p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2023. № 1
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2023. № 1

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

UDC 004.4’24, 004.896

PARAMETER-DRIVEN GENERATION OF EVALUATION PROGRAM FOR A NEUROEVOLUTION ALGORITHM ON A BINARY MULTIPLEXER EXAMPLE

Doroshenko A. Yu. – Dr. Sc., Professor, Head of the Computing Theory Department, Institute of Software Systems of the National Academy of Sciences of Ukraine, Kyiv, Ukraine.

Achour I. Z. – Post-graduate student of the Department of Information Systems and Technologies, National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

Yatsenko O. A. – PhD, Senior Researcher of the Computing Theory Department, Institute of Software Systems of the National Academy of Sciences of Ukraine, Kyiv, Ukraine.

ABSTRACT

Context. The problem of automated development of evaluation programs for the neuroevolution of augmenting topologies. Neuroevolution algorithms apply mechanisms of mutation, recombination, and selection to find neural networks with behavior that satisfies the conditions of a certain formally defined problem. An example of such a problem is finding a neural network that implements a certain digital logic.

Objective. The goal of the work is the automated design and generation of an evaluation program for a sample neuroevolution problem (binary multiplexer).

Method. The methods and tools of Glushkov’s algebra of algorithms and hyperscheme algebra are applied for the parameter-driven generation of a neuroevolution evaluation program for a binary multiplexer. Glushkov’s algebra is the basis of the algorithmic language intended for multilevel structural design and documentation of sequential and parallel algorithms and programs in a form close to a natural language. Hyperschemes are high-level parameterized specifications intended for solving a certain class of problems. Setting parameter values and subsequent interpretation of hyperschemes allows obtaining algorithms adapted to specific conditions of their use.

Results. The facilities of hyperschemes were implemented in the developed integrated toolkit for the automated design and synthesis of programs. Based on algorithm schemes, the system generates programs in a target programming language. The advantage of the system is the possibility of describing algorithm schemes in a natural-linguistic form. An experiment was conducted consisting in the execution of the generated program for the problem of evaluating a binary multiplexer on a distributed cloud platform. The multiplexer example is included in SharpNEAT, an open-source framework that implements the genetic neuroevolution algorithm NEAT for the .NET platform. The parallel distributed implementation of the SharpNEAT was proposed in the previous work of the authors.

Conclusions. The conducted experiments demonstrated the possibility of the developed distributed system to perform evaluations on 64 cloud clients-executors and obtain an increase in 60–100% of the maximum capabilities of a single-processor local implementation.

KEYWORDS: algebra of algorithms, automated program design, cloud computing, hyperscheme, neuroevolution, neural network, parallel programming.

ABBREVIATIONS

IDS is an Integrated toolkit for software Design and Synthesis;

NEAT is NeuroEvolution of Augmenting Topologies;

SAA is a system of algorithmic algebra;

SharpNEAT is an open-source framework written in C# that implements the genetic neuroevolution algorithm NEAT.

NOMENCLATURE

A is a nonterminal operator from set R_T;

A_j is an operator from set $\overline{Op}$;

AHS is algebra of hyperschemes;

e is an empty word;

E_3 is a three-valued logic;

E_4 is a four-valued logic;

$F(A, p)$ is a function that specifies the generation method for operations of AHS signature;

GA is Glushkov’s algebra (system of algorithmic algebra);

IS is a set of states (an information set) of the operational automaton of the abstract automaton model of a computer;

$\overline{L}$ is a set of states of tape $\tilde{L}$;

$\tilde{L}$ is a tape of operational automaton $\overline{\Phi}$;

DOI 10.15588/1607-3274-2023-1-8

m is a number of address inputs of a multiplexer;

$\overline{M}$ is a set of states of operational automaton $\overline{\Phi}$;

$\tilde{M}$ is a stack of control automaton $\overline{\Psi}$;

n is a number of data inputs of a multiplexer;

Op is a set of operators of GA;

$\overline{Op}$ is a set of operators of AHS;

p is an element of set $\overline{P}$;

$\overline{P}$ is a set of states (an information set) of the operational automaton of the abstract automaton model of the parameter-driven generator of texts;

P_0 is an array length;

P_1 is a number of address inputs of a multiplexer (hyperscheme parameter);

P_2 is a number of information inputs of a multiplexer (hyperscheme parameter);

P_3 is the total number of inputs of a multiplexer (hyperscheme parameter);

P_q is a hyperscheme parameter with number q;

Pr is a set of predicates of GA;

$\overline{Pr}$ is a set of predicates of AHS;

R_N is a set of nonterminal operators of AHS;

R_T is a set of terminal operators of AHS;

s_j is address input of a multiplexer;

u_k is a predicate from set $\overline{Pr}$;

x_i is a data input of a multiplexer;

y is an output of a multiplexer;

η is “not computed” value;

μ is “undefined” value;

$\overline{\Phi}$ is an operational automaton;

$\overline{\Psi}$ is a control automaton;

Ω_{AHS} is a signature of operations of AHS;

Ω_{GA} is a signature of operations of GA;

Ω_1 is a set of logic operations included in Ω_{GA};

Ω_2 is a set of operator operations included in Ω_{GA}.

INTRODUCTION

One of the promising directions in the development and research of parallel and distributed computing systems is the construction of software abstractions in the form of algebraic-algorithmic languages and models, which aims to develop architecture- and language-independent programming tools for multiprocessor computing systems and networks. In [1], authors proposed a theory, methodology, and software tools for the automated design of parallel programs based on high-level algebraic formalization and automation of program transformations based on rewriting rules. In particular, an instrumental system of programming automation called the integrated toolkit for software design and synthesis (IDS) was developed. One of the important problems within the algebra-algorithmic approach is increasing the adaptability of programs to the specific conditions of their use. In particular, it can be solved by using the method of parameter-driven generation of algorithm schemes based on higher-level specifications named hyperschemes.

In this paper, the developed algebra-algorithmic facilities are applied to the field of neuroevolution algorithms. Neuroevolution is a promising approach for solving complex problems of machine learning, the development of artificial neural networks, adaptive control, multi-agent systems, and evolutionary robotics [2]. The main advantage of neuroevolution is that it can be used more widely than supervised learning algorithms, which require a program of correct input-output pairs. Neuroevolution only requires evaluating the performance of the network when performing a task. It uses evolutionary algorithms to train a neural network and belongs to the category of reinforcement learning methods. All evolutionary algorithms develop a set (“population”) of solutions (“individuals”). Individuals are represented by their genotype, which is expressed in the form of a phenotype, with which quality, “adaptability” is associated. There are a large number of neuroevolutionary algorithms, divided into two groups. The first includes algorithms that perform the evolution of weights for a given network topology, the second includes algorithms that, in addition to the evolution of weights, also perform the evolution of the network topology. Evolutionary algorithms manipulate a set of genotypes, which are a representation of a neural network. In a direct coding scheme, the genotype is equivalent to the phenotype, and neurons and connections are directly specified in the genotype. Conversely, in the scheme with indirect coding, the rules and structures for creating a neural network are specified in the genotype.

The object of study is the automated development of evolutionary algorithms.

One of the implementations of evolutionary algorithms is SharpNEAT [3], an open-source framework developed in the C# language. It implements the genetic neuroevolution algorithm NEAT (NeuroEvolution of Augmenting Topologies) for the .NET platform. The algorithm uses the evolutionary mechanisms of mutation, recombination, and selection to find neural networks with behavior that satisfies the conditions of a certain formally defined problem. Examples of such problems are controlling the movements of a robot’s limbs, flying a rocket, or finding a neural network that implements a certain digital logic (for example, a multiplexer).

Despite the strengths of the NEAT method, such as the possibility of its application in tasks where it is difficult to choose the cost function and neural network topology, one of the problems is the slow convergence to optimal results, especially in the case of complex environments. The distributed implementation of the NEAT evaluation method was proposed in the previous work of the authors [5]. It allows to radically speed up finding optimal configurations of neural networks in the presence of sufficient computing resources.

DOI 10.15588/1607-3274-2023-1-8

The subject of study is the automated design and generation of evaluation programs for neuroevolution algorithms.

The purpose of the work is to apply algorithm algebra and hyperschemes [1, 6] for the parameter-driven generation of an evaluation program for a sample neuroevolution problem.

Hyperschemes are parameterized specifications intended for solving a certain class of problems. Setting specific values of parameters and subsequent interpretation of hyperschemes allows obtaining algorithms adapted to specific conditions of their use. The generator of algorithms based on hyperschemes is one of the components of the above-mentioned IDS toolkit [1]. Algorithm schemes being designed in the toolkit are presented in Glushkov's system of algorithmic algebra (SAA).

The approach to the parameter-driven generation of programs is illustrated on generating the source code of the evaluation procedure for the binary multiplexer problem example included in SharpNEAT [4]. The results of the execution of multi-threaded and distributed versions of the generated procedure on a multicore processor and a cloud platform are given.

1 PROBLEM STATEMENT

The problem consists in designing a high-level parameterized specification in the algebra of hyperschemes [1, 6] that is applied to generate classes of evaluation schemes for a binary multiplexer (Binary MultiplexerEvaluator) example [4] depending on the multiplexer parameters, followed by the automated synthesis of code in C# language for the SharpNEAT framework.

A multiplexer is a device that has several data inputs x_i $(i = 0,...,n-1)$, address inputs s_j $(j = 0,...,m-1)$, and one output y. The device transmits a signal from one of the data inputs to the output; at the same time, the selection of the desired input is carried out by applying the appropriate combination of control signals to the address inputs. The number of data inputs n and the number of address inputs m are related by the ratio: $n = 2^m$. The conditional scheme of the multiplexer with 11 inputs is shown in Fig. 1.

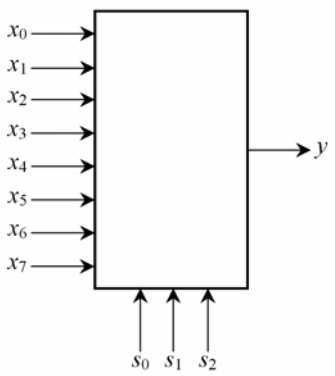


Figure 1 – The conditional scheme of a multiplexer with 11 inputs

The parameters of the hyperscheme are the number P_1 of address inputs of the multiplexer, the number P_2 of its information inputs and the total number of inputs $P_3 = P_1 + P_2$. All inputs accept binary values (0 or 1). A binary address is applied to the address inputs, representing the selection of one of the input values for data. The evaluation consists of exhaustively testing the neural network on each of the 2^{P3} possible input combinations [4]. The output value of the neural network must match the value of one of the data inputs, which is represented by a binary address from the address inputs. An output value less than 0.5 is considered a binary zero, and an output value greater than or equal to 0.5 is a binary one. The value of the assessment (suitability) is calculated additively as a result of the comprehensive check.

Depending on the values of the hyperscheme parameters, a specific scheme of an algorithm in SAA [1] is to be generated, representing a multiplexer evaluation scheme with a specific number of inputs. The examples of parameter values are shown in Table 1. The SAA schemes are the basis for the generation of C# programming code.

Table 1 – The values of the hyperscheme parameters (P_1–P_3) for generating multiplexer evaluation schemes

	Corresponding values of hyperscheme parameters		
Number of multiplexer inputs	P_1	P_2	P_3
3	1	2	3
6	2	4	6
11	3	8	11

2 REVIEW OF THE LITERATURE

This paper is related to works on the automated generation of programs from specifications [7–10] and neuroevolution of augmenting topologies [2, 11, 12].

In particular, paper [7] presents a tool for the automatic generation of C++ programs from Isabelle (high-order logic theorem prover) specifications. In [8], a combination of code and test generation based on the specification language of the Temporal Logic of Actions (TLA) is proposed. Work [9] presents a tool for generating C++ code from abstract state machine models. Paper [10] proposes an automated technique that generates executable tests from structured natural language specifications.

The peculiar feature of our approach to program generation consists in using algebra of algorithms and hyperschemes [6]. Algorithms and programs are constructed using high-level algebra-algorithmic schemes represented in a natural linguistic form. The developed tools provide automated generation of sequential and parallel code in C++ and Java languages from the schemes. In this paper, we apply these algebra-algorithmic facilities for the automated design of an evaluation procedure for a neuroevolution algorithm.

Neuroevolution of augmenting topologies is a genetic algorithm for finding artificial neural networks through evolution (neuroevolutionary method) [2]. HyperNEAT (Hypercube-based NeuroEvolution of Augmenting To-

DOI 10.15588/1607-3274-2023-1-8

pologies) is an extension of NEAT that uses a form of indirect encoding called Compositional Pattern-Producing Networks (CPPNs) [11]. The implementations of NEAT and HyperNEAT are part of a package called SharpNEAT developed in C# by Colin Green [12]. The peculiarity of NEAT and SharpNEAT is that they search both the structure of the neural network (nodes and connections) and the weight parameters of connections between nodes. The parallel distributed version of SharpNEAT was proposed by the authors in [5].

In this work, the distributed version is applied for evaluating the performance of the code generated for binary multiplexer problem example on a cloud platform.

3 MATERIALS AND METHODS

In this section, we consider the system of algorithmic algebra and hyperschemes, which are the basis of the algebra-algorithmic approach to algorithm design and synthesis. The software tools for the automated generation of algorithm schemes and programs are also described.

Glushkov's SAA is focused on the analytical form of algorithm representation and formalized transformation of these specifications, in particular, with the aim of optimizing the algorithms according to specified criteria [1]. SAA is the two-sorted algebra $GA = \langle \{Pr, Op\}; \Omega_{GA} \rangle$, where sorts are a set Pr of predicates and a set Op of operators defined on information set IS. The operators are mappings (possibly partial) of IS to itself. The predicates take values of the three-valued logic $E_3 = \{0, 1, \mu\}$. The signature $\Omega_{GA} = \Omega_1 \cup \Omega_2$ consists of system Ω_1 of logic operations (conjunction, disjunction, negation, and prognosis) that take values in set Pr, and system Ω_2 of operator operations (composition, branching, loop, and other) that take values in set Op and are considered further in more detail.

SAA is the basis of the algorithmic language SAA/1, designed for multilevel structural design and documentation of sequential and parallel algorithms and programs. The advantage of its use is the possibility of describing algorithms in a natural-linguistic form. The operators represented in the SAA/1 language are called SAA schemes. Identifiers of predicates in this language are enclosed in single quotes, and operators – in double ones. Predicates and operators in SAA/1 can be basic or compound. Basic elements are elementary atomic abstractions in algorithm schemes. Compound conditions and operators are built from basic ones using the operations from the SAA signature.

Some operator operations of SAA used in this paper are the following (represented in a natural-linguistic form):

– composition (sequential execution) of operators: "operator1"; "operator2";

– branching: IF 'condition' THEN "operator1" ELSE "operator2" END IF;

– for loop: FOR (counter FROM start TO fin) "operator" END OF LOOP;

– parallel processing of a list: PARALLEL FOR EACH (elem IN list)("operator(elem)").

The algebraic facilities for generation of algorithm schemes are based on SAA and the abstract automaton model of the parameter-driven text generator [1, 6]. The generator works according to a feedback principle (see Fig. 2). The automaton $\overline{\Psi}$ with stack $\tilde{M}$ is used as a control automaton, and the automaton $\overline{\Phi}$ with tape $\tilde{L}$ is used as an operational one. Tape $\tilde{L}$ is intended for writing the text of an SAA scheme being generated.

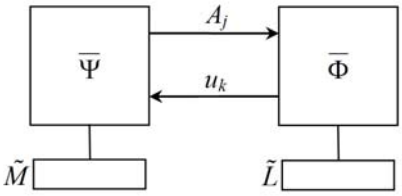


Figure 2 – The abstract automaton model of the parameter-driven text generator

Set $\overline{M}$ of states of automaton $\overline{\Phi}$ is associated with parameters that control the generation of schemes. The elements of the information set $\overline{P} = \overline{M} \times \overline{L}$ are called the states of the operational structure. At each step of the automaton's work, a set of values of logical conditions $\overline{Pr} = \{u_k\}$ defined on set $\overline{P}$ is sent from the operational to the control automaton. Depending on these values and contents of stack $\tilde{M}$, the control automaton initiates the execution of some operator. The set of operators $\overline{Op} = \{A_j\}$ is divided into two disjoint sets – terminal operators R_T and nonterminal operators R_N. Execution of the terminal operator from set R_T consists in changing the current state of the operational structure, which, in particular, can be writing some text on tape $\tilde{L}$. The execution of operator $A \in R_N$ at current state $p \in \overline{P}$ consists in writing some term $F(A, p)$ to stack $\tilde{M}$ and its further interpretation by the control automaton. The term $F(A, p)$ is an analog of the concepts of macro definition, procedure, routine, etc. Stack $\tilde{M}$ is used at processing nested and recursive terms. The generated text is the content of tape $\tilde{L}$ in the final state of the operational structure

The considered abstract automaton model is matched with the algebra of hyperschemes intended for the formalization of algorithms for the parameter-driven generation of SAA schemes [6]. It is the two-sorted algebra $AHS = \langle \{\overline{Pr}, \overline{Op}\}; \Omega_{AHS} \rangle$, where predicates from set $\overline{Pr}$ are defined on information set $\overline{P}$ and take values of the four-valued logic $E_4 = \{0, 1, \mu, \eta\}$; operators from set $\overline{Op}$ are defined on and take values in set $\overline{P}$.

DOI 10.15588/1607-3274-2023-1-8

The set of predicates is associated with parameters that control the process of SAA scheme generation. The operations of the signature Ω_{AHS} are similar to the SAA operations. The difference from SAA is that the predicates from set $\overline{Pr}$ map information set $\overline{P}$ to set E_4 with additional value η, which is used to indicate that the value of a predicate cannot be computed due to a lack of information about the values of hyperscheme parameters.

The application of operator $A \in \overline{Op}$ at state $p \in \overline{P}$ leads to the transition of operational structure $\overline{\Phi}$ to a new state $A(p) \in \overline{P}$ and writing some (possibly empty) fragment $F(A, p)$ of a scheme being generated to tape $\tilde{L}$. The function $F(A, p)$ specifies the generation method for all operations of the algebra of hyperschemes and is defined in detail in [6].

In particular, function $F(A, p)$ for operation "operator1"; "operator2" generates the composition operation without changes.

For the operation of branching, the generation function is

$$F(A, p) = \begin{cases} \text{"operator1", if 'condition'} = 1; \\ \text{"operator 2", if 'condition'} = 0; \\ \text{IF 'condition' THEN "operator1" ELSE} \\ \text{"operator2" END, if 'condition'} = \eta; \\ e, \text{if 'condition'} = \mu, \end{cases}$$

where e is an empty word.

The result of the interpretation of this operation is the text of the first operator at the true value of the condition, and the text of the second operator at the false value. The whole text of the branch operation is generated at a not computed value of the condition. An empty text is a result in the case if there was an error during the interpretation process.

Representations of operators in AHS are called hyperschemes. Each hyperscheme A applied at state $p \in \overline{P}$ generates an SAA scheme $F(A, p)$. Hyperscheme A defines the class of SAA schemes $\{F(A, p) \mid p \in \overline{P}\}$.

The processing of basic conditions and operators of a hyperscheme consists in computing expressions with hyperscheme parameters and substituting them into the text of these basic elements.

The considered approach to the generation of algorithm schemes is implemented in the integrated toolkit for software design and synthesis [1]. Hyperschemes are designed in an automated way by detailing the language constructs of the hyperscheme algebra. The constructs are selected from a list and added to the algorithm design tree. At each step of the design process, the system offers a list of algebra operations depending on the type of tree node selected. A hyperscheme is used for further generation of an SAA scheme of an algorithm (see Fig. 3) and synthesis of a program in a target programming language (C, C++, Java, and other).

To facilitate processing, the parameters are written in the text of the basic and other elements of a hyperscheme in the form P_q $(q = 0, 1, 2, ...)$. Expressions with hyperscheme parameters are enclosed in square or curly brackets.

Example. Consider the use of the hyperscheme facilities for designing a fragment of the hybrid sorting algorithm.

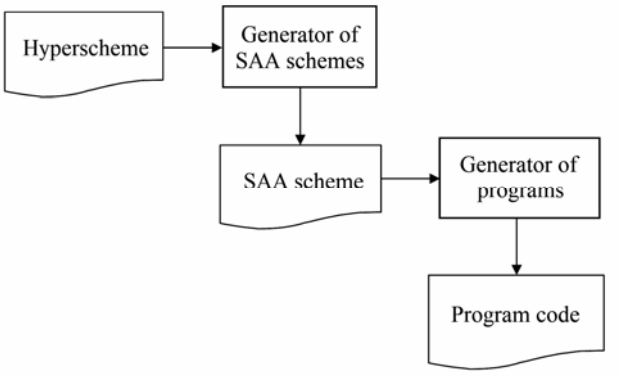


Figure 3 – The sequence of generation of algorithms and programs in the IDS toolkit

In the SAA scheme below, one of the sorting algorithms (insertion, sequential, or parallel merge sort) is selected depending on the length P_0 of the array.

```
"Hybrid sort (array)" ===
= IF '[P0 <= 200]'
  THEN
    "Insertion sort (array)"
  ELSE
    IF '[P0 <= 1000]'
    THEN
      "Sequential merge sort (array)"
    ELSE
      "Parallel merge sort (array)"
    END IF
  END IF
```

Let it be known in advance that the algorithm represented by the scheme will be applied in conditions when $P_0 \geq 500$, then the given SAA scheme becomes redundant. Considering it as a hyperscheme, we can assume that at the stage of generation of an SAA scheme, the condition '[P_0 <= 200]' takes the value "false", while '[P_0 <= 1000]' takes the value "not computed". As a result of the generation of text according to the hyperscheme, we will get the shortened SAA scheme:

```
"Hybrid sort (array)" ===
= IF '[P0 <= 1000]' THEN
    "Sequential merge sort (array)"
  ELSE
    "Parallel merge sort (array)"
  END IF
```


DOI 10.15588/1607-3274-2023-1-8

4 EXPERIMENTS

In this paper, we apply the facilities of hyperschemes for generating classes of SAA schemes intended for the evaluation of a binary multiplexer (BinaryMultiplexer Evaluator) [4].

The hyperscheme constructed using the IDS toolkit is given below. Its parameters P_1, P_2, P_3 were described in Section 1. In the scheme, curly brackets $\{P_3\}$ indicate the parameter that needs to be replaced with the corresponding number written in words, that is, if the value of $P_3 = 11$, the text "Eleven" will be inserted. Square brackets (for example, $[P_1]$ or $[P_3 – 1]$) indicate parameters or arithmetic expressions to be replaced by the corresponding number. So, for the loop FOR (i FROM 0 TO [Pow(2, P_3) – 1]) at the value of the parameter $P_3 = 11$, the text FOR (i FROM 0 TO 2047) will be generated.

```
SCHEME
BINARY{P3}MULTIPLEXEREVALUATOR ===
    "Binary {P3}-multiplexer evaluator scheme"
  END OF COMMENTS

"Binary{P3}MultiplexerEvaluator" ===
= NAME SPACE
  SharpNeat.Domains.Binary{P3}Multiplexer
  (
    CLASS Binary{P3}MultiplexerEvaluator OF
    TYPE public INHERITS
    IPhenomeEvaluator<IBlackBox>

      "Declare a constant (StopFitness) of type
       (double) = (10E + [P1])";
      "Declare a variable (_evalCount) of type
       (ulong)";
      "Declare a variable (_stopConditionSatisfied)
       of type (bool)";

      REGION IPhenomeEvaluator<IBlackBox>
             Members

        PROPERTY public ulong EvaluationCount
        GET
        (
          "Return value (_evalCount)"
        )
        END OF PROPERTY

        PROPERTY public bool
                 StopConditionSatisfied
        GET
        (
          "Return value (_stopConditionSatisfied)"
        )
        END OF PROPERTY

        METHOD public FitnessInfo
        Evaluate(IBlackBox box)
          "Declare a variable (fitness) of type
           (double) = (0.0)";
          "Declare a variable (success) of type
           (bool) = (true)";
          "Declare a variable (output) of type
           (double)";
          "Declare a variable (inputArr) of type
            (ISignalArray) = (box.InputSignalArray)";
          "Declare a variable (outputArr) of type
            (ISignalArray) = (box.OutputSignalArray)";
          "Increase (_evalCount) by (1)";
          FOR (i FROM 0 TO [Pow(2, P3) – 1])
          LOOP
            "Declare a variable (tmp) of type
             (int) = (i)";
            FOR (j FROM 0 TO [P3 – 1])
            LOOP
              (inputArr[j] := tmp&0x1);
              (tmp := tmp >> 1)
            END OF LOOP;
            "Activate the black box (box)";
            "Read output signal (output)(outputArr)";
            IF (((1 << ([P1] + (i&0x[P2 – 1])))&i) != 0)
            THEN
              (fitness := fitness + 1.0 – ((1.0 – output) *
                          (1.0 – output)));
              IF (output < 0.5)
              THEN (success := false)
              END IF
            ELSE
              (fitness := fitness + 1.0 – (output *
                          output));
                IF (output >= 0.5)
                THEN (success := false)
                END IF
            END IF;
            "Reset black box state ready for next test
             case (box)"
          END OF LOOP;
          IF success
          THEN (fitness := fitness + 10E + [P1])
          END IF;
          IF (fitness >= StopFitness)
          THEN (_stopConditionSatisfied := true)
          END IF;
          "Return value (new
                  FitnessInfo(fitness, fitness))"
        END OF METHOD

        METHOD public void Reset()
          "Empty operator"
        END OF METHOD

      END OF REGION

    END OF CLASS
  )

END OF SCHEME
```


DOI 10.15588/1607-3274-2023-1-8

Based on the hyperscheme, SAA schemes for evaluating multiplexers with three, six, and 11 inputs were generated using the IDS toolkit. Further, C# program code for the SharpNEAT framework was generated according to the schemes.

The scheme of the parallel multi-threaded evaluation procedure for the multiplexer example, implemented in SharpNEAT, looks like this:

```
METHOD private void
  Evaluate_Caching(IList<TGenome> genomeList)
  PARALLEL FOR EACH (genome IN genomeList)
  (
    “Get (phenome) for (genome)”;
    IF (phenome = null)
    THEN “Decode the (phenome) and store a
              reference against the (genome)”
    END IF;
    IF (phenome = null)
    THEN
       “Set (genome) fitness to (0.0)”;
       “Set (genome) auxiliary fitness info to (null)”
    ELSE
       “Evaluate (phenome) and get fitness
        (fitnessInfo)”;
      “Set (genome) fitness to (fitnessInfo._fitness)”;
      “Set (genome) auxiliary fitness info to
       (fitnessInfo._auxFitnessArr)”
    END IF
  )
END OF METHOD
```

In [5], the distributed version of this procedure was developed for execution on a cloud computing platform.

In this work, the experiments with multithreaded and distributed implementations of neuroevolution of augmenting topology were carried out. A multiplexer with 11 inputs was selected as an example.

The following configurations were chosen as the execution environments for single-process and distributed implementation:

1) local environment, Intel Core i9-9900K CPU (3.60 GHz – 5.00 GHz), 8 cores, 16 logic processors, 32.0 GB RAM, one process, 16 threads;

2) local environment, Intel Core i9-9900K CPU (3.60 GHz – 5.00 GHz), 8 cores, 16 logic processors, 32.0 GB RAM, distributed implementation, 16 local clients-executors;

3) cloud environment, 3rd Gen AMD EPYC Amazon EC2 C6a.large, 3.60 GHz, 2 cores, 4.0 GB RAM, up to 12.5 Gbit/s of network bandwidth, and up to 6600 Mbit/s of storage bandwidth, distributed implementation, 16 local clients-executors;

4) the same cloud environment, but with 32 cloud client executors;

5) the same cloud environment but with 64 cloud client executors.

5 RESULTS

This section gives the results of executing the multiplexer example in the computing environments described above.

Fig. 4 shows the graph of the dependence of the evaluation speed (the number of evaluations per second) on the generation number for local configurations of the environment.

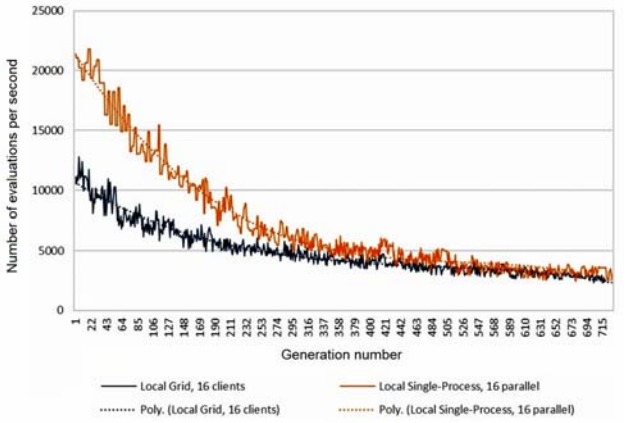


Figure 4 – The graph of the dependence of the evaluation speed on the generation number for local environment configurations

Fig. 5 shows a graph of the evaluation speed on the generation number for the cloud-based environment configurations.

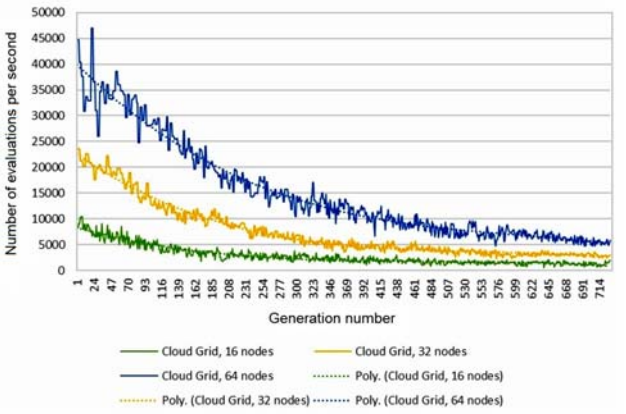


Figure 5 – The graph of the dependence of the evaluation speed on the generation number for cloud environment configurations

6 DISCUSSION

As seen from the graph in Fig. 4, the distributed implementation is expected to show worse results compared to the single-process implementation due to the overhead of interaction between processes. As the complexity of the evaluation task increases (the size of the generated neural network increases), the efficiency of the single-process and local distributed implementation is leveled off, since the overhead costs of computing resources become prohibitively lower than the evaluation costs.

As shown in Fig. 5, the distributed cloud implementation is expected to show worse results (for the

DOI 10.15588/1607-3274-2023-1-8

same number of client-executors) compared to the single-process and local distributed implementation due to the overhead of interaction between the processors of many computers, clients-executors. However, with the growth of the number of executors, we can neglect the constant value of the overhead and obtain a linear increase in the efficiency of the distributed system.

The results of the experiment demonstrated the ability of the distributed system to conduct evaluations on 64 cloud clients-executors and obtain an increase of 60–100% from the maximum capabilities of a single-processor local implementation.

CONCLUSIONS

The scientific novelty of obtained results is that the facilities of hyperscheme algebra are firstly applied for the automated generation of parametric neuroevolution evaluation algorithms on the example of the evaluation problem for a binary multiplexer. A hyperscheme is a high-level parameterized algorithm for solving a certain class of problems. Setting parameter values and subsequent interpretation of the hyperscheme allows obtaining algorithm schemes adapted to specific conditions of their use.

The practical significance of obtained results is that the means of hyperschemes are implemented in the developed integrated toolkit of automated design and synthesis of programs. Based on algorithm schemes, the system generates programs in a target programming language. The advantage of the system is the possibility of describing algorithm schemes in a natural-linguistic form. An experiment was conducted consisting in execution of the generated program for the problem of evaluating a binary multiplexer on a distributed cloud platform, which demonstrated the possibility of the developed distributed system to perform evaluations on 64 cloud clients-executors and obtain an increase in 60–100% of the maximum capabilities of a single-processor local implementation.

Prospects for further research are to apply the algebra-algorithmic method and tools for the automated development of the parallel implementation of evolutionary algorithm evaluation procedure on a graphics processing unit.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of the Institute of Software Systems of the National Academy of Sciences of Ukraine “Development of methods, technologies and tools for automating parallel programming using methods of computational intelligence” (state registration number 0122U002282).

