Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Proposed Arabic Information Retrieval System Using Cat Swarm Optimization

Alia karim Abdul Hassan

Department of Computer Science, University Of Technology-Iraq. Email : 110018@uotechnology.edu.iq

ARTICLE INFO

Article history:
Received: 28 /02/2019
Rrevised form: 17 /03/2019
Accepted : 29 /04/2019
Available online: 12 /06/2019

Keywords:

Keywords:Information, Retrieval, Arabic, NLEL, Cat, Optimization, Feature, Selection.

ABSTRACT

According to the recently published research, the developed Information Retrieval systems are concerned with English language documents compared to all others in Arabic language. Ÿthe morphological difficulty of Arabic language increases the concerns for the availability of Arabic test copora. Therefore, This paper presents an Arabic information retrieval system for text documents. The proposed algorithm uses Cat Swarm Optimization to select the most important features with the cosine similarity. In addition, it finds the most relevant document to user query. The simulation results in using the standard NLEL of Arabic dataset corpus. The proposed algorithm for Arabic document retrieval uses swarm optimization with cosine similarity which provides effectively with accuracy 81.4%.

1 . Introduction

Large amounts of generating online documents and profiles over the social networks require accurate approaches to retrieve the most related information to the user query. Feature selection is one method in datamining that help to find important features in and help obtain more effective results in extract the required information. In most researches to get more precise results the Feature selection is applied as a primary stage, as in medical field for Breast cancer and Al Zheimer [7,4]. The most used approaches to find the finest set of features are wrapper, filter and the embedded approach. Bio inspired algorithms might be considered as frequently used algorithms in solving computational and hard problems. A few of these algorithms were utilized efficiently in Arabic language such as Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Firefly Algorithm (FA) [13]. Web Mining might be defined as applying the methods of data mining for exploring and understanding the usage and patterns related to data on the web to achieve the requirements of applications on the web. Usage data capture the origin or identity and the browsing behavior of the user on the website [12].

The two terms “seeking” and “tracing” are the modes which represent the two major behaviors on which the Cat Swarm Optimization CSO, is based. In application the CSO in the optimization problem, the first step is to determine the number of cats to use and its (cat) has its own M-dimensional position, velocities for each dimension. To identify whether the cat is in seeking or tracing mode a fitness value which represent the accommodation of the cat to the fitness function and a seeking/tracing flag. The best position for one of the cats represent the best solution,

Corresponding author : Alia karim Abdul Hassan

Email addresses: 110018@uotechnology.edu.iq

Communicated by Qusuay Hatim Egaar

which is kept by the CSO until the end of all iterations [5]. In this paper a proposed information retrieval of Arabic text document CSO with cosine similarity algorithm to discover and fetch the more relevant documents.

2. Related work

The most related works in Arabic document Information retrieval that used the swarm intelligence as a feature selection are:

Duwairi et al. (2007) [6], they use two techniques stemming and light stemming as feature selection, applied to Arabic corpus and then compared the result. The used Arabic text document dataset were manually collected and prepared from internet sites. In experiments the light stemming as a feature selection method give better results than using stemming.

Harrag et al. (2010) [8] proposed three methods to feature a selection of documents in Arabic language are used are namely Document Frequency, Latent Semantic Analyses, and Term Frequency Inverse Document Frequency . Term Frequency Inverse Document Frequency proved its effectiveness at the experiments results that performed on an Arabic dataset.

Aisha Adel [1], the author use a combination of two techniques for feature selection methods based on the average weight of the features for Arabic language. The experiments are conducted to classify a published Arabic corpus using Naïve Bayes and Support Vector Machine classifiers. The information Gain method gets the best results. The results also show that the combination of multiple feature selection techniques outperforms the best results obtain by the individual methods.

In Abdul Hassan A. and Abdul Ameer Z. [3] Modify the chicken swarm optimization algorithm to select the best feature to retrieve the most relevant Arabic documents. Experimental results using ZAD corpus data set show that the proposed algorithm improves the accuracy of retrieval results.

