International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 08, August-2015

Literature Survey on Topic Focused Multi-Document Summarization

R. Narmatha M.E.CSE
Department of CSE,
SNS College of Technology,
Coimbatore, India.

S. Ranjitha M.E.CSE
Department of CSE,
SNS College of Technology,
Coimbatore, India.

V. Praveena M.E, (Ph.D)
Department of CSE,
SNS College of Technology,
Coimbatore, India.

***Abstract*-Multi-document summarization is the way of generating summaries that are highly related to human generated summaries from multiple documents that are retaining the most important characteristics of the original document. It's a process of summarizing a text by computer where a text is given to the computer as input and the output is a shorter and less redundant. There are two approaches in multi document summarization: abstractive based summarization and extractive based summarization. In this paper, we compare various techniques done for multi-document summarization. So that in future one can gets significant instruction for further analysis.**

Keywords-Abstractive Method, Extractive Method, Cluster Based, Rank Based, Time Based, Graph Based Approach.

I. INTRODUCTION

The huge amount of documents available on the Internet brings the difficulty of finding out whether a single document can meet a user's need. To solve this difficulty multi-document summarization [9], [2], reduce the length of a collection of documents. There is huge amount of data available in structured and unstructured form and it is difficult to read all data or information. It is a need to get information within less time. Hence we need a system that automatically retrieves and summarize the documents as per user need in time limit.

Document Summarizer is one of the feasible solutions to this problem. Summarizer is a tool used to get the information in an efficient way from data sets. Summarization is the process of extracting the important content from the original document. The document summarization is a difficult task to build a summary from multiple documents. In general, the summaries are defined in two ways. They are Single Document Summarization and Multiple Document Summarization. The summary which is extracted and created from single document is called as Single Document Summarization whereas Multiple Document Summarization is a method for the extracting and generating a summary from multiple documents. There are many issues in multi-document summarization when compared to the single document summarization. Issues are redundancy elimination, passage selection, formulation of summary.

Text summarization can be categorized into two approaches: extractive and abstractive. Extractive summarization method is used to select the most important sentence based on some ranking strategy. Abstractive summarization may create a new sentence, undetected in the original sources. Abstractive approaches require deep NLP such as semantic information, inference and natural language generation.

Extractive multi-document summarization is a method of extract and creates a summary from multiple documents about the same topic or different topic. This survey covers Term Frequency based approach, Cluster based approach, Time based approach, Graph based approach and Rank based approach.

II. REVIEW ON TOPIC FOCUSED MULTI-DOCUMENT SUMMARIZATION

Many techn9iques have been implemented for multi-document summarization. In this literature we focus on different type of techniques.

A. Time Based Approach

Time based approach is a enhancement of Graph based ranking approach.

TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization in 2007 [13], proposes a TimedTextRank algorithm based on graph based rank approach. Latest documents are more important than the earlier documents. This algorithm uses the temporal dimension information of document. This is based on the relationship between the two sentences and chooses the sentence according to the vote or cost or recommendation.

B. Term Frequency Based Approach

Multi-Document Summarization Using Document Set Type Classification in 2004 [7], to propose a multiple document summarization technique that applies simple strategy to generate a summary by using TF*IDF based sentence extraction for single document summarization and

use of single document summarization for multiple documents.

TF= (No. of occurrences of a word in a file/Total no. of unique words in that file)
IDF= (1+log (total no. of docs/no. of docs with given term)

This system automatically classifies the document set into three types:

i. One type: the first document will describe the one topic or event and following document will also describe same topic or event
ii. Multi type: set of document will describe the same topic or same event. For example first document will describe the new printer of one company and next document will describe the new printer of some other company
iii. Other type: set of documents are related to each other

This technique does not work well. Because of some bugs in classification of document sets.

