Relevant Content Extraction and Text Summarization

[1]Yashashvi Sharma, [2]Ashutosh Dixit
[1,2]Dept. of Computer Engineering, Y.M.C.A. University of Science & Technology, Faridabad, India

Abstract
Text Summarization is a process of extracting or collecting important information from original text and providing that information in the form of summary. It is an important research area in today's era of the fast growing information age. As information is growing day by day on the internet, it is difficult for users to identify the relevant information. Users have to read the whole document to determine whether the given document is relevant or not. With the help of text summarization a shorter version of large text documents by keeping relevant information from the original text document can be generated. In this work, the focus is on the comparison of clustering technique and novelty detection technique used in generating summary of the documents.

Keywords
Text Summarization, Clustering, Extractive Summarization, Abstractive Summarization, Vector Space Model

I. Introduction

Text Summarization is a technique used to produce a concise summary of one or more texts. As information is growing day by day on the internet, it is difficult for users to identify the relevant information. Users have to read the whole document to determine that whether the given document is relevant or not. Text summarization generates a shorter version of large text documents which is known as the summary, by keeping relevant information from the original text document. By reading the summary of the document, the user can easily decide that the given document is relevant or not. The user doesn't have to read the complete document thus saving the time.

Based on the method of summary generation, text summarization systems are of two kinds: extractive and abstractive.

In extractive text summarization, firstly sentences are scored based on certain criteria and then sentences with the higher score are considered as important to include in the summary. The number of sentences included in the summary depends on the length of the summary to be generated.

While in abstractive text summarization, instead of selecting sentences as it is from the original text document, first the original text is interpreted using some linguistic methods and then the summary is generated using natural language generation methods. Summary generation using abstractive methods is complex than the extractive methods.

II. Related Work

Past literature that use the various summarization techniques are cited in this section. Most of the researchers concentrate on sentence extraction rather than generation for text summarization. The most widely used method for summarization is based on statistical features of the sentence which produce extractive summaries.

Luhn [4] proposed that the most frequent words represent the most important concept of the text. His idea was to give the score to each sentence based on number of occurrences of the words and then choose the sentence which is having the highest score. Edmunson proposed methods based on location, title and cue words. He stated that initial few sentences of a document or first paragraph contains the topic information and that should be included in summary. One of the limitation of statistical approach is they do not consider semantic relationship among sentences. Goldstein [2] proposed a query-based summarization to generate a summary by extracting relevant sentences from a document based on the query fired. The criterion for extraction is given as a query. The probability of being included in a summary increases according to the number of words co occurred in the query and a sentence. Goldstein [1] [2] also studied news article summarization and used statistical and linguistic features to rank sentences in the document.

One of the approach for summarization can be done by sentence extraction and clustering. ZHANG Pei-ying & LI Cun [5] suggested that sentences are clustered based on the semantic distance among sentences and then calculates the accumulative sentence similarity between the clusters and finally chooses the sentences based on extraction rules. The method used to cluster the sentences is k-means algorithm[5].

The concept of lexical chain was first introduced by Morris and Hirst [7][9]. Lexical chains [7] exploit the cohesion among an arbitrary number of related words. Lexical chains are created by grouping set of words that are semantically related. Barzilay and Elhadad [6][8] constructed lexical chain by calculating semantic distance between words using WordNet. Strong lexical chains are selected and the sentences related to these strong chains are chosen as a summary.

A. Similarity Measures

A similarity measure gives us the degree of similarity between two objects. Summarization techniques often use similarity measures to find the similarity between the sentences in the text. The two methods that are to be implemented use similarity measures to identify the more informative parts of the document from the less informative parts. We used two similarity ranking algorithms in this project.

1. Cosine Similarity Measure

Cosine similarity measure is based on Bhattacharya's distance, which is an inner product of the two vectors divided by the product of their length. Given two vectors, we calculate the similarity between these two vectors by comparing the angle between them. The smaller the angle, the more similar the vectors.

