An Integrated Approach for Compendium Generator using Customized Algorithms

M. Suman, Tharun Maddu, M. Mohan
Department of Electronics and Computer Engineering, K.L. University, India

ABSTRACT

Text Summarization is a process that is to give the shorter version of a text document. For many research scholars who want to do their research on a specific domain has to search a lot of documents on that topic related to a specific domain. It is also difficult to go through the lot of the research papers present in that particular domain which takes a lot of time at this moment of time there are lots of chances in missing some key words present in those research papers. So that Summarizer is used to give the summary of a paper. The aim of our project is to reduce the body of the text and maintaining coherence and avoiding redundancy. Winnowing is an algorithm that gives the coherence between the multiple papers when multiple papers are given as the input. Redundancy that is the repeated words or sentences can be avoided using the MMR algorithm.

Copyright © 2015 Institute of Advanced Engineering and Science. All rights reserved.

Corresponding Author:
M. Suman,
Department of Electronics and Computer Engineering,
K.L. University,
Vaddeswaram, Guntur.
Email: suman.maloji@kluniversity.in

1. INTRODUCTION

The rapid growth of the Internet yielded a massive increase of the amount of information available, especially regarding text documents (e.g. news articles, electronic books, scientific papers, blogs, etc.). Due to the huge volume of information in the Internet, it has become unfeasible to efficiently sieve useful information from the huge mass of documents. Thus, it is necessary to use automatic methods to understand, index, classify and present all information in a clear and concise way, allowing users to save time and resources. The need for a tool that takes a text and shortens it into a brief and succinct summary has never been greater than now. With the huge amount of information on the internet and the necessity to get the essential of this information in a short time, the need for summarizers becomes everyday pressing, especially, for people with special needs like blind or elderly people. For those people it is vital to go directly to the essential information rather than having to read through many passages. One solution is use text summarization techniques. Text summarization (TS) is the process of automatically creating a compressed version of one or more documents. It attempts to get the meaning of documents. Essentially, TS techniques are classified as Extractive and Abstractive. Extractive summaries produce a set of the most significant sentences from a document, exactly as they appear. Abstractive summaries attempt to improve the coherence among sentences by eliminating redundancies and clarifying the contest of sentences. It may even produce new sentences to the summary. Currently, the extractive summaries are commonly used because they are easier to create. Extractive methods are usually performed in three steps.

1. Create an intermediate representation of the original text,
2. Sentence scoring,
3. Select high scores sentences to the summary.
The first step creates a representation of the document. Usually, it divides the text into paragraphs, sentences, and tokens. Sometimes some preprocessing, such as stop word removal is also performed. The second step tries to determine which sentences are important to the document or to which extent it combines information about different topics, by sentence scoring. The score should be a measure of how significant a sentence is to the understanding of the text as a whole. The last step combines the score provided by the previous steps and generates a summary.

In order to be able to make going through IEEE papers a lot easier and a lot more effective, the compendium generator analyses the paper and shows the user details for him/her and comprehend what the paper is about. It allows the user to save this short summary in case multiple papers are being referred to. This makes it simple to keep a track of all references. Using an algorithm that combines TF/IDF, Cue-Phrases, and Resemblance to title, results are proven to be most effective. The order of the sentences are kept intact. The tool also allows the user to compare two or more papers giving an output of a joint non redundant summary, which can form the basis for a new paper. It helps us to determine coherence or how strongly the papers pertaining to the same domain are linked.

Fingerprints are generated to check how strong the relevance between two documents is. Winnowing algorithm is used to determine this. These are methods used to determine plagiarism, with a degree of modification it has been used to determine degree of relevance.

2. LITERATURE SURVEY

There are plenty of summarizers available. The online summarizers do not prove to be very effective as only sentences with more no of words are chosen, not necessarily the sentences with keywords or important sentences that resemble the title of the document. 'A Context Based Text Summarization System', explains how combining algorithms can provide more effective results [2]. Depending on the context, however, some techniques may yield better results than some others. 'Assessing sentence scoring techniques for extractive text summarization’ proposes a new summarization system that easily combines different sentence scoring methods in order to obtain the best summaries depending on the context [4]. The fifteen sentence scoring methods most widely used and referenced in the technical literature in the last 10 years are applied to single document summarization. Both quantitative and qualitative measures are used to evaluate which combination of the sentence scoring methods yield better results for each context. Combining 3 to 5 specific sentences scoring methods in a certain context provides much better quality results.

