Swarm and Evolutionary Computation 1 (2011) 213–222

Contents lists available at SciVerse ScienceDirect

Swarm and Evolutionary Computation

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

Regular paper

Sentence selection for generic document summarization using an adaptive differential evolution algorithm

Rasim M. Alguliev, Ramiz M. Aliguliyev*, Chingiz A. Mehdiyev

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

ARTICLE INFO

Article history:
Received 8 November 2010
Received in revised form
1 June 2011
Accepted 2 June 2011
Available online 2 July 2011

Keywords:
Document summarization
Redundancy
Content coverage
Optimization model
Adaptive differential evolution

ABSTRACT

For effective multi-document summarization, it is important to reduce redundant information in the summaries and extract sentences, which are common to given documents. This paper presents a document summarization model which extracts key sentences from given documents while reducing redundant information in the summaries. An innovative aspect of our model lies in its ability to remove redundancy while selecting representative sentences. The model is represented as a discrete optimization problem. To solve the discrete optimization problem in this study an adaptive DE algorithm is created. We implemented our model on multi-document summarization task. Experiments have shown that the proposed model is to be preferred over summarization systems. We also showed that the resulting summarization system based on the proposed optimization approach is competitive on the DUC2002 and DUC2004 datasets.

1. Introduction

Interest in text mining started with the advent of online publishing, the increased impact of the Internet and the rapid development of electronic government (e-government). With the exponential growth of the information–communication technologies a huge amount of electronic documents are available online. This explosion of electronic documents has made it difficult for users to extract useful information from them. In this case, the user due to the large amount of information [1] does not read many relevant and interesting documents.

The text mining approach is feasible and powerful for e-government digital archives. Digital archives have been built up in almost every level of e-government hierarchy. Digital archives in the domain of e-government involve various medium formats, such as video, audio and scanned document. In fact, governmental documents are the most important production of e-government, which contain the majority information of government affairs. The text mining approach described in [2] targets the text in the scanned documents. The mined knowledge helps a lot in policy making, emergency decision support, and government routines for civil servants. The successful application of the system to archives testifies the correctness and soundness of this approach [2].

As the Internet is growing exponentially, huge amounts of information are available online. It is difficult to identify the relevant information to satisfy the information needs of users. The problem of information overloading can be reduced by automatic document summarization together with conventional information search engines to efficiently access the relevance of retrieved documents [1]. Present search engines usually provide a short summary for each retrieved document in order that users can quickly skim through the main content of the page. Therefore, it saves users' time and improves the search engine's service quality [3,4]. That is why the necessity of tools that automatically generate summaries arises. These tools are not just for professionals who need to find the information in a short time but also for large searching engines such as Google, Yahoo!, AltaVista, and others, which could obtain benefits in its results if they use automatic generated summaries. After that, the user only will require the interesting documents, reducing the information flow [1].

Depending on the number of documents to be summarized, the summary can be a single-document or a multi-document [5,6]. Single-document summarization can only condense one document into a shorter representation, whereas multi-document summarization can condense a set of documents into a summary. Multi-document summarization can be considered as an extension of single-document summarization and used for precisely describing the information contained in a cluster of documents and facilitate users to understand the document cluster. Since it combines and integrates the information across documents, it performs knowledge synthesis and knowledge discovery, and can be used for knowledge acquisition [6].

* Corresponding address: Institute of Information Technology of NAS of Azerbaijan 9, F.Agayev street, Baku, Azerbaijan, AZ1141, Azerbaijan. Tel.: +994 12 5390167; fax: +994 12 5396121.

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

2210-6502/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.swevo.2011.06.006

2. Related work

Many summarization methods have been proposed in the literature [4,7–11]. Generally, document summarization methods can be divided into two categories: abstractive and extractive [8,12]. Extractive summarization is a simple but robust method for text summarization and it involves assigning saliency scores to some textual units of the documents and extracting those with highest scores. Abstraction can be described as reading and understanding the text to recognize its content, which is then compiled in a concise text. In general, an *abstract* can be described as a summary comprising concepts/ideas taken from the source, which are then reinterpreted and presented, in a different form, whilst an *extract* is a summary consisting of units of text taken from the source and presented verbatim [13].

