ISSN: 0067-2904

Search Result Enhancement For Arabic Datasets Using Modified Chicken Swarm

Alia Karim Abdul Hassan*, Zainab AbdulAmeer
Computer Science Department, University of Technology, Baghdad, Iraq

Abstract

The need for information web-searching is needed by many users nowadays. They use the search engines to input their query or question and wait for the answer or best search results. As results to user query the search engines many times may be return irrelevant pages or not related to information need. This paper presents a proposed model to provide the user with efficient and effective result through search engine, based on modified chicken swarm algorithm and cosine similarity to eliminate and delete irrelevant pages(outliers) from the ranked list results, and to improve the results of the user's query . The proposed model is applied to Arabic dataset and use the ZAD corpus dataset for 27300 document. The experimental result shows that the proposed model improves the precision, recall, and accuracy. Thus the result produced by this method improves accuracy.

Keywords:-search result, dissimilar pattern, outlier ,chicken algorithm, swarm, redundant, relevant, rank list

العنوان (تحسين نتائج البحث لقاعده بيانات عربيه باستخدام خوارزميه الدجاج المحدثة)

علياء كريم عبد الحسن*، زينب عبد الامير شمال
قسم علوم الحاسبات، الجامعه التكنولوجيا، بغداد، العراق

الخلاصه

هناك حاجة إلى البحث عن المعلومات على شبكة الإنترنت من قبل العديد من المستخدمين في الوقت الحاضر.حيث يستخدمون محركات البحث لإدخال استفساراتهم أو سؤالهم وانتظار الإجابة أو أفضل نتائج للبحث. قد تؤدي النتائج التي يتم إجراؤها إلى طلب بحث المستخدم لمحركات البحث عدة مرات إلى إرجاع صفحات غير ملائمة أو لا تتعلق بحاجةالمستخدم إلى المعلومات المطلوبه. نقدم في هذا البحث نموذجًا مقترحًا لتزويد المستخدم بنتيجة فعالة وملائمه من خلال محرك البحث ، استنادًا إلى خوارزمية سرب الدجاج المعدلة وتشابه جيب التمام لإزالة وحذف الصفحات غير الملائمة (المتطرفة) من نتائج القائمة المرتبة ، ولتحسين نتائج طلب المستخدم. يتم تطبيق النموذج المقترح على مجموعة البيانات العربية واستخدام مجموعة بيانات ZAD corpus لمستند 27300. اظهرت النتائج التجريبية أن النموذج المقترح يحسّن الدقة . وبالتالي فإن النتيجة التي تنتجها هذه الطريقة تعمل على تحسين الدقة.

1.Introduction

The internet is considered as a primary source of information; producing tons of electronics with the rise in various technologies [1]. The data in the web is mostly unstructured or semi-structured, which contain a mix of video, audio, text and image; where it is required to mine the information related to the user's specific needs. Outliers are the observations whose actual value is various than the

*Email: hassanalia2000@yahoo.com

remain of the observed value of the data documents and of all types such as web pages, research papers, e-mails, audio and video documents [2]. The process of locating web content outliers in large web data is called Web content outlier mining. Traditional algorithms for outliers mining are designed solely to one type of data such as numeric data sets, however; the outliers mining algorithms that deal with the web must be utilized, for different types of data like html tags, video, hypertext, image, audio[3]. A large scale of web pages in the (WWW) makes result that provide from search engines , many of irrelevant information hold it Thus, the information finding that the user needs has become complicated and hard, therefore; retrieving information in a proper way has become more important. These challenges can be solved by modern techniques like Differential Evaluation (DE), Genetic Algorithm (GA), Artificial Potential Field (APF), Ant Colony Optimization (ACO), Neural Network (NN), and Chicken Swarm Algorithm that can be more effective method to solve the information retrieval problems [4,5].

