International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Text summarization using cosine similarity and clustering approach

Miss.Sushma Pawar and Prof. Dr.Sunil Rathod

Department of Computer Engineering DYPSOE,Lohegaon Pune Savitribai Phule Pune University

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

The document summarization is becoming essential as lots of information getting generated every day. Instead of going through the entire text document, it is easy to understand the text document fast and easily by a relevant summary. Text summarization is the method of explicitly making a shorter version of one or more text documents. It is a significant method of detecting related material from huge text libraries or from the Internet. It is also essential to extract the information in such a way that the content should be of user's interest. Text summarization is conducted using two main methods extractive summarization and abstractive summarization. When method select sentences from word document and rank them on basis of their weight to generate summary then that method is called extractive summarization. Abstractive summarization method focuses on main concepts of the document and then expresses those concepts in natural language. Many techniques have been developed for summarization on the basis of these two methods. There are many methods those only work for specific language. Here we discuss various techniques based on abstractive and extractive text summarization methods and shortcomings of different methods.

Keywords*: Text Summarization, extractive summary, information extraction*

Introduction

With increasing amount of data it becomes more and more difficult for users to derive material of interest, to search efficiently for specific content or to gain an overview of influential, important and relevant material. In today's information technology number of people are searching for informative data on web, but every time it is not possible that they could get all relevant data in single document, or on a single web page. They could get number of web pages as a search result [5]. This problem has given the new solution that is associated to data mining and machine learning which returns query specific information from large set of offline documents and represents as a single document to the user. So, automated summarization is an important area in Natural Language Processing (NLP) research. Automated summarization provides single document summarization and multi-document summarization [3].

A. Multi-Document Merger

The merging of data from multiple documents is called multidocument merger. Data is found in unstructured or structured form and many times we have to generate summary from multiple files in less time, so, multi-document merger technique is useful. Multi-document summarization generates information reports that are both concise and comprehensive. With different opinions being put together, every topic is described from multiple perspectives within a single document. The goal of a brief summary is to simplify information search and save the time by pointing to the most relevant information. Text summarization is gaining much importance currently. One reason for this is, due to the rapid growth in material, requirement for involuntary text summarization has enlarged. It is very difficult for human beings to manually summarize big text documents. There is a profusion of text material available on the internet. However, usually the Internet offers more material than is required. Therefore a problem of repetition is encountered: examining for similar kind of documents through a large amount of documents is very tedious task [3]. The aim of text summarization is to reduce the source text into a shorter form preserving its information content and overall meaning. If sentences in a text document were of equivalent significance, creating a summary would not be very effective. With different opinions being put together & outlined, every topic is seen and described from multiple perspectives within a single document. While the main aim of a brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, multidocument summary should itself contain the required information, hence limiting the need for accessing original files to cases when refinement is required. In this study various techniques for sentence based

extractive summarization has been encountered also various similarity measures and their comparisons.
Extractive summarizer aims at selecting the foremost relevant sentences within the document whereas maintaining a reduced redundancy within the outline. It is created by reusing portion (word, sentences etc.) of input text verbatim.

Example: Search engines typically generate extractive summaries from web pages.

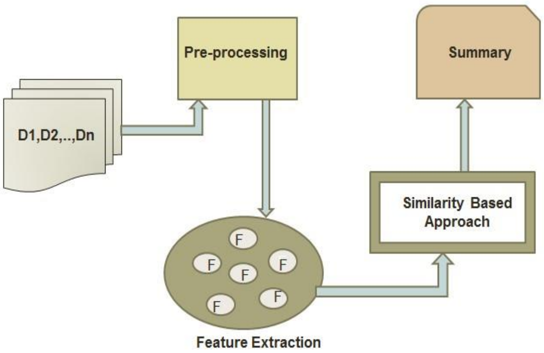

Fig.1. Generalized structure of document summarization

Abstractive text summarization methods employ more powerful natural language processing techniques to interpret text and generate new summary text, as opposed to selecting the most representative existing excerpts to perform the summarization. In this method, information from source text is re-phrased. But it is harder to use because it provides allied problems such as semantic representations.

