A Method for Text Summarization by Bacterial Foraging Optimization Algorithm

Morteza Dastkhosh Nikoo[1], Ahmad Faraahi[2], Seyyed Mohsen Hashemi[3], Seyyed Hossein Erfani[4]

[1] **Computer Engineering Department, Science and Research Branch, Islamic Azad University**
Tehran, Iran

[2] **Computer Engineering and Information Technology Department, Payam Noor University**
Tehran, Iran

[3] **Computer Engineering Department, Science and Research Branch, Islamic Azad University**
Tehran, Iran

[4] **Computer Engineering Department, Science and Research Branch, Islamic Azad University**
Tehran, Iran

Abstract

Due to rapid and increasingly growth of electronic texts and documents, we need some techniques for integration, communication and appropriate utilization of these texts. Summarizing the literature is one of the most fundamental tasks for integrating and taking advantages of these gathered texts. Selecting key words and then integrating them as a summary text, is the most common method in text summarization.

In this paper we present a new method of automatic text summarization, with bacterial foraging optimization. The main idea of this method, is weighting words, then valuing the sentences, and finally extracting key sentences from the text, as the summarized text. It should be noted that, here we used the weighting term TF-IDF method, to determine weight for each text. Also, the bacterial foraging optimization used to converge the solutions is obtained from each bacteria, and finally the best candidate summarized text is given.

Keywords: *Text Summarization, Word, Sentence, Weight of sentence, Bacterial foraging optimization algorithm.*

1. Introduction

Summary document is a presentation of the contents of a compressed text. A summary of the text refers to the views on the important points that must contain the key phrases in the text.

The summary should not duplicate the text, and should be as concise as possible. Some words and phrases may be repeated several times in a text, but summarized text should be as pressed as possible.

The summary text should refer to important parts of the text. Although different versions of summarization of a given text which is done by several people, may differ from each other because of their point of views, but if it is done properly, the summarized text will involves original text contents and titles; so the summarization done by computers, should be involve important sentences of original text.

In general Edward Hovy, defines the summary as follows: the summary is a text of one (or more) document has been prepared and includes the most important information of primary document (or documents) and also is not more than the half size of the original document (documents). [1]

Summary text can be in any format. Furthermore, considering the size of summary, it is comprehended that summary of the original text, even if contains a brief or key and referring words, should be a set of some more important words in the text, so that the overall content of the original text can be realized.

Another batch of summarized texts is the extractive summarized literatures. An extractive summary is a summary which its sentences, phrases and words are the same as the original text; they are extracted without any changes and used in the summary.

In order to summarize, summary extractive instruments must measure the weight of sentences and words, and then based on the matter of importance, decide on whether any of the sentences must be in the summary or not. [2]

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

There are generally two types of summarization, Single and multiple document summarization. Single text summarization, create a text summary of a single document, While in multiple text summarization, text summary is created of the relevant texts in several documents. [3]

Today, machine learning methods [4, 5, 6, 7, 8, 9, 10] are used for solving different problems and also text summarization. One of the methods that have recently been used in text summarization is the PSO algorithm. [11]

Another method that regarded today as swarm intelligence is bacterial foraging optimization algorithm. This method has attracted a lot of attention in the implementation of algorithms and computational modeling, and industrial systems.

Recently some models have been created to mimic the behavior of swarm, to solve some minor problems [12, 13, 14]. BFO has been successfully implemented on many engineering problems such as optimal control [15], consistent estimates [16], reducing the transmission loss [17] and machine learning.

Sentence classification in multiple text summarization [18] and applications of text summarization in image processing field [19] as well are some of the works which have been done in text summarization recently.

The rest of the paper is organized as follows: Section 2, introduces the bacterial foraging optimization algorithm (BFO). In section 3, identifying key words from the less important words, and ultimately identifying important terms will be described. In Section 4, the bit mapping words, to solve the problem of text summarization will be shown. In Section 5, the problem of text summarization, by BFO are explained. In Section 6, Rouge-N method is introduced for evaluating the summarized literature. In Section 7, results are shown and finally in Section 8, conclusions and future work will be explained.

