An Analytical Framework for Multi-Document Summarization

Jayabharathy[1], Kanmani[2] and Buvana[3]

**[1] Pondicherry Engineering College
Pondicherry-605014**

**[2] Pondicherry Engineering College
Pondicherry-605014**

**[3] Pondicherry Engineering College
Pondicherry-605014**

Abstract

Growth of information in the web leads to drastic increase in field of information retrieval. Information retrieval is the process of searching and extracting the required information from the web. The main purpose of the automated information retrieval system is to reduce the overload of document retrieval. Today's retrieval system presents vast information, which suffers from redundancy and irrelevance. There arises a need to provide high quality summary in order to allow the user to quickly locate the desired and concise information ase number of documents available on user's desktops and internet increases. This paper provides the complete survey, which gives a comparative study about the existing multi-Document summarization techniques. This study gives an overall view about the current research issues, recent methods for summarization, data set and metrics suitable for summarization. This frame work also investigates about the performance competence of the existing techniques.

Keywords:*Multi-Document Summarization, Generic Summary, Query Based Summary*.

1. Introduction

Document Summarization is an automated technique, which reduces the size of the documents and gives the outline and concise information about the given document. That is the summarization process extracts the most important content from the document. In general, the summaries are created in two ways. They are generic summary and query based summary. The generic summary refines overall content of the input document given by the user whereas the query based one retrieves the information that more relevant to the user query. Document summarizations are of two types, they are single document summarization and Multi-document summarization. The summary that is extracted and created from a single document is known as Single Document Summarization, whereas Multi-document Summarization is an automatic procedure for the extraction of information from multiple sources.

The purpose of a brief summary is to shorten the information search and to minimize the time by spotting the most relevant source documents. Widespread multi-document summary itself hold the required information, hence limiting the need for accessing original files to some cases when refinement is required. Automated summaries give the extracted information from multiple sources algorithmically.

The remainder of this paper is organized as follows: Section 2 provides the Classification of various summarization techniques and describes about the related works in field of generic based and query based summary generation. The general framework for extracting summary from documents sources and steps involved in this process of summary extraction are described in section 3. Section 4 gives the detailed discussion about the framework for analyzing existing summarization techniques. The paper is concluded with a brief discussion in section 5.

2. Classification of Summarization Techniques

This chapter gives an overview about various summarization techniques. The summarization techniques are classified into two major groups Generic and Query based summary creation. The generic summary refines overall content of the input document given by the user whereas the query based one retrieves the information that

is more relevant to the user query. The classification of multi-document summarization is shown in the figure 1. The brief description about each technique is stated below.

2.1 Generic Summary Extraction Techniques

The RANDOM based technique [9] is the simplest technique, which randomly selects lines from the input source documents. Depending upon the compression rate i.e. the size of the summary, the randomly selected lines will be included to the summary. In this technique, a random value between 0 and 1 is assigned to each sentence of the document. A threshold value for length of the sentence is provided in general. The score of 0 to 1 is assigned to all sentences that do not meet assigned length cut-off. Finally, required sentences are chosen according to assigned highest score for desired summary.

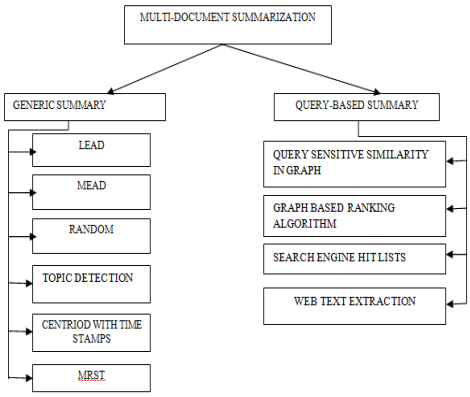

Fig.1 Classification of summarization techniques

LEAD based technique is one where first or first and last sentence of the paragraph are chosen depending upon the compression rate (CR) and it is suitable for news articles. It can be reasonable that n% sentences are chosen from beginning of the text e.g. selecting the first sentence in all the document, then the second sentence of each, etc. until the desired summary is constructed. This method is called LEAD [9] based method for summarization. In this technique a score of 1/n to each sentence is assigned, where n is the sentence number in the corresponding document file. This means that the first sentence in each document will have the same scores; the second sentence in each document will have the same scores, and so on. The length value is also provided as a threshold .The sentences with less length than the specified threshold value are thrown out.