Example: Book Reviews-if we want a summary of book The Lord of The Rings then by using this method we can make summary from it.

Literature Survey

An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms proposes an improved extractive text summarization method for documents by enhancing the conventional lexical chain method to produce better relevant information. Then, Author firstly investigated the approaches to extract sentences from the document(s) based on the distribution of lexical chains then built a transition probability distribution generator(TPDG) for n-gram keywords which learns the characteristics of the assigned keywords from the training data set. A new method of automatic keyword extraction also featured in the system based on the Markov chains process. Among the extracted n-gram keywords, only unigrams are selected to construct the lexical chain. [1]
Top- K ensemble ranking algorithm is used to rank sentences TF-IDF (term frequency-inverse document frequency) is used to word count and word level feature extraction. In paper [2] author first extracted multiple candidate summaries by proposing several schemes for improving the upper-bound quality of the summaries. Then, they proposed a new ensemble ranking method for ranking the candidate summaries by making use of bilingual features. Extensive experiments
have been conducted on a benchmark dataset.[2]

Automatic text summarization within big data framework demonstrates how to process large data sets in parallel to address the volume problems associated with big data and generate summary using sentence ranking. TF-IDF is used for document feature extraction. MapReduce and Hadoop is used
to process big data.[3]

Extractive document summarization based on hierarchical GRU (Gated Recurrent Unit) proposes two stage structure: 1) Key sentence extraction using Levenshtein distance formula. 2) Recurrent neural network for summarization of documents. In extraction phase system conceives a hybrid sentence similarity measure by combining sentence vector and Levenshtein distance and integrates into graph model to extract key sentences. In the second phase it constructs GRU as basic block, and put the representation of entire document based on LDA (Latent Dirichlet Allocation) as a feature to support summarization.[4]

Extractive algorithm of English text summarization for English teaching is based on semantic association rules. To summarize documents semantic association rule vectors is used. In this paper relative features are mined among English text phrases and sentences, the semantic relevance analysis and feature extraction of keywords in English abstracts are realized. [5]

Fairness of extractive text summarization is the first work that introduces the concept of fairness of text summarization algorithms. Author shows that while summarizing datasets having an associated sensitive attribute, one needs to verify the fairness of the summary. Especially, with the advent of neural network-based summarization algorithms (which involve super-wised learning), the question of fairness becomes even more critical. Author believe that this work will open up interesting research problems, e.g., on developing algorithms that will ensure some degree of fairness in the summaries. [6]

Automatic text summarization by local scoring and ranking for improving coherence approach provides automatic feature based extractive heading wise text summarizer to improve the coherence thereby improving the understandability of the summary text. It summarizes the given input document using local scoring and local ranking that is it provides heading wise summary. Headings of a document give contextual information and permit visual scanning of the document to find the search contents. The outcomes of the experiment clearly show that heading wise summarizer provides better precision, recall and f-measure over the main summarizer, MS-word summarizer, free summarizer and Auto summarizer [7]. A paper on data merging by Van Britsom proposed a technique based on use of NEWSUM Algorithm. It is a type of clustering algorithm which divides a set of document into subsets and then generates a summary of correlated texts. It contains three phases: topic identification, transformation and summarization by using different clusters [8]. A novel technique for efficient text document summarization as a service by Anusha Banalkotkar, represents the different techniques that explain as the main two fundamental techniques are identified to automatically summarize texts i.e. abstractive summarization and extractive summarization. Complex summarization technique (cohesive, readable, intelligible, multi-disciplinary approaches, machine learning) all are coming under this paper.[9] Multi-document summarization using sentence clustering by Virendra Kumar Gupta states that sentences from single document summaries are clustered and top most sentences from each cluster are used for creating multi-document summary. The model contains the steps as preprocessing, noise removal, tokenization, stop words, stemming, sentence splitting and feature extraction. After performing these steps, important sentences are extracted from each cluster.[10]

System Architecture

Proposed system involves the different module to generate the summary for given multiple documents. Previous system has some drawback such that it can take only text file as input. If we give other files such as PDF or word file as input then it cannot accept that file and shows the message only text files are allowed. To overcome these problems we proposed a new system that takes the input as text, PDF and word files. The system involves the following basic three phases.