REFERENCES

1. Doroshenko A. Formal and adaptive methods for automation of parallel programs construction: emerging research and opportunities / A. Doroshenko, O. Yatsenko. – Hershey : IGI Global, 2021. – 279 p. DOI: 10.4018/978-1-5225-9384-3
2. Designing neural networks through neuroevolution / [K. O. Stanley, J. Clune, J. Lehman et al.] // Nature Machine Intelligence. – 2019. – Vol. 1. – P. 24–35. DOI: 10.1038/S42256-018-0006-Z
3. SharpNEAT – Evolution of Neural Networks [Electronic resource]. – Access mode: https://github.com/colgreen/sharpneat
4. BinaryElevenMultiplexerEvaluator [Electronic resource]. – Access mode: https://github.com/colgreen/sharpneat/blob/master/src/SharpNeatDomains/BinaryElevenMultiplexer/BinaryElevenMultiplexerEvaluator.cs
5. Achour I. Z. Distributed implementation of neuroevolution of augmenting topologies method / I. Z. Achour, A. Yu. Doroshenko // Problems in Programming. – 2021. – № 3. – P. 3–15. DOI: 10.15407/pp2021.03.003
6. Yushchenko K. L. Algebraic-grammatical specifications and synthesis of structured program schemas / K. L. Yushchenko, G. O. Tseytlin, A. V. Galushka // Cybernetics and Systems Analysis. – 1989. – Vol. 25, № 6. – P. 713–727. DOI: 10.1007/BF01069770
7. Jiang D. Generation of C++ Code from Isabelle/HOL Specification / D. Jiang, B. Xu // International Journal of Software Engineering and Knowledge Engineering. – 2022. – Vol. 32, № 07. – P. 1043–1069. DOI: 10.1142/S0218194022500401
8. Moreira G. Fully-tested code generation from TLA+ specifications / G. Moreira, C. Vasconcellos, J. Kniess // Systematic and Automated Software Testing : 7th Brazilian Symposium SAST’22, Uberlandia, 3–7 October 2022 : proceedings. – New York : ACM, 2022. – P. 19–28. DOI: 10.1145/3559744.3559747
9. Bonfanti S. Design and validation of a C++ code generator from abstract state machines specifications / S. Bonfanti, A. Gargantini, A. Mashkoor // Journal of Software: Evolution and Process. – 2020. – Vol. 32, № 2. – P. 1–27. DOI: 10.1002/smr.2205
10. Motwani M. Automatically generating precise oracles from structured natural language specifications / M. Motwani, Y. Brun // Software Engineering : 41st IEEE/ACM International Conference ICSE’2019, Montreal, 25–31 May 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 188–199. DOI: 10.1109/ICSE.2019.00035
11. Kenneth O. S. A hypercube-based encoding for evolving large-scale neural networks / O. S. Kenneth, D. Ambrosio, J. Gauci // Artificial Life. – 2009. – Vol. 15, № 2. – P. 185–212. DOI: 10.1162/artl.2009.15.2.15202

Received 15.11.2022.
Accepted 11.02.2023.

УДК 004.4’24, 004.896

ПАРАМЕТРИЧНО-КЕРОВАНА ГЕНЕРАЦІЯ ПРОГРАМИ ОЦІНКИ ДЛЯ АЛГОРИТМУ НЕЙРОЕВОЛЮЦІЇ НА ПРИКЛАДІ ДВІЙКОВОГО МУЛЬТИПЛЕКСОРА

Дорошенко А. Ю. – д-р фіз.-мат. наук, професор, завідувач відділу теорії комп’ютерних обчислень Інституту програмних систем НАН України, Київ, Україна.

DOI 10.15588/1607-3274-2023-1-8

Ашур І. З. – аспірант кафедри інформаційних систем та технологій Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

Яценко О. А. – канд. фіз.-мат. наук, старший науковий співробітник відділу теорії комп'ютерних обчислень Інституту програмних систем НАН України, Київ, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу автоматизованої розробки програм оцінки для алгоритмів нейроеволюції наростаючої топології. Еволюційні алгоритми застосовують механізми мутації, рекомбінації та селекції для пошуку нейронних мереж з поведінкою, яка задовольняє умовам певної формально визначеної задачі. Прикладом такої задачі є знаходження нейронної мережі, що реалізує певну цифрову логіку.

Мета роботи – автоматизоване проектування та генерація програми оцінки для задачі нейроеволюції на прикладі двійкового мультиплексора.

Метод. Методи та інструментальні засоби алгебри алгоритмів Глушкова та алгебри гіперсхем застосовано для параметрично-керованої генерації програми оцінки алгоритму нейроеволюції для бінарного мультиплексора. Алгебра Глушкова покладена в основу алгоритмічної мови, призначеної для багаторівневого структурного проектування та документування послідовних і паралельних алгоритмів та програм у формі, наближеній до природної мови. Гіперсхеми є параметризованими високорівневими специфікаціями, призначеними для вирішення певного класу задач. Задавання значень параметрів і подальша інтерпретація гіперсхем дозволяє отримати алгоритми, адаптовані до конкретних умов їх використання.

Результати. Засоби гіперсхем реалізовано в розробленому інтегрованому інструментарії автоматизованого проектування та синтезу програм. На основі схем алгоритмів система генерує програми цільовою мовою програмування. Перевагою інструментарію є можливість опису схем алгоритмів у природно-лінгвістичній формі. Проведено експеримент з виконання згенерованої програми для задачі оцінки двійкового мультиплексора на розподіленій хмарній платформі. Згадана програма входить до складу SharpNEAT – системи з відкритим кодом, що реалізує алгоритм генетичної нейроеволюції NEAT для платформи .NET. Паралельна розподілена реалізація SharpNEAT була запропонована в попередній роботі авторів.

Висновки. Результати проведених експериментів продемонстрували можливість розробленої розподіленої системи виконувати оцінювання на 64 хмарних клієнтах-виконувачах та отримувати приріст у 60–100 % від максимальних можливостей однопроцесорної локальної реалізації.

КЛЮЧОВІ СЛОВА: алгебра алгоритмів, автоматизоване проектування програм, хмарні обчислення, гіперсхема, нейроеволюція, нейронна мережа, паралельне програмування.

ЛІТЕРАТУРА

1. Doroshenko A. Formal and adaptive methods for automation of parallel programs construction: emerging research and opportunities / A. Doroshenko, O. Yatsenko. – Hershey : IGI Global, 2021. – 279 p. DOI: 10.4018/978-1-5225-9384-3
2. Designing neural networks through neuroevolution / [K. O. Stanley, J. Clune, J. Lehman et al.] // Nature Machine Intelligence. – 2019. – Vol. 1. – P. 24–35. DOI: 10.1038/S42256-018-0006-Z
3. SharpNEAT – Evolution of Neural Networks [Електронний ресурс]. – Режим доступу: https://github.com/colgreen/sharpneat
4. BinaryElevenMultiplexerEvaluator [Електронний ресурс]. – Режим доступу: https://github.com/colgreen/sharpneat/blob/master/src/SharpNeatDomains/BinaryElevenMultiplexer/ BinaryElevenMultiplexerEvaluator.cs
5. Ашур І. З. Розподілена реалізація методу нейроеволюції наростаючої топології / І. З. Ашур, А. Ю. Дорошенко // Проблеми програмування. – 2021. – № 3. – С. 3–15. DOI: 10.15407/pp2021.03.003
6. Yushchenko K. L. Algebraic-grammatical specifications and synthesis of structured program schemas / K. L. Yushchenko, G. O. Tseytlin, A. V. Galushka // Cybernetics and Systems Analysis. – 1989. – Vol. 25, № 6. – P. 713–727. DOI: 10.1007/BF01069770
7. Jiang D. Generation of C++ Code from Isabelle/HOL Specification / D. Jiang, B. Xu // International Journal of Software Engineering and Knowledge Engineering. – 2022. – Vol. 32, № 07. – P. 1043–1069. DOI: 10.1142/S0218194022500401
8. Moreira G. Fully-tested code generation from TLA+ specifications / G. Moreira, C. Vasconcellos, J. Kniess // Systematic and Automated Software Testing : 7th Brazilian Symposium SAST'22, Uberlandia, 3–7 October 2022 : proceedings. – New York : ACM, 2022. – P. 19–28. DOI: 10.1145/3559744.3559747
9. Bonfanti S. Design and validation of a C++ code generator from abstract state machines specifications / S. Bonfanti, A. Gargantini, A. Mashkoor // Journal of Software: Evolution and Process. – 2020. – Vol. 32, № 2. – P. 1–27. DOI: 10.1002/smr.2205
10. Motwani M. Automatically generating precise oracles from structured natural language specifications / M. Motwani, Y. Brun // Software Engineering : 41st IEEE/ACM International Conference ICSE'2019, Montreal, 25–31 May 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 188–199. DOI: 10.1109/ICSE.2019.00035
11. Kenneth O. S. A hypercube-based encoding for evolving large-scale neural networks / O. S. Kenneth, D. Ambrosio, J. Gauci // Artificial Life. – 2009. – Vol. 15, № 2. – P. 185–212. DOI: 10.1162/artl.2009.15.2.15202

DOI 10.15588/1607-3274-2023-1-8

UDC 004.93

THE CURVE ARC AS A STRUCTURE ELEMENT OF AN OBJECT CONTOUR IN THE IMAGE TO BE RECOGNIZED

Kalmykov V. G. – PhD, Senior Reseacher of the Institute of Mathematical Machines and Systems Problems, Kyiv Ukraine.

Sharypanov A. V. – PhD, Senior Reseacher of the Institute of Mathematical Machines and Systems Problems, Kyiv Ukraine.

Vishnevskey V. V. – PhD, Leading Reseacher of the Institute of Mathematical Machines and Systems Problems, Kyiv Ukraine.

ABSTRACT

Context. The proposed article relates to the field of visual information processing in a computer environment, more precisely to the determination the parameters of the interest object in the image, in particular, the contour of the interest object In most cases, the contour of an object is a simply connected sequence of curve arcs.

Objective. The purpose and subject of the study is to find and to propose such a definition of the digital curve arc, as the most important element of the object contour in the recognizable image, which does not contradict modern neurophysiological conceptions about visual perception, and to recognize the object contour as a sequence of the digital curve arcs.

Method. The representation of the image in the form of a structural model is used, one of the structural elements of which is the contour of the object, consisting of digital curve arcs. Also, the image is considered as a cellular complex which corresponds to modern ideas about human visual perception.

Results. The new definition for arc of a digital curve as a sequence of digital straight segments is proposed, which does not contradict to modern concepts of neurophysiology. In contrast to the known definitions of a curve arc, the proposed definition of a digital curve arc makes it possible to determine the start and end points of the arc. According to the description of the contour of an object as a simply connected closed sequence of line segments, it is proposed to construct a description of the contour as a sequence of arcs of digital curves.

Conclusions. The use of the proposed definition of the digital curve arc in image processing makes it possible to recognize the contour of an object in an image and present it in a form close to visual perception. For best results, the use of variable resolution in image processing algorithms is recommended.

KEYWORDS: image, contour, curve arc, straight line segment, cellular complex, neurons, receptive field.

NOMENCLATURE

AB is the curve arc, corresponds to $x = \varphi(\tau)$, $y = \phi(\tau)$ at $l \le \tau \le u$;

a is an example of 0-cell;

b is an example of 0-cell;

c is an example of 0-cell;

C is an abstract cell complex;

C_m is an m-dimensional Cartesian complex;

d is the pixel size;

dim is a dimension function;

E is the set of abstract elements (cells);

e is an example of 0-cell;

e' is formal cell example;

e'' is formal cell example;

f is an example of 0-cell;

F is a binary relation;

g is an example of 0-cell;

I is the set of non-negative integers

h is the height of the arc segment;

h_s is the height of the s-th arc segment;

l is lower bound of the parameter τ definition area;

m is a dimension;

n is the cell number in the cell sequence;

N is the cell quantity in the cell sequence;

p is an example of 0-cell;

P is a set of points;

r is a brightness function;

s is the straight segment number in the sequence;

S is the straight segment quantity in the sequence;

t is the cell number in the cell sequence of the straight line segment;

T_bT_e is a straight line segment;

T_b is a begin point x_b,y_b of the straight line segment T_bT_e;

T_e is an end point x_e,y_e of the straight line segment T_bT_e;

u is upper bound of the parameter τ definition area;

x_n is the cell abscissa with number n in the sequence;

y_n is the cell ordinate with number n in the sequence;

x_s is the abscissa of the boundary common point (0-cell) of adjacent line segments, number s in the sequence;

y_s is the ordinate of the boundary common point (0-cell) of adjacent line segments, number s in the sequence;

$\underline{x}_s$ is the boundary point (0-cell) number s abscissa of line segment belonging to curve arc;

$\underline{y}_s$ is the boundary point (0-cell) number s ordinate of line segment belonging to curve arc, in the sequence;

x_t is the cell abscissa with number t in the sequence of the straight line segment;

y_n is the cell ordinate with number t in the sequence of the straight line segment;

DOI 10.15588/1607-3274-2023-1-9

φ is continuous parametrically defined function;
ψ is continuous parametrically defined function;
τ is the parameter on the segment [l,u].

INTRODUCTION

Many tasks of image analysis and processing consist in detecting an object of interest and determining its parameters. Typically, visual information and, in particular, a grayscale image is represented as a set of pixels, densely, without gaps, filling the field of the image. The boundaries (contour) of the detected object are often not defined as the result of image processing, as, for example, for statistical [1, 2, 3, 4] or neural network [5, 6, 7] approaches to image processing.

At the same time, other approaches are known for solving image recognition tasks, in particular, structural recognition [8], which involve the representation of an image in the form of a hierarchical structure.

The object of study is the arc of the curve as an structural element of the image while its processing.

The subject of study is the development of the curve definition arc, suitable for the analysis and processing of an image in its discrete representation. The well-known mathematical definitions of the curve arc are not very suitable for use in discrete image processing, since they define abstract curves in a continuous space.

The purpose of the work is to develop a definition of the curve arc, which is suitable for finding the arc of the curve in the image in the process of recognizing the contour of an object.

1 PROBLEM STATEMENT

Let $\{x_n,y_n\}$, $1\le n\le N$, be a sequence of points (0-cells) describing the contour of an object. Let the contour of the object be approximated by line segments and $\{x_s,y_s\}$, $1\le s\le S$, be a sequence representing boundary points (0-cells) of adjacent line segments. That is, each adjacent two line segments are represented in the sequence by pairs of points (0-cells): $(x_{s-1},y_{s-1}; x_s,y_s)$ and $(x_s,y_s; x_{s+1},y_{s+1})$, respectively.

The problem: whether there are in a sequence of line segments such that they form an arc of a curve can be reduced to checking the condition of belonging or not belonging to an curve arc of two adjacent line segments, followed by checking all pairs of line segments in the sequence.

Then it is necessary and sufficient to find the condition of belonging or not belonging to the arc of the curve of two adjacent line segments in the sequence and apply it to all pairs of adjacent segments in the sequence.

2 REVIEW OF THE LITERATURE

Consider the halftone image structure. The first hierarchical structure of an image is discussed in [8]. A more general structural image model, compared to [8], is presented here. In most cases, a grayscale image can be considered as a realization of an unknown brightness function $r = f(x,y)$ depending on two spatial variables x, y. This function defines a piecewise smooth surface. Objects in the field of view are regular pieces of a piecewise-smooth surface. The projections of the surface pieces contours onto the image plane coincide with the definition domains boundaries of the unknown functions that define the surface pieces, and are the objects contours in the field of view. The objects forming the background are the objects in the field of view, the contours of which partially or completely coincide with the boundaries of the image. Possible objects of interest may include objects in the field of view that do not have common boundaries with the boundaries of the image. Visual information should be presented taking into account the physiological characteristics of visual perception so that optimal results can be achieved with its automatic processing. In particular, one of the most important and natural features of human visual perception is its ability to segment the visual field into objects that differ from the background in brightness, color, texture. The main characteristic of any object is its shape, determined by the contour – the boundary between the object and the background. The contour, in turn, is perceived by a person as a sequence of straight line segments and curve arcs. The shape of halftone objects is also determined by the brightness function based on the color, texture within each of the objects.

These features of human visual perception are reflected in the structural model of the image. The structural model makes it possible to represent arbitrary images uniformly in form, invariant with respect to affine transformations – position in the field of view, scale, rotation. The problem of reduction to a structural model of arbitrary images given in a raster form, distorted by noise, in the general case, has not yet been solved. However, the transformation of images into a structural model can significantly increase the speed and quality of visual information processing in some rather numerous cases. The basis for the structural analysis of a halftone image is a model that determines its structural elements (Fig. 1). In particular, the objects of interest and the background of the image are such structural elements according to the known ideas about the mechanisms of visual perception. Objects, in turn, are defined by bounding contours and a three-dimensional brightness function within the object.

Contours are closed sequences formed by segments of straight lines and arcs of curved lines. Representation of grayscale images in the form of such or similar model is invariant to affine transformations of objects.

Consider the image as a cellular complex. There are many problems in image analysis that cannot be solved based on classical Euclidean geometry. The reason is that in classical geometry the assumption of space continuity is used. That is, each point in space contains in its neighborhood an infinite number of points, no matter how small this neighborhood is. According to the topological foundations of classical geometry, even the smallest neighborhood of each point contains an infinite number of other points. Thus, classical geometry has no means for

DOI 10.15588/1607-3274-2023-1-9

processing discrete images, because discrete image is presented as the set of isolated points, sufficiently small neighborhoods of which do not contain points at all, except for the isolated point itself. But then the discrete image can be described in classical geometry very approximately, with an accuracy of several distinctly small spatial elements.

It was proposed to use the mathematical apparatus of abstract cellular complexes to describe the image [9, 10]. An abstract cellular complex $C = (E, F, \dim)$ is understood as a set E of abstract elements (cells) that are in an antisymmetric, irreflexive and transitive binary relation $F \subset E \times E$, which is called the limit relation, and with a dimension function $\dim: E \to I$ with E on the set I of non-negative integers, such as $\dim(e') < \dim(e'')$ for all pairs $(e', e'') \in B$. If $(e', e'') \in B$ then one usually writes $e' < e''$ or says that cells e' limit cells e''. Such cells are called incident to each other.

Fig. 2a shows an arbitrary halftone image, which can be described as an abstract two-dimensional cellular complex. First of all, let's note the 0-dimensional cells: *a*,*b*,*c*,*d*,*e*,*f*,*g*. 0-dimensional cells correspond to points in the two-dimensional Euclidean space. The attributes of each point are coordinates. Lines correspond to 1-dimensional cells in Euclidean space: *ab*, *bc*, *cd*, *de*, *ef*, *fg*, *ga*. These lines in Euclidean space correspond to an interval – an open set of points, the closure of which is the points corresponding to 0-dimensional cells. For example, a 1-dimensional cell *ab* is bounded by 0-dimensional cells (corresponding to points in the two-dimensional Euclidean space) *a*,*b*. One-dimensional cells have no thickness. The attributes of each line are the coordinates of its boundary points. The pieces of the plane correspond to 2-dimensional cells *abg*, *bcfg*, *cdf*, *def* in the Euclidean space. The sets of points that form pieces of the plane are open. 2-dimensional cells are bounded by the corresponding 1-dimensional and 0-dimensional cells. The sets of points corresponding to 1-dimensional and 0-dimensional cells, bounding the 2-dimensional cells, are the closures of the point sets of these 2-dimensional cells. For example, a 2-dimensional cell *abg* is bounded by 1-dimensional cells *ab*, *bg*, *ga* and 0-dimensional cells *a*, *b*, *g*. The attributes of each piece of plane are the brightness value of its points, the coordinates of the bounding points and lines.

The discretized image can be represented as a Cartesian two-dimensional cell complex [10]. Each coordinate axes can be considered as a sequence of 0-cells and alternating 1-cells in the space where the image is presented (Fig. 2b). 0-cells are assigned to the points of intersection of the grid lines with the axis, 1-cells are assigned to the

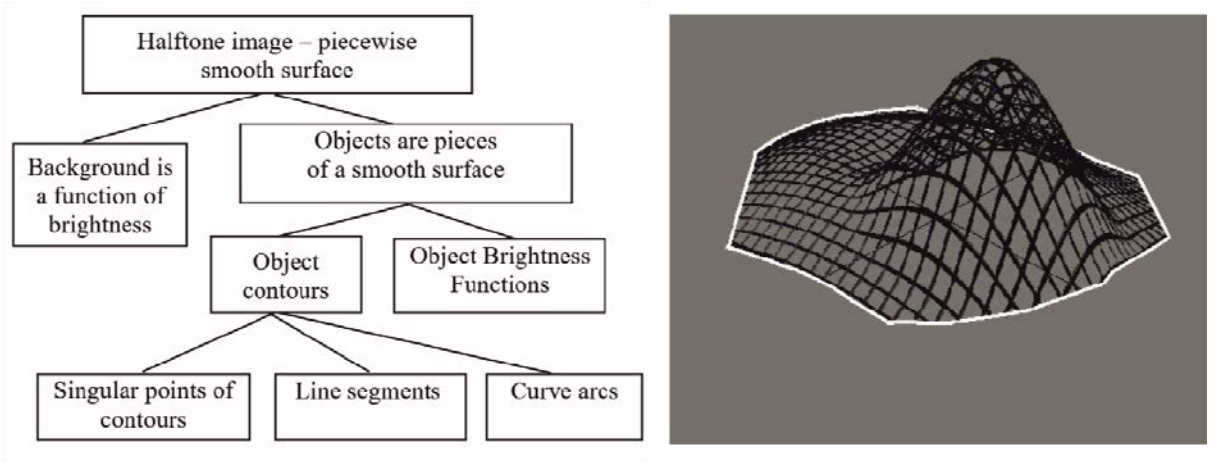


Figure 1 – The structural model of halftone image

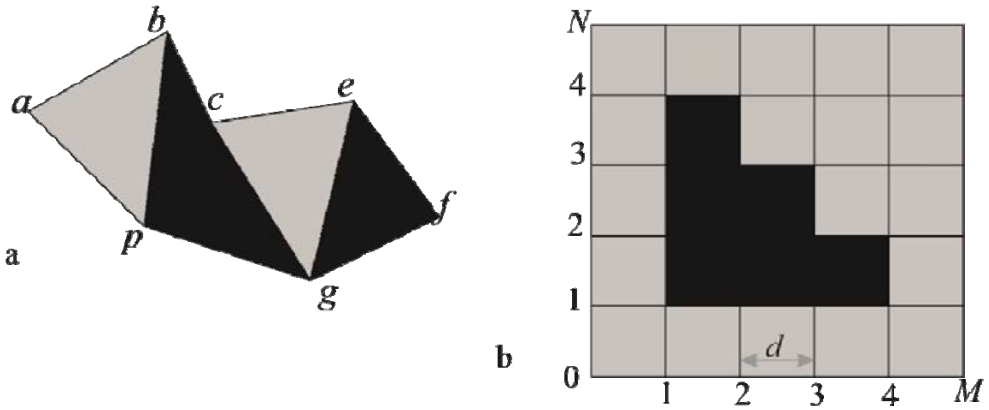


Figure 2 – Image as a: a – cell complex; b – Cartesian cell complex

DOI 10.15588/1607-3274-2023-1-9

segments between adjacent points of intersection of the grid lines with the axis. The set of cells of a two-dimensional Cartesian complex is the Cartesian product of sets of axes cells. This means that a cell of the *m*-dimensional Cartesian complex C_m is an *m*-tuple of axis cells. The bound ratio in C_m is derived from the boundary ratios of the axes. The dimension of the cell C_m is the sum of dimensions of its factors. 2-dimensional cells or grid cells correspond to pixels.

In the object in an image description the attributes of 0-cell are its coordinates, the attributes of 1-cell are its initial and final coordinates. The attributes of a 2-cell are the coordinates of its boundary 0-cell closest to the origin, as well as the brightness value of the corresponding image pixel. Each 2-cell corresponding to an image pixel is bounded by four 1-cells – cracks and four 0-cells – dots. Each 1-cell is bounded by two 0-cells. The most important feature of the object on an image is its boundary – contour. Contour of the object, when presented as a cellular complex, is a closed sequence of 0-cells and 1-cells. Fig. 2b shows the object of the halftone image and the contour of the object in the form of sequence of 0-cells and 1-cells. So, the contours are represented by lines without thickness. Thus, the description of a halftone image as a Cartesian cell complex contains sets of 0-cells, 1-cells, and 2-cells. These sets can be represented in computer memory as separate arrays.

Populations of neurons [11] have been found in the striate cortex, whose neuronal responses form the contours of figures in the visual field if their receptive fields correspond to the boundary segments of contrasting objects. The properties of these neurons' responses correspond to properties of 1-dimensional cells of cell complexes. These neurophysiological studies were carried out completely independently of mathematical work in the field of image description by methods of discrete topology.

The functioning of neurons comparable to 1-dimensional cells is described below. Most cortical cells (neurons) respond poorly to diffuse illumination and respond well to contrasting borders with a suitable orientation. For an arbitrary figure contrasting with the background, such a cell will respond if and only if a boundary segment with a certain orientation intersects its receptive field. "The same cells, whose receptive fields are located inside the boundaries of the figure, will not react in any way – they will continue to give a spontaneous impulse discharge regardless of the presence or absence of this figure. However, to excite a simple cell, it is not enough that the boundary section corresponds to the optimal orientation – the contour must also almost exactly hit the edge of the inhibitory and excitatory zones of the receptive field, because it is necessary for the answer that the light falls on the excitatory zone, but does not spread to the inhibitory one. If you the section of the contour was shifted even slightly without changing its orientation, the stimulation of this cell will turn out to be insufficient, and now another population of simple cells will begin to excite". It is known that each point of the visual field corresponds to a set of neurons (cells) with different orientational selectivity. The cells whose receptive fields and orientation coincide with the contour of the figure will answer to stimulus. The set of responses of such cells completely describes the contour of the figure. The following is also noteworthy: the output signals of cells whose receptive fields coincide with the contour of the figure and are oriented accordingly correspond to sections of the contour, one-dimensional segments without line width. That is, a material object – a neuron and its receptive field, corresponding to a part of the image, put in correspondence with an intangible object of a straight line segment – an element of the contour. Each of these contour elements can be considered as a 1-dimensional cell of the cellular complex. Here is another quote from [11]: "Another type of cells is found in the striate cortex. Typically, simple and complex cells are characterized by spatial summation – the longer the stimulus line, the better the response. However, the response only intensifies until the length of the line reaches the size of the receptive field: further lengthening of the line does not lead to a more vigorous response. In contrast, in cells that respond to line ends (end stopped cells), lengthening the line to a certain limit continues to improve the response, and if the line goes beyond this limit (in one or both directions), then the response weakens. Some cells, which we call "completely end stopped cells", do not respond at all to the presentation of a stimulus in the form of a long line. The zone from which a cell response can be elicited is called the activation zone (or excitatory zone), and the zones located at one or both ends are called inhibition zones (inhibitory zones). Thus, the entire receptive field of such a cell consists of an excitatory zone and an inhibitory zone (or zones) at the edges. A stimulus of optimal orientation, activating a cell from the excitatory zone, causes maximum inhibition outside this zone (on one or both sides)." From the standpoint of representing an image as a cellular complex, the above example can be considered as an experimental confirmation of the implementation of 0-dimensional cells in the visual system of a living organism.

Consider the mathematical definitions of the curve arc. Usually the objects contours are presented for further analysis as a closed sequence of curves arcs. The concept of a curve arc (meaning a continuous curve) is used in various fields of science and technology, in particular, recently in the processing of visual information. The concept of a continuous curve is one of the concepts that seems intuitively simple, but is actually very difficult to define. The greatest mathematicians defined the continuous curve in various ways at different periods in the development of this field of knowledge. Each new definition proceeded from the needs of human practical activity and the mathematical knowledge level of the corresponding era. The most modern definitions closely related to set theory are as follows [12].

Definition (according to Jordan). A plane curve is a set of points in a plane whose coordinates are determined by two equations:

DOI 10.15588/1607-3274-2023-1-9

$$x = \varphi(\tau),$$
$$y = \psi(\tau),$$

where τ is defined on some segment $[l,u]$. The choice of the segment values does not violate the generality of the definition.

Definition (according to Cantor) [13]. A curve in a plane is any connected, compact set P of points in the plane that does not contain any interior point.

Definition (according to Urysohn) [14]. A curve is a one-dimensional connected and at the same time compact set.

The following difficulties occur when using the above definitions while processing the visual information.

1. It is assumed in the given definitions of continuous curves that the exact boundaries of the segment or data set to be examined for suitability or non-suitability with these definitions are known. But this part of the total data set that corresponds to one or another curve arc is not known in advance, when processing signals and visual information. This part of the data set can only be formed during processing. For example, when processing an image, it is not known which part of pixels belongs to the object of interest, or to its boundaries. Often it is obtained as the result of image processing.

2. It is essential to consider the difference between curve arcs and line segments when processing visual information. This difference is not provided in the above definitions.

3. The above definitions are found and fulfilled for continuous curves in a continuous Euclidean space. But in the modern view, visual information, signals, as well as visual perception, are inherently discrete. That is, we are talking about a discrete space, where each point is isolated.

4. Traditionally, the points of a discrete curve arc in a real image are represented by pixels that have real dimensions, while the points of mathematical curves refer to infinitesimal values.

It should be recognized that another definition is required for the visual information and signals processing, taking into account the above considerations, while fully recognizing the great scientific and practical significance of the known definitions of a continuous curve.

3 MATERIALS AND METHODS

In general, the task of object contour recognition as finding simply connected closed sequence of 0-dimensional and 1-dimensional cells includes processing of grayscale image using variable resolution. The presentation of this problem and its solution are beyond the scope of this work. At the same time, the solution of this problem for a binary image is given in [15]. In the same way, the construction of an object contour for a large number of halftone images is reduced to the mentioned problem if it is possible to binarize a halftone image using a specially selected threshold. We will assume that for the object of interest in a grayscale image, a contour is constructed as a simply connected closed sequence of 0-dimensional and 1-dimensional cells. But 0-dimensional and 1-dimensional cells are elements of the contour, commensurate with the pixel, which in turn is not structural semantic, meaningful element of the contour.

As already mentioned, in accordance with the structural model of a halftone image, the contour is a simply connected closed sequence of structural elements – straight line segments and arcs of curve, which are formed from parts of a sequence of 0-dimensional and 1-dimensional cells. The task is to represent the object contour as a closed simply connected sequence of line segments and arcs of curve using the mentioned closed simply connected sequence of 0-dimensional and 1-dimensional cells as input data.

First of all, the simply connected closed sequence of line segments must be calculated for the sequence of 0-dimensional and 1-dimensional cells. Each segment corresponds to a certain simply connected part of the original sequence. The definition of line segments must be performed by 0-dimensional cells sequentially, starting from the first 0-dimensional cell of the sequence. Let $\{x_n,y_n\}$, $n=\overline{1,N}$ be a sequence of coordinates of 0-dimensional cells.

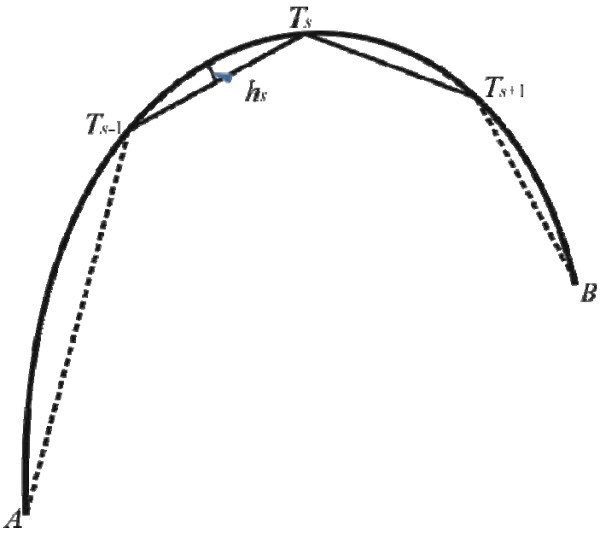


Figure 3 – The curve arc with an inscribe polyline

The possibility of representing the next part of a sequence of 0-dimensional cells $\{(x_b,y_b),\dots (x_t,y_t)\dots,(x_e,y_e)\}$, between the begin point x_b,y_b and end point x_e,y_e, as a segment of a straight line with the begin point $T_b=(x_b,y_b)$ and end point $T_e=(x_e,y_e)$, is determined by the condition [16] that

$$\max_{\forall(b\le t\le e)} dist(T_bT_e,(x_t,y_t)) \le d/2 ,$$

here $dist(T_bT_e,(x_t,y_t))$ – distance of the cell (x_t,y_t) to a line segment T_bT_e. That is, the distance from any 0-dimensional cell belonging to segment T_bT_e should not

DOI 10.15588/1607-3274-2023-1-9

exceed half the length of a 1-dimensional cell. Consistent application of the above condition allows us to represent the entire sequence as a simply connected sequence of line segments or $\{(x_1,y_1),\ldots, (x_s,y_s),\ldots, (x_S,y_S)\}$ – a sequence of boundary common 0-cells (points) of adjacent line segments.Thus, the contour of the object is presented as a simply connected, closed sequence of digital straight segments.

The definition of the arc of a digital curve proposed below makes it possible to establish or reject the fact that a sequence of digital straight segments is received due to that some arc of the curve has been discretized. We will assume, that the arcs of the curve used in graphic images represent segments of smooth functions and correspond to Jordan curves. The arcs of an arbitrary curve [9], are given by the equations $x = \varphi(\tau)$, $y = \phi(\tau)$, without multiple points or simple arcs, that is, such that for any two different values τ' and τ'' the corresponding points on the plane $[\varphi(\tau'), \phi(\tau')]$ and $[\varphi(\tau''), \phi(\tau'')]$ are different. Let $x = \varphi(\tau)$, $y = \phi(\tau)$, where the parameter τ defined on the segment $[l,u]$. As τ increases from l to u, the point with coordinates x, y describes the arc AB (Fig. 4). Consider a partition of the segment $[l, u]$ by division points

$$l=t_0<\ldots t_{s-1} < t_s < t_{s+1}\ldots < t_S = u,$$

and let these points of division correspond to the points of the curve $A,\ldots, T_{s-1}, T_s, T_{s+1}, \ldots, B$.

