature for summarization. This extractive method is based on parameters such as sentence length, sentence position and word occurrence in the document. It also uses fuzzy logic and semantic analysis. The fuzzy logic helps to extract significant sentence and word features from the given document and decision module determines the degree of importance of each sentence based on its feature scores. The decision making is based on rule-base module of fuzzy system. The dataset consists of documents or articles in a particular domain.

As shown in Fig.3.1, multi-document summarization system consists of six major steps which includes preprocessing, page rank, similarity measure, expectation maximization and clustering and extraction of final summary from the set of clusters formed.

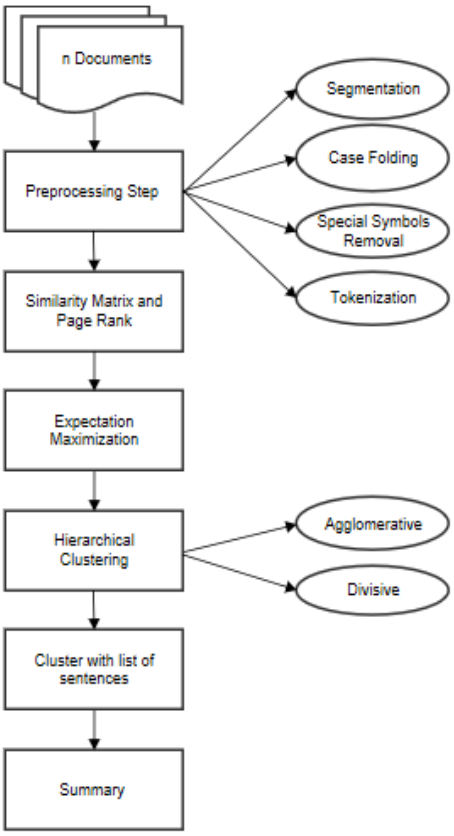

Fig.3.1 Architectural Diagram

A. Document selection

Firstly, documents should be selected for summarization. Single or Multiple documents can be selected. This documents to be selected can be in multiple formats.

B. Preprocessing

Various preprocessing steps included are:

a) **Sentence Segmentation:-**It is the process of dividing a string of written language into its component sentences. This can be achieved by analyzing question marks (?) and periods (.).

b) **Tokenization**:-It is the task of chopping sentences into pieces, called tokens , perhaps at the same time throwing away certain characters, such as punctuation.

c) **Case Folding:-**The process of converting all the characters in a document into the same case, either all upper case or lower case, in order to speed up comparisons during the indexing process.

d) **Word Count:-**Returns number of words present in the document.

C. Similarity Measure and Page Rank

a) Statistical methods can be used in the task of determining novelty in a sentence (or better: comparing the content of two sentences), in the following way: firstly, creation of vocabulary consisting of all different (stemmed) words existing in the given text is done, then representation of each sentence as a vector in the vocabulary space as a bag-of-words should take place and then comparison between two sentence is done using a similarity measure.

b) Euclidean Distance is used as a similarity measure.

c) This feature finds the similarity between the sentences. For each sentence S, similarity between S and every other sentence is computed by the method of token matching. The two dimensional matrix is formed of the size [N] [N] where N is Number of sentences in the document. In this matrix diagonal elements should be assigned 0 values as sentence should not get compared with itself. $Sim(Si,Sj)=TM[(ti)^{n_1}, (tj)^{m_1}]$ where TM is token matching method. The score for sentence to sentence similarity is calculated as ratio of summary of similarity of sentence S with every other sentence over the maximum summary.

$F5=\sum[Sim(Si,(Sj))^{N_1}]\div Max=\sum[Sim(Si,(Sj))^{N_1}])^{N_{i=1}}$

d) The premise of the PageRank algorithm is “prestige”. PageRank provides us with a means of identifying which amongst these nodes is the most prestigious.

D. Expectation Maximization

a) An Expectation–maximization (EM) algorithm is an iterative method to find maximum likelihood in presence of unknown, hidden or missing data.

b) The Expectation Maximization (EM) algorithm can be used to generate the best supposition for the distributional parameters of some multi-modal data. Parameter estimation is done in this step.

c) Each iteration of EM algorithm consists of two process the E-step and the M-step.

d) In the E-step, the missing data is estimated from the given observed data, In the M-step the likelihood function is maximized under the assumption that missing data is known.

E. Hierarchical clustering

It is a method of clustering where hierarchy of clusters are built. Two types of Hierarchical clustering used are:

a) Divisive: This is a "top down" approach: All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

b) Agglomerative: This is a "bottom up" approach. Each observation is assigned to its own cluster. According to the similarity (e.g. distance) between each of the clusters, they are merged as one moves up the hierarchy.

Fig 3.2 shows both agglomerative and divisive clustering.

