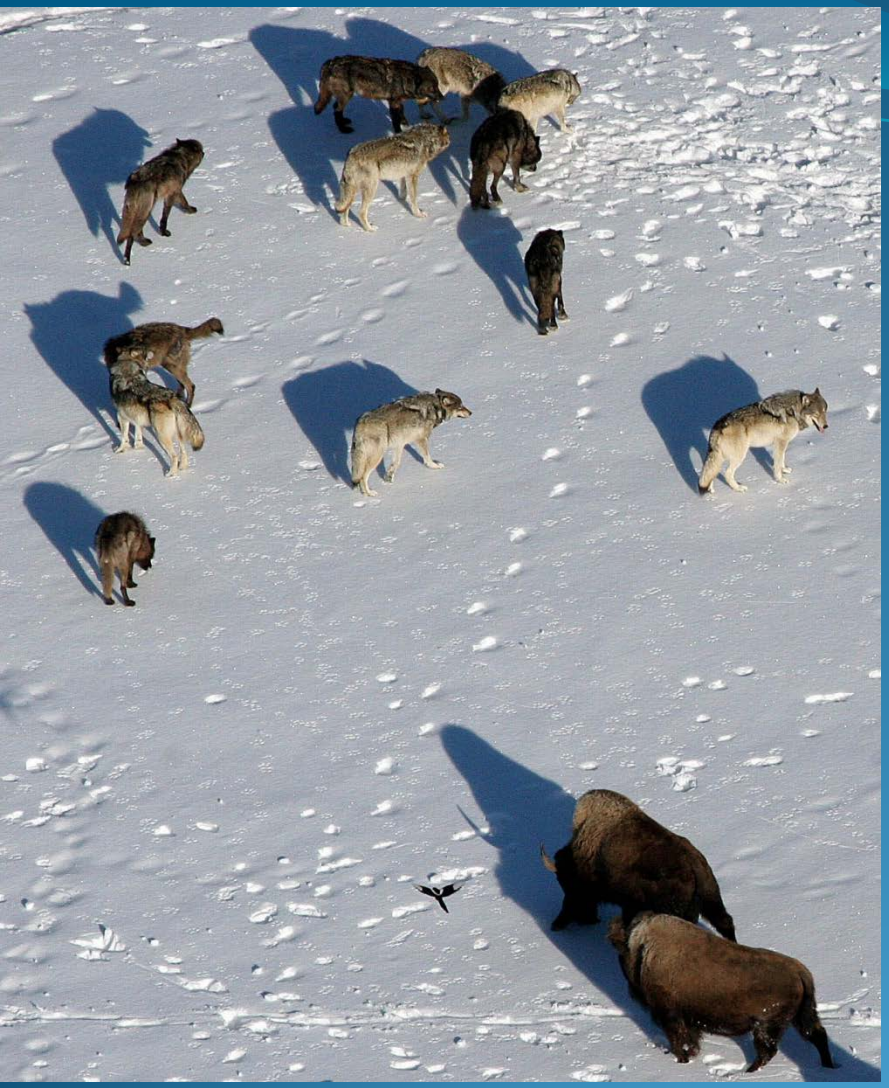


Extractive Text Summarization with Grey Wolf Optimization Algorithm

Ebru DUDAK, Pakize ERDOGMUS
Duzce University, Computer Engineering

Overview

- Today, text summarizing techniques play a major role in extracting relevant information from big data. The present article tests the success of single text summarization of the intuitive Gray Wolf Optimization Algorithm (GWO).
- The results were measured with the ROUGE evaluation metric by combining statistical keyword extraction methods such as sentence ranking and word length with cluster extraction of GWO.
- In the study we tested with the BBC News dataset, GWO experimentally demonstrates that it performs well in the inferential text summarization techniques in line with the results obtained. The results also reveal the cluster success of GWO with an average of 57.27 for ROUGE 1 and an average of 45.62 for ROUGE 2.

INTRODUCTION

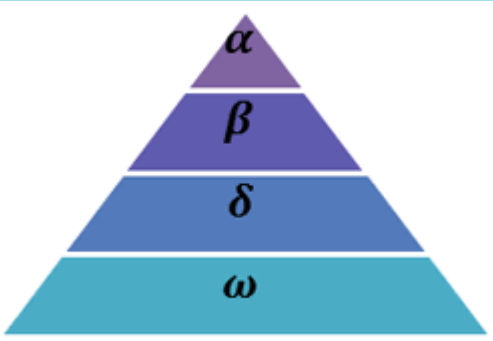
- Text summarization is the process of automatically creating a compressed version of the text document that represents the main idea of the text and the essence [1].
- The need for text summarization increases with the rapid and continuous increase in textual information sources in many fields. It can be useful for various applications such as automatic text summarization, document indexing, question answering systems, help systems and document classification.

