International Journal of
Intelligent Systems and
Applications in Engineering
ISSN:2147-6799 www.atscience.org/IJISAE
Advanced Technology and Science
Original Research Paper

Avoiding Premature Convergence of Genetic Algorithm in Informational Retrieval Systems

Ammar Sami Aldallal[*1]

Received 06th October 2014, Accepted 08th October 2014

Abstract: Genetic algorithm is been adopted to implement information retrieval systems by many researchers to retrieve optimal document 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 convergent 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]*Ahlia University – Bahrain*
**Corresponding Author: Email: aaldallal@ahlia.edu.bh*
Note:This paper has been presented at the International Conference on Advanced Technology & Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.

1.1. 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 $S^+ \subseteq D$ 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 *i* 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 *i* is chosen to be two, and this method is called *binary tournament selection*. To further enhance this selection, *i* is selected to be 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_i to produce one offspring O of population p_{i+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 [O_0, .., O_{cp}] are copied from the candidate parent that has the greatest gene's value at position L_0, 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).

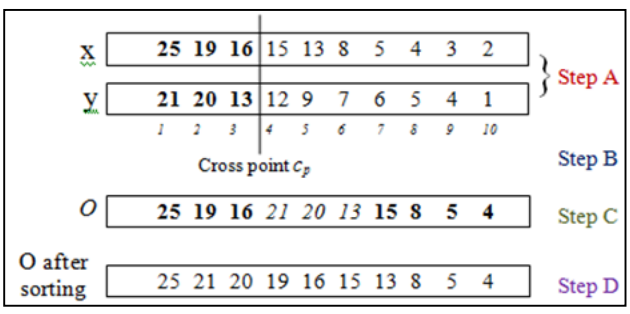

Figure 1. Example on hybrid crossover technique

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 queries. Table 1 shows some statistics for the documents and queries used to test GABIR.

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

Parameter Name	Value
Number of documents	8344
Number of queries	100
Number of unique indexed terms	128213
Average number of terms per query	2.69
Average number of relevant documents per query	16.82
Average number of indexed terms per document	410.28

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.

One of the popular measures used to evaluate the IR systems is precision at Rank N (P@N), where N is multiples of 10 [3, 11]. Rank N here means the top N ranked documents of the retrieved documents. In this method, the retrieved documents are ranked in descending order based on the fitness value (relevance to the query) and the average of precision is calculated. Therefore, this measure evaluates the system based on the percentage of the total retrieved documents.

When the maximum value of N is 100, this measure is called 11-point average precision [31, 32] and it is widely used to evaluate IR models, since it measures the performance at the points 0, 10, 20, 30 up to 100 top ranked retrieved documents, where point 0 means the first retrieved document or the top ranked document. However, this work applies the measure P@10 for seek of assessing the convergence rate rather than assessing the overall retrieved documents.

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 $\leq$ 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

Parameter Description	Value
Population size	Fixed at 125
Maximum number of generations	50
Chromosome length	125
Crossover rate	1
The number of best individuals copied to the next generation (Elitism)	1
Mutation rate	0.7

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

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

7. Conclusion

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

[1] J. Jing and M. Lidong, "The Strategy of Improving Convergence of Genetic Algorithm'. TELKOMNIKA, Vol.10, No.8, December 2012, pp. 2063~2068

[2] E. S. Nicoară "Mechanisms to Avoid the Premature Convergence of Genetic Algorithms." Petroleum-Gas University of Ploiesti Bulletin, Mathematics-Informatics-Physics Series 61.1 (2009).

[3] A. A. Radwan, B. A. Abdel Latef, A. A. Ali, and O. A. Sadeq, 'Using genetic algorithm to improve information retrieval systems. Proceedings of world academy of science, engineering and technology", 2006, vol. 17, pp. 6-12.

[4] M. H. Marghny and A. F. Ali, "Web mining based on genetic algorithm". AIML 05 Conference. Cicc, Cairo, Egypt. 2005.

[5] B. Klabbankoh and O. Pinngern, "Applied Genetic Algorithms in Information Retrieval". Retrieved Aug 22, 2009, from http://www.ils.unc.edu/~losee/gene1.pdf

[6] S. Kim and B-T. Zhang, "Genetic mining of html structures for effective web-document retrieval". Applied Intelligence, 2003, vol.18, no.3, pp.243-256.

[7] S. A. Kazarlis, S. E. Papadakis, J. B. Theocharis and V. Petridis, "Microgenetic algorithms as generalized hill-climbing operators for GA optimization," IEEE Trans. Evol. Comput., vol.5, pp.204-217, Jun. 2001.

[8] L. Ming, Y. Wang, and Y. M. Cheung, "On convergence rate of a class of genetic algorithms". In Automation Congress, 2006. WAC'06. World (pp. 1-6). IEEE.

[9] S. Kim, B.-T. Zhang, "Web-Document Retrieval by Genetic Learning of Importance Factors for HTML Tags". In Proceedings of PRICAI Workshop on Text and Web Mining'2000. pp.13~23

[10] D. Vrajitoru, "Natural Selection and Mating Constraints with Genetic Algorithms". To appear in the International Journal of Modeling and Simulation. 2007

[11] R. M. Losee, "Learning Syntactic Rules and Tags with Genetic Algorithms for Information Retrieval and Filtering: An Empirical Basis for Grammatical Rules", Information Processing & Management, 1996, 32(2), pp. 185-197.

