APPLYING GENETIC ALGORITHMS TO INFORMATION RETRIEVAL USING VECTOR SPACE MODEL

Laith Mohammad Qasim Abualigah

Al-abayt University, Mafraq, Jordan

Essam S. Hanandeh

Zarqa University,Zarqa, Jordan

ABSTRACT

Genetic algorithms are usually used in information retrieval systems (IRs) to enhance the information retrieval process, and to increase the efficiency of the optimal information retrieval in order to meet the users' needs and help them find what they want exactly among the growing numbers of available information. The improvement of adaptive genetic algorithms helps to retrieve the information needed by the user accurately, reduces the retrieved relevant files and excludes irrelevant files. In this study, the researcher explored the problems embedded in this process, attempted to find solutions such as the way of choosing mutation probability and fitness function, and chose Cranfield English Corpus test collection on mathematics. Such collection was conducted by Cyrial Cleverdon and used at the University of Cranfield in 1960 containing 1400 documents, and 225 queries for simulation purposes. The researcher also used cosine similarity and jaccards to compute similarity between the query and documents, and used two proposed adaptive fitness function, mutation operators as well as adaptive crossover. The process aimed at evaluating the effectiveness of results according to the measures of precision and recall. Finally, the study concluded that we might have several improvements when using adaptive genetic algorithms.

***KEYWORDS**:*

Information Retrieval, genetic Algorithm, Adaptive genetic Algorithm, Vector Space Model, Precision, Recall

1. INTRODUCTION

Due to the increasing number of information and documents created by millions of authors and organizations on the Internet, information can be retrieved through using information retrieval system. However, users may encounter several problems during the process of information retrieval system. Thus, there were several researches tackling these problems that had an effect on the accuracy of the system in order to find the best solutions in which researchers attempted to increase the efficiency and accuracy of the system [13].

In this study, results showed some improvements in information retrieval system performance using adaptive genetic algorithms, through implementing some queries, using several methods in order to obtain relevant information, sorting such queries and ranking them depending on similarity measure [13].

DOI : 10.5121/ijcsea.2015.5102

As it should be clear now, the study aims at investigating the information retrieval models. In this study, the researcher used two models: Vector Space Model and Extended Boolean Model to compute the similarity between the query and documents [5].

The researcher also proposed two fitness functions: cosine and jaccard's, and used a variable ratio of mutation operators in order to have better results by comparing first fitness function (Cosine) with variable probability mutations and variable probability crossover, as well as comparing the second fitness function (Jaccard's) with variable probability mutations and variable probability crossover [5].

The corpus of the study consists of 1400 English documents on Mathematics and 255 queries to evaluate the effectiveness of the results according to the measures of precision and recall [4] [7].

2. INFORMATION RETRIEVAL

There are fundamental processes in information retrieval systems to get the information that meets the needs of the user by finding similarity between the query and the existing documents. In this chapter, we shall show some of the basic processes of matching the query with documents [9], depending on Vector Space Model (VSM) in which both documents and queries are represented as vectors [10].

There are several processes of information retrieval system as follows:

1- Removing punctuation marks from each document.
2- Removing stop words from each document, such as prepositions, articles and other words that appear frequently in the document without adding any meaning to it [7] [8].
3- Stemming the words using porter stemmer that is the most commonly used [8].
4- Inverted index, giving each document a unique serial number, known as the document identifier and connecting them with elements in any document (doc ID) [7].

3. GENETIC ALGORITHM

Genetic algorithm is an important technique in random searches for the best solution among a group of solutions in the available data [8].In a genetic algorithm, a population of candidate solutions to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes) which can be mutated and altered; traditionally, solutions are represented in binary system as strings of 0s and 1s, but other encodings are also possible.[2]

Genetic algorithm uses encoded representation of solutions equivalent to those chromosomes of the individuals in nature. It is assumed that a potential solution to a problem may be represented as a chromosome [7].

3.1 Proposed Fitness Function

A proposed fitness function after modification:

1- F (cosine) =

$$((2\sum(i=1)^{t}[wi,j*wi,q])/\sqrt{(2\sum(i=1)^{t}[[wi,j]^{(2)}*\sum(i=1)^{t}[wi,q]^{(2)}])}+(2\sum(i=1)^{t}[wi,q/t])/2$$

---------- (1)

- Where $w_{i,j}$ =weights term I in document j.
- $W_{i.k}$ = weights term I in query k.

