International Journal of Advance Research in Science and Engineering
Vol. No.5, Special Issue No. 01, March 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

# ENGLISH TEXT SUMMARIZATION USING CST AND LSA APPROACH

**Deepak A[1], Gauri Patil[2], Rutuja Patil[3], Shradha Patil[4]**

*[1]Assistant Professor, [2,3,4]Student, BE, Dept of Computer Engineering,*

*St John College of Engineering and Technology (India)*

**ABSTRACT**

*Summarization is a technique which is used to get the concept of large documents in a short and brief way.It is very difficult to read huge documents for understanding the concept of given documents,and summarization is a way which will save time and make the concept understand in a better way.Cross document structure theory (CST) is used to give the relation between the different texts form a similar topic in the documents.While summarizing the sentences it does not consider numerical data,multimedia files like graphs or images.English text is identified and summarized.Latent semantic analysis (LSA) withtf-idf matrix is method used for summarization.*

***Index terms: Cross document structure theory (CST), Latentsemantic analysis (LSA), Singular value Decomposition (SVD),***

## I. INTRODUCTION

Summarization of text documents is done under Natural Language Processing (NLP).Automatic text summarization is used to compress the given text to its necessary contents, based upon user's choice of shortness.Summarization can be classified assingle document text summarization and multi document text summarization.In single document text summarisation summary is obtained from a single text document.While in multi-document text summarisation a single summary is obtained from multiple documents. After summarization of documents CST is applied to obtain the relations among the sentences.The ancient technique used for summarizing the text was LSA.But because of some drawbacks of LSA like the model is not humanly readable that is similar words are found through latent analysis which is tedious for human so now new method called LSA with tf-idf matrix is used.

The summarization process works in several phases.Phases for obtaining summary are pre-processing phase,sentence selection phase,knowledge extraction etc. The very first step in pre-processing is tokenisation.In this step sentences are broken into tokens, individual words.The preceding step is stop word elimination,which eliminates all the stopwords from the document.Stopwords are words like "and","in","the","or","if", etc. words which do not add meaning to sentences in document.After eliminating stopwords from the sentences it checks for synonyms means words which have similar meaning and the words with similar meaning are removed,this comes under stemming.The next phase which comes is sentence selection where SVD and LSA with tf-idf method is used.

The sentences are decomposed in this phase.The final phase is knowledge extraction,in this phase similarities between sentences are removed.CST can be used to obtain the relation between the similar sentences on same topic.The previous approaches that were used for summarization were graph based method, fuzzy based method etc.

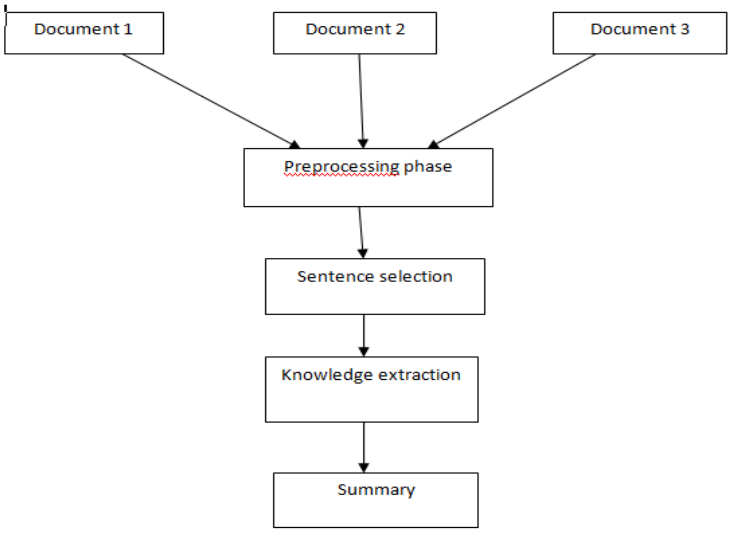


**Fig No:1Architecture Diagram of Proposed System**

But the recent work has evolved LSA with tf-idf matrix for summarization.CST approach is used in summarization for finding relations among sentences.

For representing CST multi-document cubes and multi-document graphs are used. The previous method used was cosine distance to find the similarity between the words. But this method was tedious with many mathematical formulas thus CST is used over cosine distance. Relationship types for multi-document graphs are identity, equivalence, translation etc. they are used to find relations between the sentences.CST identification algorithms are used for identifying the relations and similarities.

## II. RELATED WORK

LSA using tf-idf matrix for summarization was proposed in "Multi-document English Text Summarization uses Latent Semantic Analysis" by SoniyaPatil, Ashish T. Bhole.It has limitation that using cosine distance it is complex to compute the level of similarity of words in the matrix constructed[1].

Natural language processing began in earnest in 1950 when Alan Turing published paper entitled "Computing Machinery and Intelligence". Later on the so-called Turing Test emerged on the basis of this paper[2].

