Increasing the Reliability of Texts of Electronic Documents Based on Soft Calculations under Parametric Uncertainty

Kholmonov Sunatillo Makhmudovich
PhD in Technical Sciences, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

Nomozov Abror Ismoilovich
Graduate student, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

ABSTRACT

The main approaches, principles, models, and algorithms for increasing the reliability of texts of electronic documents in automated document management systems based on the synthesis of fuzzy, evolutionary modeling methods, genetic operators with template selection mechanisms, and a suitable individual are proposed. Options for the development of GA with parametrization of operators to changing situations, as well as schemes for finding solutions that reduce labor-intensive combined search procedures, are developed.

KEYWORDS: *evolutionary modeling, data mining, database, knowledge base, genetic operators, reliability, modification.*

Relevance of the topic. The creation of data mining systems (DMS) can be improved on the basis of stochastic, fuzzy, evolutionary modeling, as well as the use of genetic adaptation and evolution operators. At the same time, genetic algorithms (GA) become a simplified tool for optimizing multi-parameter functions and adjusting parameters to changing situations, searching for suitable solutions to practical problems that reduce labor-intensive procedures of sequential selection and combined search [1,2].

This study is aimed at using the optimization principle to improve the reliability of texts of electronic documents (ED) based on GA in automated document management systems with mechanisms for using the properties, statistical and dynamic characteristics of information. Methods have been developed for the synthesis of computational schemes for extracting useful properties, regularities, and specific characteristics of information from a database (DB), control rules from a knowledge base (KB) based on the principles of evolution, heredity, variability, and natural selection [3].

The execution of genetic operators was analyzed, options for modifying a simple GA based on templates, selecting a suitable individual, and parameterizing the execution of the GA were developed [4].

Algorithms for applying evolutionary modeling. For the basis of building a DMS to improve the reliability of the information, a formal model of evolutionary modeling used in a simple GA was studied [5].

In them, the initial information reflecting random words in the time series of texts is presented as a random population of binary strings of length l, and the algorithm is performed in the following steps [6,7].

Step 1. Calculate fitness functions $f(x)$ for each row.

Step 2. Select (with replacement) two parent individuals from the current population with a probability proportional to the relative fitness function of each x row in the population.

Step 3. Cross with a probability of p_c parent individuals (with the choice of one random point of discontinuity) for creating two offspring. If the children are exact copies of the parents, then the crossover operator is not performed and one of the two children is selected by excluding the other.

Step 4. Performing a mutation of each bit of the selected descendant is done with probability p_m and the results are placed in a new population.

Step 5. If a new population is not formed, then go to step 2.

Step 6. If a new population is formed, then go to step 1.

It is assumed that for each crossover only one offspring is assumed to survive. According to the algorithm, each encoded string is assigned an integer number i^2 in the range from 0 to $(2^l - 1)$, and the population in the t generation is represented by two real vectors, each containing a 2^l component.

It is believed that $p_i(t)$ determines the probabilities of choosing rows of the i-th type, i.e. $s_i(t)$-th component of the first vector, and SS - the probabilities of choosing a row of the i-th type, i.e. i-th component of the second vector as a parent.

The initial population is generated randomly; the population size (number of N individuals) is fixed and does not change during the entire work; each individual is generated as a L-bit string, where L is the length of the individual's encoding and the length of the encoding for all individuals is assumed to be the same. The work of the GA consists of three stages.

In the first stage, an intermediate population (intermediate generation) is generated by selection of the current generation; in the second stage, a crossover (recombination) of an individual of the intermediate population is performed by means of a crossover; in the third stage, a new generation is formed and a new generation is mutated. An intermediate population represents a set of individuals who have acquired the right to reproduce. The fittest individuals are recorded several times, and the least fit ones are not included with a high probability.

The probability of each individual falling into the intermediate population is considered to be proportional to its fitness (proportional selection).

