International Journal of Computer Science Trends and Technology (IJCST) – Volume 3 Issue 4, Jul-Aug 2015

RESEARCH ARTICLE OPEN ACCESS

Improving the Performance for Single and Multi-document Text Summarization via LSA & FL

Ms.Pallavi.D.Patil, P.M.Mane
M.E Scholar, Assistant professor
Department of Computer Science and Engineering
Dnyanganga College of Engg. & Research,
Narhe, Pune, India

ABSTRACT
The automation of the process of summarizing documents assumes a basic part in numerous applications. Automatic Text Summarization has been deliberate on keep hold of the essential data without concerning the archive quality. This paper proposes on extraction based system for Single /multi-document summarization. It uses combination of both LSA (Latent Semantic Analysis) and FL (Fuzzy Logic) methods on fusion of various features like Preprocessing, feature extraction, classification to generate better quality summary.
Keywords:- NLP, Summarization, Latent Semantic Analysis, Fuzzy Logic.

I. INTRODUCTION

At the present period of time, where the great amount of information is increasing step by step. People comprehensively make utilization of the web to find information through web browsers or portals for example, Google, Yahoo, Bing, and so on. A considerable amount of time is wasted for searching relevant documents. A document summary keeps its main content and consequently helps users to understand and interpret large volumes of information available in the document. Text summarization is rewriting the text in a shorter compressed form to represent the original text. This task is accomplished by humans after deep reading and well understanding of the document content, selecting the most important points and paraphrasing them into short version. Employing the machine to imitate the human work in creating the summaries is called automatic text summarization. The main focus of both kinds of summarization is to gather the source text by extracting its most important content together a user's or application's needs.

There are two main approaches to the task of summarization—Extractive and Abstraction. Extraction involves concatenating extracts taken from the corpus into a summary, whereas abstraction involves generating novel sentences from information extracted from the corpus.

Text Summarization is done for:

- <u>Single document Summarization</u>: Single document provide most relevant information contained in single document to use that helps user in deciding whether the user in deciding whether the document
- <u>Multi-document Summarization</u>: Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic.

Motivation:

- With summaries people can make effective decisions in less time.
- Motivation is to build such a tool which is computationally efficient & creates summary automatically for single/multiple documents which improves the performance of text summarization.

Problem Statement: Existing system is based on LSA-based single-document summarization method. In existing some of attributes have more importance & some have less so, should have balance weight in computations. So, we proposed a tool in which is combination of LSA based, Feature Extraction and FL method. Fuzzy logic solves existing system problem by defining the membership functions for each feature. Thus, summary generated can be as informative as the full text of document with better information coverage.

II. RELATED WORK

An The first Automatic text summarization was created by Luhn in 1958[1] based on term frequency. Automatic text summarization system in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights: a) Cue Method b) Title Method c) Location Method. The Trainable Document Summarizer in 1995 performs sentence extracting task, based on a number of weighting heuristics. Following features were used and evaluated [2]:

1. Sentence Length Cut-O Feature: sentences containing less than a pre-specified number of words are not included in the abstract

2. Fixed-Phrase Feature: sentences containing certain cue words and phrases are included

3. Paragraph Feature: this is basically equivalent to Location Method feature

4. Thematic Word Feature: the most frequent words are defined as thematic words. Sentence scores are functions of the thematic words' frequencies

5. Uppercase Word Feature: upper-case words (with certain obvious exceptions) are treated as thematic words.

In [3], paper shows recent techniques and challenges on advances of automatic text summarization. Special attention is paid to the latest trends in text summarization. Author discusses the key challenges in automatic text summarization. These are inherent problem of overlapping of sets of similar text units or paragraphs; documents which contain long sentences are still a problem; another challenge is the word sense ambiguities which are inherent to natural language. The problem is that matching a system summary against the ideal summary is very difficult to establish. The problem of providing much accurate or efficient result for automatic text summarization. Author has discussed different types of summarization approaches [4] depending on what the summarization method focuses on to make the summary of the text. Automatic document summarization is extremely helpful in tackling the information overload problems. It is the technique to identify the most important pieces of information from the document, omitting irrelevant information and minimizing details to generate a compact coherent summary document. Author has given the types of summaries - Abstract vs. Extract summary, Generic vs. Query-based summary, Single vs. Multi-document summary, Indicative vs. Informative, Background vs. Just the news. Author has mentioned in detail the different approaches:Graph Theoretic

Approach, Text summarization using cluster based method.

