Avoiding Premature Convergence of Genetic Algorithm in Informational Retrieval Systems

Ammar Sami Aldallal∗1

Accepted 08th October 2014

Abstract: Genetic algorithm has been adopted to implement information retrieval systems by many researchers to retrieve optimal documents set based on user query. However, GA is been critiqued by premature convergence due to falling into local optimal solution. This paper proposes a new hybrid crossover technique that speeds up the convergence while preserving high quality of the retrieved documents. The proposed technique is applied to HTML documents and evaluated using precision measure. The results show that this technique is efficient in balancing between fast convergence and high quality outcome.

Keywords: Crossover, genetic algorithm, convergence rate, information retrieval, premature convergence.

1. Introduction

Genetic Algorithm (GA) is one of the evolution based algorithms which became an important approach to solve complex problems. When optimization problem requires little knowledge about the problem that needs to be optimized, GA is one of the best adopted approaches. It is characterized by the highly parallel, random and self-adaptive algorithm which has many merits over traditional methods such as global optimization. However, GA usually has the drawbacks such as premature convergence and slow convergence speed. Premature convergence means although GA has not reached a global or satisfactory optimum, it can’t produce better offspring which outperforms its parents. When premature convergence occurs, it is difficult for GA to get rid of a local optimum and reach a global optimum [1]. Premature convergence is the main obstacle to a genetic algorithm’s practical application. In order to overcome genetic algorithm’s premature convergence, we should first have a good understanding of convergence of GA [1]. The premature convergence of a genetic algorithm arises when the genes of some high rated individuals quickly reach to dominate the population, restricting it to converge to a local optimum. In this case, the genetic operators cannot produce any more descendent better than the parents [2]. Hence, the algorithm ability to continue searching for better solutions substantially reduced.

The genetic algorithm convergence rate is been used to judge the computational time complexity of finding a global optimal solution. Therefore, it is very important to study the convergence rate of GA. Firstly, the convergence of GA must be guaranteed before analysing and evaluating solution. Based on that, the optimization efficiency of the algorithm can be judged and its results are utilized to improve the algorithm. This work will analyse the convergence rate of GA applied to information retrieval (IR) system through what is named Genetic Algorithm-based Information Retrieval (GABIR), where the outcome of this system is measured using precision measure

1.2. What is Genetic Algorithm?

Genetic algorithm is a probabilistic algorithm used to simulate the mechanism of natural selection of living organisms and it is often used to solve problems having expensive solutions. This is basically due to the principles of selection and evolution employed to produce several solutions for a given problem. Generally speaking, GA’s search space is composed of candidate solutions (chromosomes) to the problem. Each chromosome has an objective function value known as fitness value. This measure is used to favour selection of successful parents for producing new offspring. Offspring solutions are produced from parent solutions by the application of selection, crossover and mutation operators [3]. Offspring forms the second generation of possible solution. Several generations are then created by applying these operators until one of the two following criteria is satisfied. Either no more enhancements in terms of the fitness value of the entire chromosomes are achieved compared with previous generation or the maximum predefined number of generations is created. In all cases, the system will return the optimal solution from the last generation.

1.2. GA in Information retrieval

Most studies argue that IR can be seen as a standard optimization problem [4], where it has search space S represented by the set of documents D, a set of possible solutions S’ (the possible documents related to the user query), such that S’ ⊆ D and evaluation function f to evaluate the relevance of each of these possible documents related to the user query. Finally, a search engine tries to output documents that maximize f. The optimal solution is a document or set of documents that have the maximum score returned by the function f. It is found that such an optimization problem can be solved efficiently using Genetic Algorithm [5, 7]. In addition, GA requires less
processing resources compared with IR models, since there is no need to apply the searching technique to the training set before finding the optimum solution, which delay the output process. Moreover, there is no need also to evaluate all documents in the search space in order to find the optimum solution, which is computation extensive.

The rest of the paper is organized as follows: Section 2 reviews the related work done on the convergence of GA. Section three presents the proposed technique to improve the convergence rate. Section 4 describes the document set used to evaluate GABIR, while Section 5 states the evaluation measure. The experiments conducted and analyses of the obtained results are presented in Section 6. Finally, Section 7 concludes the work and suggests methods to further enhance the convergence rate.