MEAD is a commonly used technique which can perform many different summarization tasks. It can also summarize individual documents or clusters of related documents. MEAD is the combination of two baseline summarizers: lead-based and random based. Lead-based summaries are produced by selecting the first sentence of each document, then the second sentence of each, etc. until the desired summary size is met. A random summary consists of enough randomly selected sentences (from the cluster) to produce a summary of the desired size. MEAD is a centriod-based extractive summarizer that scores sentences based on sentence-level and inter-sentence features that indicate the quality of the sentence as a summary sentence .It then chooses the top-ranked sentences for inclusion in the output summary. MEAD extractive summaries score the sentences according to certain sentence features – Centriod [9], Position [9], and Length [9].

Dragomir R. Radev [1] et al proposed a multi-document text summarizer, called MEAD. The proposed system creates the summary based on cluster centroids. Centroid is the set of words that are most important to the cluster. In addition to the Centroid, position and first sentence overlap values are involved in the score calculation. Two new techniques namely cluster based relative utility and cross sentence information subsumption were applied to the evaluation of both single and multiple document summaries. Cluster base relative utility refers to the degree of relevance of a particular sentence to the general topic of the cluster. Summarization evaluation methods used could be divided into two categories: intrinsic and extrinsic. Intrinsic evaluation method measures the quality of multi-document summaries in a direct manner. Extrinsic evaluation methods measure how sucessfully the summaries help in performing a particular task. The extrinsic evaluation in terms called task-based evaluation. The new utility-based technique called CBSU was used for the evaluation of MEAD and of summarizers in general. It was found that MEAD produces summaries that are similar in quality to the ones produced by humans. MEAD's performance was compared to an alternative method, multi-document lead and showed how MEAD's sentence scoring weights can be modified to produce summaries significantly better than the alternatives.

Afnan Ullah Khan [3] et al proposed a new technique for information summarization, which is the combination of the rhetorical structure theory and MEAD summarizer. In general MEAD summarizer is totally based on mathematical calculation and lack a knowledge base. Rhetorical structure theory is used to overcome this weakness. The new summarizer system is evaluated against the original MEAD summarizer system. The proposed summarizer tool was exploited mainly in two areas of information that are Financial Articles and PubMed abstracts. The experimental results show that MEAD produces successful summaries 75% time for both

short and long documents whereas MRST produces successful summaries for short documents 70% of the time and long documents summaries 65% of the time, as the size of the document increases the performance of MRST deteriorates.

The two-stage sentence selection approach proposed by Zhang Shu [4] et al is based on deleting sentences in a candidate sentence set to generate summary. The two stages are (1) acquisition of a candidate sentence set and (2) the optimum selection of sentence. The candidate sentence set is obtained by redundancy-based sentence selection approach at the first stage where as in the second stage, optimum selection of sentences technique is used to delete sentences in the candidate sentence set according to its contribution to the whole set until desired summary length is met. With a test corpus, the ROUGE value obtained for the proposed approach proves its validity, compared to the traditional method of sentence selection. The influence of the chosen token in the two-stage sentence selection approach on the quality of the generated summaries is analysed. It differs from the traditional method of adding sentences to create summary by deleting the sentences in a set of candidate sentences to create the summary. With the test corpus used in DUC 2004, and compared to the redundancy based sentence selection, the experiments show that the two-stage sentence selection approach increases the ROUGE value of the summaries, which proves the validity of the proposed approach.