In [2], paper author has concentrating on extractive summarization methods. An extractive summary is selection of important sentences from the original text. The importance

of sentences is decided based on statistical and linguistic features of sentences. There are two broad methods of text summarization extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. An abstractive summarization method consists of understanding the original text and retelling it in fewer words. Paper [1], defines the most important criteria for a summary and different methods of text summarization as well as main steps for summarization process is discussed. There are different approaches for sentence selection are presented in order to generate a summary from a text. Author has explained detailed steps for summary of document, which are topic identification, interpretation and summary generation. Also author has given

the different approaches for scoring and selecting sentences.

In paper [5], it is said that sentence scoring is the technique most used for extractive text summarization. Paper describes 15 sentence scoring algorithm and performs a quantitative and qualitative assessment of these 15 algorithms. Example Sentence length: It works as follows: (i) Calculate the largest sentence length; (ii) Penalize sentences larger than 80 percent of the largest sentence length; (iii) Calculate the Sentence Length Score for all other sentences. Author has used three different datasets (News, Blogs and Article contexts) . It is a single document summarization.

In this [6] paper author proposed a sentences clustering based summarization approach. The proposed approach consists of three steps: first clusters the sentences based on the semantic distance among sentences in the document, and then on each cluster calculates the accumulative sentence similarity based on the multi features combination method, at last chooses the topic sentences by some extraction rules. The text summarization result is not only depends the sentence features, but also depends on the sentence similarity measure.

This paper [7], proposed an automatic text summarization approach based on sentence extraction using fuzzy logic, genetic algorithm, semantic role labelling and their combinations to generate high quality summaries. GA used in text summarization. Author has mentioned the benefits of the genetic algorithm in the optimization problem in for feature selection.

In [8] paper, author proposed a method of personalized text summarization which improves the conventional automatic text summarization methods by taking into account the differences in readers characteristics. A method of personalized summarization which extracts from the document information that we assume to be the most important or interesting for a particular user. Annotations (e.g. highlights) can indicate a user as interest in the specific parts of the document .Author used them as one of the sources of personalization.

This [9] paper describes and performs a quantitative and qualitative assessment of 15 algorithms for sentence scoring available in the literature. Three different datasets (News, Blogs and Article contexts) were evaluated. Each of the 15 scoring methods is described and implemented. A quantitative and qualitative assessment of those methods using three different datasets (news, blogs, and articles context) is performed.

In this [10] paper, the investigation of the correlation between ROUGE and human evaluation of extractive meeting summaries is carried out. Both human and system generated summaries are used. The human evaluation of different summaries and calculated ROUGE scores are examined with their correlation .The better correlation can be achieved between the ROUGE scores and human evaluation. In text summarization, ROUGE has been correlate well with human evaluation when measuring match of content units.

In recent years, algebraic methods are used for text summarization. Most well-known algebraic algorithm is Latent Semantic Analysis (LSA) (Landauer et al., 1998). This algorithm finds similarity of sentences and similarity of words using an algebraic method, namely Singular Value Decomposition (SVD). Besides text summarization, the LSA algorithm is also used for document clustering and information filtering.

