TFIDF Model based Text Summerization

Dr. Geetha C Megharaj[1], Ms. Varsha Jituri[2]
[1]Professor and Head of AIML Department, Sri Krishna Institute of Technology, B'lore-560090, India
[2]Assistant Professor, CSE Department, Sri Krishna Institute of Technology, B'lore-560090, India

***Abstract*- Text summarization is a process of automatically generating abstract summery having important sentences present in original text document. Text summarization makes a shorter version of the text document. The main strategies to perform Text summarization are extractive summarization and abstractive summarization. Extractive summary of the document is produced by selecting important sentences and word with high rank from the text document. On the other hand the sentences and words that are present in the summary produced through Abstractive method may not present in the original document.**

Text summarization has become the major requirement of many applications such as News Editing, Search Engine, Business Analysis, Market Review etc. Summarization helps to collect the required information in reduced time. This paper is an attempt to summarize the contents of webpage with the help of an extractive summarization approach viz. Term Frequency-Inverse Document Frequency (TFIDF)

Keywords- Term Frequency-Inverse Document Frequency (TFIDF), Stemming, Tokenization, Text Summarization, Natural Language Processing.

1. INTRODUCTION

With so much data, it's difficult for users to find what they are looking for, to scrutinize it thoroughly for exact content, and to get a sense of what's noteworthy, vital, and relevant. A vast number of individuals are searching the web for beneficial information in today's information technology, but it is uncommon that they will obtain all necessary information in a single document or web page. As a search result, they could acquire a large number of web sites [1]. Document summarizing is the process of mechanically rewriting a document into its smallest form while maintaining its important content. Over the years, many document summarizing models have been examined.

There are two broad categories of text summarization. Viz

1. Extractive
2. Abstractive

The Fig. 1.1 shows various summarization techniques. Extractive methods generate summaries by extracting elements of the original content, whereas abstractive methods may generate new words or phrases that were not present in the original source.

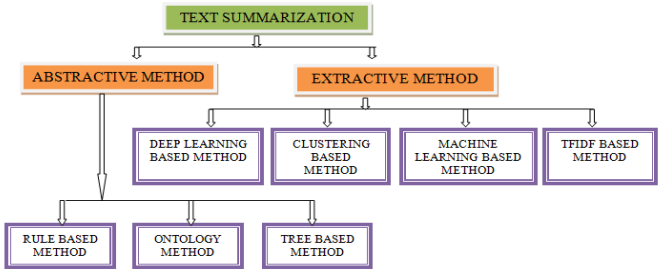


Fig 1.1 Summarization Techniques

1.1 EXTRACTIVE TEXT SUMMARIZATION

The goal of an extractive summarizer is to select the most important sentences from a document while reducing redundancy in the outline. It is made by verbatim reusing portions of input text.

a. Term Frequency-Inverse Document Frequency (TFIDF) approaches:

In TFIDF (Term Frequency-Inverse Document Frequency) approach, The bag-of-words model is created at the sentence level, using classic term-frequency and sentence frequency techniques, where sentence frequency refers to the number of sentences in the document that contain that term, and question words are terms that appear frequently in the document. These words produce generic summaries since they express the document's theme. For sentences, term frequency is usually zero or one.

b. Clustering based approach:

Poor treatment term frequency and inverse document frequency (TF-IDF) of various words make up documents. The average range of existences of similar kinds of documents over the cluster is referred to as term frequency in this context. As input, the summarizer accepts clustered documents. The subject is represented in each cluster by terms with high ranking term frequency, inverse document frequency (TF-IDF) scores. The resemblance of the sentences to the cluster's theme influences sentence selection.

c. Machine Learning Approach:

The summarizing algorithms are displayed as a classification problem in a group of documents and their extractive summaries: phrases are classed as outline sentences and non-summary sentences supported the options that they preserve. There are numerous machine learning approaches that can be utilised for document summarization, and the categorization likelihood is learned statistically from the given information using Bayes' rule.

3. LITERATURE REVIEW

The extractive algorithm [1] for summarizing English texts in English classes is based on semantic association principles. Semantic association rule vectors are used to summarize documents. The semantic relevance analysis and feature extraction of keywords in English abstracts are accomplished in this study by mining relative features among English text phrases and sentences.

