opción

Revista de Antropología, Ciencias de la Comunicación y de la Información, Filosofía, Lingüística y Semiótica, Problemas del Desarrollo, la Ciencia y la Tecnología

89

Año 35, mayo 2019 Nº

Revista de Ciencias Humanas y Sociales
ISSN: 1012-1587 ISSN-e: 2477-9385
Depósito Legal: pp: 198502ZU45

Universidad del Zulia
Facultad Experimental de Ciencias
Departamento de Ciencias Humanas
Maracaibo - Venezuela

Opcion, Año 35, Nº 89 (2019):*2899-2921*
ISSN 1012-1587/ISSNe: 2477-9385

Text Summarization using Gravitational Search Algorithm

Dr. Hayder Mahmood Salman
Al-Turath University College / Computer Science Department

Abstract
Extractive summarization is the process of precisely selecting a set of sentences from a corpus which can essentially be a demonstrative of the original corpus in a limited space. This paper presents a model which based on computing the coherence and the readability of each sentence in the corpus. The Gravitational Search Algorithm utilized to select the most suitable sentences to be included in the generated summary. The proposed model is evaluated using Doc. 2002 dataset The performance of the proposed model is evaluated using Rouge software. The calculated results support the success of the proposed model compared with other similar methods.

Keywords: Text summarization, coherence, readability, Gravitational Search Algorithm.

Resumen De Texto Mediante Algoritmo De Búsqueda Gravitacional

Resumen
El resumen extractivo es el proceso de seleccionar con precisión un conjunto de oraciones de un corpus que esencialmente puede ser una demostración del corpus original en un espacio limitado. Este artículo presenta un modelo basado en el cálculo de la coherencia y la legibilidad de cada oración en el corpus. El algoritmo de búsqueda gravitacional utilizado para seleccionar las oraciones más adecuadas para incluir en el resumen generado. El modelo propuesto se evalúa con el Doc. Conjunto de datos de 2002 El rendimiento del modelo propuesto se evalúa utilizando el software Rouge. Los resultados calculados respaldan el éxito del modelo propuesto en comparación con otros métodos similares.

Palabras clave: resumen de texto, coherencia, legibilidad, algoritmo de búsqueda gravitacional.

1. Introduction

Presently, the quantity of information on the Internet increased rapidly on any subject. Users need to get just the most pertinent information on a exact subject as fast as possible.. Automatic Text summarization (ATS) is the process to extract this significant information. ATS is the process of producing a single document from one document or more than one document with keeping the main ideas of the summarized document(s)[1]. Based on the number of document to be summarized ATS can be classified as a Single Document Summarization(SDS) or a Multi Document Summarization(MDS). In a SDS only one document can be summarized into shorter ones, whereas in MDS a set of related documents with the same topic is summarized into one shorter summary[2]. ATS techniques can also be categorized as abstractive summarization and extractive summarization. Abstractive summarization required deep natural language processing techniques, whereas extractive summarization does not require[3].
In this paper, a model for extracting generic ATS is proposed. The proposed model is based on sentences weighting using Term-Frequency and Iverse Document Frequency(TFIDF), then the coherence and the readability calculated for each sentence. Finally GSA used to select the most important sentences in the corpus.

2. Related Works

In this section some of TC methods will be investigated.

In[4] at 2014 the authors proposed an approach for extractive text summarization, after a preprocessing step a weighting method that's based on TFIDF is assigned to extracted sentences. The Cosine similarity applied to compute the similarity matrix between sentences and the keywords. The cost function is calculated by computing sentences coherence and readability. Finally a Multi-agent Partical Swarm optimization (MAPSO) used to find the most relevant sentences. In [5] at 2015 the authors proposed an approach for extractive text summarization, after a preprocessing step a weighting method that's based on TFIDF is assigned to the extracted sentences. The second step based on computing the similarity matrix between the sentences and the keywords. Finally a Cuckoo Search Optimization Algorithm (CSOA) used to obtain the most significant sentences to be involved in the final summary. In [6] at 2018 the proposed method based on representing each sentence as a bag of words. Cosine similarity used to compute the overlap between each sentence. Each of the sentences is treated as vertices of a similarity graph. The edges between sentences assigned a weight according to the similarity between them. The sentences are clustered by computing the average of the cluster. From each of cluster a sentence is choosing until reaching the summary length.

3. Problem Statement

To produce a good summary for any ATS system three issues must be considered. These issues are

A. Coherence: summary should include important sentences that talk about the same topic.

B. Readability: That indicates the selected sentences must relate to each other with a high degree of similarity.

C. Redundancy: The generated summary should include less redundant information to cover most of the relevant topics.

Formally, given a corpus which consists of many clusters, each cluster contains a set of documents called D with the same topic. The set D can be defined as D= {d1, d2,…, dn} where n is the number of distinct document in D. Each D can be represented by a set of sentences called Si, i.e D= {Si | 1<=i<=M}where M represents the total number of sentences in the set D. Our goal is to find a subset of set D called A i.e. A □ D that satisfies coherence, readability and should be less redundant as possible.

