ISSN: 2229-6959 (ONLINE)
DOI: 10.21917/ijsc.2017.0200

A HYBRID OPTIMIZATION TECHNIQUE FOR EFFECTIVE DOCUMENT CLUSTERING IN QUESTION ANSWERING SYSTEM

K. Karpagam[1] and A. Saradha[2]

[1]*Department of Master of Computer Applications, Dr. Mahalingam College of Engineering and Technology, India*
[2]*Department of Computer Science and Engineering, Institute of Road and Transport Technology, India*

Abstract

Today, the information is growing enormously and it is difficult and tedious task to retrieve the necessary information from that pool. The main area for retrieving relevant answers is called intelligent information retrieval. To achieve this, question and answering system is used. This question and answering plays a major role in user query processing, information retrieval and extracting related information from the information pool. Recently, number of optimization algorithms is introduced to obtain the accurate and better results. Genetic Algorithm and Cuckoo Search are nature inspired meta-heuristic optimization algorithms. In this paper, combination of Genetic Algorithm with Cuckoo Search is applied to the question and answering system. The proposed algorithm is tested with the Amazon review, Trip Advisor and 20newsgroup datasets. The results are compared with Genetic Algorithm and Cuckoo Search algorithms.

Keywords:

Document Clustering, Cuckoo Search, Genetic Algorithm, Information Retrieval, Question and Answering

1. INTRODUCTION

The best information source around the world is Internet and it grows enormous in size and it leads to difficult in manage and retrieve the relevant and desired information To overcome this, Information retrieval system has evolved and it helps to acquire the relevant information from various resources like web, datasets and repositories. To obtain the desired data from the lengthy of documents is the extensive and tiresome process but the user wants to fulfill their needs with less effort and short response time. Also, in the user's point of view, expect the exact answers than full document as the result. Recently, the researchers are working on the area of searching techniques and answer generation to make better in the intelligent information retrieval systems. The main goal of the information retrieval system is to improve the accuracy in the relevant retrieved documents and ranking the results to the user query.

The Question Answering system is one of the major application areas of information retrieval techniques. The input to the QA system is acquired as natural language question based on the user requirements and the answers are extracted from the web data source / large corpus. The QA system reduces the stress and searching time of the user, but the reliability and trust worthiness of the answer. It can be overcome from the trusted sources like bench mark datasets like Wikipedia, TREC, UCI etc. and the various search engine such as Google, Altavista, Bing etc. The Intelligent retrieval system have three modules includes question processing, information processing and information extraction. The question processing has to analysis the query given by the user and retrieves the relevant documents from the web source repositories. The information extraction starts within the documents and proceeds with the extraction of accurate answers from the paragraph within the documents.

Recently, the number of optimization algorithms are introduced for obtain the global optimum solutions. Some of the optimization algorithms are Genetic Algorithm (GA), Ant Colony Optimization (ACO), Differential Evolution (DE) and Particle Swarm Optimization (PSO), Cuckoo Search (CS), Artificial Bee Colony (ABC), etc. The optimization algorithms are plays an important role in obtain the best accurate results in the retrieved answers.

Genetic Algorithm (GA) is proposed by Holland in 1975 [1]. This algorithm maintains a population and it is encoded in the form of chromosomes. For each generation, three genetic operators' i.e. natural selection, crossover and mutation operators are applied with the current generation to obtain the new population. For each population have the fitness value depends on the objective function. Cuckoo Search (CS) is proposed by Yang and Deb in 2009 [2]. This algorithm is based on brood parasitic behavior of cuckoo species and the levy flight behavior of flies and birds. Cuckoos are fascinating birds, it makes beautiful sounds and lay their eggs in communal nests and may remove the others eggs to increase the hatching probability of their own eggs. If the host bird discovers the eggs that are not their own, it will either throw the alien eggs or simply abandon the nest and build new nest somewhere.

In this paper, the combination of Genetic Algorithm and Cuckoo Search is applied for Question and Answering system.