The choice of those methods depend on context of the document. 'Get Only the Essential information: Text summarizer based on implicit data’ was used to experiment and determine the best possible combination to summarize papers [1]. Thereby creating a customized algorithm including, Cue-Phrases, Resemblance to title and TF/IDF drastically improves accuracy. This helps us to summarize a single document without missing any important sentences and the context of the paper is also preserved. Recent research in multi-document summarization has focused on removing redundancy and statistic approaches in machine learning and language modeling to find important sentences and words in multiple documents. 'A Contextual Query Expansion Based Multi-document Summarizer for Smart Learning’, provides insight on how redundancy can be removed using a technique called Maximum Marginal Relevance (MMR) [6]. This technique is proposed as a relatively better approach to tackle redundancy. [3] ‘A survey of text summarization techniques’ explains that Precision is defined as the percentage of the relevant items in the returned set and Recall is the percentage of the relevant items in the returned set compared to those in the collection. If the whole collection is retrieved, then the Recall is maximum, but Precision is low. Most search engines suffer from this problem (high Recall and low Precision).

If search engines search only a documents primary ideas, instead of every word, then Recall will likely not be decreased but Precision will likely improve. Hence, an automated facility for summarizing documents to improve productivity is desirable. A good summarization system should include only sentences that are most important to a documents theme; it must also cover all documents topics. Using a summary instead of the whole documents as a representative of what the documents are about would mean processing a fraction (20 percent or less) of the documents text, yet yield better precision and lesser processing time. In order to determine the requirements of a good summarization system, many text summarization approaches were reviewed. An in-depth review of text summarization literature was conducted and results from this study along with a description of each algorithm. Coherence 'Winnowing: Local Algorithms for Document Fingerprinting’ provides insight on plagiarism detection techniques. A technique to generate unique values for chunks of text [5].

An Integrated Approach for Compendium Generator using Customized Algorithms (M. Suman)
3. PROPOSED SYSTEM

To design a compendium generator there are some specifications such as functional specifications and program specifications.

3.1. Functional Specifications

1. The compendium generator mainly aims to generate important sentences after passing through the document. Also when two or more academic papers are given as input then a combined non redundant summary is generated.
2. By creating a customized algorithm that drastically improves accuracy of the summary. This helps us summarize a single document without missing any important sentences and preserving the context of the paper.
3. Maintaining correlation with the main idea, is key to providing the ideal summary. Thus multiple documents belonging to the same domain can be summarized.

3.2. Program Specifications

3.2.1. Tokenizer

1. Every word needs to be split into individual tokens, every word becomes a token.
2. PUNKT module in NLTK is used for this.

3.2.2. Stop Removal

1. NLTK stopwords package is used to remove stop words.
2. This helps improve calculation of word frequency.

3.2.3. Stemmer and Lemmatizer

1. An inbuilt lemmatizer called Wordnet is used.
2. The Stemmer used is Snowball stemmer.

3.2.4. Cue-Phrase

1. A corpus of cue phrases that are most commonly used in research papers is created.
2. In summary, in conclusion, our investigation, the paper describes, etc. are a few examples.

3.2.5. Resemblance to Title

1. A list that stores the title is created and sentences that have resemblance to these words are ranked higher.
2. This helps maintain the core essence of the paper.

3.2.6. TF-IDF

1. A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus
2. It uses the most no of occurrences as an upper end value. The other frequencies are compared to this value.
3. A custom combination of these three algorithms ranks sentences aptly for academic research papers.

3.2.7. Sentence Selection

The sentences which have a rank above the threshold rank are selected.

3.2.8. Redundancy Removal

1. Maximum Marginal Relevance algorithm is used to remove redundancy.
2. A combined non redundant summary is generated for multiple documents.

3.2.9. Fingerprinting

1. Created a hash value function using length of fingerprint as 20. This is an ideal number as it is low enough to provide accurate results. It is large enough to be computable.
2. A formula from the paper is used to generate unique fingerprints.

3.2.10. Winnowing

An algorithm primarily used to detect plagiarism modified to determine relevance between documents. Used to identify level of coherence between documents based on the fingerprints matched.
4. IMPLEMENTATION
4.1. Text Segmentation
Three main processes take place in this module.