In this paper a proposed model uses the chicken swarm algorithm (CSA), which is a meta-heuristic algorithm depends on groups that is represent the behaviors of the CSA. The search space is divided into multi groups, each group of swarm has one rooster and multi hens and chicks. The various groups of chickens follow up and use various laws to motion. The hierarchy of order, which is represent the competition among the different sub collections [6]. Based on the CSA parameters analysis the proposed model modify CSA to enhance the search result. Cosine similarity with a threshold is used to discover and fetch the more relevant documents.

2.Related work

The most recently work on search enhancement for information retrieval systems which are based on detecting the dissimilar pattern(outlier)**,** S. Sathya Bama [3] proposed an algorithm to remove the outlier (redundant and irrelevant) webpage use a mathematical approach based on correlation method.

The strength of this approach is appear in the results that obtained where its more accurate than other method but in this work need to explore further outlier from webpage. Poonkuzhali Sugumaran1[7] used a statistical method which depend on correlation method is developed for retrieving relevant web document through outlier detection technique. In addition, this method also identifies the redundant web documents. Removal of both redundant and outlaid documents improves the quality of search results catering to the user needs. Evaluation of the correlation method using Normalized Discounted Cumulative Gain method (NDCG) gives search results above 90%. W.R. Wan Zulkifeli,[2] introduced a study to mining and remove the outlier by using the classical term weighting (tf-idf)(term frequency-inverse document frequency) in dissimilarity measure. This research used maximum frequency normalization and applied a traditional term weighting method in IR to use the value of less frequent terms among documents which are considered as more discriminative than frequent terms. This study use 20 newspaper as a dataset. The experimental result show the effective of this approach and the accuracy found is (91.10). Abu Kausar, [4] review several algorithms such that ant colony algorithm, artificial neural network, genetic algorithm, ant algorithm and differential evaluation, To resolve the problems information that are not relevant or irrelevant. This reviews proved swarm algorithms have many attractive applications in information retrieval and make an information retrieval system more powerful. A. Jenneth, K. Thangavel [8] this work use K-Means clustering algorithm to separate the whole dataset into K clusters. The centroid point is computed from the data set instead of choosing the centroid point randomly . For each test document and the centroid point the distance is computed. The cluster has minimum distance is taken and the remaining cluster is put onto the stack. Experimental results show proposed work system is efficient. Khushboo Bhatt[9] apply the firefly algorithm and the Naïve Bayes classifier to improve the selection of the best feature in webpage as aimed to reduce the webpage classifier problem. The classifier optimization is to select the best feature in each webpage to reduce the feature space of webpage classifier problem return best result to user. The firefly algorithm depends on the concept of cluster in benchmark problem hence the cluster analyzed is a way to define for each homogenous group of data together. The firefly algorithm is found most efficient than the Artificial Bee Colony (ABC) ,Particle Swarm Optimization (PSO), and other algorithms of swarm that used. As a result the measure of test analyzed is F-measure =0.962 by using the classifier (NB) and the accuracy is 98.90 .

Most of IR (or DIR) researches are concerned to manipulate the documents that are written in the English language. In contrast, there is a few IR (or DIR) researches are concerned with the Arabic language [10].

From the most recently related work it can be found that the objective of their works was to remove the dissimilar pattern to eliminating the irrelevant and not related document to user query or use the swarm algorithms to enhance the search results for English data sets.

In this paper the proposed model remove the dissimilar pattern to eliminating the irrelevant and the related document to user query or use the chicken swarm algorithms to enhance the rank list to user query for Arabic datasets.

3.Methodology

Term weighting is refer to the weight vector of a document will form $<w_{d,1}, w_{d,2}, w_{d,3}, \dots\dots\dots, w_{d,n},>$, use equation 1 [11]:

$$w_d = \frac{tf_d \times idf}{\sum (tf_d \times idf)^2} \quad \dots (1)$$

Weight vector of a query will form $< w_{q,1}, w_{q,2}, w_{q,3}, \dots, w_{q,n}, >$, between a query q and a document d, equation-2 [11]:

$$w_{q=\left(0.5+0.5\times\frac{tf_d}{\max tf_d}\right)\times idf} \quad \dots(2)$$

Where tf is term frequency and idf is the inverse document frequency and common it formulated a log ($^{N}/_{df}$), where N is the size of the document collection and df is the document frequency[12]. The factor is normalized by the maximum in the query vector [11].

