To Provide Acuuracy For Document Summerization In Contextual Documents

P.V.Hedaoo[1], K.N.Hande[2], M.S. Chaudhari[3]
[1] Dept of CSE
[2, 3] Assistant Professor, Dept of CSE
[1, 2, 3] Priydarshani Bhagwati College Of Engineering, Nagpur.

Abstract- *The increase in the performance and fast accessing of web resources has made a new challenge of browsing among huge data on Internet. Since digitally stored information is more and more available, users need suitable tools able to select, filter, and extract only relevant information. Text summarization can be classified into two approaches: extraction and abstraction. This paper focuses on extraction approach. The goal of text summarization based on extraction approach is sentence selection. One of the methods to obtain the suitable sentences is to assign some numerical measure of a sentence for the summary called sentence weighting and then select the best ones. Therefore, text summarization techniques are currently adopted in several fields of information retrieval and filtering, such as, information extraction, text mining, document classification, recommender systems, and contextual advertising. A summary text is a derivative of a source text condensed by selection and/or generalization on important content. Query-focused summaries enable users to find more relevant documents more accurately, with less need to consult the full text of the document. Extractive summarization methods try to find out the most important topics of an input document and select sentences that are related to these chosen concepts to create the summary. This paper is a Comparative study of four techniques used for extractive summarization, namely, Neural Network, Graph Theoretic, Fuzzy based method and Cluster based method.*

I. INTRODUCTION

With the tremendous increase of digitized information, the mining task has become a crucial tool for aiding and understanding the information. This includes clustering, classification, categorization and summarization. The major challenge is to find relevant information from large amount of data. Summaries are often necessary to enable timely relevancy assessments, information extraction, or information analysis from source material. Text summarization is an effective technique that is used in combination with Information Retrieval and Information filtering systems to save the user time. [1]

Today the size of the repository of information is much larger than one can manage, easily and efficiently. This includes business transactions, news reports, satellite data, digital media, text reports and memos and biological information. [2] Moreover in today's life everyone wants to gain more and more in less time. Thus reading long documents and then gaining the insight of the document is not a good idea. It will be more beneficial if one go through the summary of the long document and still gaining the theme or core information present in the document. In this way more and more information can be gathered in less time. Thus the demand for efficient data mining techniques is increasing day by day.

Now-a-days there are plenty of online news websites overwhelmed with news articles. The most important tasks of news engines are Collecting News, News Retrieval, Categorizing Search Result, Summarization and Automatic Event Detection. The quality of each of the tasks depends on the quality of several other tasks. [3] This paper focuses on simple technique to take query from user and receive the ranked news related to the user's query from web. The irrelevant news articles are discarded and user gets refined data. The technique used to produce extractive summary for single news article that carries most important information is keyword extraction technique.

II. RELATED WORK

Many previous works on extractive summarization ranks sentences based on simple features such as their position in the text, frequency of the words they contain, or some key phrases indicating the importance of the sentences [11][12][13] and select top n sentences based on the compression ratio. In another approach to multi-document summarization, information extraction is used to identify similarities and differences across the documents in the set.

In the centroid based multi-document summarization [2][3], the sentences are ranked based on its similarity to the cluster centroid and a number of top-ranked sentences are selected based on the compression ratio. The centroid is

defined as a pseudo-document consisting of words with TF*IDF scores greater than a predefined threshold.

An alternative approach to ensure good coverage and avoid redundancy is the clustering based approach that groups the similar textual units (paragraphs, sentences) into multiple clusters to identify themes of common information and selects text units one by one from clusters in to the final summary [5][6][7][8][9]. Centroid based summarization method [2][3] can be thought to be a single cluster based approach since it groups the sentences closest to the centroid in to a single cluster.

While sentences are extracted from multiple source documents, picking sentences out of context may result in incoherent summary. Ensuring coherence is difficult, because this requires some understanding of the content of each passage and knowledge about the structure of discourse. Practically, most systems follow time order and text order.

Compared to creating an extract, generation of abstract is relatively harder since the latter requires: (1) semantic representation of text units (sentences or paragraphs) in the text, (2) reformulation of two or more text units and (3) rendering the new representation in natural language. Abstractive approaches have used template based information extraction, information fusion and compression.