2. RELATED WORKS

In [3] proposed a QA system which uses the surface pattern which automatically learns to check the question and answer patterns and evaluated with the TREC 2005 and 2007 dataset to extract the answer to user questions.

The unsatisfactory results of the Boolean information retrieval model have moves the researchers to the new way of solving the issues using evolutionary algorithms. The scores of the candidate's answers are calculated by mapping the question and answer patterns types on par with the length and number of mutual keywords. In [4] discussed on the semantic similarity between mutual words was analyzed for finding the semantic relationship between the words in the question, paragraph /sentences using the standard formulas and Wordnet dictionary. In [5] suggested as to representing the information as population and select well indexes for the comparison using genetic algorithms.

In [6] Discussed using GAs with user feedback to choose weights for search terms in a query along with the population size and the attainment of improvement when combining with the PSO technique.

Civicioglu [7], Fister [8], Bhuvaneswari [9] and Gupta [10] converse about comparison about various optimization algorithms like Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms, Artificial Intelligent and all available nature Inspired Algorithms.

Alam [11] discussed the role and efficiency of *k*-means Clustering based on Cuckoo Search and Consensus Clustering for the attainment of Web Search Result and the results are compared with the cuckoo search, k-means and Bayesian Information Criterion (BIC) method proposed by [1].

Sethilnath [12] and Xin-She [2] proposed the technique for Clustering of the relevant documents one of the cuckoo search methods using Levy flight a new meta-heuristic algorithm based on the parasitic behavior of bird cuckoo species with some fruit flies.

Kartik and Dhavachelvan [13] proposed a hybrid algorithm for solving job scheduling problem using genetic and cuckoo search algorithm and result give an idea about that as the size of the problem and result in retrieval time also increases.

In [14], the author discussed the positive outcomes of applying the Hybrid Genetic and Cuckoo search Algorithm for job Scheduling which yields the result in reducing the time in allocating the job /tasks to the number of machines.

As the outcome of the survey and related works done by the various researchers, the advantages and disadvantages of algorithms have been taken for consideration and we proposed the hybrid algorithm using advantages of Genetic and cuckoo for the question answering system to make it as an intelligent interactive system.

3. GENETIC ALGORITHM

The Genetic Algorithm (GA) is a meta-heuristic optimization algorithm that is used to solve the optimization problems in nature ways. It produces the best global solution in a large search space by natural selection of ways to solve problems. The available information is representations as binary tree or directed graph of the internal representation of candidate solutions for the given problem.

The GA chromosomes are displayed as the string of bits to represent the information knowledge is explore and thought process by reaches GA to greater heights [17]. The fitness function takes the solution as the input to be optimized, compute and produces the accurate output. All the associated information's and a subset of all possible solutions related to the problem are grouped and called as population. The genetic operators such as selection cross over and mutation are applied for further processing.

4. CUCKOO SEARCH

This algorithm is based on the brood parasitism of cuckoo species which contains three idealized rules: (i) each cuckoo lays one egg at a time, and dumps it in randomly chosen nest (ii) The best nest with high quality eggs will be carried over to the next generations (iii) The number of available nests is fixed, and the egg laid by the cuckoo is discovered by the host bird with a probability pa in [0,1]. In this case, the host bird can either get rid of the egg, or simply abandon the nest and build a completely new nest.

The simple random walk is replaced with the Lévy flight behavior can be used to increase the performance of the Cuckoo Search. The following formula can describe Lévy flight behavior when generating new solutions.

$$x_i(t+1) = x_i(t) + \alpha \oplus \text{levy}(\lambda) \quad (1)$$

where, $\alpha > 0$ is the final size that has to be related to the problem of interest scale, and the product $\oplus$ refers to an entry-wise multiplication.

5. PROPOSED SYSTEM

The authors formulates the Cuckoo search breeding behavior and applied to various optimization problems [3] and their 3 categories are as follows:

i. Each cuckoo selects one sentence a time and leaves it in a randomly chosen relevance nest population for next possible answer sentence.
ii. The best population with high quality of sentence will carry to the next generations.
iii. The number of available host nest population is fixed, if a current possible sentence possibly identifies the sentence for their relevance with the probability of $pa = 0, 1$ then the low quality answers are discarded or pass to the new built generation.