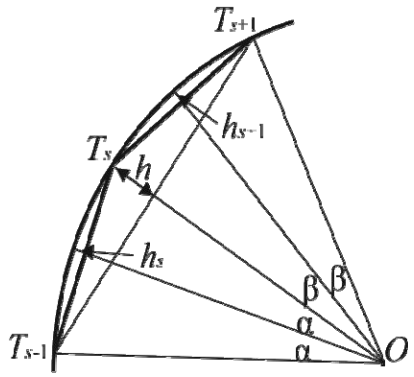


Figure 4 – The correspondence checking of straight line segments pairs to the definition of an curve arc

A polyline, inscribed in the arc AB, will be constructed if we connect successively point A with point T_1, point T_1 with… point T_{s-1}, point T_{s-1} with point T_s, point T_s with point T_{s+1}, point T_{s+1} with … point B by segments of straight lines. The figure bounded by the segment of the polyline T_s,T_{s+1} and the corresponding arc link $\cap T_s,T_{s+1}$ will be called the segment of the arc T_s,T_{s+1}, and the maximum length of the line between the segment T_s,T_{s+1} and $\cap T_s,T_{s+1}$, perpendicular to the segment T_s,T_{s+1} is the height of the arc segment h_s. Let be

$$\beta = \max_{s=0,1,\ldots,S-1} l(T_s,T_{s+1}).$$

If β tends to zero with a corresponding increase in s, then the length of any of the links of the inscribed polyline will tend to zero, as well as the height of each segment of the arc, due to the continuity of the functions $\varphi(t)$, $\phi(t)$.

While an arc and an inscribed polyline represent in a discrete space of discreteness d, segments of the inscribed polyline are displayed as line segments. Since the coordinate values take integer multiples of d in a discrete space, then objects smaller than half the discreteness the heights of the segments, in particular, will not be displayed in this space, their lengths will become equal to zero, starting from the moment when $h_s < d/2$. So, the discrete mappings of the parts of the arc will coincide with the corresponding links of the inscribed polyline – segments of digital lines for $h_s < d/2$. Thus, the contour, which consists of straight segments and arcs of arbitrary curves, after discretization is defined as a sequence of digital straight segments. Sequences of digital straight segments that correspond to arcs of curves can be considered as polylines inscribed in these arcs of curves. Such inscribed polylines will be called arcs of digital curves. The contour can include both individual segments of straight lines, and sequences of such segments – polylines that are not arcs of digital curves.

Consider pairs of adjacent segments of digital lines in sequence. In general, many curves can be drawn through three points defined by a pair of line segments. Nevertheless, as already noted, the lengths of the arcs segments heights that correspond to the segments of the inscribed polyline should not exceed the value of the space discreteness $d/2$. Thus, in order to consider pairs of segments of digital lines $T_{s-1}T_s$, T_sT_{s+1} as part of an arc of a digital curve, it is necessary to establish the existence of a curve that passes through the points T_{s-1}, T_s, T_{s+1}, such that the condition is satisfied: $(h_s<d/2)\&(h_{s+1}<d/2)$ (Fig. 4).

The curvature of a plane curve is usually identified with the curvature of a contacting circle [9]. The contacting circle of a plane curve at the point T_s is the limiting position of the circle passing through two neighboring points T_{s-1} and T_{s+1} as T_{s-1} and T_{s+1} tend to T_s. We can formulate the following definition, based on the above considerations.

Under the arc of a digital curve in a two-dimensional discrete space of discreteness d we mean such a sequence of straight line segments that through the three end points of each pair of adjacent segments it is possible to draw such a circle that the heights of the circle segments do not exceed d/2.

This definition is valid to the extent that it is legitimate to identify a segment of an arc of an arbitrary curve that corresponds to a pair of neighboring segments with an arc of a contacting circle.

Having constructed a circle in accordance with the definition of the digital curve arc for the points T_{s-1} ,T_s,T_{s+1}, let us estimate the distance of the common point T_s of the pair of segments $(T_{s-1}$,$T_s)$,(T_s,T_{s+1}) to the segment T_{s-1}, T_{s+1}, that is, the height h (Fig. 4). The

DOI 10.15588/1607-3274-2023-1-9

lengths of each segment in this pair cannot differ significantly, since this would contradict the smoothness condition – that is, $l(T_{s-1}, T_s) \sim l(T_s, T_{s+1})$. As already noted, the maximum distance between the points of the arc lines and the corresponding segment of the digital straight line is $h_s = h_{s+1} = d/2$. At the same time

$$h_s = OT_{s-1} - OT_{s-1} \times \cos\alpha = r - r\cos\alpha = r(1 - \cos\alpha).$$
$$h = OT_{s-1} - OT_{s-1} \times \cos 2\alpha = r - r\cos 2\alpha =$$
$$= r(1 - \cos 2\alpha) = 2r(1 - \cos^2\alpha).$$

$h/h_{s-1} = 2(1 + \cos\alpha)$; or $h = 2(1 + \cos\alpha) \times h_{s-1}$. If $h_{s-1} \approx d/2$ and $\alpha \le 10°$, that is $\cos\alpha \approx 1$, then the height of the triangle $(T_{s-1}T_s\ T_{s+1})$ $h \approx 2d$. Using the value h_s instead of h_{s-1} will not affect the result, since both $\beta \le 10°$ and $\cos\beta \approx 1$.

This means that in order to be related to the digital arc of curve for the considered pair of segments, it is necessary that the value of the maximum deviation of h does not exceed $2d$. The minimum deviation is $h > d/2$, since at a smaller deviation the directions of the segments $T_{s-1}T_s$ and T_sT_{s+1} are indistinguishable, and a pair of segments of different directions turns into one straight segment. If $h > 2d$, then the segments under consideration are segments of a polyline. Thus, taking into account the above considerations, the sequence of common points of adjacent segments takes the form: $\{(x_1,y_1),\ldots, (\underline{x_{s-1},y_{s-1}}),(\underline{x_s,y_s}), (\underline{x_{s+1},y_{s+1}}),\ldots, (x_S,y_S)\}$, where the points belonging to the digital curve arc are underlined.

4 EXPERIMENTS

Experimental verification of the proposed method consists in representing the contour of the interest object as a sequence of digital curve arcs and segments of digital straight lines. Moreover, the contour elements sequence of the object must be the same for various affine transformations – the rotation of the interest object, changing the position in the image field. For comparison, the representation the contour of the interest object as a sequence of digital curve arcs and segments of digital straight lines was performed using a well-known tool – the graphical editor Corel Draw.

For the experiment, binary images of object contours that were not distorted by noise were used, since noise filtering, as well as recognition of the object contours in a grayscale image, are separate tasks that must be solved by appropriate means. Separate works will be devoted to solving these problems.

An example of the image used in the experiment is shown in Fig. 5. Objects in the image are the identical sectors of the ellipse, differing in space position and angle of rotation. The result of processing each object is its contour, represented by a closed sequence of line segments in the form of boundary points (0-cells) of adjacent line segments. Some parts of the line segments sequence are defined as arcs of curves with indication of their boundary points (0-cells).

Figure 5 – The example of the image for the experiment to determine the contour as a sequence of curve arcs and line segments

The experimental program was executed in the Visual C environment. The required RAM is no more than 512 MB.

The main blocks of the program:

1. Represent the image as a cell complex and determine the sequences of boundary 0-cells and 1-cells for each object.

2. Define the contour of each object as a sequence of line segments in the form of a sequence of boundary points of neighboring segments.

3. Determine for the contour of each object the parts of the line segments sequence that form the arcs of the curves, indicating the boundary points of the arcs of the curves. The use of curve arcs as the structural elements of image contours description would approach its description to intuitional, natural representation of images by a man, substantially would shorten the expenses of memory for storage of image and image processing time. As an example we will consider description of contours of binary images which are got with the use of tools of widespread graphics editor of Corel Draw. On Fig. 5a the contours of three identical objects are represented, not to be distorted by noises. Each of objects contains the arc of ellipse and differs from the other objects by spatial position and rotation angle. The boundary points which divide contours into the curve arcs and the straight segments are marked by the squares. Identical with each other arcs, to belong to different objects, are represented by sequences containing the different amount of different arcs of curves. Each of identical objects in the image is represented with the different elements. Such description of objects can not be directly used in the intelligence systems for interpretation of images, as supposes enough hard processing. The represented example shows existence of the problem even at the images not distorted by noises and actuality to solve the problem.A special program that implements the proposed method and algorithms has been developed.

The example of contour recognition as the sequences of digital straight segments and of digital curve arcs by the program is demonstrated on a Fig. 6b. The image is used from the Fig. 6a, but the offered algorithms are implemented in the program. Unlike the contours of Fig. 6a, got by means of the program of Corel Trace, the arcs of contours are represented without laying out by intermediate points on a few arcs regardless of different spatial positions and rotation angles for the each of objects.

DOI 10.15588/1607-3274-2023-1-9

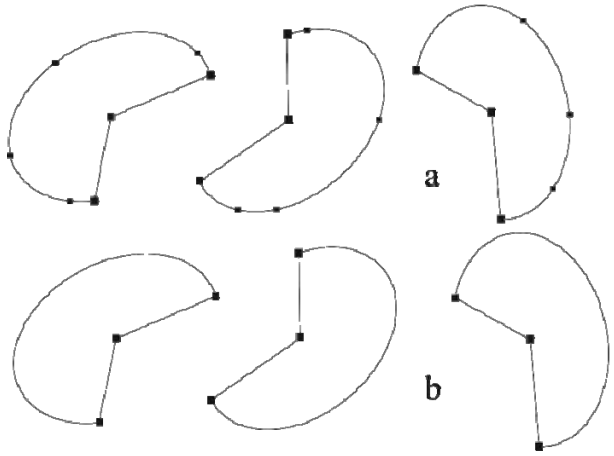


Figure 6 – a – Contour recognition in three identical, rotated in relation to each other objects by facilities of Corel Draw;
b – Contour recognition by the program, using the proposed method and algorithms

5 RESULTS

The main result, apparently, should be considered the ability to determine the arcs of curves in sequences of 0-cells and 1-cells that form the contour of an object. The same contour configurations must correspond to the same sequences of curve arcs, regardless of affine transformations – shift, rotation angle, scale.

The significance of the result obtained is clear only when comparing the processing of the same object by known and proposed methods. The result of the experiment is shown in Fig. 6. The contours of objects from Fig. 5, recognized by means of the Corel Draw graphic editor are shown in Fig. 6a. Each of objects contains the arc of ellipse and differs from the other objects by spatial position and rotation angle. The boundary points which divide contours into the curve arcs and the straight segments are marked by the squares. Identical with each other arcs, to belong to different objects, are represented by sequences containing the different amount of different arcs of curves. Each of identical objects in the image is represented with the different elements. The arc of each ellipse is represented by several unequal arcs. Such representation of processing results cannot be used in artificial intelligence tasks, in particular, in recognition tasks.

The proposed method is free from these defects. Recognition of the same objects by the developed experimental program is shown in Fig. 6b

6 DISCUSSION

The representation of image within the framework of the structural model, on the one hand, is natural for visual perception, on the other hand, it is fully consistent with the theory of cell complexes. The image object is a 2-dimensional cell. The contour of the object, its boundary, is, most often, a simply connected, closed sequence of 1-dimensional cells that form segments of straight lines and arcs of curve. The boundaries of line segments and arcs of curve, i.e. singular points correspond to 0-dimensional cells. Compared to recent imaging concepts, an important advantage to using cell complexes in image processing is the following [16].

One of the simplest tasks of image processing is to encode the object contour of the binary image as a single-connected closed sequence of the object boundary elements in the image. The contour of the object is a closed curve that divides the image into two parts: the object itself and the other part of the image. Traditionally, for a binary discretized image, contour pixels bounding with background pixels or, conversely, background pixels bounding with object pixels are used as contour elements. In order to construct a single-connected closed sequence from boundary pixels that corresponds to the contour of the image object, the concept of connectivity – the pixels neighborhood – must be defined. The following ideas about the connectivity of pixels in two-dimensional discrete space are generally accepted:

1. Pixels are considered adjacent if they have a common side. In this case, each pixel has four adjacent pixels.

2. Pixels are considered adjacent if they have a common side or a common point. In this case, each pixel has eight adjacent pixels. Examples of closed lines and contours of objects formed by boundary pixels are shown in Fig. 6. As follows from the above examples, this representation of contour has significant disadvantages [13].

Fig. 7a shows a closed line drawn according to the rules of 4-neighborhoods. As a result of the use of 4-neighborhoods, the contour line divides the image field not into two areas, as it should be, but into three.

A closed simply connected line drawn according to the rules of 8-neighborhoods is shown on Fig. 7b. But, due to 8 neighborhoods, the space inside the closed line and outside of it is not divided: there are connections between pixels inside and outside the line.

Fig. 7c shows an attempt to construct a closed contour of the object using the boundary pixels adjacent to the pixels of the object according to the rules of 4-neighborhoods: the contour line is not closed.

Fig. 7d illustrates the case of constructing a closed contour of an object using boundary pixels adjacent to the pixels of the object according to the 8-neighborhood rules: the contour line is not simply connected.

A curved line in continuous space, as follows from its definition, has no thickness. That is, each of the infinitely large set of points that form a line is an infinitesimal value. This also applies to the closed curve of the object contour line (boundary). At the same time, traditionally used representations of a curved line in a discrete space assume that the curve consists of minimal elements of this space – points. But the point in this case corresponds to a pixel – the minimum element of the image that has finite dimensions. It is this difference that is the reason for the above paradoxes of representing lines as a sequence of pixels. That is, the correct representation of initially continuous images in discrete space is possible using the theory of cell complexes.

DOI 10.15588/1607-3274-2023-1-9

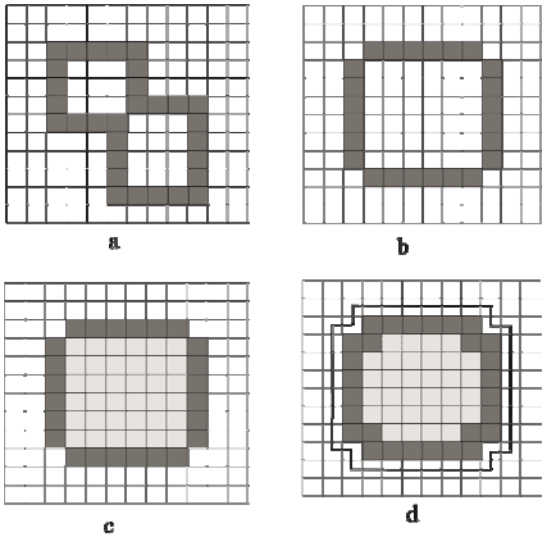


Figure 7 – Closed lines formed by pixels: problems and paradoxes

In the primary visual (striate) cortex, neurons were found that generate signals – responses to extended pieces at the border of contrast areas of the visual field, and those and only those neurons are excited whose receptive fields match the border and their orientation match the orientation of the corresponding sections of the contrast area border. That is, the excited neurons respond to the boundary segments of the straight line in certain orientations, which can be considered as 1-cells if the contrast area is considered as a cellular complex. That is, the representation of an image as a cellular complex can be considered as an approximation to the implementation of the mechanisms of visual perception. It can also be assumed that the signals of a curved line are formed by pairs of segments of the corresponding directions, as proposed in this paper.

6 CONCLUSIONS

The paper considers the arc of the curve as a structural element of the image, more precisely as a structural element of the interest object contour in the image, and the image is presented as a cellular complex.

The scientific novelty is that the arc of a digital curve is defined in the discrete space of a digital image, in contrast to the known definitions of continuous curves, which are oriented to use in a continuous space.

The practical significance is that the interest objects contours are presented as sequences of line segments and arcs of digital curves. This representation of the object contour does not depend on affine transformations, such as position in the field of view and rotation, which greatly simplifies image processing.

Prospects for further research are as follows. The successful result of object contour recognition in the form of a sequence of straight line segments and arcs of a digital curve depends on the choice of d – the resolution value when sampling the image. The task of determining the most appropriate resolution for a particular image has not been solved. It is all the more possible that different parts of the same image must be processed at different resolutions. Therefore, the recognition of line segments and arcs of curves using variable resolution will be considered in subsequent publications.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of the Institute of Mathematical Machines and Systems Problems “Structural methods of processing cyclic biomedical signals and cloud services based on them” (state registration number 0121U110584).

REFERENCES

1. Pavlidis T. Algorithms for Graphics and Image Processing. Berlin, Springer-Verlag, 1982, 400 p.
2. Gonzalez R. C., Woods R. E., Eddins S. L. Digital Image Processing using MATLAB. New York, Pearson Education, 2004, 616 p.
3. Pratt W. K. Digital Image Processing. New York, John Wiley & Sons, Inc, 1982, 738 p.
4. Schlesinger M., Hlavac V. Ten Lectures on Statistical and Structural Pattern Recognition. Dordrecht / Boston / London, Computational Imaging and Vision Kluwer Academic Publishers, 2002. 520 p.
5. Ivakhnenko A. G., Lapa V. G. Cybernetics and Forecasting Techniques. New York, American Elsevier Publishing Company, 1967, 168 p.
6. LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D. Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation*, 1989, Vo1. 1, No. 4, pp. 541–551. doi:10.1162/neco.1989.1.4.541. S2CID 41312633.
7. Maitra D. S., Bhattacharya U., Parui S. K. CNN based common approach to handwritten character recognition of multiple scripts, *13th International Conference on Document Analysis and Recognition (ICDAR): 23-26 August 2015: proceedings.* Tunis, IEEE 2015, pp. 1021–1025. doi:10.1109/ICDAR.2015.7333916. ISBN 978-1-4799-1805-8. S2CID 25739012.
8. Fu K. S. Syntactic Methods in Pattern Recognition. New York and London, Academic Press, 1974, 511 p.
9. Aleksandrov P. S. Combinatorial Topology. Rochester, Graylock Press, 1956, 656 p.
10. Kovalevsky V. Finite Topology as Applied to Image Analysis, *Computer Vision, Graphics and Image Processing*, 1989, Vol. 46, No. 2, pp. 141–161.
11. Hubel D. H. Eye, brain, and vision. New York, Scientific American Library, Distributed by W.H. Freeman, 1988, 240 p.
12. Berg G. O., Julian W., Mines R., Richman F. The constructive Jordan curve theorem, *Rocky Mountain Journal of Mathematics*, 1975, Vol. 5, № 2, pp. 225–236. DOI: 10.1216/RMJ-1975-5-2-225, ISSN 0035-7596, MR 0410701
13. Dovgoshey O., Martio O., Ryazanov V., Vuorinen M. The Cantor function, *Expositiones Mathematicae. Elsevier BV*, 2006, Vol. 24, № 1, pp. 1–37. DOI: 10.1016/j.exmath.2005.05.002. ISSN 0723-0869. MR 2195181
14. Alexandrov A. D., Reshetnyak Yu. G. General theory of irregular curves, *Mathematics and its Applications (Soviet Series)*, 29. Kluwer, Academic Publishers Group, Dordrecht, 1989, 288 p. ISBN: 90-277-2811-9
15. Schlesinger M. I. Mathematical Tools of Picture Processing. Kyiv, Naukowa Dumka, 1989, 117 p.

DOI 10.15588/1607-3274-2023-1-9

16. Kovalevsky V. A. Applications of Digital Straight Segments to Economical Image Encoding, *7th International Workshop, DGCI'97*. Montpellier, France, December 3–5 1997, proceedings, Springer 1997, pp. 51–62.

Accepted 09.11.2022.
Received 05.01.2023.

УДК 004.93

ДУГА КРИВОЇ ЯК СТРУКТУРНИЙ ЕЛЕМЕНТ ЗОБРАЖЕННЯ, ЩО МАЄ БУТИ РОЗПІЗНАНЕ

Калмиков В. Г. – канд. техн. наук, старший науковий співробітник Інституту проблем математичних машин і систем, Київ, Україна.

Шарипанов А. В. – канд. техн. наук, старший науковий співробітник Інституту проблем математичних машин і систем, Київ, Україна.

Вишневський В. В. – канд. техн. наук, провідний науковий співробітник Інституту проблем математичних машин і систем, Київ, Україна.

АНОТАЦІЯ

Актуальність. Пропонована стаття стосується галузі обробки візуальної інформації в комп'ютерному середовищі, а саме визначення параметрів об'єкта інтересу на зображенні, зокрема контуру об'єкта інтересу. У більшості випадків контур об'єкта інтересу є однозв'язна послідовність дуг кривих.

Мета. Мета і предмет дослідження – знайти і запропонувати таке визначення дуги цифрової кривої, як найважливішого елемента контуру об'єкта в розпізнаваному образі, яке не суперечить сучасним нейрофізіологічним уявленням про зорове сприйняття, і розпізнати контур об'єкта як послідовність дуг цифрових кривих.

В якості **методу** використовується подання зображення у вигляді структурної моделі, одним із структурних елементів якої є контур об'єкта, що складається з цифрових дуг кривих. Також зображення розглядається як клітинний комплекс, що відповідає сучасним уявленням про зорове сприйняття людини.

Результати. Запропоновано нове визначення дуги цифрової кривої як послідовності відрізків цифрових прямих, що не суперечить сучасним уявленням нейрофізіології. На відміну від відомих визначень дуги кривої, запропоноване визначення дуги цифрової кривої дає можливість визначити початкову та кінцеву точки дуги. За описом контуру об'єкта як однозв'язної замкнутої послідовності відрізків пропонується побудувати опис контуру як послідовності дуг цифрових кривих.

Висновки. Використання запропонованого визначення дуги цифрової кривої при обробці зображень дає змогу розпізнати контур об'єкта на зображенні та представити його у формі, наближеній до зорового сприйняття. Для досягнення найкращих результатів рекомендується використовувати змінну роздільну здатність в алгоритмах обробки зображень.

КЛЮЧОВІ СЛОВА: зображення, контур, дуга кривої, відрізок прямої, клітинний комплекс, нейрони, рецептивне поле.

ЛІТЕРАТУРА

1. Pavlidis T. Algorithms for Graphics and Image Processing / T. Pavlidis. – Berlin : Springer-Verlag, 1982. – 400 p.
2. Gonzalez R. C. Digital Image Processing using MATLAB / R. C. Gonzalez, R. E. Woods, S. L. Eddins. – New York : Pearson Education, 2004. – 616 p.
3. Pratt W. K. Digital Image Processing / W. K. Pratt. – New York : John Wiley & Sons, Inc, 1982. – 738 p.
4. Schlesinger M. Ten Lectures on Statistical and Structural Pattern Recognition / M. Schlesinger, V. Hlavac. – Dordrecht / Boston / London: Computational Imaging and Vision Kluwer Academic Publishers, 2002. – 520 p.
5. Ivakhnenko A.G. Cybernetics and Forecasting Techniques / A.G. Ivakhnenko, V. G. Lapa. – New York : American Elsevier Publishing Company, 1967. – 168 p.
6. Backpropagation Applied to Handwritten Zip Code Recognition / [Y. LeCun, B. Boser, J.S. Denker et al.] // Neural Computation. – 1989. – Vol. 1, № 4. – P. 541–551. DOI: 10.1162/neco.1989.1.4.541. S2CID 41312633.
7. Maitra D. S. CNN based common approach to handwritten character recognition of multiple scripts / D. S. Maitra, U. Bhattacharya, S. K. Parui// 13th International Conference on Document Analysis and Recognition (ICDAR): 23–26 August 2015: proceedings. – Tunis: IEEE 2015. – P. 1021–1025. DOI: 10.1109/ICDAR.2015.7333916. ISBN 978-1-4799-1805-8. S2CID 25739012.
8. Fu K. S. Syntactic Methods in Pattern Recognition / K. S. Fu. – New York and London: Academic Press, 1974. – 511 p.
9. Aleksandrov P. S. Combinatorial Topology / P. S. Aleksandrov. – Rochester : Graylock Press, 1956. – 656 p.
10. Kovalevsky V. Finite Topology as Applied to Image Analysis / V. Kovalevsky // Computer Vision, Graphics and Image Processing. – 1989. – Vol. 46, No. 2. – P. 141–161.
11. Hubel D. H. Eye, brain, and vision / D. H. Hubel. – New York : Scientific American Library, Distributed by W.H. Freeman, 1988. – 240 p.
12. The constructive Jordan curve theorem / [G. O. Berg, W. Julian, R. Mines, F. Richman] // Rocky Mountain Journal of Mathematics. – 1975. – Vol. 5, № 2. – P. 225–236. DOI: 10.1216/RMJ-1975-5-2-225, ISSN 0035-7596, MR 0410701
13. The Cantor function / [O. Dovgoshey, O. Martio, V. Ryazanov, M. Vuorinen] // Expositiones Mathematicae. Elsevier BV. – 2006. – Vol. 24, № 1. – P. 1–37. DOI:10.1016/j.exmath.2005.05.002. ISSN 0723-0869. MR 2195181
14. Alexandrov A. D. General theory of irregular curves / A. D. Alexandrov, Yu. G. Reshetnyak // Mathematics and its Applications (Soviet Series), 29. – Kluwer Academic Publishers Group, Dordrecht. – 1989. – 288 p. ISBN: 90-277-2811-9
15. Schlesinger M. I. Mathematical Tools of Picture Processing / M. I. Schlesinger. – Kyiv: Naukowa Dumka, 1989. – 117 p.
16. Kovalevsky V. A. Applications of Digital Straight Segments to Economical Image Encoding / V. A. Kovalevsky // 7th International Workshop, DGCI'97 – Montpellier, France, December 3–5 1997 – proceedings. – Springer 1997. – P. 51–62.

DOI 10.15588/1607-3274-2023-1-9

UDC 004.93

APPLICATION OF TWO-DIMENSIONAL PADÉ-TYPE APPROXIMATIONS FOR IMAGE PROCESSING

Olevskyi V. I. – Dr. Sc., Professor, Professor of the Information Technology and Computer Engineering Department, Dnipro University of Technology, Dnipro, Ukraine.

Hnatushenko V. V. – Dr. Sc., Professor, Head of the Department of Information Technology and Computer Engineering Department, Dnipro University of Technology, Dnipro, Ukraine.

Korotenko G. M.– Dr. Sc., Associate Professor, Professor of the Information Technology and Computer Engineering Department, Dnipro University of Technology, Dnipro, Ukraine.

Olevska Yu. B. – PhD, Associate Professor, Associate Professor of the Applied Mathematics Department, Dnipro University of Technology, Dnipro, Ukraine.

Obydennyi Ye. O. – Assistant of the of Information Technology and Computer Engineering Department, Dnipro University of Technology, Dnipro, Ukraine.

ABSTRACT

Context. The Gibbs phenomenon introduces significant distortions for most popular 2D graphics standards because they use a finite sum of harmonics when image processing by expansion of the signal into a two-dimensional Fourier series is used in order to reduce the size of the graphical file. Thus, the reduction of this phenomenon is a very important problem.

Objective. The aim of the current work is the application of two-dimensional Padé-type approximations with the aim of elimination of the Gibbs phenomenon in image processing and reduction of the size of the resulting image file.

Method. We use the two-dimensional Padé-type approximants method which we have developed earlier to reduce the Gibbs phenomenon for the harmonic two-dimensional Fourier series. A definition of a Padé-type functional is proposed. For this purpose, we use the generalized two-dimensional Padé approximation proposed by Chisholm when the range of the frequency values on the integer grid is selected according to the Vavilov method. The proposed scheme makes it possible to determine a set of series coefficients necessary and sufficient for construction of a Padé-type approximation with a given structure of the numerator and denominator. We consider some examples of Padé approximants application to simple discontinuous template functions for both formulaic and discrete representation.

Results. The study gives us an opportunity to make some conclusions about practical usage of the Padé-type approximation and about its advantages. They demonstrate effective elimination of distortions inherent to Gibbs phenomena for the Padé-type approximant. It is well seen that Padé-type approximant is significantly more visually appropriate than Fourier one. Application of the Padé-type approximation also leads to sufficient decrease of approximants' parameter number without the loss of precision.

Conclusions. The applicability of the technique and the possibility of its application to improve the accuracy of calculations are demonstrated. The study gives us an opportunity to make conclusions about the advantages of the Padé-type approximation practical usage.

KEYWORDS: Padé-type approximants, Gibbs phenomenon, size of the image file.

ABBREVIATIONS

DCT is a discrete cosine transform;

2D DCT is a two-dimensional discrete cosine transform;

NOMENCLATURE

$L_2[\ ]$ is a metric Hilbert space with a square measure;

a_i are minimal levels of variables;

b_i are maximum levels of variables;

x_i are complex variables on the interval (a_i, b_i);

$f()$ is an arbitrary function in the space under consideration;

a_{kp} are coefficients of harmonics for the 2D Fourier series of f;

B_i are countable sets of basic functions;

e_{ik} are basic functions;

B is a basis of the space;

$P[\ ]$ is a two-dimensional Padé approximant;

m_i are maximum powers of x_1 for numerator and denominator of P;

n_i are maximum powers of x_2 for numerator and denominator of P;

S is a two-dimensional power series;

GS is a generalized power series;

GP_{GS} is a Padé-type functional;

N is a number of harmonics for the Fourier series in x_1 direction, and the maximum power of x_i in both numerator and denominator of P in the Chisholm approximation;

M is a number of harmonics for the Fourier series in x_2 direction;

λ_i are frequencies of 2D Fourier series;

n_F is a number of parameters in the Fourier series;

n_P is a number of parameters in the Padé-type approximation;

F is a 2D DCT for f;

$f_{N,N}$ is the Chisholm approximation of N-th order;

p is a matrix of the Chisholm approximation numerator coefficients;

p_{kp} are elements of p;

DOI 10.15588/1607-3274-2023-1-10

q is a matrix of the Chisholm approximation denominator coefficients;

q_{kp} are elements of q.

INTRODUCTION

Image processing by expansion of the signal into a two-dimensional Fourier series in order to reduce the size of the graphical file often leads to significant image distortion, in particular, due to the Gibbs effect. The Gibbs phenomenon is the property of the one-dimensional Fourier series which manifests in the decomposition of a discontinuous periodic function when they are truncated and a finite number of members are used [1–3]. In the case of truncation a distortion occurs near the discontinuity points which cannot be eliminated by increasing the finite number of terms of the series. In two-dimensional case the Gibbs phenomenon significantly reduces the quality of the processed images for most popular graphic standards because they use a finite sum of harmonics. Distortion occurs on the borders of sharp contrast change and leads to the appearance of false optical shadows. It negatively influences the analysis quality when processing the results of x-ray and sonar studies.

The object of study is the image processing.

The subject of study is the generalized sum of a two-dimensional Fourier series obtained by image processing using the fractional rational approximants.

The purpose of the work is the application of the two-dimensional Padé-type approximations for the purpose of elimination of the Gibbs phenomenon for image processing and the reduction of the size of the resulting image file.

1 PROBLEM STATEMENT

Let the original monochrome image be given in the form of a two-dimensional function of tone *f* or a set of its values for rectangle $(a_2,b_2)\times(a_2,b_2)$. The aim is to find an appropriate two-dimensional Padé-type approximation $P[m_1,n_1/m_2,n_2](x_1,x_2)$ to eliminate the Gibbs phenomenon for image processing and reduce the size of an image file.

2 REVIEW OF THE LITERATURE

The Gibbs phenomenon also exists in the two-dimensional case, and it significantly reduces the quality of images processing for most popular graphic standards because they use a finite sum of harmonics [2]. Distortion occurs on the borders of sharp contrast change and leads to the appearance of false optical shadows [3, 4]. It is detrimental for the analysis quality when processing the results of the x-ray and sonar studies. Gibbs phenomenon is a type of MRI artifact, which leads to a series of lines in the MRI image parallel to abrupt and intense changes in the object, such as the CSF-spinal cord and the skull-brain interface [4, 5].

The theory of approximation of mathematical physics functions is the most rapidly developing field of mathematics [1, 6, 7]. Traditionally, only the approximation techniques which use polynomials [1, 8, 9] and trigonometric functions [6, 7, 10] are considered. The most successful type of such approximation techniques is the approximation by fractional rational functions [11, 12]. The interest in the theory of fractional-rational approximations has been steadily increasing due to their wide application in various studies in the field of theoretical physics, applied mechanics, geophysics, etc. [8, 11, 12] due to them allowing for a generalized summation of series and the extension of the function to be approximated into the meromorphic domain.