4. The Proposed Method

The proposed model consists of four main stages. In the first stage the preprocessing must be done to the input documents. The weight assigned to sentences based on the TFIDF in the second stage, while the coherence and readability are computed in the third stage. The final stage the Gravitational Search Algorithm (GSA) applied to choose, the important sentences to be included in the generated summary.

4.1 Preprocessing

Consists of the following steps:

A- Sentence segmentation: which can be performed by splitting sentences based on the dot between them.

B- Tokenization: The main goal of tokenization is separate sentences into words..

C- Stop Words Removal: is the manner of removing words that appear many times in the text and don't offer the required information for recognizing an important sense of the document. There are many strategies utilized for indicating such stop words list. Now, various English stop word list is generally utilized to the TC procedure.

D- Stemming: is the method of generating origin of the word. In this paper word stemming is done using Porter's stemming algorithm[7].

4.2 Sentence weighting

In this stage a weight is given to every sentence based on Eq.(1) and Eq.(2) as follows.

$$TF_{i,j} = \frac{Freq_{i,j}}{Max_i Freq_{i,j}} \quad (1)$$

$$IDF_i = log \frac{N}{n_i} \quad (2)$$

Where

$TF_{i,j}$ is the term frequency computed for every word i in the sentence j.

$Freq_{i,j}$ represents the frequency of word i in the sentence j.

$Max_i Freq_{i,j}$ is the maximum number of frequency for word i in the sentence j.

N represent the total number of sentences.

n_i represent the number of sentences in which word i occurs.

Next the similarity matrix is computed according to Eq.(3). This matrix is used to calculate the similarity between keywords and sentences.

$$sim(Si,k) = \frac{\sum_{i=1}^{t} W_{i,j} W_{i,k}}{\sqrt{\sum_{i=1}^{t} (W_{i,j})^2} * \sqrt{\sum_{i=1}^{t} (W_{i,k})^2}} \quad (3)$$

Where $W_{i,j}$ and $W_{i,k}$ are sentence weight and keywords weight respectively[5].

4.3 Sentence Coherence and Readability

Sentence coherence is a proportion of the measure of information that is normal to the arrangement of sentences that are as of now chosen and the new sentence that will be chosen. The readability indicates how relate the selected sentences to each others. The coherence and readability can be computed according to Eq.(4) and Eq.(5) respectively.

$$C_s = \frac{\sum_{v_i,v_j \in \text{Summary subgraph}} W(v_i,v_j)}{N_s}$$

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

Where

C_s represent average similarity of the sentences.

M is the maximum weight of sentences.

$$R_s = \sum_{0 \le i < s} W(s_i, s_{i+1})$$

$$RF_s \frac{R_s}{Max_{\forall i} Ri} \quad (5)$$

RFs represents the readability factor of a summary with length S [4].

4.4 Gravitational Search Algorithm based Text Summarization

GSA can be considered as one of the modern nature-inspired algorithms which can be used to solve optimization problems. It is built on the Newton's law of gravity and the Newton's second law of motion. Basically, inspired by the newton theory of gravitational interaction between masses. GSA is sophisticated by Rashedi et al. in 2009 which classified as a population based process that include various masses. . Built on the gravitational force, the masses are participation their acquaintance to make the search towards the best position in the search solution [8].

In GSA, the agent has four parameters which are position, inertial mass, active gravitational mass, and passive gravitational mass. The solution of the problem represented as the position of the mass, fitness function used to specify the gravitational and inertial masses. The algorithm is navigated by setting the gravitational and inertial masses, while each mass makes a solution the heaviest mass will be attracted. Therefore, an optimal solution is presented the heaviest mass in the search space [9].

Eq.(6) represent the Newton's law of gravity while Eq.(7) represent Newton law of motion.

$$F = G\frac{M_1 M_2}{R^2} \qquad (6)$$

$$a = \frac{F}{M} \qquad (7)$$

Where
M1 is the active mass, M2 is the passive mass,
r represent the distance between masses.
G is the gravitational constant which is changing through a course of time.

Summarized and important sentences should be extracted from the main text by GSA. Algorithm (1) shows the main steps of GSA as a text summarization problem.