The problem of mapping the user query to find the answer among the long list of documents, the search space is more complex and increases in response time. The system gets the query from the user in natural language through a interface, the question is analyzed using a POS-tagger for the question type such as evaluative question, hypothetical question, confirmative/ rhetorical question, non- factoid question from the learning model.

The learning model is trained with the set of question pairs of grammar arrangement of questions in a phrase tree format. The question is preprocessed for removing the stop word and stemming the words to extract the keywords from the question.

$$\text{Keyword Extraction: } Kw_N = \bigcup_{i=1}^{N} Ext\left(Pos\_noun\left(d_i\right)\right) \quad (2)$$

The datasets are processed by grouping the documents with respect to the related domain context on the keyword basis. The semantic meaning of the keyword is tested using wordnet Rita, an online dictionary with thesaurus.

After the processing of the dataset documents with the finding of the word sense and word order, the knowledge base is built.

The query keyword is matched with the documents in the knowledgebase and retrieves the most relevant document from the knowledgebase with the semantic similarity between them.

Similarity computation between keywords in the document is carried out by using the Eq.(3)

$$Sim(x, y) = \frac{1}{3}\left(\frac{m}{l_1} + \frac{m}{l_2} + \frac{m-n}{m}\right) \quad (3)$$

where, *m* is the matching characters, *n* is the misplaced characters, l_1, l_2 is the length of the two words. As a subsequent activity the related document to be ranked for the extraction, there is a chance of more than one document contain the keyword in group and then the normalized rank of document calculated using the formula

$$\text{Document rank} = \frac{\text{Rank of the document}}{\text{No. of retrieved documents}} \quad (4)$$

After ranking the document, learning model maintains a list of documents with rank based on the query keyword along with the occurrence of the keyword in the list of the documents.

The retrieved documents is taken for the extraction of related paragraph and sentence, the answer relevance score is calculated by using the minimum number of words between a keywords and candidate answer and string distance metrics. After obtaining the list of n sentences are reflecting as the population and apply to HGCSO algorithm for the answers.

The process flow of the proposed system is depicted in the Fig.1.

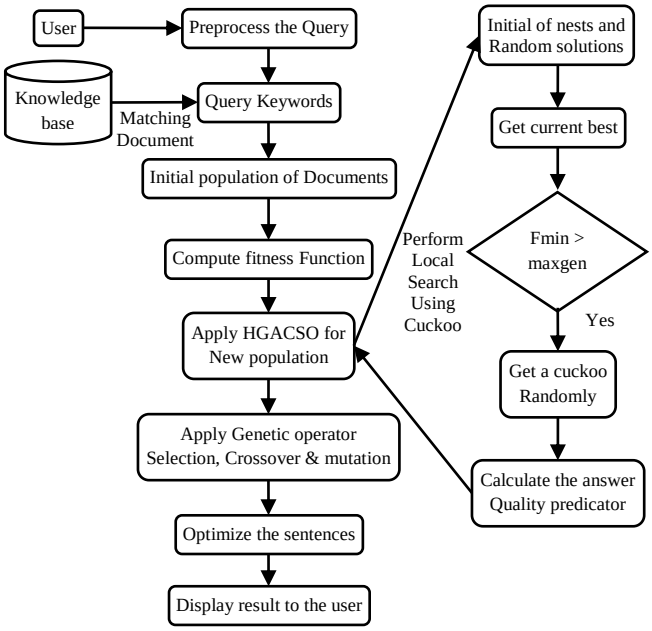

Fig.1. System Architecture of HGACSO

- User enter query into the interface of the system
- Process the query by removing the stop words and stemming of the words
- Retrieve the new list of the sentence with top most accuracy.

The proposed algorithm has combined the advantage of genetic and cuckoo search algorithm to give the better result for convergence in local and global optima.