III. IMPLEMENTATION

Figure 1 illustrates an overview of the proposed approach for multi-document summarization system. The input to the system is a collection of documents. The output is a concise cluster-wise summary providing the condensed information of the input documents. The proposed approach produces an extractive summary by selecting salient sentences from the documents cluster wise. All the relevant documents are grouped together into clusters by using threshold-based document clustering approach. Based on feature profile salient sentences from each cluster are identified and ranked according to their weights of importance. Based on the ranking of sentences, sentences are selected and ordered. The system then iteratively extracts one sentence at a time, until the required summary length is met for each cluster.

The proposed approach can be decomposed into five sub processes:

1. Preprocessing
2. Documents Representation and Clustering
3. Sentence Score Calculation based on Feature Profile
4. Cluster wise Sentence Ranking and Ordering
5. Summary Generation

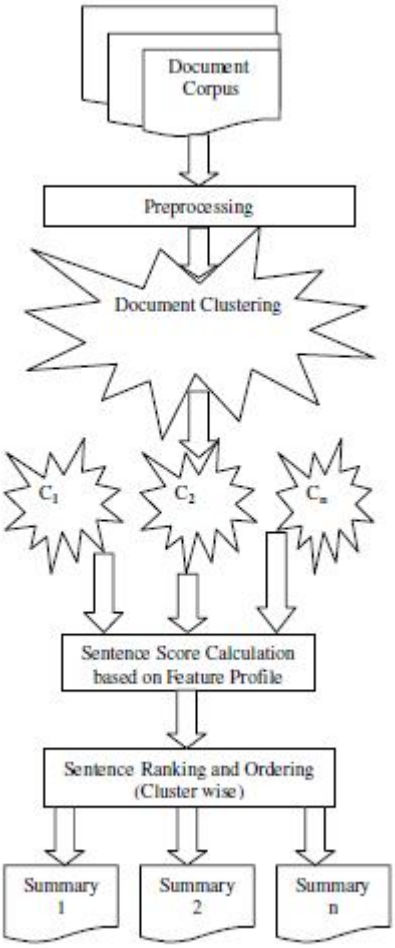

Fig 1. Summarization Working

The working of these component modules is explained below.

3.1 Corpus Builder

This module collects news pages retrieved from the web corresponding to the query terms. The user enters query terms through user interface. This query is passed to news collector that retrieves the resultant news pages from the web. The headlines of the retrieved news pages are tokenized so as to find the headline similarity. All the news pages that are having similarity value above some set threshold say Θh are added to the corpus.

3.2 Summarizer

News page corresponding to selected headline is pre-processed first. This module takes the document and performs some sort of pre-processing so as to obtain an intermediate representation. Then keywords are extracted and weighted which are then later used by sentence ranker to calculate

sentence relevancy. The sentences are then ranked and k top most ranked sentences are presented as summary to user. The summarizer includes: pre-processing, keyword extraction, sentence ranking and sentence filtering.

The working of different components of summarizer is as follows:

3.2.1 Pre-processing: In pre-processing, first main article is extracted from the news page, stopwords are eliminated, and light stemming is performed (i.e. only plural forms are stemmed).

3.2.2 Keyword Extraction: Keywords are the index terms that contain most important information. Kaur and Gupta [12] discussed the different approaches to identify keywords. The quality of summary highly depends on keyword extracted which is only possible when several features are combined. The proposed system identifies the keywords using the following approaches:

IV. EXPERIMENTAL RESULT

The following snaps gives the idea about the document summarization and how it work.

