Comparison of Crossover Types to Build Improved Queries Using Adaptive Genetic Algorithm

Khaled Almakadmeh, Assistant Professor
Department of Software Engineering
Hashemite University
P.O. Box 330136 Zarqa (13115) Jordan
Email: khaled.almakadmeh@hu.edu.jo

Wafa' Alma'aitah, Lecturer
Department of Basic Sciences
Hashemite University
P.O. Box 330136 Zarqa (13115) Jordan
Email: wafaa_maitah@hu.edu.jo

***Abstract*— this paper presents an information retrieval system that use genetic algorithm to improve information retrieval efficiency and vector space model to measure similarity between query and documents retrieved. Therefore, documents with high similarity to query retrieved first. Using the genetic algorithm, each query represented by a chromosome, these chromosomes are fed into genetic operator process: selection, crossover, and mutation until a query chromosome generated for document retrieval. The proposed approach is experimented using a data set of (242) proceedings abstracts collected from a Saudi Arabian national conference. Experimental results show that information retrieval with adaptive crossover probability set to two-point type crossover and roulette wheel as selection type yields the highest recall.**

Keywords*— *Adaptive genetic algorithm; vector space model; Cosine similarity, Crossover.

I. INTRODUCTION

A. Information Retrieval Systems

Information retrieval is a field of study that helps the user to find needed information from a collection of large documents. Retrieving information simply means finding a set of documents that is relevant to the user query [1]. A ranking of these documents is also performed in accordance to their relevance scores to the query. The user with information need issues a query to the retrieval system through the query operational module. Information retrieval systems deal with documentary bases containing textual, pictorial or vocal information and process user queries trying to allow the users to access the relevant information within an acceptable time interval [2].

An IRS consists of three basic components: documentary database, query subsystem, matching mechanism [3]. The documentary database stores documents along with representation of their information content. It is typically associated with the indexer module, which automatically generates a representation of each document by extracting the document contents [3].

The query subsystem allows the user to specify their information needs and presents the relevant documents retrieved by the system. The efficiency of an information retrieval system significantly depends upon query formation [4].

The matching mechanism evaluates the degree to which documents are relevant to user query giving a retrieval status value (RSV) for each document. The relevant document is ranked based on this value [4].

B. Vector Space Model

In vector space model, a document is viewed as a vector in n-dimensional document space (where n is the number of distinguishing terms used to describe contents of the documents in a collection) and each term represents one dimension in the document space. A query is also treated in the same way and constructed from the terms and weights provided in the user request. Document retrieval is based on the measurement of the similarity between the query and the documents [5]. This means that documents with a higher similarity to the query are judged more relevant to it and should be retrieved by the information retrieval system in a higher position in the list of retrieved documents. Using this model, the retrieved documents can be orderly presented to the user with respect to their relevance to the query [6].

C. Genetic Algorithms

A genetic algorithm is a probabilistic search algorithm, which is used for optimization of difficult problem. It is based on Darwinian principle of natural selection. It exploits and explores the document search space [7]. The basic operators used by genetic algorithm are selection, crossover and mutation. By using these operators, complex problems can be easily solved. Genetic Algorithm basic components are [8] [9]:

- Chromosome Representation: chromosomes are the initial input given to GA. All the documents and query are first converted into chromosome. This is given as input to the genetic algorithm [10].
- Fitness Function: gives a value which is used to calculate the similarity between query and document. Based on this value chromosome is selected for selection mechanism [10].
- Selection operator: selection is the process in which chromosomes are selected for next step or generation in genetic algorithm based on fitness value of chromosomes. Poor chromosome or lowest fitness chromosome selected few or not at all [11].

- Crossover operator: is one of the basic operators of Genetic algorithm. The performance of GA depends on them. In crossover two or more parent chromosomes is selected and a pair of genes are interchanging with each other [12].
- Mutation operator: is a process in which gene of the chromosome is changed. In one point mutation if gene is 0 then change it into 1 and if gene is one (1) then change it into zero (0) [12].