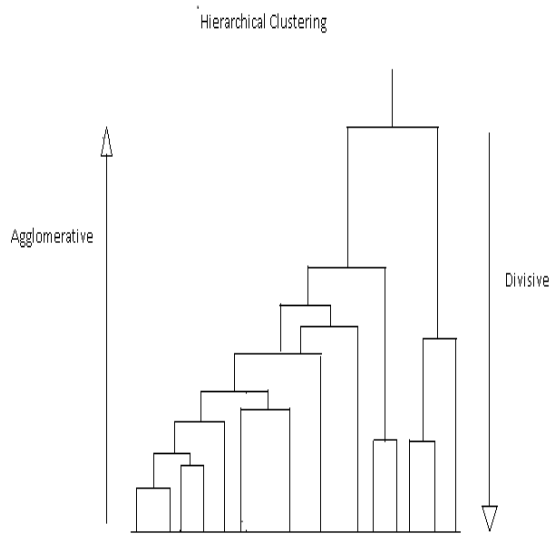

Fig.3.2 Types of Hierarchical Clustering

F. Summary Extraction

Summary will be extracted from the list of clusters. This Summary is the final summary of the single document or multiple documents.

IV. EXPERIMENTAL RESULTS

- This section reports the experimental results of our evaluation.
- The following graph shows the comparison between actual word count before obtaining the summary and the summarized word count.
- The summarized content contains the sentences from the selected documents with highest parameter measures.
- Multiple formats have been taken into consideration to perform the experimental result.

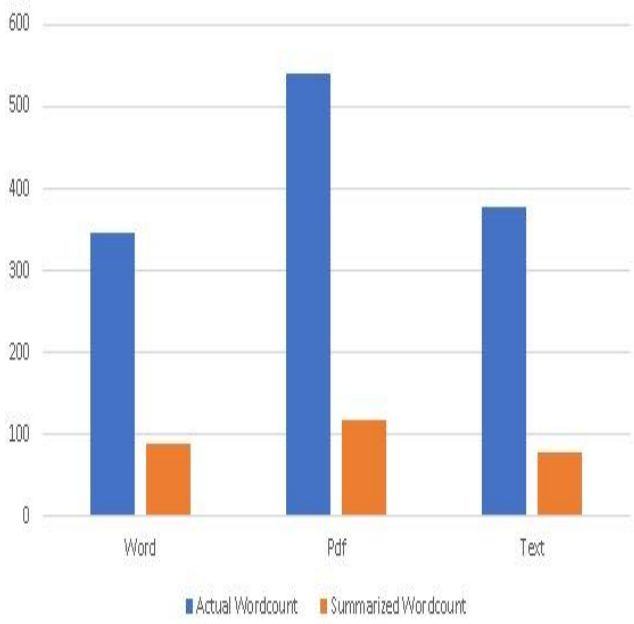

Fig.4.1 Wordcount Comparison from Fuzzy Relational Eigenvector Centrality based Clustering Algorithm and Hierarchical Clustering among different formats

V. CONCLUSION

This system proposes multiple text documents summarized using extractive method. The summarization system which is based on HFRECCA method will improve the quality of the summary. Based on different similarity measures and ranking procedures, presentation of model for performing consensus of multiple summarization algorithms that perform extraction based text summarization has been done. In the Proposed system most of the used resources such as list of nouns, stop word list and Cue word list, list of special symbols etc are considered. It offers a possibility of finding important points of texts from multiple text documents and so user can spend comparatively less time of finding important or precise information.

VI. REFERENCES

[1] Muhammad Azhari, Yogan Jaya Kumar, improving text summarization using neuro-fuzzy approach, taylor & Francis, Vol. 1, No. 4, 367-379, 2017.

[2] Taeho Jo, K Nearest neighbor for text summarization using feature similarity, ICCCCEE, 2017.

[3] Kalal Gayatri Pradip and D.R.Patil, Summarization of Sentences using Fuzzy and Hierarchical Clustering Approach,IEEE,2016.

[4] Saif alZahir and Qandeel Fatima, Martin Cenek: New Graph-Based Text Summarization Method, IEEE, 2015, 978-1-4673-7788-1/15.

[5] Sebastian Suarez Benjumea, Elizabeth Leon Guzman: Genetic Clustering Algorithm for Extractive Text Summarization, 978-1-4799-7560-0/15 $31.00 © 2015 IEEE, DOI 10.1109/SSCI.2015.139

[6] Aik, L.E.(2008). A study of neuro –fuzzy system in approximation based problems. Neuro – fuzzy system ANFIS: Adaptive Neuro – Fuzzy Inference System 24(2), 113-130.

[7] Mr. Sarda A. T., Mrs. Kulkarni A.R.: Text Summarization using neural networks and Rhetorical Structure Theory, IJARCCE.

[8] Rucha S. Dixit, Prof. S. S. Apte, Improvement of Text Summarization using Fuzzy logic based method, IOSR Journal of Computer Engineering.

[9] Atif Khan, Naomie Salim, Haleem Farman, Clustered Genetic Semantic Graph Approach for Multi-document Abstractive Summarization, 978-1-4673-8753-8/16/$31.00 ©2016 IEEE.