Dingding Wang [7] et al proposed a summarization system which is mainly based on sentence-level semantic analysis and non-negative matrix factorization. The sentence-sentence similarity is calculated by using the semantic analysis and the similarity matrix is constructed. Then the symmetric matrix factorization process is used to group the similar documents into clusters. The experimental result on DUC2005 and DUC2006 datasets achieves the higher performance.

Ben Hachey [8] proposed a generic relation extraction based summarization system. A GRE system builds the systems for relation identification and characterization which can be transferred across domains and tasks without any modification in model parameters. Relation identification is the extraction of relation forming entity mention pairs whereas relation characterization is the assignment of types of relation mentions. An experimental result shows that the proposed system's performance is slightly superior when compared to the existing system.

Md. Mohsin Ali [9] et al proposed two techniques for both single and multi document text summarization. The first technique is adding a new feature called SimWithFirst (Similarity with First Sentence) with MEAD (Combination of Centroid, Position, and Length Features) called CPSL and second is the combination of LEAD and CPSL called LESM. In general LEAD is the summarization technique in which first or first and last sentence of the paragraph are chosen depending upon the compression rate (CR). The results of proposed techniques are compared with conventional methods called MEAD with respect to some evaluation techniques. The results demonstrate that CPSL shows better performance for short summarization than MEAD and for remaining cases it is almost similar to MEAD and LESM also shows better performance for short summarization than MEAD but for remaining cases it does not show better performance than MEAD.

Shu Gong [11] et al proposed a Subtopic-based Multi-documents Summarization (SubTMS) method. This method adopts probabilistic topic model to find out the subtopic information inside each and every sentence and uses a hierarchical subtopic structure to explain both the whole documents collection and all sentences inside it. here the sentences represented as subtopic vectors, it assess the semantic distances of sentences from the documents collection's main subtopics and selects sentences which have short distance as the final summary. They have found that, training a topic's documents collection with some other topics' documents collections as background knowledge, this approach achieves fairly better ROUGH scores compared to other peer systems in the experimental results on DUC2007 dataset.

A.Kogilavani [12] et al proposed an approach to cluster multiple documents by using document clustering approach and to produce cluster wise summary based on feature profile oriented sentence extraction strategy. Most similar documents are grouped into same cluster using document clustering algorithm. Feature profile is generated which mainly includes the word weight, sentence position, sentence length, and sentence centrality, proper nouns in the sentence and numerical data in the sentence. Based on this feature profile sentence score is calculated for each and every sentence in the cluster of similar documents. According to different compression ratio sentences are extracted from each cluster and ranked. Then the sentences are extracted and included in the summary. Extracted sentences are arranged in chronological order as in input documents and with the help of this, cluster wise summary will be generated. An experimental result shows that the proposed clustering algorithm is efficient and feature profile is used to extract most important sentences from multiple documents. The summary generated using the proposed method is compared with human summary created manually 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.

2.2 Query Based Summary Techniques

Dragomir R. Radev [2] et al designed a prototype system called SNS, which is pronounced as “essence”. This mainly integrates natural language processing and information retrieval techniques in order to perform automatic customized summarization of search engine results. The proposed system actually retrieves documents related to an unrestricted user query and summarizes a subset of them as selected by the user Task-based extrinsic evaluation showed that the system is of reasonably high quality.

Furu [5] et al proposed a graph based query oriented summarization based on query sensitive similarity measure. For the evaluation of sentence-sentence edges the similarity measure incorporates the query influence technique. Graph modeling and graph based ranking algorithm is used for finding the similarity between the sentences. Then sentences which are more similar to the user query will be retrieved. The experimental results on DUC 2005 shows that it improves ROUGH score.
Xiao [6] et al designed and proposed a system to automate the multi-document summarization. The proposed system retrieves the documents related to the query given by the user. The sentence score is calculated based on relevant value and in-formativeness value. These values are realized by word sentence overlap and semantic graph techniques. Then the sentences with the highest score are included to the summary. The investigational result shows that the proposed system achieves better quality.

