Expert Systems with Applications 40 (2013) 1675–1689

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Multiple documents summarization based on evolutionary optimization algorithm

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

Institute of Information Technology of Azerbaijan National Academy of Sciences, 9, B. Vahabzade Street, Baku AZ1141, Azerbaijan

ARTICLE INFO

Keywords:
Multi-document summarization
Diversity
Content coverage
Optimization model
Differential evolution algorithm
Self-adaptive crossover

ABSTRACT

This paper proposes an optimization-based model for generic document summarization. The model generates a summary by extracting salient sentences from documents. This approach uses the sentence-to-document collection, the summary-to-document collection and the sentence-to-sentence relations to select salient sentences from given document collection and reduce redundancy in the summary. To solve the optimization problem has been created an improved differential evolution algorithm. The algorithm can adjust crossover rate adaptively according to the fitness of individuals. We implemented the proposed model on multi-document summarization task. Experiments have been performed on DUC2002 and DUC2004 data sets. The experimental results provide strong evidence that the proposed optimization-based approach is a viable method for document summarization.

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). With the exponential growing 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. While the Internet has increased access to text collections on a variety of topics, consumers now face a considerable amount of redundancy in the texts that they encounter online. In this case, the user due to the large amount of information does not read many relevant and interesting documents. 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 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 Dong, Yu, and Jiang (2009) targets the text in the scanned documents. The mined knowledge helps a lot in policymaking, 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.

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). Boydell and Smyth (2010) focus on the role of snippets in collaborative web search and describe a technique for summarizing search results that harnesses the collaborative search behavior of communities of like-minded searchers to produce snippets that are more focused on the preferences of the searchers. They go on to show how this so-called *social summarization* technique can generate summaries that are significantly better adapted to searcher preferences and describe a novel personalized search interface that combines result recommendation with social summarization.

Depending on the number of documents, summarization techniques can be classified into two classes: single-document and multi-document (Fattah & Ren, 2009; Zajic, Dorr, & Lin, 2008). Single-document summarization can only condense one

* 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), r.aliguliyev@gmail.com, a.ramiz@science.az, 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.2012.09.014

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 (Zajic et al., 2008). In addition to single document summarization, which has been first studied in this field for years, researchers have started to work on multi-document summarization whose goal is to generate a summary from multiple documents. The multi-document summarization task has turned out to be much more complex than summarizing a single document, even a very large one. This difficulty arises from inevitable thematic diversity within a large set of documents. A multi-document summary can be used to concisely describe the information contained in a cluster of documents and to facilitate the users to understand the document cluster.

2. Related work

Multi-document summarization has been widely studied recently. Researchers all over the world working on multi-document summarization are trying different directions to see methods that provide the best results (Tao, Zhou, Lam, & Guan, 2008; Wan, 2008; Wang, Li, Zhu, & Ding, 2008; Wang, Zhu, Li, Chi, & Gong, 2011; Wang, Li, Zhu, & Ding, 2009). In general, document summarization can be divided into extractive summarization and abstractive summarization. Extractive summarization produces summaries by choosing a subset of the sentences in the original document(s). This contrasts with abstractive summarization, where the information in the text is rephrased. An extract summary consists of sentences extracted from the document, while an abstract summary employs words and phrases not appearing in the original document (Mani & Maybury, 1999). 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 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 (Kutlu, Cigir, & Cicekli, 2010). Although an abstractive summary could be more concise, it requires deep natural language processing techniques. Thus, an extractive summary is more feasible and has become the standard in document summarization. In this paper, we focus on extractive multi-document summarization. There are several most widely used extractive summarization methods as follows.

Summaries can be generic or query-focused (Dunlavy, O'Leary, Conroy, & Schlesinger, 2007; Gong & Liu, 2001; Ouyang, Li, Li, & Lu, 2011; 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. Zhao, Wu, and Huang (2009) propose a query expansion algorithm used in the graph-based ranking approach for query-focused multi-document summarization. This algorithm makes use of both sentence-to-sentence relationships and sentence-to-word relationships to select expansion words from the documents. By this method, the expansion words satisfy both information richness and query relevance. The problem of using topic representations for multi-document summarization has received considerable attention recently. Several topic representations have been employed for producing informative and coherent summaries. The work presented in Harabagiu and Lacatusu (2010) has two main goals. First, it introduces two novel topic representations that leverage sets of automatically generated topic themes for multi-document summarization. It shows how these new topic representations can be integrated into a state-of-the-art multi-document summarization system. Second, it presents eight different methods of generating multi-document summaries.

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 et al., 2008, 2009) 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.

The graph-based ranking algorithms such as PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1999) have also been used in generic multi-document summarization. The major concerns in graph-based summarization researches include how to model the documents using text graph and how to transform existing web page ranking algorithms to their variations that could accommodate various summarization requirements (Wenjie, Furu, Qin, & Yanxiang, 2008). A similarity graph is produced for the sentences in the document collection. In the graph, each node represents a sentence. The edges between nodes measure the cosine similarity between the respective pair of sentences where each sentence is represented as a vector of term specific weights. An algorithm called LexRank (Erkan & Radev, 2004), adapted from PageRank, was applied to calculate sentence significance, which was then used as the criterion to rank and select summary sentences. In Chali, Hasan, and Joty (2011), authors extensively study the impact of syntactic and semantic information in measuring similarity between the sentences in the random walk framework for answering complex questions. They apply the tree kernel functions and Extended String Subsequence Kernel (ESSK) to include syntactic and semantic information. Ordering extracted sentences into a coherent summary is a non-trivial task. Bollegala, Okazaki, and Ishizuka (2010) presented a bottom-up approach to arrange