# AUTOMATED DOCUMENT SUMMARY GENERATION USING UNSUPERVISED MACHINE LEARNING TECHNIQUES

Divya Chandran[1], ShobhaRani G[2]
DM, BSTC, BEL, Bangalore, India[1]
DGM, BSTC, BEL, Bangalore, India[2]

***Abstract***:

Currently we are in digital information era and entire world is running towards capturing and analyzing the enormous data feeds from various sources. This competition is towards deriving the maximum and unique intelligence that can discover unknown facts. Document summary extraction by manual method i.e. human analysis is very difficult task, as it dependents on individual perceptions, frequency of feeding documents as well as priority assigned to various sub topics of the document. Therefore automatic summarization of the incoming information becomes predominant in this digital world. As the system automation enables us to extract the required information on the flow and also synthesize with already available data, this reduces the effort to compress the original document and extract only essential information. This paper has evolved out of the study done on various methods / techniques of solving the text summarization and information extraction from document stock. Main focus of this paper is to apply one of the unsupervised learning techniques, called 'clustering' with file pre-processing and result storage. This method helps to process the data set through certain number of clusters and find the categorized data sets.

***Keywords*—Text Summarization, Tokenization, TF- IDF**

## I. INTRODUCTION

Precise information at precise time is the key for success in any domain. With huge volume of data flowing in real time, extracting the relevant information is a challenge for domain users. Also in the near real / real war scenario, plethora of inputs coming into the systems creates a confusion scenario to the commanders. In military domain the intelligence comes from various sources like human intelligence, electronic intelligence, open source intelligence etc. and in various formats like docs, pdf, images. Intelligence from images can be obtained using image processing techniques (which is out of scope for this paper). With system automation, we can extract the most important information from a document, within no time helping the commanders to take quick decisions, which is called *text summarization*.

Text summarization is the process of automatically creating a compressed version of a given document which provides the most useful information for the user. Basically two methods of summarization is available Extraction and Abstraction

Extraction: - Choosing the set of important sentences from the document there by creating a subset of the original document.

Abstraction: - Summary generated by reframing the sentences available from the main document.

When a military domain is concerned the data format is mostly fixed, and are not allowed to perform any alteration/ reframing of the sentence.

So in this paper, we are using extraction based text summarization technique to generate summaries for the documents uploaded. The generated summaries will be saved in data base with time stamp for further analysis. We have used K-means clustering algorithm for extracting the document summary. The text summarization consist of 4 steps

1. Uploading a file
2. File preprocessing
3. Sentence score calculation
4. Applying clustering algorithm for summarization

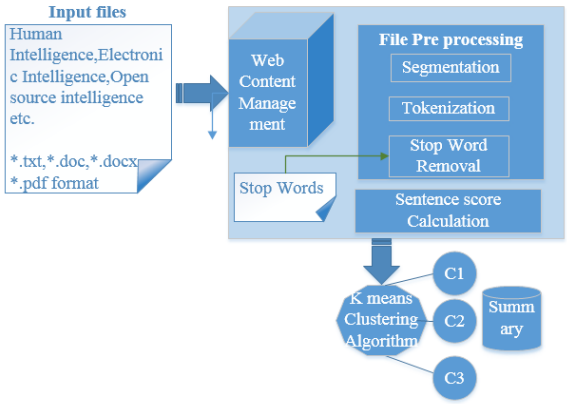


**Fig 1: Summarization Frame work**

## II. UPLOADING FILE

Web content management is a system for organizing and facilitating collaborative creation of documents and other content. As and when the file is selected, which will be saved to Web content management for future reference and the summarization algorithm will be invoked. User can select *.txt,*. pdf,*.doc,*.docx format file for extracting summary. The actual file will be available in content management system and the summary generated will be saved in data base.

## III. FILE PREPROCESSING

Once the file is uploaded and summary generation algorithm is invoked, a set of preprocessing operations is performed on the file, which is Tokenization and Stop Word removal.

Tokenization is the extraction of individual tokens or words from the document. Each sentence is separated by splitting, using a full stop (.), question mark (?) or an exclamatory mark (!).The sentences are maintained in the same order, in which it appear in the original document for future reference.

