Text Summarization using Clustering Technique

Anjali R. Deshpande [#1], Lobo L. M. R. J. [*2]

[#1] *Student M. E.(C. S. E.) Part–II, Walchand Institute of Technology, Solapur University, Solapur, India.*

[#2] *Professor & Head, Department of InformationTechnology, Walchand Institute of Technology, Solapur University, Solapur, India.*

***Abstract*— A summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Due to the problem of information overload, access to sound and correctly-developed summaries is necessary. Text summarization is the most challenging task in information retrieval. Data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. This paper presents a combined approach to document and sentence clustering as an extractive technique of summarization.**

***Keywords*— Extractive Summarization, Abstractive Summarization, Multi-document summarization, Document Clustering, Vector space model, Cosine similarity.**

I. INTRODUCTION

Text Summarization has become significant from the early sixties but the need was different in those days, as the storage capacity was very limited. Summaries were needed to be stored instead of large documents. As compared to that period large number of storage devices are now available which are fairly cheap but due to the wide availability of huge amount of data, the need for converting such data into useful information and knowledge is required. When a user enters a query into the system, the response which is obtained consists of several Web pages with much information, which is almost impossible for a reader to read it out completely. Research in automatic text summarization has received considerable attention in today's fast growing age due to the exponential growth in the quantity & complexity of information sources on the internet. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration. A summary is a text that is produced from one or more documents, which contains a fundamental portion of the information from the original document, and that is no longer than even half of the original text. The stages of Text Summarization include topic identification, interpretation, and summary generation. The summaries may be Generic or Query relevant summaries (query based summaries).

There are 2 types of Summarization: Extractive & Abstractive [1]. Human generated summaries reflect the own understanding of the topic, synthesis of concepts, evaluation, and other processing. Since the result is something new, not explicitly contained in the input, this requires the access to knowledge separate from the input.

Since computers do not yet have the language capabilities of human beings, alternative methods must be considered. In case of automatic text summarization, selection-based approach has so far been the dominant strategy. In this approach, summaries are generated by extracting key text segments from the text, based on analysis of features such as term frequency and location of sentence etc. to locate the sentences to be extracted.

A. Multi-document Summarization

It is an automatic procedure designed to extract the information from multiple text documents written about the same topic. Resulting summary allows individual users, such as professional information consumers, to quickly familiarize themselves with information contained in a large collection of documents.

The multi-document summarization task is much more complex than summarizing a document & it's a very large task. This difficulty arises from thematic diversity within a large set of documents. A good summarization technology aims to combine the main themes with completeness, readability, and conciseness [2].

The present implementation includes development of an extractive technique that combines first clustering of documents & then clustering of sentences in these documents. The results achieved are remarkable in terms of efficiency & reducing redundancy to a great extent.

II. LITERATURE REVIEW

The approach presented in [3] is to cluster multiple documents by using document clustering approach and to produce cluster wise summary based on feature profile oriented sentence extraction strategy. The related documents are grouped into same cluster using threshold-based document clustering algorithm. Feature profile is generated by considering word weight, sentence position, sentence length, sentence centrality, proper nouns and numerical data in the sentence. Based on the feature profile a sentence score is calculated for each sentence. This system adopts Term Synonym Frequency-Inverse Sentence Frequency (TSF-ISF) for calculating individual word weight. According to different compression rates sentences are extracted from each cluster and ranked in order of importance based on sentence score. Extracted sentences are arranged in chronological order as in original documents and from this, cluster wise summary is generated. The output is a concise cluster-wise summary providing the condensed information of the input documents.

Kamal Sarkar presented an approach to Sentence Clustering-based Summarization of Multiple Text Documents in [4]. Here three important factors considered are: (1) clustering sentences (2) cluster ordering (3) selection of representative sentences from the clusters. For the sentence clustering the similarity histogram based incremental clustering method is used. This clustering approach is fully unsupervised & is an incremental dynamic method of building the sentence clusters. The importance of a cluster is measured based on the number of important words it contains. After ordering the clusters in decreasing order of their importance, top n clusters are selected. One representative sentence is selected from each cluster and included in to the summary. Selection of sentences is continued until a predefined summary size is reached.

A query based document summarizer based on similarity of sentences and word frequency is presented in [5].The summarizer uses Vector Space Model for finding similar sentences to the query and Sum Focus to find word frequency. In this paper they propose a query based summarizer which being based on grouping similar

sentences and word frequency removes redundancy. In the proposed system first the query is processed and the summarizer collects required documents & finally produces summary.

After Pre-processing, producing the summary involves the following steps:

1. Calculating similarity of sentences present in documents with user query.
2. After calculating similarity, group sentences based on their similarity values.
3. Calculating sentence score using word frequency and sentence location feature.
4. Picking the best scored sentences from each group and putting it in summary.
5. Reducing summary length to exact 100 words.

