Text Summarization Using Cuckoo Search Optimization Algorithm

Seyed Hossein Mirshojaei*, Behrooz Masoomi

Department of Computer Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Abstract

Today, with rapid growth of the World Wide Web and creation of Internet sites and online text resources, text summarization issue is highly attended by various researchers. Extractive-based text summarization is an important summarization method which is included of selecting the top representative sentences from the input document. When, we are facing into large data volume documents, the extractive-based text summarization seems to be an unsolvable problem. Therefore, to deal with such problems, meta-heuristic techniques are applied as a solution. In this paper, we used Cuckoo Search Optimization Algorithm (CSOA) to improve performance of extractive-based summarization method. The proposed approach is examined on Doc. 2002 standard documents and analyzed by Rouge evaluation software. The obtained results indicate better performance of proposed method compared with other similar techniques.

Keywords: Text summarization, Extractive method, Cuckoo Search Optimization algorithm.

1. Introduction

With the growing world of information and widespread access to the Internet, sites and online text resources are developed more than ever. Investigating and studying of the required contents force us to focus on text summarization. It is a proper way that enables users to overview all related text with their favorite issues and help them to make next decisions.

Text Summarization methods can be classified into extractive and abstractive summarization (Hovy, 2005). However, another classification method consists of summarization based on the number of input documents, which is including of multi-document and single-document summarization. Classification based on the purpose of summarization is another type which is involving of public query methods (Gupta, 2005). Extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form, which is improved in this study. Abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

* Corresponding author. Email: mirshojaee.h@gmail.com

Several methods have been proposed for extractive text summarization, which one of the main of them is Term-Frequency Inverse Document-Frequency (TFIDF) technique (Ledeneva, 2008 , Garsia, 2009).In this method, weighting is done based on the words frequency and inverse iteration of sentences. The term 'sentences iteration' is called to number of sentences in the document which are including of those words. Although this method is simple but, it may lead to a deviation in summarization process due to iteration of some unimportant words.

Documents are usually written and organized so that different subjects brought one after another. These subjects are written in the form of sentences, explicitly or implicitly. This type of organization must be applied for summarized structure. Using of text summarization based on sentences clustering methods are another common ways (Zang, 2009). In clustering, each cluster represents a topic. Here, the similarities between sentences are investigated according to a set of parameters and then, similar phrases are placed in a cluster. Finally, in each cluster sentences that were most similar to the cluster topic and earned high scores could be selected for summarizing. The main advantage of this method is that the topic of each text could be well recognized but, since the number of clusters is important and it might to be too much or too low, the results of summarization could be affected. In the other words, selection of optimal number of clusters is a difficult activity, which is considered the disadvantages of this method.

In other extractive summarization method, the data sets which are tagged by a man is used as a tool for summarization. In the other words, we have the same set of input text and its summary text. This method, first, sentences are broken into segments by a special cue markers. Each segment is represented by a set of predefined features (e.g. location of the segment, number of iteration and title words in the segment) (Chaung, 2000). Then, a supervised learning method are used to train the summarizer to extract important sentence segments, based on the feature vector. Some of these methods are: decision tree, Bayes theorem, neural networks, and fuzzy logic (Song, 2011, Suanmali, 2009). Decreasing the precision and speed of operation in big documents are the main disadvantages of these methods. The main problem of fuzzy logic is that if the rules are not properly defined, precision comes down. Meta-heuristic methods are different types of summarization, which the main purpose of them is to find high scored sentences. Genetic algorithms (Fattah, 2009), Particle Swarm Optimization (PSO) (Oi-Mean, 2011), and Bacterial Foraging Optimization Algorithm (BFOA) (Asgari, 2013) are the most known of meta-heuristic approaches. Stopping in local minimum or maximum and obtaining the wrong results is the common disadvantage of these methods.