2. Related work

Many researchers and experts have made a thorough and extensive study on convergence rate of GA. Apart from IR systems, [1] and [8] developed techniques based on Markov chain to improve the GA convergence. The approach of [1] suggests a study on calculation of the first expected hitting time based on the absorbing Markov chain. The premature convergence was avoided through self-adaptive crossover and mutation probability and close relative breeding avoidance method. [8] formulated a model for a class of genetic algorithms, which analyses the convergence rate of this class of genetic algorithms in a different way, and proved that the convergence rate is linear based on property of Markov chain. They showed that this technique is applicable to arbitrary coding, arbitrary crossover, arbitrary mutation and arbitrary selection. However, these two studies established a theoretical mathematical model of Markov chain and were lagging in providing experimental results to measure their efficiency.

The authors in [2] proposed two techniques to prevent premature convergence in GA. One of them is the dynamic application of many genetic operators based on the average progress, and the second one is the population partial re-initialization.

Generally speaking, the number of generations required to produce the final result reflects the speed of convergence. When applying GA to IR systems it is found that number of generations is mentioned as a parameter of the implemented GA models where it ranges between 20 [9] and 500 [10]. The speed of convergence is not discussed separately as a factor to enhance the system performance. Nevertheless, their researches will be considered here to highlight the speed of convergence compared with results achieved. They present the number of generations in their results either graphically [9, 11] or in a tabular format [6, 10, 11]. Represented the results graphically and the maximum number of generations was 90 for the first experiment and 60 for the second one, whereas the number of generations in [6] is 30 when examining the HTML tag weight using GA. On the other hand, [9] plotted the results of average fitness per generation and show that the maximum number of generations is 20 which somehow indicates the speed of convergence for this model. This small number reflects fast convergence but the quality of the results in terms of precision and recall are lower. Nevertheless, there is no special importance given to this factor. Therefore these figures will not reflect the actual speed of convergence but will be utilized when compared with the proposed model. [17] produces recall of 1 when the number of generation is very huge where it is 1200. While [3, 12] converges within 100 generations. Much apart from these figures is the number of generations required to divert by the approach proposed in [4] which is 12000 generations. Obviously, this gives an idea about the slow performance of such approach which aims to produce high average mean quality of the retrieved documents. However, the achieved quality for this approach was only 25%.

One of the main operators that heavily affects the convergence is crossover. In literature, many techniques of crossover have been developed and analyzed to study their effect on the output from several perspectives [13]. The most common crossover technique is one-point crossover [3, 5, 12, 14-22]. It chooses single point randomly within the chromosome and copies the values of parents 1 and 2 before or after this point to the same locations in the new offspring 1 and 2. Then, the values after or before this point are exchanged by copying them to the new offspring such that genes of parent 1 are copied to offspring 2 and that of 2 are copied to offspring 1.

Another well-known crossover techniques are two-point crossover [23, 24], and inversion crossover [25]. In one-point and two-point crossover two offspring are produced from two parents. These offspring inherit mixed properties from the two parents which may or may not perform better than the parents. Consequently, the convergence may slow down. Close to this performance is the inversion crossover. In this technique the order of genes between 2 randomly chosen positions within the chromosome is reversed. One offspring is produced from two parents in this technique; hence, its performance differs only in the arrangement of genes within the chromosome, and if the order is not important then this technique has no effect of overall performance of the chromosome and as a result, this technique may lead to premature convergence.

While many approaches are developed by researchers to enhance the convergence speed in GA, this work is featured by introducing a new crossover technique as well as adopting specific GA operator techniques that are expected to enhance convergence speed while maintaining the high performance of the IR system.

3. The GABIR approach and its effect on convergence

GA is controlled by set of operators. These operators are: selection, crossover and mutation, while the fitness function is used to evaluate each chromosome during selection and after producing new generation. The technique of implementing each operator actually influences the convergence of the GA system. This works examines several techniques of implementing each operator on IR system and studies the convergence of GA using each technique and analyses its performance.