- Text summarization techniques are divided into two as extractive and abstractive according to the output of the system. Text summarization techniques are divided into two as single and multiple document summaries according to the number of entries of the system.
- Single and extractive text summarization technique was used in the study. In the article, a summarizing technique has been developed using the sentence clustering feature of GWO.

RELATED WORKS

The earliest work on text summarization is based on sentence analysis and various approaches have been tried, including statistical learning approaches. Summarizing success of heuristic algorithms and classification techniques has been tested in many studies.

GWO



- Gray Wolf Optimization Algorithm is one of the population-based optimization algorithms developed with inspiration from nature. Simulating the hunting behavior of gray wolves, GWO was developed by Mirjalili in 2014[7].

- GWO, which is used in the solution of continuous and discrete optimization problems, has also been studied by combining it with other heuristic algorithms. [8].It is also presented as a new solution approach for nonlinear systems of equations [9]. GWO has been successfully applied for the solution of many engineering problems such as classification, system definition and filter design, feature selection.

- GWO starts with random solutions as much as the number of solutions in the given ranges of variables. Each solution represents the position of the wolves in the search space. The variable number of the problem corresponds to the size of the search space.
- Using the three best solutions in iteration, the average value of these three solutions and the positions of all wolves are updated. It is repeated until it reaches the maximum iteration. In the beginning, it starts to search with the solution of random wolves (search agent) in the given ranges of the variables.
- Each wolf's behavior examines three main steps: searching for prey, encircling prey, and attacking prey. Gray wolves are divided into alpha (a), beta (b), delta (d) and omega (w). The pseudo code of the gray wolf optimization algorithm is given in Figure 1.

Fig. 1. Pseudo code of the GWO algorithm

```
Start the Gray Wolf Populations
Assign parameters A, a, C
Calculate the suitability value of each wolf
Identify first, second and third best solutions
while (t <maximum iteration)
        for each agent
Update position
        end
        Update a, A, C values
        Update the fitness values of each wolf
       Update first, second and third best solutions
        t = t + 1;
end
```

MATERIAL AND METHOD

A. Data Set

In this article, 2004-2005 BBC News data set was used as data set [10]. The data set consists of 2225 English documents from the BBC News website corresponding to news in five current areas from 2004 to 2005. The news is divided into five classes: Business, Entertainment, Politics, Sports and Technology. In the study, a total of 30 news items randomly selected from each news branch were .

Start the Gray Wolf Populations Assign parameters A, a, C Calculate the suitability value of each wolf Identify first, second and third best solutions while (t <maximum iteration) for each agent Update position end Update a, A, C values Update the fitness values of each wolf Update first, second and third best solutions t = t + 1; end
tested.

B. Preprocessing

Before creating the summary, a series of steps are required to process the documents in advance. These steps consist of dividing the document into sentences, removing stop words, uppercase / lowercase transformation (lowercase transformation), and punctuation. All documents have been passed through these preprocessing stages.

C. Scoring Text Sentences

Four features are used to calculate the scores of the relevant text sentences. These are sentence position, sentence length, frequency of terms, and similarity to the title. When the necessary calculations for each feature are applied separately, the system collects them and calculates one point for each sentence and sorts them in descending order.

D. Sentences Selection with GWO

The sentences whose weights are calculated are clustered with the parameters entered with the help of GWO and a summary is created. In the study, the reduction ratio of the texts was determined as 60%. This rate is the closest reduction rate to the summary created by the human eye in the data set. This rate also affects the number of clusters to be created.

The values of other parameters used in the study are as follows. SearchAgents_no: 5, Max_iteration = 300, alpha = 0.1, n_grid = 10, beta = 4, gamma = 2 (Recommended values for the algorithm were used.)

.EXPERIMENTS ANS RESULTS

The Ngram association statistic (ROUGE) was used to evaluate the similarity of the sample summary obtained from the experiments and the human sample summary and the success of the study. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was developed by Lin in 2004 using the Perl programming language [11].

