ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# ENRICH FRAMEWORK FOR MULTI-DOCUMENT SUMMARIZATION USING TEXT FEATURES AND FUZZY LOGIC

[1]SACHIN PATIL, [2] RAHUL JOSHI
[1, 2] Symbiosis Institute of Technology, Department of Computer science, Pune
Affiliated to Symbiosis International University (SIU), Pune
E-mail: [1]sachin.patil@sitpune.edu.in, [2]rahulj@sitpune.edu.in

## ABSTRACT

The rapid growth of Information Technology triggers collection of documents in massive form, so to find the important information from multiple document is a complex task. The multiple documents summarization is task of producing assured summary from these document set. There are other summarization techniques like sentence clustering, term weight etc. However, these techniques use only two or three feature of text to find the importance of considered sentence. In this paper, we put forward an idea of text summarization which considers multiple extracted features by applying natural language processing (NLP) protocol. The ten feature of text are extracted and these feature classified on the basis of fuzzy logic to get the best documents summary. The key features are preprocessing, feature scoring, inference engine, and fuzzy logic**.**

**Keywords:** *Preprocessing, Feature Scoring, Normal Distribution, Inference Engine, Fuzzy Logic.*

## 1. INTRODUCTION

Over the past several years, there has been much interest developed in the area of multi-document summarization. Multi-document summarization is a increasingly necessary task as document collections grow larger due to technological advancements. So, there is a greater need to summarize these documents to help users to quickly find either the most important information over-all or the most relevant information to the user. For example, the areas where multi-document summarization is helpful like in news, email threads, blogs, reviews, and search results. With the rapid growth of online information, and there might be possibility that many documents may be covering the same topic, the summarization of information from these different sources into an informative summary helps to reduce overhead in finding specified information.

The natural language processing (NLP) is a skilled system to process the natural language like English instead of any specialized computer language like C, C++ and Java. The text is the largest repository of human knowledge and it growing faster like e-mails, web pages, technical documents, news articles, PDF files or general information documents[1]. The aim of NLP developer is to design a system that can understand and manipulate the natural language to perform specified task.

The summary is a text that is produced from one or more document which preserve the meaning of original document and shorter than the original length of considered documents. It is mandatory that produced summary is pointer to some part of original document. In this paper, automatic summarization takes place from multiple source documents as input, then preprocessing of these documents i.e. removal of stopword and stemming is done. The output of data preprocessing goes as an input to feature extraction. Here, ten feature of each sentence are extracted to find importance of that sentence in the document. Finally, by using the fuzzy logic and normal distribution, decision about the importance of sentence is taken, then by using if-then rule the best sentence is picked from documents as a summary of multiple documents.

As shown in fig.1 the proposed system consists of three parts viz., 1) preprocessing of document, 2) extraction of text or sentence features, and 3) fuzzy logic.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

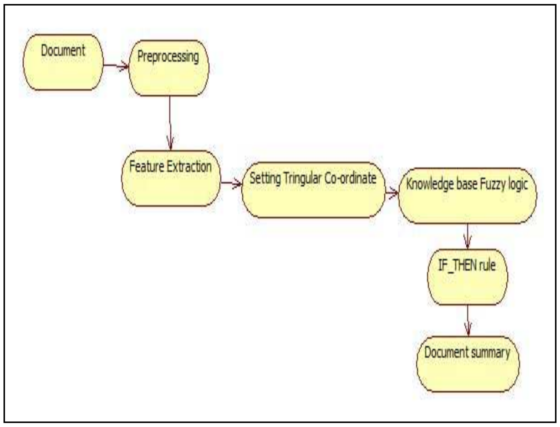


*Figure 1: Mind Map of Proposed System*

The document pre-processing consists of Removal of stopwords (i.e. to remove the frequently occurring but meaningless words in term information retrieval), stemming of words, and tokenization of sentence. Then, scores are find out for title feature, sentence position, sentence length, term weight, sentence to sentence similarity, proper noun, Numerical data, thematic word, positive words and negative words. Each of these scores of every sentence is to be considered as an input to fuzzy inference engine. The Fuzzy inference engine uses fuzzy logic and normal distribution to classify the sentences. At the last uses if-then rule is used to check the importance of sentences in the document and retrieval of the high important sentences from documents is done as a summary.

In this paper, section-II shows related work, section-III explains proposed system and section-IV lists out conclusions.

.