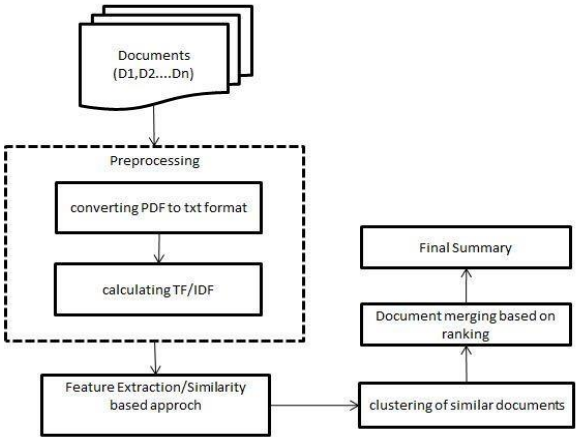
Documents (D1,D2....Dn)
Preprocessing
converting PDF to txt format
calculating TF/IDF
Feature Extraction/Similarity based approch
clustering of similar documents
Document merging based on ranking
Final Summary

Fig 2. System design for proposed system

A. Module 1: Processing of Input multiple documents Tokenization

a) Tokenization:

It breaks the text into separate lexical words that are separated by white space, comma, dash, dot etc. [3]

b) Stop Word Removal:

Stop words are words which are filtered out before or after processing of natural language data (text).Stop word removal is helpful is keyword searching. [2]

c) Stemming Suffixes:

Here enlisted suffixes are removed for topic detection Example

Plays and playing = Play where “s” and “ing” are suffixes added to topic play need to be removed for accuracy purpose.

B. Module 2: Similarity based documents extraction from multiple documents

Cosine Similarity Approach:

Cosine Similarity measures the similarity between two sentences or documents in terms of the value within the range of [-1, 1] whichever you want to measure. That is the Cosine Similarity. Cosine Similarity extracted TF and IDF by using following formulae:

TFIDF:

TF (term, document) = Frequency of term / No of Terms

$$tfi = (\frac{ni}{\sum knk}) \quad \text{............ (i)}$$

IDF (inverse document frequency) calculates whether the word is rare or common in all documents. IDF (term, document) is obtained by dividing total number of Documents by the number of documents containing that term and taking log of that.

IDF (term, document) = log (Total No of Document / No of Document containing term)

$$idfi = \log(\frac{D}{\{d:t\epsilon d\}}) \quad \text{.............(ii)}$$

TF-IDF is the multiplication of the value of TF and IDF for a particular word. The value of TF-IDF increases with the number of occurrences within a document and with rarity of the term across the corps

$$tfidf = tf * idf$$

f .... (iii)

Thus, the Tf-idf weight is the product of these quantities TF-IDF = 0.03 * 1 = 0.03.

Given a document containing terms with given frequencies:

A(3), B(2), C(1) and total number of terms in document are 15.

Assume collection contains 10,000 documents and document Frequencies of these terms are:

A(50), B(1300), C(250).

Then, using above equation (i), (ii) and (iii) we calculate the tf-idf for terms A, B and C.

A: $tf = \frac{3}{15} = 0.2$; $idf = \log\left(\frac{10000}{50}\right) = 7.6$;
$tfidf = 0.2 * 7.6 = 1.52$

B: $tf = \frac{2}{15} = 0.133$; $idf = \log\left(\frac{10000}{1300}\right) = 2.9$
$tfidf = 0.133 * 2.9 = 0.38$

C:$tf = \frac{1}{15} = 0.066$;$idf = \log\left(\frac{10000}{250}\right) = 5.3$;
$tfidf = 0.066 * 5.3 = 0.35$

;

C. Module 3: Summary Generation

After checking similarity based approach and relevancy of documents, relevant sentences are extracted and merge the relevant sentences into one by using cosine similarity approach. Thus after merging the data it generates a final summary.