The vector space model refer to the inner product between these two vectors is computed to find the similarity between a query and a document use equation- 3 [11]:

$$\text{VSM } (d,q) = \sum_{t=1}^{n} (w_{d,t\times W_q,t}) \quad \dots(3)$$

the cosine similarity is a value reflecting the similarity between two documents xi and yj by usingequation-4[11]:

$$\cos(xi, yj) = \frac{xi \times yj}{\|xi\| \times \|yj\|} \quad \dots(4)$$

4. Chicken Swarm Algorithm

Chicken swarm Algorithm (CSA)is consider as new intelligent meta heuristic algorithm proposed according to various behaviors of rooster , hens, and chicks in the process of searching food. In this algorithm, chicken swarm in searching space is mapped as specific particle individual. *Rooster particle swarm*, *hen particle swarm*, and *chicken particle swarm* are sorted according to fitness value of particle, and each sub swarm uses different searching mode. In chicken algorithm , multi particles that has best fitness are selected as rooster particle swarm, which is given by the CSA simulate the hierarchal order in the chicken swarm and the manner of the chicken swarm, which stems from the manner of the birds behavior. The CSA algorithm can be partitioned to multi groups , each one of them consists of a single cock(rooster) and chicken and a group of chicks. In the process of finding food note that it is always the preference for the roosters, either chickens follow the cocks to find their food while the chicks follow their mother in the search for their food. The various individuals within the chicken population follow the various laws of movement. There is a competition among the various squad members under their hierarchical system where the position of each individual within the chicken swarm is a suitable solution to many problems of improvement [6].

CSA basic variables are, R, H , C and M are the number of roosters, hens, chicks and mother hens, respectively; N is the number of the whole chicken swarm, D is the dimension of the search space; and $x_{ij(t)(i\in[1,\dots\dots\dots,N],j\in[1,\dots\dots\dots,D])}$ is the position of each individuality at time t. Best R chickens would be supposed to be the roosters, while the worst C ones would be assumed as the chicks. The rest of the chicken swarm is observed as the hens [5,13]. Roosters with best fitting values have primacy for food access than the ones with worse fitting values. Equation-5 used to update rooster location.

$$x_{ij}^{(t+1)} = x_{ij}^{(t)} + (1+ \text{Ran}(0,\sigma^2)), \quad \dots\dots\dots(5)$$

Where

$$\sigma^2 = \begin{cases} 1, & f1 \geq f_{k,} \quad k \in [1,N], K \neq i, \\ \exp\left(\frac{fk - fi}{|fi| + \varepsilon}\right), & other\ wise, \end{cases}$$

And Ran(0,σ^2) represent a random number of Gaussian distribution with mean 0 and (σ, ϵ) is refer to small constant used to avoid zero-division error, and k is the index of rooster , which selected random from the roosters group (k≠ i), fi is the fitness value of particle i. As for the hens, they can follow up their group-mate roosters to seeking for food and randomly theft the food found by other the individuals. Equation-6 used to update hens position[12].

$$x_{ij}^{t+1} = x_{ij}^{t} + \text{C1 Ran}(x_{r1,j}^{t} - x_{i,j}^{t}) + \text{C2 Ran}(x_{r2,j}^{t} - x_{i,j}^{t}) \quad \ldots(6)$$

$$x_{ij}^{t+1} = x_{ij}^{t} + \text{C1 Rand}(x_{r1j}^{t} + x_{ij}^{t}) + \text{C2 Rand}(x_{r2j}^{t} + x_{ij}^{t}) \quad \ldots(7)$$