4.1.1. Tokenization
Splitting a sentence into individual words. NLTK PUNKT is used.

4.1.2. Lemmatization
Converting a word to its root form. E.g. says, said, saying will all map to root form – say.

4.1.3. Stemmer
It is similar to a lemmatize, but it stems a word rather than get to the root form. eg. Laughed, laughing will stem to laugh. However, said, saying will map to sa - which is not particularly enlightening in terms of what, “sa” means. Stop word removal also takes place where constantly repeated words are removed.

4.2. Sentence Ranking
Since the words are tokenized, they are now ranked according to Cue Phrase, Sentence Position and Resemblance to title algorithms.

4.2.1. Cue Phrase
Cue-Phrases: In general, the sentences started by in summary, in conclusion, our investigation, the paper describes and emphasizes such as the best, the most important, according to the study, significantly, important, in particular, hardly, impossible as well as domain-specific bonus phrases terms can be good indicators of significant content of a text document.

4.2.2. TF-IDF
TFIDF, short for term frequency inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It uses the most no of occurrences as an upper end value. The other frequencies are compared to this value.

4.3. Sentence Selection
Sentences with rank above threshold frequency are selected.

4.4. Redundancy Removal
As multiple documents are being summarized, some documents may have points that are repeated. When a combined summary of all the documents is being displayed this redundancy continues. MMR algorithm is used to get rid of this redundancy.

4.5. Fingerprinting
Fingerprinting is a technique used to detect Plagiarism in academic documents. This method forms representative digests of documents by selecting a set of multiple substrings (n-grams) from them. So the first step is to do a text segmentation as matches should be unaffected by extra space, capitals and punctuation, etc. Then k-grams are formed where k is 20. It is found to be the ideal value.

4.6. Winnowing
This helps understand how strongly various papers pertaining to a single domain are linked. It gives us a good perspective of how the data can be organized and used. Level of similarity that needs to be matched is given a value. A lower threshold would be a noise threshold that determines if there’s some amount of similarity between the documents being compared. From there on thresholds are set at custom points that determine similarity.

5. RESULTS
5.1. Module 1
Summarization for the single or multiple IEEE papers. Enter the number of papers to summarize.

Inputs:
Figure 1. To enter the number of papers

Figure 2. IEEE paper 1 as Input

Figure 3. IEEE paper 2 as Input
AAS
ISSN: 2252-8814

An Integrated Approach for Compendium Generator using Customized Algorithms (M. Suman)

Output:

This method used a term-based sentence aggregation approach to investigate the effects of using dynamic unigrams and bigrams for sentence scoring. Results indicated that using both unigrams and bigrams significantly improves the accuracy of the compendium generator. The results also showed that the use of dynamic unigrams and bigrams improves the diversity of the generated sentences.

5.2. Module 2

To check the coherence for the multiple IEEE papers.

Input:

Paper 1

With the explosion of the World Wide Web and the abundance of text available on the Internet, the need to provide high-quality summaries in order to allow the user to quickly locate the documents of interest is a major challenge. Two generalizable techniques that were developed to address this challenge are summarization and text classification.

Paper 2

This paper presents an integrated approach for document summarization. A major challenge related to document summarization is the selection of features that best capture the important aspects of a document. The focus of this research is on developing an automatic text summarization system that can effectively capture the important aspects of a document. The system is based on a combination of several techniques, including feature selection, document classification, and text ranking.

Figure 5. IEEE paper 2 as input

Figure 6. IEEE paper 2 as input
Output:

Figure 7. Output for Coherence

6. EVALUATION

Rogue method will be used to evaluate the summarizer. The official evaluation toolkit for text summarization in DUC, to evaluate the performance of our summarization system. It involves manually summarizing a document and then compare it with the automated summary. Also involves manually determining coherence between documents, and comparing it with the documents.

REFERENCES


BIOGRAPHIES OF AUTHORS

Dr. M. Suman professor (Signals and Systems) in department of Electronics and Computer Engineering (ECM) has extended his services as HOD in ECM department, K L University. He was awarded with Ph.D. from JNTUH, Hyderabad for the thesis entitled “ENHANCEMENT OF COMPRESSED NOISY SPEECH SIGNAL”. He is also the life member of Computer Society of India (CSI).

Tharun Maddu student of Electronics and Computer Engineering (ECM) pursuing 4th year of B.TECH in K L University. My previous research works are based on data mining. The present work is related to NLTK on which the present paper research is done.