Rouge is a measurement method based on the number of common words of the two documents to be compared. ROUGE has five different measurement modes: ROUGE-N, ROUGE-L, ROUGE-S, ROUGE-W, ROUGE-SU. ROUGE-N. In this study, ROUGE-1 and ROUGE-2 measurements were applied. In order to evaluate the cluster success of GWO's text summaries, all operations were applied with k-means. Comparison of the results obtained is made in Table 1.

	GWO		K-MEANS	
Document Number	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
1	49.71	47.50	43.53	45.81
2	54.05	49.38	68.92	4135
3	87.52	62.96	67.50	54.72
4	49.09	42.62	58.18	46.68
5	59.28	54.32	62.32	47.27
6	61.54	50.23	45.23	41.12
7	53.04	48.98	58.34	47.17
8	53.17	41.12	53.47	40.04
9	68.99	51.38	74.29	56.12
10	68.75	57.83	50.12	32.18
11	65.23	53.33	46.67	38.46
12	41.82	33.78	50.20	40.66
13	60.45	55.06	56.67	43.06
14	52.34	38.55	48.45	31.54
15	58.14	41.24	68.04	51.28
16	52.83	45.76	60.00	34.49
17	58.70	47.17	54.35	45.28
18	64.52	59.38	64.52	58.33
19	50.47	41.18	54.13	43.02
20	67.12	54.24	63.12	53.17
21	57.47	51.58	60.12	51.56
22	58.67	39.73	48.38	35.54
23	54.12	37.35	48.57	39.76
24	54.87	40.91	51.68	33.46
25	56.23	42.29	51.69	41.37
26	58.12	44.78	56.72	33.12
27	52.12	33.57	48.19	37.45
28	48.34	38.22	54.15	41.29
29	58.45	47.23	48.99	43.27
30	53.45	40.12	50.17	38.23
Average	57.27	45.62	55.64	40.76

- According to the ttest results, although there is no significant difference in text summarization using GWO with Kmeans in ROUGE1, this difference is significant in Rouge 2 ($p = 0.005$).
- Results The GWO has been shown to be a successful tool for text summarization. Also in the study, a successful clustering was made even with the minimum number of solutions. In future studies, it is planned to test with different feature weights and solution numbers in order to increase the success rate.

REFERENCES

- [1] Hovy E, Lin CY. Automated text summarization and the SUMMARIST system. In: Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. Association for Computational Linguistics; 1998. p. 197–214.
- [2] J.B. Wang, P. Hong, J.S. Hu, "Automatic Keyphrases Extraction from Document using Neural Network", Springer-Verlag LNAI 3930, pp.633-641, 2006.
- [3] M. Mendoza, S. Bonilla, C. Noguera, C. Cobos, E. León, Extractive singledocument summarization based on genetic operators and guided local search, Expert Syst. Appl. 41 (9) (2014) 4158–4169.
- [4] Sinha, A., Yadav, A., Gahlot, A., 2018. Extractive text summarization using neural networks. arXiv preprint arXiv:1802.10137.
- [5] T. Young, D. Hazarika, S. Poria, E. Cambria Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag., 13 (3) (2018), pp. 55-75
- [6] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks.pages 93–98.
- [7] Seyedali Mirjalili, Seyed Mohammad Mirjalili, Andrew Lewis, Grey Wolf Optimizer, In Advances in Engineering Software, Volume 69, 2014, Pages 46-61, ISSN 0965-9978.
- [8] Chao Lu, Shengqiang Xiao, Xinyu Li, Liang Gao, An effective multi-objective discrete grey wolf optimizer for a real-world scheduling problem in welding production, In Advances in Engineering Software, Volume 99, 2016, Pages 161-176, ISSN 0965-9978.
- [9] Erdogmus, Pakize. (2019). A New Solution Approach for Non-Linear Equation Systems with Grey Wolf Optimizer. Sakarya University Journal of Computer and Information Sciences. 1. 1-11. 10.35377/saucis.01.03.475565.
- [10] http//mlg.ucd.ie/datasets/bbc.html
- [11] C. Lin, Rouge: A package for automatic evaluation of summaries, Text Summarization Branches Out: Workshop Held in Conjunction with ACL'04, pp. 74-81, ACL Press, Stroudsburg,PA,2004

Thank you…

- For further questions please ebrududak@hotmail.com