2. Bacterial Foraging Optimization Algorithm

Bacterial foraging optimization algorithm (BFO) is an optimization algorithm that acts based on the social behavior of E.coli bacteria in the body.

Coordination and switching between two states, swim and tumble, enables the bacteria to run and orientation, for looking for food to survive. [20]
In general, BFO algorithm has three stages which are defined as follows: [20]

Chemotaxis: This step represents the movement of bacteria and is calculated as Eq. (1).

$$\theta^{i}(j+1,k,l) = \theta^{i}(j,k,l) + C(i)\frac{\Delta(i)}{\sqrt{\Delta^{T}(\acute{\iota})\Delta(\iota)}} \qquad (1)$$

Words θ^{i} (j, k, l), represents the ith bacterium in jth chemotaxis, kth reproduction, and lth elimination-dispersal step, and C(i) is the size of chemotaxis.

Proliferation: Health status of each bacterium in his life will be considered as all appropriate steps that it can be formulated as Eq. (2). [20]

$$\sum_{j=1}^{N_c} J(i,j,k,l) \qquad (2)$$

Here N_c is the maximum number of chemotaxis's steps. All bacteria are ordering descending based on health status. According to The Proliferation step, only the upper half of the list of bacteria can survive. Then each of the surviving bacteria has to multiply the two bacteria and Placed in the same place. However, the number of bacteria remained constant.

Elimination-dispersal: According to the possibility rules and considering the different positions of bacteria, It is possible that bacteria sticks in the first place and unable to navigates the entire search space; In this case, by using the elimination-dispersal step, these bacteria can be removed from the cycle of searching, or distributed their accumulation in the area. The elimination-dispersal of bacteria can be regular or random. [20]

3. Weighting the Words and Sentences

To summarize the text and identify important sentences from less important ones, we need a method for rating the text.
With recognizing that every word in a text is how important and useful, some values can be given to words (based on its position in the text). When the value of each word in the text determined, each sentence's value of the text can be easily identified. However, recognizing the important sentences from useless sentences is easy.

To achieve this goal, a system should be used to weighting the words of the text.
In this paper we used the TF-IDF method, for weighting each word in the text. [3]

$$G(t_{ij}) = Log\,(N / n_j) + 1 \qquad (3)$$

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

In the above equation, N is the total number of sentences of the text, and n_j is the number of sentences that has the word j.

Sentences are made of words; to calculate the weight of each sentence, we need the weight of each term and sum of the weights of a sentence. So, more important sentences can be recognized among the less important sentences.

4. Bit-Mapping the Words

In this report, we used a bit string to use and display the words. Each bit that corresponds a word in the text, can take only zero or one.

The first bit represents the first term, the second bit represents the second term, and the process continues as the same until the last word. When a bit sets to one, it means the choice of corresponding word, and zero value, indicates the un choice of corresponding word in a candidate summary.

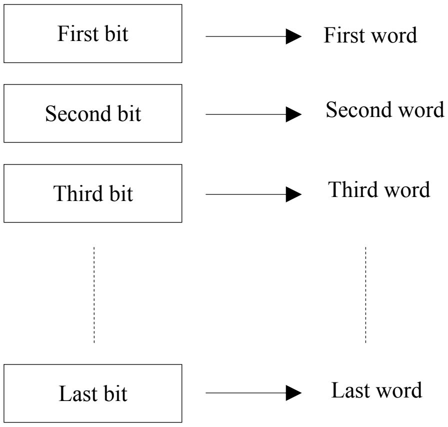

Fig. 1 Bit-mapping the words.