Recently, great attention has been given to the expansion of the classical theory of approximation by fractional-rational functions to various types of basis functions and different methods of constructing approximants – to Padé-type approximations [12–19]. A special choice of the constructing method for the approximation makes it possible in many cases to achieve a significant improvement in the useful properties of the approximants for certain distinct classes of functions [12–14].

We suggest the application of the two-dimensional Padé-type approximants method which we have developed earlier [12] for the purpose of reduction of the Gibbs phenomenon in the harmonic two-dimensional Fourier series.

3 MATERIALS AND METHODS

Let's provide some basic principles of construction of the Padé-type approximants for the harmonic two-dimensional Fourier series as a subset of power series. According to our approach [12], we consider the separable space $L_2[(a_2,b_2)\times(a_2,b_2)]$ of two-dimensional complex functions $f(x_1,x_2)$, which are integrable on this rectangle. The boundaries of the rectangle can be finite or infinite. We can choose a countable set of functions $B_1=\{e_{1k}, k=\overline{1,\infty}\}$ and $B_2=\{e_{2j}, j=\overline{1,\infty}\}$ as a basis of space with respect to individual coordinates of the form

$$B=\{e_{1k}e_{2j}, k=\overline{1,\infty}, j=\overline{1,\infty}\}. \quad (1)$$

The basic functions in the case of trigonometric functions can be represented in the form

$$e_{nk}=\left(e^{ix_n}\right)^k=e^{ikx_n},\ n=1,2. \quad (2)$$

The expansion of an arbitrary function in the space under consideration with respect to the basis (2) can be regarded as a two-dimensional generalized power series of the form

$$f=\sum_{k,p=1}^{\infty} a_{kp}(e_{11})^k(e_{21})^p. \quad (3)$$

DOI 10.15588/1607-3274-2023-1-10

In [12] we have proposed the definition of the functional of the Padé-type in following form.

Definition. Suppose a two-dimensional power series $S=\sum_{k,p=1}^{\infty}a_{kp}(x_1)^k(x_2)^p$ of complex variables x_1 and x_2 and the associated Padé approximant $P[m_1,n_1/m_2,n_2](x_1,x_2)$ in the proper sense are given. The Padé-type functional $GP_{GS}[m_1,n_1/m_2,n_2](f_1,f_2)$ associated with the given generalized power series $GS=\sum_{k,p=1}^{\infty}a_{kp}(f_1)^k(f_2)^p$ for the complex functions of these variables is defined as

$$GP_{GS}[m_1,n_1/m_2,n_2](f_1,f_2)= \\ = P[m_1,n_1/m_2,n_2](x_1,x_2)\big|_{x_1=f_1,\,x_2=f_2}. \quad (4)$$

The following sequence details the construction process:

1. The types of bases for the individual variables B_1, B_2 and the basis of the space B (1) are chosen.

2. The function f to be approximated is represented in the form (3).

3. For the power series of two complex variables x_1 and x_2 with coefficients coinciding with (3), a Padé approximant $P[m_1,n_1/m_2,n_2](x_1,x_2)$ is constructed in the proper sense.

4. A substitution of basis functions into a functional of Padé-type (4) is performed.

The proposed scheme makes it possible to determine the set of coefficients of a series that is necessary and sufficient for the construction of the Padé-type approximant with a given structure of the numerator and denominator.

If we have a function $f(x_1,x_2)$ which represents brightness of monochrome image point in the range of [0,1] on the rectangle $(a_2,b_2)\times(a_2,b_2)$ in form of truncated Fourier series

$$f(x_1,x_2)\approx\sum_{m=0}^{M}\sum_{n=0}^{N}f_{mn}\cos(m\lambda_1x_1)\cos(n\lambda_2x_2),$$

and it's Padé approximant $P[m_1,n_1/m_2,n_2](x_1,x_2)$, then to obtain cosine part of two-dimensional exponent, we use the following equality:

$$\cos x\cos y=\frac{1}{2}\mathrm{Re}\left(e^{ix}e^{iy}+e^{ix}e^{-iy}\right).$$

As was considered in [18], in this case

$$f(x_1,x_2)\approx\frac{1}{2}\mathrm{Re}\left[P(x_1,x_2)+P(x_1,-x_2)\right]. \quad (5)$$

If the image is stored as a two-dimensional array of points (for example, as a bmp file), discrete Fourier transform procedures can be used to implement Padé approximation [2]. DFT is the basis of many image and video compression algorithms, especially the basic jpeg and mpeg standards for compressing both still and video images.

The input image is decomposed into spectral components using a two-dimensional discrete cosine transform (2D DCT). 2D DCT can be calculated by applying a one-dimensional DCT algorithm for each row or column of a two-dimensional matrix of the input signal, since DCT is a separable function. The direct two-dimensional DCT of a $M\times N$ matrix of a two-dimensional signal $f(x,y)$ can be written as

$$F(u,v)=\frac{2}{\sqrt{MN}}C(u)C(v)\times \\ \times\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}f(n,m)\cos\left(\frac{\pi(2n+1)u}{2N}\right)\cos\left(\frac{\pi(2m+1)v}{2M}\right). \quad (6)$$

where

$$C(u)=\begin{cases}\frac{1}{\sqrt{2}}, \text{ for } u=0,\\ 1, \text{ for } u\neq 0.\end{cases}$$

Coefficients of a truncated cosine Fourier series for $f(x_1,x_2)$ can be obtained as the values of $F(u,v)$ (6), divided by the step.

Next, the range of the frequency values on the integer grid is selected according to the Vavilov method [13]. The size of this range directly determines the number of equations that must be generated.

For this purpose, we use the generalized two-dimensional Padé approximation for case $N=M$ proposed by Chisholm [11]. If $f(x,y)$ is a function of two variables with a two-dimensional expansion into a power series of the form

$$f(x,y)=\sum_{k,p=1}^{\infty}a_{kp}x^k y^p.$$

then the N-th Chisholm approximation can be written as

$$f_{N,N}(x,y)=\frac{\sum_{k,p=1}^{N}p_{kp}x^k y^p}{\sum_{k,p=1}^{N}q_{kp}x^k y^p}.$$

If we use a set of power values bounded by a right triangle with the axes being its legs, then the coefficients p_{kp} and q_{kp} can be calculated with the help of the following equations

DOI 10.15588/1607-3274-2023-1-10

$$\sum_{\sigma=0}^{\gamma}\sum_{r=0}^{\delta} q_{\sigma r} a_{\gamma-\sigma,\delta-r} = p_{\gamma\delta},$$
$$(\gamma,\delta = 0,1,\ldots,2N, 1 \le \delta+\gamma \le 2N),$$
$$\sum_{\sigma=0}^{\gamma}\sum_{r=0}^{\delta}\left(q_{\sigma r} a_{\gamma-\sigma,\delta-r} + q_{r\sigma} a_{\delta-r,\gamma-\sigma}\right) = 0,$$
$$(\gamma = 1,2,\ldots,2N, \delta+\gamma = 2N),$$
$$p_{00} = 1.$$

The solution of this system of equations are matrixes of coefficients p and q.

The algorithm for compression of two-dimensional signals using two-dimensional discrete cosine transformation is shown in Fig. 1.

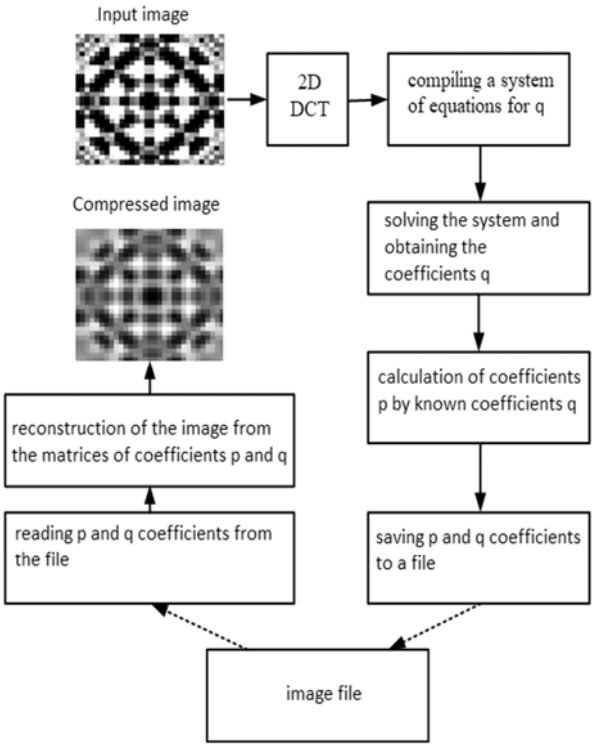


Figure 1 – Example of training samples formation from the original sample for the Fischer problem in the space

4 EXPERIMENTS

Here we consider some examples of Padé approximants application to simple discontinuous template functions for both formulaic and discrete representation. In general case images could be also represented using such series [2].

Let's consider a periodic template function with the period of 2π in the form

$$f(x_1,x_2) = \begin{cases} 1, x_1^2 + x_2^2 \le \pi, \\ 0, x_1^2 + x_2^2 > \pi. \end{cases} \quad (7)$$

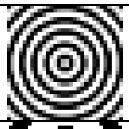
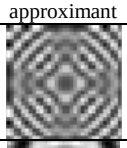
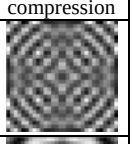
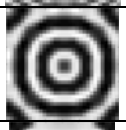
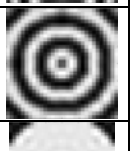
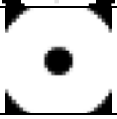
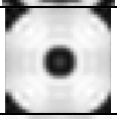
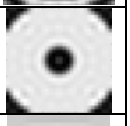
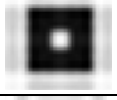
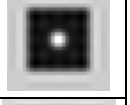
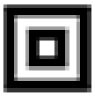
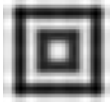
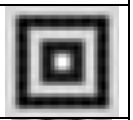
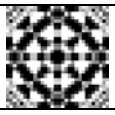
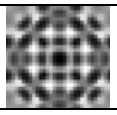
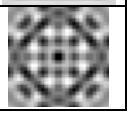
This function is symmetrical about both x_1 and x_2 axes. Thus, the most suitable way of its approximation is a truncated Fourier series in the form

$$f(x_1,x_2) \approx \sum_{m=0}^{4}\sum_{n=0}^{4} f_{mn}\cos(mx_1)\cos(nx_2). \quad (8)$$

We have applied Padé-type approximation to (6) with structure $[2,2/2,2]$, using subset of previously estimated Fourier coefficients (6) and no additional information. To obtain the cosine part of the two-dimensional exponent, we use the converse transformation (5) for Padé-type approximant $P[2,2/2,2](x_1,x_2)$ to obtain the desired real approximation. Transformation which is used for cosine series is the same as the one used in radio physics [10].

We also consider the six monochrome symmetrical bitmap test images from the digital library [20]. These were also used as input data and can be seen in Table 1. In the case of an asymmetric input signal, the image can be artificially expanded to a symmetrical one.

Table 1 – Initial and restored images with the number of harmonics N=8

No	Initial image	Image Padé-type approximant	Image with compression
1			
2			
3			
4			
5			
6			

When using the jpeg standard, insignificant decomposition coefficients are excluded in order to reduce the file size. This procedure was used by processing the results of 2D DCP of the input image, considering the rapid decline of harmonic amplitudes, and it was this image that was used for comparison with the quality of the one compressed by the proposed method. Standardized root mean square error and normalized mean absolute error were used as comparison criteria.

DOI 10.15588/1607-3274-2023-1-10

5 RESULTS

Two-dimensional grayscale images used the template periodic function (7). The resulting truncated Fourier series (8) and it's Padé-type approximant are represented on Fig. 2a, 2b and 2c respectively. For the Fourier series the image demonstrates distortions inherent to Gibbs phenomena, and their effective absence for the Padé-type approximant. It is well seen that the Padé-type approximant is much more visually appropriate than Fourier one.

In order to assess the accuracy of the Fourier series method of and the Padé-type approximation, Fig. 3 presents their one-dimensional sections for comparison with the template function. This also demonstrates the advantage of the Padé approximation.

The size of the area on the integer grid was chosen in the range between 2 and 8, while the number of coefficients by which the reconstruction of the compressed image was performed, was gradually increased. For the compressed image this number (and, therefore, the volume of the graphic file) was approximately half of the value. Two extreme cases are schematically shown in Fig. 4.

A subjective assessment of the quality of the restored image can be obtained from Tab. 1, which shows the input and the reconstructed images.

6 DISCUSSIONS

Analyzing the graphs of the mean square error, one can notice a sharp decrease in the mean square error when the image approaches the psycho-visual similarity to the original. The results of the criteria calculations showed that each type of image has its own lower limit when the reconstructed image visually correlates with the initial one (Table 2).

Table 2 – The number of harmonics for minimum error

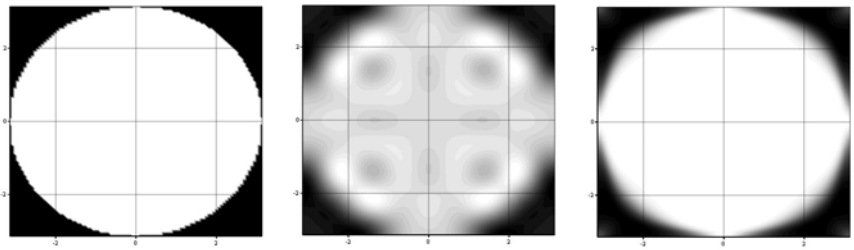
image *No*	1	2	3	4	5	6
the number of harmonics *N*	7	6	4	4	5	8

The study makes it possible to draw some conclusions about the practical use of the Padé-type approximation method and its advantages.

First of all, one can see the low level of noise for the Padé approximation (Fig. 5b) compared to the Fourier series for the cosine (Fig. 5a).

Secondly, the use of Padé-type approximation leads to a sharp decrease in the number of approximant parameters without the loss of accuracy (and even with its increase). Indeed, when using Fourier series with an equal number of *N* harmonics in both directions, the following number of parameters $n_F = N^2$ is obtained. Using the Padé-type approximation with the same powers of the numerator and denominator equal to *N/2* gives the following number of parameters:

$$n_P = 2\left(\frac{N}{2}\right)^2 - 1 = \frac{N^2}{2} - 1 .$$



a b c

Figure 2 – Images of a – the template function, b – Fourier series 4×4, c – Padé Approximant $[2,2/2,2]$.

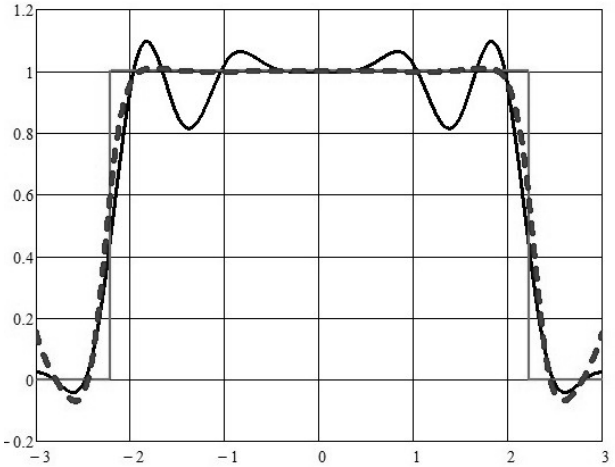
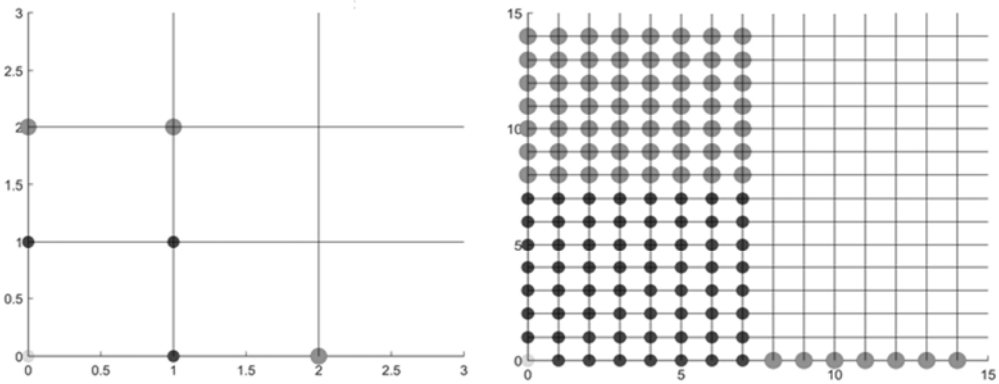


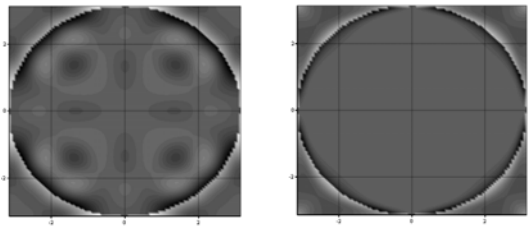
Figure 3 – Cross-sections of approximants and template along line $x_1 = x_2$. Grey line – the template, black – Fourier series, dashed line – Padé approximant.

DOI 10.15588/1607-3274-2023-1-10



a b

Figure 4 – An integer grid of power values for the two extreme cases: a – N=2, b – N=8.



a b

Figure 5 – Distortions between template functions and their approximations for a – Fourier cosine series, b – Padé approximation of cosine series

Thus, the number of the parameters is more than halved:

$$\frac{n_F}{n_P} > 2 .$$

This is very important for the purpose of saving the images in digital signal processing and can provide a theoretical basis for building a new effective image format similar to the well-known jpeg format [1, 2, 10].

CONCLUSIONS

An important problem of applied mathematics is solved in order to reduce the Gibbs phenomenon for the harmonic two-dimensional Fourier series.

The scientific novelty of obtained results is that they demonstrate effective absence of distortions inherent to Gibbs phenomena for the Padé-type approximant. It is well seen that the Padé-type approximant is much more visually appropriate than Fourier one. Application of the Padé-type approximation also leads to the sufficient decrease of the approximants' parameter number without the loss of precision.

The practical significance of obtained results is that the software implementing the proposed method is fit for practical use along with the estimation of the appropriate application conditions.

Prospects for further research are to study the proposed method as a theoretical basis for building a new effective image format similar to the well-known jpeg format.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Dnipro University of Technology "Models and information technologies of data processing and analysis in complex computer systems and networks" (state registration number 0121U114523).

REFERENCES

1. Timan A. F. Theory of approximation of functions of a real variable. New York, MacMillan, 1963, 631 p. https://doi.org/10.1016/c2013-0-05307-8
2. Mitra S. K. Digital Signal Processing: A Computer-Based Approach. New York, McGraw-Hill, 2001, 866 p. https://doi.org/10.1016/s0026-2692(98)00072-x
3. Helmberg G. Localization of a Corner-Point Gibbs Phenomenon for Fourier Series in Two Dimensions, *Journal of Fourier Analysis and Applications*, 2002, Vol. 8(1), pp. 29–42. DOI: 10.1007/s00041-002-0002-9
4. Archibald R. and Gelb A. A method to reduce the Gibbs ringing artifact in MRI scans while keeping tissue boundary integrity, *IEE Transactions of Medical Imaging*, 2002, Vol. 21(4), pp. 305–319. DOI: 10.1109/TMI.2002.1000255
5. Veraart J., Fieremans E., Jelescu I. O., Knoll F., and Novikov D. S. Gibbs ringing in diffusion MRI, *Magn. Reson. Med.*, 2016, Vol. 76, pp. 301–314. DOI: 10.1002/mrm.25866
6. Serov V. Fourier Series, Fourier Transform and Their Applications to Mathematical Physics. New York, Springer International Publishing, 2017, 534 p. DOI: 10.1007/978-3-319-65262-7.

DOI 10.15588/1607-3274-2023-1-10

7. Maggioli F., Melzi S., Ovsjanikov M., Bronstein M. M., Rodolà E. Orthogonalized fourier polynomials for signal approximation and transfer, *Computer Graphics Forum*, 2021, Vol. 40(2), pp. 435–447. https://doi.org/10.1111/cgf.142645
8. Andrianov I., Awrejcewicz J., Danishevskyy V., Ivankov A. Asymptotic Methods in the Theory of Plates with Mixed Boundary Conditions. New York, John Wiley & Sons, 2014, 288 p. DOI: 201410.1002/9781118725184.
9. Olevska Yu. B., Olevskyi V. I., Olevskyi O. V. Using of fuzzy mathematical models in automated systems for recognition of high molecular substances, *Application of Mathematics in Technical and Natural Sciences: 10th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences – AMiTaNS'18, Albena, 20–25 June: proceedings*. New York, American Institute of Physics, Melville, NY, 2018, pp. 060003-1–060003-9. (AIP Conference Proceedings, Vol. 2025(1)). https://doi.org/10.1063/1.5064911
10. Prots'ko I. O., Kuzminskij R. D., Teslyuk V. M. Efficient computation of the integer DCT-II for compressing images, *Radio Electronics, Computer Science, Control*, 2019, No. 2, pp. 151–157. https://doi.org/10.15588/1607-3274-2019-2-16
11. Baker J. A., Jr. and Graves-Morris P. Padé approximants. New York, Cambridge University Press, 1996, 746 p. https://doi.org/10.1017/cbo9780511530074
12. Andrianov I. V., Olevskyi V. I., Shapka I. V., Naumenko T. S. Technique of Padé-type multidimensional approximations application for solving some problems in mathematical physics, *Application of Mathematics in Technical and Natural Sciences: 10th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences – AMiTaNS'18, Albena, 20–25 June, 2018: proceedings.* New York, American Institute of Physics, Melville, NY, 2018, pp. 040002-1–040002-9. (AIP Conference Proceedings, Vol. 2025 (1)). DOI: 10.1063/1.5064886
13. Bosuwan N., López Lagomasino G. Inverse Theorem on Row Sequences of Linear Padé-orthogonal Approximation, *Comput. Methods Funct. Theory*, 2015, Vol. 15, pp. 529–554. https://doi.org/10.1007/s40315-015-0121-3
14. Labych Yu. A., Starovoitov A. P. Trigonometric Padé approximants for functions with regularly decreasing Fourier coefficients, *Sb. Math*, 2009, Vol. 200(7), pp. 1051–1074. DOI: 10.1070/SM2009v200n07ABEH004027
15. Buslaev V. I., Suetin S. P. On the existence of compacta of minimal capacity in the theory of rational approximation of multi-valued analytic functions, *J. Approx. Theory*, 2016, Vol. 206, pp. 48–67. DOI: 10.1016/j.jat.2015.08.002
16. Sablonniere P. Padé-Type Approximants for Multivariate Series of Functions, *Lecture Notes in Mathematics*, 1984, Vol. 1071, pp. 238–251. https://doi.org/10.1007/bfb0099622
17. Kida S. Padé-type and Padé approximants in several variables, *Appl. Numer. Math*, 1989/90, Vol. 6, pp. 371–391. https://doi.org/10.1016/0168-9274(90)90027-D
18. Olevska Yu. B., Olevskyi V. I., Shapka I. V., and Naumenko T. S. Application of two-dimensional Padé-type approximants for reducing the Gibbs phenomenon, *Application of Mathematics in Technical and Natural Sciences: 11th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences – AMiTaNS'19, Albena, 20–25 June, 2019: proceedings*. New York, American Institute of Physics, Melville, NY, 2018, pp. 060014-1–060014-8. (AIP Conference Proceedings, Vol. 2164). https://doi.org/10.1063/1.5130816
19. Daras N. J. The convergence of Padé-type approximants to holomorphic functions of several complex variables, *Appl. Numer. Math*, 1989/90, Vol. 6, pp. 341–360. https://doi.org/10.1016/0168-9274(90)90025-B
20. TESTIMAGES free collection of digital images for testing [Electronic resource]. Access mode: https://testimages.org/Received 00.00.2023.

Received 04.01.2023.
Accepted 11.02.2023.

УДК 004.93

ЗАСТОСУВАННЯ ДВОВИМІРНИХ АПРОКСИМАЦІЙ ТИПУ ПАДЕ ДЛЯ ОБРОБКИ ЗОБРАЖЕНЬ

Олевський В. І. – д-р. техн. наук, професор, професор кафедри інформаційних технологій та комп'ютерної інженерії НТУ «Дніпровська політехніка», Дніпро, Україна.

Гнатушенко В. В. – д-р. техн. наук, професор, завідувач кафедри інформаційних технологій та комп'ютерної інженерії НТУ «Дніпровська політехніка», Дніпро, Україна.

Коротенко Г. М. – д-р. техн. наук, доцент, професор кафедри інформаційних технологій та комп'ютерної інженерії НТУ «Дніпровська політехніка», Дніпро, Україна.

Олевська Ю. Б. – канд. фіз.-мат. наук, доцент, доцент кафедри прикладної математики НТУ «Дніпровська політехніка», Дніпро, Україна.

Обиденний Є.О. – асистент кафедри інформаційних технологій та комп'ютерної інженерії НТУ «Дніпровська політехніка», Дніпро, Україна.

АНОТАЦІЯ

Актуальність. У двовимірному випадку феномен Гіббса значно погіршує обробку зображень для більшості популярних графічних стандартів, оскільки вони використовують кінцеву суму гармонік коли використовується обробка зображення шляхом розкладання сигналу в двовимірний ряд Фур'є з метою зменшення розміру графічного файлу. Тому зменшення цього явища є дуже важливою проблемою.

Мета роботи. Метою роботи є використання двовимірних апроксимацій типу Паде для усунення феномену Гіббса під час обробки зображень та зменшення розміру файлу зображення.

Метод. Ми використовуємо метод двовимірних апроксимацій типу Паде, який ми розробили раніше, щоб зменшити феномен Гіббса для гармонійного двовимірного ряду Фур'є. Запропоновано визначення функціонала типу Паде. Для цього використовується узагальнена двовимірна апроксимація Паде, запропонована Чізхолмом, при цьому діапазон значень частоти на цілочисельній сітці вибирається за методом Вавілова. Запропонована схема дає змогу визначити набір коефіцієнтів

DOI 10.15588/1607-3274-2023-1-10

ряду, необхідний і достатній для побудови апроксимації типу Паде із заданою структурою чисельника та знаменника. Розглядаються деякі приклади застосування апроксимацій Паде до простих розривних шаблонних функцій як для аналітичного, так і для дискретного представлення.

Результати. Наше дослідження дає можливість зробити деякі висновки щодо практичного використання апроксимації типу Паде та її переваг. Вони демонструють практичну відсутність спотворень для апроксиманти типу Паде, властивої саме явищам Гіббса. Добре видно, що апроксимація типу Паде є набагато зручнішою візуально, ніж апроксимація Фур'є. Використання апроксимації типу Паде також призводить до значного зменшення кількості параметрів апроксимантів без втрати точності.

Висновки. Продемонстровано працездатність методики та можливість її застосування для підвищення точності розрахунків. Дослідження дає можливість зробити висновки про переваги практичного використання апроксимації типу Паде.

КЛЮЧОВІ СЛОВА: апроксимації типу Паде, феномен Гіббса, розмір файлу зображення.

ЛІТЕРАТУРА

1. Timan A. F. Theory of approximation of functions of a real variable / A. F. Timan. – New York : MacMillan, 1963. – 631 p. https://doi.org/10.1016/c2013-0-05307-8
2. Mitra S. K. Digital Signal Processing: A Computer-Based Approach / S. K. Mitra. – New York : McGraw-Hill, 2001. – 866 p. https://doi.org/10.1016/s0026-2692(98)00072-x
3. Helmberg G. Localization of a Corner-Point Gibbs Phenomenon for Fourier Series in Two Dimensions / G. Helmberg // Journal of Fourier Analysis and Applications. – 2002. – Vol. 8(1). – P. 29–42. DOI: 10.1007/s00041-002-0002-9
4. Archibald R. A method to reduce the Gibbs ringing artifact in MRI scans while keeping tissue boundary in-tegrity / R. Archibald and A. Gelb // IEE Transactions of Medical Imaging. – 2002. – Vol. 21(4). – P. 305–319. DOI: 10.1109/TMI.2002.1000255
5. Gibbs ringing in diffusion MRI / [J. Veraart, E. Fieremans, I. O. Jelescu et al.] // Magn. Reson. Med. – 2016. – Vol. 76. – P. 301–314. DOI: 10.1002/mrm.25866
6. Serov V. Fourier Series, Fourier Transform and Their Applications to Mathematical Physics / V. Serov. – New York : Springer International Publishing, 2017. – 534 p. DOI: 10.1007/978-3-319-65262-7.
7. Maggioli F. Orthogonalized fourier polynomials for sig-nal approximation and transfer / F. Maggioli, S. Melzi, M. Ovsjanikov et al.] // Computer Graphics Forum. – 2021. – Vol. 40(2). – P. 435–447. https://doi.org/10.1111/cgf.142645
8. Andrianov I. Asymptotic Methods in the Theory of Plates with Mixed Boundary Conditions / I. Andrianov J. Awrejcewicz, V. Danishevskyy, A. Ivankov. – New York: John Wiley & Sons, 2014. – 288 p. DOI: 201410.1002/9781118725184.
9. Olevska Yu. B. Using of fuzzy mathematical models in automated systems for recognition of high molecular substances / Yu. B. Olevska, V. I. Olevskyi, O. V. Olevskyi // Application of Mathematics in Technical and Natural Sciences: 10th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences – AMiTaNS'18, Albena, 20–25 June: proceedings. – New York: American Institute of Physics, Melville, NY, 2018. – P. 060003-1–060003-9. – (AIP Conference Proceedings, Vol. 2025(1)). https://doi.org/10.1063/1.5064911
10. Prots'ko I. O. Efficient computation of the integer DCT-II for compressing images / I. O. Prots'ko, R. D. Kuzminskij, V. M. Teslyuk, // Radio Electronics, Computer Science, Control. – 2019. – No. 2. – P. 151–157. https://doi.org/10.15588/1607-3274-2019-2-16
11. Baker J. A., Jr. Padé approximants / J. A. Baker, Jr. and P. Graves-Morris. – New York: Cambridge University Press, 1996. – 746 p. https://doi.org/10.1017/cbo9780511530074
12. Technique of Padé-type multidimensional approximations application for solving some problems in mathematical physics / [I. V. Andrianov, V. I. Olevskyi, I. V. Shapka, T. S. Naumenko] // Application of Mathematics in Technical and Natural Sciences: 10th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences – AMiTaNS'18, Albena, 20–25 June, 2018: proceedings. – New York : American Institute of Physics, Melville, NY, 2018. – P. 040002-1–040002-9. – (AIP Conference Pro-ceedings, Vol. 2025 (1)). DOI: 10.1063/1.5064886
13. Bosuwan N. Inverse Theorem on Row Sequences of Linear Padé-orthogonal Approximation / N. Bosuwan, López Lagomasino G. // Comput. Methods Funct. Theory. – 2015. – Vol. 15. – P. 529–554. https://doi.org/10.1007/s40315-015-0121-3
14. Labych Yu. A. Trigonometric Padé approximants for functions with regularly decreasing Fourier coefficients / Yu. A. Labych, A. P. Starovoitov // Sb. Math. – 2009. – Vol. 200(7). – P. 1051–1074. DOI: 10.1070/SM2009v200n07ABEH004027
15. Buslaev V. I. On the existence of compacta of minimal capacity in the theory of rational approximation of multivalued analytic functions / V. I. Buslaev, S. P. Suetin // J. Approx. Theory. – 2016. – Vol. 206. – P. 48–67. DOI: 10.1016/j.jat.2015.08.002
16. Sablonniere P. Padé-Type Approximants for Multivariate Series of Functions / P. Sablonniere // Lecture Notes in Mathematics. – 1984. – Vol. 1071. – P. 238–251. https://doi.org/10.1007/bfb0099622
17. Kida S. Padé-type and Padé approximants in several variables / S. Kida // Appl. Numer. Math. – 1989/90. – Vol. 6. – P. 371–391. https://doi.org/10.1016/0168-9274(90)90027-D
18. Application of two-dimensional Padé-type approximants for reducing the Gibbs phenomenon / [Yu. B. Olevska, V. I. Olevskyi, I. V. Shapka, and T. S. Naumenko] / Application of Mathematics in Technical and Natural Sciences: 11th International Conference for Promoting the Application of Mathematics in Technical and Natural Sciences – AMiTaNS'19, Albena, 20–25 June, 2019: proceedings. – New York: American Institute of Physics, Melville, NY, 2018. – P. 060014-1–060014-8. – (AIP Conference Proceedings, Vol. 2164). https://doi.org/10.1063/1.5130816
19. Daras N. J. The convergence of Padé-type approximants to holomorphic functions of several complex variables / N. J. Daras // Appl. Numer. Math. – 1989/90. – Vol. 6. – P. 341–360. https://doi.org/10.1016/0168-9274(90)90025-B
20. TESTIMAGES free collection of digital images for test-ing [Electronic resource]. – Access mode: https://testimages.org/

DOI 10.15588/1607-3274-2023-1-10

UDC 004.93

THE METHOD OF ASSESSING THE VALUE OF INFORMATION

Pilkevych I. A. – Dr. Sc., Professor, Professor of the Department of Computer Information Technologies, Korolov Zhytomyr Military Institute, Zhytomyr, Ukraine.