The author of paper [2] first extracted many candidate summaries by providing several techniques for improving the summaries upper-bound quality. Then, using bilingual features, suggested a new ensemble ranking method for rating candidate summaries. A benchmark dataset was used to undertake extensive tests. The system is created for multiple language document summarizations in paper [2] however the quality of the summarizing is not up to standard; according to

paper [2] the accuracy for summarization is 60% with a sophisticated execution framework.

The first paper to introduce the concept of fairness in text summarizing algorithms is Fairness of Extractive Text Summarization. The author demonstrates that when summarizing datasets with a sensitive attribute, the fairness of the summary must be verified. The issue of fairness becomes even more important with the introduction of neural network-based summarization techniques (which use super-wised learning). The authors anticipate that this research will lead to intriguing research questions, such as designing algorithms to assure some degree of impartiality in the summaries [3].

The proposed technique [4] is based on WordNet, which are theoretically domain agnostic and the author has used Wikipedia for some of the words that do not appear in WordNet. The author attempted to employ more cohesion clues for summary than prior lexical chain based summarization systems. The evaluated results were comparable to those of previous summarizing algorithms and yielded satisfactory outcomes.

The author of this study [5] introduced the PASCAL algorithm, which is a fresh optimization of the well-known Apriori algorithm. PASCAL is one of the most efficient algorithms for mining common patterns, according to a new method called pattern counting inference, which the three algorithms Apriori, Close, and Max-Miner demonstrate.

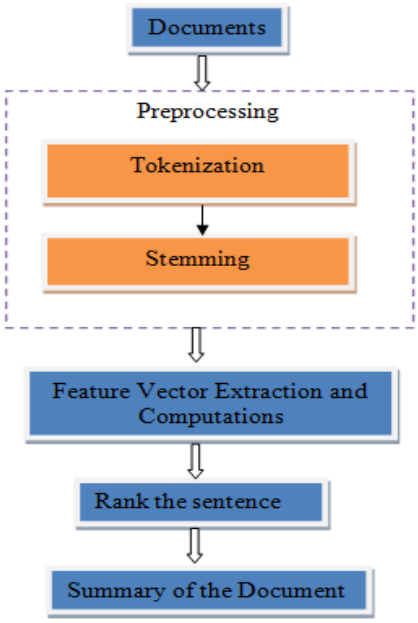
The Semantic Link Network method reinforcement ranking can be applied to any structural text and provides various summarization services such as automatically generating the Mind Map of a scientific paper, slides for a given paper, and an extended abstract for a long scientific paper or book to give readers a quick impression of the core content [6].

Text mining pattern finding technique is used to address low-frequency and misunderstanding difficulties. To refine the identified patterns in text documents, the proposed technique [7] employs two processes: pattern deploying and pattern evolving. According to the experimental results, the suggested model outperforms not only existing pure data mining-based approaches and the concept-based model, but also term-based state-of-the-art models such as BM25 and SVM-based models.

4. METHODOLOGY

The internet now has a huge number of electronic collections, many of which carry high content. However, the Internet frequently gives more information than is required. Use needs to choose the optimum data collection for a specific information demand in the shortest time possible. One of the applications of information retrieval is script summarization, which is the way of compressing the input text into a shorter version while maintaining its information content and overall meaning. The internet now has a huge number of electronic collections, many of which contain high-quality content. However, the Internet frequently gives more information than is required. In the shortest amount of time, the user wishes to select the best collection of facts for a certain information need.

Proposed system is shown in the following fig 1.2.



PageRank algorithm aids in the ranking of online pages by search engines such as Google. We looked at the PageRank algorithm to see how it may be utilised to rank text rather than web pages. This can be accomplished by shifting perspectives and employing the PageRank style matrix as a similarity score, rather than links between pages.

A. *Algorithm For Proposed System*

The code behind the extraction summarization technique is described in the following way:

Step 1: Data parsing:

Data parsing is making a single block of text out of all the text in the source document.