When selecting an individual, the intermediate population is randomly divided into pairs, which are then crossed with a certain probability. As a result, two descendants are obtained, which are recorded in the new generation and do not interbreed. The couple themselves are enrolled in the new generation. After selection and crossing, the mutation operator is applied to the new generation, which is necessary to “knock out” the population from the local extremum. Mutation probabilities are determined by options $1/L$ or $1/N$. The final solution of the problem is the fittest individual of the last generation.

The stop criterion is the creation of a population with a given number of generations (convergence).

Algorithm based on GA with template selection mechanism. The template (schema) defines a string with character length 0,1 and *("don't care" character). A string is a representative of a pattern

if all characters except * match. The pattern order is determined by the number of bits fixed in it. The length of the pattern is the distance between its extreme fixed bits.

The fitness of a template is determined by the average fitness of rows from the population that are its representatives. This value depends on the population and changes over time.

At each generation, the number of representatives of the template changes in accordance with its current fitness. "Good" templates have, on average, more adapted representatives; they are more often selected into the intermediate population. "Bad" patterns are more likely to die out. When selecting one line, a whole set of templates is selected at once. The number of representatives of the $M(H,t)$ template in the t generation in the intermediate generation is defined as

$$M(H,\ t+\text{int}\,ermediate) = M(H,t)\frac{f(H,t)}{< f(t) >},$$

where $f(H,t)$ is the fitness of the H template in the t generation;

$< f(t) >$ is the average fitness of the t generation.

The crossover destroys the pattern and none of the children is considered to be a representative of the pattern in question.

The probability of destruction is less than $\frac{\Delta H}{L-1}\left(1-P(H,t)\frac{f(H,t)}{< f(t) >}\right)$, where $P(H,t)$ is the proportion of representatives of the H template in the t generation.

The first multiplier of the product is equal to the probability of hitting the dividing point between the fixed bits of the template, and the second is the probability of choosing a representative of another template in the pair.

To estimate the proportion of representatives, the following inequality is used:

$$P(H,t+1) \geq P(H,t)\frac{f(H,t)}{< f(t) >}\left(1-p_c\frac{Д(H)}{(L-1)}\left(1-P(H,t)\frac{f(H,t)}{< f(t) >}\right)\right).$$

To assess the impact of a mutation in a pattern of $o(H)$ fixed bits, we use the inequality

$$P(H,t+1) \geq P(H,t)\frac{f(H,t)}{< f(t) >}\left(1-p_c\frac{\Delta(H)}{(L-1)}\left(1-P(H,t)\frac{f(H,t)}{< f(t) >}\right)\right)(1-p_m)o(H).$$

The resulting inequality describes the situation for the next generation.

Algorithm based on GA with a mechanism for selecting a suitable individual. Four mechanisms for selecting an individual are proposed.

Rank selection (rank selection). It is believed that for each individual, its probability of falling into the intermediate population is proportional to its serial number in the population sorted by increasing fitness. This type of selection does not depend on the average fitness of the population.

Tournament selection (tourna, 3mentselection). t individuals are randomly selected from the population and the best of them is placed in the intermediate population. This process is repeated N times until the intermediate population is full.

Truncation selection (truncation selection). The population is sorted by fitness, then a given proportion of the best is taken and an N individual is randomly selected for further development. When modifying the crossover operator, the following models are considered.

Homogeneous crossover. By which one of the descendants inherits each bit with probability p_0 from the first parent and with probability $(1-p_0)$ from the second. The second descendant receives not inherited by the first the bits. Usually $p_0 = 0.5$.

The Genitor (Whitley) model uses a specific selection strategy. It is believed that at each step only one pair of random parents creates only one child. This child does not replace a parent, but one of the worst individuals in the population. Only one individual is updated in the population.

Model CHC (Eshelman). For the new generation, N the best different individuals among parents and children. Duplication lines are not allowed. For crossing, all individuals are divided into pairs, but only those pairs are crossed, between which the Hamming distance is greater than some threshold.

Model HUX-operator (Half Uniform Crossover). Crossing consists of using a kind of uniform crossover, where exactly half of the bits of each parent pass to each descendant. The algorithm converges quickly due to the elimination of the mutation operator.