Algorithm: GSA for Text summarization
Input: set of document collection D **Output:** summary of n sentences
Step1: collect a set of document D Set summary={} Step2: Apply preprocessing to each document D_i in the D Step3:Calculate the similarity matrix as in Eq.(3) Step4:Calculate the coherence as in Eq.(4) Step5: Calculate the readability as in Eq.(5). Step 6: generate initial GSA population using random sentences. Step7:repeat Step 8:while not(stop condition) do Step8.1: evaluate the fitness for each agent. Step8.2:update the G best. Step8.3: for each agent calculate the a as in Eq.(7). Step8.4: update position and velocity Step8.6 return best solution End while Step 9:Compare best solution with all summary sentences using cosine similarity if similarity<threshold add best solution to the summary. Else ignore best solution {this step the remove redundancy} Until (reach maximum summary length)

5. Dataset and Evaluation Metrics

The dataset utilized in this model experiment is Doc. 2002 standard documents. ROUGE will be utilized to evaluate the performance of the proposed model. ROUGE package produces three numbers representing: Precision, Recall and F−score[10]. They are formulated as follows.

$$\text{Precision} = \frac{\text{system summary sentences} \cap \text{ideal summary sentences}}{\text{number of sentences in the system summary}} \quad (10)$$

$$\text{Recall} = \frac{\text{system summary sentences} \cap \text{ideal summary sentences}}{\text{number of sentences in the ideal summary}} \quad (11)$$

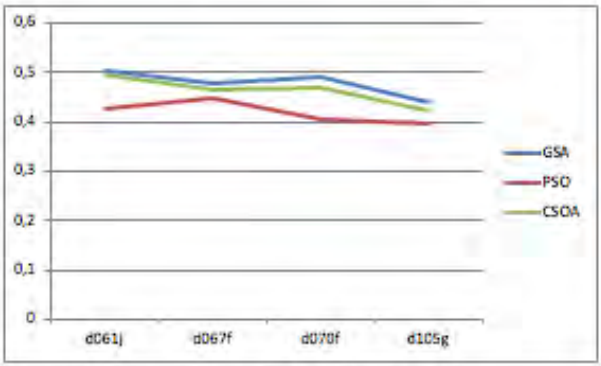
$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (12)$$

6.Experimental Results

Table-1 together with Figure-1 present the comparison results of Doc. 2002 standard such as d105, d070f, d067f, and d061j are used and respective results are examined through Rouge. The results of the proposed model compared with similar models such as PSO and CSOA. The calculated results support the success of the proposed model.

Table-1 Methods comparison based on the F-score results

Doc.2002	GSA	PSO	CSOA
d061j	0.50212	0.42869	0.49761
d067f	0.48011	0.44637	0.46476
d070f	0.49212	0.40616	0.47126
d105g	0.43825	0.39517	0.42391



Figure(1). Comparison of GSA, PSO and CSOA.

7. Conclusions

The need for effective ATS methods to get the significant information from a corpus becomes of necessity. A good summary should include sentences that are more coherent and readable to each others. In this paper a method for ATS are developed based on GSA. The proposed method computes the coherence and readability for every sentence, then GSA algorithm used to obtain best sentences to be involved in the final generated summary. The results show the success of the proposed model.

8. References

1- Jesus M. Sanchez-Gomez , Miguel A. Vega-Rodríguez , Carlos J. Pérez.(2018). " Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach", Knowledge-Base d Systems VOl. 159 PP.1-8.

2- Kumar, R. & Chandrakal,D. (2016)." A survey on text summarization using optimization algorithm," ELK Asia Pacific Journals, Vol. 2, No. 1.

3- Nallapati, R., Xiang, B. & Santos,C. (2016)." Abstractive Text summarization using Sequence-to-sequence RNNs and Beyond", Conference on Computational Natural Language Learning (CoNLL).PP.280-290

4- Asgari, H., Masoumi, B., & Sheijani, O. S. (2014). Automatic text summarization based on multi-agent particle swarm optimization. 2014 Iranian Conference on Intelligent Systems (ICIS

5- Mirshojaei,S & Masoumi, B. (2015) "Text summarization using Cuckoo Search optimization Algorithm". Journal of Computer & Robotics. Vol. 8 ,No. 2,PP.19-24.

6- Dutta M.,Das, A.K. Mallick, C., Sarkar, A.,&Das, A.K.(2018)." A Graph Based Approach on Extractive summarization". Emerging Technologies in Data Mining and Information Security.

7- Porter stemming algorithm: http://www.tartarus.org/martin/PorterStemmer/.

8- Rashedi,E, pour,H, & Saryazdi,S.(2009)." GSA: A Gravitational Search Algorithm"." Information Sciences. Vol. 179. No. 13. pp. 2232–2248.

9- Sabri,N, PutehM, Mohamad, R.(2013). "An overview of Gravitational Search Algorithm utilization in optimization problems" , IEEE 3rd International Conference on System Engineering and Technology.

10- Chin-Yew,L. (2004). “Rouge: A package for automatic evaluation of summaries.” In Text summarization branches out: Proceedings of the ACL-04 workshop, Vol. 8

UNIVERSIDAD DEL ZULIA

opción
Revista de Ciencias Humanas y Sociales

Año 35, N° 89, (2019)

Esta revista fue editada en formato digital por el personal de la Oficina de Publicaciones Científicas de la Facultad Experimental de Ciencias, Universidad del Zulia.
Maracaibo - Venezuela

www.luz.edu.ve www.serbi.luz.edu.ve

produccioncientifica.luz.edu.ve