Lei Huang [10] et al considers document summarization as a multi-objective optimization problem involving four objective functions, namely information coverage, significance, redundancy and text coherence. These functions measure the possible summaries based on the identified core terms and main topics (i.e. a cluster of semantically or statistically related core terms). The datasets namely DUC 2005 and 2006 have been chosen for query-oriented summarization tasks to test the proposed model. The experimental results indicate that the multi-objective optimization based framework for document summarization is truly a promising research direction. It is valuable to note that a real optimization based summarization method is different from the existing non-optimization based methods in two noteworthy aspects. First, it ranks summaries instead of ranking individual sentences. Second, though ignored in the previous literature, the approach to rank summaries should not directly rely on the approach to rank sentences. Otherwise, the optimization solutions will degenerate to the traditional non-optimization based (e.g. MMR like) methods.

3. General Procedure for Document Summarization

Usually document sources are of unstructured format, transforming these unstructured documents to structured format requires some pre-processing steps. Fig.1 presents the sequence of steps involved in document Summarization. Some commonly used pre-processing steps are

Sentence Decomposition: The given input document is decomposed into sentences.

Stop words removal: Stop words are typical frequently occurring words that have little or no discriminating power, such as \a", \about", \all", etc., or other domain-dependent words. Stop words are often removed.

Stemming: Removes the affixes in the words and produces the root word known as the stem [13]. Typically, the stemming process is performed so that the words are transformed into their root form. For example connected, connecting and connection would be transformed into ‘connect’. Most widely used stemming algorithms are Porter [17], Paice stemmer [16], Lovins [15], S-removal [14]

Feature Vector Construction: Feature vector is constructed based on term frequency (TF-DF) and inverse document frequency (TF-IDF).

After applying the preprocessing techniques, the processed documents are clustered using a clustering algorithm in order to group the similar documents. Cluster analysis or clustering is the assignments of a set of observations into subsets (called clusters) so that observations of same cluster are similar in some sense. Some of the famous types of clustering are described below.

Hierarchical algorithms find consecutive clusters using previous clusters. They are of two types namely agglomerative ("bottom-up") and divisive ("top-down"). The first type begins with each element as a individual cluster and merge them into larger clusters. Divisive algorithms start with the whole document set and divide it into smaller clusters.

Partitional algorithms typically resolve all clusters at once, but can be used as divisive algorithms in the hierarchical clustering.

After the clustering process the summary is created for the clustered documents. We have discussed the variety of summary creation techniques in the previous section.

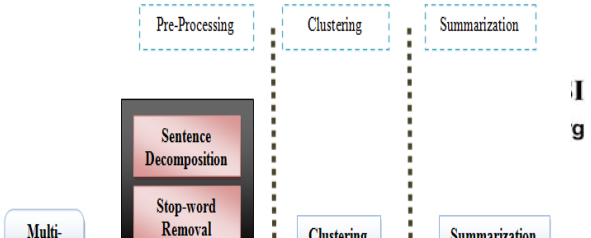

Fig. 2 General Procedure for Document Summarization

4. A FRAMEWORK FOR ANALYZING DOCUMENT SUMMARIZATION

This study mainly highlights the recent research work in the field of multi-document summarization. This paper primarily focuses about the proposed frame work for comparing various multi-document summarization techniques. The comparative study is based on the survey, which is made by analyzing the existing algorithms, considering the characteristic factors like Document summarization technique, Data set used for experiments, Performance metrics and detail about the performance of the proposed technique. Column one in this table presents the title of the related papers. The Frame work, algorithm and techniques which are discussed in the existing papers are stated in column two. The third column gives the details of the data set which are considered for conducting the experiments. Metrics considered by the authors for performance evaluation are given in column four. The concise details about the performance of the proposed techniques are listed in column five.