Where the value of C1, C2 as follows:-

$$\text{C1} = \exp((f_i - f_{r1})/(|f_i| + \varepsilon)), \quad \ldots(8)$$

$$\text{C2} = \exp((f_{r2} - f_i)), \quad \ldots(9)$$

Where Ran is a random number in [0,1], r1 is an index of the rooster, which is the i th hen's group-mate, and r2 is the index of the chicken (rooster or hen), which is randomly chosen from the chicken swarm(r1≠ r2). With respect to the chicks, they follow their mother to forage for food. The position of chicks swarm is update equation-9.

$$x_{i,j}^{t+1} = x_{i,j}^{t} + \text{F}(x_{m,j}^{t} - x_{i,j}^{t}) . \quad ...(9)$$

$$\text{x}_{\text{ij}}^{\text{t+1}} = \text{x}_{\text{ij}}^{\text{t}} + \text{f}(\text{x}_{\text{mj+}}^{\text{t}}\text{x}_{\text{ij}}^{\text{t}}) \quad \ldots .(10)$$

Where $X_{m;j}$ (t) is referring to the chick's mother position, Fϵ[0,2] is the flow coefficient, which marks that the chick follows its mother to forage food[12].

The fittenes function in this application use the (cosine similarity) value as we see later in the details of the algorithm2.

5.Proposed Information Retrieval Model for Arabic Datasets Using Modified CSA

The proposed model information retrieval model for Arabic Datasets using modified CSA to enhance and refine the results retrieved from the search engine by eliminating undesirable values and relying only on retrieval of documents and results that are relevant to the user query which provide more effort to the user. The detailed description of the proposal as the following:

Step1:Document preprocessing step in both (offline and online), in which extracts the basic words which are meaningful and useful. Document preprocessing includes Tokenization, stop word removal, stemming and normalization. Tokenization is the process by which the entire files and documents are converted into separate words, which are referred to as tokens. Stop word removal is remove popular words that are not useful in the search process such as(.....وحتى ،ان، اما ،وكان), etc. Stemming is the process of removing the word derivatives and return the word to its root and the normalization often removes punctuation, diacritics (primarily weak vowels) and non-letters.

Step2: compute term frequency and assign weights to each term in every document. Assign weight for a term in a document, use equation-1. Assign weight for a term in a query use equation-2

step3: Document representation by vector space model use equation-3.

Step4: Dissimilar patterns finding use proposed algorithm(algorthem-1). In algorithem-1 for each document (xi) or (yj) in database collection there is two list are the **outlier list** and another list is **relevant list**, the two list constructed using equation-4 with threshold determined by experiment. For the outlier list ,the cosine similarity the threshold value with in [0,0.02], while for the relevant list the cosine similarity within [0.03,0.09]. Algorithem-1 return two lists relevant and outline keep it on file.

Step5 : use algorithm -2 to retrieve the document. In algorithm-2 randomly select N documents from data set. For each d in N use equation-4 to compute the cosine similarity between document and query put the L, list contain the documents fitness values.The control variable values taken from L by a different rooster document as highest values using equation-5. Then determine the best hens

which have global best searching food mechanism fitness value, the roosters, hens and chicks are arranged in the ascending order their (sound echoes) and the first hen's will be the candidate with the best food searching candidate (minimum cost)and give best global index value using a modified(by experiment) version of equation -6 described in equation -7. While the new rooster document is generated around the global best swarm by adding/subtracting a normal random number using a modified(by experiment) version of equation-9 described in equation -10 . finaly the rooster ,hen , chick documents with the best global fitness value until this process take place with the number of iterations, the steps 3-6 are repeated until stopping criteria has not been achieved.the output of this algorithm is relevant document list to query.

$$x_{ij}^{t+1} = x_{ij}^{t} + C1\ \text{Rand}(x_{r1j}^{t} + x_{ij}^{t}) + C2\ \text{Rand}(x_{r2j}^{t} + x_{ij}^{t}) \quad \ldots.7$$

. $$x_{ij}^{t+1} = x_{ij}^{t} + f(x_{mj+}^{t} x_{ij}^{t}) \quad \ldots.8$$

Step8: Remove the outlier list from retrieved documents from algorithem.