Vakaliuk T. A. – Dr. Sc., Professor, Professor of the Department of Software Engineering, Zhytomyr Polytechnic State University, Zhytomyr, Ukraine.

Boichenko O. S. – PhD, Head of the research department of the scientific center, Korolov Zhytomyr Military Institute, Zhytomyr, Ukraine.

ABSTRACT

Context. The task of assessing the value of the institution's information as one of the objects of protection of the information security model is considered.

Objective. The goal of the work is the creation of a method of assessing the value of information, which takes into account the time of the final aging of information.

Method. The results of the analysis of methods for evaluating the value of information showed that modern approaches are conventionally divided into two directions. In the first direction, the value of information is calculated as the amount of information in bytes. In the second direction, the value of information is calculated in monetary terms. It is shown that modern approaches do not take into account the influence of time on the value of information. A method of assessing the value of information is proposed, which takes into account such characteristics as the term of final aging of information, the level of its access restriction, importance, and form of ownership. The value of information is presented as a quantitative measure that determines the degree of its usefulness for the owner. It is proposed to calculate the value of the initial value of information during its creation or acquisition by calculating the normalized weight of the coefficients according to the formula of the arithmetic mean. It was shown that the current value of information has a functional dependence on the time of existence of information and the time of its final aging.

Results. The results of the experiment confirm that the value of information has a nonlinear functional dependence on the time of final aging of information.

Conclusions. The conducted experiments confirmed the efficiency of the proposed method of evaluating the value of information and allow recommending it for use in practice to protect the institution's information. Prospects for further research may include the creation of a methodology for assessing the value of an institution's information, taking into account the aging of information and subsequent adjustment of measures to protect it.

KEYWORDS: aging of information, the importance of information, restriction of access to information, value of information.

ABBREVIATIONS

ISO is an International Organization for Standardization;

TMIP is a technical means of information processing;

VOI is a value of information.

NOMENCLATURE

n is the number of information characteristics that affect the value of information;

k_i is a coefficient that characterizes the quantitative measure of the impact of the characteristics of information on its value;

k_1 is a coefficient that characterizes the influence of the level of restriction of access to information on the value of information;

k_2 is a coefficient that characterizes the influence of the time of final aging of information on the value of information;

k_3 is a coefficient that characterizes the influence of the importance of information on the value of information;

k_4 is a coefficient that characterizes the influence of the form of ownership of information on the value of information;

k_5 is a coefficient that characterizes the influence of the method of storing information on the value of information;

L is the number of levels of restriction of access to information in the institution;

l_i is a quantitative assessment of the level of restriction of access to information in the institution;

I the number of levels of importance of the institution's information

i_j is a quantitative assessment of the appropriate level of importance of information of the institution;

F the number of forms of ownership of information in the institution;

f_j is a quantitative assessment of the appropriate form of ownership of information of the institution;

S the number of methods of storing information used in the institution;

s_j quantitative assessment of the method of information storage;

VOI_0 the initial value of the information of the institution (at the time of its occurrence or receipt);

t is a time from the moment of information occurrence to the moment of determining its value;

t_1 is a time from the moment of information occurrence to the moment of its final aging.

DOI 10.15588/1607-3274-2023-1-11

INTRODUCTION

In modern society, the role of information and information resources has significantly increased in all spheres of human life. The transformation of information into a product that has a certain value and corresponding value has led to the emergence of a new object of security – information and information resources. Previously, information security consisted of protecting information and information resources from unauthorized actions. At present, there is a need to protect people, society, or the state from threats that may pose information and information resources. Thus, today the threats that information and information resources can carry affect such aspects of human life as economic, financial, military, technical, and political. The emergence of such threats leads to the improvement of the conceptual model of information security.

The conceptual model of information security conditionally consists of security objects, a model of threats, a model of violators, and an information security system [1–2]. Recently, institutions have increased the cost of developing their information security model. However, not all institutions separately assess the value of information of the institution as one of the objects of security. It also does not take into account the fact that over time, the information loses its value and its further protection becomes impractical.

The object of study is the process of assessing the value of information.

The subject of study is the method of assessing the value of information.

Known methods [8–10] do not take into account the aging time of information, which has a direct impact on the value of information.

The purpose of the work is to develop a method for assessing the value of information, which takes into account the time of final aging of information.

1 PROBLEM STATEMENT

Suppose a set of coefficients $<k_i>$ is specified, which characterize the quantitative measure of the influence of information characteristics on its value and take values from the interval [0…1], $i = 1, 2, \ldots, N$.

For a given set of coefficients $<k_i>$, the problem of evaluating the value of information can be represented as the problem of calculating the normalized weight of the coefficients according to the formula of the arithmetic mean.

2 REVIEW OF THE LITERATURE

The authors in the scientific works [3–7] reflected on the results of the analysis of the implementation of information security in institutions and proposed to use the international standard of information security ISO/IEC 27001 to manage information-related risks.

In the scientific paper [8] the authors defined the multiple criteria value of information and demonstrated the potential application when conservation issues conflict with monetary issues.

In the scientific work [9] the authors proposed a new approach to assessing the value of information based on the theory of pattern recognition, which expands its scope and can be successfully implemented using declarative programming languages or universal modeling languages.

An approach to calculating the value of information obtained from the example of a fuzzy mathematical model of the queuing system is presented in a scientific paper [10].

In scientific works [8–10] assessment of the value of information is realized using the methods of information theory, where the main property of information is its quantity.

The method of assessing the information potential with the use of coefficient, cost, and effective methods makes it possible to identify reserves of rational information support of enterprises, to determine the cost of information resources and the effectiveness of their use. This technique is proposed in a scientific paper [11] and can be used by the information security service of institutions to determine the cost of information security of the institution.

The authors in the scientific work [12] proposed to use the coefficient of permissible change, which characterizes the losses of the institution that will not lead to its bankruptcy. This ratio is characterized by the ratio of the growth/decline of the capital of the institution to the growth/decline of the amount of damage from the implementation of information threats.

In the results of the research, which reflected in the scientific work [13], the authors proposed a model for assessing the level of protection of information in the social network from external influences on the information social resource. The result is an assessment of the economic feasibility of implementing an appropriate mechanism of technical means of information protection in social networks, depending on the value of information.

Therefore, today there are two known approaches to assessing the value of information: calculating the amount of information in bytes and calculating the value of information in monetary terms. The first approach makes it possible to protect information without taking into account the level of restriction of access, the level of importance, or the form of ownership of information. The second approach implements the protection of individuals, institutions, and the state from damage that may occur in the event of unauthorized actions with information. The second approach also ensures the integrity, confidentiality, and accessibility of information.

The above approaches do not take into account the influence of time on the relevant properties of information. It is a known fact that the time of existence of information affects its value.

DOI 10.15588/1607-3274-2023-1-11

3 MATERIALS AND METHODS

The value of information in this study should be understood as a quantitative measure that determines the degree of its usefulness to the owner of the information. The functional dependence of the value of information on its characteristics is reflected in the expression [14–17]:

$$VOI = \frac{\sum_{i=1}^{n} k_i}{n}. \quad (1)$$

Assessing the value of information based on the methods and techniques of the modern theory of systems analysis provides tools for determining the appropriate coefficients based on the characteristics of information. The coefficients take values from the segment [0 … 1].

The following characteristics of the information are chosen to assess the value of the institution's information.

1. Level of restriction of access to information.

Restricted information, confidential and/or public information may be processed in the institution. The impact of the level of restriction of access to information on its value is determined by the coefficient k_1:

$$k_1 = \frac{l_i}{L}, \quad (2)$$

where $i = [1...L]$.

The level of restriction of access to information is determined by documents on the organization of information security.

2. Period of final aging of information.

The time of final aging of information depends on the level of access restriction and cannot be longer than the time during which the level of access restriction exists. The impact of the time of final aging of information on its value is determined by the coefficient k_2:

$$k_2 = \frac{Ol_i}{OL}, \quad (3)$$

where $i = [1...OL]$.

3. Importance of information.

The importance of information of the institution should be understood as such information, the loss or unauthorized access to which will cause great damage to the institution or completely stop its work [17]. It is advisable to have several levels of importance of information in the institution. In this case, all information in an institution must have its level of importance (rank). A Group of experts determines the level of importance of information.

The impact of the importance of information on its value is determined by the coefficient k_3:

$$k_3 = \frac{i_j}{I}. \quad (4)$$

where $j = [1...I]$.

4. Form of ownership of information.

The information circulating in the institution may have the following forms of ownership: private, collective, and public. The form of ownership of information must have its level of importance in terms of information security. It is expedient to introduce in the institution an additional division of the collective form of ownership of information into collective information of the institution and collective information of structural units of the institution. The impact of the form of ownership of information on its value is determined by the coefficient k_4:

$$k_4 = \frac{f_j}{F}, \quad (5)$$

where $j = [1...F]$.

A Group of experts determines the form of ownership of information.

5. Method of storing information.

The method of storing information determines on which media it is stored. Depending on the level of restriction of access to information, its importance is determined by the method of storing information. A Group of experts determines the method of storing information.

The impact of the method of storing information on its value is determined by the coefficient k_5:

$$k_5 = \frac{s_j}{S}, \quad (6)$$

where $j = [1...S]$.

The value of information changes over time. As a rule, the value of information decreases over time. The dependence of the value of information on time is determined [14–16]:

$$VOI(t) = VOI_0 \cdot 10^{-\frac{t}{t_1}}. \quad (7)$$

4 EXPERIMENTS

An example is considered to verify the method of assessing the value of information of the institution. The institution processes information with the following levels of access: unclassified, confidential, secret, and top secret. According to expression (2), the values of the coefficients are obtained, which are shown in Table 1.

Table 1 – The value of the coefficient k_1

Level of restriction of access to information	Rank	Value
Unclassified	4	0.25
Confidential	3	0.33
Secret	2	0.5
Top secret	1	1

DOI 10.15588/1607-3274-2023-1-11

The time of final aging of information according to the appropriate levels of access is set in the institution. According to expression (3), the values of the coefficients are obtained, which are shown in Table 2.

Table 2 – The value of the coefficient k_2

Level of restriction of access to information	Period of final aging	Rank	Value
Unclassified	362	4	0.25
Confidential	1086	3	0.33
Secret	1810	2	0.5
Top secret	3620	1	1

The importance of information in the institution is classified according to the following levels: insignificant, useful, important, and very important [17]. According to expression (4), the values of the coefficients obtained, are shown in Table 3.

Table 3 – The value of the coefficient k_3

Level of restriction of access to information	Period of final aging	Rank	Value
Unclassified	362	4	0.25
Confidential	1086	3	0.33
Secret	1810	2	0.5
Top secret	3620	1	1

The following forms of ownership of information are established in the institution: personal, department of the institution, institution, and state [18]. According to expression (5), the values of the coefficients are obtained, which are shown in Table 4.

Table 4 – The value of the coefficient k_4

Level of restriction of access to information	Rank	Value
Personal	4	0.25
Department of the institution	3	0.33
Institution	2	0.5
State	1	1

The following technical solutions are used to store information in the institution: server, technical means of information processing, and portable storage. According to expression (6), the values of the coefficients are obtained, which are shown in Table 5.

Table 5 – The value of the coefficient k_5

Method of storing information	Rank	Value
Server	3	0.33
Technical means of information processing (TMIP)	2	0.5
Portable storage	1	1

The institution has compiled a list of information in need of protection, which is shown in Table 6.

Substituting the data from tables 1–5 to expression (7) we obtain the initial value of information and the value of information as of 27.01.2022. Calculated values of coefficients that affect the value of information, the initial value of information, and the current value of information are shown in Table 7.

Table 6 – The list information of the institution

Title of the document	Date of create	Period of final aging	Access restriction stamp	Importance	Form of ownership of information	Method of storing information
Activity plan of the department of the institution	05.01.2022	362	Unclassified	Insignificant	Department of the institution	TMIP
A personal plan of the employee of the institution	10.01.2022	362	Unclassified	Insignificant	Personal	TMIP
Enterprise development strategy	05.01.2020	1810	Secret	Important	Institution	Portable storage
Industry development plan	09.01.2021	3620	Top secret	Very important	State	Portable storage
Report on the results of department 1 of the institution for 2021	28.12.2021	1086	Confidential	Useful	Department of the institution	Server
Report on the results of department 2 of the institution for 2021	28.12.2021	3620	Secret	Important	Department of the institution	TMIP
Report on the results of department 1 of the institution for 2020	29.12.2020	1086	Confidential	Useful	Department of the institution	server
Report on the results of department 1 of the institution for 2019	28.12.2019	1086	Confidential	Useful	Department of the institution	Server

Table 7 – Initial and current value of information

Title of the document	k_1	k_2	k_3	k_4	k_5	Initial value	Current value
Activity plan of the department of the institution	0.1	0.25	0.25	0.33	0.5	0.29	0.25
A personal plan of the employee of the institution	0.1	0.25	0.25	0.25	0.5	0.27	0.24
Enterprise development strategy	0.5	0.5	0.5	0.5	1	0.6	0.23
Industry development plan	1	1	1	1	1	1	0.78
Report on the results of department 1 of the institution for 2021	0.33	0.33	0.33	0.33	0.33	0.33	0.31
Report on the results of department 2 of the institution for 2021	1	0.5	0.5	0.5	0.5	0.5	0.56
Report on the results of department 1 of the institution for 2020	0.33	0.33	0.33	0.33	0.33	0.33	0.14
Report on the results of department 1 of the institution for 2019	0.33	0.33	0.33	0.33	0.33	0.33	0.065

DOI 10.15588/1607-3274-2023-1-11

5 RESULTS

The study yielded the following results:

1. Assessment of the initial value of information carried out by calculating the average value of the sum of the relevant coefficients. Each of the selected coefficients was calculated using the ranking method. The rank of the relevant information characteristics is determined from the guidance documents on the organization of information security or a specially created group of experts.

The assessment of the current value of information is carried out by taking into account the time of final aging of information and the date of creation (receipt) of relevant information. The value of the current value of information is calculated as the product of the initial value of information by the power factor, which characterizes the aging process of information.

2. In the method of estimating the value of information, a coefficient is calculated that takes into account the place of storage of relevant information. The need to introduce this factor was to take into account additional organizational measures aimed at restricting access to information. Thus, to obtain information stored on removable media, an attacker must obtain it from the information protection service. This approach introduces additional controls and limits the list of people who can carry out an insider threat.

3. Each institution has a list of information that needs protection. According to the level of restriction of access to information, the time of final aging of information is set, as well as its form of ownership and level of importance for the owner. For information from this list, its initial value is calculated.

4. The results of the experiment showed that for information that has the same values as the initial value, over time, the value of the current value of the information decreases. For the information contained in the documents of the first department of the institution, namely in the activity reports for the year, the initial value of the information is 0.33. A year later, the value of information decreased almost 2.4 times, and after 2 years almost 5 times. The results of the experiment confirm that the value of information has a nonlinear functional dependence on the time of final aging of information.

6 DISCUSSION

The results of data analysis in Table 7 show that the initial values of information value for a typical document of one department of the institution with identical details, except for the date of creation, are the same. Also, the initial value of information has the same values. At the same time, the current value of the value of information decreases as the time of existence of this information increases. This is explained by the fact that when the time of existence of information approaches the time of final aging of information, the importance of this information is lost.

Taking into account the time of final aging of information for different levels of access to information, (Table 3) will make it possible to evaluate the value of information and provide information to the head of the institution to make a decision on the feasibility of further expenses for the protection of relevant information.

The method of assessing the value of information consists in finding the initial value of the value of information by calculating the average arithmetic value of the coefficients characterizing the quantitative measure of the influence of the characteristics of the information on its value. When the number of such coefficients increases, the adequacy of the model, which is used to calculate the value of information, will also increase.

CONCLUSIONS

The scientific novelty of obtained results is that the method of assessing the value of information was improved. This method allows obtaining a quantitative value of information, taking into account the level of restriction of access to information, the level of importance of information, period of final aging, method, and place of storage, as well as the form of ownership of information. The proposed method provides an opportunity to automate the process of assessing the value of information on the current date using the mathematical apparatus of the modern theory of systems analysis.

The mathematical model used in the method of assessing the value of information provides an opportunity to investigate the value of information that belongs to the person of the institution and the staff of the institution.

The practical orientation of the study is to use the developed method in the information security service of the institution to assess the value of information of the institution and in deciding on the choice of an adequate method of protection of relevant information.

Prospects for further research are to study the impact of the proposed set of coefficients based on the characteristics of information for a broad class of practical problems in information security.

ACKNOWLEDGEMENTS

The author expresses gratitude to Ruslan Hryshchuk, Doctor of technical science, professor for research support and a fruitful paper discussion.

REFERENCES

1. Pevnev V., Tsuranov M., Zemlianko H., Amelina O. Conceptual Model of Information Security, *Integrated Computer Technologies in Mechanical Engineering*, 2020, Vol. № 188, pp. 158–168. DOI: 10.1007/978-3-030-66717-7_14.
2. Onyshchenko S., Yanko A., Hlushko A., Sivitska S. Conceptual Principles of Providing the Information Security of the National Economy of Ukraine in the Conditions of Digitalization, *International Journal of Managemen*, 2020, № 11(12), pp. 1709–1726. DOI: 10.34218/IJM.11.12.2020.157.
3. Hasan Shaikha, Ali Mazen, Kurnia Sherah, Thurasamy Ramayah Evaluating the cyber security readiness of organizations and its influence on performance, *Journal of Information Security and Applications*, 2021, Vol. 58, P. 102726. DOI:10.1016/j.jisa.2020.102726.

DOI 10.15588/1607-3274-2023-1-11

4. Palko D., Myrutenko L., Babenko T., Big-dan A. Model of Information Security Critical Incident Risk Assessment, *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&*T), 2020, pp. 157–161. DOI: 10.1109/PICST51311.2020.9468107.
5. Fazlida M. R., Said Jamaliah Information Security: Risk, Governance and Implementation Setback, *Procedia Economics and Finance*, 2015, Vol. 28, pp. 243–248. DOI: doi.org/10.1016/S2212-5671(15)01106-5.
6. Mirtsch Mona, Blind Knut, Koch Claudia, Dudek Gabriele Information security management in ICT and non-ICT sector companies: A preventive innovation perspective, *Computers & Security*, 2021, Vol. 109, P. 102383. DOI: doi.org/10.1016/j.cose.2021.102383.
7. Aven T. Risk assessment and risk management: Review of recent advances on their foundation, *European Journal of Operational Research*, 2016, Vol. 253, Issue 1, pp. 1–13. DOI: doi.org/10.1016/j.ejor.2015.12.023.
8. Eyvindson K., Hakanen J., Mönkkönen M., Juutinen A., Karvanen J. Value of information in multiple criteria decision making: an application to forest conservation, *Stochastic Environmental Research and Risk Assessment*, 2019, № 33, pp. 2007–2018. DOI: 10.1007/s00477-019-01745-4.
9. Zaiats V. M. and Zaiats M. M. The figurative approach to calculate the amount of information and estimates its values, Visnyk Natsionalnoho universytetu "Lvivska politekhnika", 2017, 872, pp. 93–100. (Serie: Informatsiini systemy ta merezhi).
10. Zaiats V. M., Rybytska O. M., Zaiats M. M. An approach to evaluating the values and quantity of information in queueing systems based on pattern recognition and fuzzy sets theories, *Kibernetika ta sistemnij analiz*, 2019, Vol. 55, № 4. pp. 133–144.
11. Pererva P.G. Informational activity of the enterprise: management, price and marketing composition, Bulletin of the *National Technical University "KhPI"* (economic sciences), 2018, № 37 (1313), pp. 120–125.
12. Mokhor V., Davydiuk A. Approach of the information properties destruction risks assessing based on the color scale, *Information Technology and Security*, 2020, Volume 8, Issue 2, pp. 216–223. DOI: doi.org/10.20535/2411-1031.2020.8.2.222608.
13. Laptiev O., V. Sobchuk, A. Sobchuk, S. Laptiev, T. Laptieva Improved model of estimating economic expenditures on the information protection system in social networks, *Electronic Professional Scientific Edition "Cybersecurity: Education, Science, Technique"*, 2020, № 4(12), pp. 19–28. DOI: 10.28925/2663-4023.2021.12.1928.
14. Sawatnatee A., Prakancharoen S. Insider Threat Detection and Prevention Protocol: ITDP, *International Journal of Online and Biomedical Engineering*, 2021, Vol. 17, № 02, pp. 69–89. DOI: doi.org/10.3991/ijoe.v17i02.18297
15. Hmelevskoy R. Research on information security threat assessment of information activity objects, *Modern Information Security*, 2016, № 4, pp. 65–70.
16. Gulak G. M. Metodolohiia zakhystu informatsii. Aspekty kiberbezpeky: pidruchnyk, Kyiv, Vydavnytstvo NA SB Ukrainy, 2020, P. 256.
17. Korchenko O. H., Arkhypov O. Ye., Dreis Yu. O. Otsiniuvannia shkody natsionalnii bezpetsi Ukrainy u razi vytoku derzhavnoi taiemnytsi: Monohrafiia. Kyiv, Nauk.-vyd. tsentr NA SB Ukrainy, 2014, P. 332.
18. Horne C. A., Maynard S. B., Ahmad A. Information security strategy in organisations: review, discussion and future research, *Australasian Journal of Information Systems*, 2014, Vol. 21. DOI: doi.org/10.3127/ajis.v21i0.1427

Received 22.11.2022.
Accepted 05.02.2023.

УДК 004.93

МЕТОД ОЦІНЮВАННЯ ЦІННОСТІ ІНФОРМАЦІЇ

Пількевич І. А. – д-р техн. наук, професор, професор кафедри комп'ютерних інформаційних технологій Житомирського військового інституту імені С. П. Корольова, Житомир, Україна.

Вакалюк Т. А. – д-р педагогічних наук, професор, професор кафедри інженерії програмного забезпечення Державного університету «Житомирська політехніка», Житомир, Україна.

Бойченко О. С. – канд. техн. наук, начальник науково-дослідного відділу наукового центру Житомирського військового інституту імені С. П. Корольова, Житомир, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу оцінювання цінності інформації установи, як одного з об'єктів захисту моделі інформаційної безпеки.

Мета роботи – створення методу оцінки цінності інформації, що враховує час остаточного старіння інформації.

Метод. Результати аналізу методів оцінювання цінності інформації показали, що сучасні підходи умовно поділяються на два напрямки. У першому напрямку цінність інформації обчислюється як кількість інформації в байтах. У другому напрямку цінність інформації обчислюється в грошовому еквіваленті. Показано, що сучасні підходи не враховують вплив часу на цінність інформації. Запропоновано метод оцінки цінності інформації, який враховує такі характеристики, як термін остаточного старіння інформації, рівень її обмеження доступу, важливість і форма власності. Цінність інформації представлена як кількісна міра, яка визначає ступінь її корисності для власника. Пропонується розраховувати величину початкової вартості інформації під час її створення чи отримання шляхом розрахунку нормованої ваги коефіцієнтів за формулою середнього арифметичного. Показано, що поточна цінність інформації має функціональну залежність від часу існування інформації та часу її остаточного старіння.

Результати. Результати експерименту підтверджують, що цінність інформації має нелінійну функціональну залежність від часу остаточного старіння інформації.

Висновки. Проведені експерименти підтвердили працездатність запропонованого методу оцінювання цінності інформації та дозволяють рекомендувати його для використання на практиці для захисту інформації установи. Перспективи подальших досліджень можуть включати створення методології оцінювання цінності інформації установи з урахуванням часу старіння інформації та подальшим коректуванням заходів із її захисту.

DOI 10.15588/1607-3274-2023-1-11

КЛЮЧОВІ СЛОВА: старіння інформації, важливість інформації, обмеження доступу до інформації, цінність інформації.

ЛІТЕРАТУРА

1. Conceptual Model of Information Security / [V. Pevnev, M. Tsuranov, H. Zemlianko, O. Amelina] //Integrated Computer Technologies in Mechanical Engineering – 2020. – Vol. № 188. – P. 158–168. DOI: 10.1007/978-3-030-66717-7_14.
2. Conceptual Principles of Providing the Information Security of the National Economy of Ukraine in the Conditions of Digitalization / [S.Onyshchenko, A. Yanko, A.Hlushko, S. Sivitska] // International Journal of Management – 2020. – № 11(12). – P. 1709–1726. DOI: 10.34218/IJM.11.12.2020.157.
3. Evaluating the cyber security readiness of organizations and its influence on performance / [Shaikha Hasan, Mazen Ali, Sherah Kurnia, Ramayah Thurasamy] // Journal of Information Security and Applications. – 2021. – Vol. 58. – P. 102726. DOI:10.1016/j.jisa.2020.102726.
4. Model of Information Security Critical Incident Risk Assessment / [D. Palko, L. Myrutenko, T. Babenko, A. Bigdan] // 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), 2020. – P. 157–161. DOI: 10.1109/PICST51311.2020.9468107.
5. Fazlida M. R. Information Security: Risk, Governance and Implementation Setback / M. R. Fazlida, Jamaliah Said // Procedia Economics and Finance. – 2015. – Vol. 28. – P. 243–248. DOI: doi.org/10.1016/S2212-5671(15)01106-5.
6. Information security management in ICT and non-ICT sector companies: A preventive innovation perspective / [Mona Mirtsch, Knut Blind, Claudia Koch, Gabriele Dudek] // Computers & Security. – 2021. – Vol. 109. – P. 102383. DOI: doi.org/10.1016/j.cose.2021.102383.
7. Aven T. Risk assessment and risk management: Review of recent advances on their foundation / T. Aven // European Journal of Operational Research. – 2016. – Vol. 253, Issue 1. – P. 1–13. DOI: doi.org/10.1016/j.ejor.2015.12.023.
8. Value of information in multiple criteria decision making: an application to forest conservation / [K. Eyvindson, J. Hakanen, M. Mönkkönen et al.] // Stochastic Environmental Research and Risk Assessment. – 2019. – № 33. – P. 2007–2018. DOI: 10.1007/s00477-019-01745-4.
9. Заяць В. М. Образний підхід до обчислення кількості інформації та оцінки її цінності / В. М. Заяць, М. М. Заяць // Вісник Національного університету «Львівська політехніка», 2017– № 872. – С. 93–100. – (Серія: Інформаційні системи та мережі).
10. Заяць В. М. Підхід до оцінювання цінності та кількості інформації в системах масового обслуговування на основі теорії розпізнавання образів та нечітких множин / В. М. Заяць, О. М. Рибицька, М. М. Заяць // Кибернетика и системнй анализ. – 2019. – Том 55, № 4. – С. 133–144.
11. Перерва П. Г. Інформаційна діяльність підприємства: управлінська, цінова та маркетингова складові / П. Г. Перерва // Вісник НТУ «ХПІ». Серія: Економічні науки. – 2018. – № 37 (1313). – С. 120–125.
12. Мохор В. Спосіб оцінювання ризиків порушення властивостей інформації за колірною шкалою / В. Мохор, А. Давидюк // Information Technology and Security. – 2020. – Vol. 8, Issue 2. – P. 216–223. DOI: doi.org/10.20535/2411-1031.2020.8.2.222608.
13. Удосконалена модель оцінювання економічних витрат на систему захисту інформації в соціальних мережах / [О. А. Лаптєв, В. В. Собчук, А. В. Собчук та ін.]// Кібербезпека: освіта, наука і техніка. – 2020. – № 4(12). – С. 19–28. DOI: 10.28925/2663-4023.2021.12.1928.
14. Sawatnatee A. Insider Threat Detection and Prevention Protocol: ITDP / A. Sawatnatee, S. Prakancharoen // International Journal of Online and Biomedical Engineering. – 2021. – Vol. 17, № 02. – P. 69–89. DOI: doi.org/10.3991/ijoe.v17i02.18297
15. Хмелевський Р. М. Дослідження оцінки загроз інформаційній безпеці об’єктів інформаційної діяльності / Р. М. Хмелевський // Сучасний захист інформації. – 2016. – № 4. – С. 65–70.
16. Гулак Г. М. Методологія захисту інформації. Аспекти кібербезпеки : підручник / Г. М. Гулак. – К. : Видавництво НА СБ України, 2020. – 256 с.
17. Корченко О. Г. Оцінювання шкоди національній безпеці України у разі витоку державної таємниці : монографія / О. Г. Корченко, О. Є. Архипов, Ю. О. Дрейс. – К. : Наук.-вид. центр НА СБ України, 2014. – 332 с.
18. Horne C. A. Information security strategy in organisations: review, discussion and future research / C. A. Horne, S. B. Maynard, A. Ahmad // Australasian Journal of Information Systems. – 2014. – Vol. 21. DOI: doi.org/10.3127/ajis.v21i0.1427

DOI 10.15588/1607-3274-2023-1-11

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2023. № 1
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2023. № 1

УДК 004.9

ТЕХНОЛОГІЯ ВИПРАВЛЕННЯ ГРАМАТИЧНИХ ПОМИЛОК В УКРАЇНОМОВНОМУ ТЕКСТОВОМУ КОНТЕНТІ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Холодна Н. М. – магістр кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

Висоцька В. А. – канд. техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Більшість досліджень у напрямі виправлення граматичних та стилістичних помилок зосереджені на корекції помилок в англомовному текстовому контенті. Завдяки наявності великих наборів даних досягнуто суттєвого підвищення точності корекції граматики англійської мови. На жаль, досліджень інших мов мало. Системи в для англійської мови постійно розвиваються і наразі активно використовують методи машинного навчання: класифікацію (sequence tagging) та машинний переклад. Для створення якісної моделі машинного навчання для корекції граматичних/стилістичних помилок у текстах тих мов, які є складними морфологічно, необхідна велика кількість паралельних або вручну розмічених даних. Ручна анотація даних вимагає багато зусиль професійних лінгвістів, що робить створення корпусів текстів, особливо морфологічно багатих мов, зокрема, української, часо- та ресурсозатратним процесом.

Мета – є розроблення технології виправлення помилок в українськомовних текстах на основі методів машинного навчання з використанням невеликого набору анотованих паралельних даних.

Метод. Для даного дослідження при розробці системи корекції помилок в україномовних текстах із застосуванням оптимального конвеєру (pipeline), що включає в себе попереднє опрацювання текстового контенту, вибір та генерування ознак, обрані алгоритми машинного навчання, в умовах наявності невеликих за обсягом корпусів анотованих даних. Застосування нейронних мереж з новою архітектурою, огляд state-of-the-art методів та порівняння різних етапів конвеєру дасть змогу визначити таку їх комбінацію, яка дозволить отримати якісну модель корекції помилок в україномовних текстах.

Результати. Розроблено модель машинного навчання для корекції помилок в україномовних текстах. Запропоновано універсальну схему розробки системи корекції помилок для різних мов. Відповідно до отриманих результатів, нейронна мережа має здатність виправляти прості речення, написані українською, однак розроблення повноцінної системи вимагатиме застосування перевірки орфографії за допомогою словників і перевірки правил, як простих, так і заснованих на результаті парсингу залежностей або інших ознак. З-поміж трьох моделей, найкращі показники має попередньо навчена модель нейронного перекладу mT5. З метою економії обчислювальних ресурсів можливим також є застосування попередньо навченої нейронної мережі типу BERT, використовуючи її як у якості енкодера, так і декодера. Така нейронна мережа має вдвічі менше параметрів, ніж інші попередньо навчені моделі машинного перекладу, і показує задовільні результати при виправленні граматичних та стилістичних помилок.

Висновки. Створена модель показує відмінні результати класифікації на тестових даних. Розраховані метрики якості машинного перекладу дають змогу лише частково порівняти моделі, оскільки більшість слів і словосполучень у початковому та виправленому реченні співпадають. Найкраще значення як BLEU (0.908) , так і METEOR (0.956) отримано для mT5, що співпадає із аналізом прикладів, у якому найбільш точні виправлення помилок без зміни початкового значення речення отримані для такої нейронної мережі. M2M100 має більшу оцінку BLEU (0.847), ніж “Ukrainian Roberta” Encoder-Decoder (0.697), однак, суб’єктивно оцінюючи результати виправлення прикладів, M2M100 значно гірше справляється із подібним завданням, ніж дві інші моделі. Для METEOR також M2M100 (0.925) має більшу оцінку, ніж “Ukrainian Roberta” Encoder-Decoder (0.876).

КЛЮЧОВІ СЛОВА: NLP, text pre-processing, корекція помилок, виправлення граматичних помилок, машинне навчання, глибинне навчання, аналіз тексту, класифікація тексту, нейронна мережа.

АБРЕВІАТУРА

БД – база даних;
ІС – інтелектуальна система;
ІТ – інформаційна технологія;
ПЗ – програмне забезпечення;
ПО – предметна область;
BERT – bidirectional encoder representations from transformers;
GEC – grammatical error correction;
GECS – grammatical error correction system;
LSTM – long short-term memory;
MT – machine translation;
ML – machine learning;
NLP – natural language processing;
NN – neural network;
TPP – text pre-processing.