3.1. Initial Generation Creation

When looking at the methods of creating initial generation, there is a trade-off between creating initial generation in a fast way with low quality or slower way but with high quality. Fast creation is done by selecting individuals randomly without any selection criteria.
However, this method may stick at a local optimum solution causing the results to be less effective due to fast convergence [12]. To overcome this drawback, random selective technique is applied to select individuals based on some criteria. Although this method slows down the creation of initial generation, it provides a higher probability to find optimal solutions rapidly and avoids fast convergence with local optima.

Several well-known selection techniques where proposed and heavily applied in GA such as heuristic creation operator [4] and random selection [12, 14, 16, 22, 23, 26, 27].

3.2. Parent selection

Next operator of GA is parent selection. The most popular one is simple random sampling selection also called proportional selection. It has been applied by many researches [3, 6, 14, 18, 23, 28, 34]. This method performs roulette-wheel selection, where each individual is represented by a space that proportionally corresponds to its fitness. Stochastic sampling is used to choose individuals by repeatedly spinning the roulette wheel. This method may speed up the convergence with small fraction and avoids early premature convergence since good individuals have high probability of being selected for crossover.

In tournament selection [16], a group of individuals are randomly chosen from the population. This group takes part in a tournament and an individual with highest fitness value wins. In many cases, selecting two individuals is chosen to be two, and this method is called binary tournament selection. To further enhance this selection, selecting four so each time from each 4 individuals, the best is selected.

3.3. Design of Hybrid Crossover

The proposed crossover operator to be implemented here is a combination of reordering crossover [18], fusion crossover [18] and one-point crossover. When genes within a chromosome are ordered based on their fitness value and the order is important, then the crossover applied to such chromosomes is called a reordering crossover. In fact, the order of genes in the GABIR is important as it represents the ranked documents that will be displayed to the user. If one offspring is to be produced from the crossover process rather than two, then it is called a fusion crossover [29]. Combining these two techniques together and applying a one-point crossover on them forms the proposed hybrid crossover suggested in this paper.

The hybrid crossover operates in the following manner. Suppose there are two parents x and y of length L. These two individuals are selected randomly using binary tournament selection from current population p to produce one offspring O of population p+1. Firstly, the chromosome’s genes are ordered based on their fitness value from higher to lower from the previous generation. Then a one-point crossover is applied by choosing cross point cp randomly over the range [1..L]. The selected cross point divides the chromosomes into two parts. The first O’s genes [O0, .., Ocp] are copied from the candidate parent that has the greatest gene’s value at position L0, suppose it is x in this example. The remaining genes of O are copied from the second parent starting from the leftmost position until the offspring O is filled up or until it reaches the specified location cp. Through the process of copying the remaining genes from the parents, the uniqueness of the copied gene must be considered, i.e., each gene can occur only once in the new offspring O. This is implemented by excluding the genes that already exist in O. When O is not filled up to the specified length, the fitness values of other genes in both parents are compared starting from location cp+1. The gene that has a higher fitness value contributes to O. This is done in order to generate offspring with appropriate genes from each parent and to guarantee that the length of O is maintained at L.

In the one-point crossover, GA selects one point randomly to perform exchange of genes. A reordering crossover is applied to chromosomes having their genes ordered based on their fitness value from higher to lower. The rationale behind using the ordered crossover technique over other techniques is the need to inherit the good genes and maintain the good building blocks while passing them to the resulting offspring.

In fusion crossover, only one offspring is generated from the two selected parents. In this technique, the offspring inherits the genes from one of the parents with a probability according to its performance. The advantage of this technique is that the good genes of both parents are inherited simultaneously to the offspring, producing high quality offspring and increasing the speed of convergence.

Combining the three techniques of crossover into one process expected to allow fast convergence with high quality offspring. The ordered technique gathers the good genes into one side of the chromosome. Then the one-point crossover copies these gathered genes from the heavy side of both parents to one offspring only. This results in an offspring having the best genes of both parents.

Graphical illustration of hybrid crossover is shown in Fig. 1 in which numbers in each chromosome represent the fitness value of the gene at that position. The crossover will take place between two previously selected candidates x and y (Step A). The cross point cp is selected randomly to perform a one-point crossover (Step B). In this example, cp=3. Because the first gene of x has a greater fitness value than the first gene of y, x’s genes to the left of the cross point are copied to the offspring. To complete the genes of O, y’s genes to the left of the cross point are copied to the offspring as well. Then a competition between the genes to the right of the cross point in both x and y is done to decide which parent’s genes will be copied the remaining space in the offspring. Because the gene at position cp+1 in y has a greater value than that of x’s, then y’s genes are copied into O (the right bold set of genes in step C). Once all positions in the offspring are populated with genes, these genes are ordered from higher to lower based on their fitness value (step D).

