Modified Text Summarization Based On Information Retrieval

Miss Anjali R. Deshpande
Student M. E. (C. S. E.) Part –II
Walchand Institute of Technology, Solapur

Prof. Lobo L. M. R. J.
Professor & Head
Department of Information Technology,
Walchand Institute of Technology, Solapur

Abstract

Summarization system consists of reducing a text document into a short set of words or paragraph that conveys the key meaning of the text. The observable fact of information overload means that access to coherent and correctly-developed summaries is essential. Text summarization is the most challenging task in information retrieval systems. Data reduction helps user to find required information quickly without wasting time and effort in reading the whole document.

1. Introduction

Automated information retrieval systems are used to reduce "Information Overload". Information overload refers to the difficulty a person can have understanding an issue & making decisions that can be caused by presence of too much information. Web Search Engines are the most noticeable IR applications. An information retrieval process begins when a user enters a query into the system. Now a days, it is very common that a keyword-based search on the internet returns hundreds, or even thousands of hits, by which the user is often confused. The problem is the lack of an efficient and effective method to find the required information. Therefore, there is an increasing need of new technologies that can help the user to go through large volumes of information & to quickly identify the most relevant documents as well as absorbing a large quantity of relevant information. It is very difficult for human beings to manually summarize large documents of text. Research in automatic text summarization has received considerable attention in the past few years due to the exponential growth in the quantity & complexity of information sources on the internet.

IJERT

Automatic summarization is the creation of short version of text by computer program.

The product of this procedure still contains the most important points of the original document. Most important advantage of using a summary is its reduced reading time & the link between a text element in the summary & its position in the original document can be easily established.

Summaries can be indicative or informative. Users can make use of indicative summaries before referring to the source e.g. to judge the relevance of the document. A compact & concise summary enables the user to quickly get a rough idea of the documents content & to efficiently identify the documents that are most relevant to his/her needs. On the other hand, users may use summaries in place of the source text these are informative summaries. Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. In the practice of automatic text summarization, selection-based approach has so far been the dominant strategy. In this approach, summaries are formulated by extracting key text segments i.e. sentences or paragraphs from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency and location etc. to locate the sentences to be extracted. The “most important” content is treated as the “most frequent” or the “most favorably positioned” content. Such an approach thus avoids any efforts on deep text understanding. Text summarization tends to be an important task in content extraction during web mining. Such summarizations are divided into two main categories – Extractive & Abstractive.

Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original document.[1] The state-of-the-art abstractive methods are still quite weak, so most research has focused on extractive methods, and this is what we will cover.

Extractive techniques merely copy the information deemed most important by the system to the summary e.g. key clauses, sentences or paragraphs, while abstraction involves paraphrasing sections of the source document.

In case of abstraction techniques, programs that can do this are harder to develop as they require the use of natural language generation technology, which itself is a growing field.

A good summary system should extract the diverse topics of the document while keeping redundancy to a minimum. Microsoft Word's AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary.

IJERT

Summary generation by an automatic procedure has advantages as: (i) reduced reading time. (ii) the size of the summary can be controlled (iii) its content is deterministic and (iv) the link between a text element in the summary and its position in the original text can be easily established.

With large texts, text summarization software processes and summarizes the document in the time probably it would take the user to read the first paragraph. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning. Producing abstractive summary is very difficult at present. It may take some time to reach a level where machines can fully understand documents. An extractive summary, in contrast, is composed with a selection of important sentences from the original text. Extractive text summarization process can be divided into two steps: 1) Pre Processing step and 2) Processing step. Pre Processing is structured representation of the original text. In Processing step, features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary.