The research in [6] is done in three phases: a text document collection phase, training phase and testing phase. They need the document text files in Indonesian language.

The training phase was divided into three main parts: a summary document, text features, and genetic algorithm modelling. As a part of this, document was manually summarized by three different people. Text feature extraction is an extraction process to get the text of the document. At the stage of modelling genetic algorithms, genetic algorithm serves as a search method for the optimal weighting on each text feature extraction. Stages, summary and manual extraction of text features are used to calculate the fitness function that serves to evaluate the chromosomes.

Testing phase used 50 documents (documents used at this stage were different from the documents used in the training phase). The next process is the extraction of text features. This process is similar to that done in the text feature extraction stage of training. The process of summarizing text automatically based on models that have been created in the training stage. Selection of the sentence serves to generate a summary.

A novel technique for summarizing news articles using neural networks is presented in [7]. There are three phases in the process: neural network training, feature fusion, and sentence selection. The input to the neural network is either real or binary vectors. The first step involved training a neural network to recognize the selection of the summary sentences in terms of their importance by learning the relevant features of sentences that should be included in the summary of the article. It is trained on a corpus of articles.

The neural network is then modified to generalize and combine the relevant features apparent in summary sentences. Once the network has learned the features that must exist in summary sentences, it is needed to discover the trends and relationships among the features that are inherent in the majority of Sentences. This is accomplished by the feature fusion phase, which consists of 2 steps: Eliminating uncommon features & Collapsing the effects of common features.

Through feature fusion, the network discovers the importance (and unimportance) of various features used to determine the summary-worthiness of each sentence.

Finally, the modified neural network is used as a filter to summarize news articles.

In [8] the concept of OpenNLP tool for natural language processing of text for word matching is introduced. In order to extract meaningful and query dependent information from large set of offline documents, data mining document clustering algorithms are adopted. In this paper the framework is represented where user uploads the document files, and then paragraphs are obtained. Every paragraph is considered as node. The syntactic relationship among every node is measured using traditional edge weight formula. This structure is called as document graph where every node is a paragraph and every edge between two nodes represents the association (similarity) between two nodes, to which the edge is connecting to. The document graph is input to the clustering algorithm & after accepting a query from user, effective clustering algorithm K-Means is applied, which groups the nodes according to their edge weights and builds the clusters. Every node in the cluster maintains association with every other node.

III. METHODOLOGY

This paper proposes a new approach to multi-document summarization. The method ensures good coverage and avoids redundancy. It is the clustering based approach that groups first, the similar documents into clusters & then sentences from every document cluster are clustered into sentence clusters. And best scoring sentences from sentence clusters are selected in to the final summary.

We find similarity between each sentence & query. To find similarity "cosine similarity measure" is used.

Given two vectors of attributes, *A* and *B*, the cosine similarity, θ, is represented using a dot product and magnitude as:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

[9]

The important part is generation of vectors. As per above equation both vectors must be of same size because $\sum$ is from 1 to n for both sentence & query.

So we have merged sentence & query. Then we take each word from the merged sentence & check whether that word appears in sentence & query both. If yes then we have used the weight (tf*idf) of the word from document & placed that value in vector of sentence for the i^{th} location in vector, & term frequency of the term is placed in vector of query.

If the word appears either in sentence or query then the weight of the word is placed in appropriate vector & 0 is placed in the vector which doesn't contain the word.

After following these steps the above formula is used to calculate cos Θ.

A. Algorithm:

1. The user selected collection of documents & query is the input to the summarizer.
2. We have maintained a list of maps where each term from document collection is stored in a map with its number of occurrences. A map contains all the synonyms, of the term from the document collection. We have used WordNet dictionary to find synonyms.
3. Query modification technique is used. It works as follows:
   3.1. Split a query into tokens & find the synonym for each token. We will get the synonym from list of maps if the token or synonym exists in a document collection & append the most frequent synonym of the query term to query.
   3.2. We have generated a corpus for strengthening the query. The most frequently occurred words from corpus are selected & those words are appended to the query. So the query is strengthened.
4. Some features are used to calculate sentence score. The features are listed as follows : Noun Feature, Cue Phrase Feature, Sentence Length Feature, Numerical data Feature, Sentence

Position, Sentence centrality (similarity with other sentences), Upper Case word feature, Sentence similarity with user query, Term frequency & Inverse Document Frequency is also used to score the sentences.

5. The documents are clustered by using, cosine similarity as a similarity measure to generate the appropriate document clusters.
6. Then from every document cluster, sentences are clustered based on their similarity values.
7. Calculate the score of each group (sentence cluster).
8. Sort sentence clusters, in reverse order of group score.
9. Pick the best scored sentences from each sentence cluster and add it to the summary.
10. We have decided the number of sentences to be selected depending on sentence clusters size.