Results And Discussion

The performance analysis will be evaluated to prove the effectiveness of the proposed methodology in terms of the comparison with the existing system.

For result analysis we are going to create our own dataset and used one standard dataset OpinosisDataset1.2.

Below graph1 shows the comparative analysis between two clustering methods, first is WeightSum method which generates the cluster based on the weight of documents, and other one is threshold clustering. Threshold clustering is proposed clustering algorithm. Here result and analysis is performed on OpinosisDataset1.2.

Table 1: Accuracy of clustering methods

Clustering methods	**Accuracy (%)**
WeightSum	80
Threshold	91

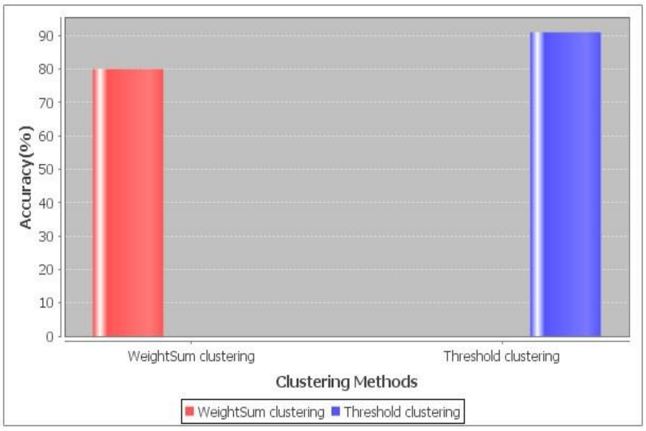

Fig. 3: Accuracy of clustering methods

Accuracy: It is the degree to which the result of a measurement, calculation, or specification conforms to the correct value (true). The proposed system gives maximum 83% Accuracy. To calculate the accuracy in percentage we can multiply by 100 to the result.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where,
TP (True Positive) – Correctly Identified.
FP (False Positive) – Incorrectly Identified.
TN (True Negative) – Correctly Rejected.
FN (False Negative) – Incorrectly Rejected.

Below graph 8.2 shows the accuracy of a proposed system on variable number of documents. The size of all documents is between 2 kb to 10kb. Here for calculating result we took first dataset that contains total 6 documents in that 1 file is irrelevant and 5 files are relevant and system merged total 4 documents.

TP – 4, FP – 0, TN – 1, FN – 1

$$Accuracy = \frac{4 + 1}{4 + 1 + 0 + 1} = 0.8333$$

Table 2: Accuracy with variable number of documents

Number of documents	**Accuracy (%)**
6	83.33
12	83.33
18	77.77

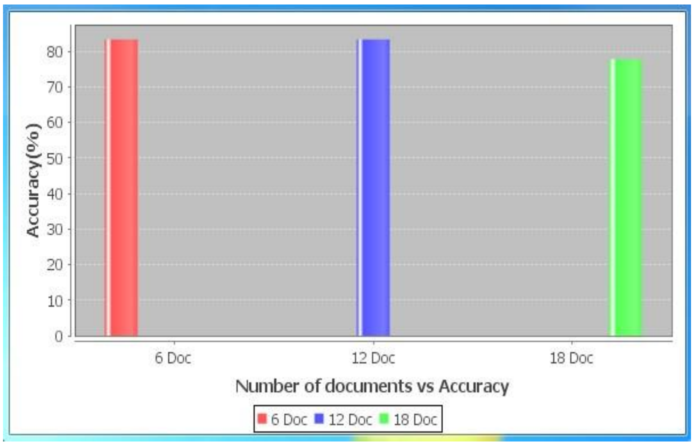

Fig 4: Accuracy with variable number of documents

Conclusion

The data can be retrieved by using the background knowledge for generalization. Now a day the growth of data increases in structured or unstructured form and we want a summary from that data in less time. To overcome the drawback of previous model we propose a new system. The work is under implementation and focuses on using different aspects of text mining to come up with an efficient approach that will aid the document creator with the draft format of the contents essentially conveying the important concepts of the text. So, multi-document merger summarization is used. It reduces our time and gives efficient output.