2. Literature Review

2.1. Query based extractive text summarization

In query based text summarization system, the sentences in a given document are scored based on the frequency counts of terms (words or phrases). The sentences containing the query phrases are given higher scores than the ones containing single query words. Then, the sentences with highest scores are incorporated into the output summary together with their structural context. [2][3]

2.2. Graph theoretic approach:

After the common pre-processing steps, namely, stop word removal and stemming, sentences in the documents are represented as nodes in an undirected graph. There is a node for every sentence. Two sentences are connected with an edge if the two sentences share some common words, or in other words, their (cosine, or such) similarity is above some threshold. This representation yields two results: The partitions contained in the graph (that is those sub-graphs that are unconnected to the other sub graphs), form distinct topics covered in the documents. This allows a choice of coverage in the summary. Query-specific summaries, sentences may be selected only from the pertinent sub graph, while for generic summaries, representative sentences may be chosen from each of the sub-graphs. The second result yielded by the graph-theoretic method is the identification of the important sentences in the document. The nodes with high cardinality (number of edges connected to that node), are the important sentences in the partition, and hence carry higher preference to be included in the summary.[4][5]

2.3 Automatic text summarization based on fuzzy logic :

This method considers each characteristic of a text such as sentence length, similarity to little, similarity to key word and etc. as the input of fuzzy system. Then, it enters all the rules needed for summarization, in the knowledge base of system. Then, a value from zero to one is obtained for each sentence in the output based on sentence characteristics and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary. The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using IF-THEN rules according to the feature criteria.

Fuzzy logic system design usually implicates selecting fuzzy rules and membership function. The selection of fuzzy rules and membership functions directly affect the performance of the fuzzy logic system. The fuzzy logic system consists of four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, crisp inputs are translated into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IFTHEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final crisp values by the defuzzifier using membership function for representing the final sentence score.[6][7]

2.4 Cluster based method:

Documents are represented using term frequency- inverse document frequency (TF-IDF) of scores of words. Term frequency used in this context is the average number of occurrences (per document) over the cluster. IDF value is computed based on the entire corpus. The summarizer takes already clustered documents as input. Each cluster is considered a theme. The theme is represented by words with top ranking term frequency, inverse document frequency (TF-IDF) scores in that cluster. Sentence selection is based on similarity of the sentences to the theme of the cluster Ci. The next factor that is considered for sentence selection is the location of the sentence in the document. In the context of newswire articles, the closer to the beginning a sentence appears, the higher its weightage for inclusion in summary. The last factor that increases the score of a sentence is its similarity to the first sentence in the document to which it belongs. Once the documents are clustered, sentence selection
from within the cluster to form its summary is local to the documents in the cluster. [8]

IJERT

2.5 Text summarization with neural networks

This method involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included. [9]

3. Comparison between technologies:

The Extractive Techniques for Text Summarization are compared as shown in the table:

Table1: Comparison between extractive techniques.

Sr. No.	Cluster Based method	Graph Based Approach	Fuzzy Logic	Genetic Algorithm	Query Based Summarizer
1.	In clustering, the sentences are grouped into clusters. In clustering based summarization, performance heavily depends on 3 important factors: Clustering sentences, cluster ordering, & selection of representative sentences from the clusters.	Each Node is a sentence. An edge exists between two nodes if their similarity is above a threshold.	It is reasoning with uncertainty. i.e. instead of 2 valued logic(true & false), there are multiple values (true, false, maybe)	This is an example of evolutionary computing methods & is optimization- type algorithm.	Generic Summaries can be constructed solely from the content in the original text (static summary). In contrast to this, Summaries should be dynamic, reflecting the user's interest.
2.	Set of sentences.	Set of sentences with their links to other sentences	The features are extracted & feature score is given as input to Fuzzifier.	Initial population is a set of initial chromosomes which are randomly generated within a generation. A chromosome is represented as a weighted combination of all features such as: W1 | - | - | - | Wn	Set of sentences.
3.	The groups are not predefined. Instead the grouping is accomplished by finding similarities between sentences according to characteristics found.	Based on nodes sub graphs will be created depending on the weights attached to neighborhood nodes.	It uses rules & membership functions to estimate a continuous function.	A fitness function is used to determine the best individuals in the population	Vector Space Model is used. Clustering can be used to group similar sentences.
4.	Can't say about redundancy.	Can't say about redundancy.	Nothing special is done to reduce redundancy.	Nothing special is done to reduce redundancy.	Redundancy can be reduced by grouping similar sentences & then best sentences are picked from each group to generate summary.

