Generating Multi-Document Summarization using Data Merging Technique

G.V. Garje, PhD S.V. Khaladkar A.N. Khengare J.M. Pawar M.S. Vidhate
Department of Computer Engineering, PVG's College of Engineering and Technology, Savitribai Phule Pune University, Pune, Maharashtra, India.

ABSTRACT
In this paper we propose a summarization method to summarize a set of co-referent documents that has been clustered using hard clustering techniques. The main focus of this paper lies on the weighted optimal merge function (f_β). This weighted optimal merge function uses the weighted harmonic mean to find the balance between precision and recall. The global precision and recall measures are defined by triangular norm. Triangular norm receives local precision and recall values as input, in order to generate a multi-set of key concepts that we use to generate summarization.

General Terms
Text summarization, Data mining

Keywords
Content Selection, Weighted optimal merge function, Multi-document summarization.

1. INTRODUCTION
Due to fast growth and wide spread of World Wide Web the data is getting generated in exponential proportion. With an estimated amount of about 50 billion web pages indexed on Google. Today, it is quite difficult (sometimes impossible) to get or find the desired information without the use of some kind of software tool. That is why it gradually became more common to summarize textual data into briefer versions that will still reflect the relevant information using automated summarization tools.

Text summarization is a mechanism which deals with the compression of large document into shorter version of text. Text summarizations choose the most important part of text and create coherent summaries that state the main purpose of the given document.

Multi-document summarization problem can be seen as general data merging problem. When we deal with texts, database records or aggregation the issue of merging data always arises. When we try to generate a multi document summarization, we enable the possibility to use data merging techniques. We will be working with merge function.

The main focus of this paper is a merge function that tries to optimize the weighted harmonic mean between correctness and completeness of the suggested solution with respect to the source. Both the correctness and completeness are calculated by using triangular norms.

2. EXISTING SYSTEMS
Jade Goldstein, Vibhu Mittal, Jaime Carbonell, Mark Kantrowitz (2000) proposed a Multi-Document summarization system. This summarization system uses sentence extraction approach for multi-document summarization which is built on a single document summarization method. It uses additional available information about the document set as a whole and the relationships between the documents. It uses domain independent techniques based mainly on statistical processing and a metric (for reducing redundancy and maximizing diversity in the selected passages). It attempts to maximize the novelty of the information being selected, and different genres or corpora characteristics can be taken into account easily.

Jun ichi Fukumoto (2004) proposed a summarization system which automatically classifies type of document set and summarizes a document set with its appropriate summarization mechanism. This system classifies a document set into three types, a series of events, a set of the same events and related events, by using information of high frequency nouns and named entities. The unnecessary parts are deleted from each summarized document and generates multi document summary. Author has used single document summarization mechanism for each document of a document set and removed similar parts between summarized documents for generation of a target summary. They applied a Term Frequency (TF)/(Inverse Document Frequency (IDF) based sentence extraction approach for single document summarization and use of single document summarization for multi-document summarization.

Summarization system proposed by Judith D. Schlesinger, Dianne P. O'Leary and John M. Conroy (2008) used CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) system architecture to summarize document. CLASSY is an automatic, extract generating summarization system that uses linguistic trimming and statistical methods to generate generic or topic/query-driven summaries for single document or clusters of documents. CLASSY used trimming rules to shorten sentences, identify sentences, select sentences and organize the selected sentences for the final summary.

3. PROPOSED SYSTEM
The focus of our idea is on merging co-referent items. Co-referent items is a set of documents related to the same topic that one wants to summarize which are ready to be merged in the data merging problem. A document is decomposed into a multi-set of concepts. After decomposition of the documents into multi-set of concepts a weighted optimal merge function is applied. The multi-set of concepts thus obtained is considered as a set of key concepts. For summary generation a basic adaptation of the NEWSUM algorithm is introduced. It is a summarization technique that uses sentence extraction approach in order to generate summarizations.

