Expert Systems with Applications 41 (2014) 4158–4169

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Extractive single-document summarization based on genetic operators and guided local search

Martha Mendoza[a,b,*], Susana Bonilla[a], Clara Noguera[a], Carlos Cobos[a,b], Elizabeth León[c]

[a] *Information Technology Research Group (GTI), Universidad del Cauca, Sector Tulcán Office 450, Popayán, Colombia*
[b] *Computer Science Department, Electronic and Telecommunications Engineering Faculty, Universidad del Cauca, Colombia*
[c] *Data Mining Research Group (MIDAS), Engineering Faculty, Universidad Nacional de Colombia, Bogotá, Colombia*

ARTICLE INFO

Keywords:
Extractive summarization
Single document
Memetic algorithm
Guided local search

ABSTRACT

Due to the exponential growth of textual information available on the Web, end users need to be able to access information in summary form – and without losing the most important information in the document when generating the summaries. Automatic generation of extractive summaries from a single document has traditionally been given the task of extracting the most relevant sentences from the original document. The methods employed generally allocate a score to each sentence in the document, taking into account certain features. The most relevant sentences are then selected, according to the score obtained for each sentence. These features include the position of the sentence in the document, its similarity to the title, the sentence length, and the frequency of the terms in the sentence. However, it has still not been possible to achieve a quality of summary that matches that performed by humans and therefore methods continue to be brought forward that aim to improve on the results. This paper addresses the generation of extractive summaries from a single document as a binary optimization problem where the quality (fitness) of the solutions is based on the weighting of individual statistical features of each sentence – such as position, sentence length and the relationship of the summary to the title, combined with group features of similarity between candidate sentences in the summary and the original document, and among the candidate sentences of the summary. This paper proposes a method of extractive single-document summarization based on genetic operators and guided local search, called MA-SingleDocSum. A memetic algorithm is used to integrate the own-population-based search of evolutionary algorithms with a guided local search strategy. The proposed method was compared with the state of the art methods UnifiedRank, DE, FEOM, NetSum, CRF, QCS, SVM, and Manifold Ranking, using ROUGE measures on the datasets DUC2001 and DUC2002. The results showed that MA-SingleDocSum outperforms the state of the art methods.

1. Introduction

Due to the exponential growth of textual information available on the Web and the access to information by the users through new portable devices, it is necessary that the end user can access the information in summary form and without losing the most important aspects presented therein. Some of the application areas of the generation of extractive summaries from a single document are the summaries of web pages presented on the search engines (Porselvi & Gunasundari, 2013); the assignation of the labels to groups generated in the web document clustering (Carpineto, Osinski, Romano, & Weiss, 2009); and in the E-learning context is used to select the most important information from a text (Kumaresh & Ramakrishnan, 2012). The automatic generation of text summaries has been tasked with addressing this problem for many years, seeking to obtain short texts that present the most relevant ideas in a document (Lloret & Palomar, 2012; Nenkova & McKeown, 2012; Spärck Jones, 2007). To achieve this, several methods have been developed that summarize one or multiple documents, with the aim that the user select and review in the shortest time those documents that really meet their information needs.

Different taxonomies for the summaries exist (Ježek & Steinberger, 2008; Lloret & Palomar, 2012; Nenkova & McKeown, 2012), based on the way the summary is generated, the target audience of the summary, the number of documents to be summarized, and so on.

According to the way in which it is generated, the summary may represent either an extraction or an abstraction

[*] Corresponding author at: Information Technology Research Group (GTI), Universidad del Cauca, Sector Tulcán Office 450, Popayán, Colombia. Tel.: +57 28366524; fax: +57 28209810.

E-mail addresses: mmendoza@unicauca.edu.co, mendoza.martha.eliana@gmail.com (M. Mendoza).

0957-4174/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.042

(Ježek & Steinberger, 2008; Lloret & Palomar, 2012; Nenkova & McKeown, 2012). Extraction summaries are formed from the reuse of portions of the original text. Abstraction based summaries, on the other hand, are rather more complex, requiring linguistic analysis tools to construct new sentences from those previously extracted.

Depending on the target audience, summaries may be (Ježek & Steinberger, 2008; Lloret & Palomar, 2012; Nenkova & McKeown, 2012) generic, query-based, user-focused or topic-focused. Generic summaries do not depend on the audience for whom the summary is intended. Query-based summaries respond to a query made by the user. User-focused ones generate summaries to tailor the interests of a particular user, while topic-focused summaries emphasize those summaries on specific topics of documents.

With regard to the number of documents that are processed, summaries (Ježek & Steinberger, 2008; Lloret & Palomar, 2012; Nenkova & McKeown, 2012) can be either single document or multiple document. In addition, as regards the language of the document, they may be monolingual or multilingual, and regarding document genre may be scientific article, news, blogs, and so on.