![Figure 1. Example on hybrid crossover technique](image-url)
3.4. Fitness function

Among several fitness functions developed by researcher Proximity fitness function proposed in [33] is been adopted in this work due to its high efficiency in IR systems.

4. Document Set Description

The documents that will be evaluated by GABIR can be either plain text, semi-structured (i.e., HTML (HyperText Markup Language) documents) or structured. Because most of web-documents are written in HTML [6], this format is adopted for implementing GABIR.

In similar studies, researchers tend to use ready-made data sets which use vector space indexing models such as TREC and ACAM data sets. These sets include documents, vector space index, queries and their results. However, these sets are not suitable for the proposed model because of the indexing model on the one hand, and due to the additional data that need to be included in the index which is not supported by these data sets on the other hand.

The document set or search space used in this work is a set of HTML web documents. This set is the Carnegie Mellon University data set (WebKB). It is a set of HTML documents from the departments of computer science at various universities collected in January 1997 by the World Wide Knowledge Base project of the CMU text learning group. It consists of 8284 documents [35] and used by several researches [30]. This set consists of seven categories, named: course, department, faculty, project, staff, student and others, in additional to another 60 web documents downloaded from the Web by passing different keywords to the Google search engine. Hence, the total number of HTML documents in the set is 8344. Table 1 shows the categories of the document set as well as the number of documents in each category. This document set is expected to be reasonable to analyze the proposed model since this size is in the range of document size used in similar researches. In the literature, the data set used to test most GA-based IR systems is CISI [3, 12]. This data set consists of 1460 documents and was tested against 76 to 112 documents from the departments of computer science at several universities.

Average number of relevant documents per query

When evaluating the convergence rate of GA system, there is a tradeoff between premature fast convergence, and slow convergence with high performance. Hence, in order to balance between these outcomes, the results of the proposed system are evaluated by using precision measures. Precision is defined as the percentage of relevant retrieved documents to the total number of retrieved documents.

5. Evaluation Measure

When evaluating the convergence rate of GA system, there is a tradeoff between premature fast convergence, and slow convergence with high performance. Hence, in order to balance between these outcomes, the results of the proposed system are evaluated by using precision measures. Precision is defined as the percentage of relevant retrieved documents to the total number of retrieved documents.

Table 1. Statistics of the test collection used in the proposed model

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of documents</td>
<td>8344</td>
</tr>
<tr>
<td>Number of queries</td>
<td>100</td>
</tr>
<tr>
<td>Number of unique indexed terms</td>
<td>128213</td>
</tr>
<tr>
<td>Average number of terms per query</td>
<td>2.69</td>
</tr>
<tr>
<td>Average number of relevant documents per query</td>
<td>16.82</td>
</tr>
<tr>
<td>Average number of indexed terms per document</td>
<td>410.28</td>
</tr>
</tbody>
</table>

6. Experiments and Results

In order to investigate the speed of convergence of GABIR, several experiments were conducted. In each experiment, all operators along with their parameters are fixed except for the one under consideration. These operators and parameters are listed in Table 2.

To evaluate each outcome, 10 runs are executed and the average result is considered in the analysis. Since GA consists of four operators: initial generation creation, parent selection, crossover, and mutation, the following experiments are conducted: first experiment applies the hybrid crossover on two selection methods: random selection and random selection with selective criterion. The applied selective criterion is to consider the documents that consist of the queried keywords. The second experiment applies the hybrid crossover on two parent selection methods: binary tournament selection in which the best one is always selected and the second one is binary tournament selection which favours better parent with probability ≤ 0.75. By fixing the selection method, third experiment compares between three crossover techniques. The first one is hybrid crossover. The second one is one-point crossover which has non-ordered genes and produces two offspring. The last one is a two–point crossover that produced one offspring. By adopting hybrid crossover technique, the fourth experiment utilizes different fitness functions. These fitness functions are term proximity fitness function [33], Okapi-BM25 [14] and Bayesian inference network model [6]. For details about these fitness functions, reader is recommended to review author’s previous work [33]. The average number of generations formed by GABIR was 22.26 which demonstrate its fast convergence.