C. Cluster Based Approach

Multi-document summarization by sentence extraction in 2000 [5], proposed a statistical method for generating multi-document summaries based on extractive based approach. In previous work for single document summarization differ from multi-document summarization. In multi document summarization there is a need of eliminating redundant information from multiple documents and attain better compression ratio. The steps involve in statistical methods are:

- *Clustering:* groups of similar objects such that the objects in a group will be related to one another and unrelated to the objects in other groups
- *Coverage:* to find and extract the most important points from source document
- *Anti-redundancy:* to minimize the redundancy in the summary
- *Summary generation:* by using cluster based and rank based algorithm to generate a summary

This algorithm will minimize the redundancy; maximize the diversity in the passage selection and do not generate coherent summary (should be readable and relevant one).

Multi-Document Summarization Using Cluster-based Link Analysis in 2008 [12], presents the two methods namely Cluster-based Conditional Markov Random Walk Model (ClusterCMRW) and the Cluster-based HITS Model (ClusterHITS).

In ClusterCMRW uses a two level layer relationship (cluster level information into a graph link).Upper layer represents a theme cluster and Lower layer represents a link between the sentences; dashed line between the upper and lower layer represents the relationship between the sentences and theme clusters.

ClusterHITS uses the sentence to cluster relationship. Sentences are represented as a authority in lower layer and theme clusters are represented as a hubs in upper layer. Calculate the hub score and authority score in a reinforcement way.

This approach gives better effectiveness and more robustness. But do not guarantee the quality of the cluster.

Towards Multi-Document Summarization of Scientific Articles: Making Interesting Comparisons with SciSumm in 2011 [10], proposed a G-Flow algorithm for joining the Sentence selection and sentence re-ordering methods. It is used to generate a summary in query oriented manner.

Steps in G-Flow algorithm: pre-processing, clustering, ranking, summary generation. This algorithm generates a coherent summary by joining the sentence selection and re-ordering with salience information.

D. Graph Based Approach

Multi-Document Summarization by Graph Search and Matching in 1997 [4], proposed a Spreading activation algorithm for summarizing similar and differences in a related document sets by using graph based approach. Spreading activation algorithm is used to determine the nodes that are related to core theme in each document. Activated graphs in each document are then matched based on similarity or difference between the pair of documents.

E. Rank Based Approach

Ranking through Clustering: An Integrated Approach to Multi-Document summarization in 2013 [11], proposes a novel based approach. In existing cluster based approach to apply a clustering approach and ranking approach in isolation manner which lead to incomplete result. There are three methods to be proposed for this work:

- Definition of basic ranking functions: Three ranking functions are there:Local Ranking, Global Ranking, Conditional Ranking
- Reinforcement Between Within-Cluster Ranking and Clustering: This module has three subdivisions:

✓ A sentence mixture model considers term conditional rank distributions across K theme clusters to establish a relationship between a sentence and a term set.
✓ The EM (Expectation Maximization) Algorithm is used to estimate the parameter matrix Θ repetitively until it converges to local maximum. This generates the component coefficients of the mixture model.
✓ The similarity measure between a sentence and a cluster is calculated as the cosine similarity between them, which is used to adjust the clusters that the

sentences belong to and in turn modify within-cluster ranking for the sentences in the updated clusters.

- Ensemble Ranking and Summary Generation:Ensemble ranking function (f) of sentence is defined as an ensemble of all the sentence ranking score on the K clusters.

There are three models can be implemented:

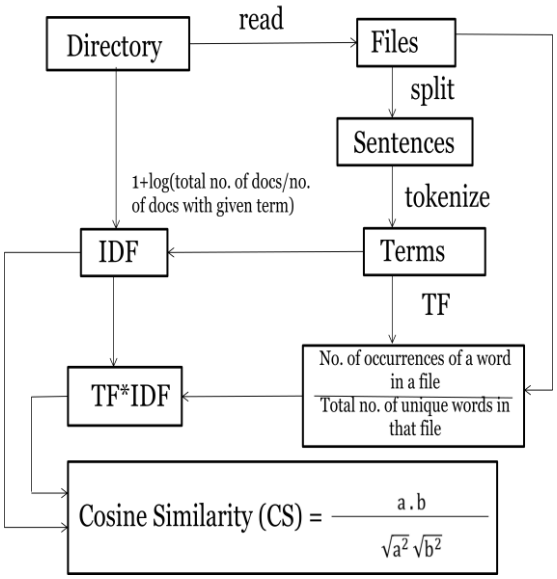
- Calculation of Cosine Similarity (CS)

The formula to calculate Cosine Similarity (CS) is given by:

$$CS = \cos\theta = \frac{a.b}{\sqrt{a^2}\sqrt{b^2}}$$

where a and b are Term Frequency (TF) vectors of two sentences

Block Diagram for calculating a cosine similarity between the two sentences



- Ranking and Clustering

✓ The procedure to generate a rank based on the adjacency matrix and similarity values is given as follows:

1.Initialise SV to 0 and score to 1.