3. Cat Swarm Optimization[5,10]

CSO first proposed in 2007. It mimics the cat's behavior with two modes are seeking and tracing. For each 'cat' define a position and direction of movement(velocity)[6]. The CSO algorithm has several parameters are:

- *The initial feasible solutions is N*
- *The ratio of cats in the tracing process called ´Mixture Rating (MR).*
- *The best candidate solutions in the seeking process called Seeking- Memory- Pool (SMP).*
- *The state of the candidate solution determined by Self-Position-Consideration (SPC), its value is either true or false ; is true if the candidate solution stays without changing seeking- process.*
- *The seeking range of the selected position in seeking- mode is the Seeking- Range-Dimension SRD.*
- *The number count of dimensions in seeking process (the dimensions to be changed) is the Count -Dimension to Change CDC.*
- *D represent the number of selected features ´Dimention and the tracing process have a constant value is C*

In Seeking mode several copies of ´current cats where generated all of the them represent a candidate solution. In case SPC has value true, one of the candidates stays in the same position (stationary), while the other positions not changed. The position of each selected dimension then is to be changed with a value SRD-´ percent current position. The position value will either increase or decrease randomly. Equation-1, used to compute the probability of each candidate cat.

$$pi = \frac{|FT_i - F|}{FT_{MAX} - FT_{MIN}} \quad ...(1)$$

where

P_i :selected probability´candidate *i*;

FT_{max} : maximum value fitness of all candidates;

FT_{min} :minimum value fitness of all candidates.

FT_i :fitness of candidate *i*.

F : FT_{min} if the goal is to find the maximum fitness, and is FT_{max}´, if the goal is to find the minimum fitness.

Each candidate cat has (P_i´), the original copy will be replaced with the selected candidate cat[11].

In tracing – mode equation -2& 3 are used to compute best cat (fitness) in this mode, then change the velocity of each cat′

$$v_{nw} = v^{t}{}_{k,d} + R \times c1 \times \left(x^{t}{}_{bst,d} - x^{t}{}_{k,d}\right), \quad \dots(2)$$

$$x_{nw} = x^{t}{}_{k,d} + v_{nw} \quad \dots(3)$$

Where

$v^{t}{}_{k,d}$: the cat_ k velocity in t _ iterations;

d: dimension to be changed (d value in [1..D]);

$x^{t}{}_{d}$: the cat position with best value in t- iteration; and $x^{t}{}_{k,d}$ is cat$k$ *position* .

R: randomly generated number in[0..1]

c1 : constant value

v_{nw} : new velocity of cat k

x_{nw} : new position of cat k.

4. Information Retrieval &Document Representation

The process of organizing, storing, representing and searching information items is defined as Information retrieval (IR). Information should be organized in a way that guarantees retrieving related information. Vector space model (VSM) is very commonly utilized in IR to retrieve the documents. In VSM the documents and queries are stored in weights vectors [2]. The weight vector of a document will form $<w_{d,1}, w_{d,2}, w_{d,3}, \dots, w_{d,n},>$, using equation-4. The weight vector of a query will form $< w_{q,1}, w_{q,2}, w_{q,3}, \dots\dots\dots, w_{q,n}, >$, between a query q and a document d using equation-5 [10].

$$w_d = \frac{tf_d \times idf}{\sum (tf_d \times idf)^2} \quad \dots(4)$$

$$w_{q=\left(0.5+0.5\times\frac{tf_d}{\max tf_d}\right)\times idf} \quad \dots(5)$$

Where tf is term frequency and idf is the inverse document frequency and common it formulated as log ($N/_{df}$), N might be defined as the size of document collection, while df could be defined as the document's frequency[8]. The factor is normalized by the maximum in the query vector [10]. Vector space model can be constructed using equation-6.

$$\text{VSM } (d, q) = \sum_{t=1}^{n} (w_{d,t \times W_q,t}) \quad \dots(6)$$

5 . Data Set

One of the available Data set that used for Arabic information retrieval systems is the Arabic Wikipedia corpus, (11638 Arabic _Wikipedia documents in S.G.M.L format, 193 queries), that built by Benajiba and others, which NLEL University of Valencia publish it. Table-1 display sample of queries from this data set [2].

