ABSTRACT

Document indexing is an active domain, which is interesting a lot of researchers. Generally, it is used in the information retrieval systems. Document indexing encompasses a set of approaches that can be applied to index a document using a corpus. This treatment has several advantages, like accelerating the research process, finding the pertinent contains related to a query, reducing storage space, etc. The use of the entire document in the indexing process affects several parameters, such as indexing time, research time, storage space of treatment, etc. The focus of this chapter is to improve all parameters (cited above) related to the indexing process by proposing a new indexing approach. The goal of proposed approach is to use a summarization to minimize the size of documents without affecting the meaning.
INTRODUCTION

There is an enormous amount of textual material, and it is growing every moment and time. Think of the internet comprised of web pages, news articles, status updates, blogs and so much more. The data is unstructured and the best that we can do to navigate it is to use search and skim the results.

There is a great need to reduce much of the text data to shorter and focused summaries that capture the salient details. So, we can navigate it more effectively as well as check whether the larger documents contain the information that we are looking for. We cannot possibly create summaries of all of the text manually; there is a great need for automatic methods.

They are many reasons why we need automatic text summarization tools. Here are some of them:

- Summaries reduce reading time.
- When researching documents, summaries make the selection process easier.
- Automatic summarization improves the effectiveness of indexing.
- Automatic summarization algorithms are less biased than human summarizes.
- Personalized summaries are useful in question-answering systems as they provide personalized information.
- Using automatic or semi-automatic summarization systems enable abstract commercial services to increase the number of texts, they are able to process (Torres & Juan, 2014).

The rest of the chapter is organized as follows. Section 2 gives a description of the automatic text summarization. Section 3 is dedicated to present the principal of indexing document and its steps. Section 4 proposes a new approach of indexing based on summarization to reduce the size of the document preserving the meaning. Section 5 is devoted to the experimental results obtained and criteria used in evaluation. Finally, the conclusion is given in Section 6.

AUTOMATIC TEXT SUMMARIZATION

Automatic text summarization is the process of creating a short and coherent version of a longer document. We are generally good at this type of task as it involves first understanding the meaning of the source document and then distilling the meaning and capturing salient details in the new description. As such, the goal of automatically creating summaries of text is to have the resulting summaries as good as those written by humans.

It is not enough to just generate words and phrases that capture the gist of the source document. The summary should be accurate and should read fluently as a new standalone document. The different dimensions of text summarization can be generally categorized based on its input type (single or multi document), purpose (generic, domain specific, or query-based) and output type (extractive or abstractive) (Kumar, Goh, Basiron, Choon, & Suppiah, 2016).

There are two main approaches to summarize text documents: Extractive Methods and Abstractive Methods. Extractive text summarization (Gupta & Lehal, 2010) involves the selection of phrases and sentences from the source document to make up the new summary. Techniques involve ranking the relevance of phrases in order to choose only those most relevant to the meaning of the source.