2. For i = 0 to n do

For j = 0 to n do

SV += score*CS(i,j)

score = (α * SV) + [(1-α) / n]

Where SV= sentence value, α = 0.85 is the damping factor

- Summary Generation: In Multi-document summarization to summarize very large amount of documents. In single document summarization there is no problem of redundancy elimination. In case of multi-document summarization redundancy elimination becomes big issue. So, necessary to eliminate the redundancy. Eliminating the redundancy we are using Maximal Marginal Relevance (MMR) [6] and clustering [8], [3].

To form a cluster based on theme or query. Apply a ranking function to the cluster and choose the first sentence into the summary. Then go to next sentence and compare with sentences already in summary. Not similar sentences to be added to summary (cosine similarity between the sentences).

III.CONCLUSION AND FUTURE WORK

In this paper, examine the recent technology in document summarization and NLP is used to generate a coherent summary that are highly related to human generated summary. We reported a literature survey for extractive multi-document summarization based on theme or topics. Extractive multi-document is a selecting a highly scored sentence by using some ranking strategy from multiple documents. After generating a summary the information will not be a duplicate and less redundancy and maximum diversity in formation of summary.

In future, we consider the semantic role information of the sentence. Also using graph based dependency parser in Mate tool for parsing the sentences in each document. And also eliminate the redundant information by using Maximal Marginal Relevance (MMR) and clustering algorithm.

REFFERNCES

[1] Su Yan,Xiaojun Wan, "SRRank:Leveraging Semantic Roles for Extractive Multi-Document Summarization," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, december 2014.

[2] C. D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[3] D. R. Radev, H. Y. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," Inform Process Manag, vol. 40,no. 6, pp. 919–938, 2004.

[4] Inderjeet Mani and Eric Bloedorn, "Multi-document summarization by graph search and matching," AAAI/IAAI, vol. cmplg/9712004, pp. 622-628, 1997.

[5] Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell, "Multi-Document Summarization by Sentence Extraction," ANLP/NAACL Workshops. Association for Computational Linguistics, New Jersey, pp. 40-48,2000.

[6] J. G. Corbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in Proc. 21st SIGIR Conf., pp. 335–336, 1998.

[7] Jun'ichi Fukumoto, "Multi-Document Summarization Using Document Set Type Classification," Proceedings of NTCIR-4,Tokyo, pp. 412-416, 2004.

[8] K. R. Mckeown, J. L. Kalvans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, "Towards multi-document summarization by reformulation: Progress and prospects," in Proc. 13th AAAI Conf., pp. 121–128, 1999.

[9] K. S. Jones, “Automatic summarising: The state of the art,” Inf. Process Manag., vol. 43, no. 6, pp. 1449–1481, 2007.
[10] Nitin Agarwal, Gvr Kiran, Ravi Shankar Reddy and Carolyn Penstein Rose,“Towards Multi-Document Summarization of Scientific Articles:Making Interesting Comparisons with SciSumm,” Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, pp. 8–15, 2011.
[11] Xiaoyan Cai and Wenjie Li, ”Ranking Through Clustering: An Integrated Approach to Multi-Document Summarization,” IEEE Transactions on audio, speech, and language processing, vol. 21, no. 7, july 2013.
[12] Xiaojun Wan and Jianwu Yang, “Multi-Document Summarization Using Cluster-based Link Analysis,” Proceedings of ACM SIGIR conference on Research and development in information retrieval, Singapore, pp.299-306, 2008.
[13] Xiaojun Wan, “TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization,” Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, pp. 867- 868, 2007.