CHM model (cataclysmic mutation). All strings, except for the fittest one, are heavily mutated. At the same time, GAs are robust and allow finding a good solution

The key tasks are the description of GA parameterization methods at the stages of selection, crossing, mutation and reduction of individuals in a population.

Unlike statistical methods of control according to the rules of three sigma, in the proposed approach, the boundaries of control are calculated by selecting suitable parameters to adapt to a changing situation.

Algorithm based on GA with parameterization mechanism. Let us propose the result of forecast accuracy control, which in statistical modeling is estimated by the standard deviation criterion [8,9], in evolutionary modeling it is estimated by the fitness function F, represented by a two-dimensional matrix, such that $F_{i,j} = 0$ for $i \neq j$ and $F_{i,i} = f(i)$ [10,11].

All elements of the F matrix, except for the diagonal (i,i), are equal to zero and equal to the fitness values of the rows of the i-th type. The $\vec{p}(t)$ vector determines the composition of the population in the t generation, and the $\vec{s}(t)$ vector specifies the probabilities of selecting rows for crossbreeding [12,13].

For proportional selection, the vector $\vec{s}(t)$ is defined as

$$\vec{s}(t) = \frac{F\,\vec{p}(t)}{\sum_{j=0}^{2^l-1} F_{jj}\,\vec{p}_j(t)}\ . \qquad (1)$$

An "operator" G is being built, as a result of which the $\vec{s}(t)$ will correspond to the launch of the GA in the t generation to form a population in the $t+1$ generation [14]:

$$\vec{s}(t+1) = G\,\vec{s}(t)\,. \qquad (2)$$

Let $E(x)$ denote the expected share of the string x. Since $\bar{s}_i(t)$ is equal to the probability of choosing a row of the i-th type at each selection step, then

$$E(\vec{p}(t+1)) = \vec{s}(t).$$

From equation (1) we have $\vec{s}(t+1) \sim F\ \vec{p}(t+1)$, which leads to the following relation $E(\vec{s}(t+1)) \sim F\vec{s}(t)$.

If $G = F$, then this means using only the selection operator without crossover and mutation.

Now G is defined as the composition of the fitness matrix F and the "recombination operator" Γ.

The Γ operator determines the $r_{i,j}(k)$ with the probability of obtaining the k-th row as a result of the recombination of the i and j rows.

When $r_{i,j}(k)$ is known, then it is calculated

$$E(p_k(t+1)) = \sum_{i,j} s_i(t)s_j(t)r_{i,j}(k).$$

The expected share of the k row in the $t+1$ generation is equal to the probability of obtaining it for a given pair of parents, multiplied by the probability of their choice and summed over all possible pairs of parents.

However, the definition of $r_{i,j}(k)$ and Γ is a difficult task. To resolve this issue, a simpler matrix F is first determined, the elements of which $F_{i,j}$ are equal to the probability $r_{i,j}(0)$, i.e. recombination of strings i and j gives string 0 (containing only zeros) [15].

Once the elements of $r_{i,j}(0)$ are defined, they are used to consider the general case. The expression for the definition of $r_{i,j}(0)$ consists of the sums of two products:

- first, the probability that the crossover between rows i and j does not happen, and the chosen descendant (row i or j) mutates to the zero row;
- second, the probabilities that the crossover will happen and the chosen descendant will mutate to a string of all zeros.

If rows i and j are chosen for crossing, then the probability of a crossover is p_c, and the probability of the opposite event is $1 - p_c$.

The probability that a string i mutates to zero is defined as

$$p_m^{|i|}(1-p_m)^{l-|i|},$$

where $|i|$ is the number of single digits in the string i of length l;

p_m is the probability of mutation of each bit of the selected descendant,

$1 - p_m$ is the probability that the mutation will not occur.