This paper is organized as follows: section 2 presents the review of the literature; section 3 presents the proposed adaptive genetic algorithm; section 4 presents the experimental settings; section 5 presents an analysis experimental results; and finally, section 6 presents conclusion and future work.

II. Literature Review

There are several studies that used genetic algorithm in information retrieval system to optimize the user query. An efficient adaptive genetic algorithm for vector space model was discussed by Alma'aitah and Almakadmeh [18]: this study used an Adaptive Genetic Algorithm (AGA) aimed to enhance the performance of information retrieval under Vector Space Model (VSM) in both (Cosine and Dice similarity) with used two-point crossover as the crossover operator. Using the algorithm aimed to improve the quality of the results of user's query and generate improved queries that fit searcher's needs.

Another Information Retrieval approach of using similarity function of Horng and Yeh Coefficient applied in [2]; in this paper, cosine and Horng & Yeh similarity function used to increase the efficiency of Information Retrieval System. Horng and Yeh similarity function is applied using Genetic Algorithm with set of parameters: Probability of Crossover (Pc): Pc = [0.7, 0.8, 0.9] and Probability of mutation (Pm): Pm = [0.1, 0.05, 0.01].

Query optimization using genetic algorithm in the vector space model described in [15]; this study used genetic algorithm to build improved solution for difficult problem. Different similarity measures in vector space model are used and for each similarity measure different genetic algorithm, using different crossover and mutation technique are compared.

Nassar & Al-Mashagba [19] proposed an optimization technique to optimize user query in Arabic data. To optimize query they used genetic algorithm with different fitness, different crossover and mutation technique. All this technique in Boolean model are applied.

Korejo and Khuhro [20] apply a genetic algorithm using adaptive mutation operator as numerical optimization function. In this study, authors used adaptive mutation, proposed four operators in the genetic algorithm to determine the operator mutation despite the difficulty of the matter in the application process, and then proposed a solution to the problem by adapting the mutation percentage of mutation. Each operator mutation based on the behavior of the initial population of each generation.

Comparison of selection methods and crossover operations using steady state genetic-based intrusion detection system was studied by Alabsi and Naoum [21]; the authors described a genetic algorithm to improve the information retrieval systems, the fitness function based on the frequency of words in the looked-up document.

An enhanced genetic algorithm (EGA) to reduce text dimensionality is proposed in [22] by improving the crossover and mutation operators. The crossover operation is performed based on chromosome (feature subset) partitioning with term and document frequencies of chromosome entries (features), while the mutation is performed based on the classifier performance of the original parents and feature importance. Thus, the crossover and mutation operations are performed based on useful information instead of using probability and random selection.

A novel method is proposed in [23] using hybrid of Genetic Algorithm (GA) and Back Propagation (BP) Artificial Neural Network (ANN) for learning of classification of user queries to cluster for effective Personalized Web Search. The GA- BP ANN has been trained offline for classification of input queries and user query session profiles to a specific cluster based on clustered web query sessions.

It can be observed from the literature that genetic algorithms can be applied in many information retrieval systems using genetic operators like selection, crossover and mutation. This paper will present using adaptive genetic algorithms to build improved queries for retrieved documents using different crossover types.

III. Proposed Adaptive Genetic Algorithm

A. Representation of chromosomes

The chromosome represents initial population in genetic algorithms. Each chromosome consists of several genes and each chromosome is calculated based on the number of query terms that the researcher considers ten (10) bits, which represents the number of maximum terms queries. If the term query exists in the document, it is represented as one (1). Otherwise, it is represented as zero (0).

B. Initial population

The proposed adaptive genetic algorithm populated with an initial population that consist of chromosomes correspond to the top fifteen (15) documents retrieved from traditional IR with respect to that query.

C. Fitness Function: cosine similarity

The proposed adaptive genetic algorithm used the cosine similarity as fitness function [13]. Ranking documents according to angle with query: take a document d and append it to itself. Call this document d'. d' is twice as long as d. "Semantically" d and d' have the same content. The angle between the two documents is zero, corresponding to maximal similarity. The following two notions are equivalent:

- Rank documents according to the angle between query and document in decreasing order.