Table 2. Parameter setting of GABIR

<table>
<thead>
<tr>
<th>Parameter Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>Fixed at 125</td>
</tr>
<tr>
<td>Maximum number of generations</td>
<td>50</td>
</tr>
<tr>
<td>Chromosome length</td>
<td>125</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>1</td>
</tr>
<tr>
<td>The number of best individuals copied to the next generation (Elitism)</td>
<td>1</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Meanwhile, the quality of the retrieved documents in terms of precision measure is very high compared to that of other techniques. The main reason of the combination of the low
number of generation and high precision is the way the GABIR’s hybrid crossover is implemented along with the operators’ techniques adopted. As illustrated in Table 3, the main contribution to the high convergence (smaller number of generations) is coming from the hybrid crossover technique. It converges 20.5% faster than non-ordered crossover but it was slower than 2-point crossover by 63%. However, the performance of these two crossover techniques (non-ordered crossover and 2-point crossover) was very poor in terms of precision. The convergence of 2-point crossover provides an example of premature convergence, where it converges very fast (only 13 generations) but the quality of the retrieved results in terms of precision was too low which is 40%. The reason of why hybrid crossover achieved best results is influenced by nature of the technique. It pushes the high quality genes towards the left of the chromosome by ordering the genes and then it combines and passes the best genes of both parents to one offspring.

Considering other GA operators, it is found that each operator technique has an effect on the convergence. The first effect comes from the initial selection technique. When applying selective random selection to GABIR, the speed of convergence is faster than applying pure random selection by 10.8%.

Smaller effect on convergence results from the parent selection, where binary tournament parent selection-100 is 2.8% faster than binary tournament parent selection-75. Another great improvement in convergence speed is achieved from the adopted fitness function. The term proximity fitness function was faster than both OKPI-BM25 and Bayesian network inference by 46.6% and 5.6% respectively.

**Table 3. The average convergence of each GA operator technique**

<table>
<thead>
<tr>
<th>Operator Name</th>
<th>Technique Name</th>
<th>Average Convergence</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial selection</td>
<td>Initial generation with selective criterion (selective random selection)</td>
<td>22.26</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Random selection of initial generation (pure random selection)</td>
<td>24.96</td>
<td>0.6</td>
</tr>
<tr>
<td>Parent selection</td>
<td>Binary tournament selection which favours better parent (Parent Selection-100)</td>
<td>22.26</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Binary tournament selection which favours better parent with p ≤ 0.75 (Parent Selection-75)</td>
<td>22.90</td>
<td>0.83</td>
</tr>
<tr>
<td>Crossover method</td>
<td>Hybrid crossover</td>
<td>22.26</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Non-ordered crossover representation</td>
<td>28</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Two-point crossover producing one offspring</td>
<td>13.65</td>
<td>0.4</td>
</tr>
<tr>
<td>Fitness function</td>
<td>Term proximity fitness function</td>
<td>22.26</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Okapi-BM25</td>
<td>41.66</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Bayesian inference network model</td>
<td>23.58</td>
<td>0.49</td>
</tr>
</tbody>
</table>

This paper proposed GA-based information retrieval (GABIR) system that combines the fast convergence and high performance. The key feature of this system is the proposed hybrid crossover technique. It is constructed by applying 1-point crossover to the ordered chromosome to produce one offspring which combines the best genes of both parents. This technique is applied as part of GABIR to retrieve HTML documents based on user query. Several experiments were conducted to examine the performance of GABIR. These experiments study the influence of convergence on the quality of the retrieved results in terms of precision measure.

The performance of GABIR that adopted the hybrid crossover outperforms other operators. It managed to enhance the convergence rate by up to 63% for some operators such as Okapi-BM25 fitness functions with enhancement in the quality by 55%.

To generalize the results and further demonstrate its efficiency in the IR domain, it needs to be compared with additional crossover techniques such as the uniform crossover, and need to be applied to larger document set. This work is implemented on chromosome with fixed length and can be further improved by examining the performance on variable length chromosomes in terms of precision and convergence speed.

**References**