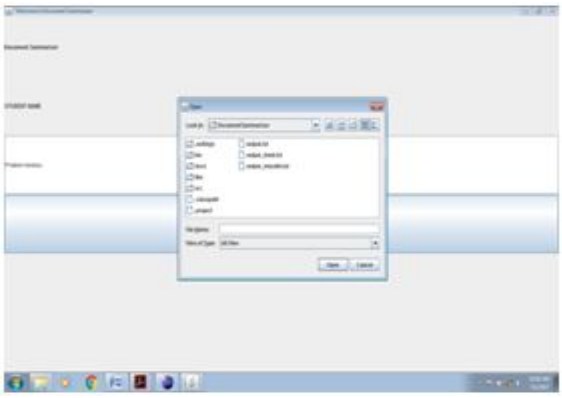

Fig 2. Asking to select documents

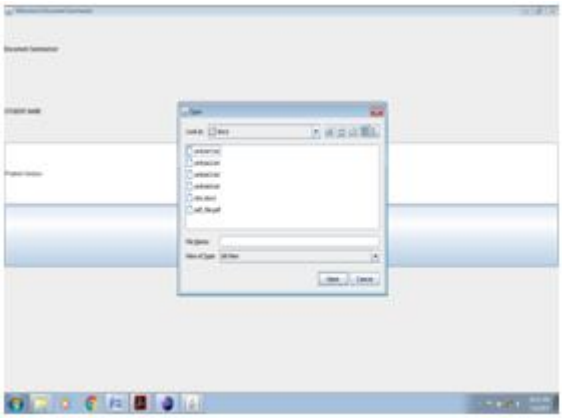

Fig 3. Option to select text,word or pdf file

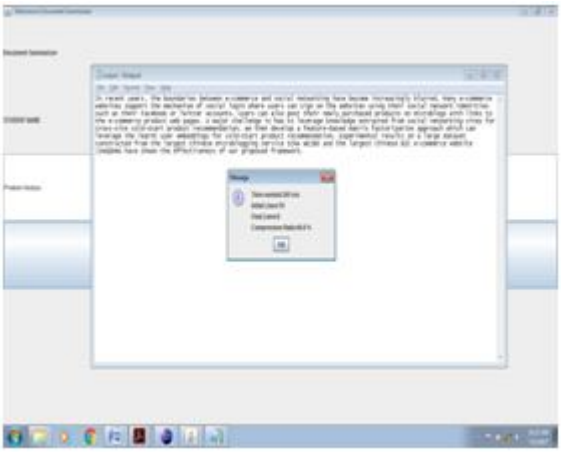

Fig 4. Output – Summarize text, time required and compression ratio

Table 1. Result evalution & analysis

Sr. No.	Document Type	Initial Lines	Final Lines	Delay (in ms)	Compression Ratio (%)
1.	Text	10	06	125	40
2.	Word	25	14	560	44
3.	Text	39	26	252	33.33
4.	pdf	46	30	313	34.78
5.	word	103	35	1298	66.02
6.	pdf	424	289	2764	31.84

V. CONCLUSION

The proposed method discusses about grouping related documents using document clustering and cluster-wise summary generation using feature profile oriented sentence extraction strategy. Accuracy is improved by employing TSF-ISF measure. The summary generated using the proposed method is compared with human summary and its performance has been evaluated and the result shows that the machine generated summary coincides with the human intuition for the selected dataset of documents.

REFERENCES

[1] Mani, I., Maybury, M. T., Ed. Advances in Automatic Text Summarization. The MIT Press, 1999.

[2] Hovy, E., Lin, C. Y., Zhou, L., et al. Automated Summarization Evaluation with Basic Elements. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC). Genoa, Italy, 2006.

[3] Jackson, P., Moulinier, I. Natural language processing for online applications. John Benjamins Publishing Company, 2002.

[4] Padr´o Cirera, L., Fuentes, M.J., Alonso, L., et al. Approaches to Text Summarization: Questions and

Answers. Revista Iberoamericana de Inteligencia Artificial, ISSN 1137-3601, (22):79–102, 2004.

[5] E. H. Hovy, Automated Text Summarization, The Oxford Handbook of Computational Linguistics, chapter 32, Oxford University Press, Oxford, 2005, pp. 583-598.

[6] M. T. Maybury, Generating summaries from event data, Information Processing and Management, Elsevier, vol. 31, no. 5, 1995, pp. 735-751

[7] X Wan and J Yang, Multi-document summarization using cluster-based link analysis, In Proceedings of '08 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), ACM. Singapore, 2008, pp. 299-306.

[8] F EI-Ghannam and T EI-Shishtawy, Multi-topic multidocument summarizer, International Journal of Computer Science & Information Technology (IJCSIT), vol. 5, no. 6, 2013, pp. 77-90.

[9] M. Damova and I. Koychev, Query-based summarization: a survey, In International Conferenceon Software, Services and Semantic Technologies (S3T '2010), Varna, Bulgaria, 2010, pp. 142-146

[10] Wang, W., et al. Exploring hypergraph-based semisupervised ranking for query-oriented summarization, Information Sciences, Elsevier, vol. 237, 2013, pp. 271-286.

[11] E. Lloret, A. Balahur, M. Palomar, and A. Montoyo, "TowardsBuilding a Competitive Opinion Summarization System: Challenges and Keys," Proc. Human Language Technologies: The 2009 Ann. Conference of the North Am. Ch. Assoc. for Computational Linguistics.

[12] Lin C.Y. and Hovy E. 2000. The automated acquisition of topic signatures for text summarization. In Proc. of the 18th conference on Computational linguistics - Volume 1.

[13] Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1.