The Disadvantage of genetic algorithm is that it can be easily trapped in local optima is overcome with the help of cuckoo search. Local optima mean it provides different results for same parameters on different runs. To overcome this difficulty of the genetic algorithm, Cuckoo Search algorithm it uses only a single parameter for searching which leads to very fast and efficient concert. The hybrid algorithm performs the process with the initial population on the needed parameters which performs the searching operation on the basis of their category with maximum generations. The matched relevant paragraph /sentences and the query keyword is the input parameters to the algorithm.

The Genetic algorithms fitness function is calculated for find the similarity between question keyword and sentences and applies the genetic operators like selection, crossover and mutation operation on the population to get the global and local best solutions with genetic diversity.

The initial generation is defined and applies the cuckoo search concept on the local search on the sentences. The bird's nest is initialized with the small set of paragraph or sentences, find the fitness value for the similarity between the sentence and the keyword. If the fitness value of cuckoo *i* is greater than the fitness value of cuckoo *j*, replace the cuckoo *j* by the new solutions. The nest with low fitness values are discarded, the last fittest solution is stored and passed to the next generation.

The genetic operator selection is used for selecting the fittest solution every time it's generated the crossover does the job of produce the new child from the populations and the mutation operator is to maintain the genetic diversity between the sentences.

The result of the first iteration is passed to the next iteration as the population, the process repeats until the global best is obtained.

Pseudo Code for Genetic Algorithm-Cuckoo Search

Define Objective function $f(x)$, $x=(x_1, x_2,\ldots,x_d)$

Initial a population of n host nests $x_i(i=1,2,\ldots,d)$, t

Define the cuckoo search parameters and genetic algorithm parameters pc, pm

While ($t < MaxGen$) or (*stop*);

$i=0$;

Generate Initial population $P(0)$

Evaluate $P(0)$ fitness

While ($t < MaxGen$) or (*stop*); do

$i=i+1$;

Select $P(i)$ from $P(i\text{-}1)$

Recombine $P(i)$ with crossover probability *pc*

Mutate $P(i)$ with mutation probability *pm*

Evaluate $P(i)$ fitness

End while

Rank the chromosomes find the current best and save

Post process results

For $i=1:n$

Get a cuckoo (say i) randomly and generate a new solution by Levy flights;

Evaluate its quality /fitness; *Fi*

Choose a nest among *n* (say *j*) randomly;

If ($Fi > Fj$)

Replace j by the new solution;

Keep the fittest solution

Forward the current best solution to the future generation.

End

6. EXPERIMENTAL RESULTS

The evaluation of the proposed method is done with the help of benchmark 20newsgroup.Amazon review and Trip advisor dataset, in which raw data of 200 documents taken as the training and test set. These datasets are considered for training and testing the hybrid algorithm for their consistency in mapping the question and retrieving the relevant candidate answers.

The 20 question set is framed and tested aligned with the language learning model proposed on each type with user and expert question. By using the POS-tagger based question pattern analysis model, the proposed system come up with the enhancement in the results of identifying the question patterns and train the system by positive and negative tags.

After the successful execution of the proposed system, we test our system with the test dataset to analysis with the help of 100 questions. The non-relevant document with different questions and non-relevant sentences are eliminated to enhance the speed up of the result to increase response time by the proposed HGACSO algorithm.

The Mean Average Precision (MAP) is one of the popular performance measures in the field of information retrieval which is used to evaluate of ranked relevant documents retrieved with the average precision values.

$$MAP = \frac{1}{n} + \sum_{Q_i} \frac{1}{R_i} \sum_{D_j \subseteq R_i} \frac{j}{r_{ij}} \quad (5)$$

where, r is the rank of the j^{th} relevant document in Q_i, n is the number of test questions and R is the relevant document for Q_i.