## 2. RELATED WORK

In this section, we scrutinized different methods of multi-document summarization. In general, summarization is divided into two categories [2] viz., Abstractive and Extractive. The abstractive method is human generated summary which attempts to develop understanding of concept and explain into a simple natural language. Since, yet a computer does not have language capabilities as that of human being's extractive method for automatic text summarization. This method considers extraction of important sentences and merging them into a shorter form without changing the meaning of original document.

The recent work on summarization is mainly focused on term weight [2].The proposed system calculates frequency of term based on occurrence of that term. Then assign weight to sentence by adding all the term weights of a term of that sentence. Finally, extraction of a highly ranked sentence as a summary is done. Because of only consideration of term weight, it may be possible that word which occurs frequently in documents are not related or not have high importance, and because of this, important words are neglected. So, in this paper an idea is put forward to consider more features which paramount's to measure importance of sentence.

A sentence similarity based summarization is presented in [4]. In this case, they measure the similarity between sentence by calculating the word similarity, word order and word semantic. Here, only one feature of text (i.e. sentence similarity) is measured while there are many other features of text that are essential to generate the summary.

Document clustering is used for improving the performance of Information Retrieval System. Many clustering methods have been presented for browsing documents or organizing the retrieved results for viewing them easily [5]. Some researchers have also applied agglomerative clustering method in which each document considers as separate cluster. In next step two similar documents are merged and make it as one cluster, this process is repeated until the required number of cluster obtain. In this method any individual property of cluster is neglected, so when noise is present, there is possibility of wrong merging.

. Most of the existing systems are graph-based ranking algorithms and they treat a sentence or text as a bag of words and leverage only literal or syntactic information in text documents. Also, it ignores the features of text which are so important while generating the summary like sentence similarity, term weight, numerical data etc. So, the output summary may not efficient or it may contain true positive and false negative data. So, the accuracy of summary falls down in this case. Hence, the proposed system uses the features of text to generate the fluent summary. Here, use of the ten features of text which are very useful to produce efficient summary is done.

Genetic algorithm (GA) is also used for multiple document summarization. As in genetic algorithm, use of historic data to perform the present task is done and for the same, training data set is required. The training data includes manually extracted summary which is used in genetic modeling to calculate the fitness function. Here, it may be possible that the documents used in training data set are not related to user input documents [7].

Journal of Theoretical and Applied Information Technology
30th June 2016. Vol.88. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

## 3. APPROCH USED IN PRAPOSED SYSTEM

The system is divided into three phases viz., 1) Preprocessing, 2) Extraction of features, and 3) Generation of summary using fuzzy logic.

### 3.1 Preprocessing

The preprocessing part contains removal of stopwords, stemming and tokenization of words.

#### 3.1.1 Stopwords Removal

To remove the stopwords from document we maintain the array list of stopwords. Stopwords are removed by comparing the each word of document with words of array list.

#### 3.1.2 Stemming

Stemming of words means to find the root of words. To find the root of words of document we store list of words which mostly come in beginning or in ending of words. So by checking the words we can find the root of words.

#### 3.1.3 Tokenization

After removing stopwords and stemming of words, the indexes is given to output sentences.

### 3.2 Text Feature Scoring

#### 3.2.1 Sentence position

Sentence position is a sentence location in a paragraph. We assumed that the first sentence of each paragraph is define general meaning of paragraph and hence it is most important sentence. Therefore, we sort the sentence based on its position.

$$Score\,(f1) = \sum_{\substack{5 \geq i \geq 1 \\ 1 \leq j \leq 5}} \left(\frac{i}{5}\right) \; for\ j^{th}\ sentence \quad (1)$$

#### 3.2.2 Sentence to sentence similarity

Sentence similarity is the similar vocabulary words between sentence and other sentences in the document. Similarity between sentence is computed by the cosine similarity.

$$Score\ (f2) = \frac{Sum\ of\ similarity\ between\ S\ and\ other\ sentence}{Maximum\ Sentence\ similarity} \quad (2)$$

#### 3.2.3 Proper Noun

Usually the sentence that contains more proper nouns is important than other general sentences and it is most probably included in the document summary.

$$Score\,(f3) = \frac{Total\ No.of\ Proper\ Noun\ in\ S}{Length\ of\ S} \quad (3)$$

#### 3.2.4 Numerical data

As numerical data contain count of specific thing, it more likely to be included in summary. The sentence that contains numerical data is an important and usually included in the document summary.

$$Score\,(f4) = \frac{Total\ No.\ of\ Numerical\ data\ in\ S}{Length\ of\ S} \quad (4)$$

#### 3.2.5 Sentence length

Sentences that are too short are not expected to belong to the summary. Longer sentence contain more information hence longer sentence has higher importance.