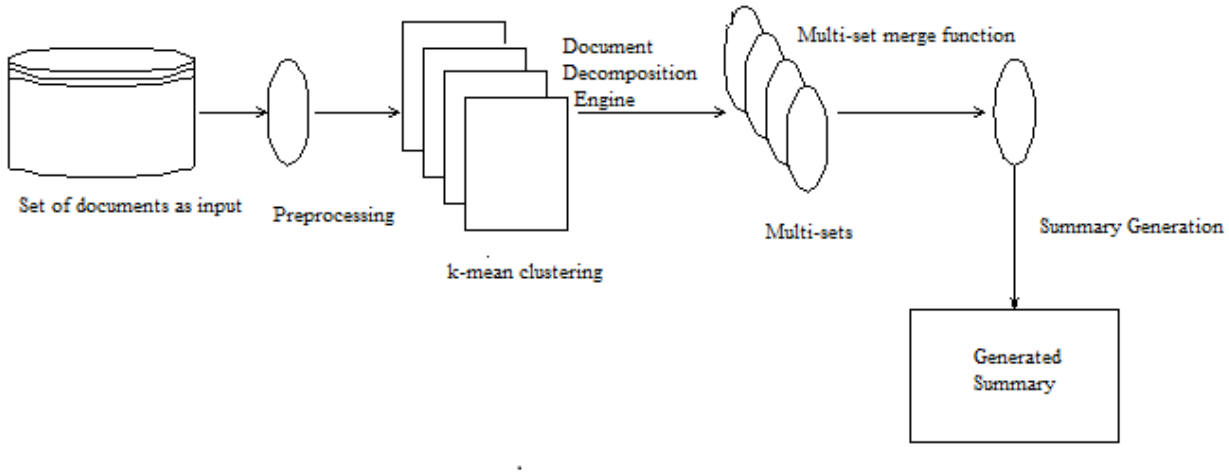

Fig.1 System Architecture

The proposed system consisting of following modules as depicted in Fig.1:
1. Preprocessor
2. Clustering
3. Merging
4 .Summary generator

3.1 Preprocessor

1. Segmentation: It is a process of dividing a given document into sentences.
2. Removal of Stop Words: Stop Words are frequently occurring words such as 'a' an', the' that provides less meaning and considered as a noise in the sentence.
3. Tokenization: breaks a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.
4. Word Stemming: converts every word into its root form by removing its prefix and suffix.

3.2 Clustering

In this phase clustering is done using k-means clustering algorithm which work based on TF/IDF and cosine similarity (this metric is used to determine similarity between two documents). The Term Frequency is the number of times a word occurs in a document (stop-words have been eliminated earlier itself and will not figure in this calculation). Inverse Document Frequency is the number of documents in the document set which contains that word.

Then based on cosine similarity document set is clustered into various clusters.

3.3 Merging

This phase works with weighted optimal merge function. Important keyword selection is done in this phase. This function uses the 'Local Precision, Local Recall' to select the words with high importance.

Weighted optimal merge function:

$$\varpi^*(M) = \arg\max_{\mathscr{S} \in \mathcal{M}(U)} f_\beta(\mathscr{S}|M)$$

$$= \arg\max_{\mathscr{S} \in \mathcal{M}(U)} \left(\frac{(1+\beta^2) \cdot p(\mathscr{S}|M) \cdot r(\mathscr{S}|M)}{\beta^2 \cdot p(\mathscr{S}|M) + r(\mathscr{S}|M)} \right)$$

When β <1 a preference is given to precision. When β >1 a preference is given to recall.

3.4 Summary Generator

Basic adaptation of the NEWSUM algorithm (a summarization technique) will be applied that uses sentence extraction approach in order to generate summarizations.

NEWSUM Algorithm

```
SUMMARIZER(Cluster, char *K[])
    {
    while (size_of (K) != 0)
       {
                Rate all sentences in Cluster by key concepts K
Select sentence 's' with highest score and add to final summary (S)
          }
Return(S)
 }
```

4. CONCLUSION

It has been observed from the literature review that multi-document summarization involves generating summary from multiple documents which will be readable for user. The system will make use of preprocessing techniques like stop-word removal & stemming as well as k-means algorithm for clustering, weighted optimal merge function & NEWSUM algorithm to generate summary of better quality. The proposed system can produce better quality summary. Sometimes there may be loss of important information but still our system can provide an abstract understanding of particular concept from the summary.

5. ACKNOWLEDGMENTS

We would like to take this opportunity to express our profound gratitude and deep regard to Dr.G. V. Garje for his exemplary guidance. His valuable suggestions were of immense help throughout our project work.

6. REFERENCES

[1] D.Van Britson, A. Bronselaer and G. De. Tre, (2015) "Using data merging techniques for generating multi-document summarizations", in IEEE Transactions on Fuzzy System, Vol - 23, NO.3, pp. 576 – 592.

[2] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, Mark Kantrowitz, (2000) "Multi Document Summarization by Sentence Extraction", in proceedings of NAACL-ANLP Workshop on Automatic summarization, Volume 4, pp.40-48.

[3] Jun ichi Fukumoto, (2004) "Multi-Summarization using document set type classification", in Proceedings of NTCIR- 4, Tokyo.

[4] Judith D. Schlesinger, Dianne P. O'Leary, and John M. Conroy, (2008) "Arabic/English Multi-document Summarization with CLASSY-The Past and the Future" in the proceedings of Computational Linguistics and Intelligent Text Processing, 9th International Conference, Haifa, Israel, Springer, pp 568–581.

[5] Pallavi D. Patil and N. J. Kulkarni, (2014) "Text Summarization using Fuzzy-Logic", in International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 3.

[6] Fatma El-Ghanna and Tarek El-Shishtawy. (2013) "Multi-Topic Multi-Document Summarizer" in International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 6.