- Rank documents according to cosine (query, document) in increasing order.

Cosine is a monotonically decreasing function of the angle for the interval [0, 180]

$$\frac{\sum_{k=1}^{t} (d_{ik} * q_k)}{\sqrt{\sum_{k=1}^{t} d_{ik}^2 * \sum_{k=1}^{t} q_k^2}} \quad (1)$$

Where d_{ik} is the weight of term i in document k and q_k is the weight of term i in the query.

D. Selection

Two selection methods are applied:

- Tournament selection in which for selecting each parent three-candidate chromosomes are used [14].
- Roulette Wheel weighting in which the rank weighting technique is used.

E. Crossover

Three types of crossover are applied [14] in order to compare effectiveness of each query.

- Single point
- Two point
- Uniform crossover

F. Mutation

The proposed adaptive genetic algorithm implemented as a random process. A real random number is generated in a given interval, in our case [0, 1], and that number is taken as the new value for the gene that has to mutate [15].

G. Control parameters

Crossover probability *pc* and mutation probability *Pm* play an important role in genetic algorithm. Crossover causes a randomized exchange of genetic material between chromosomes. Crossover occurs only with some probability *pc* which controls the rate at which chromosome is subjected to crossover [16, 17].

The larger value P_c is, the faster is the new chromosome introduced into the population. The smaller value P_c is, the lower the searching process is leading to stagnation. Typical initial value of P_c is in the range of [0.5-1.0]. The mutation probability P_m varied according to the generations.

The initial pm is larger for the global search, and in some generations, it is smaller for the local search. Finally, it is larger again for avoidance of local optimum. Typical initial value of pm is in the range [0.005-0.05].

We put forward adaptive varied values of P_c and P_m as follows [16] [17]:

$$p_c = \begin{cases} p_{c1} - \dfrac{(p_{c1} - p_{c2}) * (f' - f_{avg})}{f_{\max} - f_{avg}}, f' \geq f_{avg}, \\ p_{c1}, f' \prec f_{avg} \end{cases} \quad (2)$$

$$p_m = \begin{cases} p_{m1} - \dfrac{(p_{m1} - p_{m2}) * (f - f_{avg})}{f_{\max} - f_{avg}}, f' \geq f_{avg}, \\ p_{m1}, f' \prec f_{avg} \end{cases} \quad (3)$$

where, f_{max} is the maximum fitness function of current generation, f_{avg} is the average fitness function of current generation, f' is larger fitness function of the two crossover chromosomes selected, f is the fitness function of mutation chromosome selected, p_{c1}; p_{c2} is crossover probability, and p_{m1}, p_{m2} is mutation probability. The study experimental parameters include p_{c1} = 0:9; p_{c2} = 0:6, p_{m1} = 0:1; p_{m2} = 0:001.

IV. Experimental Settings

The test database is 242 proceedings abstracts (in Arabic) collected from the Saudi Arabian national conference. The study applied fifty-nine (59) queries and we choose these queries according to each query do not retrieve 15 relevant documents for our IR system. The information retrieval system built for this study using vector space model with Cosine similarity. Different AGA strategies are used in this study; the strategies are as follows:

- Strategy1:AGA1: with adaptive crossover probability and single point type crossover and roulette wheel as selection type.
- Strategy2:AGA2: with adaptive crossover probability and two point type crossover and roulette wheel as selection type.
- Strategy3-AGA3: with adaptive crossover probability and uniform type crossover and roulette wheel as selection type 3.

The termination Condition to end the generation process. If we have no sufficient improvement in two or more consecutive generations, the number of iteration used in this study is 75 iterations. The standard metrics used in evaluating the performance of an IR system are Precision, Recall, Mean Average Precision, and F-Measure. Mean Average Precision (MAP) measure is used, and for each run, the top ten documents retrieved are considered in the evaluation.

V. Experimental Results

The top fifteen (15) documents retrieved from vector space information system considered as initial population to adaptive genetic algorithm strategies as shown in Figure 1.