$$Score\ (f5) = \frac{Total\ No.of\ words\ in\ S}{No.of\ words\ of\ long\,est\ sent\,ence} \quad (5)$$

#### 3.2.6 Title feature

The title is the term overlap between sentence and the document title. Title feature measure by counting number of matches between content word in sentence and word in title. As maximum number of matches, sentence is more related to topic.

$$Score\ (f6) = \frac{Number\ of\ similar\ words\ between\ title\ and\ sentence}{Total\ numb\,er\ of\ words\ in\ title} \quad (6)$$

#### 3.2.7 Term weight

Term weight is nothing but the occurrence of word in particular document. The term occur

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

frequently in document means it is more informative or it has more importance.

$$Score(f7) = \frac{Sum\ of\ TF\text{-}IDF\ in\ S}{Max\ sum\ of\ TF\ IDF} \quad (7)$$

### 3.2.8 Thematic word

Word that occurs more frequently is more related to subject or title of document. So it is very important to include such word in summary. We can consider top 10 word as thematic words.

$$Score(f8) = \frac{No.\ of\ Thematic\ words\ in\ S}{Total\ No.\ of\ Thematic\ words\ in\ documents} \quad (8)$$

### 3.2.9 Positive Keyword

Positive words are keywords shows the positive attitude towards things. Positive keyword are words which mostly included in summary.

$$Score\ (f9) = \frac{Number\ of\ positive\ words\ in\ S}{Length\ of\ S} \quad (9)$$

### 3.2.10 Negative keyword

Negative words are keywords shows the negative attitude towards things. Negative keyword are words which mostly included in summary.

$$Score\ (f10) = \frac{Number\ of\ Negative\ words\ in\ S}{Length\ of\ S} \quad (10)$$

## 3.3 NORMAL OR GAUSSIAN DISTRIBUTION

Information may be "disbursed" (spread out) in distinctive methods. It can be spread out more on the left or more on the right or it may be all jumbled up[11]. So before applying fuzzy rule it is very essential to make feature score tends to be around central value with no bias left or right. The Normal Distribution has imply, median, mode and symmetry.

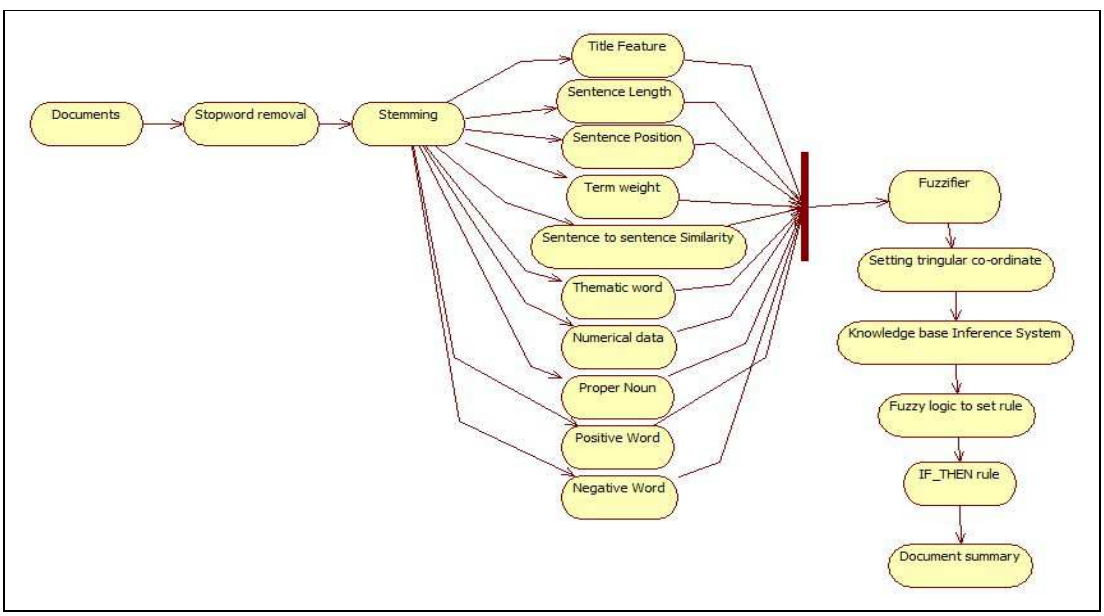


*Figure 2: Architecture of System*

Journal of Theoretical and Applied Information Technology
30th June 2016. Vol.88. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