Algorithm -1: dissimilar pattern finding
Input: Database
Output: List of Relevant document and list outlier document

Step1: for each document in Database use vector space model representation use equation-3.
Step2: for each pair of document in Database do
- Compute Cosine similarity(Cos) using equation-4
If Cos value within [0, 0.02] then the result of set document with document D is outlier list
Else
If Cos value within [0.03, 0.09] then the result of set document with document D is relevant list
Step3: return(Relevant document list, outlier document list).

Algorithm-2: Chicken Swarm based Document retrieval
Input: data set D, N population size, query
Output: rooster list.

Step1: Randomly select N documents from data set.
Step2: For each d in N use equation-4 to compute the cosine similarity(fiteness function) between document and query put the L. //L list of documents fitness values
Step 3: **determine the best values as rooster group in L** *by a different rooster document as highest values using equation-5*
Step 4: the second highest value of fiteness represent the hens document which follow the rooster document to get fooding using equation -7.
Step 5: the remainder individual in chicken swarm represent the chick document which has the minimum value and it follow her hens or mother using equation -10.
Step 6:resort of the value as rooster ,hens, chicks document which one is given the best global fitness value until this process take place with number of iteration.

Step7: the steps 3-6 are repeated until stopping criteria has not been achieved.
Step8:- for each rooster list check it with result of algorthim1 and remove the outlier document in
Ste9: return rooster list // relevant document list.

6.Experimental Result

The proposed model was implemented using python programming language -----. The proposed model was experimented Zad-Al-Ma'ad corpus, name as ZAD and contained (2730 Arabic documents, 25 Arabic queries, supported by relevance judgments), this dataset is written by the Islamic scholar "Ibn Al-Qyyim". Redis database server were used to provide a very simple client protocol similar to Telnet and to draw a simple simulation to a client-server database structure.

For comparative study traditional model for information retrieval were implemented use the Inverted-Index, VSM, and the two lists of dissimilar patterns. Table-1 show comparative result of the two models for Arabic Information retrieval proposed one dissimilar patterns and the traditional using the Precision and recall evaluation metrics. The samples of eight quires (q1, q2, q3, q4, q5,q6,q7,q8 in ZAD collection) are selected depending on the ownership of largest inverted lists sizes. It can be the effectiveness of the proposed model due to the nature of the random search and with dissimilar patterns.

Table 1-Experimental Result

Query	Query	Traditional		Chicken Swarm		Dissimilar Pattern		The CSO and dissimilar pattern together	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Query 2	أحكام الجنائز	0.0	0.0	0.04	0.05	0.04	0.02	0.04	0.05
Query 6	أحداث وأحكام غزوة بدر الكبرى	0.0	0.0	0.56	0.23	0.125	0.01	0.42	0.375
Query 9	الحبة السوداء	0.1	0.2	0.1	0.5	0.1	0.5	0.09	0.5
Query 12	موعد ووقت صلاة الجمعة	0.0	0.0	0.08	0.4	0.10	0.2	0.02	0.1
Query 14	حكم صلاة الضحى	0.7	0.5	0.02	0.07	0.0	0.0	0.24	0.85
Query 18	أحكام الاذان والاقامة	0.5	0.1	0.2	0.31	0.8	0.5	0.14	0.21
Query 22	موقف أبي سفيان بن حرب في فتح مكة	0.0	0.0	0.0	0.0	0.08	0.07	0.02	0.07
Query 23	نبي الله شعيب (عليه السلام)	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.125

7.Conclusion

A proposed model were presented in this paper.eproposed method uses the chicken aswarm algorithm with dissmilier patterns for Arabic data sets to improve the rank list result to user query . The chicken swarm algorithem were modified to search and find the most relevant document to user query. Dissimilar patterns were used to remove the irrelevant and noise documents. The proposed system is compared with traditional system which depend on cosine similarity only .The experimental work test and used the ZAD corpus to find the result and test the system performance. The dissimilar pattern increases the effectiveness and the chicken swarm optimization increase the efficiency.