This section outlines related work done in summarization particularly extracting sentences from a document. In fact majority of research have been focused on summary extraction, which selects the pieces such as keywords, sentences or even paragraph from the source to generate a summary. In this paper, we also focus on extraction-based methods. To date, various extraction-based methods have been proposed for generic document summarization. The centroid-based method is one of the popular extractive summarization methods [14]. Gong et al. [15] propose a method using latent semantic analysis (LSA) to select highly ranked sentences for summarization. Other methods include NMF-based topic specification [9,10,16] and CRF-based summarization [3]. Paper [9] proposes a framework based on sentence-level semantic analysis and symmetric NMF (Non-negative Matrix Factorization). In [17], text summarization modeled as a maximum coverage problem that aims at covering as many conceptual units as possible by selecting some sentences. McDonald [18] formalized text summarization as a knapsack problem and obtained the global solution and its approximate solutions. In [19], Takamura and Okamura represented text summarization as a maximum coverage problem with the knapsack constraint (MCKP). Shen et al. [3] presented a Conditional Random Fields (CRF) based framework for generic summarization and reported that CRF performed better than many existing models, such as HMM and SVM. A common feature of all these works is that they all relied on classification models to rank sentences. In [20], text summarization formalized as a budgeted median problem. This model covers the whole document cluster through sentence assignment. An advantage of this method is that it can incorporate asymmetric relations between sentences in a natural manner. The work [10] proposes a Bayesian sentence-based topic model (BSTM) for multi-document summarization by making use of both the term-document and term-sentence associations. It models the probability distributions of selecting sentences given topics and provides a principled way for the summarization task. Huang et al. [21] consider document summarization as a multiobjective optimization problem. In particular, they formulate four objective functions, namely information coverage, significance, redundancy and text coherence.

We model text summarization task as an optimization problem. One of the advantages of this approach is that it directly discovers key sentences in the given collection and covers the main content of the original source(s). Other advantage of our model is that it can reduce redundancy in the summary. In this paper, an adaptive differential evolution algorithm is created to solve the optimization problem. The performance of the proposed approach is tested on the standard DUC2002 and DUC2004 datasets and is compared with baseline systems. The effectiveness of the proposed approach is demonstrated.

The rest of this paper is organized as follows. Sentence selection problem for text summarization is introduced in Section 3. This problem is formulated as an optimization problem. Section 4 briefly describes the basics of differential evolution algorithm. Section 5 describes a modified DE algorithm for solving the optimization problem. The numerical experiments and results are given in Section 6. Finally, we conclude our paper in Section 7.

3. Formulation of sentence selection problem

In this section, we formalize sentence-extraction-based summarization of multiple documents as an optimization problem.

3.1. Problem statement

We present our approach toward all of the three aspects of summarization, namely: (1) *content coverage*, summary should contain salient sentences that cover the main content of the documents, (2) *redundancy*, summaries should not contain multiple sentences that convey (carry) the same information, and (3) *length*, summary should be bounded in length. Optimizing all three properties jointly is a challenging task and is an example of a global summarization problem. That is why the inclusion of relevant textual units relies not only on properties of the units themselves, but also on properties of every other textual unit in the summary [1].

3.2. Mathematical formulation of optimization problem

Given a document collection $D = \{d_1, d_2, \ldots, d_N\}$, where N is the number of documents. For simplicity, we represent the document collection simply as the set of all sentences from all the documents in the collection, i.e. $D = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes the ith sentence in D, n is the number of sentences in the document collection. We attempt to find a subset of the sentences $D = \{s_1, s_2, \ldots, s_n\}$ that covers the main content of the document collection while reducing the redundancy in the summary. If we let $s_i \in D$ be the sentence constituting a summary, then the similarity between the set of sentences and the sentence is going to be $\mathrm{sim}(D, s_i)$, which we would like to maximize. On the other hand, for redundancy avoiding in the summary we choose those sentences which similarity between them was minimum. Then the text summarization task can be formulated as follows:

$$\text{maximize} \quad f(X) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} [\mathrm{sim}(s_i, O) + \mathrm{sim}(s_j, O)]x_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{sim}(s_i, s_j)x_{ij}}, \tag{1}$$

$$\text{subject to} \quad L - \varepsilon \le \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (l_i + l_j)x_{ij} \le L + \varepsilon, \tag{2}$$

$$x_{ij} \in \{0, 1\}, \quad \forall i, j. \tag{3}$$

x_{ij} denotes a variable which is 1 if pair of sentences s_i and s_j are selected to be included to the summary, otherwise 0. According to this definition, we have that $x_{ij} = x_{ji}$. L is a length of summary, l_i denotes the length of sentence s_i, O is the mean vector of the collection $D = \{s_1, s_2, \ldots, s_n\}$ which will be defined below. It is known that it is not possible to precisely form a summary with the given length. Therefore, in this model a tolerance ε is introduced, which we define as $\varepsilon = \max_{i=1,\ldots,n} \mathrm{len}\,(s_i) - \min_{i=1,\ldots,n} \mathrm{len}\,(s_i)$. The number of words or in bytes measures the lengths of summary and sentence.

From [14] we know that the center of the document collection reflects the main content of document collection. Thus, in Eq. (1) the numerator evaluates the importance of the sentence s_i and s_j by measuring their similarity to the center O of document collection D. In Eq. (1), the denominator evaluates the correlation between the sentences s_i and s_j. The numerator provides the covering of the