Swarm and Evolutionary Computation ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Swarm and Evolutionary Computation

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

Regular Paper

Fuzzy evolutionary cellular learning automata model for text summarization

Q1 Razieh Abbasi-ghalehtaki [a,*], Hassan Khotanlou [b], Mansour Esmaeilpour [c]

[a] Department of Computer Engineering, Hamedan Branch, Islamic Azad University, Hamedan, Iran
[b] Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran
[c] Department of Computer Engineering, Hamedan Branch, Islamic Azad University, Hamedan, Iran

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form
25 January 2016
Accepted 1 March 2016

Keywords:
Text summarization
Cellular learning automata
Artificial bee colony
Particle swarm optimization
Genetic algorithm
Fuzzy Logic

ABSTRACT

Text summarization is the automatic process of creating a short form of an original text. The main goal of an automatic text summarization system is production of a summary which satisfies the user's needs. In this paper, a new model for automatic text summarization is introduced which is based on fuzzy logic system, evolutionary algorithms and cellular learning automata. First, the most important features including word features, similarity measure, and the position and the length of a sentence are extracted. A linear combination of these features shows the importance of each sentence. To calculate similarity measure, a combined method based on artificial bee colony algorithm and cellular learning automata are used. In this method, joint n-grams among sentences are extracted by cellular learning automata and then an artificial bee colony algorithm classifies n-friends in order to extract data and optimize the similarity measure as fitness function. Moreover, a new approach is proposed to adjust the best weights of the text features using particle swarm optimization and genetic algorithm. This method discovers more important and less important text features and then assigns fair weights to them. At last, a fuzzy logic system is used to perform the final scoring. The results of the proposed approach were compared with the other methods including Msword, System19, System21, System28, System31, FSDH, FEOM, NetSum, CRF, SVM, DE, MA-SingleDocSum, Unified Rank and Manifold Ranking using ROUGE-l and ROUGE-2 measures on the DUC2002 dataset. The results show that proposed method outperforms the aforementioned methods.

1. Introduction

Q2 Nowadays owing to the large amount of data on the Web, it is necessary that users access the information in concise form without losing any beneficial information. Hence, automatic text summarization has attracted great attention in the last few years, considering that it is the best method to deal with huge amounts of data. The aim of automatic text summarization is extracting the most important content which an end user needs [42]. In this method an explanatory tree structure is constructed to show non-structural features with explanatory relations among different sentence segments of the documents. Some of the works use search engines for the generation of extractive summaries for a single document on web pages [52]. Other works on single documents can be mentioned E-learning for extracting the most important features [36,2] and allocating the labels to set web document clustering [14]. Machine learning approaches [21,49,68] are utilized for text summarization because of their high ability in text mining. In recent years, fuzzy logic is used in text summarization [10,34,37,60] and also, combination of the fuzzy logic with swarm intelligence [10] and cellular learning automata (CLA) [38] have a good result.

However, most of the researches used evolutionary algorithms in extractive summarization; [11,12,22,26,3,40,53,56,62]. Evolutionary algorithms with local search heuristics (i.e. mimetic algorithm) [43,48] and Genetic Algorithm (GA) are used as the tools of extracting sentences in text summarization. Moreover, they are used for customized weights of factors that give a score to each sentence in a document [22,26,40,53]. Particle Swarm Optimization (PSO) is used [33] for features selection problem in text summarization. [10,9]. Also, PSO has good applications in classification and data clustering [16,44]. Term frequency is used as an approach to identify important sentences along with reduction in information redundancy [20,47,58].

Recent applications of natural language processing underline a need for an effective method to compute the similarity between sentences of a document [45]. An example can be a conversational agent system with script strategies [5]. In text mining, sentence

* Corresponding author.
E-mail address: r.razieabbasi.a@gmail.com (R. Abbasi-ghalehtaki).

http://dx.doi.org/10.1016/j.swevo.2016.03.004
2210-6502/

similarity is used as a measure to discover unseen knowledge from textual databases [6]. In addition, the incorporation of short-text similarity is beneficial to applications such as text summarization [18], text categorization [35] and machine translation [69]. These exemplar shows that the computing sentence similarity has become a generic element of discovering text-related knowledge representation. As the size of digital information grows exponentially, large volumes of raw data need to be extracted. Nowadays, there are several methods to employ data according to end user's needs. The most common method is Data Mining (DM). DM is used for extracting useful knowledge from large amount of data [61]. The extracted knowledge must be accurate, easy and readable. The algorithms which are used in this case can be swarm based approaches like PSO (Kennedy & Eberhart, 2007) and Artificial Bee Colony (ABC) [51]. According to previous studies, fuzzy logic system, PSO, GA, ABC and CLA can be having a good performance in text summarization. Proposed method, called FPGAC, takes advantage of all of them. It uses feature selection to extract the most important features, CLA to bring out joint n-grams in short time, ABC is used to classify n-friends and optimize similarity measure, PSO-GA to allocate the fair weights to the text features and fuzzy logic system is utilized to score sentences and extract the summary. The rest of this paper is organized as follows: Section 2 introduces related works. Section 3 presents proposed method. Section 4 describes the experimental design. Section 5 presents the experimental results and evaluation. Finally, Section 7 draws a conclusion.