The researcher modification of traditional fitness function (Cosine) was through adding new equation to traditional cosine equation as $\left(\sum_{i=1}^{t}[wi,\frac{q}{t}]\right)$ to compute average weights of all term query, adding them to the main equation and dividing them by two. This addition in the equation gave more weight to term query to get relevant documents.

2- F (jaccard's) =

$$\left(\left(\sum_{i=1}^{t}(w_{i,j} * w_{i,q})\right) / \left(\sum_{i=1}^{t}(w_{i,j})^2 + \sum_{i=1}^{t}(w_{i,q})^2 - \sum_{i=1}^{t}(w_{i,j} * w_{i,q})\right)\right) + \left(\sum_{i=1}^{t}(w_{i,q}/t)\right)/2 \quad \text{--- (2)}$$

- Where $w_{i,j}$ =weights term I in document j.
- $W_{i.k}$ = weights term I in query k.

The researcher modification of traditional fitness function (Jaccard's) was through adding new equation to traditional jaccard's equation as $\left(\sum_{i=1}^{t}(w_{i,\frac{q}{i}})\right)$ to compute average weights of all term query, adding them to the main equation and dividing them by two, this addition in equation gave more weight to term query.

3.2 Equation of crossover probability

The researcher used probability crossover (Pc) equation (3), to find the variable value (not fixed) of Crossover probability using many ratio probabilities, to get the best solution, though getting better ratio when applying crossover to chromosome, and to produce new chromosome (original offspring) by swaping some genes, to get new chromosome (original offspring) better than original chromosome [11].

$$Pc = \begin{cases} \dfrac{K_1(f_{max} - f')}{f_{max} - f_{avg}}, & f \geq f_{avg} \\ K_2, & f < f_{avg} \end{cases} \quad \text{----------(3)}$$

- Where k1 =0.5 ,k2=0.9 is crossover probabilty depending on many experiences in the system.
- Fmax = fitness function maximam in chosen chromosome.
- F' = function fitness in chromocome.
- Favg = avarge fitness function for all chromosome fitness.

3.3 Adaptive mutation

The probabilities of crossover (pc) and mutation (pm) greatly determine the degree of solution accuracy and the convergence speed that genetic algorithms can reach. In this study, the researcher got the best results by choosing some of the ratios used. So, the researcher used many ratios mutation probabilities, but we have chosen (0.2, 0.001) depending on many experiences in the system to calculate the ratios probability with fitness function, and then the more accurate class is determined for use in the future in the next research [2] [5]**.**

3.4 Equation of mutation operator probability

The researcher used probability mutation (Pm) equation (4), to find the variable value (not fixed) of mutation probability using many ratios probabilities, and get the best solution, though getting a good ratio when applying mutation to a chromosome produced from crossover as new chromium (mutated offspring) by flipping some genes,in order to get a new chromosome (mutated offspring) better than the original chromosome (original offspring) [11].

$$Pm = \begin{cases} \dfrac{K_3(f_{max} - f')}{f_{max} - f_{avg}}, & f \geq f_{avg} \\ K_4, & f < f_{avg} \end{cases} \quad \text{----------(4)}$$

4. LITERATURE REVIEW

In (2013), Wafa Zaal: In this paper, she worked on the adaptation of genetic algorithm using some models, such as Vector Space Model, the logical model and the language model. She used crossover and mutation probability variable instead of using a fixed value in traditional genetic algorithms. And she expanded both crossover and Mutation to get best results. This thesis used Arabic corpus. Her study has showed that using Vector Space Model with cosine similarity was the best solution and the improvement rate was 13.0% [10].

In (2013), Korejo and Khuhro: In this paper, the author studied adaptive mutation and proposed four operators in the genetic algorithm to determine the operator mutation in spite of the difficulty of the matter in the application process. The author proposed a solution to the problem by adapting the mutation percentage of mutation. He chose each operator mutation according to the behavior of the initial population of each generation. Finally, this study has showed that the work of adaptation mutation gave the best result at work [2].

In (2012), Ammar Sami: In this paper, he proposed a research method based on genetic algorithm to improve information retrieval system from websites online, and to apply information retrieval using a genetic algorithm to divide the work into two units, document indexing unit and genetic algorithm unit by the use of crossover, mutation operator and specialized fitness function. Finally, it obtained an improvement rate up to 90% [13].