Step 2: Preprocessing:

Preprocessing is the process of initial data processing in order to prepare it for processing and production or additional analysis. In this step stemming is performed, system defines sentences by checking for punctuation signs like the period (.), the question mark (?), and the exclamation mark (!) and removing all the symbols available in the document. Then the text document is separated into sentences.

Step 3: Feature Vector Extraction

This step prepares vector representations of individual sentences available in the document. It is now necessary to comprehend vector representations.

Word embedding is a sort of word representation in which words with related meanings are mathematically described.

In reality, this is a broad category of algorithms that encode words as real-valued vectors in a predetermined space. Each word is represented by a multi-dimensional real-valued vector.

Step 4: Sentence Matrix

Now the system has a vector representation for given words, we can extend the approach to represent full sentences as vectors. To do so, system can get the vector representations of the terms that make up words in a sentence, and then take the mean of those vectors to get the phrase's consolidated vector.

Step 5: Sentence Matrix
At this stage, each individual sentence has a vector representation. The cosine similarity approach can now be used to assess similarities between phrases. The cosine similarity of the phrases can then be used to fill an empty matrix.
Step 6: Optimal Feature Vector Set Generation
Now that we have a matrix containing the cosine similarity between the sentences, system can fill it in. This matrix can be converted into a graph, with nodes representing sentences and edges representing similarity between sentences. To get at the sentence rating, system applies the helpful PageRank algorithm on this graph.
Step 7: Summary and Ranking of Sentence
All of the sentences in the input document are rated in the order of significance. To produce a summary of the document, top N sentences from the list are extracted.

5. RESULT

Proposed TFIDF based Text Summarization is demonstrated on the contents of web page with the below link 'https://www.ibm.com/cloud/learn/load-balancing'.
The summarized information is as below.

Summery: Load balancing handles these concurrent sessions to avoid any performance and availability issues. Network load balancing also provides network redundancy and failover. Network load balancers use the TCP/IP protocol to distribute traffic across wide area network (WAN) links. By routing user requests evenly across a group of servers, load balancers minimize the likelihood of downtime.This approach speeds up the load balancing process but it makes no accommodation for servers with different levels of capacity. Without load balancing, applications, websites, databases, and online services would likely fail when demand gets too high.
https://indianexpress.com/article/cities/mumbai/only-54-per-cent-students-in-bmc-run-schools-attending-offline-classes-data-7775645/
Summery : As many as 916 teachers have completed only the first dose yet. In the case of students' vaccination, BMC has made good progress. Out of 37,759 students who are eligible for Covid vaccine, parental consent was acquired for 27,907 students. These students, with the consent of their parents, were taken to nearby vaccination centres by the respective schools. Out of a total of 17,616 teachers, including school principals, 16,553 have been vaccinated with two doses whereas 2,852 have completed booster doses too. More so, with online mode of learning still being made available as per the directives, many are preferring it.

6. CONCLUSION

Text summarization is developing as sub – branch of Natural Language Processing an interesting research topic that helps generate compressive, meaningful, abstract of information available on net. Concise information helps to search more and efficiently and effectively. Despite the fact that text summarizing research began many years ago, there is still a long way to go and many more questions to be answered.
Text summarization has its importance in both commercial as well as research community. This paper proposes an extractive method based text summarization approach

REFERENCES

[1] Lili Wan "Extractive Algorithm of English Text Summarization for English Teaching" IEEE 2018.
[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] AnuragShandilya, KripabandhuGhosh, SaptarshiGhosh "Fairness of Extractive Text Summarization", ACM 2018.
[4] Mohsen Pourvali and Mohammad SanieeAbadeh, "Automated Text Summarization Base on Lexicales Chain and graph Using of WordNet and Wikipedia Knowledge Base" , IJCSI, Vol. 9, Issue 1,2012.
[5] Yang Gao,YueXu, Yuefengli, "Pattern-based Topics for Document Modeling in Information Filtering" in IEEE Transaction on Knoweledge and Data Engineering, vol.27,No.6,June 2015.
[6] Xiaoping SunandHaiZhuge, "Summarization of Scientific Paper through Reinforcement Ranking on Semantic Link Network" ,IEEE ,2018
[7] 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.