A huge diversity is to be found among the methods of automatic generation of extractive summaries from a single document. These are mainly based on the handling of basic statistical features such as sentence position and terms frequency (Edmundson, 1969), based on machine learning techniques (Aone, Okurowski, Gorlinsky, & Larsen, 1999; Conroy & O'leary, 2001; Dunlavy, O'Leary, Conroy, & Schlesinger, 2007; Kupiec, Pedersen, & Chen, 1995; Shen, Sun, Li, Yang, & Chen, 2007; Svore, Vanderwende, & Burges, 2007; Wong, Wu, & Li, 2008), connectivity of texts (Barzilay & Elhadad, 1997; Louis, Joshi, & Nenkova, 2010; Marcu, 1998; Ono, Sumita, & Miike, 1994), graphs (Mihalcea & Tarau, 2004) (Wan, 2010), algebraic reduction (Gong, 2001; Lee, Park, Ahn, & Kim, 2009; Steinberger & Jezek, 2004; Steinberger & Ježek, 2006; Yeh, Ke, Yang, & Meng, 2005) and evolutionary models (Abuobieda, Salim, Kumar, & Osman, 2013; Aliguliyev, 2009a; Binwahlan, Salim, & Suanmali, 2009, 2010; Dehkordi, Kumarci, & Khosravi, 2009; Fattah & Ren, 2009; García-Hernández & Ledeneva, 2013; Litvak, Last, & Friedman, 2010; Qazvinian, Sharif, & Halavati, 2008; Shareghi & Hassanabadi, 2008; Steinberger & Ježek, 2006).

Evolutionary algorithms have traditionally shown good results in solving the problem of extractive summarization (Aliguliyev, 2009a; Binwahlan et al., 2009, 2010; Fattah & Ren, 2009; Litvak et al., 2010; Qazvinian et al., 2008; Shareghi & Hassanabadi, 2008; Steinberger & Ježek, 2006), while memetic algorithms (evolutionary algorithms with local search heuristics) have contributed to the successful resolution of different combinatory optimization problems (Cobos, Montealegre, Mejía, Mendoza, & León, 2010; Neri & Cotta, 2012). Nevertheless, memetic algorithms have not until now been used for solving the specific problem of extractive single-document summarization. In this paper, therefore, we propose a method of generic-extractive summarization for a monolingual document of any genre, based on memetic algorithms. In this case, the evaluation was done using news items.

This algorithm, called MA-SingleDocSum, defines the quality of a solution based on the weighting of individual statistical features of each sentence, such as position, sentence length and the relation of the summary to the title, combined with group features based on the similarity between candidate sentences in the summary and the original document, and the similarity among the sentences in the summary in order to obtain coverage of the summary and cohesion of summary sentences. The algorithm consists of rank-based and roulette wheel parent selection, one-point crossover, multi-bit mutation, guided search-based local optimization, and restricted competition replacement.

The rest of the paper is organized as follows: Section 2 introduces work related to automatic generation of the extractive summaries from a single document; document representation, similarity measures, and features of the objective function proposed are presented in Section 3; the strategies for selection, crossover, mutation, local search and replacement that make up the proposed memetic algorithm are described in Section 4; while the results of evaluation using data sets, along with a comparison and analysis with other state of the art methods, are presented in Section 5; and finally, Section 6 presents the conclusions and future work.

2. Related work

Early research suggests as relevant factors for the score of a sentence and its inclusion in the summary the use of the frequency of occurrence of a term in a text, the position of the sentences in the document, and the presence of keywords or words from the document title in the sentences (Edmundson, 1969).

Using the machine learning approach, Bayes' Theorem has been applied to develop a function that estimates the probability that a sentence be included in a summary (Aone et al., 1999; Kupiec et al., 1995). As such, an approach is proposed based on the Hidden Markov Model (HMM), whose main feature is the recognition of local dependencies between sentences through a sequential model (Conroy & O'leary, 2001; Dunlavy, O'Leary, Conroy, & Schlesinger, 2007). Neural networks (Svore et al., 2007) and Conditional Random Fields (Shen et al., 2007) are also used. More recently, the Probabilistic Support Vector Machine (PSVM) and Naïve Bayesian Classifier were used in an semi-supervised learning approach (Wong et al., 2008)

Other works have applied approaches based on text connectivity, in order to establish the connections that may exist between different parts of a text to try to achieve more coherent and more understandable summaries (Marcu, 1998; Ono et al., 1994). Highlighted among these is the use of lexical chains. This approach starts with the segmentation of the original text and continues with the construction of lexical chains, the identifying the strongest chains and extracting the most significant sentences, completing the process of the production of the summary (Barzilay & Elhadad, 1997). More recently, the rhetorical structure theory approach has also been employed (Louis et al., 2010).

In addition, the graphs have been adapted for the automatic generation of extractive summaries (Mihalcea & Tarau, 2004), where the sequence of one or more lexical units extracted from a text and the relationships between them are the vertices and edges of the graph, respectively. A particular focus based on graphs is that proposed by Wan (2010), in which the automatic summarization of one and of multiple documents is carried out at the same time, making use of a local importance that indicates the relevance of a sentence within a document to generate the summary of a single document; and of a global importance, that indicates the relevance of the same sentence but at the level of the entire set of documents to generate the summary of multiple documents.

In the case of algebraic reduction, the most widely used method for extractive summarization is that based on Latent Semantic Analysis (LSA), which allows the extracting, representing and comparing the meaning of words using the algebraic-statistical analysis of a text, the basic assumption for which is that the meaning of a word is determined by its frequent occurrence next to other words. Gong (2001) proposed using LSA for automatic generation of generic summaries, applying Singular Value Decomposition (SVD). The semantic analysis process consists of two steps. The first is the creation of a terms by sentence matrix $A = [A_1, A_2, \ldots, A_n]$, where each column A_i represents the weight vector, based on the frequency of terms from the sentence i in the document. The next step consists of applying SVD to matrix A. To generate a summary,