IJERT

"A good summarizer system should extract the information from document while keeping redundancy to a minimum." So we propose a Query based summarizer technique with some improvement.

4. Methodology :

After studying the methods explained in previous section a need was found that the query based summarizer technique would perform better when multiple new queries were generated by selecting new words from the corpus. This could be done in generations instead of single iteration. This discussion shows that the modified method that we propose for query based summarizer technique strongly recommends that an evolutionary computational method can be used to develop a better result.
A recommended method is suggested in the steps of the algorithm given below:

4.1. Algorithm :

1. Calculate similarity of sentences present in documents with user query.
2. Group similar sentences.
3. Calculate sentence score.
4. Compute the word weight of Wi words appearing in a document
5. Compute the Sentence score and Location score.
6. Calculate the score of each group.
7. Arrange the groups in ascending order depending on their group scores.
8. From the best groups, pick the best scored sentences and put it in summary.
   Here the summary is generated.
9. Form the new query by selecting set of words related to the original query from the corpus.
10. Use new query for summarization & repeat the above steps.

For the application of an evolutionary algorithm after step 5 the basic population can be formed & based upon the features of the words the genes & chromosomes can be formed. Considering the score function as a fitness function & using appropriate evolutionary computational operators better summarized sentences can be achieved.

5. Conclusion

We propose a modified approach for query-based summarizer. We extend the approach by using evolutionary computation to generate robust sentences. In future, we would like to improve the system by adding sentence simplification techniques to construct a summary. We can add sentence simplification feature to simplify the complex and very large sentences.

IJERT

6. References

[1] Vishal Gupta, Gurpreet Singh Lehal , "A Survey of Text Summarization Extractive Techniques", *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3,* AUGUST 2010

[2] A. P. Siva kumar, Dr. P. Premchand, Dr. A. Govardhan, "Query-Based Summarizer Based on Similarity of Sentences and Word Frequency", *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3,* May 2011

[3] E1-Haj, M.O., Hammo, B.H., "Evaluation of Query-Based Arabic Text Summarization System", *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on 19-22 Oct. 2008*

[4] Thakkar, K.S., Dharaskar, R.V.; Chandak, M.B., "Graph-Based Algorithms for Text Summarization", *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on 19-21 Nov. 2010*

[5] Rada Mihalcea, "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization"

[6] Kyoomarsi, F., "Optimizing Text Summarization Based on Fuzzy Logic", *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on 14-16 May 2008*

[7] Ladda Suanmali, Naomie Salim and Mohammed Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization" , *(IJCSIS) International Journal of Computer Science and Information Security,Vol. 2, No. 1, 2009*

[8] Kamal Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents", *TECHNIA – International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1, July 2009. (ISSN 0974-3375)*

[9] Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", *SECOND IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS, JUNE 2004*

Ms. Deshpande Anjali Ramkrishna received B. E. degree in Information Technology in 2009 from Solapur University, Solapur, India and persuing the M.E. degree in Computer Science & Engineering in Walchand Institute of Technology, Solapur, India.

Mr. Lobo L. M. R. J received the B.E degree in Computer Engineering in 1989 from Shivaji University, Kolhapur, India and the M. Tech degree in Computer and Information Technology in 1997 from IIT, Kharagpur, India .

He has registered for Ph.D in Computer Science and Engineering at SGGS, Nanded of Sant Ramanand Teerth Marathawada University, Nanded, India. Under the guidance of Dr. R. S. Bichkar. He is presently working as a Professor of Information Technology department with an undergraduate and postgraduate institute, Walchand Institute of Technology, Solapur, Maharashtra, India. His research interests include Evolutionary Computation, Genetic Algorithms and Data Mining.

IJERT