Table(1) sample of queries

No	Query
.1	1. كم هي العوائد السنوية لغوغل؟
.2	2. كم ينفق على التدخين سنويا في الولايات المتحدة الامريكية؟
.3	3. كم تبلغ القيمة المادية لجائزة المغرب للكتاب؟
.4	4. كم يبلغ طول ضلع بركة الحمام؟
.5	5. كم يبلغ ارتفاع برج ايفل؟
.6	6. كم كان عمر سنفستر عندما بدا مسيرته الفنية؟
.7	7. كم كان عمر الامير سعود عبد العزيز بن متعب الرشيد عندما تولى الحكم؟
.8	8. كم عدد موتى اعصار 18 سبتمبر 1906 بهونج كونج؟
.9	9. كم كانت قوة زلزال هوكايدو؟
.10	10. كم من كتاب الفه وجالينوس في الطب والصيدلة؟

6. The Proposed System

The proposed system to retrieve Arabic text documents by using cat swarm optimization have the following main steps

Step1: Preprocessing the user query and the documents in database, to extract the basic words which are meaningful and useful. Preprocessing includes Tokenization, stop word removal, stemming and normalization. In Tokenization the entire files and documents are converted into separate words, then normalization often removes punctuation, diacritics (primarily weak vowels) and non-letters. Stemming remove the word derivatives and return the word to its root.

Step2: Document Representation :Vector space model representation for both preprocessed query and documents is made using equations 4,5,and 6. Now each document in the database was represented in a vector of (tf_idf) which represent the document feature.

Step3:Feature selection using CSO: To represent each document with the best feature CSO is used to optimize the document features to best feature vector as described in algorithm-1. Table-2display the initial values for the CSO parameters. With D-dimensional space Algorithm-1 will randomly generate N- solution sets , represented as cats. Then find min (tf-idf), max(tf-idf) for each document and query. For each solution set a fitness is computed using equation (1). The best cats value will copied in (Xg) which represent the value of best solution. Step for each value (min-max) obtains weight between them. This occurs in each document or query depend on fitness value. On MR the seeking/tracing mode will assigned to cats. The search operation will be made according to the selected mode. If termination conditions are satisfied, output the best subset; otherwise, return to step3. The best solution is a feature vector its length is (10).

Algorithm 1*: best feature construction using CSO*
Input: f*eature-vector(tf-idf) for preprocessed-document*
Output: *best- feature subset for each document.*

Step1: cat population Xi (i = 1, 2, … , n), v, and SPC
Step2:While (the stop criterion is not satisfied or i < imax)
- Calculate the fitness function values for all cats each cat position represent the value of this fitness
- measure the probability by using equation-1 and sort them
- Xg= cat with the best solution

- *For i = 1: n*
If SPC = 1 then Start seeking mode
Else Start tracing mode
- *End for i*

- *End while*

Step3:end

Table(2) Prameter Setting

Parameter name	value
Dimension	10
Iteration	100
MR (number of cats that hunt)	10
SMP (seeking memory pool)	2
SPC(self-position considering)	False
CDC (counts of dimension to change)	1
SRD(seeking range of the selected dimension	0.1
w(constant)	0.1

Step4:Similarity Computation: The Similarity computation between the documents best feature vector output of CSO and query feature vector as described in algorithm-2.

Step5:Rank list of the most relevant retrieved documents. According the similarity value sort the documents in a list in which the most related documents will be at the top of list.

11. Algorithm-2: Relevant Documents list
Input: Query best feature vector , Documents-best feature vectors
Output: rank- list

Step1: for each vector (*FQ*),(*FD*) compute the validate value which represents weight value (w)
Step2: similarity computation
12. For each weight in (FQ),weight in (FD) do
13. If (weight -value!=null && weight > 0) And (sum - weight < size -weight) then
14. Find the max -weight (FQ,FD)
15. End if
16. Next
Add FD with Abs (max (FQ, FD)) to the ranke-d list
Step4: *Sort Ran-klist*

Step5: *Return Rank- list*

7. Experimental Results

The Proposed system simulated using NLEL Arabic Wikipedia corpora. The proposed system experimented using the sample of queries listed in table-1. Accuracy is used which is used for measuring the effectiveness of learning algorithms. Accuracy indicates the total number of runs, which are correctly returned in list, (see equation- 7)[9].

$$\textbf{Accuracy} = \frac{TP + TN}{ALL} \quad \dots (7)$$

Where,
TP: the total number of flows which are relevant returned
TN: the total number of flows which are irrelevant returned
From table-3 the proposed system achieves 81.4% as average accuracy while, if the retrieval algorithm applied with only the cosine similarity, the average accuracy of the same input is 75.4%.