Since the early 1960s, several summarization methods have been proposed by researchers.The most prominent study was conducted by H.P. Luhnusing term frequencies and word collections from "The Automatic Creation of Literature Abstracts". The research aimed to obtain generic abstracts for several research papers[4].Thisapproach was able to handle only single documents with less than 4000 words total.

Another important approach in the early stages was proposed by Edmundson [5] using term frequencies and emphasizing the location of the sentences. Sentences at the beginning and end were given priority over other

sentences. In the recent past, several methods have been proposed based on statistical, graph based, and machine learning approaches

M.Sanderson proposed a query-relevant summarizer that divides documents in to equally sized overlapping passages.In the last ten years a lot of new approaches have appeared as a result of the information overload[6]. Recently, new approaches for LSA have been developed.

LSA for multi-document for text summarization for Persian language proposed by Asef, Kahani, Yazdi and Kamyar in 2011. Its limitation is that it is used for Persian language and hence it differs from English language semantically [7].

## III. PROPOSED WORK

This section introduces the advance method of LSA with tf-idfwhich overcomes the limitations of earlier LSA. The Multi-Document summarization can have single or multiple documents.Generally multiple documents are taken as input. The process begins with the initial step of collecting documents and forming numerical dataset. The input dataset is fit into a vector space model, where each unique word in a sentence is represented by the number of times it occurs in the document.

LSA with tf-idfmethod obtains summaries in three stages: Input matrix, Singular value decomposition(SVD) and sentence selection.Input Matrix transforms the given sentences in to matrix form. Singular value decomposition (SVD) is used to decompose input matrix in to obtain and clear summary.Sentence Selection extracts the important concepts from the documents.But prior to these stages the most important step is Pre-processing.

### 3.1Pre-processing stage

The documents given as input are text documents. The first step of pre-processing phase is segmentation. Segmentation is the process of breaking the paragraphs in to sentences which is called segmentation.

The tokenization is the next step in pre-processing phase. Thetokenization is the method of dividing sentences in to words or different tokens. The next step in pre-processing stage is removal of stopwords. Stopwords are the words that do not add to the individual meaning and they are used for forming or joining of sentences for example,“a”,”the”.

The final step is to remove redundancy of words or words with similar meaning.

### 3.2 Input Matrix:

The input matrix is constructed using a bag of words obtained from pre-processing phase.

**Table 1**

| | Sentence1 | Sentence 2 | Sentence 3 |
|---|---|---|---|
| The | 1 | 1 | 1 |
| Unicorn | 1 | 1 | 1 |
| Is | 1 | 2 | 1 |
| Magical | 1 | 0 | 1 |
| Mythical | 0 | 1 | 1 |
| When | 0 | 1 | 0 |
| It | 0 | 1 | 0 |
| Horned | 0 | 1 | 0 |
| And | 0 | 0 | 1 |

For example, consider the following dataset:

"The unicorn is magical"

"The unicorn is mythical when it is horned"

"The unicorn is mythical and magical"

The above dataset has input matrix (Table 1) which is constructed on the basis of how many types a unique word has appeared in the sentence.

The input matrix for summarization constitutes an identity matrix representing the numberof documents and word frequency (W) matrix shown above. The identity matrix for these documents is shown below;

$$D= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

To multiply the3x3 identity matrix, term frequency is transposed to match the number of columns in the matrix.

The input matrix for decomposition is WT :

$$A \leftarrow WT = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The matrix representation of the words in the dataset can be subjected.

### 3.3 Sentence Selection

This phase is related to extraction of the actual concept of document .The main concept of document is extracted using LSA.SVD is similar to Eigen value decomposition, but the difference between SVD and Eigen value decomposition is that it can be applied on rectangular matrices, but Eigen decomposition is strictly used for square matrices.

Singular vector decomposition (SVD) is used to reduce the number of rows in matrix and convert the matrix into orthogonal factors which represent term and document. Matrix vectorization is performed using SVD.

There is function in SVD used to recognise the document related to particular file. SVD posses feature which has the ability to identify the relationship between words or terms which is further used to cluster terms as well as sentences semantically. The process of sentence selection consists of three step sentence formulation, sentence selectionand combination.

Step1: Formulation of document :In SVD there is input matrix A where rows consist of the number of wordsor terms and column consist of number of sentences in whole document.

Step2: Sentence selection

- Let S be the summary set. S is set to null.
- Calculate the index value for each sentence and determine the highest index value. Each sentence has particular pattern of words.If this pattern is present in the document frequently then it is represented using Singular vector followed by calculating the index value which gives the degree of importance of that pattern. If any pattern is similar to that word pattern then index value is calculated for that pattern and pattern with highest index value is taken
- U and V matrix are left singular and right singular vectors respectively.$\sum$ represent the scalar matrix whose diagonal elements are non-zero. The relationship between them is given as: $A= U \sum V^T$

Step3:Combination

- Delete the terms appear in both matrix
- If S is less than M where M is the maximum number of sentences then go to step 2.

