Expert Systems with Applications 40 (2013) 458–465

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

CDDS: Constraint-driven document summarization models

Rasim M. Alguliev, Ramiz M. Aliguliyev *, Nijat R. Isazade

Institute of Information Technology of Azerbaijan National Academy of Sciences, Azerbaijan

ARTICLE INFO

Keywords:
Constraint-driven summarization
Coverage-driven summarization
Diversity-driven summarization
Quadratic integer programming
Particle swarm optimization

ABSTRACT

This paper proposes a constraint-driven document summarization approach emphasizing the following two requirements: (1) diversity in summarization, which seeks to reduce redundancy among sentences in the summary and (2) sufficient coverage, which focuses on avoiding the loss of the document's main information when generating the summary. The constraint-driven document summarization models with tuning the constraint parameters can drive content coverage and diversity in a summary. The models are formulated as a quadratic integer programming (QIP) problem. To solve the QIP problem we used a discrete PSO algorithm. The models are implemented on multi-document summarization task. The comparative results showed that the proposed models outperform other methods on DUC2005 and DUC2007 datasets.

1. Introduction

Interest in text mining started with advent of on-line publishing, the increased impact of the Internet and the rapid development of electronic government (e-government). The Internet and e-government do potentially accessible huge information volumes and thereby puts new problems – effective work with such volumes. With the rapid increasing of online information and fast development of science and technology, many research efforts have been made on web mining, text mining, information extraction, knowledge discovery and information retrieval. However, the conventional information retrieval technologies are becoming more and insufficient for obtaining useful information effectively. This makes how to summarize documents with all kinds of information increasingly urgent. Therefore, in a situation of 'an information overload' especially actual there are automatic processing methods with great volumes of the information, in particular – the compressed representation methods of text documents – summaries. Automatic document summarization techniques are one way of helping people find information effectively and efficiently. The goal of automatic summarization is to take a source document, extract content from it, and present the most important content to the user in a condensed form. Thus, now more than ever, consumers need access to robust text summarization systems, which can effectively condense information found in several documents into a short, readable synopsis, or summary (Harabagiu & Lacatusu, 2010; Yang & Wang, 2008).

Text summarization is a good way to condense a large amount of information into a concise form by selecting the most important and discarding the redundant information. According to Mani and Maybury (1999), automatic text summarization takes a partially structured source text from multiple texts written about the same topic, extracts information content from it, and presents the most important content to the user in a manner sensitive to the user's needs. Nowadays, without browsing the large volume of documents, search engines such as Google, Yahoo!, AltaVista, and others provide users with the clusters of documents they are interested in and present a summary of each document briefly which facilitates the task of finding the desired documents (Boydell & Smyth, 2010; Shen, Sun, Li, Yang, & Chen, 2007; Song, Choi, Park, & Ding, 2011; Yang & Wang, 2008).

In Mani and Maybury (1999) it is underlined two basic approaches to automatic summarization: abstractive and extractive. The first approaches, which promise to produce summaries that are more like what a human might generate but are limited by the progress of natural language understanding. Abstractive approaches involve generating novel sentences from information extracted from the source(s) and assume use of more refined methods of the linguistic and semantic analysis. The more widely used extractive approaches are focused on passage extraction (usually – sentences, from here the general designation of the approach – sentence extraction) which rank sentences, extract sentences with highest scores, and then compose the summary. The result of the summarization systems based on sentence extraction is far from an ideal – the coherent summary made by the qualified expert. However, better summarization systems demand the difficult software, have lower productivity and often impose essential restrictions on style and subjects of the source text.

* Corresponding author. Address: 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan. Fax: +994 12 539 61 21.

E-mail addresses: rasim@science.az (R.M. Alguliev), a.ramiz@science.az, r.aliguliyev@gmail.com, aramiz@iit.ab.az (R.M. Aliguliyev), depart13@iit.ab.az (N.R. Isazade).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.07.049

Besides, summaries can be generic or query-focused (Ouyang, Li, Li, & Lu, 2011; Teng, Xiong, He, Yang, & Liu, 2010; Wan, 2008). A query-focused summary presents the information that is most relevant to the given queries, while a generic summary gives an overall sense of the document's content. As compared to generic summarization that must contain the core information central to the source documents, the main goal of query-focused multi-document summarization is to create from the documents a summary that can answer the need for information expressed in the topic or explain the topic.

Further document summarization is divided into single- and multi-document depending on the number of summarized documents. The multi-document summarization is to produce a single summary from a set of related documents whereas single-document summarization is intended to summarize only one document (Fattah & Ren, 2009; Zajic, Dorr, & Lin, 2008).