НОМЕНКЛАТУРА

S – система граматичної корекції;
I – множина вхідних даних;
O – множина вихідних даних;
R – основні правила опрацювання потоку вхідних даних в ІС граматичної корекції;
U – параметри опрацювання вхідних даних;
N – нейронна мережа;
α – оператор скачування вхідних даних;
β – оператор опрацювання вхідних даних;
γ – оператор збереження вхідних даних;

DOI 10.15588/1607-3274-2023-1-12

μ – TPP-оператор;

χ – оператор пошуку помилок;

ω – оператор машинного навчання ІС на достовірних текстових даних;

λ – оператор граматичної корекції тексту;

i_1 – множина даних ідентифікації;

i_2 – множина вхідного текстового контенту;

i_3 – множина шаблонів/правил помилок;

i_4 – підтвердження правки від автора/користувача;

o_1 – маркований/тегований текст з помилками;

o_2 – колекція пропозицій корекції тексту;

o_3 – множина підтверджених автором правок;

r_1 – правила алгоритму взаємодії;

r_2 – NLP-правила;

r_3 – правила алгоритму нейронної мережі;

r_4 – правила алгоритму корекції помилок;

u_1 – множина рівнів доступу;

u_2 – множина вимог доступу;

u_3 – множина NLP-вимог;

u_4 – множина метрик машинного навчання;

u_5 – множина вимог корекції помилок.

ВСТУП

GEC – задача ідентифікації та усунення помилок у вхідному тексті. GEC застосовують в різних сферах, включаючи виправлення пошукових запитів і MT-результатів, TPP, перевірку правопису в браузерах і текстових процесорах тощо. GEC-методи поділяють на категорії: методи, засновані на правилах; методи, засновані на синтаксичному аналізі речень; статистичне моделювання; класичні ML-методи; MT на основі глибинного навчання.

Перевірка на основі правил використовує набір попередньо визначених шаблонів помилок для відповідності тексту. Всі правила розробляють зазвичай вручну. Текст є помилковим, якщо відповідає одному з правил [1]. Переваги підходу: швидкодія, інтерпретація результатів, можливості ітеративного розвитку ІС. Однак метод має недоліки: складність ІС збільшується в міру появи різних типів помилок, створення правил є ресурсозатратним і вимагає експертних знань з ПО, зокрема лінгвістики. Одночасно для покриття усіх можливих випадків необхідна величезна кількість правил.

При перевірці на основі синтаксису повністю аналізується морфологія та синтаксис тексту. Для цього потрібна лексична БД, морфологічний і синтаксичний аналізатори (парсери). Залежно від граматики мови, синтаксичний парсер визначає синтаксичну структуру кожного речення у вигляді дерева. Якщо повний аналіз не був успішним, тоді текст є помилковим [1]. Недоліком синтаксичного підходу є необхідність розробки додаткових правил для уточнення необхідних виправлень. Ці правила мають покривати усі можливі варіанти помилок.

Для автоматичного отримання правил із великої кількості тексту використовують статистичні моделі, які навчаються на великій кількості речень і можуть призначати ймовірність новій послідовності слів на основі кількості спостережуваних сполучень слів у навчальному корпусі. Поширені та більш вірогідні послідовності, які часто зустрічаються в корпусі, вважаюьб правильними, тоді як рідкі послідовності можуть містити помилки [2]. Переваги: автоматичне створення правил та відсутність необхідності застосування експертних знань для опрацювання даних або створення ознак. Недоліки: складна інтерпретація результатів, залежність точності моделі від якості навчального набору даних, необхідність застосування великого за обсягом набору даних.

При класифікації модель навчається передбачувати виправлення для кожного слова/тегу, що позначають дію над певним токеном, яку потрібно виконати для виправлення вхідного речення (sequence labelling). Для класифікації використовують складні системи із рекурентними NN або трансформерами, а також класичні ML-методи: наївний Байєсів класифікатор, метод опорних векторів, випадковий ліс тощо. Останні вимагають ручного створення ознак, як-от POS-теги слів у реченні, парсинг залежностей, відмінки слів, головних і другорядних членів речення тощо. Окрім необхідності застосування експертних знань при побудові моделі, головним недоліком є припущення про незалежність помилок у реченні або контекст слова не містить помилок. Цей метод не дає змоги одночасно виправити кілька співзалежних помилок.

Глибинне навчання архітектури рекурентних NN використовують для багатьох NLP-завдань. На відміну від методу класифікації, моделі глибокого навчання не вимагають розробки ознак, оскільки NN можуть досліджувати їх автоматично. Це є великою перевагою, оскільки генерація ознак вимагає експертних знань з лінгвістики. Особливо популярною варіацією рекурентних NN є LSTM. Рекурентні NN використовують для передбачення тегу, що позначає необхідну дію над токеном для виправлення речення. Дана NN-архітектура використовується у GEC-задачі для MT тексту, що містить помилки, у його правильний варіант. Для створення кращої MT-моделі використовують велику кількість пар речень (паралельних корпусів), точність системи у цьому випадку буде залежати від якості набору даних. Окрім того, рекурентні NN сприймають токени послідовно, що сповільнює час навчання та передбачення, унеможливлює паралельне опрацювання даних. У випадку комерційного застосування ІС час очікування є критичним, тому все більше досліджень наразі спрямовані на застосування іншої NN-архітектури, що називається трансформер.

Трансформер – модель глибинного навчання, яка замінює механізм рекурентності на механізм уваги, що забезпечує контекст для будь-якого положення токену у вхідній послідовності. Ця властивість надає змогу розпаралелювати набагато більше процесів у порівнянні з рекурентними нейронними мережами, і відтак знижує тривалість навчання [3]. Трансформери

DOI 10.15588/1607-3274-2023-1-12

швидко стали домінуючою архітектурою для NLP [4], випереджаючи такі альтернативи, як згорткові та рекуррентні NN, особливо для завдань розуміння мови (класифікація, переказ і узагальнення тексту, МТ) та її генерації. Архітектура масштабується відповідно до навчальних даних і фіксує особливості тексту на великій відстані. Попереднє навчання моделі дозволяє навчати трансформери на великих відкритих неанотованих корпусах і згодом легко налаштовувати їх до конкретних завдань, отримуючи в результаті високу якість системи.

Підвид трансформерів BERT призначений для попереднього навчання глибоких двонаправлених представлень з неанотованого набору даних. В результаті попередньо навчена модель BERT може бути налаштована лише одним додатковим вихідним шаром для різних NLP-задач без істотних модифікацій внутрішньої архітектури [5].

Метою дослідження є проектування та створення GEC-системи в текстах українською мовою за допомогою ML-методів з використанням невеликого набору анотованих паралельних даних. До задач, які необхідно вирішити для досягнення поставленої мети, належать:

– опис функціональності та вимог проектованої ІС;

– порівняння state-of-the-art методів для GEC;

– проектування і застосування нейронних мереж з різною архітектурою;

– вибір найбільш оптимальної моделі у TPP-контексті, векторного вкладення або векторизації, вибору та генерування ознак, ML-алгоритму та його параметрів.

Об’єкт дослідження – процеси ідентифікації та корекції граматичних та стилістичних помилок в україномовному текстовому контенті. Предмет дослідження – методи та засоби виправлення граматичних/стилістичних помилок в україномовних текстах із застосуванням оптимального конвеєру (pipeline) на основі TPP, вибору та генерування ознак, алгоритмів машинного навчання, в умовах наявності невеликих за обсягом корпусів анотованих даних. Наукова новизна – застосування нейронних мереж з новою архітектурою, огляд state-of-the-art методів та порівняння різних етапів конвеєру (pipeline) дасть змогу визначити таку їх комбінацію, яка дозволить отримати якісну модель виправлення граматичних помилок в україномовних текстах.

1 ПОСТАНОВКА ПРОБЛЕМИ

GEC-систему S подано кортежем:

$$S = < I, O, R, U, N, \alpha, \beta, \gamma >,$$

де $I = \{i_1, i_2, i_3, i_4\}$, $O = \{o_1, o_2, o_3\}$, $R = \{r_1, r_2, r_3, r_4\}$, $U = \{u_1, u_2, u_3, u_4, u_5\}$.

Основними процесами ІС граматичної корекції є «TPP», «Пошук помилок», «Машине навчання» та «Виправлення граматичних помилок». TPP-процес ІС граматичної корекції опишемо суперпозицією:

$$C_{AU} = \mu \circ \beta \circ \alpha,\ C_{AU} = \mu(\beta(\alpha(i_1, i_2, i_4), r_1, u_1), u_2).$$

Процес «Пошук помилок» ІС граматичної корекції опишемо суперпозицією: $C_{CU} = \chi \circ \beta \circ \alpha$, тобто

$$C_{CU} = \chi(\beta(\alpha(C_{AU}, i_2, i_3, i_4), r_1, u_3), r_2).$$

Процес машинного навчання на достовірних даних ІС граматичної корекції опишемо суперпозицією:

$$C_{UL} = \omega \circ \gamma \circ \beta \circ \alpha,\ C_{UL} = \omega(\gamma(\beta(\alpha(C_{CU}, i_2), i_3), u_4), r_3).$$

Процес «Виправлення граматичних помилок» ІС граматичної корекції на основі GEC-методів опишемо суперпозицією:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha,\ C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, i_2), i_4), u_5), r_4).$$

GEC-методи, засновані на правилах, синтаксичному аналізі або статистичному моделюванні, описані у дослідженнях, що спрямовані на побудову системи перевірки та виправлення текстів, написаних данською [7], грецькою [8], латвійською [9], слов’янської [10], пенджабі [11], філіппінською [12] та арабською [13] мовами. Системи GEC для англійської постійно розвиваються і наразі активно використовують ML-методи: класифікацію (sequence tagging) та МТ. Для створення якісної ML-моделі для GEC у текстах тих мов, які є складними морфологічно, необхідна велика кількість паралельних або вручну розмічених даних. Проблему отримання анотованих даних без ручного маркування частково вирішують за допомогою алгоритмів, що змінюють початковий текст, додаючи у нього «шум» (noise injection). Окрім того, використовують зворотний переклад безпомилкових текстів для того, щоб отримати їх неграматичні відповідники. Таким чином автоматично згенеровують певну кількість розмічених корпусів паралельних текстів [14].

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

Основними аналогами розроблюваної ІС є два додатки: Grammarly та LanguageTool.

Grammarly – комерційна онлайн-платформа компанії Grammarly Inc., яка перевіряє й виправляє не тільки граматичні помилки, а й пропонує рекомендації щодо чіткості (стислість та зрозумілість), захопливості (словниковий запас та розмаїття) та тону повідомлення (формальність, ввічливість і впевненість) [16–17]. Для покращення перевірки і точності пропозицій платформа використовує ML-алгоритми і методи глибинного навчання [18]. 30 мільйонів людей і 30 тисяч команд використовують Grammarly щодня [18]. Платформа потрапила до рейтингу Time «Time100 Most Influential Companies of 2022» та FastCompany «The 10 most innovative companies in artificial intelligence of 2022», також до списку Forbes «Cloud 100» та The Software

DOI 10.15588/1607-3274-2023-1-12

Report «Top 100 Software Companies» [18]. Наразі Grammarly підтримує лише англійську мови та її діалекти: американський, британський, канадський, австралійський [19].

LanguageTool – GEC-програма із відкритим кодом, що у своїй основі використовує перевірку на основі правил. Усього система більше 30 мов, зокрема і українську [20]. LanguageTool розвивається з 2003 року і наразі налічує 5 415 правил для англійської мови, 1 022 – для української [21]. Щодня користувачі перевіряють більше 20 мільйонів текстів за допомогою цієї платформи [22]. Для перевірки правильності написання україномовних текстів доступні правила таких категорій: варваризми (наприклад, приймати участь – брати), великі літери, граматика, логічні помилки (наприклад, неправильна дата), орфографія, оформлення, пунктуація, стиль, типографія [21]. Недоліком LanguageTool для перевірки україномовних текстів є невелика навіть у порівнянні з англійською мовою кількість правил, які не можуть покрити усі можливі граматичні помилки.

Sai Muralidhar Jayanthi, Danish Pruthi, Graham Neubig розробили систему NeuSpell [23] для виправлення орфографічних помилок англомовних текстів, що заснована на методах глибинного навчання. Система складається з 10 різних нейронних мереж, які дають змогу фіксувати контекст навколо орфографічних помилок. Для навчання нейронних мереж використані синтетичні навчальні дані, створені за допомогою кількох стратегій додання «шуму». NeuSpell є програмою із відкритим кодом і наразі підтримує лише англійську мову.

Hunspell [24] – система для виправлення орфографічних помилок, створена для мов з багатою морфологією, складеними словами і різним кодуваннями символів. Використовує спеціальний формат словника, який визначає, які стеми та афікси є дійсними в певній мові. Система забезпечує токенізацію, стемінг і перевірку орфографії для майже будь-якої мови чи алфавіту.

В [6] запропоновано підхід перевірки граматичної правильності текстів на основі «мінімального навчання з вчителем» (minimal supervision). Даний метод полягає у застосуванні невеликого анотованого набору даних і додання штучних помилок до корпусу новин, художньої літератури та інших жанрів (усього 18 мільйонів слів). Дослідники розробили класифікатори для кількох поширених типів граматичних помилок: прийменники, відмінок іменника, форма дієслова та узгодження з дієсловом (розподіл на число та рід). Окрім того, вони застосували метод нейронного МТ тексту, що містить помилки, у його правильний варіант. Alla Rozovskaya та Dan Roth зауважили, що МТ-точність є занадто низькою у зв'язку з недостатньою кількістю навчальних даних. Запропонований метод «мінімального навчання з вчителем» збільшує як точність класифікаторів, так і МТ-якість. За допомогою експериментів показали, що МТ-метод показує незадовільні метрики якості (F-score = 10.6) з використанням набору даних відповідних текстів, що містить приблизно 200 тис. вручну анотованих і виправлених слів. Цю МТ-систему значно перевершує запропонований у [6] ML-підхід (різновидність навчання з учителем, яке використовує немарковані дані для тренування – зазвичай невелику кількість маркованих даних та велику кількість немаркованих).

Дослідники Oleksiy Syvokon, Olena Nahorna презентували набір даних – корпус текстів [15], що професійно анотований для GEC та вільного редагування українською мовою. Дослідники зібрали тексти з помилками (20 715 речень – 328 779 токенів) від різноманітних авторів, у тому числі носіїв мови. Дані охоплюють найрізноманітніші сфери письма, від текстових чатів і есе до офіційного письма. Професійні коректори виправляли та анотували корпус на помилки, пов'язані з вільним мовленням, граматикою, пунктуацією та орфографією. За [1, 15], цей корпус можна використовувати для розробки та оцінки систем GEC українською мовою і дослідження морфологічно багатих мов. За [1], найбільш очевидною проблемою, пов'язаною з GEC-системами, є потреба у високоякісних навчальних корпусах, що містять велику кількість навчальних прикладів.

В [25] розроблено морфологічний парсер pymorphy2 для української мови у форматі бібліотеки для мови програмування Python. pymorphy2 аналізує частину мови, число, відмінок, час, лему та стем заданого слова. Морфологічний парсер заснований на словниках OpenCorpora, що конвертовані до формату XML. Користувачі також мають змогу додавати власні слова та правила, це дає можливість проводити морфологічний аналіз текстів певної ПО без зміни вихідного коду pymorphy2 та адаптувати pymorphy2 для роботи з іншими мовами.

В [26] розроблено систему ТРР, морфологічного та синтаксичного аналізу україномовних текстів. Для токенізації, розподілу на речення та пошуку email-адрес дослідники використали бібліотеку NLTK мови програмування Python та регулярні вирази, для видалення стоп-слів та пошуку іменованих сутностей – відповідні словники. Для морфологічного аналізу слів, їх лематизації або стемінгу була використана бібліотека pymorphy2, що підтримує українську мову. Окрім того, дослідники реалізували графічний інтерфейс додатку за допомогою бібліотеки PyQt.

В [27] проаналізовано існуючі методи пошуку іменних груп в англомовних і україномовних текстах та розробили метод детектування іменних груп на основі дерева залежностей речення і моделі розпізнавання іменованих сутностей. Довели, що методи аналізу англомовного тексту не можуть бути використані для україномовних документів, адже вони створені з урахуванням особливостей структури побудови речень тільки в англомовних текстах.

В [28] адаптовано алгоритм стемінгу Портера для ТРР україномовних текстів. Окрім того, дослідники реалізували алгоритм на мові програмування PHP і

DOI 10.15588/1607-3274-2023-1-12

виклали його у вільний доступ. Довели, що кращого результату можна досягти через лематизацію, однак для цього потрібно провести додаткові дослідження.

В [29] розроблено GEC-систему в англомовних текстах за допомогою спеціальної моделі глибинного навчання – Transformer [3]. Виправлення помилок відбувається завдяки використанню даної моделі для задачі перекладу неправильного тексту у його відкоригований відповідник. Особливістю даної системи є вибір TPP-методів: автоматичне виправлення орфографічних помилок, токенізація [30], метод кодування ознак Byte Pair Encoding. Довели, що дана ІС має достатню точність (F-score = 60.93 на завданні CoNLL-2014 [31]) і швидкість (10 слів за секунду) для подальшого її впровадження і застосування на онлайн-платформах.

В [32] застосовано нейронну мережу архітектури encoder-decoder, що складається із шарів згортки та механізму уваги, для GEC у англомовних текстах за допомогою МТ. Для процесу декодування використані ансамблі однакових моделей, що ініціалізовані випадковим чином. Для фінального вибору речень з-поміж можливих кандидатів оцінки їх ймовірностей змінені за допомогою додатково створених ознак. ІС значно перевершила якість попередньо існуючих МТ-систем, окрім того і на основі рекурентних нейронних мереж.

В [33] розроблено GEC-систему у англомовних текстах, що заснована на підході анотації послідовностей (sequence tagging). Окрім стандартних тегів (keep, delete, append, replace), що позначають дію, яку необхідно виконати над токеном для виправлення речення, також запропоновані так звані g-transformations: зміна регістру першої літери, об’єднання або розділення токену, зміна числа іменника або форми дієслова. Для виправлень токенів застосований ітеративний підхід, у якості ML-методу – різні моделі сімейства Transformer. Такий підхід дозволив отримати найкращі на момент публікації статті оцінки якості системи та збільшення швидкості опрацювання даних у 10 разів у порівнянні із іншими ІС, заснованих на архітектурі Transformer.

В [34] описано GEC-систему, засновану на підході анотації послідовностей за допомогою класичного ML-алгоритму – наївного класифікатора Байєса. Усього використано п’ять моделей – відповідно до наявних у корпусі типів граматичних помилок. Попри те, що система була однією з найкращих на завданні CoNLL [35], аналіз результатів показує, що досягнення дуже високої точності при граматичній корекції вимагає більш складних NLP-методів.

В [9] розроблено засновану на правилах програму перевірки граматики для латвійської мови. Усього реалізовано дві групи правил: правила, що описують правильні речення, і правила, що описують граматичні помилки. На корпусі текстів, написаних людьми, що не є носіями мови, система досягла F1-score 62,4% і 40,2% на корпусі студентських робіт. Недолік: не перевіряє правильність слова в залежності від його контексту. Відповідно, правильно написані та узгоджені слова, що не належать реченню, будуть маркуватися як безпомилкові.

В [8] описано ІС для перевірки правильності текстів, написаних на грецькій мові. Система заснована на правилах та парсингу синтаксису речення. ІС аналізує текст користувача та надає виправлення, опис помилок, а також правила щодо стилю та семантичної інформації в тексті. Системне оцінювання проводилося як паралельне виправлення одних і тих самих текстів програмою та людиною. Система перевірки грецької граматики наблизилася до 90% виправлення людського. Однак ця ІС не може опрацювати всі можливі граматичні помилки.

В [36] розроблено універсальну ML-модель для GEC у текстах, написаних на різних мовах. Дослідники стверджують, що для отримання якісної багатомовної GEC-системи мають бути виконані такі два кроки: автоматична генерація достатньої кількості навчальних даних і використання ML-моделей сімейства Transformer із величезною кількістю параметрів (до 11 мільярдів). Таким чином, автори досягнули високої точності виправлення помилок для чотирьох мов: англійської, чеської, німецької.

В [37] досліджено здатність GEC-моделей до узагальнення граматичних правил для корекції нових помилок у тексті. ML-модель основі Transformer протестована з використанням невідомих їй прикладів. Отримано незадовільні результати навіть у випадку простих правил і зменшеного словнику, що може свідчити про те, що алгоритму бракує можливості узагальнення, необхідної для виправлення нових помилок у наданих тестових прикладах. Для більш якісного GEC за меншого обсягу навчального корпусу існуючі ІС на основі ML варто поєднати з окремою перевіркою певних правил.

В [38] запропоновано мовно-незалежну стратегію для розробки багатомовної GEC-системи. Для її імплементації необхідні лише попередньо навчена МТ-модель та корпус паралельних навчальних даних для перекладу з англійської на обрану мову. Перш за все, МТ-модель генерує нові синтетичні дані на обраній мові, що застосовують у якості помилкових вхідних текстів. Отриманий корпус використовують для попереднього навчання моделі, після чого її потрібно додатково налаштувати, використовуючи анотований набір даних для GEC у текстах обраної мови. Досить гарні показники точності виправлення досягнуті для німецької, китайської мов.

В [39] адаптовано фреймворк Break-It-Fix-It (BIFI), що оригінально використовувався для виправлення коду програм за умови відсутності ідеальних зразків, до завдання перевірки граматичної правильності речень. Дослідники використали попередньо навчену мовну модель для розробки бінарного класифікатора LM-Critic, який виконує завдання попереднього сортування речень і вказує, чи містить певний текст граматичні помилки. LM-Critic визначає речення граматично правильним, якщо йому відповідає

DOI 10.15588/1607-3274-2023-1-12

більший числовий показник ймовірності, ніж його неправильним «сусідам». Утворені пари правильних і неправильних речень використовують для навчання «коректора», що опрацьовує лише речення, позначені LM-Critic як ті, що містять граматичні помилки.

В [40] для вирішення проблеми необхідності застосування великих за обсягом корпусів навчальних даних згенерували помилкові версії великих неанотованих наборів текстів за допомогою запропонованої функції шуму. Отримані паралельні корпуси згодом використовують для попереднього навчання моделей на основі архітектури Transformer. Потім необхідно додатково налаштувати ML-модель відповідно до ПО та стилю датасету. Дана GEC-система ускладнена використанням нейронної перевірки орфографії, що сортує запропоновані варіанти виправлень в залежності від контексту.

В [41] змінили архітектуру ML-моделі Transformer, впровадивши новий механізм уваги, що заснований на дереві залежностей слів у реченні. Так як неправильно побудоване речення може спричинити помилку у парсингу дерева залежностей, дослідники також навчили NN виправляти графи, що містять помилки. Додатково застосувавши запропонований метод аугментації даних, отримали гарні показники GEC-точності навіть без попереднього навчання NN.

В [42] застосували нейронну модель sequence-to-sequence на рівні символів для того, щоб уникнути проблем зі словами OOV (out of vocabulary words, відсутні у словнику). У якості енкодера застосовують двонапрямлену рекурентну NN, декодер – також рекурентна нейронна мережа, поєднана із механізмом уваги. Декодер генерує вихідне речення посимвольно. Навіть попри те, що такий підхід допомагає нейронній мережі опрацьовувати невідомі їй слова, вона не може ефективно опрацьовувати інформацію на рівні слова: ця модель отримала оцінку F-score = 40,56 на тестовому наборі CoNLL [31].

В [43] застосували новий підхід до автоматичної генерації навчальних анотованих прикладів та GEC за допомогою генеративних змагальних мереж (GANs). Генеративні змагальні мережі складаються з генератора, що постійно генерує навчальні приклади, і дискримінатора, що класифікує отримані дані. У [43] генератором є NN Grammatical Error Labeler, що навчається додавати помилки, аналогічні помилкам у навчальній вибірці текстів, дискримінатор – NN Grammatical Error Detector, що навчається розпізнавати правильну мітку, що позначає дію, яку необхідно виконати над певним токеном.

В [44] розробили дизайн платформи для GEC і створили відповідну ML-модель, поєднану з перевіркою правил. Система складається з трьох модулів: виправлення помилок, адміністрування системи, фільтрування відгуків. GEC-Модуль включає TPP. На TPP-етапі також визначається валідність введених користувачем даних. Модуль фільтрування відгуків дозволяє користувачам скорегувати неправильні виправлення системи. Такий підхід дає змогу виправити можливі помилки, наявні у наборі даних, оскільки отримані відгуки користувачів можуть використовуватись для додаткового навчання нейронної мережі.

В [45] презентували корпус анотованих даних для створення GEC-системи та редагування текстів, написаних чеською мовою. Усього набір даних містить 42 210 речень. Окрім того, дослідники використали MT-підхід для GEC. Нейронна мережа архітектури Transformer попередньо навчена на додатково синтезованих з головного набору даних реченнях. Окрім чеської мови, автори статті застосували даний підхід для виправлення помилок в німецькомовних текстах.

В [46] запропонували та реалізували новий метод автоматичного створення навчальних даних із застосуванням двох різних за якістю моделей перекладу. «Погана» система перекладу є статистичною моделлю на рівні речень, якісна – навченою нейронною мережею. Даний метод дав змогу отримати 10 мільйонів паралельних речень для навчання нейронної мережі архітектури Transformer. Даний метод дасть змогу отримати великі датасети навчальних даних для інших мов, для яких великі корпуси анотованих даних не є доступними.

В [47] використали генеративні змагальні мережі для створення системи автоматичного виправлення помилок у англомовних текстах. На відміну від алгоритму в [43], в даному дослідженні генератор виправляє помилки, «перекладаючи» помилкові речення у правильні, а дискримінатор навчається оцінювати якість перекладу та наявність помилок у опрацьованому генератором реченні. Дискримінатор, що приймає на вхід два речення, заснований на сіамських рекурентних або згорткових NN.

3 МАТЕРІАЛИ ТА МЕТОДИ

Система перевірки граматичної правильності речень може використовуватись для перегляду запропонованих змін та їх пояснень, а також для автоматичного виправлення правильності речень.

Така система може використовуватись як і індивідуальним користувачем для перевірки власних текстів, так і сторонньою NLP-програмою для попереднього або фінального опрацювання речень.

Зовнішніми сутностями є: користувач, NLP-програма, адміністратор системи.

Зацікавлені особи прецеденту та їх вимоги:

– користувач створює обліковий запис, завантажує або відкриває документи, перевіряє граматичну правильність речень, застосовує запропоновані зміни;

– стороння NLP-програма відправляє текст на перевірку за допомогою API системи, отримує змінений текст або перелік можливих виправлень;

– адміністратор системи виконує її налаштування, навчання нейронної мережі, перевірку запропонованих користувачем виправлень.

DOI 10.15588/1607-3274-2023-1-12

Користувач ПС: 1) фізична особа, що використовує систему для перевірки власних текстів; 2) NLP-програма.

Передумови прецеденту:

– комп'ютер, за допомогою якого здійснюватиметься аналіз, підключений до Інтернету та має встановлене необхідне ПЗ;

– користувач має бути успішно авторизованим у системі;

– словник (якщо використовується перевірка орфографії) містить достатню кількість слів і (або) правил їх утворення, є доступним;

– реалізовані усі ключові модулі, що забезпечують основний функціонал системи;

– штучна нейронна мережа (якщо застосовуються ML-методи) є попередньо натренованою на виконання необхідного завдання і має достатньо високі показники якості.

Основний успішний сценарій:

– користувач завантажує програму або розширення, або відкриває онлайн-додаток у браузері;

– користувач завантажує або створює документ і заповнює його;

– система виконує перевірку тексту;

– система відображує необхідні виправлення;

– користувач підтверджує або скасовує застосування запропонованих виправлень;

– користувач зберігає документ. Зберігається також і історія змін.

Альтернативні потоки:

– Власний файл з даними не відповідає типу файлу, який може відкрити програма.

1. Програма повідомляє користувача про помилку і скасовує відкриття файлу.

2. Користувач обирає файл з правильним розширенням.

3. Програма відкриває файл з правильним розширенням, зчитує дані, проводить їх попередню обробку (точка повернення в основний сценарій).

– Помилка у роботі програми:

1. Користувач звертається до служби технічної підтримки (розробника) і повідомляє про помилку у програмі.

2. Розробник усуває помилку і надає нову версію або надає інформацію про способи її самостійного усунення (точка повернення).

Пост-умови:

– Користувач запропонував власний GEC-приклад у реченні у випадку отриманих незадовільних результатів;

– Запропоновані користувачем виправлення переглянуті адміністратором для подальшого налаштування системи або збережені для персоніфікації подальших виправлень;

– Дані, налаштування і файли користувача занесені до бази даних.

Спеціальні СВ:

– ІС повинна відповідати бажаному рівню точності, яка перевіряється спочатку на навчальних даних, потім – на тестових даних, у режимі перевірки власного файлу точність не перевіряється;

– система має підсвічувати неправильні слова та словосполучення, надавати обґрунтування запропонованих змін;

– для доступу до функціоналу системи сторонній NLP-додаток використовує її API-сервіс.

Список необхідних ІТ та додаткових пристроїв:

– користувач повинен мати комп'ютер із встановленим ПЗ та підключенням до мережі Інтернет для отримання даних;

– система використовує розподілену нереляційну базу даних для зберігання даних користувачів;

– для додаткового опрацювання текстів використовуються попередньо навчені ML-моделі.

Для побудови діаграми варіантів використання для системи автоматичного виправлення помилок україномовних текстів (рис. 1) спочатку необхідно визначити основних акторів: Користувач (User) і NLP-Програма (NLP Application), а також основні варіанти використання як Apply correction (застосувати корекцію), Check text (перевірити текст), Suggest correction (запропонуйте виправлення), Upload text (завантажити текст), Check text for all types of errors (перевірити текст на всі види помилок), View suggestions (переглянути пропозиції).

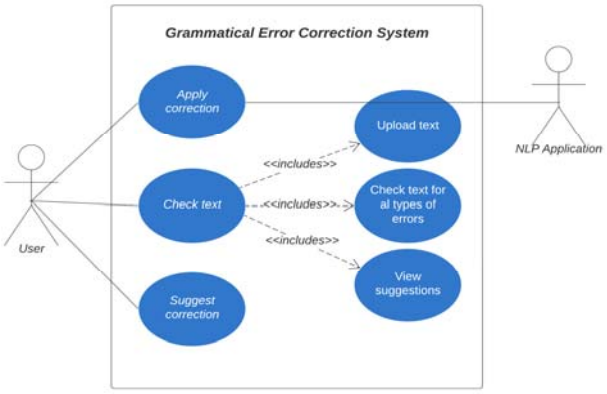


Рисунок 1 – Діаграма варіантів використання GECS

GECS – це десктопний / мобільний додаток, онлайн-платформу або модуль-доповнення до певного ПЗ, наприклад, браузеру або текстового редактору. ПЗ для аналізу природньої мови у свою чергу може використовувати API сервісу автоматичного виправлення речень, обмінюючись повідомленнями за допомогою архітектури REST взаємодії між додатками. GEC-система включає в себе також сервер опрацювання даних, оскільки за умови застосування ML-методів, у тому числі і мовних моделей на основі архітектури Transformer з мільйонами параметрів, додаток потребуватиме багато обчислювальних ресурсів для обробки текстів. Збільшення кількості користувачів, і, відповідно, збільшення обсягу даних також потребуватимуть додаткових комп'ютерних потужностей. До складу вищезазначеної ІС також входить сервер БД користувачів, що зберігає документи у обліковому

DOI 10.15588/1607-3274-2023-1-12

записі. Взаємодія із серверами опрацювання та зберігання даних відбувається через графічний інтерфейс ІС, онлайн-платформи або доповнення. Альтернативним варіантом є встановлення запропонованого ПЗ на потужному комп'ютері, що має достатньо ресурсів для опрацювання даних локально. У іншому випадку, локальна версія ПЗ може мати спрощений функціонал. На рис. 2 подана розширена діаграма варіантів використання із уточненням функціоналу ІС з відповідними варіантами використання як Apply correction (застосувати корекцію), Check text (перевірити текст), Suggest correction (запропонуйте виправлення), Create a document (створити документ), Login/Signup (вхід/вихід), Upload text (завантажити текст), Check text for all types of errors (перевірити текст на всі види помилок), View suggestions (переглянути пропозиції), Upload a document (завантажити документ), Send request (відправляти запит), Get list of suggested corrected (отримати список запропонованих виправлень).