The main challenge of extractive- based summarization approaches are the high volume of documents' information and large search space, which become the problem unsolvable. In large texts that are composed of a large number of words, ratings and more importantly, selecting of sentences are very difficult and yield to low precision and speed. In these situations, use of meta-heuristic approaches can be helpful in solving the problem and achieving to optimum and effective solution. Since, optimization methods stop in local maximum or minimum, the cuckoo search optimization method is adopted in this paper. The results indicate better performance of proposed method compare with the pervious techniques.

The rest of paper is organized as follows: definition of basic concepts is presented in section 2. Our proposed methodology is described in section 3 and finally, the implemented experiments and comparison of the results with other similar methods are concluded in section 4.

2. Basic Concepts and Definitions

Text summarization is the technique which automatically creates an abstract or summary of a text for specific users (Uddin, 2007). Extractive

summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. The most important advantage of this method is: simplicity, high speed of summarization process, and above all, reduction of users' study time. However, this method also has some disadvantages like: the length of extracted sentences may be too short or medium size, as well as relevant and important information may be broadcast between other sentences and extractive method cannot detect them.

The overall architecture of extractive summarization process is made of two pre-processing and processing phases. In the pre-process phase, the end of sentences is identified, words that have no meaning are removed, and words' stemming is done. Effects and relationship of sentences with the main topic are identified in process phase. Then, a weight is assigned to each of them and finally, sentences with the highest scores are selected for ultimate summarized text.

2.1. Cuckoo Search Optimization Algorithm

Cuckoo search is a meta-heuristic optimization method that proposed by Yang and Deb (2009). This method has an evolutionary approach to search optimal solution. Cuckoos have a belligerent reproduction tactic that involves the female laying her fertilized eggs in the nest of another species so that the surrogate parents unwittingly raise her brood. Sometimes the cuckoo's egg in the nest is revealed and the surrogate parents throw it out or dump the nest and start their own brood elsewhere (Samiksha, 2011). The cuckoo search optimization algorithm considered various design parameters and constraints, the three main idealized rules on which it is based are as follows:

- Each cuckoo lays one egg at a time, and dumps its egg in randomly chosen nest;
- The best nests with high quality of eggs will carry over to the next generations;
- The number of available host nests is fixed, and the egg laid by a cuckoo is discovered by the host bird, most probably. In this case, the host bird can either throw the egg away or abandon the nest, and build a completely new nest. For simplicity, the number of nests can be replaced by new nests (solutions). In addition, each nest can represent a set of solutions;

The CS technique has been demonstrated successfully on some benchmark functions and its precision and success are far better, than other approaches like PSO.

3. Proposed Method

In our proposed method, a pre-processing must be first done on the input text. Then, based on the TFIDF weighting method a weight is assigned to extracted sentences in pre-processing phase. The weighting process is done by applying *Eq.(1)*and *Eq.(2)*as follows:

$$tf_{i,j} = \frac{freq_{i,j}}{max_l freq_{l,j}} \qquad (1)$$

$$idf_i = log\frac{N}{n_i} \qquad (2)$$

Where,$freq_{i,j}$ is number of frequency of i^{th} word in sentence j^{th},$max_l freq_{l,j}$ is maximum number of frequency of i^{th} word in sentence j^{th}, N is total number of sentences, and $n_i$is number of sentences in which word *i* occurs.

After weighting sentences, it is necessary to obtain the similarity matrix by *Eq.(3)*. This matrix is used to calculate available similarities between sentences and keywords.

$$sim(s_i, q) = \frac{\sum_{i=1}^{t} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \qquad (3)$$

Where, *sim* (s_i,q) is sentence similarity matrix, $w_{i,q}$ and $w_{i,j}$ are key words and title weight, and the weight of each word, respectively.

Now, summarized and important sentences should be extracted from main text by CSOA. The proposed method has a series of parameters that must be initialized to start. For applying this method, number of birds and iteration loop are assumed to be 50 and 100, respectively. Then, total number of sentences, number of summarized sentences and similarity matrix are considered as input parameters of proposed method. After the procedure is completed, high scored sentences are selected and displayed as summarized text. Overall steps of CSOA are described as follows:

Step1- initialization of CSOA's parameters;

Step 2- random assignment of sentences to birds;

Step3-birds assessment based on the cost function;

Step4- updating birds' position;

Step 5- If the end condition of the loop is satisfied, then the algorithm is finished. Otherwise, return to step 3.