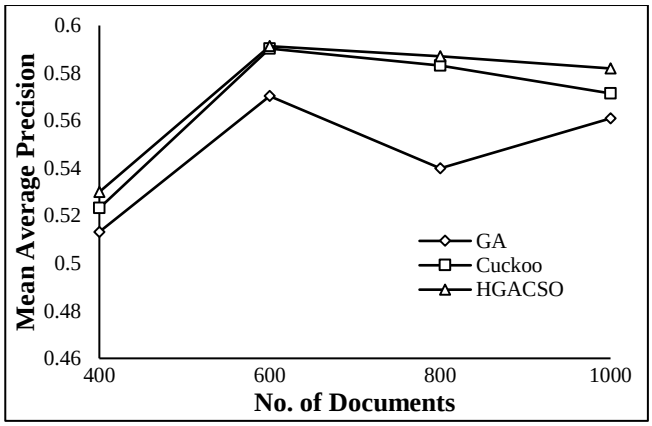

Fig.2. Mean Average Precision for the Sentence ranking for 20newsgroup

The Mean average precision is also calculated for various datasets like Amazon product review, Trip advisor with the number of documents for the given user query is exhibit in the Table.1.

The Table.2 depicts the comparison between the number of relevant documents on par with the retrieved documents in the datasets for the given test questions using HGACSO algorithm.

Table.1. Results of sentence retrieval for various datasets

No. of Documents	MAP based on Sentence retrieval		
	Amazon	Trip	20
	Review	Advisor	Newsgroup
10	0.375	0.378	0.392
50	0.378	0.389	0.395
100	0.39	0.381	0.395
100	0.385	0.4	0.413
200	0.39	0.405	0.413

Table.2. Performance analysis of HGACSO

Datasets	No. of Documents	No. of Questions (all Question types)	Relevant Documents	HGACSO Retrieved Documents
20 Newsgroup	200	30	40	39
Trip Advisor	200	30	35	32
Amazon Review	200	30	37	34

7. PERFORMANCE ANALYSIS

Lomiyets (2011) [15] and Gunnar Schröder (2011) [16] discussed in brief about the metrics used for the text mining techniques to verify the correctness of the results achieved based on the standard metrics like Precision, Recall, F-Score, Fallout and miss rates. The impact of features on the answer extractions with the considerations of the parameters analyzed are true positive, true negative, false positive, false negative, true and false positive rates.

The precision is also called as positive predictive rate is the used to find the relevant sentences from the retrieved one by relevant sentences intersect with the retrieved sentences to the retrieved sentences.

$$\text{Precision} = \frac{\text{Relevant Sentences} \cap \text{Retrieved Sentences}}{\text{Retrieved Sentences}}$$

$$\text{Precision} = \frac{tp}{tp + fp} \quad (6)$$

Recall is the used to find the relevant sentences from the retrieved one by relevant intersect with the retrieved sentences of the relevant sentences.

$$\text{Precision} = \frac{tp}{tp + fn} \quad (7)$$

The F-Score is the combination of precision and recall into a single score by calculating different types of means of both metrics.

$$F1 = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}} \quad (8)$$

Fallout or false positive rate is calculated as the ratio of relevant sentences that are irrelevant to the total number of irrelevant sentences.

$$\text{Fallout} = \frac{fp}{fp + tn} \quad (9)$$

Miss rate or false negative rate is calculated as the ratio of items not recommended but actually relevant to the total number of relevant sentences.

$$\text{Miss rate} = \frac{fn}{tp + fn} \quad (10)$$

Experimental results show that by using information from the external corpora, HGACSO models produce imperative improvements on question pattern, document clustering based on domain context classification tasks and fallout & miss out rate are decreased, especially on datasets with few or short documents.

8. CONCLUSION

In this paper, combination of Genetic Algorithm with Cuckoo Search is applied to the question and answering system. The proposed algorithm HGACSO is tested with the Amazon review, Trip Advisor and 20newsgroup datasets. The results are compared with Genetic Algorithm and Cuckoo Search algorithms. The HGACSO algorithm outperforms compared to Genetic Algorithm and Cuckoo Search optimization algorithms. By using the HGACSO algorithm, the efficiency of the answer retrieval system is improved by 10 %. The system provide the exact solution in less time, the future work is to analysis the user behavior and create a FAQ knowledge base for frequently and unanswered question.

