Applied Computing and Informatics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Applied Computing and Informatics

journal homepage: www.sciencedirect.com

An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA

Rasmita Rautray [a],*, Rakesh Chandra Balabantaray [b]

[a] Department of Computer Science and Engineering, Siksha 'O' Anusandhan University, Bhubaneswar 751030, Odisha, India
[b] Department of Computer Science, IIIT, Bhubaneswar, Odisha, India

ARTICLE INFO

Article history:
Received 21 February 2017
Revised 1 May 2017
Accepted 11 May 2017
Available online xxxx

Keywords:
Text summarizer
Multi-document summarization
Extractive summary
Cuckoo search

ABSTRACT

In today's scenario the rate of growth of information is expanding exponentially in the World Wide Web. As a result, extracting valid and useful information from a huge data has become a challenging issue. Recently text summarization is recognized as one of the solution to extract relevant information from large documents. Based on number of documents considered for summarization, the summarization task is categorized as single document or multi-document summarization. Rather than single document, multi-document summarization is more challenging for the researchers to find accurate summary from multiple documents. Hence in this study, a novel Cuckoo search based multi-document summarizer (MDSCSA) is proposed to address the problem of multi-document summarization. The proposed MDSCSA is also compared with two other nature inspired based summarization techniques such as Particle Swarm Optimization based summarization (PSOS) and Cat Swarm Optimization based summarization (CSOS). With respect to the benchmark dataset Document Understanding Conference (DUC) datasets, the performance of all algorithms are compared in terms of ROUGE score, inter sentence similarity and readability metric to validate non-redundancy, cohesiveness and readability of the summary respectively. The experimental analysis clearly reveals that the proposed approach outperforms the other summarizers included in this study.

1. Introduction

Now a day's the rate of information growth is expanding exponentially in the World Wide Web, which creates information overload problem. One solution to this problem is shortening of information, called text summarization (TS). Text summarization is the process of creating shorter version of original text without losing main contents [1] called summary. The summary provides a quick guide to create interest on information, helps in making decision on document whether it is readable or not as well as it is served as a time saver for users [2]. The way in which summary is generated either is an extraction or an abstraction method [3,4]. Extraction based summaries are generated by selecting the important portions of the original text. Whereas, abstraction based summaries requires linguistic analysis to construct new sentences from the original text [5,6]. Based on dimension, extraction based summaries can be categorized into two ways i.e., generic or query dependent [7]. Generic summary reflects the major content of the documents without any additional information. But, Query-dependent summary focuses on the information expressed in the given queries [8,9].

Number of documents considered for generating summary, can classify the summarization problem as single document or multi-document summarization [10,11]. When a document is condensed into a shorter version, it is called single document summarization, whereas condensing a set of documents into a summary is called multi-document summarization. Therefore, summarization of multiple documents can be considered as an extension of summarization of single document [12]. In multi-document summarization, search space is larger compared to single document summarization, which makes it more challenging for extracting important sentences. In that context, multi-document summarization can be considered as an optimization problem with the objective of producing optimal summary containing informative

* Corresponding author.
E-mail addresses: rashmitaroutray@soauniversity.ac.in (R. Rautray), rakesh@iiit-bh.ac.in (R.C. Balabantaray).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2017.05.003
2210-8327/© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.

sentences of the original documents. Nature inspired optimization based approaches are the suitable choices to address this optimization problem. In literature several meta heuristic techniques such as particle swarm optimization (PSO), differential evolution (DE), harmony search (HS), Cuckoo search (CS) and genetic algorithm (GA) are applied in single or multi-document summarization.

Being inspired by the application of Cuckoo search in other optimization problems [13–34], in this study a novel Cuckoo search algorithm based summarizer is presented for multi-document summarization. Though single document using Cuckoo search algorithm is present in literature [35] but, multi-document summarizer using Cuckoo search is new to this area. Further the model is also compared with Particle Swarm Optimization based summarizer and Cat Swarm Optimization based summarizer. The performance of such models are analyzed over DUC datasets with respect to few summary evaluation metrics such as ROUGE score, inter sentence similarity and readability metric. These evaluation metrics are considered to validate the non-redundancy, cohesiveness and readability of the generated summary.

The structure of paper is organized as follows. Section 2 briefly describes the related works on text summarization problem using global optimization techniques. Section 3 introduces the proposed extractive summarization model. Section 4 presents Cuckoo search based summarizer for solving summarization problem. Next, Section 5 details the numeric calculation for objective function, Section 6 elaborates on experiments and result analysis and finally Section 7 addresses the conclusions.

2. Related works

In this section, a theoretical study of evolutionary algorithms based text summarization and various applications of Cuckoo search algorithm is discussed.

