# Application of Genetic Algorithms for Effective Choice of Information

Hristo I. Toshev[1], Stefan L. Koynov[2] and Chavdar D. Korsemov[3]

*Abstract*: - **The application of genetic algorithms for an effective choice of information sources is introduced. The object is a combined GA as a probabilistic choice of information sources in the search of quasioptimal solutions. The general structure of the algorithm and its operation is presented. The comparative research under equal conditions and the presented graphs show that the increase of the problem dimensionality lead to a nonlinear improvement related to the method of the random search and also that it is better than a previously cited algorithm. This contributes to the more effective solution of the problem.**

*Keywords* – **genetic algorithms, optimization, random search, selection, crossover, mutation.**

## I. INTRODUCTION

Genetic algorithms (GA)are a method for search based on the selection of the best species in the population in analogy to the theory of evolution of Ch. Darwin.

Their origin is based on the model of biological evolution and the methods of random search. From the bibliographical sources [1]-[4] it is evident that the random search appeared as a realization of the simplest evolutionary model when the random mutations are modeled during random phases of searching the optimal solution and the selection is modeled as "removal" of the unfeasible versions.

The main goal of GA-s is twofold:

- abstract and formal explanation of the adaptation processes in evolutionary systems;

- modelling natural evolutionary processes for efficient solution of determined class of optimization and other problems.

The continuously growing number of publications and also of the practical implementations during the last years is a stable proof of the growing expansion of the scientific and application research in the domain of GA

In order to give a general fancy for the type of applications, they could be classified in four main directions [5]: science, engineering, industry and various other directions (miscellaneous applications). Some specific areas inside any of these directions are discussed below.

Scientific applications [5]-[9] – of chemical, analysis of spectroscopy, medical image reconstruction [8], computer aided diagnosis, machine-learning in highly dimensional data, the analysis of promoters in biological sequences in the problem to deal with [9] etc.

[1] Hristo I. Toshev, [2] Stefan L. Koynov and [3] Chavdar D. Korsemov are with the Institute of Information Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev str., bl. 29A, 1113 Sofia, Bulgaria, E-mail: toshev@iinf.bas.bg, , slk@iinf.bas.bg, chkorsemov@iinf.bas.bg

Engineering applications [4], [5], [10], [11] – electrical, hydraulic, structural, aeronautical, robotics and control etc.

Industrial applications [4], [5], [12]-[14] – design, manufacture, scheduling, management etc.

Miscellaneous applications [5], [15], [16] – problem of attribute selection in data mining, decisions support system, finance, optimization a forecast model, forest management etc.

We propose in the rest of the paper the usage of the GA for an effective choice of information sources.

## II. PROBLEM FORMULATION

The modern practice becomes more and more bounded to the process of solving different search problems of exactly defined information from huge data bases (DBs), its representation and visualization included. There is a certain number of concepts that have been well-developed and which offer tools to solve the above-introduced problems. Some of the are:

KRAFT – Knowledge Reuse and Fusion / Transformation [17]. The main aim of this project is to enable sharing and reuse of constraints embedded in heterogeneous databases and knowledge systems. It has a hierarchy of shared ontologies for local resource ontology translation.

OBSERVER – Ontology Based System Enhanced with Relationship for Vocabulary heterogeneity Resolution [18] is a system for information retrieving from information sources (IS). The main aim is to retrieve information from heterogeneous data bases without knowledge of their structure, location and existence of the requested information.

EXPECT [19] is a framework for knowledge based systems developed to support knowledge acquisition and explanation.

Disciple-RKF [20] is aimed at development and experimental validation of a collaborative assistant for rapid data basis formation and reasoning to enable a team of subject matter experts that do not have prior knowledge engineering experience, to rapidly construct, update and extend a high quality integrated base for a complex application.

ODM - Ontology-Driven Methodology [21] Smirnov's approach is designed as a combination of the discussed systems. It is an integrated structure for choices of information sources.

The idea of such problems is the choice of certain concepts which shall be applied in the search and the processing of users' requests in the most effective way and according to predefined criteria – costs, time, etc.

The formulation of such problems requires to define the components of their complex elements [21] (applied ontology

– AO, information sources – IS, IS ontologies, user requests, requests' ontologies, etc.). Different relations among the set of elements are established below, the goal-function definition included.

AO *A* contains some ontology elements (OE – $\{a_j\}$), i.e. classes *O*, attributes *Q*, domains *D*, and constraints *C* of the application domain.

$$A = (O,Q,D,C) = \{a_j\},\ j = 1,...,n \quad (1)$$

where *n* is the number of OEs.