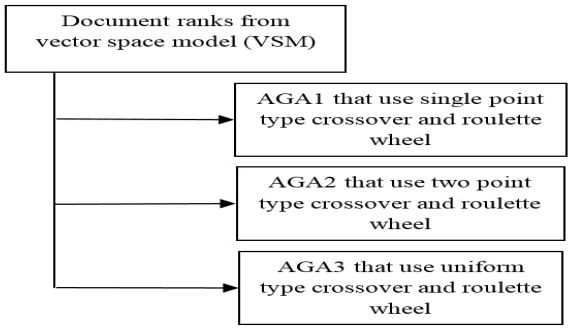

Fig. 1 Vector Space information system (VSM) with AGA strategies

The experimental results of three strategies are presented in Table 4 and in figure 2. The mean average precision is used as an evaluation measure for the three strategies. In table 4, column 1 presents the recall points that range from 0.1 to 0.9. Further, columns 2, 3, and 4 presents the corresponding mean average precision at different recall points for the three adaptive genetic algorithm (AGA) strategies. It can be noticed that the second AGA strategy (i.e. AGA-2) that uses two-point as crossover type and yields the highest mean average precision (i.e. 0.324) compared to other strategies (AGA-1) and (AGA-3) that yield mean average precision of 0.301 and 0.315 respectively.

TABLE I. MAP OF THREE DIFFERENT AGA STRATEGIES

Mean Average Precision			
Recall point	**AGA1**	**AGA2**	**AGA3**
0.1	0.136	0.144	0.134
0.2	0.187	0.191	0.191
0.3	0.214	0.298	0.288
0.4	0.244	0.282	0.242
0.5	0.363	0.376	0.376
0.6	0.372	0.384	0.384
0.7	0.387	0.389	0.388
0.8	0.400	0.421	0.401
0.9	0.412	0.432	0.433
Average	0.301	0.324	0.315

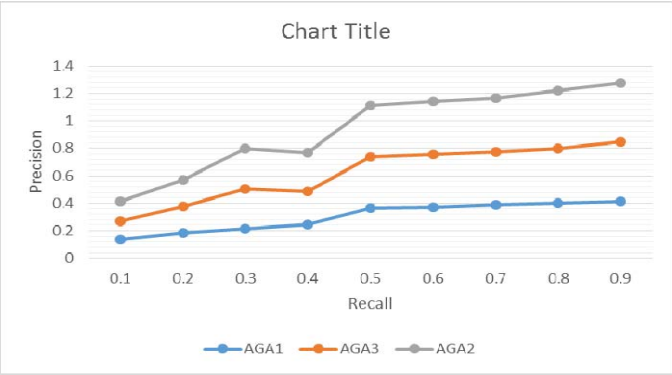

Fig 2 Average Precision curves for AGA1, AGA2 and AGA3

VI. CONCLUSION

This paper presented an information retrieval system using an adaptive genetic algorithm by representing each query by a chromosome; such a chromosome is fed into a genetic operator process to yield in an improved chromosome that improves the information retrieval process.

The proposed retrieval system tested using a data set that consists of conference proceedings' abstracts that written in Arabic. Fifty-nine (59) queries executed against this data set as proof of concept using three (3) strategies. Experimental results show that strategy-AGA2 which is built with adaptive crossover probability as two point type crossover and roulette wheel as selection type, yields better average recall than the other two strategies. Future work devoted to verify the proposed retrieval system using different types of data sets that contain larger amount of data elements in order to generalize the findings presented in this paper.

REFERENCES

[1] D. Hiemstra. Using language models for information retrieval, Ph.D. thesis, University of Twente, Netherlands, 2000.

[2] M. Chahal, J. Singh. Effective information retrieval using similarity function: horng and yeh coefficient, International Journal of Advanced Research in Computer Science & Software Engineering, Vol. 3, Issue 8, 2013.

[3] D. Grossman, O. Frieder. Information retrieval: algorithm and heuristic. Kulwar Academic Press, USA: 1998.

[4] G. Kanaan E. Hanandeh. Evaluation of different information retrieval models and different indexing methods on arabic documents, Ph.D. thesis, Arab Academy, Jordan, 2008.