References

[1]. HtetMyet Lynn 1 , Chang Choi 2 , Pankoo Kim “An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms”, Springer-Verlag Berlin Heidelberg 2017

[2]. Xiaojun Wan 1 , FuliLuo 2 , Xue Sun Songfang Huang3 , Jin-ge Yao “Cross-language document summarization via extraction and ranking of multiple summaries” Springer-Verlag London 2018

[3]. Andrew Mackey and Israel Cuevas “automatic text summarization within big data frameworks”, acm

[4]. 2018

[5]. Yong Zhang, Jinzhi Liao, Jiyuyang Tang “Extractive Document Summarization based on hierarchical GRU”, International Conference on Robots & Intelligent System IEEE 2018

[6]. Lili Wan “Extractive Algorithm of English Text Summarization for English Teaching” IEEE 2018

[7]. Anurag Shandilya, Kripabandhu Ghosh, Saptarshi Ghosh “Fairness of Extractive Text Summarization”, ACM 2018

[8]. P.Krishnaveni, Dr. S. R. Balasundaram “Automatic Text Summarization by Local Scoring and Ranking for Improving Coherence”, Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication

[9]. Bagalkotkar, A., Kandelwal, A., Pandey, S., &Kamath, S. S. (2013,

[10]. August). “A Novel Technique for Efficient Text Document Summarization as a Service”, In Advances in Computing and Communications (ICACC), 2013 Third International Conference on (pp. 50-53). IEEE.

[11]. Ferreira, Rafael, Luciano de Souza Cabral, Rafael DueireLins, Gabriel Pereira e Silva, Fred Freitas, George DC Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. "Assessing sentence scoring techniques for extractive text summarization." Expert systems with applications 40, no. 14 (2013): 5755-5764.

[12]. Gupta, V. K., &Siddiqui, T. J. (2012, December). “Multi-document summarization using sentence clustering”, In Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on (pp. 1-5). IEEE.

[13]. Min-Yuh Day Department of Information Management Tamkang University New Taipei City, Taiwan myday@mail.tku.edu.tw Chao-Yu Chen Department of Information Management Tamkang University New Taipei City, Taiwan Intelligence for Automatic Text Summarization”,2018 IEEE International Conference on Information Reuse and Integration for Data Science

[14]. Xiaoping SunandHaiZhuge*, Senior Member, IEEE Laboratory of Cyber-Physical-Social Intelligence, Guangzhou University, China Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, University of Chinese Academy of Sciences, Chinese Academy of Sciences, China System Analytics Research Institute, Aston University, UK “Summarization of Scientific Paper through Reinforcement Ranking on Semantic Link Network” ,IEEE 2018

[15]. Ahmad T. Al-Taani (PhD, MSc, BSc) Professor of Computer Science (Artificial Intelligence) Faculty of Information Technology and

[16]. Computer Sciences Yarmouk University, Jordan. ahmadta@yu.edu.jo“Automatic Text Summarization Approaches” ,IEEE 2017

[17]. AlokRanjan Pal Dept. of Computer Science and Engineering College of Engineering and Management, KolaghatKolaghat, India chhaandasik@gmail.com DigantaSaha Dept. of Computer Science and

[18]. Engineering Jadavpur University Kolkata, India neruda0101@yahoo.com“An Approach to Automatic Text Summarization using WordNet”, IEEE 2014

[19]. Yue Hu and Xiaojun Wan “PPSGen: Learning-Based Presentation Slides Generation for Academic Papers” , IEEE transactions on knowledge and data engineering, vol. 27, no. 4, april 2015

[20]. Daan Van Britsom, AntoonBronselaer, and Guy De Tre “Using Data merging techniques for generating multidocument summarizations” , ieee transactions on fuzzy systems, vol. 23, no. 3, june 2015

[21]. NingZhong, Yuefeng Li, and Sheng-Tang Wu “Effective Pattern discovery for text mining”, ieee transactions on knowledge and data engineering, vol. 24, no. 1, january 2012