In **Existing System LSA** is used for Text Summarization . LSA is difficult to handle polysemy, also in LSA some of attributes have more importance and some have less. To balance weight in computations we use fuzzy logic. LSA doesn't require any training or external knowledge. It has ability to collect all trends \& patterns from all documents and is based on context input document .Summary generated by fuzzy logic can be informative but the hidden semantic between sentences in the text is missing. So for Removing the drawbacks of LSA and Fuzzy Logic, we build such a tool which is fusion of LSA and Fuzzy Logic Method and it improves the performance of text summarization. This tool is used for single and Multi-document Summarization. For Multi-document Summarization Clustering is used.

III. TECHNIQUES

A. Latent Semantic Analysis

LSA is the most conspicuous algebraic learning algorithm utilized for Information Retrieval (IR) from text based information. LSA is generally utilized as a part of different applications for the measurement dimension of extensive multi-dimensional information.

LSA strategy utilizes Singular Value Decomposition (SVD) for discovering semantically comparable words and sentences. SVD is a strategy that models connections among words and sentences. It has the ability of commotion lessening, which prompts a change in precision.

LSA Algorithm:

Step 1 - Creating the Count Matrix
Step 2 - Modify the Counts with TFIDF
Step 3 - Using the Singular Value Decomposition
Step 4 - Sentence Selection for summary.
Sentence Selection based on LSA[10]
Input: Document *D*, Matrix *U*, Matrix *VT*, *M*
Output: Set *S*
1. **Initialize** $S=\phi$, $k=1$
2. **while** $|S|<M$
3. get *l* in *vk*, S=S ∪{ *sentl* }, update *VT*, $Nk=1$
4. get *p*, *q*, *s* in *uk*, *T*={ *term p*, *term q*, *terms* }
5. $T0=T \cap sentl$, $T=T-T0$
6. **while** $(T\neq\phi)$
7. **if** ($Nk<3$ and $|S|<M$)
8 .get *l* in *vk*, S=S ∪{ *sentl* }, update *VT*, $Nk = Nk+1$
9. $T0=T \cap sentl$, $T=T-T0$
10.**else** $T=\phi$
11. **end while**
12. $k=k+1$
13. **end w**

B. Feature Extraction

The text document is represented by set, D= {S_1, S_2,- - - , S_k} where, Si signifies a sentence contained in the document D .The document is subjected to feature extraction. The important word and sentence features to be used are decided .This work uses features such as Sentence length, Sentence position, numerical data, Term weight, sentence and proper Nouns

Sentence Length: We dispose of the sentences which are excessively short, for example, datelines or creator names. For each sentence the standardized length of sentence is ascertained as,

F1= Number of words in sentence/Number of words in the longest sentence

Sentence Position: The sentences happening first in the section have most astounding score. Assume a section has n sentences then the score of each sentence for this peculiarity is computed as takes after:

F2(S1) = n/n; F2(S2)=4/5; F2(S3)=3/5; F2(S4)=2/5; and so on.

Numerical data: The sentences having numerical information can reflect imperative insights of the archive and may be chosen for rundown. Its score is computed as

F3 (Si) = Number of numerical data in sentence Si /Sentence Length

Term weight: The frequency of term occurrences within a document has often been used for calculating the importance of sentence. The score of a sentence can be calculated as the sum of the score of words in the sentence. The score of important score w_i of word i can be calculated by word cloud.

Proper Nouns: The sentence that contains maximum number of proper nouns is considered to be important. Its score is given by,

F5=Number of proper nouns in the sentences/Sentence Length(s)

C. Fuzzy Logic

Fuzzy Logic-FL is used for solving uncertainties in a given problem. In short Fuzzy system consists of a formulation of the mapping from a given input set to and output using fuzzy logic, which consists of the following five steps:

\item \begin{itemize}

Step 1: Fuzzification of input variables, defining the control objectives and criteria.

Step 2: application of fuzzy operators (AND, OR, NOT) in the IF (antecedent) part of the rule. Determine the output and input relationships and choose a minimum number of variables for input to the fuzzy logic engine.

Step 3: implication from antecedent to the consequent (THEN part of the rule) for the desired system output response for a given system input conditions.

Step 4: aggregation of the consequents across the rules by creating fuzzy logic membership functions that define the meaning (values) of input/output terms used in the rule.