**Algorithm**
// Input : Document Set **D= { $D_1$, $D_2$, $D_3$……..$D_n$}**
// Output : Summary **S**
Step 0: **Start**
Step 1: **FOR** i=0 to **D** size
Step 2: Initiate String **cont** to empty
Step 3: **$F_c$**= $D_i$(Each File Content)
Step 4: **cont**=**cont**+$F_c$
Step 5: **END FOR**
Step 6: Extract sentence from **cont** and add in vector **SENT**
Step 7: **FOR i=0 to SENT size**
Step 8: Get all the Features **F**= {$F_1$, $F_2$, $F_3$……….$F_{10}$}
Step 9: **END FOR**
Step 10: **FOR** i=0 to SENT size
Step 11: Get a Feature $F_i$
Step 12: Get mean μ and Standard deviation δ of the feature
Step 13: calculate Gaussian function **g(x)** for the Random value x
Step 14: add all Fi values and **g(x)** in avector **Temp**
Step 15: **END FOR**
Step 16: Identify Centroid **C**, Small **S** and Big **B**
Step 17: Based on **S**, **C** and **B** create Fuzzy Crisp Values like VL, L, M, H, and VH
Step 18: **Set Protocols for Ideal Sentence**
Step 19: Apply Fuzzy **IF – THEN** Rules
Step 20: Extract IDEAL SENTENCE and add into Set **S** as Summarized Sentences
Step 21: **Stop**

Gaussian Equation

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

**For Step 17:**
**VL= Very Low , L= Low , M= Medium , H= High, VH= Very High**

## 4. RESULTS AND DISCUSSIONS

To show the effectiveness of the proposed system which includes ten features as mentioned in the prior section. Many experiments are conducted on java based windows machine using Netbeans as IDE which includes data in doc, pdf and txt format files. We use the human generated summaries by experts, for the measurement of our experiment results. We evaluate the summaries generated by the program with human generated summaries. The human generated summaries are the gold standard summaries, as the humans can capture and relate deep meanings of the text.

Table I shows the feature scores of all sentences of a sample document which contains 10 sentences. All the feature scores in the above table are between 0 and 1.From the ten feature values of each sentence; one value for each sentence is obtained using fuzzy logic method. We used 10 news based text documents as an input to the text summarizer. We applied the number of features to these input documents in the increasing order (application of 4 features, application of 6 features ,application of 8 features and finally application of 10 features) and obtained different resultant summaries as shown in Table II.

Journal of Theoretical and Applied Information Technology
30th June 2016. Vol.88. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

*Table I: Ten Feature Scores For Each Sentence Of A Document*

| Sentence No | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | F10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | 0.4736 | 0.0 | 1.0 | 0.4444 | 0.4 | 0.5882 | 0.5882 | 0.32 | 0 |
| 2 | 0.11 | 0.7368 | 0.0 | 0.8 | 0.4285 | 0.0 | 0.1764 | 0.1764 | 0.42 | 0.4 |
| 3 | 0.0 | 0.2631 | 0.0 | 0.6 | 0.2 | 0.0 | 0.0 | 0.0 | 0.212 | 0.1 |
| 4 | 0.0 | 0.3157 | 0.0 | 0.4 | 0.1666 | 0.2 | 0.4411 | 0.4411 | 0.19 | 0 |
| 5 | 0.44 | 1.0 | 0.0526 | 0.2 | 0.2105 | 1.0 | 1.0 | 1.0 | 0.22 | 0 |
| 6 | 0.33 | 0.6842 | 0.0 | 0.0 | 0.0769 | 0.8 | 0.8823 | 0.8823 | 0.08 | 0 |
| 7 | 0.22 | 0.7368 | 0.0 | 0.0 | 0.2142 | 0.8 | 1.0 | 1.0 | 0.78 | 0 |
| 8 | 0.11 | 0.6315 | 0.0833 | 0.0 | 0.1666 | 0.6 | 0.7352 | 0.7352 | 0.001 | 0 |
| 9 | 0.0 | 0.4210 | 0.0 | 0.0 | 0.0 | 0.2 | 0.1764 | 0.1764 | 0.0 | 0 |
| 10 | 0.11 | 0.5263 | 0.0 | 0.0 | 0.4 | 0.0 | 0.2058 | 0.2058 | 0.0 | 0.3 |

*Table II: Fuzzy Summarizer For Different Number Of Features*

| No. of features | Fuzzy Summarizer |
|---|---|
| 4 | 45% |
| 6 | 60% |
| 8 | 75% |
| 10 | 79% |

From the results observed in Table II, it is clear that use of all the ten features in the calculation of summary yields better summary.