[12] A. Aly, "Applying genetic algorithm in query improvement problem". Information Technologies and Knowledge, 2007, vol.1, pp. 309-316.

[13] A. Al-Dallal, "The Effect of Hybrid Crossover Technique on Enhancing Recall and Precision in Information Retrieval", Proceedings of The World Congress on Engineering 2013, Vol. III, WCE 2013, 3 - 5 July, pp1571-1576, London, UK.

[14] C. Lopez-Pujalte, V. P. Guerrero-Bote and F. de Moya-Anegon, "Genetic algorithms in relevance feedback: a second test and new contributions". Information Processing and Management, 2003, vol. 39, pp. 669–687.

[15] W. Song, and S. C. Park, "Genetic algorithm for text clustering based on latent semantic indexing". Computers and Mathematics with Applications, 2009, vol. 57, no.11, pp. 1901-1907.

[16] J.-Y. Yeh, J.-Y. Lin, H.-R. Keyword and W.-P. Yang, "Learning to rank for information retrieval using genetic programming". In Proceedings of ACM SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR '07), pp. 41-48. Amsterdam, Netherlands.

[17] D. Húsek, V. Snášel, J. Owais, and P. Krömer, "Using genetic algorithms for Boolean queries optimization". Proceeding of the Ninth IASTED International Conference internet and multimedia systems and applications, 2005, pp. 178-184. Honolulu, Hawaii, USA.

[18] D. Vrajitoru, "Large population or many generations for genetic algorithms? Implications in information retrieval". In F. Crestani, and G. Pasi (Ed.), Soft Computing in Information Retrieval. Techniques and Applications, 2000, pp. 199-222. Physica-Verlag, Heidelberg.

[19] G. Desjardins, R. Godin and R. A. Proulx, "Genetic algorithm for text mining". Proceedings of the 6th international conference on data mining, text mining and their business applications, 2005, vol. 35, pp. 133-142.

[20] J. Carroll and T. Lee, "A genetic algorithm for segmentation and information retrieval of SEC regulatory filings", *Proceedings of the 2008 international conference on Digital government research*, Publisher: Digital Government Society of North America.

[21] P. Simon, and S.S. Sathya, "Genetic algorithm for information retrieval", *International Conference on Intelligent Agent & Multi-Agent Systems. IAMA 2009*. pp. 1 – 6, IEEE Conference Publications.

[22] H. Drias, I. Khennak, and A. Boukhedra, "A hybrid genetic algorithm for large scale information retrieval", *International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009*. IEEE vol: 1, pp. 842 - 846 IEEE Conference Publications

[23] P. Pathak, M. Gordon, and W. Fan, "Effective information retrieval using genetic algorithms based matching functions adaption". 33rd hawaii international conference on science (HICS). Hawaii, USA. 2000.

[24] W. M. Spears, and K. A. De Jong, "An analysis of multipoint crossover", in Foundations of Genetic Algorithms, G. Rawlins, Ed. San Mateo, CA: Morgan Kaufman, 1991, pp. 301–315.

[25] D. E. Goldberg, "Genetic algorithms in search, optimization, and machine learning". Addison-Wesley, 1989.

[26] D. Beasley, D. R. Bull, R. R. Martin, "An Overview of Genetic Algorithms: Part 1", Fundamentals University Computing 15 (2), 58-69, 1993.

[27] X. Zhang, K. Wei, and X. Meng, "A XML query results ranking approach based on probabilistic information retrieval model", *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012*, pp.: 915 – 919. IEEE Conference Publications

[28] H. M. Pandey, A. Dixitand, and D. Mehrotra "Genetic algorithms: concepts, issues and a case study of grammar induction", September 2012 , *CUBE '12: Proceedings of the CUBE International Information Technology Conference*

[29] D. Vrajitoru: "Crossover Improvement for the Genetic Algorithm in Information Retrieval". Information Processing and Management, 1998, 34(4), 405-415.

[30] H. Dong, F. K. Hussain, E. and Chang, E. "A survey in traditional information retrieval models". Second IEEE International conference on digital ecosystems and technologies, 2008, pp. 397 - 402.

[31] S.M. Alzahrani and N. Salim, "On the use of fuzzy information retrieval for gauging similarity of Arabic documents", Applications of Digital Information and Web Technologies, ICADIWT '09. Second International Conference on the Digital Object, 2009, pp.: 539 – 544. IEEE Conference Publications.

[32] Manning, C. D., Raghavan, P., and Schütze, H. "An introduction to information retrieval". Cambridge, England: Cambridge University Press, 2009.

[33] A. Al-Dallal, R. S. Abdul-Wahab, "Achieving High Recall and Precision with HTLM Documents: An Innovation Approach in Information Retrieval", Proceedings of the World Congress on Engineering 2011 Vol. III. pp1883-1888, WCE 2011, 6 - 8 July, 2011, London, U.K.

[34] M. Saini, D. Sharma, P. K. Gupta, "Enhancing information retrieval efficiency using semantic-based-combined-similarity-measure". International Conference on Image Information Processing (ICIIP), 2011, pp. 1 - 4. IEEE Conference Publications.

[35] The 4 Universities Data Set. [online]. Available at: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/ [Accessed 12/11/2009]