A good summary is expected to preserve the topic information contained in the documents as much as possible, and at the same time to contain as little redundancy as possible, known as information richness and diversity, respectively. In automatic document summarization, the selection process of the distinct ideas included in the document is called diversity. The diversity is very important to control the redundancy in the summarized document and produce more appropriate summary. Li, Zhou, Xue, Zha, and Yu (2009) argue that an effective summarization method should properly consider the following three key requirements:

- *Coverage*: The summary should contain every important aspects of the document. By considering coverage, the information loss in summarization can be minimized.
- *Diversity*: A good document summary should be concise and contain as few redundant sentences as possible, i.e., two sentences providing similar information should not be both present in the summary. In practice, enforcing diversity in summarization can effectively reduce redundancy among the sentences.
- *Balance*: The summary should emphasize the various aspects of the document in a balanced way. An unbalanced summary usually leads to serious misunderstanding of the general idea of the original document.

In this paper, we propose the summarization models, constraint-driven document summarization, also referred to as CDDS, which aim at utilizing user-provided constraints to generate a summary with a maximum content coverage and a high diversity. The content coverage is measured by the similarity of the summary to the document. The diversity of the summary is measured by the sum of a pairwise similarity between the sentences in summary.

The contributions of this paper are as follows:

- We propose the CDDS problem, which incorporates a maximum content coverage and a high diversity constraint for creating actionable summary.
- We model the CDDS as a quadratic integer programming (QIP) problem.
- We develop the binary particle swarm optimization (PSO) algorithm to solve the QIP problem.
- We evaluate our CDDS model on real datasets and demonstrate the quality of the generated summary and the efficiency of the model.

The rest of this paper is organized as follows. Section 2 introduces the overview of related work. The proposed generic text summarization models are presented in Section 3. Section 4 presents the binary PSO algorithm to solve the QIP problem. Section 5 presents the experiments and analysis. Finally, Section 6 concludes the paper.

2. Related work

Up to now, various extraction-based techniques have been proposed for generic multi-document summarization. In order to implement extractive summarization, some sentence extraction techniques are utilized to identify the most important sentences, which can express the overall understanding of a given document. The centroid-based method, MEAD, is one of the popular extractive summarization methods (Radev, Jing, Stys, & Tam, 2004). MEAD uses information from the centroids of the clusters to select sentences that are most likely to be relevant to the cluster topic. Gong and Liu (2001) proposed a method using latent semantic analysis (LSA) to select highly ranked sentences for summarization. Other methods include NMF-based topic specification (Lee, Park, Ahn, & Kim, 2009; Wang, Li, Zhu, & Ding, 2008) and CRF-based summarization (Shen et al., 2007). In framework CRF (conditional random fields), input document is conveyed to sequence of sentences first, and then each sentence evaluated by CRF to represent its importance. Wang et al. (2008) proposed a framework based on sentence-level semantic analysis and symmetric NMF (non-negative matrix factorization). Wang, Li, and Ding (2010) proposed the weighed feature subset non-negative matrix factorization (WFS-NMF), which is an unsupervised approach to simultaneously cluster data points and select important features and different data points are assigned different weights indicating their importance. They applied proposed approach to document clustering, summarization, and visualization. Recently, Wang and Li (2012) proposed a novel weighted consensus summarization method to combine the results from different summarization methods, in which, the relative contribution of an individual method to the consensus is determined by its agreement with the other members of the summarization systems.

For effective document summarization, it is important to reduce redundancy and diversity, and extract sentences which are common to given documents. Redundancy represents how many terms or concepts are repeated across documents, while diversity or difference represents how many terms or concepts are different among the summarized documents. To remove redundancy, some systems select the top most sentences first and measure the similarity of a next candidate textual unit (sentence or paragraph) to that of previously selected ones and retain it only if it contains enough new (dissimilar) information (Sarkar, 2010). Many approaches to reduce redundancy, such as maximal marginal relevance (MMR), were reported in the literature. Carbonell and Goldstein (1998), which was used for reducing redundancy while maintaining query relevance in document re-ranking and text summarization, introduced the MMR approach. Unlike the MMR that uses greedy approach to sentence selection and redundancy removal, the clustering-based approaches control redundancy in the final summary by clustering sentences to identify themes of common information and selecting one or two representative sentences from each cluster into the final summary (Alguliev & Aliguliyev, 2008; Aliguliyev, 2009, 2010; Wang, Zhu, Li, Chi, & Gong, 2011). The work (Sarkar, 2010) presents a sentence compression based summarization technique that uses a number of local and global sentence-trimming rules to improve the performance of an extractive multi-document summarization system. Binwahlan, Salim, and Suanmali (2010) introduced a different hybrid model based on fuzzy logic, swarm intelligence and diversity selection for text summarization problem. The purpose of employing the swarm intelligence for producing the text features weights was to emphasize on dealing with the text features fairly based on their