Step 5: defuzzification to obtain a crisp result.

D. Proposed Algorithm for Multi-document Summarization :

Agglomerative K-means algorithm with concept Analysis :

Choose k number of clusters to be determined.

Choose k objects randomly as the initial cluster center.

3. Repeat

3.1 Assign each object to their closest cluster

3.2. Compute new clusters, i.e. Calculate mean points.

4. Until

4.1. No changes on cluster centers (i.e. Centroids do not change location any more)

4.2. No object changes its cluster (We may define stopping criteria as well.

5. Calculates TF weight for each term t of each cluster using word cloud.

6. The term which has highest tf value that term is name of that cluster.

IV. PROPOSED SYSTEM

The figure shows the system architecture for Single Document :

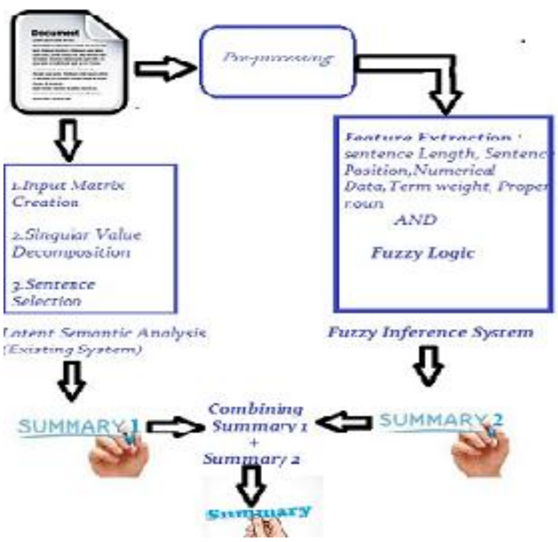

Fig. 1 :Flow for Single Document

The systemconsists of the following main steps:

1. Read the single documents into the system.
2. For pre-processing step, the system extracts the individual sentences of the original documents. Then, separate the input document into individual words. Next, remove stop words. The last step for pre-processing is word stemming.
3. In the sentence features extraction of fuzzy system each sentence is associated with vector of eight features that described in above Section, whose value are derived from the content of the sentence; In the same way to the semantic system the input matrix of term by documents is created with Cell values.
4. In the sentence features extraction of fuzz y system each sentence is associated with vector of eight features that described in above Section, whose values are derived from the content of the sentence; In the same way to the semantic system the input matrix of term by documents is created with Cell values.
5. In fuzzy system, a set of highest score sentences are extracted as document summary based on the compression rate, and in SVD system, the VT matrix cell values represents the most important sentences extracted . A higher cell value indicates the most related sentence .Thus numbers of sentences are collected into the summary based on compression rate.
6. After getting summary1 and summary2, we intersect both summaries and extract a set of common sentences and a set of uncommon sentences. From uncommon set, we extract the sentences with high sentence scoring. And final set of improved summary is obtained by union of both the sets.

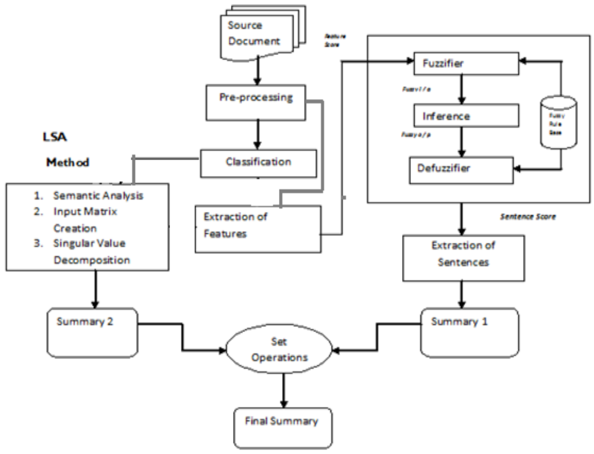

Fig. 2 :System Architecture for Multiple Documents

The figure shows the system architecture for Multiple Documents:

The systemconsists of the following main steps:

1. Read the multiple documents into the system.
2. For pre-processing step, the system extracts the individual sentences of the original documents. Then, separate the input document into individual words. Next, remove stop words. The last step for pre-processing is word stemming.
3. After pre-processing we use classification. For classification we propose Agglomerative K-means algorithm with concept Analysis in that classification is done using Agglomerative K-means after that Calculates TF weight for each term t of each cluster using word cloud. We gives TF as input to LSA.
4. The term which has highest tf value that term is gives to name of that cluster.
5. in SVD system, the VT matrix cell values represents the most important sentences extracted . A higher cell value indicates the most related sentence .Thus numbers of sentences are collected into the summary.
6. In the sentence features extraction of fuzzy system each sentence is associated with vector of eight features that described in above Section, whose value are derived from the content of the sentence; In the same way to the semantic system the input

matrix of term by documents is created with Cell values.

7. In fuzzy system, a set of highest score sentences are extracted as document summary based on the Fuzzy rule
8. After getting summary1 and summary2, we intersect both summaries and extract a set of common sentences and a set of uncommon sentences. From uncommon set, we extract the sentences with high sentence scoring. And final set of improved summary is obtained by union of both the sets.

V. RESULTS

We use online generated summary by online summarizer as a gold summary standard for evaluation that has become standards of automatic evaluation of summaries. It compares the summaries generated LSA method, Fuzzy logic based method, Combined Summary (Proposed Summary) with the human generated (gold standard) summaries. For comparison, we used accuracy statistics. Our evaluation was done using accuracy percentage which was found to have the highest correlation with human judgments, namely, at a confidence level of 95%. It is claimed that our summary correlates highly with human assessments and has high recall and precision significance test with manual evaluation results. So we choose precision, recall as the measurement of our experiment results. we calculate Precision ,Recall, F-measure for online summarizer with different data sets.we calculate Precision ,Recall, F-measure for LSA summary (old Summary)of different data sets. we calculate Precision ,Recall, F-measure for Fuzzy Summary , we calculate Precision ,Recall, F-measure for combined Summary / Proposed Summary (New Summary), We calculate Average Precision,Average Recall , Average F-Measure of different Data Set for Online ,LSA, Fuzzy and proposed Summarizer.

Comparison Table of Average Precision ,Avrege Recall ,Average F-measure:

Summary	Avg .Precision	Avg.Recall	Avg.F - measure
Online Summary	87.4	46.2	66.8
LSA summary	87.4	46.4	67
Fuzzy Summary	87.6	44.8	66.2
Combined Summary(Proposed Summary)	89	43.6	66.3

Table 1: Average Precision ,Avrege Recall ,Average F-measure:

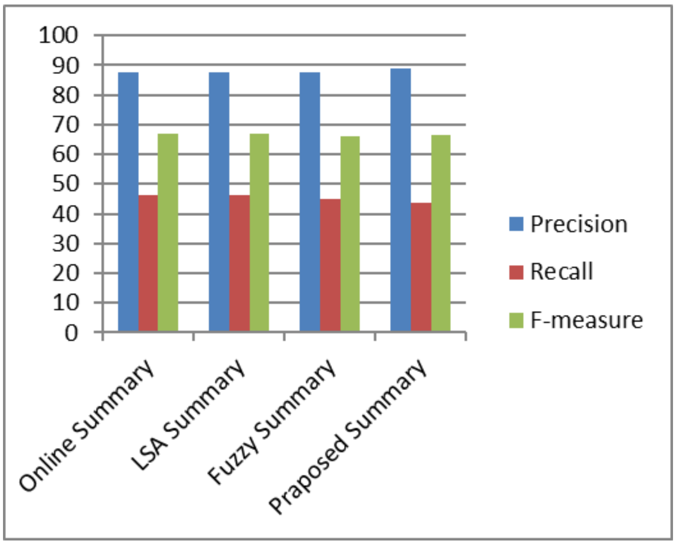

Fig 4: Comparison of Average- Recall, Precision, F-measure for Online ,Lsa ,Fuzzy, Combined Summary