The representation of the method used is as shown in Fig 1.

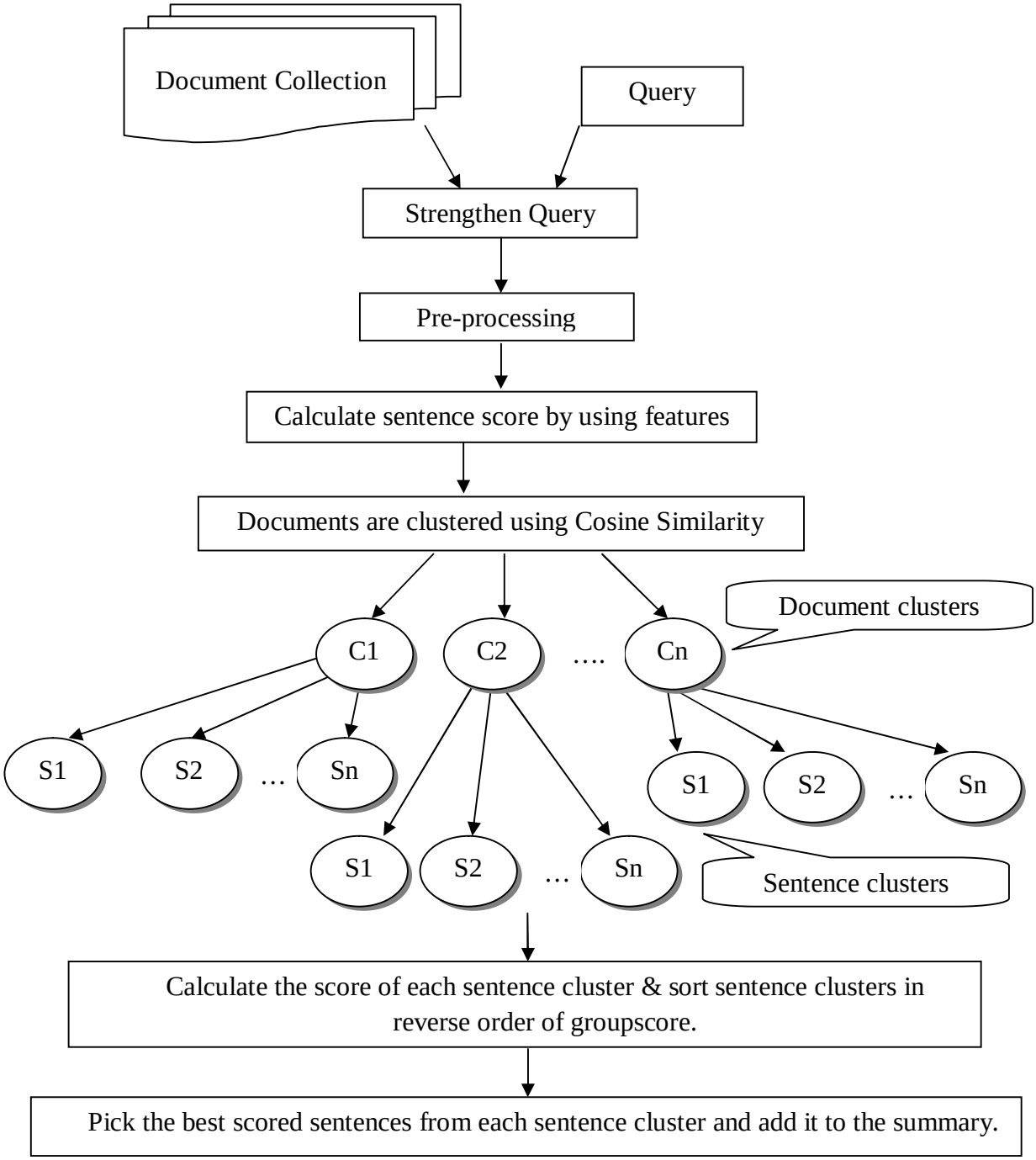

Fig. 1 Document & Sentence Clustering approach to summarization

IV. Experimental Results & Evaluation

We have compared the results generated by 3 methods which we have implemented as a part of our dissertation. And the results show that our novel method "Document & Sentence Clustering based Text Summarization" outperforms the other 2 methods. We have displayed the results obtained by performing summarization on different document collections in fig. 2 & fig 3.