5. Text Summarizing with BFO

In this paper, the BFO algorithm has been used to summarize the text. First, the words and sentences are recognized in the text. Then according to Eq. (1), the words are given a weight. The weight of sentences is determined by summing the weight of the words in the sentences. Weight of each summary is evaluated by DUC. [21]

The sentences sorted in descending order on the basis of weight, and the first N sentences (usually 20% of the original sentence), considered as a candidate summary text. Next, according to the bit-mapping the terms that described in Section 4, each bacteria selects the words and phrases to make a candidate summary text. Finally, according to Rouge-N equation, the value of each candidate summary text determined. Summary Process of the bacteria continues until the value of bacteria converging to the threshold value. Figure 2 shows this trend.

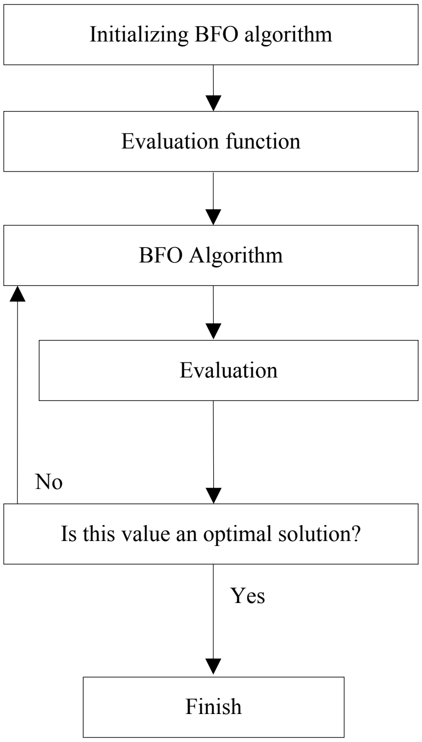
Initializing BFO algorithm
Evaluation function
BFO Algorithm
Evaluation
No
Is this value an optimal solution?
Yes
Finish

Fig. 2 General trend of text summarization using BFO.

6. Evaluation Function

In this report, the evaluation function Rouge-1 that calculated as Eq. (4), is used to evaluate each candidate summary text. [22]

$$Rouge - N = \frac{\sum_{S\in\{Refrence\ Summary\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Refrence\ Summary\}}\sum_{gram_n\in S} Count(gram_n)} \quad (4)$$

Where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

7. BFO Text Summarization Results

In this problem, the optimal value for each parameter of BFO, displays as the following diagrams.

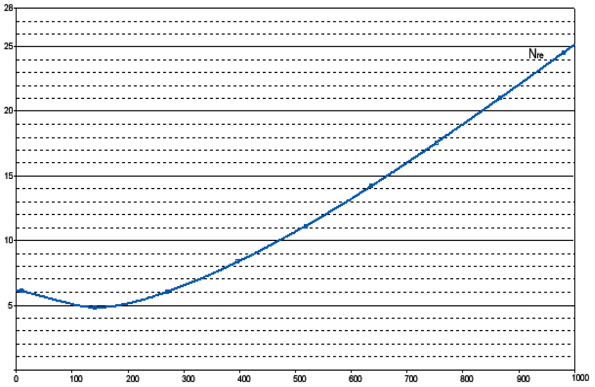

Fig. 3 Result of changing the N_{re} parameter.

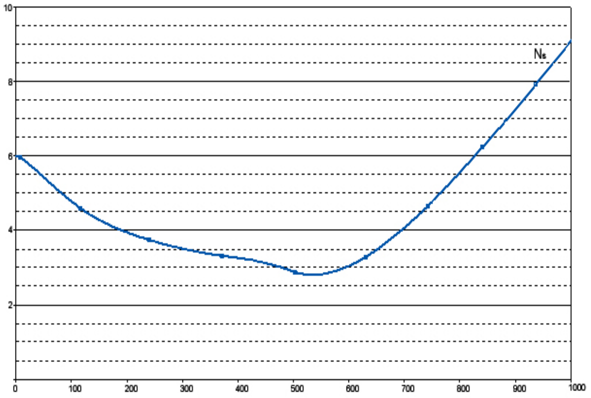

Fig. 4 Result of changing the N_s parameter.