IS $S_i$ contains some OEs $\{s_{lit}\}$ at a time instant *t*. Besides OEs, IS contains instances (information content *I*), i.e. it constitutes a constraint network $CNet(S_i)$:

$$CNet(S_{it}) = (O(S_{it}), Q(S_{it}), D(S_{it}), C(S_{it}), I(S_{it})) = \{s_{lit}\},\ i = 1,...,m,\ t = 1,...,T,\ l = 1,...,p_i \quad (2)$$

where *m* is the number of ISs in the system, *T* is the system life time, and $p_i$ is the number of OEs of $IS_i$.

Information map associates OEs of ISs with those of AO at a time instant *t*. Such association is denoted by a symbol "$\rightarrow$", and a statement "OE $a_j$ is associated with IS $S_{it}$" is denoted by $(a_j \rightarrow S_{it})$:

$$IM_t = \{(a_j \rightarrow S_{it})\},\ a_j \in A \quad (3)$$

It is considered that for each IS its parameters such as costs, availability, access time, on-line schedule, etc. are known. IS ontology will be defined as an association of IS' elements with AO's elements:

$$A(S_{it}) = \{(a_j \rightarrow s_{lit})\} \quad (4)$$

When a user request *R* is received by the system it is decomposed into a set of subrequests $r_k$, which then are associated with the AO's OEs (i.e. translated into the system's terms). This association is contained in the request ontology *A*(*R*). When these operations are completed the request translated and decomposed into subrequests associated with the AO's OEs will be obtained (denoted by R′):

$$R = \{r_k\} \quad (5)$$

$$A(R) = \{(r_k \rightarrow a_j)\},\ r_k \in R,\ a_j \in A \quad (6)$$

$$R' = \{a_j\},\ \forall a_j \in R' : \exists (r_k \rightarrow a_j) \in A(R) \quad (7)$$

When the operations above are completed a set of feasible decisions of the task $Dec_R$ can be written as:

$$Dec_R = \{dec_R\} dec_R = \{(r_k \rightarrow s_{lit})\} \quad (8)$$

Costs *Cost* , time *Time* and reliability *Reli* required for request processing can be used as criteria of the decision's effectiveness:

$$Cost = f_{cost}(dec_R) \sum_{s_{jik} \in dec_R} f_{Cost}(s_{lik}) \quad (9)$$

$$Time = f_{Time}(dec_R) \quad (10)$$

$$Reli = f_{Reli}(dec_R) \sum_{s_{jik} \in dec_R} f_{Reli}(s_{lik}) \quad (11)$$

Also, an overall index of effectiveness *Eff* including estimations of both costs and time can be considered (multicriteria optimization). For instance, normalized values of cost and time (superscript *N*) functions can be summarized using weights $w_{Cost}$ , $w_{Time}$ and $w_{Reli}$

$$\begin{aligned} Eff &= f_{Eff}(dec_R) = \\ &= f'_{Eff}(f_{Cost}(dec_R), f_{Time}(dec_R), f_{Reli}(dec_R)) = \\ &= w_{Cost} \cdot f^N_{Cost}(dec_R) + w_{Time} \cdot f^N_{Time}(dec_R) + \\ &+ w_{Reli} \cdot f^N_{Reli}(dec_R) \\ &w_{Cost} + w_{Time} + w_{Reli} = 1 \end{aligned} \quad (12)$$

Decision is considered effective (denoted by $dec_R^{eff}$ ) if the value of goal function, e.g., (11), is minimal with the constraints (1 - 8) being true:

$$dec_R^{eff} \in Dec_R,$$
$$\forall dec_R \in Dec_R,\ f_{Eff}(dec_R^{eff}) \le f_{Eff}(dec_R) \quad (13)$$

## III. The Usage of Genetic Algorithms

Genetic algorithms are intended for searching a space of possible solutions to identify the best one. The "best" solution is defined as the one optimizing a predefined numerical measure called the solution *fitness*. Although different implementations of the genetic algorithms vary in their implementation details, they usually share the following structure [22].

The algorithm works by iteratively updating a set of possible solutions, called population. On each iteration, all members of the population are evaluated according to the fitness function.

. A new population is generated by probabilistically selecting the most fit individuals from the current population.

