# Automatic Text Summarization

Trun Kumar


Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, Odisha, India

# Automatic text summarization

*Thesis report submitted in partial*
*fulfillment of the requirements for*
*the degree of*

Bachelor of Technology

*In*

Computer Science and Engineering

*By*

Trun Kumar

(Roll No. 110cs0127)

*Under the guidance of*

Prof. K. Sathya babu


Department of Computer Science and Engineering
National Institute of Technology Rourkela
Odisha, 769 008, India

Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela-769 008, India. www.nitrkl.ac.in

Dr. K. Sathya Babu
Professor

May 05, 2014

# Certicate

This is to certify that the work in the project entitled **Automatic Text summarization** by Trun Kumar being roll no 110cs0127 is a record of his work carried out under my supervision and guidance in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering.

Professor K. Sathya Babu
Department of Computer Science & Engineering

# Acknowledgment

I take this opportunity to express my gratitude and regards to my guide Prof. K. Sathya Babu for his exemplary guidance, monitoring and constant encouragement throughout the course of this project.

I also take this opportunity to express a deep sense of gratitude to my friends for their support and motivation which helped me in completing this task through its various stages.

I am obliged to the faculty members of the Department of Computer Science & Engineering at NIT Rourkela for the valuable information provided by them in their respective elds. I am grateful for their cooperation during the period of my assignment.

Lastly, I thank my parents for their constant encouragement without which this assignment would not have been possible.

Trun Kumar

# Abstract

**Automatic summarization** is the process of reducing a text Document with a computer program in order to create a summary that retains the most important points of the original document. As The problem of information overload has grown, and as the quantity of data has increased, so has interest in automatic summarization. It is very difficult for human beings to manually summarize large documents of text. Text Summarization methods can be classified into 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. The importance of sentences is decided based on statistical and linguistic features of sentences. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. The extractive summarization systems are typically based on techniques for sentence extraction and aim to cover the set of sentences that are most important for the overall understanding of a given document.

In frequency based technique obtained summary makes more meaning. But in k-means clustering due to out of order extraction, summary might not make sense.

# Contents

# 1. Introduction

## 1.1 Objective

With the growing amount of data in the world, interest in the field of automatic summarization generation has been widely increasing so as to reducing the manual effort of a person working on it. This thesis focuses on the comparison of various existent algorithms for the summarization of text passages.

# 1.2Application

Automatic summarization involves reduces a text file into a passage or paragraph that conveys the main meaning of the text. The searching of important information from a large text file is very difficult job for the users thus to automatic extract the important information or summary of the text file.

This summary helps the users to reduce time instead Of reading the whole text file and it provide quick Information from the large document. In today's world to Extract information from the World Wide Web is very easy. This extracted information is a huge text repository.

With the rapid growth of the World Wide Web (internet), information overload is becoming a problem for an increasing large number of people. Automatic summarization can be an indispensable solution to reduce the information overload problem on the web.

## 1.3 Overview of the Project

Generally, there are two approaches to automatic summarization: *extraction* and *abstraction*. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary.

First clean the text file by removing full stop, common words (conjunction, verb, adverb, preposition etc.). Then calculate the frequency of each words and select top words which have maximum frequency. This technique retrieves important sentence emphasize on high information richness in the sentence as well as high Information retrieval. These related maximum sentence generated scores are clustered to generate the summary of the document. Thus we use k-mean clustering to these maximum sentences of the document and find the relation to extract clusters with most relevant sets in the document, these helps to find the summary of the document.

## 1.4 Organization of the Thesis

In **Chapter 2**, I have given the Literature Survey which includes the review of extraction based approaches and what are the existing algorithms in these approaches.

In **Chapter 3,** I have discussed the work proposed work and algorithms used in the implementation.

In **Chapter 4,** results of the various implemented algorithms and as well as the results are analyzed.

In **Chapter 5,** I have the results conclusions drawn from the results as well as the scope for future work is discussed.

# 2 Literature Review

Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text.

## 2.1 Frequency based approach

### 2.1.1 Term Frequency:

The term frequency is very important feature. TF (term frequency) represents how many time the term appears in the document (usually a compression function such as square root or logarithm is applied) to calculate the term frequency. The term identifying sentence boundaries in a document is based on punctuation such as (. (, “, [, {, etc.) and split into sentences. These sentences are nothing but tokens.

## 2.1.2 Keyword Frequency

The keywords are the top high frequency words in term sentence frequency. After cleaning the document calculate the frequency of each world. And which words have the highest frequency these words are called keywords. The words score are chosen as keywords, based on this feature, any sentence in the document is scored by number of keywords it contains, where the sentence receives 0.1 score for each key word.

## 2.1.3 Stop word filtering:

In any document there will be many words that appear regularly but provide little or no extra meaning to the document. Words such as 'the', 'and', 'is' and 'on' are very frequent in the English language and most documents will contain many instances of them. These words are generally not very useful when searching; they are not normally what users are searching for when entering queries

# 2.2 K means clustering approach

## 2.2.1 K- means clustering

***K-means*** **clustering** is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k \le n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where $\boldsymbol{\mu}_i$ is the mean of points in $S_i$.