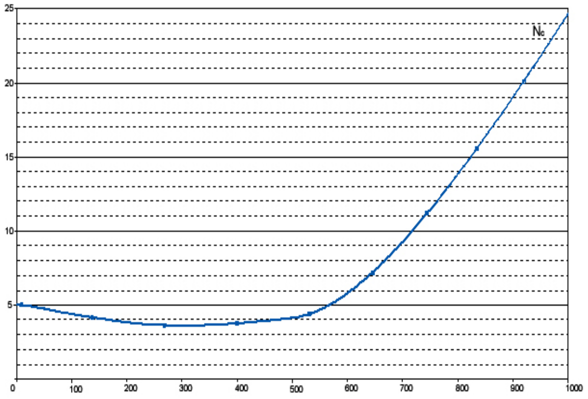

Fig. 5 Result of changing the N_c parameter.

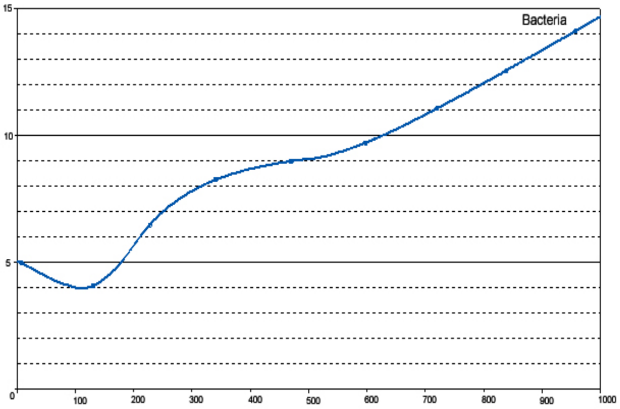

Fig. 6: Result of changing the bacteria parameter.

In figures, a threshold value for each input parameter can be seen.

8. Conclusions

We introduced a model to summarize the text, by using the bacterial foraging optimization algorithm. This model works based on scoring the words and sentences. Each component which is a bacteria, attempts to summarize the text and improves its position each time. Summary Process of the bacteria, continues while the value of bacteria converging to the threshold value.

As future works, this model can be used for summarizing documents in other languages, other features of text can also be used to summarize the texts.

References

[1] E. Hovy, “Automated Text Summarization,” *chapter The Oxford Handbook of Computational Linguistics*, pp. 583–598, 2005.

[2] H. Jing, “Sentence Reduction for Automatic Text Summarization,” *In Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, USA, pp. 310–315, 2000.

[3] G. Ercan, “Automated text summarization and keyphrase extraction,” *Master thesis*, 2006.

[4] J. Kupiec, J. Pedersen, and F. Chen, “A trainable document summarizer,” *In Proceedings of the ACM. SIGIR conference*. New York, USA, pp. 68-73, July 1995.

[5] C. Y. Lin, and E. Hovy, “Identifying topics by Position,” *In Proceedings of the Fifth conference on Applied natural language processing*, San Francisco, CA, USA, pp. 283-290, 1997.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

[6] C. Y. Lin, “Training a selection function for Extraction,” *In Proceedings of the Eighteenth Annual International ACM Conference on Information and Knowledge Management (CIKM). 2-6 Nov*. Kansas City, Kansas, pp. 55-62, 1999.

[7] J. M. Conroy, and D. P. O'leary, “Text summarization via hidden markov models,” *Proceedings of SIGIR '01*. New Orleans, Louisiana, USA, pp. 406-407, 9-12 September 2001.

[8] M. Osborne, “Using maximum entropy for sentence extraction,” *Proceedings of the ACL'02 Workshop on Automatic Summarization*. Morristown, NJ, USA, pp. 1-8, July 2002.

[9] K. Svore, L. Vanderwende, and C. Burges, “Enhancing single document summarization by combining Rank Net and third-party sources,” *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague: Association for Computational Linguistics*, pp. 448–457, June 2007.

[10] M. A. Fattah, and F. Ren, “GA, MR, FFNN, PNN and GMM based models for automatic text summarization,” *Computer Speech and Language*, pp. 126-144, 2008.