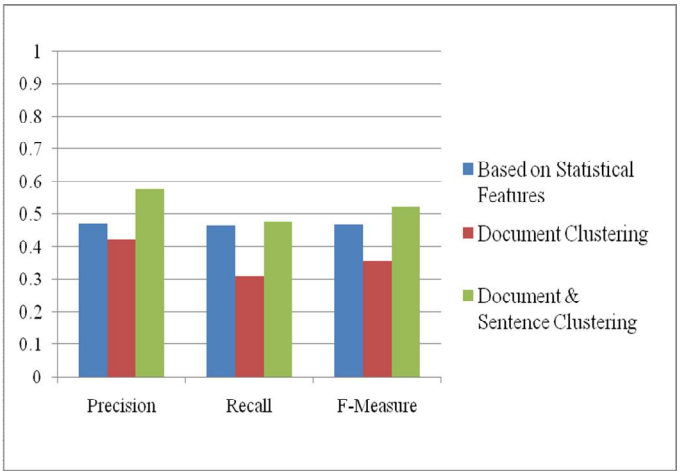

Fig. 2 Summarization performed on Document Collection D1.

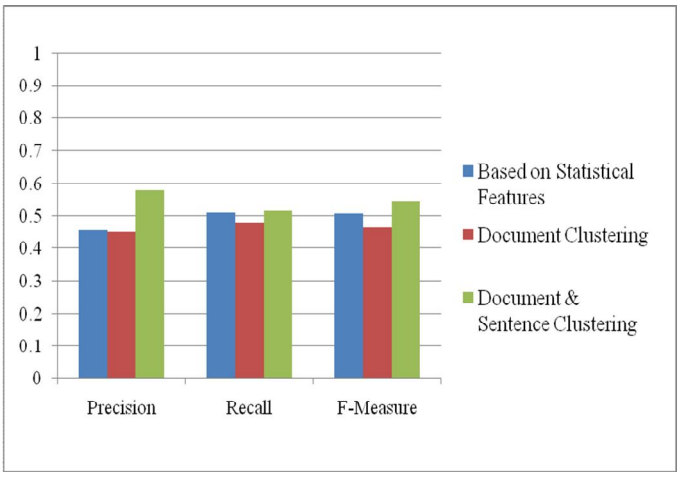

Fig 3 Summarization performed on Document Collection D2.

V. Conclusion

We have concentrated on extractive summarization techniques. We have compared the results developed by different extractive techniques by using Performance and correctness measures such as Precision, Recall and F-measure. As per results our novel method outperforms the other methods & it reduces redundancy due to clustering. In future, we would like to improve the system by adding sentence simplification technique for producing summary i.e. it can be used to simplify the sentences which are complex and very large. This approach can also be extended to multi lingual platform. We can also add paraphrasing technique to give abstractive feel to summary.

Acknowledgment

I would like to give deep gratitude to all those who have supported me to complete the implementations I have presented in this paper. A very special thanks goes to the Management, Principal, Head of department & Faculty of Walchand Institute of Technology, Solapur, for all their support & giving me the resources required to complete the development work.

References

[1] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, vol. 2, no. 3, Aug. 2010.

[2] "Multi-document summarization", *Wikipedia, the free encyclopedia*, 2012.

[3] A. Kogilavani, Dr.P.Balasubramani, "CLUSTERING AND FEATURE SPECIFIC SENTENCE EXTRACTION BASED SUMMARIZATION OF MULTIPLE DOCUMENTS", *International journal of computer science & information Technology*, vol.2, no.4, Aug. 2010.

[4] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", *TECHNIA – International Journal of Computing Science and Communication Technologies*, vol. 2, no. 1, Jul. 2009.

[5] A. P. Siva kumar, Dr. P. Premchand and Dr. A. Govardhan, "Query-Based Summarizer Based on Similarity of Sentences and Word Frequency", *International Journal of Data Mining & Knowledge Management Process*, vol.1, no.3, May 2011.

[6] Aristoteles, Yeni Herdiyeni, Ahmad Ridha and Julio Adisantoso, "Text Feature Weighting for Summarization of Document in Bahasa Indonesia Using Genetic Algorithm", *International Journal of Computer Science Issues*, vol. 9, no. 3, May 2012.

[7] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", *SECOND IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS*, June 2004.

[8] Harshal J. Jain, M. S. Bewoor and S. H. Patil, "Context Sensitive Text Summarization Using K Means Clustering Algorithm", *International Journal of Soft Computing and Engineering* ,volume-2, no.2, May 2012.

[9] "Cosine similarity", *Wikipedia, the free encyclopedia*, 2012.

Ms. Deshpande Anjali Ramkrishna received B. E. degree in Information Technology in 2009 from Solapur University, Solapur, India and persuing the M.E. degree in Computer Science & Engineering in Walchand Institute of Technology, Solapur, India.

Mr. Lobo L. M. R. J received the B. E. degree in Computer Engineering in 1989 from Shivaji University, Kolhapur, India and the M. Tech degree in Computer and Information Technology in 1997 from IIT, Kharagpur, India.