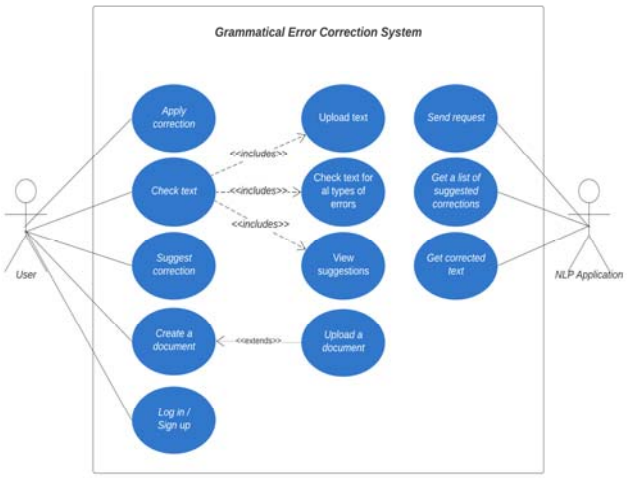


Рисунок 2 – Доповнена діаграма GECS

В залежності від виду взаємодії користувача і GECS не всі типи перевірок можуть бути доступними. Для зберігання та редагування особистих файлів користувач має бути зареєстрованим та авторизованим (варіант використання Log in / Sign up) у системі. Документи, логін, пароль та інші дані користувачів зберігаються на окремому сервері.

Варіант використання «Перевірити текст» (Check text) включає в себе три інших: Завантажити текст, Перевірити текст на всі типи помилок, Переглянути пропозиції. Завантаження тексту може відбуватись у кілька варіантів: вставлення свого тексту у textbox, відкриття файлу або створення нового документу. Створення (завантаження) документу виділено у окремий варіант використання, оскільки даний функціонал може бути реалізований лише для онлайн-платформи або локально встановленого додатку. До базових належать помилки, пов'язані з вільним мовленням, граматикою, пунктуацією та орфографією. В залежності від реалізованого функціоналу система може пропонувати заміни слів на синоніми для збільшення лексичної різноманітності тексту або заміни відповідно до певного стилю письма. Варіант використання Переглянути пропозиції позначає почерговий перегляд запропонованих виправлень разом із поясненням порушеного правила або причини заміни. Окрім того, частини тексту, що містять помилки, мають бути підсвічені відповідно до типу порушених правил. Користувач повинен мати змогу застосувати лише обрані виправлення, ті, які є необхідними у даному випадку. Якщо користувач вважає якість запропонованих виправлень незадовільною, він повинен мати змогу самостійно виправити текст і зберегти його як зразок для подальшого налаштування системи або навчання нейронної мережі. Таким чином, наступні виправлення системи матимуть кращу якість і будуть більш персоніфікованими для кожного окремого користувача. Варіант використання Застосувати виправлення дає змогу користувачеві автоматично виправити усі знайдені системою граматичні помилки без перегляду детальних пояснень порушених правил.

NLP-Додатки також можуть використовувати GEC-систему для попередньої або пост-обробки текстів. В цьому випадку додаток має відправити запит, використовуючи Application Programming Interface (API) сервісу, система – відправити результат аналізу або одразу виправлений текст. Отже, GEC може бути реалізоване як для індивідуального використання у вигляді застосунку, веб-сайту або розширення для додатків, так і як модуль у системі аналізу або генерації природньої мови.

Для побудови діаграми класів (рис. 3) у контексті інформаційної GEC-системи визначено вісім основних класів: GEC Software (Програма, що реалізує користувацький інтерфейс системи), User (Користувач), User Database (БД користувачів), Document (Документ), Data Processing Server (Сервер обробки даних), Check Result (Результат перевірки), Dictionary (Словник), ML model (ML-Модель), NLP application (NLP-Додаток).

Клас User (користувач) описує індивідуального користувача системи, який завантажує додаток, встановлює розширення або перевіряє тексти за допомогою онлайн-платформи. Цей клас містить такі атрибути як ім'я, електронна пошта і дані для авторизації у програмі – логін і пароль. Відповідно, він може авторизуватися у системі, завантажити і перевірити текст, створити і зберегти документ, застосувати виправлення, надане системою, або запропонувати власне. Також, у контексті інформаційної системи, користувач створює документ, що зберігає безпосередньо сам текст, а також мета-дані про історію внесених змін і виправлень, автора документу і дату створення. Клас User Database (БД користувача) може зберігати, записувати, оновлювати, повертати інформацію і документи, додати або верифікувати користувача.

DOI 10.15588/1607-3274-2023-1-12

Клас GEC Software (ПЗ GEC) постає посередником між користувачем, сервером обробки даних та базою даних. ПЗ може авторизувати користувача, запросити дані із сховища, відправити текст на аналіз, отримати результати та візуалізувати їх, застосувати запропоновані виправлення і зберегти результат, зберегти введені користувачем виправлення. Як вже зазначалось, альтернативним варіантом є використання локальних обчислювальних ресурсів без під'єднання до зовнішнього серверу. GEC Software також надає API сервісу для запитів на аналіз текстів та автоматичне виправлення помилок: до атрибутів цього класу належать тип запиту, адреса для запиту і тип даних, що відправляє сервер як відповідь. Даний ресурс може обробляти запити, автоматично застосовувати необхідні виправлення та відправляти правильний текст на сервер, з якого прийшов запит. Подібні сервіси визначають тип користувача за токеном.

Клас NLP Application (NLP-застосунок) позначає будь-яку програму для аналізу або генерації природньої мови, що використовує GEC-систему для попередньої або пост-обробки даних. Така програма має свою структуру і призначення, метод doItsMagic() відповідає за пряме її застосування. Ця програма може відправляти запити певного типу і обробляти пакет даних, отриманий від додатку GEC Software.

Клас Data Processing Server (сервер опрацювання даних) описує поведінку серверу обробки даних. Такі сервери мають свої технічні характеристики (кількість ядер процесора, об'єм операційної та постійної пам'яті, наявність графічного процесора тощо), які напряму пливають на швидкість обробки великого масиву даних.

Для коректного аналізу граматичної коректності речень сервер повинен спочатку виконати TPP, після цього – представити записи як вектори. Потім – парсинг дерева залежностей слів у реченні, частин мови та іменованих сутностей. Отримане дерево залежностей аналізується за допомогою правил. Для автоматичного виправлення використовується підхід передбачення міток-тегів, що вказують на дію, яку необхідно виконати над кожним токеном. Для речень із багатьма помилками, парсинг дерев залежнотей яких дає некоректні результати, застосовується MT-метод тексту із помилками у правильний його варіант.

Перевірка орфографії є першим етапом перевірки граматичної коректності речень. Словник Dictionary позначає окремий клас тому, що для перевірки орфографії використовують словники певної мови, що також можуть містити правила утворення слів, а результати ранжуються за зростанням Edit Distance або модифікаціями цієї метрики.

Клас ML model (ML-модель) описує поведінку ML-моделі, до методів якого належать тренування моделі на навчальних даних та передбачення результатів. Кратність зв'язків «0…*» та відношення агрегації вказують на те, що система не має обов'язково використовувати MT-алгоритми для перевірки граматичної правильності речень та їх автоматичного виправлення.

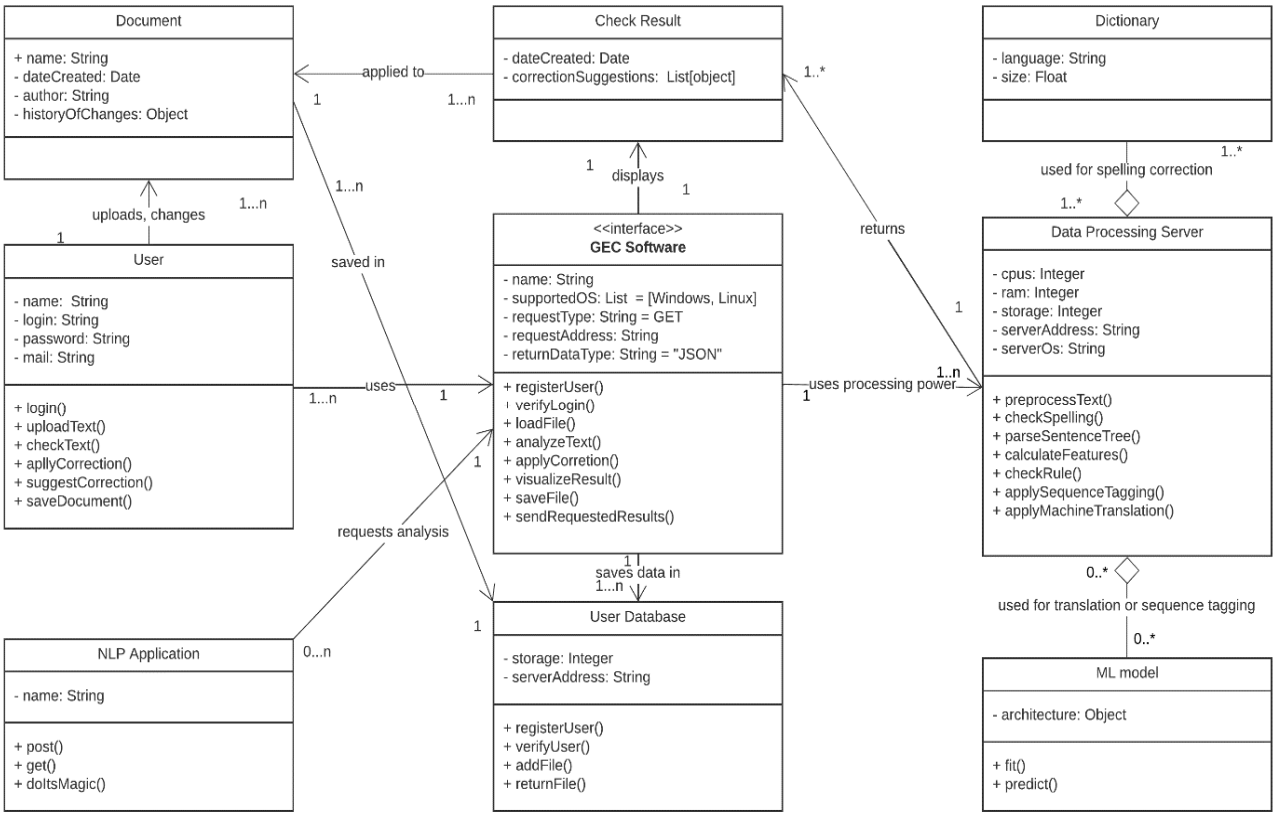


Рисунок 3 – Діаграма класів для GEC-системи

DOI 10.15588/1607-3274-2023-1-12

Клас Check Result (результат перевірки) означає результат перевірки одного тексту. Його основним атрибутом є список запропонованих виправлень. Виправлення містять індекси початкового і кінцевого символів, виправлену стрічку, пояснення порушеного правила або причини необхідної заміни. Кратності усіх зв'язків та їх назви вказані на рис. 3.

Послідовність кроків взаємодії із системою перевірки граматичної правильності текстів може бути представлена за допомогою двох діаграм послідовності відповідно для індивідуального користувача додатку та для NLP-застосунку, який використовує виправлення текстів у граматично правильний варіант як одну із складових власного конвеєру (pipeline).

На рис. 4 подано діаграму послідовності для режиму перевірки власного тексту користувача відповідними компонентами як Software (програмне забезпечення), ML-модель Processing server (сервер опрацювання), ML model (), User Database (база даних користувачів). На схемі пропущені кроки реєстрації та авторизації, а також створення і збереження документу. Основний алгоритм такий:

1. Завантажити текст (Load text).
2. Надіслати текст (Send text).
3. Перевірити орфографію (Check spelling).
4. Виправити орфографічні помилки (Correct spelling errors).
5. Проаналізувати дерево залежностей або непередбачених ситуацій (Parse dependency or contingency tree).
6. Перевірити розширені правила на основі дерева (Check advanced rules based on tree).
7. Перевірити за допомогою моделі машинного навчання (Check with machine learning model).
8. Застосувати корекцію машинного навчання (Apply machine learning correction).
9. Повернути результати (Return results).
10. Повернути пропозиції у форматі: [i_початок, i_кінець, пропозиція, пояснення] (Return suggestions in format: [i_start, i_end, suggestion, explanation]).
11. Показати пропозиції (Display suggestions).
12. Застосувати виправлення (Apply fixes).
13. Запропонувати виправлення (Suggest correction).
14. Запам'ятати вподобання (Remember preferences).

БД користувачів використовується для збереження вподобань / зразків виправлень користувачів для подальшої персоніфікації запропонованих системою виправлень. Спочатку користувач має обрати варіант завантаження даних –вставкою тексту, відкриття файлу або створення і заповнення нового документу. ПЗ, встановлене на персональному комп'ютері, або доступне у форматі онлайн-сервісу чи розширення, відправляє дані для аналізу на сервер. На цій діаграмі послідовностей наведений лише один із можливих підходів до побудови системи граматичної

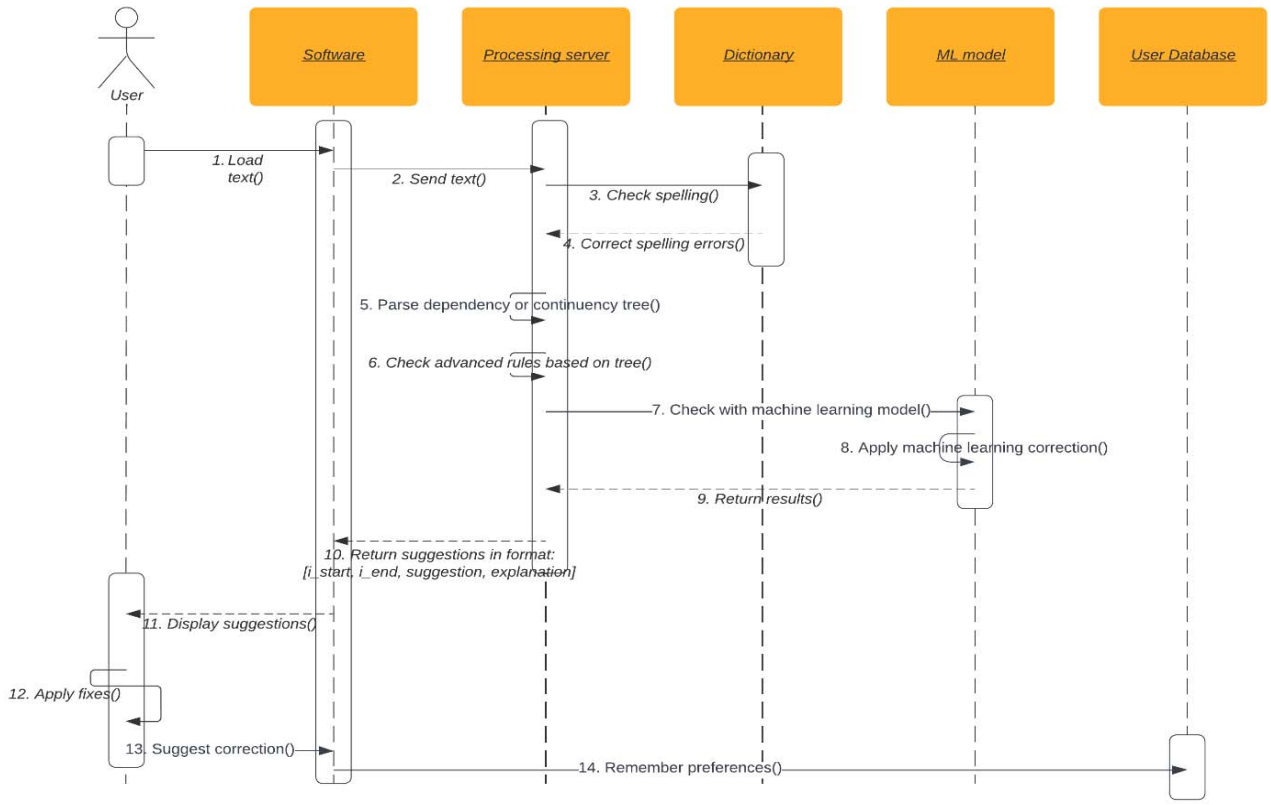


Рисунок 4 – Діаграма послідовності для першого режиму роботи системи

DOI 10.15588/1607-3274-2023-1-12

корекції, що поєднує у собі перевірку орфографії, правил, парсинг синтаксичних дерев, перевірку правил на основі дерев і ML-алгоритми. Після перевірки орфографії, що відбувається з використанням словника певної мови, система має перевірити виконання базових правил, що не потребують парсингу синтаксичних дерев, як-от чергування «у-в», правопис «пів-напів», вживання апострофа тощо. Застосування перевірки правил та парсингу синтаксичних дерев обумовлене двома факторами:

– відсутністю достатньо великих анотованих / паралельних корпусів української мови для навчання повністю автоматичних моделей, що засновані на алгоритмах глибинного навчання;

– недостатньою спроможністю мовних моделей штучного інтелекту узагальнювати правила.

У випадку, коли у системі не реалізовані усі можливі правила правопису, в тому числі і ті, що засновані на парсингу синтаксичних дерев, одним із варіантів є додаткове використання ML-методів для передбачення тегів, що позначають дію, яку необхідно виконати над токенами, або для «перекладу» тексту із помилками у правильний варіант. Надалі ПЗ візуалізує і застосовує запропоновані виправлення, зберігає документ у базі даних за необхідності.

Другий режим застосування системи – аналіз граматичної коректності отриманих за допомогою API запитів (рис. 5). На цій діаграмі послідовності зображені етапи відправки, отримання і опрацювання запитів від стороннього NLP додатку (NLP Application) до системи (GEC System), зокрема:

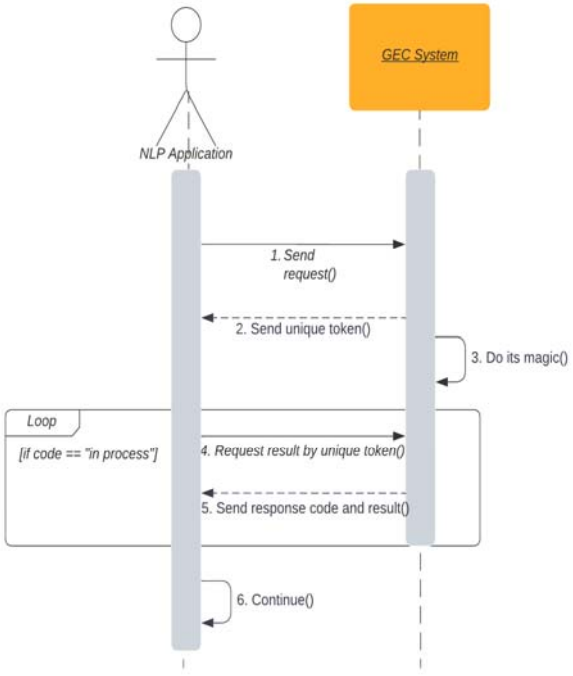


Рисунок 5 – Діаграма послідовності для другого режиму роботи

1. Надіслати запит (Send request).

2. Надіслати унікальний маркер (Send unique token).

3. Опрацювання запитів (Do its magic) з циклом (Loop) типу [якщо код == “в процесі”] ([if code == “in process”]).

4. Результат запиту за унікальним маркером (Request result by unique token)

5. Надіслати код відповіді та результат (Send response code and result).

6. Продовжити (Continue).

Об’єкт GEC System позначає сукупну систему граматичної корекції разом з усіма складовими, включаючи графічний інтерфейс, сервер для обробки даних, попередньо навчені ML-моделі. Повідомлення 3. Do its magic(), у свою чергу, позначає повний аналіз тексту та застосування необхідних виправлень.

GEC System надає кожному запиту унікальний токен, за допомогою якого можна отримати результати після завершення системою опрацювання запиту. Оскільки, в залежності від навантаження на систему, обробка даних може зайняти певний час, то NLP додаток має з певним інтервалом відправляти новий запит на отримання результатів. NLP додаток продовжує свою діяльність після отримання виправлених текстів. На діаграмі діяльності (рис. 6) зображений один із-поміж множини можливих підходів до розробки системи автоматичного виправлення граматичних помилок. Алгоритм:

1. Якщо доступний мовний словник [Language dictionary is available], то застосувати виправлення орфографії, що не запам’ятовується (Apply low-recall spelling correction).

2. Якщо мовний словник недоступний [Language dictionary is not available], то застосувати перевірку основних правил (Apply check of basic rules).

3. Якщо доступний синтаксичний аналізатор залежності/виборчої групи [Dependency/Constituency parser is available], то застосувати перевірку більш конкретних правил (Apply check of more specific rules).

4. Якщо правила можуть охоплювати всі випадки Rules can cover all cases], то використати підхід, заснований на правилах (Use a rule-based approach).

5. Якщо правила не можуть охоплювати всі випадки [rules can’t cover all cases] і/або якщо парсер недоступний [parser is not available], то перевірити або визначити розмір корпусу.

6. Якщо достатньо великі корпуси недоступні [large enough corpora is not available], то застосувати методи збільшення/генерування даних (Apply data augmentation / generation techniques).

7. доступні великі корпуси [large corpora is available], то визначити характеристики корпусу.

8. Якщо доступні великі анотовані корпуси [large annotated corpora is available], то використати підхід позначення послідовності (Use sequence tagging approach).

9. Якщо доступні великі паралельні корпуси [large parallel corpora is available], то використати підхід

DOI 10.15588/1607-3274-2023-1-12

нейронного машинного перекладу (Use neural machine translation approach).

10. Якщо доступні великі неанотовані корпуси [large unannotated corpora is available], то використати підхід на основі статистики (Use statistical-based approach).

Перш за все, якщо словник слів (або правил утворення слів) є доступним для певної мови, пропонується застосовувати його для виправлення орфографічних помилок. Варто зазначити, що модуль для автоматичного виправлення орфографічних помилок має мати низькі показники повноти (recall).

Значення повноти також можна інтерпретувати як відношення кількості правильно визначених позитивних випадків до усіх істинно позитивних випадків, тобто як частку загального числа позитивних зразків, що було знайдено.

Значення влучності (precision), у свою чергу, можна оцінити як частку правильно визначених істинно позитивних випадків.

В залежності від якості класифікатора та значення порогу прийняття рішення значення влучності буде залежним від значення повноти (рис. 7).

Низькі показники повноти (і високі – влучності) позначають те, що класифікатор (у даному випадку функція, що визначає орфографічні помилки) має бути достатньо «впевненим» у своєму «рішенні» про виправлення слова саме для того, щоб заміна не могла причинити зміну семантики речення.

Після перевірки і виправлення орфографії важливим є застосування перевірки базових правил граматики. Приклади для української мови: чергування «у-в», правопис «пів-», правопис апострофа та м’якого знака тощо.

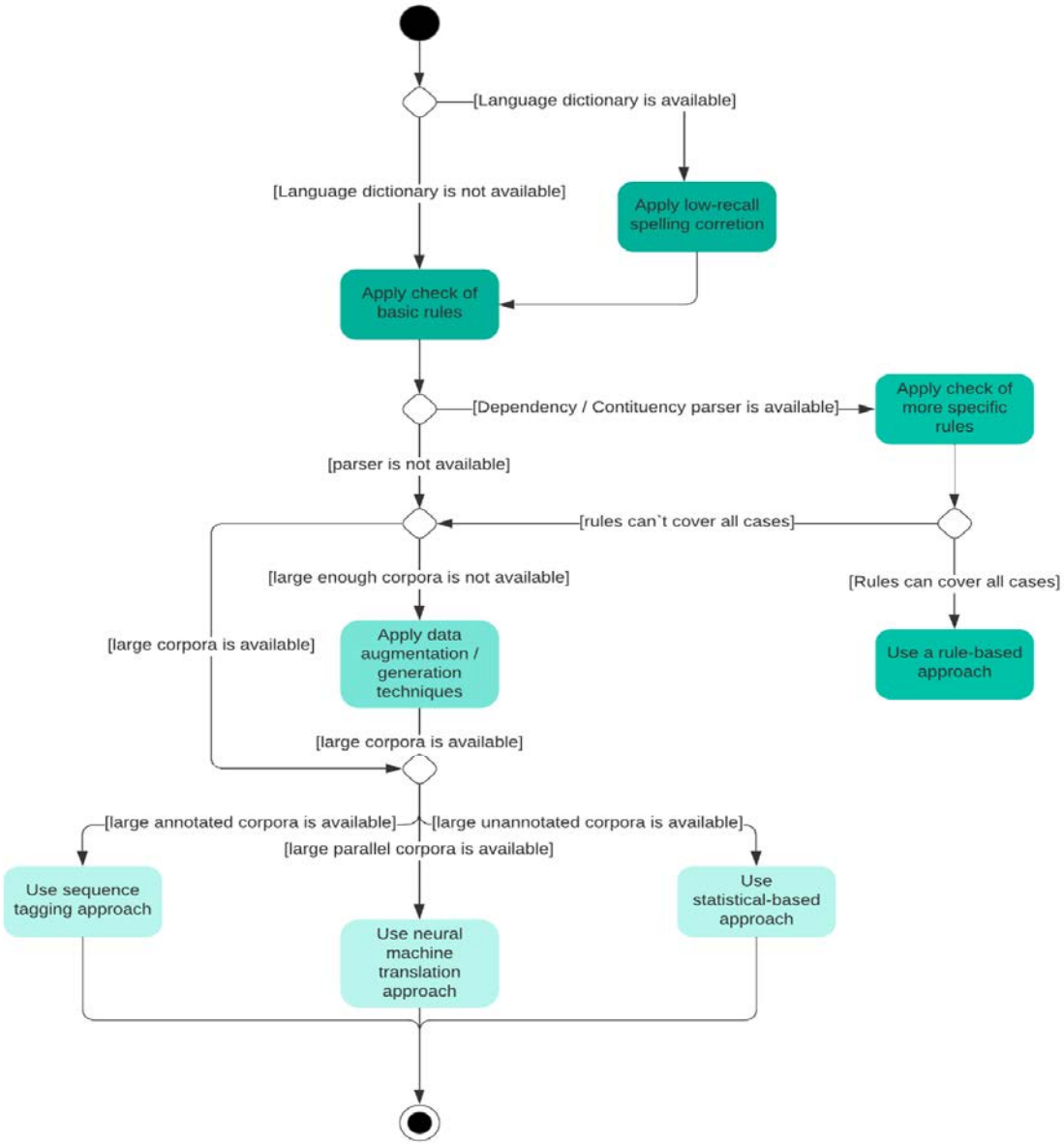


Рисунок 6 – Діаграма діяльності процесу розробки GEC-системи

DOI 10.15588/1607-3274-2023-1-12

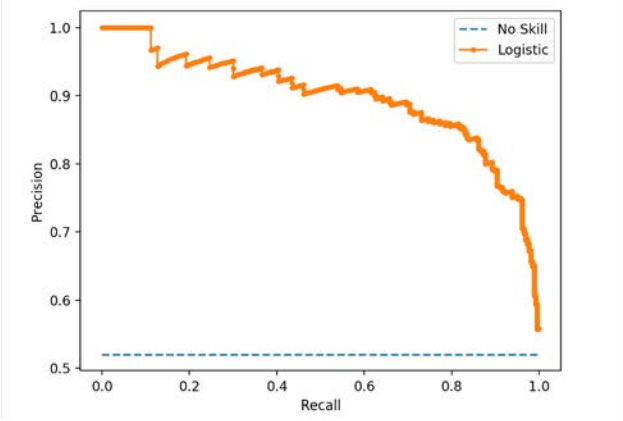


Рисунок 7 – Приклад кривої Precision-Recall

В залежності від того, чи є доступними парсери залежностей, груп слів, іменованих сутностей, частин мови тощо, можна застосовувати перевірку додаткових правил на основі отриманих результатів. Якщо теоретично можливий ітеративний розвиток системи і впроваджені правила можуть пояснити усі можливі помилки певної мови, використання методів і алгоритмів не є необхідним.

У випадку, якщо перевірки додаткових правил не є достатньо (особливо у випадку морфологічно багатих мов), для виправлення залишкових помилок можливим є застосування ML-методів. Якщо навчальних даних не є достатньо або вони взагалі не є доступними, для отримання паралельного / анотованого корпусу застосовують методи аугментації та генерації даних. Також, в залежності від того, чи є дані анотованими або корпус містить набір паралельних текстів, використовують відповідні підходи до розробки і навчання моделі. У випадку достатньо великих і якісних неанотованих корпусів також можна побудувати мовну модель на основі статистичних методів.

На діаграмі розгортання GEC-системи (рис. 8) зображено 3 основні процесори на основі трьох пристроїв/програм, зокрема : Пристрій користувача (:User Device), Сервер БД (:DB Server) та Сервер опрацювання даних (:Data Processing Server). :User Device застосовує програму/браузер (Program/ Browser). :DB Server використовує База даних NoSQL (NoSQL Database). :Data Processing Server побудований на основі реалізації ML-моделі (ML model), словника (Dictionary), модуля перевірки правил (Rule-check module) та модуля огляду користувачів (User review module).

Користувач має доступ до функціоналу системи за допомогою графічного інтерфейсу завантаженої програми, розширення або онлайн-платформи. У БД зберігається інформація про користувачів, приклади запропонованих виправлень, документи. Така БД за своїм типом не є реляційною.

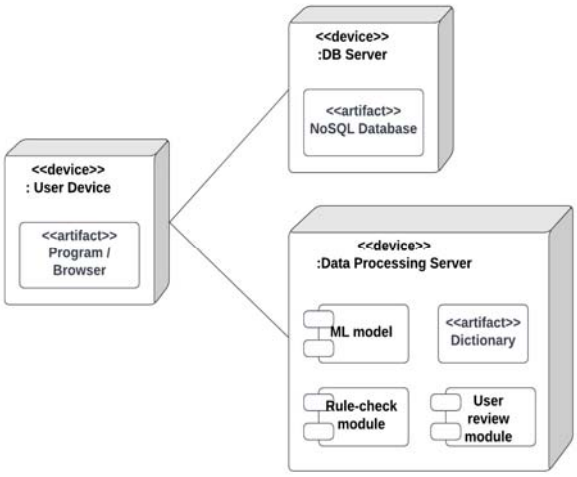


Рисунок 8 – Діаграма розгортання GEC-системи

Сервер обробки даних містить модулі ML-моделі, перевірки правил і перевірки запропонованих користувачами виправлень. Остання може застосовуватись для додаткового налаштування загальної системи перевірки та виправлення.

Альтернативним варіантом схеми розгортання є використання власних обчислювальних ресурсів персонального комп'ютера для аналізу даних, однак, з врахуванням того, що для отримання достовірного результату необхідно застосувати кілька етапів перевірки, особливо з використанням ML-методів, така архітектура не є доцільною.

Вибір набору даних. Набір україномовних UA-GEC анотованих текстів презентований дослідниками з Grammarly. Цей набір даних був професійно анотований для GEC та редагування україномовних текстів. Наскільки відомо на час публікації, це перший корпус GEC з української мови. Дослідники зібрали тексти з помилками (20 715 речень – 328 779 токенів) від різноманітних авторів, у тому числі носіїв мови. Дані охоплюють найрізноманітніші сфери письма, від текстових чатів і есе до офіційного письма. Професійні коректори виправляли та анотували корпус на помилки, пов'язані з вільним мовленням, граматикою, пунктуацією та орфографією. Даний корпус можна використовувати для розробки та оцінки систем GEC українською мовою і дослідження морфологічно багатих мов.

Методи. Основні підходи виявлення помилок, на яких зосереджені академічні дослідження, можна розділити на такі категорії: методи, засновані на правилах; методи, засновані на синтаксичному аналізі речень; статистичне моделювання; класичні ML-методи; MT на основі глибинного навчання. Для покращення якості системи використовують різні методи генерації синтетичних навчальних даних.

MT. Поява у 2018 році нової архітектури глибинного навчання Transformer, що містить у собі механізм уваги, дозволила значно спростити розробку мовних моделей, що тепер може відбуватися без

DOI 10.15588/1607-3274-2023-1-12

застосування експертних знань домену, у даному випадку – лінгвістики. Більша частина досліджень застосовує модель Transformer для впровадження нейронного МТ тексту із помилками у його правильний варіант. Значним недоліком такого підходу для розробки моделей для різних мов є необхідність у великих корпусах анотованих або паралельних даних.

Transformer – модель глибинного навчання, особливістю якої є застосування механізму уваги, що роздільно зважують важливість кожної частини даних входу.

Будівельними блоками трансформера є вузли масштабованої скалярнодобуткової уваги (scaled dot-product attention units). Коли моделі Transformer передають речення, ваги уваги між усіма лексемами обчислюються одночасно. Вузол уваги виробляє вкладення для кожної лексеми в контексті, які містять інформацію про саму лексему, разом зі зваженим поєднанням інших релевантних лексем, кожну з яких зважено за вагою уваги до неї.

Альтернативою архітектурі Transformer є згорткові або рекурентні нейронні мережі архітектури encoder-decoder.

– Згорткові нейронні мережі (CNN, Convolutional Neural Network)

CNN складається з шарів входу та виходу, а також із декількох прихованих шарів. Приховані шари CNN зазвичай складаються зі згорткових шарів, агрегувальних шарів, повноз'єднаних шарів та шарів нормалізації. Згорткові шари застосовують до вхідних даних операцію згортки, передаючи результат до наступного шару.