In (2009), Noha Marwan: In this paper, she worked on the use of four different Islands in data retrieval, and used Jaccard's and Ochiai's in fitness function. She applied each measure to islands independently. She used expanded query, showed the results and compared the results of the four islands by and without the use of expansion. She pointed out that the use of expansion improved the search results in accordance with the user needs and, showed that Jaccard's was better than Ochiai's in using random selection and that Ochiai's was better than Jaccard's in using unbiased model tournament selection [14].

5. EXPERIMENTAL RESULTS

5.1 Results

In this study, the researcher used ten queries. Every query tended to have 8 statements. In other words, the researcher had 80 results. After obtaining the results, the researcher analyzed them through using evaluation criteria. The researcher chose to show the results of query number one.

- In table 1, results showed that the adaptive genetic algorithm (AGA) is used in information retrieval system (IRs) using Vector Space Model (VSM) and cosine fitness function. It showed relevant documents in order; the researcher found that the result of using cosine in table 1 was better than the result of using proposed cosine with VSM in table 2, and that recall increased when precision decreased in all cases because of the increasing number of samples.

Table 1: value of precision and Recall for query number1 by using (VSM) and (AGA) using the cosine fitness function.

Recall	Precision (%)
0.1	85
0.2	75
0.3	72
0.4	64
0.5	50
0.6	38
0.7	32
0.8	22
0.9	20

- In table 2, results showed that the adaptive genetic algorithm (AGA) is used in information retrieval system (IRs) using Vector Space Model (VSM) and cosine fitness function. It has showed relevant documents in order; the researcher found better results in the evaluation, a better degree of similarity and an average precision greater than the traditional cosine. But, this result had topped the table with a precision rate of 91% and a recall value of 0.1 because the AGA gave more weight to term query, and has better crossover and mutation probability.

Table 2: value of precision and Recall for query number1 by using (VSM) with (AGA) using the proposed cosine fitness function.

Recall	Precision (%)
0.1	91
0.2	85
0.3	79
0.4	65
0.5	54
0.6	45
0.7	39
0.8	28
0.9	22

- In table3, results shows that the adaptive genetic algorithm (AGA) is used in information retrieval system (IRs) using the Vector Space Model (VSM) and jaccard's fitness function. The researcher found better results in table 4 when using proposed jaccard’s and VSM. And, in this table, the researcher could see that recalls increased when precision decreased in all cases. But there was a bad result that had a precision rate of 80% and a recall value of 0.1 compared to the next case.

Recall	Precision (%)
0.1	80
0.2	70
0.3	60
0.4	58
0.5	43
0.6	33
0.7	29
0.8	21
0.9	19

- In table 4, results showed that the adaptive genetic algorithm (AGA) is used in information retrieval system (IRs) using the Vector Space Model (VSM) and proposed jaccard's fitness function. Here when using the proposed jaccard's and jaccard's fitness function, the researcher found better evaluation results and a better degree of similarity compared to proposed jaccard's fitness function. The researcher had a higher average precision by the use of proposed jaccard's fitness function. But, the best case among the four cases in (VSM) was when using proposed cosine.

Table 4: value of precision and Recall for query number1 by using (VSM) with (AGA) using proposed jaccard's fitness function.

Recall	Precision (%)
0.1	87
0.2	76
0.3	71
0.4	64
0.5	50
0.6	35
0.7	30
0.8	25
0.9	21

Table 5: Average value of precision for all queries using VSM with cosine

Recall	Average precision Cosine (%)	Average precision proposed cosine (%)	AGA Improvement (%)
0.1	85	92	7
0.2	76	85	9
0.3	72	80	8
0.4	65	68	3
0.5	51	55	4
0.6	39	45	6
0.7	33	38	5
0.8	23	28	5
0.9	20	23	3
average	51.5	57.1	5.6