Each sentence is made up of a set of connecting phrases such as "and", "the", "but" etc. which will not contribute any value to the overall sentence score and need to be removed. This procedure is called *stop word removal*. A file is created with all stop words which normally appears in any statements. Each sentence from the actual document is compared with the stop words present in the file, and will be removed. After the stop word removal, the sentences will have only the tokens which will provide some meaning. Each token from the sentence will be extracted using white space.

Operations done in file pre processing

a) Separate all sentences of the document.
b) Remove stop words from the sentence.
c) Extract each token from the sentence.
d) Calculate the total length of the document with and without stop words.

## IV. SENTENCE SCORE CALCULATION

This step finds out the goodness of a sentence for to be the part of generated summary. Each sentence is given a score. Score for each sentence is calculated using TF-IDF sum. TF-IDF calculates the importance of each key word in a document.

- TF (Term frequency):- Calculate number of times a particular word appears in a document.

TF (t) = Total number of times a token appear in the document / Total length of the document.

All terms of the document are considered equally important.

- IDF (Inverse Document Frequency):- Finds out the actual importance of the token. During the calculation of TF each word frequency is divided by total length of the document. In IDF each term frequency is divided by the document length excluding the stop words.

IDF (t) = log (Length of document without stop words/ Total number of times a token appear in the document).

- TF-IDF weighting: - Combine the value of TF and IDF to produce the weightage of a particular term.

TF-IDF = TF * IDF

Summing up the TF-IDF value of all individual terms will produce the overall weightage of the sentence.

The tf-idf equation followed in our summarization procedure is

Tft = f (t, d)/ f (d)

t: Token, d: Document, f (t, d): frequency of t in d, f (d): frequency of every term in d

Idf (t) = log10 (N/f (t, d))

»N: No of sentence in the document

tf-idf (t) = Tf * Idf

Score of sentence X

$$\text{Score}(X) = \sum_t tf - idf(t)/|X|$$

|X| - sentence length

t - term in sentence.

On completion of this step, algorithm will be aware of the weightage of each sentence. The score of each sentence act as the input for K-mean clustering algorithm, and will be invoked.

## V. UNSUPERVISED MACHINE LEARNING TECHNIQUES

Machine Learning, a branch of Artificial intelligence is mainly divided in to two categories as 'Supervised learning' and 'Unsupervised learning'. Supervised learning algorithms learn from the training data set, and are used in predictive modeling. In predictive modeling we can predict the output value based on the input data. Supervised learning algorithms are further divided in to

1. Classification algorithms:- here the target output is based on some category such as "Red" or "Blue" or "Dog" or "Cat"
2. Regression algorithms: - here the target output variable is a real time value such as "Rupees", "weight" etc.

Unsupervised algorithms:- In unsupervised learning there is no target or output variable is defined, the algorithm learns by itself and presents interesting structures in the data.

Unsupervised algorithms are further grouped in to

1. Association problems:- algorithm wants to discover rules which describes large portion of data. e.g.:- buy stock X or Y.
2. Clustering: - it's the process of partitioning a set of data points in to number of clusters or simply the process of organizing objects in to a set of groups whose members are similar in some ways.

Figure 2 depicts various machine learning models.

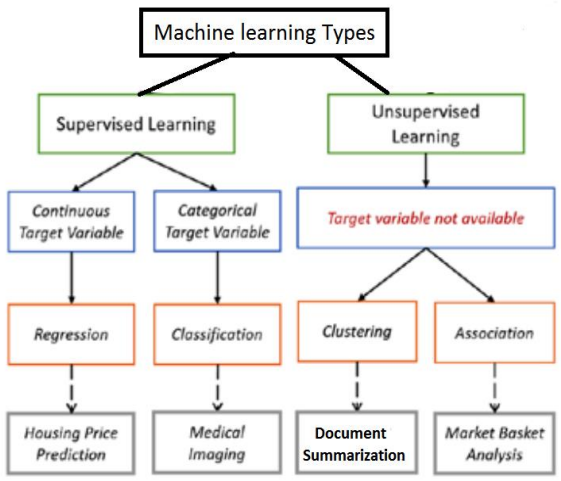


**Figure 2 Machine Learning Models**

K-means or Lloyd's algorithm is the most commonly used unsupervised clustering

algorithm. In general if we have n data points x1,x2,x3….xn we will be partitioning to k clusters S = {S1,S2,S3…Sk}. The goal is to assign a cluster to each data points in such a way that the distance to each data points from the cluster will be minimal by minimizing the within-cluster sum of squares (WCSS).