The proposed method can be presented in the form of a pseudo-code as follows:

Objective function: $f(\mathbf{x}),\ \mathbf{x}=(x_1,x_2,\ldots,x_d)$;

Generate an initial population of *n* host nests;

While (t<Max Generation) or (stop criterion)

Get a cuckoo randomly (say, i) and replace its solution by performing Lévy flights;

Evaluate its quality/fitness F_i

[For maximization, $F_i \propto f(\mathbf{x}_i)$];

Choose a nest among n (say, j) randomly;

if ($F_i > F_j$),

Replace j by the new solution;

end if

A fraction (p_a) of the worse nests are abandoned and new ones are built;

Keep the best solutions/nests;

Rank the solutions/nests and find the current best;

Pass the current best solutions to the next generation;

end while

Cost function is calculated by *Eqs*. *(4)*and*(5)*, which compute sentences coherence and readability of summarized sentences, respectively. Coherence factor makes summarized sentences to talk about the same information and readability factor indicates that, they are relate to each other with a high degree of similarity.

$$CF_s = \frac{\log(C * 9 + 1)}{\log(M * 9 + 1)} \tag{4}$$

$$C_s = \frac{\sum_{\forall s_i, s_j \in Summary\ subgraph} W(s_i s_j)}{N_s}$$

$$RF_s = \frac{R_s}{max_{\forall i} R_i} \quad , \quad R_s = \sum_{0 \le i < s} W(s_i, s_{i+1}) \tag{5}$$

Where, CF_S is coherence factor of sentences, C_s is average similarity of available sentences in summarization, and M is maximum weight of sentences. In *Eq.(5)*,RF_s represents the readability factor of a summary with length of *s*.

4. Simulation and Evaluation of Results

Four following criteria are used to evaluate summarizers in extractive methods:

1. *Evaluation based on the text quality*: this is an evaluation process that takes place on the basis of human's judgments. Here, the summarized text is scored according to a set of pre-defined measures;
2. *Evaluation based on the selection:* this process is done on the basis of matching sentences together;
3. *Evaluation based on the content:* this process is done on the basis of matching words together;
4. *Evaluation based on the task:* here, measuring quality of summarized text is done by a series of requested requirements.

In proposed method, simulation is done based on the second criteria (matching sentences together), by MATLAB software environment. Here, the extracted summary from original text is compared with ideal summary based on the precision, recall, and F-score

criteria (Abuobeieda, 20313). These criteria are defined in Eq. (5)-(7), respectively.

Precision means common between summarized extracted and ideal sentences, divided by all extracted sentences. Recall criteria means common between relevant and retrieved sentences, divided by all relevant sentences. F-score is a statistical criteria which is a combination of precision and recall criteria and determines the score of final selected sentences in produced summary.

$$Precision = \frac{Relevant\ Sentences \cap Retrieved\ Sentences}{Retrieved\ Sentences} \quad (6)$$

$$Recall = \frac{Relevant\ Sentences \cap Retrieved\ Sentences}{Relevant\ Sentences} \quad (7)$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

Since, the calculation of these criteria are relatively difficult and time consuming, the Rouge software is adopted to perform automatic computations. This software is designed according to Perl programming language and equipped with several packages for evaluating summarized texts. To evaluate proposed method, documents of Doc. 2002 standard such as d105, d070f, d067f, and d061j are used and respective results are examined through Rouge. The obtained results of comprising proposed method with other similar techniques like PSO, MS Word, and BFOA are summarized in Table 1.

Table 1

Methods comparison based on the F-score results.

