ARTICLE IN PRESS

+ MODEL

Available online at www.sciencedirect.com

ScienceDirect

Karbala International Journal of Modern Science xx (2017) 1–12
http://www.journals.elsevier.com/karbala-international-journal-of-modern-science/

Bio-inspired approaches for extractive document summarization: A comparative study

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

Received 14 March 2017; revised 17 May 2017; accepted 6 June 2017

Abstract

With the exponential growth of information in World Wide Web, extracting relevant information from huge amount of data has become a critical task. Text summarization has been appeared as one of the solution to such problem. As the main objective is to retrieve a condensed document that pertain the original information, so it can be considered as an optimization problem. In this paper, a comparative analysis of few meta-heuristic approaches such as Cuckoo Search (CS), Cat Swarm Optimization (CSO), Particle Swarm Optimization (PSO), Harmony Search (HS), and Differential Evolution (DE) algorithm is presented for single document summarization problem. The performance of all these algorithms are compared in terms of different evaluation metrics such as F score, true positive rate and positive predicate value to validate summary relevancy and non-redundancy over traditional and standard Document Understanding Conference (DUC) datasets.
© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of University of Kerbala. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Document summarization; Cuckoo Search; Extractive summary; Bio-inspired approach

1. Introduction

Impact of increasing information in World Wide Web and digital libraries led's information overload problem. In this case, extracting relevant information from massive data is a difficult task for user. To handle such task, text summarization has been used as a solution in information retrieval field. Text summarization is the process of creating a summary that is a shorter version of text from the original text documents. The problem of generating coherent summaries are of two types: extractive and abstractive [1–3]. Extractive is the method to generate summaries by selecting the important portions of the original text document. Whereas, abstractive method requires linguistic analysis such as semantic representation, inference and natural language generation to construct new sentences from the original text [3–5]. Based on dimension, extractive summarization can be categorized into two ways such as: query focused or generic. A query focused summary favors specific themes of the text in response to a users' query, while generic summary reflects the author's point of view for the input

* Corresponding author.
E-mail addresses: rashmitaroutray@soauniversity.ac.in (R. Rautray), rakesh@iiit-bh.ac.in (R.C. Balabantaray).
Peer review under responsibility of University of Kerbala.

http://dx.doi.org/10.1016/j.kijoms.2017.06.001
2405-609X/© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of University of Kerbala. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article in press as: R. Rautray, R.C. Balabantaray, Bio-inspired approaches for extractive document summarization: A comparative study, Karbala International Journal of Modern Science (2017), http://dx.doi.org/10.1016/j.kijoms.2017.06.001

text providing an equal importance to all major themes in it [6–8].

Based on language and depending on its input and output text, summary can be categorized into three types: Mono-lingual, Multi-lingual and Cross-lingual. The mono-lingual system accepts documents with specific language. Thus, the input and output languages are same. But in multi-lingual case, the systems can accept documents in different languages and the users can choose the languages of the output summary. Whereas, the output and input languages different from each other for generating cross-lingual summary [8].

Depending on the document to be summarized, the summary either can be a single document summary or multi document summary. The aim of single document summary is to produce a concise summary from single document whereas, producing a concise summary from multiple documents, is called multi document summary. This paper presents a comparative analysis of extractive mono-lingual bio-inspired algorithm based summarization systems, each of which produces a summary from single document based on sentence informative score. Though the generic model for summary generation is presented using Cuckoo Search (CS) algorithm, Cat Swarm Optimization (CSO), Particle Swarm Optimization (PSO), Harmony Search (HS), and Differential Evolution (DE) algorithm, but the cuckoo search algorithm is used to select best combination of sentences for generating summary. The performance of such models is analyzed with respect to few summary evaluation metrics such as F score, true positive rate and positive predicate value to validate summary relevancy and non-redundancy [9–11]. From the experimental result analysis over DUC datasets, it is clearly observed that the performance of CS based document summarizer is showing better result than DE, PSO, HS and CSO based summarizer.

The remaining part of this paper is arranged as follows. Section 2 briefly describes literature survey on text summarization problem using global optimization techniques. Section 3 focuses on summarization steps. Section 4 presents bio-inspired framework for single document summarization. Next, Section 5 includes detailed steps of bio-inspired framework based document summarization. Lastly experimental analysis and result discussion are given in Section 6 and Section 7 addresses the conclusions.

2. Literature survey

In this survey, a theoretical study of bio-inspired algorithm based text summarization is discussed. In literature, as text summarization is considered an optimization problem, genetic algorithm (GA) was first used to retrieve relevant document based on query and relevant judgments in Ref. [12]. López-Pujalte et al. [13] have compared the efficiency of GA with the Ide dec-hi method for relevance feedback in information retrieval problem for maintaining the document order. Later on GA based programming technique is used for fuzzy retrieval system, which is automatically learn weighted queries and modeling the user's need to extract information based on query by applying off-line adaptive process [14].

Considering the sentence score, in Ref. [15] GA has been used for text summarization. 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 by 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.

Rautray, & Balabantaray, 2015 [9] 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 Ref. [10], which has used the same objective function as described in Ref. [9], but by taking features of text as an input arguments instead of sentence weights as input arguments to the model. The summary based on PSO summarizer is also presented in Ref. [16] by considering summary features such as content coverage, readability and length. Though a number of optimization algorithms have been proposed in literature [17–22], few text summarization problems with respect to single document using bio-inspired techniques is listed in Table 1.

3. Text summarization

Text summarization is an automatic process to create a concise and comprehensive document. The entire process goes through three basic steps such as preprocessing, processing and summary generation. As a result a summary is generated. Here steps involved in the proposed summarizer for a single document is discussed in Table 2:

Please cite this article in press as: R. Rautray, R.C. Balabantaray, Bio-inspired approaches for extractive document summarization: A comparative study, Karbala International Journal of Modern Science (2017), http://dx.doi.org/10.1016/j.kijoms.2017.06.001

Download English Version:

https://daneshyari.com/en/article/6899127

Download Persian Version:

https://daneshyari.com/article/6899127

Daneshyari.com