[11] M. S. Binwahlan, N. Salim, L. Suanmali, “Swarm Based Text Summarization,” *Spring Conference - International Association of Computer Science and Information Technology*, 2009.

[12] H. J. Bremermann and R.W. Anderson, “An alternative to back-propagation: a simple rule of synaptic modification for neural net training and memory,” *Tech. Rep. PAM-483, Center for Pure and Applied Mathematics*, University of California, San Diego, Calif, USA, 1990.

[13] S. Mueller, J. Marchetto, S. Airaghi, and P. Koumoutsakos, “Optimization based on bacterial Chemotaxis,” *IEEE Transactions on volutionary Computation*, vol. 6, no. 1, pp. 16–29, 2002.

[14] K. M. Passino, “Biomimicry of bacterial foraging for distributed optimization and control,” *IEEE Control SystemsMagazine*, vol. 22, pp. 52–67, 2002.

[15] D. H. Kim and J. H. Cho, “Adaptive tuning of PID controller for multivariable system using bacterial foraging based optimization,” in *Proceedings of the 3rd International Atlantic Web Intelligence Conference (AWIC ’05)*, vol. 3528 of *Lecture Notes in Computer Science*, pp. 231–235, Lodz, Poland, June 2005.

[16] S. Mishra, “A hybrid least square-fuzzy bacterial foraging strategy for harmonic estimation,” *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 1, pp. 61–73, 2005.

[17] M. Tripathy, S. Mishra, L. L. Lai, and Q. P. Zhang, “Transmission loss reduction based on FACTS and bacteria foraging algorithm,” in *Proceedings of the Parallel Problem Solving from Nature (PPSN ’06)*, pp. 222–231, Reykjavik, Iceland, September 2006.

[18] R. Barzilay, N. Elhadad, “Inferring Strategies for Sentence Ordering in Multi document News Summarization,” *Journal Of Artificial Intelligence Research, Volume 17*, pp. 35-55, 2002-2011.

[19] L.Plaza, E.Lloret, A.Aker, “Improving Automatic Image Captioning Using Text Summarization Techniques,” *Springer-Verlag Berlin Heidelberg* 2010.

[20] J. Adler, “Chemotaxis in bacteria,” *Science, vol. 153*, pp. 708–716, 1966–2011.

[21] The Document Understanding Conference (DUC). http://duc.nist.gov.

[22] C. Lin, “Rouge: a package for automatic evaluation of summaries,” *Proceedings of the Workshop on Text Summarization Branches Out, 42nd Annual Meeting of the Association for Computational Linguistics*. pp. 25–26, Barcelona, Spain. July 2004.

Morteza Dastkhosh Nikoo
High School Diploma in Mathematics and Physics, Mofatteh High School, Tehran, Iran, 2004.
B.S. in Computer Software Engineering, University of Allameh Mohaddes, Mazandaran, Iran, 2009.
M.S. in Computer Software Engineering, Science and Research Branch, IAU University, Tehran, Iran, 2012.

Ahmad Faraahi
High School Diploma in Mathematics and Physics, Kharazmi High School, Tehran, Iran.
B.S. in Computer Science, University of Sharif, Tehran, Iran.
M.S. in Computer Software Engineering and Information Technology, Dundee University, Dundee, United Kingdom.
Ph.D. in Computer Software Engineering and Information Technology, Bradford University, Bradford, United Kingdom.

Seyyed Mohsen Hashemi
B.S. in Computer Science (Software Engineering), Teacher Training University, Tehran, Iran, 2000.
M.S. in Computer Science (Software Engineering), Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, 2003.
Ph.D. in Computer Science (Software Engineering), Science and Research Branch, IAU University, Tehran, Iran, 2009.

Seyyed Hossein Erfani
High School Diploma in Mathematics and Physics, Dr.Hesabi High School, Tehran, Iran, 2002.
B.S. in Computer Software Engineering, Science and Research Branch, IAU University, Tehran, Iran, 2007.
M.S. in Computer Software Engineering, Science and Research Branch, IAU University, Tehran, Iran, 2010.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.