[5] A. Radwan, B. Abdel Latef, A. Ali, O. Sadek. Using genetic algorithm to improve information retrieval systems, World Academy of Science Engineering and Technology Vol. 17, 2008, pp. 1307-6884

[6] A. Jain, S. Chande, P. Tiwari. Relevance of genetic algorithm strategies in query optimization in information retrieval, International Journal of Computer Science and Information Technologies, Vol. 5, 2014, pp. 5921-5927.

[7] V. Thada, V. Jaglan. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. International Journal of Innovations in Engineering and Technology, Vol. 2, No. 4, 2013, pp. 202-205.

[8] Md. Abu Kausar, S. Kumar Singh. A detailed study on information retrieval using genetic algorithm. Journal of Industrial and Intelligent Information, Vol. 1, No. 3, 2013, pp. 122-127.

[9] I. Al-Hadid, S. Afaneh, H. Al-Tarawneh, H. Al-Malahmeh. Arabic information retrieval system using the neural network model, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, No. 12, 2014, pp. 8664-8668.

[10] P. Bhatnagar, N.K. Pareek. A combined matching function based evolutionary approach for development of adaptive information retrieval system, International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 6, 2012, pp. 249-256.

[11] V. Thada, V. Jaglan. Use of genetic algorithm in web information retrieval, International Journal of Emerging Technologies in Computational and Applied Sciences, 2014, pp. 278-281.

[12] P. Mitkal. A survey on improving performance of information retrieval system using adaptive genetic algorithm, International Journal of Emerging Trends & Technology in Computer Science, Vol. 4, No. 1, 2015, pp. 205-209.

[13] K. Saravan, M. Thangamani, E. T. Venkatesh. Document retrieval system using genetic algorithm. International Journal of Scientific Engineering and Technology, Vol. 2, No.10, 2013, pp. 943-946.

[14] Y. Bade, R. Bhat, P. Borate. Optimization techniques for improving the performance of information retrieval system, International Journal of Research in Advent Technology, Vol.2, No. 2, 2014, pp 263-267.

[15] E. Al Mashagba, F. Al Mashagba. M. Nassar. Query optimization using genetic algorithms in the vector space model. International Journal of Computer Science, Vol. 8, Issue 5, No. 3, 2011, pp. 450-457.

[16] W. Lei, S. Tingzhi. An improved adaptive genetic algorithm and its application to image segmentation, Proceedings of the 5th International Conference on Artificial Neural Network and Genetic Algorithms, Vienna, 2004, pp. 112-119.

[17] M. Annunziato, S. Pizzuti. Adaptive parameterization of evolutionary algorithms driven by reproduction and competition, Proceedings of the Genetic & Evolutionary Computation Conference, Germany, 2000, pp. 1597-1598.

[18] W. Alma'aitah, K. Almakadmeh. An efficient adaptive generic algorithm for vector space model, Journal of Theoretical and Applied Information Technology, Vol.71 No.2, 2015, pp 281-286.

[19] M. Nassar, F. Al Mashagba, E Al Mashagba. Improving the user query for the boolean model using genetic algorithm, International Journal of Computer Science, Vol. 8, No. 1, 2011, pp. 66-70.

[20] Korejo and Khuhro. Genetic Algorithm Using an adaptive mutation operator for numerical optimization functions, Sindh University Research Journal, Vol.45, No. 1, 2013, pp. 41-48.

[21] F. Alabsi, R. Naoum. Comparison of selection methods and crossover operations using steady state genetic based intrusion detection system, Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 7, 2012, pp. 1053-1058.

[22] Abdullah Saeed Ghareb, Azuraliza Abu Bakar , Abdul Razak Hamdan, Hybrid feature selection based on enhanced genetic algorithm for text categorization, Expert Systems with Applications, Volume 49, 1 May 2016, Pages 31–47.

[23] Suruchi Chawla, Application of genetic algorithm and back propagation neural network for effective personalize web search-based on clustered query sessions, International Journal of Applied Evolutionary Computation (IJAEC), 2016, vol. 7, issue 1, pages 33-49.