Given two |V|-dimensional vectors $\underset{x}{\rightarrow} = \langle x_1 \rangle\ x_2, \ldots, x_{|V|}\rangle$ and $\underset{y}{\rightarrow} = \langle y_1, y_2, \ldots, y_{|V|}\rangle$

We have,

$$\underset{x}{\rightarrow} . \underset{y}{\rightarrow} = \underset{x}{\rightarrow} . \underset{|y|}{\rightarrow} . \cos\theta$$

Where $\vec{x} . \vec{y}$ represent the inner product between the vectors. This dot product is defined as

$$\vec{x} . \vec{y} = \sum_{i=1}^{|V|} x_i . y_i$$

And the length of a vector can be computed by the Euclidean distance formula

$$|\vec{x}| = \sqrt{\sum_{i=1}^{|V|} x_i^2}$$

Given the two vectors v_1 and v_2, the cosine similarity sin $(\overrightarrow{v_1}, \overrightarrow{v_2})$ is calculated as,

$$\text{Sin}\ (\overrightarrow{v_1}, \overrightarrow{v_2}) = \frac{\overrightarrow{v_1}}{|\overrightarrow{v_1}|} \cdot \frac{\overrightarrow{v_2}}{|\overrightarrow{v_2}|}$$

Cosine similarity measure value lies between 0 and 1. The higher the value, the more similar are the two vectors.

In many search engines, cosine similarity measure is used for comparing the query and documents to retrieve the documents which are similar to the query. Another use of cosine similarity measure is to get the similar pages for a particular page in the search results. In this case, we replace the query vector by document vector.

2. Tf-Idf

The vectors we use to calculate the cosine similarity contains the TF-IDF weights. Here TF is the Term Frequency. This function measures how common the term is in the document and IDF is inverse document frequency which relates the document frequency to the total number of documents in the corpus.
Formulas for calculating the TF and IDF is as follows:
TF = $\log(f_{t,d}) + 1$ if $f_{t,d} > 0$ and 0 otherwise,

$$\text{IDF} = \log(\frac{N}{N_t})$$

Where, $f_{t,d}$ is the frequency of the term t in document d and N_t represents the number of document containing the term t.
After calculating the TF and IDF, we save the TF-IDF weight score into the vector of the given document.

III. Survey on Extractive Summarization Techniques

Extractive summarizers aim at picking out the most relevant sentences in the document while also maintaining a low redundancy in the summary.

A. Cluster Based Method

Clustering is the process of discovering natural groupings or clusters and identifying interesting distributions and patterns within multidimensional data based on some similarity measure. The topic of clustering has been extensively studied in many scientific disciplines such as text mining, pattern recognition, IR etc. Document clustering is a central problem in text mining which can be defined as grouping documents into clusters according to their topics or main contents. Document clustering has many purposes including expanding a search space, generating a summary, automatic topic extraction, browsing document collections, organizing information in digital libraries and detecting topics. The surveys on the topics offer a comprehensive summary of the different applications and algorithms.
In the proposed system first the query is processed and the summarizer takes document & finally produces summary. After Pre-processing, producing the summary involves the following steps:

1. Calculating similarity of sentences present in document with user query.
2. After calculating similarity, group sentences based on their similarity values.
3. Calculating sentence score using word frequency and sentence location feature.
4. Picking the best scored sentences from each group and putting it in summary.

Implementation Steps

1. The user selected the document & query is the input to the summarizer.
2. The documents are clustered by using, cosine similarity as a similarity measure to generate the appropriate document clusters.
3. Then from the document, sentences are clustered based on their similarity values.
4. Calculate the score of each group (sentence cluster).
5. Sort sentence clusters, in reverse order of group score.
6. Pick the best scored sentences from sentence cluster and add it to the summary.
7. We have decided the number of sentences to be selected depending on sentence clusters size above the threshold value.

B. Novelty Detection Technique

To begin with, sentences are extracted from the given document. A similarity metric, such as cosine similarity, is utilized to measure the similarity between sentences. Then, relevant sentences are to be selected based on relevant threshold. Finally, novel sentences get retrieved from the relevant sentences. The threshold technique is applied to the following operation, retrieval or filter. In the following, we will discuss this approach in detail.

1. Relevant Sentence Retrieval

This problem aims to find sentences which are relevant to the query. Sentence retrieval is considered as different from document retrieval because sentences contain less text than documents [38]. Since they contain less text, it may be expected that the systems that work on sentences are not reliable. Despite this possible problem, taking sentences as the unit of retrieval enables adjusting sentence-level decisions to different levels of texts such as the aim of these workshops which is a system that helps information retrieval system users to skim through result set of a query by only seeing relevant and novel sentences.