8.Future work

For such research that aim to achieve the effectiveness of web content outlier mining through some mathematical approach and to improve the result to user through swarm algorithm through mathematical approach for all types of web documents. the research for future could be:

1- The chicken algorithm is new bio-spread algorithm and may apply most of modification to this algorithm to enhance search result to user query .

2- using another dataset for Arabic IR system.

3- By apply and modified the equations and use another approach to mining the outlier from document.

References

1. Mennatollah, M., Shaimaa, S. and Doaa, S. E**. 2017.** Improving Web Search Results by removing Outliers using Data Mining Techniques. *International Journal of Computer Applications*, (0975 – 8887) , **176**(7).
2. Zulkifel, W. R., Mustapha, N. and Mustapha, A. **2012.** Classic Term Weighting Technique for Mining Web Content Outliers. International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2012) Penang, Malaysia.
3. Bama, S. S., Ahmed, M. S. I. and Saravanan, A. **2015**. A Mathematical Approach for Mining Web Content Outliers using Term Frequency Ranking . *Indian Journal of Science and Technology*, **8**(14), DOI: 10.17485/ijst/2015/v8i14/55679.
4. Kausar, Md. A., Nasar, Md. And Sanjeev, K. S. **2013** . Information Retrieval using Soft Computing: An Overview . *International Journal of Scientific & Engineering Research*, 4(4): 388.
5. Alia, K. A. H. and Mustafa, J. H. **2016**. Sense-Based Information Retrieval Using Artificial Bee Colony Approach . *International Journal of Applied Engineering Research*, **11**(15): 8708-8713, © Research India Publications. http://www.ripublication.com.
6. Dinghui, W., Shipeng, X. and Fei, K. **2016** . Convergence Analysis and Improvement of the Chicken Swarm Optimization Algorithm . DOI 10.1109/ACCESS.2016.2604738.
7. Raheemaa, R.L. K., Ahmed, M.S. I., Riyad, A. M. **2017**. A Novel Analytical Approach for Identifying Outliers from Web Documents. *International Journal of Applied Engineering Research*, 12(22): 12156-12161, © Research India Publications. http://www.ripublication.com.
8. Jenneth, A. and Thangavel, K. **2016** . Ranking Algorithm for Documents using Clustering . *International Journal of Computer Science and Information Security (IJCSIS*), **14**(9).
9. Khushboo, B., Anju, S. and Divakar, S. **2016**. An Improved Optimized Web Page Classification using Firefly Algorithm with NB Classifier (WPCNB). *International Journal of Computer Applications*, **146**(4).
10. Alia, K. A. H. and Mustafa, J. H. **2017**. Proposed MABC-SDAIR Algorithm For Sense-Based Distributed Arabic Inormation Rertrieval . *Journal of Theoretical and Applied Information Technology*. **95**(3).
11. Drias, H. and Mosteghanemi, H. **2010** . Bees Swarm Optimization based Approach for Web Information Retrieval. IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology,1, 6-13.
12. Chiwen, Q., Shi, anZ., Yanming, F. and Wei, H. **2017** . Chicken Swarm Optimization Based on Elite Opposition-Based Learning . Hindawi Mathematical Problems in Engineering, Article ID 2734362, 20 pages https://doi.org/10.1155/2017/2734362.
13. Nursyiva, I., Aris, T. and Dian, E. W. **2017.** Chicken Swarm as a Multi Step Algorithm for Global Optimization . *International Journal of Engineering Science Invention*, **6**(1): 08-14.