# 3 Proposed Work

## 3.1 Frequency Based Approach

First clean the document means remove the stop words from the document. Then count the frequency of each word in remaining text file by comparing select world with the each word. Then select the keyword which have highest frequency. After that select the sentences which have these keywords.

### 3.1.1 Frequency detection method

In this technique, we first eliminate commonly occurring words and then find keywords according to the frequency of the occurrence of the word. This assumes that if a passage is given, more attention will be paid to the topic on which it is written, hence increasing the frequency of the occurrence of the word and words similar to it. Now we need to extract those lines in which these words occur since the other sentences wouldn't be as related to the topic as the ones containing the keywords would be. Thus, a summary is generated containing only useful sentences.

### 3.1.2 Keyword Frequency method

This algorithm takes the previous algorithm to a further level. This takes into account facts such as the first few words of an article has more weights as compared to the rest, since they represent the first paragraph generally contains a gist of what is being said in the rest of the article. Secondly, it also takes into account the frequency of occurrence of keywords obtained in the previous algorithm in a particular sentence. Higher the keyword count within a sentence, more is its relevance to the topic at hand.

## 3.2 Using k-**means clustering**

Clusters are nothing but grouping the similar sentences together. Thus we have many clustering technique, in that one of the technique is k-mean clustering. The main region to use k-mean clustering is to group all the similar set of sentences together by cumulative similarity, and divide the document into k-clusters is to fine k centroids for each cluster. These centroids are placed in different location (not arranged properly) defines different result. Thus we go to next step to place them properly according to the given data and to group the nearest centroid. Thus we repeat this step until the complete grouping is done to the entire text file.

At this point we have to re-calculate k new centroids as center of previous step clusters. These k new centroids build the new data set points of nearest new centroid. As the loop is generated the k-

centroids change their location step by step until no more changes are done.

# 3. Result and Analysis

In extraction type of summarization often addressed in the literature are key world extraction, where the goal is to select individual words to "tag" a document, and document summarization, where the goal is to select whole sentences to create a short paragraph summary.

Using k-means clustering technique we conducted multiple experiments with the sentence clustering based document summarization. For each experiment, input data sets are preprocessed by removing stop words, but no stemming is applied. Each experiment deals with a summarization method in which sentence clustering algorithm remains the same, but cluster ordering and representative selection techniques are replaced with the possible alternatives to judge whether summarization performance depends on the cluster ordering and representative selection techniques.

In frequency based technique obtained summary makes more meaning. But in k-means clustering due to out of order extraction, summary might not make sense.

In frequency based technique extraction of important sentences involving keyword is tough. But in k-means clustering extraction of keyword related topics might be one of its most important strength.

In frequency based technique complexity of implementation is low but in k-means clustering complexity of implementation is very high.

In comparison, the keyword frequency based summary generation algorithm has been found to be very simple where complexity is concerned. And generally it has found that this method provides a much better summary than the other two methods, though this will depend on the text given in hand. Meaning of the summary generated by this method is usually higher than k-means clustering algorithm for generating summary since the sentences are extracted in the same order as in the given article.

# 5. Conclusion and future work

In frequency based technique obtained summary makes more meaning. But in k-means clustering due to out of order extraction, summary might not make sense.

The effective diversity based method combined with K-mean Clustering algorithm to generating summary of the document. The clustering algorithm is used as helping factor with the method for finding the most distinct ideas in the text. The results of the method supports that employing of multiple factors can help to find the diversity in the text because the isolation of all similar sentences in one group can solve a part of the redundancy problem among the document sentences and the other part of that problem is solved by the diversity based method.

In future work abstractive methods can be implemented. In abstractive method build an internal semantic representation and then use natural language generation techniques to create a summary.

# 6. Bibliography


[1] Fang Chen, Kesong Han and Guilin Chen, “An Approach to sentence selection based text summarization", Proceedings of IEEE TENCON02, 489-493, 2002.

[2] A. Kiani –B and M. R. Akbarzadeh –T, “Automatic Text Summarization Using: Hybrid Fuzzy GA-GP”, IEEE International Conference on Fuzzy Systems, 16-21 July, Vancouver, BC, Canada,977 -983, 2006.

[3] C. Jaruskulchai and C. Kruengkrai, “Text Summarization Using Local and Global Properties”, Proceedings of the IEEE/WIC International Conference on web Intelligence, 13-17 October. Halifax, Canada: IEEE Computer Society, 201-206, 2003.

[4] D. R Radev. H. Jing, and M. Budzikowska. Centroid-based summarization of multiple documents: Sentence extraction, utility based evaluation, and user studies. *In ANLP/NAACL Workshop on Summarization*, Seattle, April, (2000).

[5] M. Osborne. Using Maximum Entropy for Sentence Extraction. *In ACL Workshop on Text Summarization*, (2002).

[6] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, "Automatic Text Summarization using a Machine Learning Approach”, Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol 2507/2002, 205-215,2002.

[7] H. P. Edmundson.,”New methods in automatic extracting”, Journal of the ACM, 16(2):264-285, April 2008.