Table(3) Accuracy Value

Query no	Query	Retrieval system with only cosine similarity	Proposed system
Q1	17. كم هي العوائد السنوية لغوغل؟	75%	85%
Q2	18. كم ينفق على التدخين سنويا في الولايات المتحدة الامريكية؟	78%	83%
Q3	19. كم تبلغ القيمة المادية لجائزة المغرب للكتاب؟	77%	84%
Q4	20. كم يبلغ طول ضلع بركة الحمام؟	77%	87%
Q5	21. كم يبلغ ارتفاع برج ايفل؟	79%	80%
Q6	22. كم كان عمر سنفستر عندما بدا مسيرته الفنية؟	74%	81%
Q7	23. كم كان عمر الامير سعود عبد العزيز بن متعب الرشيد عندما تولى الحكم؟	79%	80%
Q8	24. كم عدد موتى اعصار 18 سبتمبر 1906 بهونج كونج؟	65%	77%
Q9	25. كم كانت قوة زلزال هوكايدو؟	70%	76%
Q10	26. كم من كتاب الفه وجالينوس في الطب والصيدلة؟	80%	81%
Average		75.4%	81.4%

8.Conclusions

A proposed system for Arabic text document retrieval based on using the feature selection to enhance the retrieval results. The proposed method uses the cat swarm algorithm as a feature selection approach and the cosine similarity measure to improve the rank list result of user query. The proposed algorithm proved its effectiveness when experimented with NELE corpus Arabic Data set with accuracy 81.4%. Applying retrieval algorithm using only the cosine similarity achieved accuracy 75.4% with the same sample of the data set.

Reference

1- Aisha A., Nazlia O., Adel Al-Shabi "A Comparative Study Of Combined Feature Selection Methods For Arabic Text Classification ". Knowledge Technology Group, Centre for AI Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia,2014.

2- Alia K. Abdul Hassan, Zainab A. Ameer, 'Search Result Enhancement For Arabic Datasets Using Modified Chicken Swarm', Iraqi Journal of Science, Vol. 59, No.4A, pp: 1952-1958,2018,

3- Alia K. Abdul Hassan, Mustafa J. Hadib "Proposed MABC-SDAIR Algorithm For Sens-Based Distributed Arabic Information Retrieval ", Journal of Theoretical and Applied Information Technology, Vol.95, No.3, 2017.

4- Ays¸e Demirhan, Talia M Nir, Artemis, " Feature Selection Improves the Accuracy of Classifying Alzheimer Disease using diffusion tensor images". In Biomedical Imaging (ISBI), IEEE 12th International Symposium on, pages 126–130.IEEE, 2015.

5- Chuan S. Chu, P. wei Tsai, Shyang J. Pan, "LNAI 4099 - Cat Swarm Optimization," pp. 854 858, 2006.

6- Duwairi, R., M. Al-Refai, Khasawneh N., "Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization". Proceedings of the 4th International Conference on Innovations in Information Technology, Nov. 18-20, IEEE Xplore Press, Dubai, pp: 446-450. DOI:10.1109/IIT,4430403, 2007.

7- Girish Chandrashekar and Ferat Sahin, "A Survey On Feature Selection Methods", Computers & Electrical Engineering, 40(1):16–28, 2014.

8- Harrag, F., El-Qawasmeh E., Al-Salman A.M.S., "A Comparative Study of Statistical Feature Reduction Methods for Arabic Text Categorization". In: Networked Digital Technologies, Springer, Berlin Heidelberg, pp: 676-682. ISBN-10:978-3-642-14305-2,2010.

9- Hemanth KS. Doreswamy, "Performance Evaluation of Predictive Classifiers for Knowledge Discovery from Engineering Materials Data Sets" in arXiv preprint arXiv: 1209.2501, 2012,

10- Lin K. C., Kai Y. Zhang, Yi H. Huang, Jason C. Hung, Neil Y., "Feature Selection Based On an Improved Cat Swarm Optimization Algorithm For Big Data Classification." *The Journal of Supercomputing* 72 (2016): 3210-3221.

11- Lin KC, Chien H., "CSO-based Feature Selection and Parameter Optimization for Support Vector Machine". In: Joint conference on pervasive computing (JCPC), pp 783–788, 2009.

12- Poonkuzhali Sugumaran1, Kishore Kumar Ravi1("A Novel Algorithm for Enhancing Search Results by Detecting Dissimilar Patterns Based on Correlation Method"), The International Arab Journal of Information Technology, Vol. 14, No. 1, January 2017.

13- Souad Larabi Marie-Sainte ," Firefly Algorithm based Feature Selection for Arabic Text Classification", Journal of King Saud University – Computer and Information Sciences (2018).