We used the two summarizes namely Baseline summarizer and MS Word summarizer for comparison with our fuzzy summarizer along with the summarizer proposed in the paper [10] where author used 8 features for the summary extraction.

For measuring the performance of the system, precision and recall are used. Precision is defined as number of relevant i.e. summary obtained to the total numbers of relevant and irrelevant summary by the human judgement. Generally this entity is defined in percentage. So in general we can say that precision is used to find relative effectiveness of the system.

Recall is defined as a numbers of relevant summary obtained to the total number of relevant summary not obtained and number of irrelevant summary obtained. Absolute accuracy of the system is defined by the recall.

For deep understanding concern, following details can be used.

- A = The number of relevant summary sentences obtained,
- B = The number of relevant summary sentences not obtained
- C = The number of irrelevant summary sentences obtained

So,

Precision = (A/ (A+ C))*100

And Recall = (A/ (A+ B))*100

On comparing average precision and recall we get the following graph as mentioned below.

The below plot in Fig.3 indicates that our approach is yielding better result than all , even then the system uses 8 features by [10].This directly indicates as we are increasing number of features we will get better accuracy in summarization.

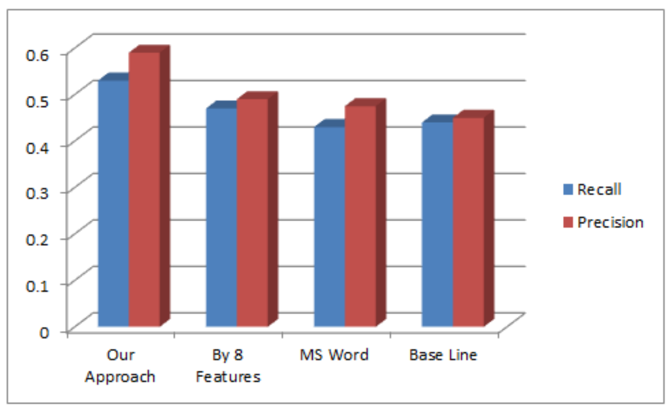


*Figure.3: Average Precision and recall comparison*

## 5. CONCLUSION

In this paper, we investigate use of the important features based on fuzzy logic; title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word, numerical data, positive words and negative words. We find the most important sentences of document using the Normal distribution, triangular membership function and fuzzy logic.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

**REFERENCES**


[1] Yue, Guangzhi Di, Yueyun Yu, Wei Wang, Huankai Shi.,"Analysis of the Combination of Natural Language Processing and Search Engine Technology." *International Workshop on Information and Electronics Engineering (IWIEE)*; Procedia Engineering, 2012; 291636-1639.

[2] R.C.Balabantaray, D.K.Sahoo, B.Sahoo, M.Swain," Text Summarization using Term Weights," *International Journal of Computer Applications (0975 – 8887) Volume 38– No.1, January 2012.*

[3] LaddaSuanmali, NaomieSalim and Mohammed Salem Binwahlan,"Feature-Based Sentence Extraction Using Fuzzy Inference rules,"*International Conference on Signal Processing Systems, 2009.*

[4] Anjali R. Deshpande , Lobo L. M. R. J.," Text Summarization using Clustering Technique,"*International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- August 2013.*

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

[6] Su Yan, Xiaojun Wan, "SRRank: Leveraging Semantic Roles for Extractive Multi-Document Summarization,"*IEEE /ACM Transaction on Audio, Speech, and Language Processing,* Vol. 22, No. 12, December 2014

[7] Aristoteles, YeniHerdiyeni, Ahmad Ridha, Julio Adisantoso, "Text Feature Weighting for Summarization of Documents in Bahasa Indonesia Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012

[8] Wikipedia, "Stemming", http://en.wikipedia.org/wiki/Stemming, last accessed 2 September 2014.

[9] ShresthaMubin ",Cosine Similarity", http://computergodzilla.blogspot.in/2012/12/what-is-cosine-similarity.html , 20 Dec. 2012.

[10] Wikipedia,"TF-IDF", https://en.wikipedia.org/wiki/Tf-idf, last modified on 11 March 2016

[11] Wikipedia, "Normal Distribution", https://en.wikipedia.org/wiki/Normal_distribution , last modified on 10 March 2016

[12] Pierce, Rod, 2016, 'Math is Fun - Maths Resources', Math Is Fun, Available at: <http://www.mathsisfun.com/index.htm>. [Accessed 14 Mar 2016]

[13] Wikipedia, "Precision and_recall", https://en.wikipedia.org/wiki/Precision_and_recall