INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

REVIEW ON MULITPLE DOCUMENT SUMMERIZER

MISS. VIJAL S. RAJA[1], DR. H R. DESHMUKH[2]

1. Student ME CSE, IBSS COE, Amravati, Maharashtra, India.
2. HOD CSE, IBSS COE, Amravati Maharashtra, India.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: We are building website for Document summarizer which will take information from the user and present the information in a summarized form as well, thereby reducing user efforts. The task of Summarization is to take an information source, extract content from it, We are going to use an intelligent algorithm, the event indexing and summarization (EIS) algorithm, for automatic document summarization, which is based on taking into account a cognitive psychological model, the event-indexing model, and the roles and importance of sentences in document understanding.

Keywords: Multiple Document Summarizers, Event Indexing Summarization, Scientific, Mathematic, User Defines Approach

Corresponding Author: MISS. VIJAL S. RAJA

PAPER-QR CODE

Access Online On:

www.ijpret.com

How to Cite This Article:

Vijal S. Raja, IJPRET, 2015; Volume 3 (9): 243-248

INTRODUCTION

Over the past few years, especially with the emergence of the Internet, the exchange of information has increased immensely, affecting all of us. On the one hand, the scientific community makes us aware instantly of its scientific breakthroughs while on the other hand, journalists present reports from around the world in real time. The growing number of electronic articles, magazines and books that are made available every day, puts more pressure on professionals from every walk of life as they struggle with information overload. In fact, nowadays, most people have to read daily papers, magazine articles, specialized literature, e-mails and Web pages during the course of their everyday activities. The majority of information exists in the form of text. Text documents, including news articles and emails, are used to convey information, share knowledge, coordinate activities, as well as to document business conducts and processes. With the increasing availability of information and the limited time people have to sort through it all, it has become more and more difficult for them, whether they are business people, journalists, lawyers, researchers or doctors to keep abreast of developments in their respective disciplines. As the number of text documents increases, it becomes more time consuming and laborious for people to quickly locate critical information in a large text collection or glean insights from such a collection. Thus searching vast web pages, heavy newspaper articles or elongated research papers for little information is a waste of time. The phenomenon of information overload has meant that access to coherent and correctly-developed summaries is vital .Hence the interest in summarization has increased over the years. The search summarization techniques used by popular search engines like GOOGLE and YAHOO are not available to the users. Therefore we aim to create our website for summarizing the content. Thus the system we are implementing will also summarize the document on various categories. Thus the user gets information which is summarized and also relevant.

2. LITERATURE REVIEW AND RELATED WORK

Automatic Document Summarization is a highly interdisciplinary research area related with computer science as well as cognitive psychology. In our project, we introduce an intelligent algorithm, the event indexing and summarization (EIS) algorithm, for automatic document summarization, which is based on taking into account a cognitive psychological model, the event-indexing model, and the roles and importance of sentences and their syntax in document understanding. The EIS algorithm involves syntactic analysis of sentences, clustering and indexing sentences with the five indices from the event-indexing model, and extracting the most prominent content by lexical analysis at phrase and clause levels.

Zwaan, Langston and Graesser have explained and tested their Event-Indexing model, which means that readers monitor five aspects or indices of the evolving situation model when they read stories: (1) 'Protagonist(s)'; (2) 'Temporality'; (3) 'Causality'; (4) 'Spatiality'; (5) 'Intention'. To make this model applicable to computing, we redefined the concepts of 'event', 'Protagonist', 'Temporality', 'Spatiality', 'Causality' and 'Intention' as below. 'Event' is a cognitive psychological concept, and can be either a story or a sentence in microstructure. Zwaan and Radvansky treated each sentence as an event, in their paper. So we also render the equal concepts of 'event' and 'sentence'. In this context, 'Protagonist' can be considered as the subject or a noun phrase that plays the role of subject of each sentence. 'Temporality' is the temporal information contained in each sentence. 'Spatiality' is the space or location information in each sentence. 'Causality' is the causal relationship of a sentence to previous sentences or contained in one sentence. 'Intention' is the relationship between a subject's goal and sentences in the document(s).

3. PROPOSED ALGORITHM

As the number of text documents increases, it becomes more time consuming and laborious for people to quickly locate critical information in a large text collection or glean insights from such a collection. The phenomenon of information overload has meant that access to coherent and correctly developed summaries is vital. Therefore, we are aiming to develop a website for document summarization to help people cope with ever increasing amount of text document.

There are existing software's that provide us the summarization of a particular text only on the basis of word count and sentence in the summarized output are in the randomized order.

The website we are aiming to build will use EIS algorithm, which is based on taking into account a cognitive psycho-logical model, the event-indexing model, and the roles and importance of sentences in document understanding.

Thus the system we are implementing will also summarize the document on various categories. Thus the user gets information which is summarized and also relevant.

We have divided the entire document into five types of indices namely

- CAUSALITY
- INTENTION
- TEMPORALITY

- SPATIALITY
- PROTAGONIST.

While parsing we ignore commonly occurring grammatical words (are, is, in, the etc.). For "Causality", we will of 'intention' (e.g. purpose, intent, aim, goal, target, etc.). For temporality we have used words which give time information (e.g. am, pm, now, etc.). For spatiality we used words which give information about space or locality. All other words are put into protagonist table. In protagonist we also have addressed the issue of pronouns by replacing it with all the appropriate nouns of its previous sentences.