To determine the multipliers, we will assume that h and k denote the descendants obtained as a

result of applying the c crossover by a gap at the c point. The first factor in the expression is:

$$p_m^{|i|} = \frac{1}{2}(1-p_c)[p_m^{|h|}(1-p_m)^{l-|h|} - p_m^{|k|}(1-p_m)^{l-|k|}].$$

In total there is an $l-1$ possible breaking point. The probability of choosing point c is equal to $1/(l-1)$. In this regard, the second multiplier is defined as

$$(1-p_m)^{l-|i|} = \frac{1}{2}\frac{p_c}{l-1}\sum_{c=1}^{l-1}[p_m^{|h|}(1-p_m)^{l-|h|} - p_m^{|k|}(1-p_m)^{l-|k|}].$$

Methods for optimizing the definition and tuning of model parameters based on modified GAs have been implemented. It is proved that the synthesis of models of evolutionary computations in the structure of DMS systems makes it possible to obtain tools for optimizing the definition and setting of parameters to ensure high reliability of the texts of electronic documents.

References

1. Jumanov, I., Djumanov, O., & Safarov, R. (2021). Improving the quality of identification and filtering of micro-object images based on neural networks. In E3S Web of Conferences (Vol. 304, p. 01007). EDP Sciences.
2. Jumanov, I. I., Djumanov, O. I., & Safarov, R. A. (2021, November). Mechanisms for optimizing the error control of micro-object images based on hybrid neural network models. In AIP Conference Proceedings (Vol. 2402, No. 1, p. 030018). AIP Publishing LLC.
3. Jumanov, I. I., Safarov, R. A., & Xurramov, L. Y. (2021, November). Optimization of micro-object identification based on detection and correction of distorted image points. In AIP Conference Proceedings (Vol. 2402, No. 1, p. 070041). AIP Publishing LLC.
4. Isroil, J., & Khusan, K. (2020, November). Increasing the Reliability of Full Text Documents Based on the Use of Mechanisms for Extraction of Statistical and Semantic Links of Elements. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.
5. Ibragimovich, J. I., Isroilovich, D. O., & Maxmudovich, X. S. (2020, November). Effective recognition of pollen grains based on parametric adaptation of the image identification model. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.
6. Djumanov, O., & Kholmonov, S. (2012). Methods and algorithms of selection the informative attributes in systems of adaptive data processing for analysis and forecasting. Applied Technologies & Innovations, 8(3).
7. Жуманов, И. И., & Каршиев, Х. Б. (2019). Основе базы электронных документов и особенностей правил контроля базы знаний. Проблемы вычислительной и прикладной математики, (3), 57-74.
8. Холмонов, С. М., & Абсаломова, Г. Б. (2020). Методы и алгоритмы повышения достоверности текстовой информации электронных документов. Science and world, 43.
9. Muminov, B., & Dauletov, A. (2021, November). Mathematical and Information Model of Electronic Document Management System. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 01-04). IEEE.

10. Jumanov, I. I., Djumanov, O. I., & Safarov, R. A. (2021, September). Methodology of Optimization of Identification of the Contour and Brightness-Color Picture of Images of Micro-Objects. In 2021 International Russian Automation Conference (RusAutoCon) (pp. 190-195). IEEE.

11. Jumanov, I. I., & Karshiev, K. B. (2019). Analysis of efficiency of software tools optimizing the information reliability of electronic documents in automated control systems. Chemical Technology, Control and Management, 2019(2), 57-66.

12. Isroil, J., & Khusan, K. (2020, November). Increasing the Reliability of Full Text Documents Based on the Use of Mechanisms for Extraction of Statistical and Semantic Links of Elements. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.

13. Israilovich, D. O., Makhmudovich, K. S., & Uktomovich, Y. F. (2021). Increasing The Credibility Of Forecasting Random Time Series Based On Fuzzy Inference Algorithms. International Journal of Progressive Sciences and Technologies, 26(1), 12-15.

14. Djumanov, O. I., Kholmonov, S. M., & Shukurov, L. E. (2021). Optimization of the credibility of information processing based on hyper semantic document search. Theoretical & Applied Science, (4), 161-164.

15. Jumanov, I. I., & Xolmonov, S. M. (2021, February). Optimization of identification of non-stationary objects due to information properties and features of models. In IOP Conference Series: Materials Science and Engineering (Vol. 1047, No. 1, p. 012064). IOP Publishing.