2. Related work

Most of the researches have used sentences scoring [41] to extract the most important sentences. There are three methods used for sentence scoring: (i) word scoring, (ii) sentences scoring, and (iii) graph scoring. In word scoring, scores are allocated to the words as their features and the weights of each sentence are obtained from the sum of its words scores. Some of the word features used in most papers are: numerical data [2,22], proper noun [22], title words [17], thematic words and keyword. Different searches use different techniques to determine the importance of the words such as word frequency [2] and TF/IDF (Sekine et al., 2001; Murdock, 2006).

In sentences scoring, sentence position [2,7,22] and sentence length [22] are two important features in a sentence. Relationship and connection among the sentences are determined in graph scoring. Text rank is a graphical model for extracting main key words [7,46]. The bushy path of the node model and aggregate similarity is another graph scoring model in which each sentence is considered as a node and a link is assigned between them to the related sentences [22].

Binwahlan et. al. [10] proposed a model based on PSO that extracts text features such as cohesion, readability and relationship with the title [56]. PSO is used to adjust the fair weights for text features and separating more and less important features. Harmony search is also used to extract sentences.

In another work, Binwahlan et al. [12] proposed a fuzzy swarm diversity hybrid model for text summarization which includes three models: (i) Maximal Marginal Importance (MMI) diversity, (ii) swarm diversity based method, and (iii) fuzzy swarm based method. In the first method, sentences are sorted in a binary tree according to their scores by diversity-based and MMI is used to select the sentences to be included in the summary. The second method which is based on PSO is used to adjust weight to the features based on their importance and MMI and diversity methods are similar to the first method. In the third method, the fuzzy algorithm is used to calculate sentence score and the input of the fuzzy system is weights that are found by PSO. In each summarizer, sentences are ranked according to their scores and then the n best sentences are extracted from each summarizer .At last, a weight is allocated to each summarizer sentence and the n best sentences are extracted. Kiyoumarsi and their coworkers (Kiyoumarsi et al. 2011) proposed a text summarization method based on Cellular Automata (CA). This method is based on CA, GA and fuzzy logic. Three methods of text summarizer are examined (i) text summarization based on the GA approach, (ii) text summarization based on fuzzy logic approach, and (iii) text summarization based on CA approach. Also, this paper examines the influence of features on the summarization and then trains features by CA, GA and fuzzy. Ramiz and Aliguliye [54] proposed a new sentence similarity measure and sentence based extractive technique for automatic text summarization. This method is based on clustering and extraction of sentences. Sentences are clustered, and then on each cluster representative sentences are defined. Also this paper demonstrated that the summarization result depends on the similarity measure.

This paper is organized similar to the other models of extractive summaries in a single document. But the differences of the proposed model from these models are expert rules for scoring to the words, calculating similarity by CLA-ABC and using PSO-GA for weighting the sentences.

3. Proposed method

Propose method contains the four main parts as Fig.1: (i) Preprocessing text, (ii) Extract text feature, (iii) Weighting to the extractive features by Particle Swarm Optimization and Genetic Algorithm and (iv) Scoring to sentences by fuzzy system. In extracting text feature step, features are divided in two groups: word features and sentence features. Word features contain six features: title word, thematic word, keyword, proper noun, numerical data and term weigh features. Sentence features include four features: word, position, length and similarity features. In the following statement, text features are extracted as Section 3.1. Moreover, Similarity features are calculated by Cellular Learning Automata and Artificial Bee Colony (CLA-ABC) as Section 3.2. After that a new method based on Particle Swarm Optimization and Genetic Algorithm (PSO-GA) is proposed to assign suitable weights to the text feature as Section 3.3. At last, Section 3.4 presents Fuzzy PSO-GA CLA-ABC based text summarization.

3.1. Text features

In this section, the most important text features are extracted from the original preprocessed text. Some features help us to extract the rich sentences and reduce redundancy. These features include word features, sentence position, sentence length and similarity feature. The linear combination of these features shows the importance of each sentence. These features are used as follows:

3.1.1. Word feature

Sentences are made of words, so the score of words of a sentence is an important factor to decide sentence importance in the document. Feature score for each word is calculated as follows:

- Title word

Sentences containing title words indicate the subject of the document. These sentences have a greater chance to be included in the summary, so title words should take a higher score. Then,

Please cite this article as: R. Abbasi-ghalehtaki, et al., Fuzzy evolutionary cellular learning automata model for text summarization, Swarm and Evolutionary Computation (2016), http://dx.doi.org/10.1016/j.swevo.2016.03.004