- **Clustering Sentences with Indices**

In this phase we associate a sentence priority with each sentence which we calculate by adding the weight (i.e. word count) of each word occurring in that sentence based on its occurrence in the document.

- **Cluster-Filtering**

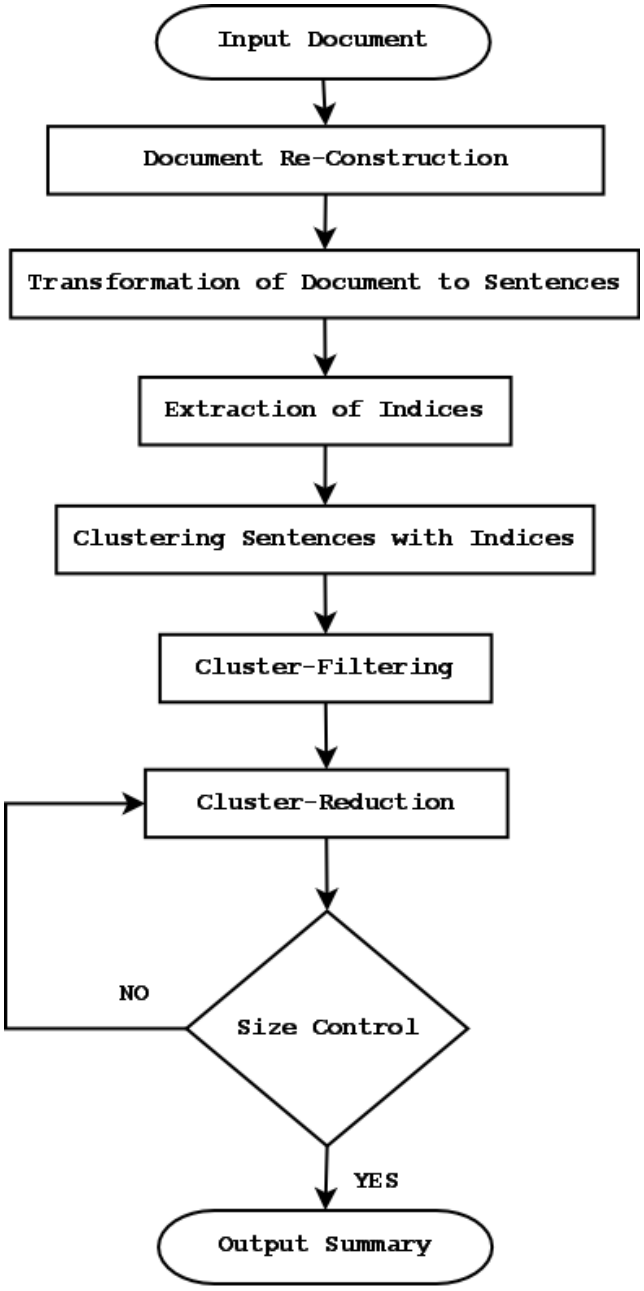
From each indices table, we first choose the largest word count and then set a threshold for each indices .

- **Indices as follows**
- Protagonist – x * largest word count from table protagonist
- Temporality – x * largest word count from table temporality
- Spatiality – x * largest word count from table spatiality
- Causality – x * largest word count from table causality
- Intention – no threshold all words are selected
- **Cluster-Reduction**

Based on the above threshold we then select the sentences with words which satisfy the threshold constraints.

- **Size Control and Output Summary**

If an extracted summary exceeds the required summary size, the summary will be returned to the module of cluster reduction for further word reduction. Once the summary is qualified at the required size, it will become an output of a formal result of summarization.



Structure of EIS Algorithm

CONCLUSION

In this paper, we proposed a new algorithm The main idea behind this project is to create a system that can take information from the user and summarize the information according to the length and criteria desired by the user. With the help of the event indexing& Summarization Algorithm. We are planning to produce a summary that gives a summary of the input text in a concise and clear manner

5. REFERENCES

1. On Finity, Countability, Cardinalities, and Cylindric Extensions of Type-2 Fuzzy Sets in Linguistic Summarization of Databases Adam Niewiadomski IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 18, NO. 3, JUNE 2010

2. YIGUO ANDGEORGE STYLIOS, "AUTOMATIC DOCUMENT SUMMARIZATION" IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 13, NO. 2, DECEMBER 2011

3. R. A. Zwaan, M. C. Langston, and A. C. Graesser, 20, "The Construction of Situation Models in Narrative Comprehension: An Event-Indexing Model" IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 10, NO. 4, DECEMBER 2012.

4. A Multicriteria Approach to Data Summarization Using Concept Ontologies Ronald R. Yager, Fellow, IEEE, and Frederick E. Petry, Fellow, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL.14, NO. 6, DECEMBER 2011

5. Using data merging techniques for generatingmulti-document summarizations Daan Van Britsom, Antoon Bronselaer, GuyDeTr´eDOI 10.1109/TFUZZ.2014.2317516,IEE Transactions on Fuzzy Systems1063-6706 (c) 2013 IEEE.