Some of the selected individuals are carried forward into the next population intact. Others are used as the basis for creating new offspring individuals by applying genetic operations such as crossover and mutation.

| | $IS_1$ | … | $IS_i$ | … | $IS_m$ |
|---|---|---|---|---|---|
| $OE_1$ | $dec^R_{1,1}$ | … | $dec^R_{1,i}$ | … | $dec^R_{1,m}$ |
| … | … | … | … | … | … |
| $OE_k$ | $dec^R_{k,1}$ | … | $dec^R_{k,i}$ | … | $dec^R_{k,m}$ |
| … | … | … | … | … | … |
| $OE_n$ | $dec^R_{n,1}$ | … | $dec^R_{n,i}$ | … | $dec^R_{n,m}$ |

Fig. 1. Structure of a feasible decision used in GA

An application of the genetic algorithm (GA) is proposed for the solution of the above-defined task. It is effective for problems of similar nature. A feasible static decision $dec_R$ represents a chromosome and has the following structure:

$$dec_R = \{dec^R_{k,i}\} \quad (14)$$

where each $dec^R_{k,i}$ is a Boolean variable equal to 1 if $IS_i$ is used for obtaining $OE_k$ or to 0 otherwise

Hence, $dec_R$ represents a binary matrix (Fig. 1), whose rows are considered as genes for GA.

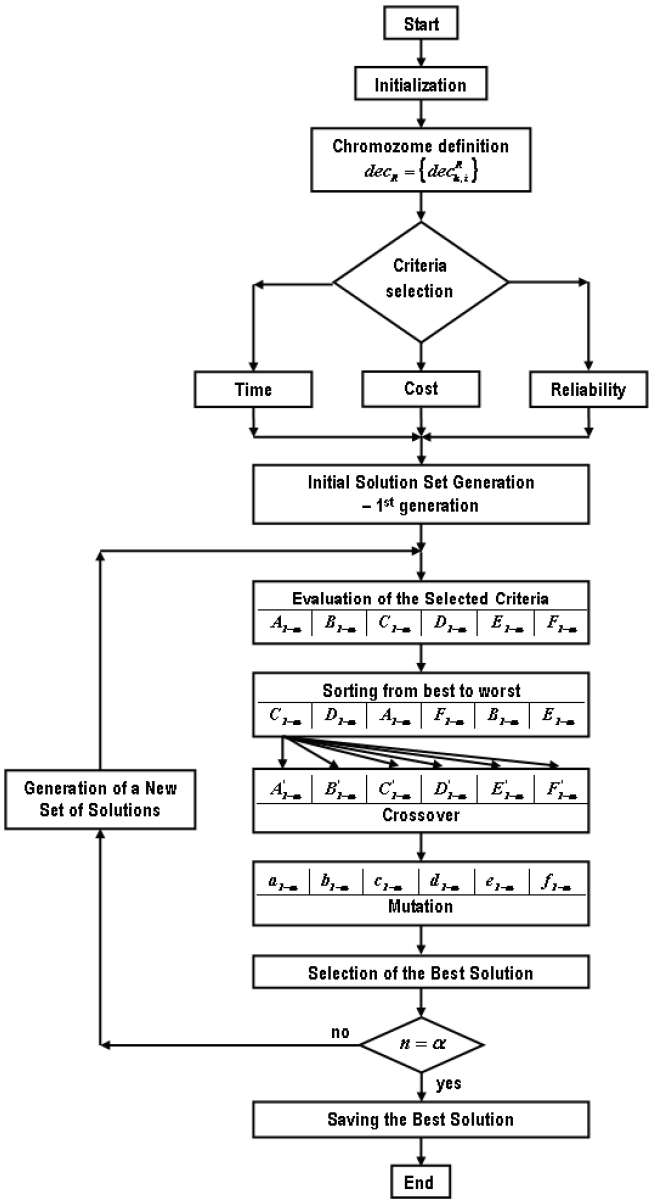


Fig. 2

The solution of the already postulated problem is via a new GA which is created on the basis of a combination of elements from algorithms of Gen [23], Falkenauer [24] and Goldberg [25] as a probabilistic approach to quasioptimal solutions, using certain parts of the algorithms, above mentioned and we have also added some supplementary elements, that allow larger choice of the criteria and better selection after the population accomplished, which leads to decrease in number of the necessary computations.. The general structure of the algorithm is presented in Fig. 2. After the starting initialization a definition of a chromosome with the following structure is defined: $dec_R = \{dec^R_{k,i}\}$, where each $dec^R_{k,i}$ is a Boolean variable equal to 1 iff $IS_i$ is used to obtain $OE_k$ or to 0 otherwise. Then a choice is made of the used criteria (Time, Cost, Reli).

First, a random set of solutions (1st generation) is generated. Then the solutions are estimated according to the selected criteria (fitting). The next step is sorting of the solutions (ranking from the best to the worst). Now mechanisms of crossover and mutation are applied for the best solution to generate new solutions. A selection of the best solution is done followed by a verification whether the defined number of populations is reached. Otherwise a random set of solutions is generated and the loop is repeated over and over till the defined number of populations is achieved. After that the best solution is saved.