In multi-document summarization, compression of multiple documents, speed of sentence extraction, redundancy between sentences and sentence selection are the critical issues in the formation of useful summaries. In the past, such issues are resolved by statistical tools. But, due to significantly poor performance of statistical tools in text extraction, from 2000 onwards a number of global optimization techniques such as particle swarm optimization (PSO) [2,11,36–38], differential evolution (DE) [1,7,11,12,36,37,39–44], and genetic algorithm (GA) [10,45–51] are proposed by several researchers for improving the performance of sentence selection in document summarization. Initially, the optimization algorithm GA was first used in test summarization problem [45] to retrieve relevant document based on query and relevant judgments. Thereafter in [46], the author has evaluates the efficiency of GA with fitness functions for relevance feedback in information retrieval problem for maintaining the document order. Later on GA based programming technique is used for fuzzy retrieval system to extract information based on query by applying off-line adaptive process [48] and in [49], the author has used GA for text summarization based on sentence score. Each sentence score is obtained through the comparison of each sentence with all other sentences as well as with the document title by cosine measure. The informative features weights are calculated using GA to influence the words relevancy. Word relevancy defines relevancy and rank of the sentences having highest score with respect to a threshold, are selected as summary sentences. A single document generic summary has been extracted based on different sentence features using GA by comparing with some other techniques and were evaluated using ROUGE score [10]. Kogilavani et al. [50] Presents a feature based multi-document generic summarization using GA & clustering to enhance the summary quality by maximizing length, coverage and informativeness while minimizing the redundancy. Whereas, genetic algorithm based document summarization has been proposed to generate optimal summary by combining article sentences and query sentence to achieve satisfied length, high coverage, high informativeness and low redundancy in summary [51,52]. However the GA is providing better result for text summarization. But GA suffers from issues of more parameter tunning [39]. To obtain better summary with less parameter tunning, the authors of [1,7,40,41] have used DE for text summarization problem. Aliguliyev [1] presents a generic document summarizer based on sentence clustering using DE. Whereas in [42], a single document summarizer focuses on sentence feature as key ingredient instead of clustering to extract summary. A summarizer for single document based on clustering has been presented and made comparison of discrete DE and conventional DE for summarization and showed comparison result by the authors of [36]. Alguliev et al. [43] have used DE algorithm to enhance sentence feature based summary by maximizing content coverage, readability and cohesion to improve text readability and informativeness of summary. As the problem of summarization is considered as discrete optimization problem in [43], to solve such problem the author has used adaptive DE to maximize informativeness of summary while reducing the redundancy of summary. In contrast, the summarization problem is considered as p-median problem and Quadratic Boolean programming problem by the authors of [7,40], for that a new variation of DE with self adaptive mutation and crossover parameters and binary DE is used. Where as in [43], adaptive crossover parameter is used for optimizing the summary result. The models discussed in [7,12,39] not only express sentence-to-sentence relationship, but also express summary-to-document and summary-to-subtopics relationships. In all the above cases, DE based summarizer is showing significantly better result than GA based summarizer both for single and multi-document summarization.

Rautray and Balabantaray [37] presents a generic summarizer for single document using particle swarm optimization algorithm, by considering content coverage and redundancy feature as key aspects of summary. For solving such problem, the objective function is designed by taking weighted average of content coverage and redundancy features. Another PSO based single document summarizer is also proposed in [11], which has used the same objective function as described in [37], but by taking features of text as an input arguments instead of sentence weights as input arguments to the model. Binwahlan et al. [2] have presented a PSO based extractive summarizer where expression of ROUGE is used as fitness functions for extraction of summary sentences. The summary based on PSO is also presented by Asgari et al. [38] considering summary features such as content coverage, readability and length. A multi-document summarization system using PSO has been presented in [36] based on the concept of clustering of sentences by calculating inter sentence similarity between sentences and sentence to document set to achieve content coverage and diversity of summary. In contrast, similarity metric also used by Alguliev et al. [44] to achieve content coverage, diversity and length of summary for multiple document sets. Rautray et al. [53] presents cat swarm optimization (CSO) algorithm based multi document summarizer, which takes content coverage, readability and cohesion as key aspects of summary. The summary is evaluated over DUC dataset and compared with two other optimization algorithms such as particle swarm optimization and harmony search algorithm, in which CSO shows competitively better result than other two algorithms.

Cobos et al. [15] have implemented Cuckoo search algorithm for web document clustering or web clustering engine. Cuckoo search uses Balanced Bayesian Information Criteria for fitness function and compared against existing clustering algorithms for web document, Suffix Tree Clustering, Lingo and Bisecting K-mean

Please cite this article in press as: R. Rautray, R.C. Balabantaray, Applied Computing and Informatics (2017), http://dx.doi.org/10.1016/j.aci.2017.05.003