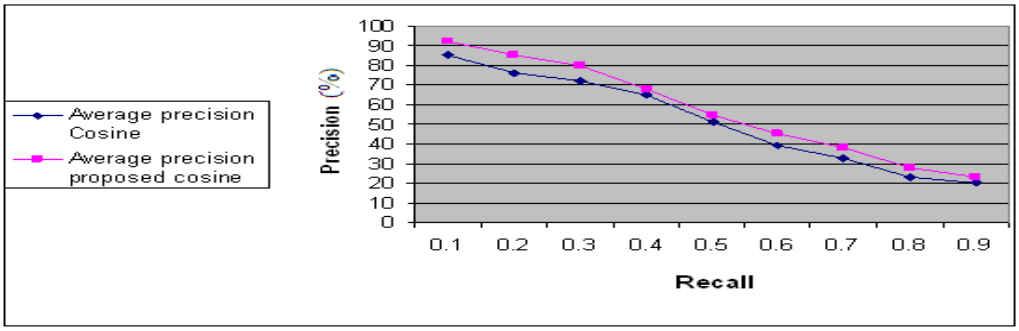

Figure 1: Average value of precision for all queries using VSM-cosine

- In this case, Vector Space Model was used with cosine and proposed cosine. As can be seen, the average of precision when using proposed cosine was greater than cosine and the precision was decreasing when the recall was increasing. Moreover, the degree of improvement was good regarding the top recall value because taking a small sample makes a precision rate and a recall value of 0.1 good for all queries. But, the recall value of 0.9 was bad. This result was good because the modification of fitness function gave weight to each term query and got better ratio probability using adaptive crossover and mutation. But as showed in figure 1, using VSM with cosine made a greater improvement than using jaccard's fitness function.

Table 6: Average value of precision for all queries using VSM with jaccard's

Recall	Average precision Jaccard's (%)	Average precision proposed Jaccard's (%)	AGA Improvement (%)
0.1	80	87	7
0.2	71	76	5
0.3	62	70	8
0.4	57	65	8
0.5	42	47	5
0.6	32	35	3
0.7	30	31	1
0.8	21	25	4
0.9	19	20	1
average	46	50.6	4.6

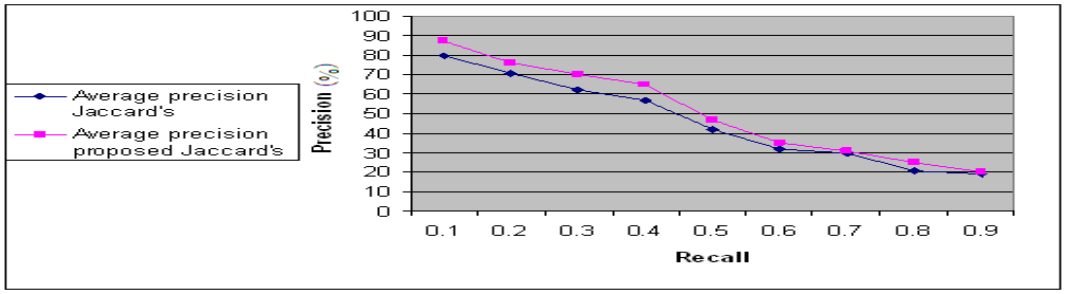

Figure 2: Average value of precision for all queries using VSM-jaccard's

- In this case, Vector Space Model with jaccard's and proposed jaccard's was used. As can be seen, the average precision, when using proposed jaccard's, was greater than when using jaccard's. Also, precision was decreasing, when recall was increasing. The degree of improvement, the top recall and the average precision rate and the recall value of 0.1 were good for all queries, whereas the recall value of 0.9 was bad. This result was good because of the modification of fitness function and the usage of adaptive crossover and mutation. But when using VSM with cosine, there was a greater improvement than when using jaccard's.

Table 7: using Vector Space Model with fitness function option.

option	Cosine	Proposed cosine	Jaccard's	Proposed Jaccard's
Average Precision(%)	51.5	57.1	46	50.6

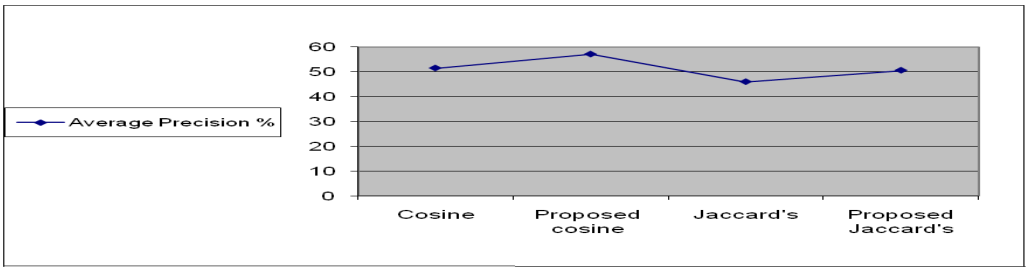

Figure 3: Using VSM with function fitness option

The researcher got the best solution when modifying fitness functions and using probability crossover and mutation operator**.**

In table 8, results showed the comparison between improvement strategies and the degree of improvements for each IR model with each fitness function.