A series of tests of the discussed GA (Fig. 2) were performed about the choice of information sources. They were based on the algorithm for problems with different dimensions. The values of the different parameters were randomly generated. Some actual problems have been solved, using six types of data bases as information sources. At that the number of the computations needed has been found at the presence of 2, 3, 4, 5 and 6 data bases.

The obtained results are shown in Table 1 with the respective number of information sources and the number of iterations for achieving the effective solution. They are visualized in Fig. 3 as the graph GA2.

Also comparative research was done under the same conditions using the methods of random search (MRS) and also using GA1, the older GA [26] applied by [21]. The obtained data are shown in Table I (MRS, GA1 and GA2), the graphs are in Fig. 3.

TABLE I

| Information Sources (IS) | Number of Computations (NC) | | |
|---|---|---|---|
| | MRS | GA1 | GA2 |
| 2 | 24 | 20 | 17 |
| 3 | 56 | 40 | 35 |
| 4 | > 100 | 71 | 63 |
| 5 | > 100 | 88 | 75 |
| 6 | > 100 | 92 | 80 |

## IV. CONCLUSION

The paper deals with the applicability of GAs under effective choice of information sources.

The results show that the introduced algorithm for the choice of information sources has much better indicators than the one of the method of random search and even better than GA1. This is due to the significantly smaller number of necessary computations to obtain the quasieffective solution compared to MRS; related to GA1 it is also better, due to the new additional elements in GA2, which enable better selection after the population realized and faster discovery of the solutions searched for. From the graphs in Fig. 3 it is evident that the increase in the dimensionality of the problem leads to a nonlinear improvement compared with MRS and with respect to GA1 there is also a certain improvement.

The conclusion is that the application of GA2 to choose information sources leads to better results thus showing that the usage of new GA-based approaches contributes to the more effective solution of certain problems.

The research in the domain can continue in pursuit of new

and better GAs for the solution of problems with complicated | more complex structures and a bigger volume like the multiobjective optimization problems.

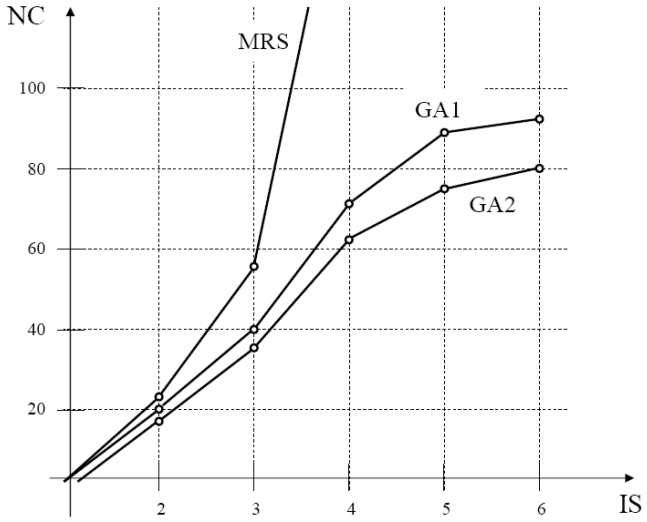


Fig. 3

## REFERENCES


[1] J. H. Holland, *"Adaptation in Natural and Artificial Systems"*, the MIT Press, 1992.

[2] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.

[3] M. Mitchell, *"An Introduction to Genetic Algorithms"*, MA: MIT Press, 1996.

[4] V. V. Emelianov, V. M. Kureichik, V. V. Kureichik, *"Theory and Practice of Evolutionary Modelling"*, Moscow, 2003. (in Russian)

[5] C. A. Coello Coello and G. B. Lamont, "Applications of Multi Objective Evolutionary Algorithms", Vol. 1, World Scientific, 2004, 761 pp. Hardcover, ISBN: 981-256-106-4.

[6] V. Guliashki, (2002) "Parallel Genetic Algorithm PGAmod, solving integer optimization problems", Proceedings of the International Scientific Conference on "Basic Technologies for E-Business'2002", September, 15-18, 2002, Albena, Bulgaria, pp. 272-277.

[7] R.M. Hubley, E. Zizler, and J.C, Roach, "Evolutionary algorithms for die selection of single nucleotide polymorphisms," UMC lliontfo: mattes, vol. 4, no. 30, July 2003.