DOC	*CSOA*	*PSO*	*MS WORD*	*BFOA*
d061j	0.49761	0.42869	0.41201	0.43543
d067f	0.46476	0.44637	0.36625	0.44126
d070f	0.47126	0.40616	0.38179	0.41765
d105g	0.42391	0.39517	0.32773	0.39121

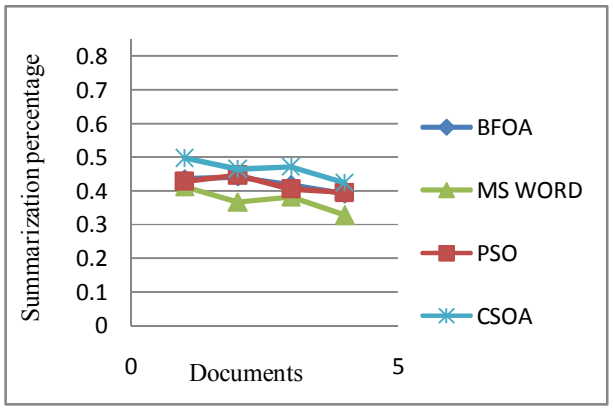

Fig. 1. Comparison of BFOA, PSO, MS Word, and CSOA methods

5. Conclusion

This paper described the extractive-based text summarization methods and discussed about their advantages and respective disadvantages. Here, cuckoo search optimization algorithm was used to summarize texts on the basis of an extractive way. Then, scoring procedure and method of selecting summarized text were described in details. Finally, the proposed method examined based on the Do. 2002 standard through Rouge evaluation software. Analyzing of obtained results and comprising them with other similar methods indicated reliability and better performance of our proposed approach.

References

[1] A. Abuobieda, N. Salim, Y. Kumar, A. Osman, "An Improved Evolutionary Algorithm for Extractive Text Summarization", in Intelligent Information and Database Systems. vol. 7803, pp. 78-89, 2013.

[2] H. Asgari, B. Masoumi, "Provide a method to improve the performance of text summarization using bacterial foraging optimization algorithm", the seventh iran data minig conference, Dec.10 2013.

[3] W. Chuang, J. Yang, "Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms", in Knowledge Discovery and Data Mining. Current Issues and New Applications.vol.1805, pp. 454-457, 2000.

[4] D. Das, A. Martins, "A Survey on Automatic Text Summarization", Language Technologies Institute Carnegie Mellon University, November 2007.

[5] A. Fattah, F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization", Computer Speech & Language, vol. 23, pp. 126-144, 2009.

[6] R. Garcia-Hernandez, Y. Ledeneva, "Word Sequence Models for Single Text Summarization", in Advances in Computer-Human Interactions, pp. 44-48, 2009.

[7] V. Gupta, "A Survey of Text Summarization Extractive Techniques ", Journal of Emerging Technologies in web Intelligence, vol. 2, August 2010.

[8] E. Hovy, Text summarization, chapter The Oxford Handbook of Computational Linguistics, 2005.

[9] Y. Ledeneva, A. Gelbukh, R. Hernández, "Terms derived from frequent sequences for extractive text summarization", in Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, pp. 593-604, 2008.

[10] F. Oi-Mean, A. Oxley, "A hybrid PSO model in Extractive Text Summarizer", in Computers & Informatics (ISCI), 2011 IEEE Symposium on, pp. 130-134, 2011.

[11] S. Punam, "Cuckoo Search Clustering Algorithm: A novel strategy of biomimicry", World Congress on Information and Communication Technologies, 2011.

[12] W. Song, L. Cheon Choi, S. Cheol Park, X. Feng Ding, "Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization", Expert Systems with Applications, vol. 38, pp. 9112-9121, 2011.

[13] L. Suanmali, N. Salim, M. Salem Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security (IJCSIS), vol. 2, Jun 2009.

[14] N. Uddin, S. Khan, "A study on text summarization techniques and implement few of them for Bangla language", in Computer and information technology, pp. 1-4, 2007.

[15] P. Zhang, C. Li,"Automatic text summarization based on sentences clustering and extraction", in Computer Science and Information Technology 2nd IEEE International Conference on, pp. 167-170, 2009.