REFERENCES

[1] John H. Holland, "*Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*", MIT Press, 1975.

[2] Xin-She Yang and Suash Deb, "Cuckoo Search via Levy Flights", *Proceedings of World Congress on Nature and Biologically Inspired Computing*, pp. 210-214, 2009.

[3] Abdessamad Echihabi, Ulf Hermjakob, Eduard Hovy, Daniel Marcu, Eric Melz and Deepak Ravichandran, "How to Select Answer String", Available at: http://www.isi.edu/natural-language/people/hovy/papers/05QAbook-answer-string-select.pdf.

[4] J. Jeon, W. Croft and J. Lee, "Finding Semantically Similar Questions based on Their Answers", *Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 617-618, 2005.

[5] P. Pathak, M. Gordon and W. Fan, "Effective Information Retrieval using Genetic Algorithms based Matching Functions Adaption", *Proceedings of 33rd Hawaii International Conference on System Sciences*, pp. 1-8, 2000.

[6] Xin-She Yang and Suash Deb, "Engineering Optimization by Cuckoo Search", *International Journal of Mathematical Modeling and Numerical Optimization*, Vol. 1, No. 4, pp. 1-17, 2010.

[7] Pinar Civicioglu and Erkan Besdok, "A Conceptual Comparison of the Cuckoo-Search, Particle Swarm Optimization, Differential Evolution and Artificial Bee Colony Algorithms", *Artificial Intelligent Reviews*, Vol. 39, No. 4, pp. 315-346, 2011.

[8] Iztok Fister Jr., Xin-She Yang, Iztok Fister, Janez Brest and Dusan Fister, "A Brief Review of Nature-Inspired Algorithms for Optimization", *Elektrotehni Ski Vestnik*, Vol. 80, No. 3, pp. 1-7, 2013.

[9] M. Bhuvaneswari, S. Hariraman, B. Anantharaj and N. Balaji, "Nature Inspired Algorithms: A Review", *International Journal of Emerging Technology in Computer Science and Electronics*, Vol. 12, No. 1, pp. 21-28, 2014.

[10] Nitisha Gupta and Sharad Sharma, "Nature-Inspired Techniques for Optimization: A Brief Review", *International Journal of Advance Research in Science and Engineering*, Vol. 5, No. 5, pp. 36-44, 2016.

[11] Mansaf Alam and Kishwar Sadaf, "Web Search Result Clustering based on Cuckoo Search and Consensus Clustering", *Indian Journal of Science and Technology*, Vol. 9, No. 15, pp. 1-18, 2016.

[12] J. Sethilnath, V. Das, S.N. Omkar and V.Maniv , "Clustering using Levy flight cuckoo search", *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 65-75, 2012.

[13] R.G. Babu Kartik and P. Dhavachelvan, "Hybrid Algorithm by the advantage of ACO and Cuckoo Search for Job Scheduling", *International Journal of Information Technology Convergence and Services*, Vol. 2, No. 4, pp. 25-34, 2012.

[14] Satyendra Singh, Jitendra Kurmi and Sudanshu Prakash Tiwari, "A Hybrid Genetic and Cuckoo Search Algorithm for Job Scheduling", *International Journal of Scientific and Research Publications*, Vol. 5, No. 6, pp. 1-4, 2015.

[15] Oleksandr Kolomiyets and Marie-Francine Moens, "A Survey on Question Answering Technology from an information Retrieval Perspective", *Information Sciences*, Vol. 181, No. 24, pp. 5412-5434, 2011.

[16] Gunnar Schroder, Maik Thiele and Wolfgang Lehne, "Setting Goals and Choosing Metrics for Recommender System Evaluations", *Proceedings of 5th ACM Conference on Dresden University of Technology Recommender Systems*, pp. 78-85, 2011

[17] Iman Khodadi and Mohammad Saniee Abadeh, "Genetic Programming-based feature Learning for Question Answering", *Information Processing and Management*, Vol. 52, No. 2, pp. 340-357, 2016.