Hybridization Between Scoring Technique and Similarity Technique for Automatic Summarization by Extraction

Mohamed Amine Boudia, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria
Amine Rahmani, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria
Mohamed Elhadi Rahmani, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria
Abdelatif Djebar, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria
Hadj Ahmed Bouarara, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria
Fatima Kabli, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria
Mohamed Guandouz, Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria

ABSTRACT

To generate a summary automatically, the theory gives three approaches: by classification, by understanding or by extraction which is the most used and easy to implement. The current literature presents three basic techniques in the extraction approach: Extraction by scoring, Extraction by similarity and last but not least extraction by prototype. In previous work, the authors have always used one technique only and after that the proposed many manner to optimize the results: by the optimization algorithm or even they introduces the bio-inspired method to optimize the performance of automatic summarizers like ants or spider socials. Each technique has of weakness and strength point. In fact, the authors proposed in this work to use two technique one after the other to compensate the weakness of each technique by the strength of the second technique. In this paper, the authors will give a short state of art that will allow them later to explain the weakness and strength of each technique, after that they will explain their approach of Hybridization we will done.

KEYWORDS

Automatic Summary Extraction, Data Mining, Hybridization, Scoring, Similarity

1. INTRODUCTION AND PROBLEMATIC

Day by day, the body of electronic textual information increases. It becomes increasingly difficult to access relevant information without using specific tools to access to the content of texts by rapid and effective means. Software engineering is more developed, we have not the same application generation problem, the hardware is also very developed, in our day personal machine are powerful. For this last has become a necessary task to find the specific method to access to the content of the texts.

A summary of a text is an effective way to represent the content of the texts, and allow quick access to their content. The proposition of an automatic summarization is to produce a short text covering the essential content of the source text. “We cannot imagine our daily life without summary” says Inderjeet Mani [Mani, 2001].

DOI: 10.4018/IJOCI.2016010101

Copyright © 2016, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Headlines, the first paragraph of a newspaper article, newsletters, weather, tables of results of sports competitions and catalogues library are just the summary. Even in research, the author of the article must accompany their scientific papers with summaries (abstract) written by them.

We can use the automatic summaries to reduce the time and find the relevant documents or to reduce processing large text by identifying key information. The suggested procedure claims on the principle that high-frequency words in a document are important words” Luhn, H. P. (1958)

The current literature presents three approaches of automatic summarization:


In this paper we worked on automatic summarization by extraction, because it is a simple method to implement and gives good results; only that in previous works the summarization produced by extraction using a single technique at a time: Score, Similarity or sentence prototype.

Scoring gives generally good results, only that his weak point is its reduced ability to eliminate the phrase that is similar, in fact, if a sentence X passes the filter scoring a sentence Y which is similar to a point X will probably have a score that allows it to also pass the filter, which produces a repetitive sentence in the summary, which is logically false; secondly, the technical similarity has the strength to eliminate repetitive sentence, but its weakness is that it cannot ensure that the sentence is to keep a high weight, actually, as long as the sentence is greater the probability to have more similar phrase increases and we know that sentence is large tend to wear more information.

This work aims to use two techniques one after another, so that each one covers the technical point of weakness of the other and brings its power to the general approach, to see the impact of this proposition we experimented our approach and compared it with the result of using one technique.

2. RELATED WORK

Automatic summarization appeared earlier as a field of research in computer science from the axis of NLP (automatic Natural Language Processing), HP Luhn [Luhn 1958] proposed in 1958 a first approach to the development of automatic abstracts from extracting phrases.

In the early 1960s, HP Edmundson and other participants in the project TRW (Thompson Ramo Wooldridge Inc) [Edmundson 1963] proposed a new system of automatic summarization where it combined several criteria to assess the relevance of sentences to extract.

These works were made to identify the fundamental ideas around the automatic summarization, such as problems caused by extraction to build summaries (problems of redundancy, incompleteness, break, etc.), the theoretical inadequacy of the use of statistics, or the difficulties to understand a text (from semantic analysis) to summarize.

Other projects were based on the same technique to build the abstracts, otherwise, they were based on the extraction of sentences from the evaluation of the relevance of each of these sentences. Consideration of the evaluation of a sentence was a plus: its relationship with the rest of the text has grown steadily, and if it is extracted, checked his link with the preceding sentences for example the presence of connector’s start of sentences.

From the 1980s, theories have emerged to describe the various treatments involved in the human cognitive system in the activities of reading and text understanding, in particular the model of Kintsch and Van Dijk explained in more construction of a summary.
Population Diversity of Particle Swarm Optimizer Solving Single and Multi-Objective Problems
Shi Cheng, Yuhui Shi and Quande Qin (2012). *International Journal of Swarm Intelligence Research* (pp. 23-60).
[www.igi-global.com/article/population-diversity-particle-swarm-optimizer/75326?camid=4v1a](www.igi-global.com/article/population-diversity-particle-swarm-optimizer/75326?camid=4v1a)

Towards a Trust Management Enabled Identity Metasystem
Weiliang Zhao and Jian Yang (2012). *Intelligent and Knowledge-Based Computing for Business and Organizational Advancements* (pp. 118-134).
[www.igi-global.com/chapter/towards-trust-management-enabled-identity/65790?camid=4v1a](www.igi-global.com/chapter/towards-trust-management-enabled-identity/65790?camid=4v1a)