[8] M. Lahanas, "Application of multiobjective evolutionarv optimization algorithms in medicine," In Carlos A. Coello Coello and Gary B. Lnmont, editors, Applications of Multi-Objective Evolutionary Algorithms, pp, 365-391. World Scientific. Singapore, 2004.

[9] R.S. Rosenberg. "Simulation of genetic populations with biochemical properties," Ph.D. thesis. University of Michigan, Ann Harbor, Michigan, 1967.

[10] D. B. Fogel, "George Friedman-Evolving Circuits for Robots", IEEE Computentional Intelligence Magazine, November 2006, pp. 52-54, 2006.

[11] V. Guliashki, Burdiek B., Mathis W., (2004) "Optimization of Test Signals for Analog Circuits", *Electronics*, ISSN 1450-5843, Vol. 8, № 1, May 2004, pp. 10-13.

[12] T. Kipouros, D. Jaeggi. B. Danes, G. Parks, and M. Savill, "Multi-objective optimization of turbomachinery blades using tabu search." In Carlos A. Coello Coello et al,, editor, Evolutionary Multi-Criterion Optimization. Third International Conference, EMO 2005. pp. 897-910. Guanajuato, Mexico. Springer. Lecture Notes in Computer Science Vol. 3410, Mar. 2005.

[13] A. Molina-Cristobal, LA. Griffin, P.J. Fleming, and D.H. Owens, "Multiobjective controller design: Optimising controller structure with genetic algorithms," In Proceedings of the 2005 IFAC World Congress on Automatic Control, Prague, Czech Republic, July 2005.

[14] M. Nicolini, "A two-level evolutionary approach to multi-criterion optimization of water supply systems," In Carlos A. Coello Coello et al, editor, Evolutionary Multi-Criterion Optimization. Third International Conference, EMO 2005, pp. 736-751, Guanajuato, Mexico, Springer, Lecture Notes in Computer Science Vol. 3410, Mar. 2005.

[15] F. Schlottmann and D. Seese. "Financial applications of multi-objective evolutionary algorithms: Recent developments and future research directions." In Carlos A. Coello Coello and Gary B. Lamont, editors. Applications of Multi-Objective Evolutionary Algorithms, pp. 627-652. World Scientific, Singapore. 2004

[16] T. Hanne and S. Nickel, "A multiobjective evolutionary algorithm for scheduling and inspection planning in software development projects." European journal of Operational Research, vol. 167, no. 3, pp. 663-678, Dec. 2005.

[17] P.R.S. Visser, D.M. Jones, M.D. Beer, T.J.M. Bench-Capon, B.M. Diaz and M.J.R. Shave, 1999. Resolving Ontological Heterogeneity in the KRAFT Project. Proceeding of the International Conference On Database and Expert System Applications (DEXA-99). Springer-Verlag, LNCS. http://www.csc.liv.ac.uk/-~kraft/publications.html. 1999.

[18] E. Mena, V. Kashyap, A. Sheth and A. Illarramendi, , 2000. OBSERVER: an approach for query processing in global information systems based on interoperation across pre-existing ontologies. Distributed and Parallel Databases 8(2). pp. 223-271. 2000.

[19] J. Blythe, J. Kim, S. Ramachandran and Y. Gil, 2001. An integrated environment for knowledge acquisition. Proceedings of the 6th international conference on Intelligent user interfaces (IUI'2001). Santa Fe, New Mexico, United States, pp. 13–20. 2001.

[20] Disciple-RKF, http://www.darpa.mil/ito/psum2001/-k176-0.html. 2002

[21] A. Smirnov, M. Pashkin, N. Chilov and T. Levashova,: Multi-agent Support of Mass Customization for Corporate Knowledge Management. Proceedings of the International Conference on Intelligent Manufacturing Systems, 2003, Budapest, Hungary, pp. 103–108. 2003.

[22] M. T. Mitchell, Machine Learning. The McGraw-Hill Companies, Inc., 1997.

[23] M. Gen and R. Cheng, Genetic Algorithms and Engineering Design, New York: Wiley, 1997.

[24] E. Falkenauer, Genetic Algorithms and Grouping Problems, New York: Wiley, 1998.

[25] D. E. Goldberg, The Design of Innovation: Lessons from and for Competent Genetic Algorithms. New York:Kluwer, 2002, vol. 7, Genetic Algorithms and Evolutionary Computation, 2002.

[26] E. Bach, E. Glaser, A. Condon, and C. Tanguay, DNA Models and Algorithms for NP-complete Problems, in: Proc. of the 11th Annual IEEE Conference on Computational Complexity (CCC'96) (Philadelphia, Pennsylvania, USA, 1996) 290—300. 1996