Step 4:Stop

The similarity of extracted document and concept of original document is determined by calculating cosine distance between two vectors,concept vector and document vector. There are so many methods for summary evaluation.The content based method is more efficient. Content based method used cosine distance formula.The similarity between two documents is calculated using this method .the formula is given below

$$Cos(x,y) = \frac{\sum x_i \times y_i}{\sqrt{\sum (x_i)^2} \times \sqrt{\sum (y_i)^2}}$$

If the cosine is distance is minimum then the document concept is nearly similarThe frequency of particular word in a particular document through the inverse proportion of the word over the entire document is calculated using term frequency inverse document frequency method.in this matrix rows represent the words or terms and column represent sentences$(tf\text{-}idf)_{ij}=tf_{ij}*\log_2(N/df_i)$

The better result can be obtained using cross document structure theory(CST)for evaluation than cosine distance. CST is used to check the relation between two documents.Three relations used are mentioned below

Identification: This relation identifies the sentences which occurs more than one document.

Subsumption: This relation gives the sentence which contains same and some extra information in other document.

Overlap: This gives non trivial components of sentences.

## IV. RESULTS

Experiment no:1

In this two constraints are considered file size and number of tokens. Number of tokens depends on number of words in file size. Non-linear graph is obtained as two constraints are compared. In some cases number of tokens decreases even if file size is larger

| File | File size | Number of tokens |
|---|---|---|
| File 1 | 1000 | 735 |
| File 2 | 450 | 237 |
| File 3 | 340 | 246 |

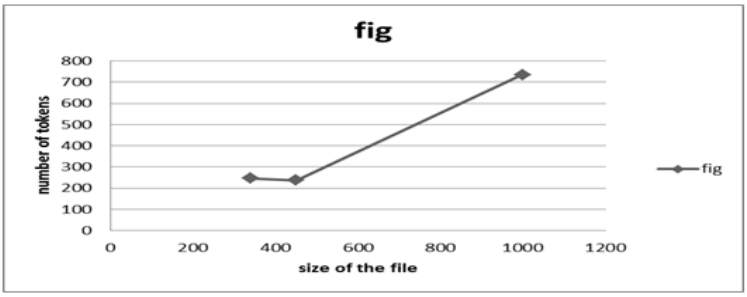


**Fig No. 2 File SizeVersus Tokens Generated**

Experiment 2:

In this two constraints are considered number of files and accuracy. Accuracy of summary depends on number file. There is inverse relationship between accuracy and number of files. As number of files increases, accuracy decreases

| Number of files | Accuracy |
|---|---|
| 2 | 85 |
| 4 | 70 |
| 6 | 65 |
| 10 | 58 |
| 14 | 49 |

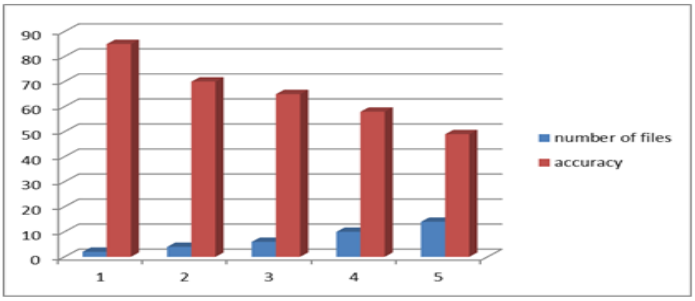


**Fig No. 3 Number of files versus Accuracy**

## V. CONCLUSION

In this LSA is used for sentence extraction along with tf-idf matrix which is used to calculate occurrences of term in document.SVD is used for decomposing the matrix.Evaluation of summary is obtained using CST relations to get better result.

## REFERENCES

[1] SoniyaPatil, Ashish T. Bhole "Multi-Document English Text Summarization using Latent Semantic Analysis", International Journal of Scientific & Engineering Research, Volume 5, Issue 12, December-2014 ISSN 2229-5518.

[2] Alan Turing "Computing Machinery and Intelligence",1950.

[3] "A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure" Dragomir R. Radev 550 E. University St. University of Michigan Ann Arbor, MI 48109.

[4] H.P.Luhn,"The automatic creation of literature abstracts". IBM Journal, 2:159–165,1958.

[5] H. P. Edmundson. "New methods in automatic extracting."J. ACM, 16(2):264-285,April 1969.

[6] M. Sanderson, \Accurate user directed summarization from existing tools," in Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM98), 1998.

[7] Asef, Kahani, Yazdi and Kamyar,\Context-Based Persian Multi-document Summarization",IEEE International Conference,2011, pp.145-149.