Table 8: Compare between improvements

	Average Improvement Proposed cosine (%)	Average Improvement Proposed jaccard's (%)
VSM	5.6	4.6

6. CONCLUSION

The researcher proposed an adaptive genetic algorithm (AGA) to enhance information retrieval systems (IRs) using several fitness functions (cosine, proposed cosine, jaccard's, proposed jaccard's) alongside Vector Space Model (VSM). Therefore, the researcher concluded things as follows:

1. The best result appeared when using adaptive crossover and mutation operator with VSM-proposed cosine to retrieve relevant document with an average precision of (57.1%).
2. Having a better generation of initial population was the best result to retrieve relevant documents.
3. When using VSM-proposed cosine, the researcher saw a high degree of similarity for each relevant document.
4. The best result of using VSM appeared when using proposed cosine with an average precision of (57.1).
5. When comparing between the use of cosine with an average precision of (51.5%) and the use of proposed cosine with an average precision of (57.1%) alongside VSM, the best result appeared when using proposed cosine with an improvement degree of 5.6%.
6. When comparing between the use of jaccard's with an average precision of (46%) and the use of jaccard's addition with an average precision of (50.6%) alongside VSM, the best result appeared when using proposed jaccard's with an improvement degree of 4.6%.

7. FUTURE WORK

After finishing the study which mainly aims at improving information retrieval using genetic algorithm and adaptive, The researcher came to the following suggestions:

1. Models, such as Extended Boolean Model and Probabilistic Model, other than Vector Space Model may be used in future studies.
2. Manners other than adaptive crossover and adaptive mutation can be used in future studies.

REFERENCES

[1] Priya Borkar and Leena Patil, "Web Information Retrieval Using Genetic algorithm Particle Swarm Optimization", international Journal of Future Computer and Communication, Vol. 2, No. 6, pp 595-599, (2013).

[2] Korejo and Khuhro, "Genetic Algorithm Using an Adaptive Mutation Operator for Numerical Optimization Functions", University of Sindh, Vol.45, pp 41- 48, (2013).

[3] Taisir Eldos, "Mutative Genetic Algorithms", Journal of Computations & Modelling, Vol.3, No. 2 , pp111-124, (2013).

[4] Mohammad Nassar , Feras AL Mshagba , Eman AL mshagba ,"Improving the User Query for the Boolean Model Using Genetic Algorithms", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No. 1 , pp66-70, (2011).

[5] Imtiaz Korejo, Shengxiang Yang, and ChangheLi, "A Comparative Study of Adaptive Mutation Operators for Genetic Algorithms", The VIII Metaheuristics International Conference, (2009).

[6] Huifang Cheng , Handan, China ,"Improved Genetic Programming algorithm ", International Asia Symposium on Intelligent Interaction and Affective Computing , ieee, pp168-177, (2009).

[7] Ahmed Radwan and Bahgat Abdel Latef and Abdel Ali , Osman Sadek, "Using Genetic Algorithm to Improve Information Retrieval Systems", World Academy of Science, Engineering and Technology, pp1021-1027, (2008).

[8] Detelin Luchev, "APPLYING GENETIC ALGORITHM IN QUERY IMPROVEMENT PROBLEM", International Journal "Information Technologies and Knowledge", Vol.1,No. 1, pp309-216, (2007).

[9] NIR OREN, "Reexamining tf.idf based information retrieval with Genetic Programming", University of the Witwatersrand, paper, pp1-10, (2002).

[10] Wafa. Maitah, Mamoun. Al-Rababaa and Ghasan. Kannan, "IMPROVING THE EFFECTIVENESS OF INFORMATION RETRIEVAL SYSTEM USING ADAPTIVE GENETIC ALGORITHM", International Journal of Computer Science & Information Technology (IJCSIT), Vol 5, No. 5 , pp91-105, (2013).

[11] Sima Uyar, Sanem Sariel, and Gulsen Eryigit, "A Gene Based Adaptive Mutation Strategy for Genetic Algorithms", Istanbul Technical University, Electrical and Electronics Faculty, Department of Computer Engineering, LNCS 3103, pp 271–281, (2004).

[12] Sima Etaner and Gulsen Cebiroglu, "An Adaptive Mutation Scheme in Genetic Algorithms for Fastening the Convergence to the Optimum", Istanbul Technical University, Computer Engineering Department, (2005).

[13] Informationretrieval"http://www.dsoergel.com/NewPublications/HCIEncyclopediaIRShortEForDS, pdf, pp1-11.

[14] Kalayanasaravan and Thangamani, "Document Retrieval System using Genetic Algorithm", Kongu Engineering College, Perundurai, Vol. 2, No.10, pp 943-946, (2013).