Table 1: Comparision Of Existing Summarization Techniques

Paper Title	Algorithm/ Technique	DataSet	Evaluation tool/Metric	Performance
Centriod based summarization of multiple documents 2003	Mead extraction algorithm	News articles	Utility based evaluation, User studies and System evaluation	Utility is very high
Automatic summarization of search engine hit lists	Centriod, Position and First sentence overlap	Global E-commerce Framework	Time, Reliability	Better Speedup in reading time, Better Reliability
MRST: a new technique FOR Information Summarization,2005	MRST	Financial Articles and pubMed abstracts	Coherence, Correctness, Compression, Overall	Existing technique such as Mead comes out more successful when compared to MRST
Two stage sentence selection approach for multi document summarization-2008	Redundancy based sentence selection	DUC2004	ROUGH	Increased ROUGH score, Proves Validity
A Query-Sensitive Graph-Based Sentence Ranking Algorithm for Query-Oriented	Graph Modeling, Graph-Based Ranking Algorithm	DUC2005	ROUGH-1, ROUGH-2, ROUGH-SU4	4.9% improvement in ROUGH-2
Multi-Document Summarization via Sentence – Level Semantic Analysis and	Semantic similarity matrix construction, Symmetric Non-negative Matrix Factorization and kernel K-means clustering	DUC2005, DUC2006	ROUGH-1,ROUGH-2,ROUGH-N(n-gram recall) ROUGH-L,ROUGH-W(ROUGH-SU(ski	Better ROUGH Scores

Symmetric Matrix Facorrization-2008			p-bigram plus unigram)	
Multi-Document Summarization Using generic Relation Extraction	GRE	DUC2001	ROUGH-SU4	Maximum ROUGH Score of 0.396 is obtained
Multi-document Text Summarization : SimWithFirst Based Features and Sentence Co-selection Based Evaluation-2009	CPSL, LESM	DUC2004	Precision, Recall, Kappa Coefficient, Cross Judge Utility agreement	Better Performance for short Summaries
Modeling Document Summarization as Multi-objective Optimization-2010	Summary Ranking	DUC2005 &2006	coverage, significance, redundancy and text coherence	Produces Optimized summary
Subtopic-based Multi-documents Summarization-2010	Subtopic vector construction and semantic distance calculation	DUC2007	ROUGH	Better ROUGH Scores

This frame work precisely states the details about the algorithms, data sets, metrics and performance. From the analysis it is understood that majority of the researchers concentrate on multi- document summarization. During the earlier stage most of the researchers concentrate on single document summarization. Multi document summarization came in picture from 2000 onwards. At earlier days the position of the sentences are considered to be important and have included to the summary like including title sentences, sentences at the mid of the paragraph etc. This method is suitable for documents which are related to news documents. But recent researchers not only concentrate on position they also give importance to the semantics of the sentences and their significance are identified and then it is added to the summary. Most of the researchers compare their proposed work with human generated summaries and justifies their work. From the survey it is concluded that MEAD is the most popular tool for Document summarization. Precision, Recall, Kappa Coefficient, F-Measure, etc are metrics used for evaluating the generated summary.

Rough score gives the measurement of sentence relevance. The Rough score are used by majority of researchers in association with DUC dataset for evaluating the quality of generated summary. In addition to that some of the document summarization uses the news articles and financial articles as the dataset. Some summarization technique ranks the sentences according the factor like position, semantic, number of nouns, length etc are included to the summary. Compression rate is considered to be one more factor for summary generation. Generic summary generation draws the attention of many researchers.

5. CONCLUSION

In this paper a frame work for analyzing existing document summarization algorithms was proposed. This framework gives the brief overview of recent research work on various algorithms in document summarization technology. Some inferences from the analytical frame work were also discussed. This gives the clear idea about the ongoing field of research in summarization. Document Summarization still has a scope in summarization in Distributed Environment and in Dynamic Multi-Document Summarization or update summarization. Automatic evaluation methods for document summarization are still

an ongoing research process. Redundancy elimination in generated summary is also an attractive area of research.

6. REFERENCES

[1] Dragomir R.Radev, Hongyan, Malgorzata Stys and Danial Tam, "Centriod- based summarization of multiple documents", Information Processing and Management, 2004.

[2] D.R.Radev, Weiguo Fan, "Automatic summarization of search engine hit lists ", *University ofMichigan Business School*.