2. Novel Sentence Retrieval

This problem aims to identify relevant sentences which contain new information with respect to the previous sentences both in the same document and the ones in the previous documents. This definition constrains novel sentence detection algorithms to run in an incremental way in which every sentence adds some knowledge which should be examined while giving decision for the next sentence. Another important point of novel sentence detection is that, it should be done over relevant sentences. Because new information contained by irrelevant sentences should not be provided to the users.

IV. Problem Definition

The volume of electronic information available on Internet is increasing day by day. As a result,dealing with such huge volume of data is creating a big problem in different real life data handling applications. Most of the research works base on finding the extracts from a given text depending on few hand tagged rules, as the position of a sentence in a text, format of words (bold,

italic etc.) in a sentence, frequency of a word in a text etc. But the drawback of this approach is, it greatly depends on the format of the text. As a result, importance of a sentence bases on its format and position in the text rather than its semantic information.
Also, extracted sentences usually tend to be longer than average. Due to this, parts of the segments that are not essential for summary also get included, consuming space.
Important or relevant information is usually spread across sentences and extractive summaries cannot capture this and produces redundancy

V. Proposed System Methodology

In this work, the focus is on the comparison of clustering technique and novelty detection technique used in generating summary of the documents.
Different parameters have been used to present the difference between them. First is to evaluate the performances of both the techniques by implementing them.
Then present their differences to know which technique is better to generate the summary than the other.
Parameters used can be:

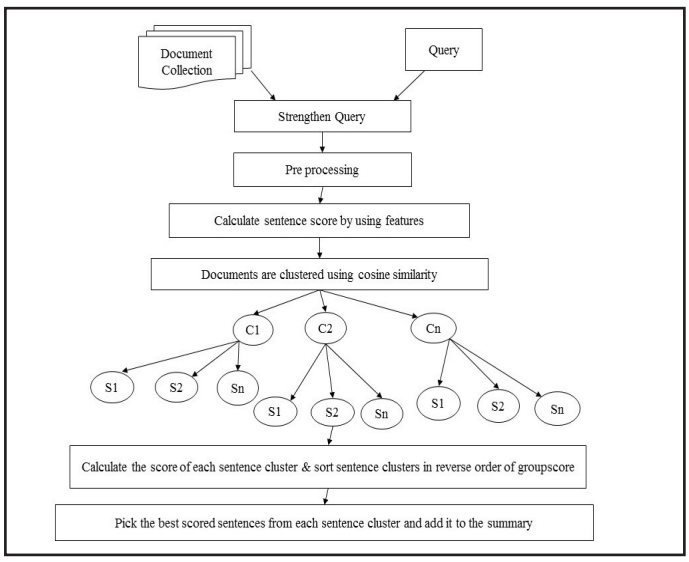
- Number of sentences included in summary generated by both the methods.
- Time required to implement them.

A. Cluster Based Technique

Algorithm:

1. The user selected collection of documents & query is the input to the summarizer.
2. The documents are clustered by using, cosine similarity as a similarity measure to generate the appropriate document clusters.
3. Then from the document, sentences are clustered based on their similarity values.
4. Calculate the score of each group (sentence cluster).
5. Sort sentence clusters, in reverse order of group score.
6. Pick the best scored sentences from sentence cluster and add it to the summary.
7. We have decided the number of sentences to be selected depending on sentence clusters size above the threshold value.

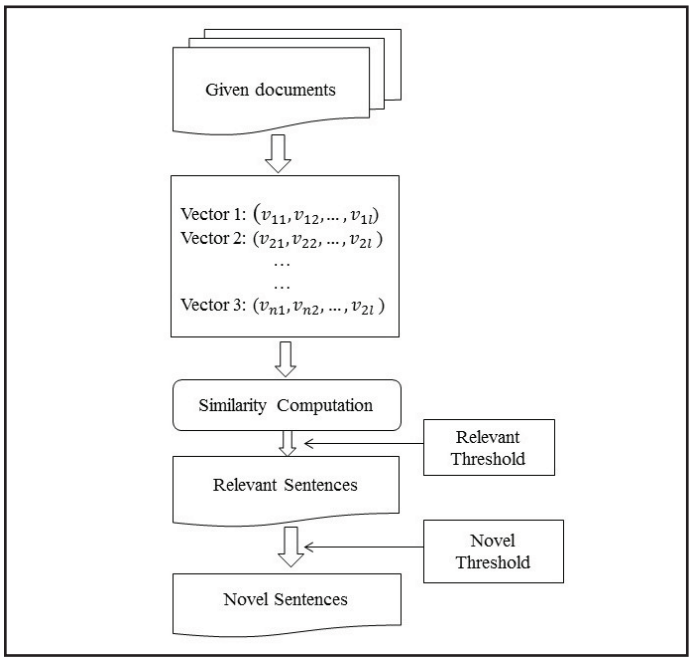
Architecture:



B. Novelty Detection Technique:

1. To begin with, sentences are extracted from the given document.
2. A similarity metric, such as cosine similarity, is utilized to measure the similarity between sentences.
3. Then, relevant sentences are to be selected based on relevant threshold.
4. Finally, novel sentences get retrieved from the relevant sentences. The threshold technique is applied to the following operation, retrieval or filter.

Architecture:



VI. Conclusion

This work is focusing on extractive summarization method's comparison. An extractive summary is the selection of important sentences from the original text.
It has been seen that without the use of NLP, the generated summary may suffer from lack of cohesion and semantics. If texts containing multiple topics, the generated summary might not be balanced.
The biggest challenge for text summarization software is to produce effective summary in less time and with least redundancy.

References

[1] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh,"A Comprehensive Survey on Text Summarization Systems", 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference.
[2] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., "Summarizing text documents: Sentence selection and evaluation metrics", In: Proc. ACM-SIGIR'99, pp. 121–128, 1999.
[3] Terrence A. Brooks,"Web Search: How the Web has changed information retrieval", Information Research, April 2003.
[4] Luhn, H.P.,"The automatic creation of literature abstracts", IBM J.Res. Develop., pp. 159–165, 1959.
[5] ZHANG Pei-ying, LI Cun-he,"Automatic text summarization based on sentences clustering and extraction".
[6] Barzilay, R., Elhadad, M.,"Using Lexical Chains for Text

Summarization", In Proc. ACL/EACL'97 Workshop on Intelligent Scalable Text summarization, Madrid, Spain,1997, pp. 10–17

[7] Youngjoong Koa, Jungyun Seo,"An effective sentence-extraction technique using contextual information and statistical approaches for text summarization", 2008.

[8] Eduard Hovy, Chin Yew Lin,"Automated text summarization in SUMMARIST", MIT Press, pp. 81–94, 1999.

[9] Morris, J., Hirst, G.,"Lexical cohesion computed by thesaural relations as an indicator of the structure of text", Computational Linguistics 17(1), pp. 21–43, 1991.

[10] Vishal Gupta, Gurpreet Singh Lehal,"A Survey of Text Summarization Extractive Techniques", Journal ofEmerging Technologies in Web Intelligence, Vol. 2, No. 3, 2010.

[11] Madaan, Rosy, A. K. Sharma, Ashutosh Dixit,"Presence Factor-Oriented Blog Summarization", Journal = {CoRR}, Vol. = {abs/1302.7131}, 2013.

[12] Rosy Madaan, A.K. Sharma, Ashutosh Dixit,"A novel architecture for relevant content extraction from Blog pages", International Journal of Scientific and Engineering research (IJSER), Vol. 4, Issue 5, May 2013.

[13] Saranyamol C S, Sindhu L,"A Survey on Automatic Text Summarization", Saranyamol C S et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), pp. 7889-7893, 2014.

[14] Nikita Munot, Sharvari S. Govilkar,"Comparative Study of Text Summarization Methods", International Journal of Computer Applications, Vol. 102, No. 12, September 2014.

[15] Rasim ALGULIEV, Ramiz ALIGULIYEV,"Evolutionary algorithm for extractive text summarization", Intelligent Information Management, 1, pp. 128-138, Indian Institute of Technology, 2009.

[16] Edmundson, H.P.,"New methods in automatic extraction", J. ACM16 (2), pp. 264–285, 1968.

[17] Ming-Feng Tsai, Ming-Hung Hsu, Hsin-Hsi Chen,"Similarity Computation in Novelty Detection", National Taiwan University, 2004.

[18] Anjali R. Deshpande, Lobo L. M. R. J,"Text Summarization Using Clustering Technique", International Journal of Engineering Trends and Technology (IJETT) - Vol. 4, Issue 8, 2013.

Ms. Yashashvi Sharma received B.Tech in computer engineering in 2014 from Manav Rachna International University, Haryana and pursuing M.Tech degree in computer engineering in YMCA University of Science and Technology, Haryana, INDIA.