VI. CONCLUSION

Text summarization systems can be categorized into various groups based on different approaches. In extraction based summarization the important part of the process is the identification of important relevant sentences of text. LSA doesn't require any training or external knowledge. It has ability to collect all trends \& patterns from all documents and is based on context input document to find out the hidden semantics to the documents. Identify important Features for sentences of document and Fuzzy logic defines those features that need to be used for text summarization Carried out semantic relation between word and sentences in text document. Drawback of LSA is difficult to handle polysemy, also in LSA some of attributes have more importance and some have less. To balance weight in computations we use fuzzy logic. Drawback of fuzzy logic is it does not have ability to find out hidden semantic words.

So, we Make combination of fuzzy logic and LSA which improves the performance of text summarization for single and Multiple.

REFERENCES

[1] Saeedeh Gholamrezazadeh ,Mohsen Amini Salehi, “A Comprehensive Survey on Text Summarization Systems ”, 978-1-4244-4946-0,2009 IEEE.

[2] Vishal Gupta and Gurpreet Singh Lehal "A survey of Text summarization techniques ",Journal of Emerging Technologies in Web Intelligence VOL 2 NO 3 August 2010.

[3] Oi Mean Foong ,Alan Oxley and Suziah Sulaiman "Challenges and Trends of Automatic Text Summarization ",International Journal of Information and Telecommunication Technology Vol.1, Issue 1, 2010.

[4] Archana AB, Sunitha. C ,"An Overview on Document Summarization Techniques" ,International Journal on Advanced Computer Theory and Engineering (IJACTE) ,ISSN (Print) : 2319 ˝U 2526, Volume-1, Issue-2, 2013 .

[5] Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva ,Fred Freitas ,George D.C. Cavalcanti ,Luciano Favaro , "Assessing sentence scoring techniques for extractive text summarization ",Expert Systems with Applications 40 (2013) 5755-5764 ,2013 Elsevier .

[6] ZHANG Pei-ying ,LI Cun-he , "Automatic text summarization based on sentences clustering and extraction ",978-1-4244-4520-2 ,2009 IEEE .

[7] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan "Fuzzy Genetic Semantic Based Text Summarization ", 2011 Ninth Ninth International Conference on Dependable, Autonomic and Secure Computing ,978-0-7695-4612-4 ,2011 IEEE .

[8] Róbert Móro, Mária Bieliková "Personalized Text Summarization Based on Important Terms Identification ",2012 23rd International Workshop on Database and Expert Sytems Applications ,1529-4188, 2012 IEEE .

[9] Rafael Ferreira ,Luciano de Souza Cabral ,Rafael Dueire Lins ,Gabriel Pereira e Silva , "Assessing sentence scoring techniques for extractive text summarization", Expert Systems with Applications 40 (2013) 5755-5764,,2013,Elsevier,Ltd.

[10] Feifan Liu and Yang Liu, Member, IEEE "Exploring Correlation Between ROUGE and Human Evaluation on Meeting Summaries ",IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 1, JANUARY 2010.

[11] Mrs.A.R.Kulkarni , Dr.Mrs.S.S.Apte "A DOMAIN-SPECIFIC AUTOMATIC TEXT SUMMARIZATION USING FUZZY LOGIC ",International Journal of Computer Engineering and Technology (IJCET), ISSN 0976- 6367(Print), ISSN 0976 - 6375(Online) Volume 4, Issue 4, July-August (2013).

[12] Yingjie Wang and Jun Ma” A Comprehensive Method for Text Summarization Based on Latent Semantic Analysis”, G. Zhou et al. (Eds.): NLPCC 2013, CCIS 400, pp. 394–401, 2013.© Springer-Verlag BerlinHeidelberg(2013).