[3] Afnan Ullah Khan, Shahzad khan and Waqar Mahmood, "MRST:A NewTechnique for Information Summarization" *World Academy of Science Engineering and Technology*, 2005.

[4] Zhang Shu ,Zhao Tiejun, Zheng Dequan& Zhao Hua , "Two stage sentence selection approach for multi-Document summarization", *Journal of electronics*, Vol.2, No.4, July 2008.

[5] Furu Wei,YanXiang He ,Wenjie Li and Qin Lu, "A Query-Sensitive Graph-Based Sentence Ranking Algorithm for Query-Oriented Multi-Document Summarization", *International Symposiums on Information Processing*, 2008.

[6] Xiao-Peng Yang and Xiao-Rong Liu," Personalized Multi-Document Summarization in Information Retrieval", *Seventh International Conference on Machine Learning and Cybernetics, Kunming*, 12-15 July2008.

[7] Dingding Wang, Tao Li, Shenghou Zhu, Chris Ding, "Multi-Document Summarization via Sentence –Level Semantic Analysis and Symmetric Matrix Factorization", *SIGIR Singapore*, July 20-24, 2008.

[8] Ben Hachey, "Multi-Document Summarization Using Generic Relation Extraction", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 420-429, 2009.

[9] Md. Mohsin Ali, Monotosh Kumar Ghosh, and Abdullah-Al-Mamun," Multi- document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation", *International Conference on Future Computer and Communication*, 2009.

[10] Lei Huang, Yanxiang He, Furu Wei, and Wenjie Li, "Modeling Document Summarization as Multi-objective Optimization", Third *International Symposium on Intelligent Information Technology and Security Informatics*, 2010.

[11] Shu Gong, Youli Qu and Shengfeng Tian, "Subtopic-based Multi-documents Summarization", *Third International Joint Conference on Computational Science and Optimization*, 2010.

[12] A.Kogilavani and Dr.P.Balasubramani, "Clustering and Feature Specific Sentence Extraction Based Summarization of Multiple Documents", *International Journal of computer science & information Technology (IJCSIT)* Vol.2, No.4, August 2010.

[13] WB Frakes, CJ Fox, "Strength and Similarity of Affix Removal Stemming Algorithms", *ACM SIGIR Forum*, 2003.

[14] Harman, D. "How Effective is Suffixing." *Journal of the American Society for Information Science 42 (1)*, 1991, 7-15.

[15] Lovins, J. B. "Development of a Stemming Algorithm", *Mechanical Translation and Computational Linguistics11*, 1968, 22-31.

[16] Paice, Chris D. "Another Stemmer.", *SIGIRForum 24 (3)*, 1990, 56-61.

[17] Porter, M. F. "An Algorithm for Suffix Stripping."*Program 14*, 1980, 130-137.

[18] Fung B, Wnag K & Ester .M, "Hierarchical Document Clustering using Frequent itemsets", *SIAM International Conference on Data Mining*, SDM '03.2003. Pp 59-70

J.Jayabharathy received her M.Tech in 1999 from Department of Computer Science and Engineering , Pondicherry University, Puducherry. She has been working as a Assistant Professor in the Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry. Currently she is working towards the Ph.D degree in Document clustering. Her areas of interest are Distributed Computing, Grid Computing, Data Mining and Document Clustering.

Dr. S. Kanmani received her B.E and M.E in Computer Science and Engineering from Bharathiyar University and Ph.D in Anna University, Chennai. She had been the faculty of Department of Computer Science and Engineering, Pondicherry Engineering College from 1992 onwards. Presently she is working as Professor in the Department of Information Technology, Pondicherry Engineering College. Her research interests are Software Engineering, Software testing, Object oriented system, and Data Mining. She is the Member of Computer Society of India, ISTE and Institute of Engineers, India. She has published about 65 papers in various International conferences and journals.

Miss Buvana Received her B.Tech(2005) in Computer Science and Engineering from Pondicheery University and Currently Doing her M.tech in Pondicherry Engineering College.