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

Where μi is the mean of points in Si.

K-means is a partition clustering approach with the following basic steps.

1. Each cluster is associated with a centroid (center point).
   Note: No of clusters, K must be specified. Initial centroids are often chosen randomly. Clusters will vary for each run.

2. Repeat
   a. Associate each point to the closest centroid. Closest point is calculated using Euclidean distance.

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2}$$

Most of the point convergence happens during the first little iteration.

   b. Re compute the centroid of each cluster.
3. Until the centroid don't change.

K-means algorithm classifies the sentences to different clusters based on the score. The densest cluster is taken as the summary because the densest cluster will have the sentences with highest score which depicts that they are very important. The summary generated will be saved in database along with time stamp.

The users will be provided with summary and a link to the actual document saved in content management system, thereby helping them to understand an overall picture of the document available without wasting precious time. If the user wishes to have a complete understanding of the document he can download the file and can update the automated summary generated by the system by adding his views and understandings about the document, which will be saved back to the data base with new time stamp.

## VI. BENEFITS OF THE CURRENT METHODOLOGY

K-means produces precise summary since the densest cluster contains the sentence with highest score.

**Evaluation:** As per the requirement the summary generated should be between 20% - 50% of compression from the parent document. The algorithm produced a summary between 20% - 35 % for large documents of size > 500 sentences and between 35% - 50% for documents having sentences up to 500.

| SI No | No. of sentence in original document | No. of sentence in generated summary | % of sentence reduction |
|---|---|---|---|
| 1 | 24 | 10 | 41 % |
| 2 | 180 | 67 | 37 % |
| 3 | 656 | 217 | 33 % |
| 4 | 2052 | 610 | 29 % |

**Table 1 Experiment Results**

Main advantage of using this technique is to accept the documents in as–is form and applying the clustering algorithm on each document input. This scenario being continuous, user need not be aware of what information is coming at what frequency from which source. Though there are various methods being tried by many researches on text summarization using clustering algorithms, this approach stands aside by finding the most distinct ideas in the text using densest cluster identification and auto collation. The advantage of this methodology is continuous extraction that is building up a pseudo-training

set, which can be applied as feedback to the system. This is directing us to move from unsupervised to supervised learning techniques.

## VII. CONCLUSION

The flexibility given to the user to view and compare the summary generated with parent document not only eases the understanding process but also develops the confidence on the summarized information. As explained in the paper, topping up the database with summarized information holding unique timestamp allows user to view the real time raw scenario inputs in the form of collation tree. The method brought out by this paper has been experimented on various sets of documents and compared the performance against each iteration. Evaluation results are compared with the benchmark methods. Future plan is to extend the work towards deep learning techniques along with Natural language processing methods, by combing the text summarization results with the information extracted from image processing.

## VIII. REFERENCES


[1] "The Document Understanding Conference (DUC)",http://duc.nist.gov

[2] S. Brin, and L. Page, "The anatomy of a large-scale hyper textual Web search engine", Computer Networks and ISDN System, 30(1–7): 107–117, 1998.

[3] W. Kraaij, M. Spitters and M. v. d. Heijden, "Combining a mixture language model and naïve bayes for multi-document summarization", Proceedings of Document Understanding Conference, 13-14 September, New Orleans, LA, 109-116, 2001.

[4] K Filippova, M. Mieskes, V. Nastase, S. P. Ponzetto and M. Strube."Cascaded Filtering for Topic-Driven Multi-Document Summarization", Proceedings of the Document Understanding Conference, 26-27 April. Rochester, N.Y., 30-35, 2007.

[5] 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.

[6] Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In Proceedings of HLT-NAACL

[7] H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008.

[8] Helmuth Späth. 1980. Cluster Analysis Algorithms for Data Reduction and Classification of Objects.Halsted Press.

[9] [A. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review",ACM Computing Surveys.31 (3), 264-323, 1999.

[10] Alsabti, K., Ranka, S., and Singh, V. An efficient K-means clustering algorithm. In Proceedings of the First Workshop on High-Performance Data Mining, Orlando, Florida, 1998; ftp://ftp.cise.ufl.edu/pub/faculty/ranka/Proceedings.