Згорткові мережі можуть включати шари локального або глобального агрегування, які об'єднують виходи кластерів нейронів одного шару до наступного шару. Наприклад, максимізаційне агрегування використовує максимальне значення з кожного з кластерів нейронів попереднього шару. Іншим прикладом є усереднене агрегування, що використовує усереднене значення з кожного кластеру нейронів попереднього шару. Повноз'єднані шари з'єднують кожен нейрон одного шару з кожним нейроном наступного шару .

– Рекурентні нейронні мережі (Recurrent Neural Networks, RNN)

Рекурентні нейронні мережі це мережі, що містять зворотні зв'язки і дозволяють зберігати інформацію. Наявність зворотного зв'язку дозволяє передавати інформацію від одного кроку навчання мережі до іншого.

Одним з різновидів RNN є LSTM-мережі. LSTM (Long Short Term Memory) – це RNN, здатні до навчання довготривалих залежностей. LSTM-мережі складаються з повторюваних елементів. Кожен такий елемент містить чотири шари і відрізняється тим, що має комірку довгої короткочасної пам'яті.

Можливість LSTM-мереж успішно вивчати дані з довготривалими залежностями робить їх оптимальним вибором для розв'язання задач, у котрих як вхідна, так і вихідна інформація представляються у вигляді послідовностей деяких елементів (наприклад, літер, слів, речень).

Sequence tagging. Окрім методу нейронного перекладу, ML-методи у задачі виявлення та виправлення орфографічних помилок також застосовують для передбачення міток, що позначають дію, яку необхідно виконати над кожним токеном. В залежності від типу помилок, кількості наявних класів міток та загальної складності завдання використовують як глибинні ML-методи, так і класичного.

До класичних ML-методів із вчителем відносяться такі алгоритми: логістична регресія, дерево рішень, наївний байєсів класифікатор, метод опорних векторів, k-найближчих сусідів, випадковий ліс тощо.

– Логістична регресія

Логістична регресія використовується для завдань бінарної класифікації, тобто коли на виході потрібно отримати відповідь, до якого з двох класів належить об'єкт. Логістична функція перетворює будь-яке значення в число в межах від 0 до 1. Так і передбачається ймовірність належності до одного чи другого класу. Тобто, на відміну від логістичної регресії, цей метод не передбачає значення числової змінної, виходячи з вибірки вихідних значень, а замість цього значенням функції є ймовірність того, що це початкове значення належить до певного класу.

– Дерево рішень

Дерево рішень будує моделі класифікації або регресії у вигляді деревоподібної структури. Метод розбиває набір даних на менші й менші підмножини. Кінцевий результат – дерево з вузлами рішення і кінцевими вузлами. Що глибше дерево, то складніші правила ухвалення рішень і точніша модель.

Однією з основних переваг дерев рішень є їхня інтуїтивність: класифікаційна модель, що представлена у вигляді дерева рішень, легко інтерпретується користувачем та спрощує розуміння завдання, яке потрібно вирішити, тому що дозволяє зрозуміти, чому конкретний об'єкт належить до відповідного класу. Ще однією важливою перевагою дерев рішень є їхня універсальність для вирішення задач класифікації та регресії. Але є в алгоритму і недоліки. По-перше, це нестабільність процесу: невеликі зміни в наборі даних можуть призводити до побудови іншого дерева. Це пов'язано з ієрархічністю дерева: зміни в вузлах на верхньому рівні можуть призвести до змін у всьому дереві нижче. По-друге, це складність контролю розміру дерева, що є критичним фактором, який зумовлює якість вирішення завдання.

– Наївний Байєсів класифікатор

Наївний метод Байєса – це набір методів класифікації, заснованих на теоремі Байєса. Модель складається з двох типів ймовірностей, які розраховуються за допомогою тренувальних даних: імовірність кожного класу і умовна ймовірність для

DOI 10.15588/1607-3274-2023-1-12

кожного класу при кожному значенні x. Наївний байєсів класифікатор називається наївним, тому що алгоритм передбачає, що кожна вхідна змінна є незалежною. Це припущення не завжди відповідає реальним даним. Проте даний алгоритм вельми ефективний для цілого ряду складних завдань на зразок класифікації спаму або розпізнавання рукописних цифр.

– Метод опорних векторів

Метод опорних векторів використовується для задач класифікації тексту, як-от призначення категорії і виявлення спаму. Метод опорних векторів – один із найбільш популярних для класичної класифікації, тому що він простий та швидкий. Виходячи з того, що об'єкт, який перебуває в N-вимірному просторі, належить до одного з двох класів, метод опорних векторів будує гіперплощину з розмірністю (N – 1), щоб всі об'єкти виявилися в одній з двох груп. Що далі від гіперплощини лежать точки даних, то більше впевненість в тому, що вони були правильно класифіковані. Що менше схожих ознак між даними, то менша ймовірність належності до одного класу.

– Метод k-найближчих сусідів

Для алгоритму k-найближчих сусідів передбачення для нової точки робиться шляхом пошуку k найближчих сусідів в наборі даних і підсумовування вихідної змінної для цих k примірників. Ідея найближчих сусідів може погано працювати з багатовимірними даними, що негативно позначиться на ефективності алгоритму при вирішенні задачі.

– Випадковий ліс

Випадковий ліс належить до сімейства ансамблевих алгоритмів. Як зрозуміло з його назви, він містить велику кількість окремих дерев рішень, які діють як ансамбль. Кожне окреме дерево у випадковому лісі обчислює прогноз класифікації, і клас із найбільшою кількістю голосів стає прогнозом моделі. Результатом обирається той варіант, який випадав найчастіше. Велика кількість некорельованих дерев рішень, тобто тих, що знаходять рішення незалежно одне від одного й діють спільно, перевершить будь-яке рішення, отримане одним деревом рішення.

Саме некорельовані моделі можуть створювати ансамблеві прогнози, які є точнішими, ніж будь-який з окремо взятих прогнозів. Причиною такого ефекту є те, що кожне дерево рішень «захищає» одне одного від індивідуальних помилок: хоча деякі дерева рішень можуть бути неправильними, більшість із них будуть правильними, тому як група дерев буде рухатися в правильному напрямку.

До переваг методу можна віднести можливість ефективно обробляти дані з великою кількістю ознак і класів та високу масштабованість. Серед недоліків можна виділити великий розмір моделей, що будуються. Що більша модель, то вища її обчислювальна складність та швидкість знаходження рішень.

– AdaBoost

Бустінг – це сімейство ансамблевих алгоритмів, суть яких полягає в створенні сильного класифікатора на основі декількох слабких. Для цього спочатку створюється одна модель, потім інша модель, яка намагається виправити помилки в першій. Моделі додаються до тих пір, поки тренувальні дані не будуть ідеально класифіковані або поки не буде перевищено максимальну кількість моделей.

AdaBoost використовують разом з короткими деревами рішень. Після створення першого дерева перевіряється його ефективність на кожному тренувальному об'єкті, щоб зрозуміти, скільки уваги має приділити наступне дерево всіх об'єктах. Тим даним, які складно передбачити, дається більшу вагу, а тим, які легко передбачити, – меншу. Моделі створюються послідовно одна за одною, і кожна з них оновлює ваги для наступного дерева. Після побудови всіх дерев робляться прогнози для нових даних, і ефективність кожного дерева залежить від того, наскільки точним воно було на тренувальних даних.

Для обрання найкращого методу класифікації для його подальшого застосування необхідно порівняти показники точності, повноти, влучності, F-міри для всіх вищезазначених алгоритмів.

Морфологічні парсери. Перевірка правил відмінювання та узгодження слів у реченні може бути заснована на парсингу дерев залежності (Dependency Parsing) або приналежності (Constituency Parsing).

Термін парсинг залежностей відноситься до процесу дослідження залежностей між фразами речення з метою визначення його граматичної структури. Речення ділиться на багато розділів здебільшого на основі цього. Процес заснований на припущенні, що між кожною мовною одиницею в реченні існує прямий зв'язок. Ці гіперпосилання називаються залежностями. На відміну від аналізу залежностей, парсинг приналежностей розподіляє речення на фрази певного типу (іменникова, дієслівна групи тощо). У цьому випадку термінальний вузол – це мовна одиниця або фраза, яка має батьківський вузол та тег частини мови.

Парсинг залежностей та приналежностей слів у реченнях певних мов можна виконати за допомогою бібліотеки для NLP-NLTK, однак вона не підтримує аналіз україномовних текстів.

Підтримка української мови. Бібліотека мови Python pymorphy2 підтримує обробку української мови. pymorphy2 аналізує частину мови, число, відмінок, час, лему та стем заданого слова. Морфологічний парсер заснований на словниках OpenCorpora, що конвертовані до формату XML. Користувачі також мають змогу додавати власні слова та правила, це дає можливість проводити морфологічний аналіз текстів певної предметної області без зміни вихідного коду pymorphy2 та адаптувати pymorphy2 для роботи з іншими мовами.

Мова і бібліотеки. Для ML найпопулярнішими мовами є Python, R, Java, C++. Перевагою Python з-

DOI 10.15588/1607-3274-2023-1-12

поміж інших мов саме для створення системи визначення емоційного забарвлення текстового контенту є його підтримка великої кількості бібліотек:

– для роботи з класичними ML-методами: Scikit-Learn;

– для створення штучних нейронних мереж, глибинного навчання: TensorFlow, Keras, PyTorch;

– для NLP: NLTK, spaCy, WordNet;

– для роботи із масивами, матрицями: NumPy;

– роботи з таблицями: pandas;

– візуалізації даних (в тому числі, й інтерактивної): Matplotlib, seaborn, Plotly.

Бібліотека Scikit-Learn підтримує TPP, зменшення розмірності даних, вибір ML-моделей для регресії, класифікації або кластерного аналізу. Однак Scikit-Learn не має комплексної підтримки для створення моделей глибинного навчання.

Для створення штучних нейронних мереж була обрана бібліотека Keras, що слугує високорівневим API для TensorFlow 2. Keras дозволяє будувати послідовні моделі у вигляді графу, вершинами якого є шари (layers) певного типу із заданою кількістю вузлів. Також, за допомогою Keras можна поєднувати результати роботи кількох окремих частин нейронної мережі для їх подальшої обробки, така структура не є лінійною.

TensorFlow дає можливість імпортувати навчену ML-модель для її подальшого використання у інших програмах. TensorFlow також підтримує виконання низькорівневих операцій із тензорами за допомогою центральних процесорів, графічних процесорів та тензорних блоків обробки.

Бібліотека NLTK у цьому дослідженні застосовується для попередньої обробки тексту: токенізації, видалення стоп-слів, стемінгу, лематизації. Також за допомогою функцій з цієї бібліотеки можна виявляти найпопулярніші n-грами і частини мови окремих токенів, розпізнавати іменовані сутності тощо.

До додаткових бібліотек, що спрощують роботу із природньою мовою, належать Regex та emoji – для використання регулярних виразів і заміни емоджі словами відповідно.

Середовище розробки. Для роботи над задачами з дослідження даних і застосування ML зручним інструментом є Jupyter Notebook, який дозволяє запускати написаний код невеликими фрагментами – комірками. Одним із онлайн-сервісів, що надає змогу використовувати Jupyter Notebook без локального встановлення, є Google Colab. Цей сервіс дає можливість використовувати графічні процесори GPU і TPU, що значно пришвидшують навчання штучних нейронних мереж.

4 ЕКСПЕРИМЕНТИ

Опрацювання набору даних. Обраний набір даних UA-GEC [15] містить 850 і 160 текстів у навчальній і тестовій вибірці відповідно. За даними авторів, набір даних містить усього 20 715 речень (як з помилками, так і без). На момент написання роботи вбудовані методи ітерації по корпусу UA-GEC дозволяють лише отримувати повні документи, а не окремі речення.

Текст анотований за наступним форматом:

```
I {like=>likes:::error_type=Grammar} turtles.
```

Для опрацювання та розділення текстів була застосована бібліотека ReGex. Регулярні вирази для розподілу текстів на речення та виявлення помилок мають наступний формат:

```
split_pattern = r'\n+'
additional_split_pattern = r'(?<=[^А-Я].[.?!(\.\.\.)…])
+(?=[А-Я“"])'
error_pattern =
r'\{(([^\{]*\(*\)*)=>([^\{]*\(*\)*):::error_type=([^\{]*))\}'
```

Правильні пари не видаляються, однак для подальшого опрацювання були обрані речення, де кількість токенів є більшою за 4 (два з них – спеціальні токени початку і кінця речення).

Усього в результаті розподілу текстів за вищезазначеними виразами отримуємо 15 599 і 2205 речень у навчальній та тестовій вибірках відповідно.

Попередньо навчені нейронні мережі.

Морфологічна складність української мови зумовлює потребу у поєднанні різних GEC-методів.

Один із модулів запропонованої системи – опрацювання речення за допомогою штучних нейронних мереж. Навчання нейронної мережі «з нуля» з випадково ініціалізованими вагами потребує паралельних корпусів великих обсягів. Єдиний на момент написання роботи україномовний набір даних містить усього тисячу текстів, що є критично мало для навчання мовної моделі для обробки морфологічно багатих мов.

Тому, навчання нейронної мережі GEC-завданню вимагає її попереднього налаштування за допомогою великих корпусів. У цьому випадку існує кілька попередньо налаштованих моделей, що підтримують українську і параметри яких є доступними у відкритому доступі: RoBERTa від YouScan [48], MT-моделі (від Google – MT5 [49], від Facebook – M2M100 [50] та mBART-50 [51]).

Модель архітектури енкодер-декодер на основі RoBERTa. У 2020 році команда дослідників із YouScan презентувала нейронну мережу «Ukrainian Roberta» [48], що була попередньо натренована на україномовних текстах загальним обсягом 2,5 мільярди слів. «Ukrainian Roberta» має 125 мільйонів параметрів, що налаштовуються, її навчання тривало 85 годин.

Складністю використання цієї нейронної мережі є те, що вона за своєю архітектурою є тільки енкодером і використовується для отримання контекстних векторних вкладень слів. «Ukrainian Roberta» була навчена вирішувати завдання передбачення замаскованого токену у реченні (рис. 9). Попереднє навчання нейронної мережі дозволяє їй отримати попереднє «знання» граматики мови.

DOI 10.15588/1607-3274-2023-1-12

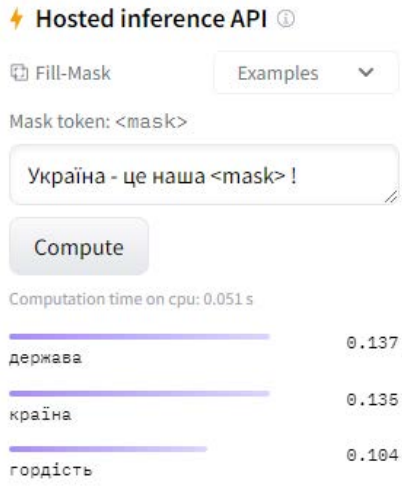


Рисунок 9 – Приклад передбачення замаскованого токену

BERT – модель глибинного навчання для NLP архітектури Transformer. Базова модель BERT має 110 мільйонів, розширена версія – 345. Особливістю архітектури Transformer є наявність механізму уваги, завдяки чому дані можуть бути оброблені одночасно (на противагу рекурентним нейронним мережам, де дані сприймаються послідовно). Окрім того, особливістю BERT є попереднє навчання нейронної мережі для вирішення двох завдань: передбачення певного слова у реченні і визначення того, чи є друге речення логічним продовженням першого. Попереднє навчання та механізм уваги дають змогу отримати контекстні векторні вкладення слів.

Попередньо навчену нейронну мережу RoBERTa можна вважати покращеним аналогом BERT: її архітектура є аналогічною до архітектури BERT, основна відмінність полягає у підборі гіперпараметрів під час попереднього навчання, збільшенні об'єму навчального корпусу, застосуванні динамічного маскування токенів для задачі передбачення слова у реченні: слово, яке необхідно передбачити у певному реченні, змінюється з кожною епохою. Окрім того, у цьому випадку відсутнє передбачення того, чи є друге речення логічним продовженням першого. Наприклад, для класифікації речень додається шар повнозв'язних нейронів, де кількість нейронів відповідає наявним класам [52].

Нейронні мережі типу денкодер утворюють констекстні векторні представлення слів у реченні фіксованої довжини незалежно від довжини вхідного повідомлення. У свою чергу, BERT є складовою нейронних мереж типу енкодер-декодер, що використовуються для завдань sequence-to-sequence. Декодери, у свою чергу, генерують токени послідовно, кожний наступний токен залежить від попередніх. Застосування попередньо навченої «Ukrainian Roberta» є можливим за аналогією до підходу, запропонованого у [53]. У випадку ініціалізації BERT у якості декодера, для передбачення наступного слова на основі контекстного векторного вкладення вхідних токенів, необхідно додати шар перехресної уваги (cross-attention layer) із випадковим чином згенерованими вагами. Окрім того, для передбачення розподілу ймовірностей наступного слова використовується LM Head (шар або шари повнозв'язних нейронів, де кількість вихідних нейронів відповідає кількості токенів у словнику). Вагові коефіцієнти LM Head ініціалізуються ваговими коефіцієнтами шару векторного вкладення BERT W_emb. Також, двонапрямлений механізм уваги BERT необхідно замінити на однонапрямлений для генерації токену в залежності лише від попередніх [52, 57]. Словник токенів, що відповідає попередньо навченій моделі «Ukrainian Roberta», містить 52 тис. елементів, де перші записи є спеціальними токенами, розділовими знаками та найбільш поширеними словами (рис. 10). Оскільки словники токенів є унікальними для кожної попередньо навченої моделі, TPP теж має свої особливості. Після токенізації текстів і речень у навчальній вибірці розподіл кількості токенів має наступний вигляд, поданий на рис. 11. Максимальна довжина речення – 100 токенів, довші речення обрізаються.

Нейронна мережа навчалась протягом 15 епох з наступними гіперпараметрами:

- n_epochs = 15
- batch_size = 8
- lr = 0.00001

Після 15-ї епохи значення функції втрат на тестовій вибірці збільшується (рис. 12), тому 15-ть епох є найбільш оптимальним значенням у даному випадку.

Налаштування попередньо навченої МТ-моделі. Facebook та Google надали відкритий доступ до ваг попередньо навчених нейронних мереж для МТ.

```
sorted(tokenizer.vocab.items(), key=lambda x: x[1])
```

```
[('<s>', 0),
 ('<pad>', 1),
 ('</s>', 2),
 ('<unk>', 3),
 ('<mask>', 4),
 ('!', 5),
 ('"', 6),
 ('#', 7),
 ('$', 8),
 ('%', 9),
 ('&', 10),
 ("'", 11),
 ('(', 12),
 (')', 13),
 ('*', 14),
 ('+', 15),
```

Рисунок 10 – Словник токенів «Ukrainian Roberta»

DOI 10.15588/1607-3274-2023-1-12

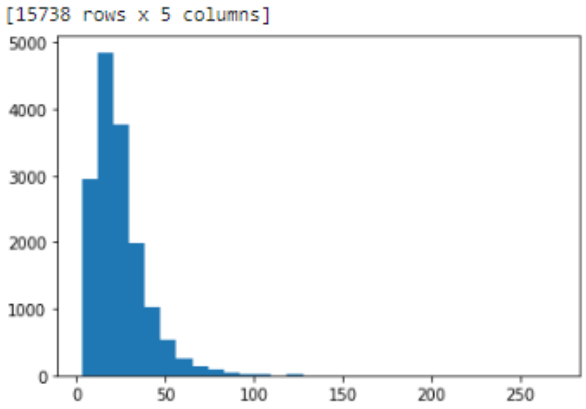


Рисунок 11 – Розподіл кількості токенів у навчальній вибірці

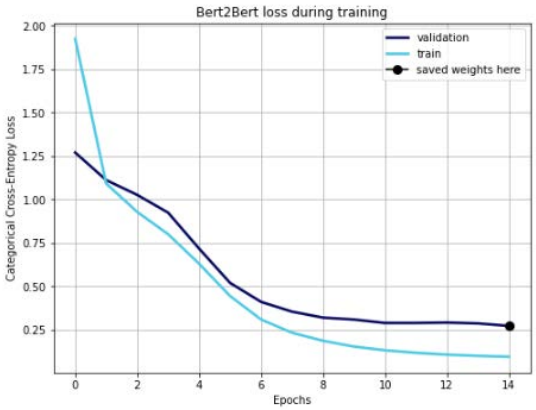


Рисунок 12 – Крива функції втрат при навчанні NN

mT5 [49] заснована на архітектурі Transformer і була попередньо навчена MT-завданню із застосуванням корпусу, заснованого на даних Common Crawl. Усього нейронна мережа була попередньо натренована на текстах, написаних 101 мовою, серед яких є і українська. Базова версія, що застосовується у цьому дослідженні, має 580 мільйонів параметрів, що налаштовується, версія mT5-XXL – 13 мільярдів параметрів.

У зв'язку з обмеженням наявних обчислювальних ресурсів, найбільша довжина речення складає 20 токенів, batch size – 2 записи. Варто зауважити, що, оскільки нейронна мережа була попередньо натренована за допомогою текстів, написаних різними мовами, то відповідний словник токенів радше містить частини україномовних слів, аніж повні слова. Тому одне слово кодується більшою кількістю токенів, а у контексті обмеженої довжини речення це може призвести до некоректного представлення навчальних прикладів і їх значення, що не є повністю завершеним. Нейронна мережа навчалась протягом 10 епох, найнижче значення функції втрат на тестовій вибірці було досягнуто на 4 епосі, тоді ж і були збережені ваги нейронної мережі.

Відповідно до [50], M2M100 має низку переваг у порівнянні із попередніми MT-моделями. За ствердженнями дослідників, попередні MT-моделі навчаються на «англо-центричних» даних, тобто для перекладу між двома мовами такі моделі навчаються перекладу з першої мови на англійську, потім – з англійської на другу мову. Такий підхід може призвести до втрати початкового значення речення, і M2M100 – перша нейронна мережа для MT, що була навчена прямому перекладу з першої мови на другу без проміжного перекладу на англійську. Такий підхід дозволив отримати найкращі результати метрики BLEU у порівнянні з іншими дослідженнями, збільшивши показник на 10 пунктів. Усього нейронна підтримує 2 200 напрямків перекладу – у 10 разів більше, ніж попередня англо-центрична мовна модель. NN попередньо натренована на тестах, написаних 100 мовами, у тому числі й українською. Базова версія, що застосовується у дослідженні, містить 418 мільйонів параметрів, найбільша – 15 мільярдів. Після розрахунку кількості токенів для текстів навчальної вибірки отримуємо наступний розподіл, поданий на рис. 13.

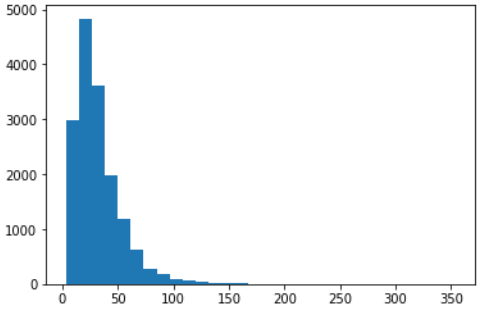


Рисунок 13 – Розподіл кількості токенів у навчальній вибірці

Аналогічно до попереднього випадку, максимальна кількість токенів є обмеженою у зв'язку з обмеженими обчислювальними ресурсами. В цьому випадку, також можливе некоректне представлення навчальних прикладів. Нейронна мережа навчалась протягом 10 епох, однак найменше значення функції втрат на тестовій вибірці було досягнуто на другій епосі (рис. 14). Для порівняння також було проведене навчання нейронної мережі протягом 10 епох, ваги були збережені після останньої.

Варто зауважити, що на якість генерації тексту впливає низка параметрів, а саме:

– Значення гіперпараметрів.
– Обсяг навчальної вибірки.
– Кількість навчальних епох.
– Архітектура і кількість параметрів нейронної мережі.
– Кількість токенів вхідного тексту.
– Навчальна вибірка (чи застосовувались валідаційні дані або був використаний весь корпус).

Приклад навчання NN MT5. Імпорт бібліотек. Основними бібліотеками у даному проекті є PyTorch, HuggingFace Transformers, Regex, pandas (рис. 15).

DOI 10.15588/1607-3274-2023-1-12

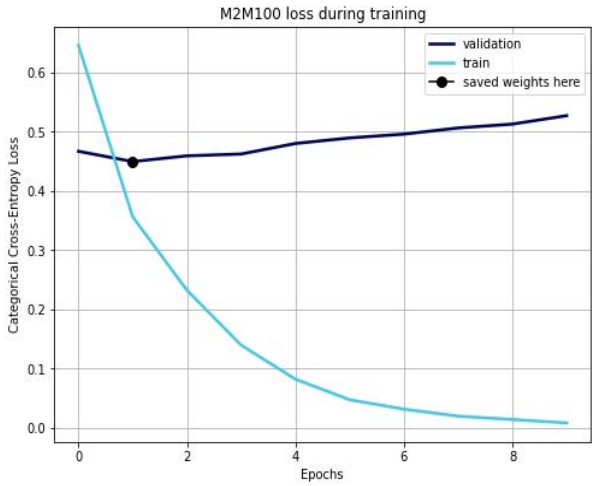


Рисунок 14 – Крива функції втрат при навчанні NN

```
################################  IMPORTS  ################################

from collections import defaultdict
from ua_gec import Corpus
from ua_gec.corpus import Document

import pandas as pd
import regex as re
import copy
import numpy as np
import matplotlib.pyplot as plt

from transformers import AdamW, AutoModelForSeq2SeqLM, AutoTokenizer
from transformers import get_linear_schedule_with_warmup
from transformers import Seq2SeqTrainer, Seq2SeqTrainingArguments

from IPython.core.display import display, HTML
from tqdm import tqdm_notebook

import torch
from torch.utils.data import Dataset, DataLoader
```

Рисунок 15 – Імпорт бібліотек

Константи на налаштування. Встановлюємо такі налаштування: регулярні виразу для розділення текстів на речення, регулярний вираз для пошуку анотації помилок, гіперпараметри (рис. 16). Завантаження попередньо навчено моделі. Завантаження попередньо навченої NN для її подальшого налаштування відбувається за допомогою кількох команд (рис. 17). Створення PyTorch Dataset. Перед створенням класу Датасет та його об’єктів необхідно провести попереднє дослідження даних, після цього агрегуючи ці етапи (рис. 18). Додаткові налаштування подані на рис. 19. При використанні PyTorch необхідно кожного разу реалізувати власну функцію навчання NN (рис. 20).

```
################################  SETTINGS  ################################

split_pattern = r'\n+'
error_pattern = r'\{{([^\{]*\(*\)*)=>([^\{]*\(*\)*):::error_type=([^\{]*)}\}'
additional_split_pattern = r'(?<=[^A-Я].[.?!(\.\.\.)…]) +(?=[A-Я""])'

regex_special_symbols = {
                '(': '\(',
                ')': '\)',
                '?': '\?',
                '[': '\[',
                ']': '\]',
                '*': '\*',
                '$': '\$'
            }

model_name = 'google/mt5-base'

n_epochs = 25
batch_size = 4
lr = 0.00001
max_token_length = 50

device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
```

Рисунок 16 – Змінні із налаштуваннями

```
#########################  LOAD PRETRAINED MODEL  #########################

tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSeq2SeqLM.from_pretrained(model_name)
model = model.to(device)
```

Рисунок 17 – Завантаження попередньо навченої NN

```
train_dataset = Parallel_Dataset('train', tokenizer, max_token_length)
test_dataset = Parallel_Dataset('test', tokenizer, max_token_length)
```

Рисунок 18 – Завантаження датасету

```
############################  ADDITIONAL  ############################

test_loader = DataLoader(test_dataset, batch_size=batch_size)
train_loader = DataLoader(train_dataset, batch_size=batch_size)

n_batches = int(np.ceil(len(train_dataset)) / batch_size)
n_batches_test = int(np.ceil(len(test_dataset)) / batch_size)

total_steps = n_epochs * n_batches

n_warmup_steps = int(total_steps * 0.01)

optimizer = torch.optim.AdamW(model.parameters(), lr=lr)
scheduler = get_linear_schedule_with_warmup(optimizer,
                                            n_warmup_steps,
                                            total_steps)
```

Рисунок 19 – Додаткові налаштування

```
model, history = train(model, optimizer, scheduler,
                       n_epochs, len(train_dataset), len(test_dataset),
                       train_loader, test_loader)
```

Рисунок 20 – Виклик функції навчання NN

Збереження ваг NN: torch.save(model, 'mt5.pt'). Отримання передбачення подано на рис. 21.

DOI 10.15588/1607-3274-2023-1-12

```
def predict(sentence):
    inputs = tokenizer(sentence, padding="max_length", truncation=True, max_length=200, return_tensors="pt")
    input_ids = inputs.input_ids.to("cuda")
    attention_mask = inputs.attention_mask.to("cuda")

    outputs = model.generate(input_ids, attention_mask=attention_mask)
    print(outputs)

    output_str = tokenizer.batch_decode(outputs, skip_special_tokens=True)

    display(HTML(f'<h3>Input sentence</h3>'))
    display(HTML(f'<p>{sentence}</p>'))
    display(HTML(f'<h3>Result</h3>'))
    display(HTML(f'<p>{output_str[0]}</p>'))

predict('учора був чудове день')
```

Рисунок 21 – Функція отримання передбачення

5 РЕЗУЛЬТАТИ

Для демонстрації роботи нейронних мереж використані власні приклади та завдання Зовнішнього Незалежного Оцінювання. Отримані результати: «Ukrainian Roberta» encoder-decoder. Нейронна мережа непогано розставляє розділові знаки у простому реченні (рис. 22).

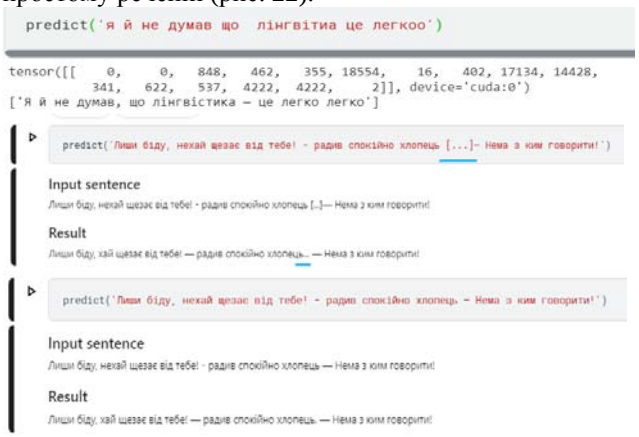


Рисунок 22 – Розставлення розділових знаків

Наступний приклад: NN правильно виправила розділові знаки, однак самостійно змінила речення, використовуючи відомі їй словосполучення (рис. 23). Інколи NN «не розуміє» суті поставленого завдання (рис. 24 – залишити правильне словосполучення у початковому вигляді). Інколи NN взагалі «не розуміє» суті поставленого завдання (рис. 25). Однак у цьому випадку речення має логічну структуру. На рис. 26 NN правильно змінила відмінок іменника з називного на кличний, однак гуаш – фарба, іменник жіночого роду, тут він не є правильно узгодженим, ймовірно тому, що аналогічних прикладів немає у навчальному наборі даних.

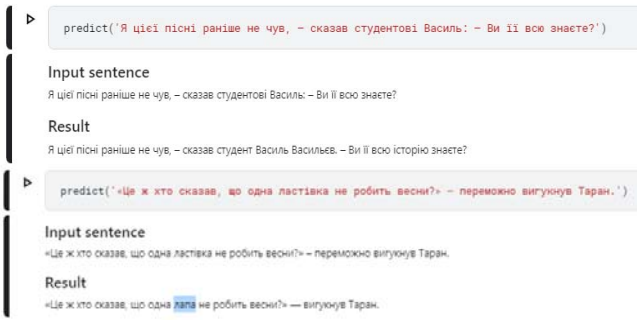


Рисунок 23 – Зміна речення ейронною мережею

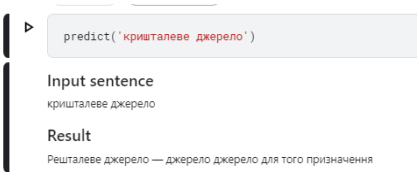


Рисунок 24 – Нерозуміння задачі NN

Рисунок 25 – Некоректна зміна речення

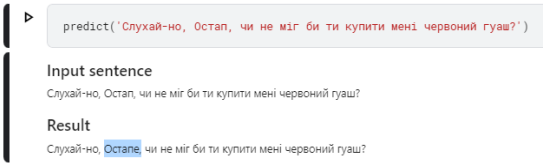


Рисунок 25 – Зміна відмінку іменника

На рис. 26 NN правильно виправила числівник, але завершила речення своєю «творчістю». У прикладах, що містять лише слова або словосполучення, нейронна мережа «намагається» завершити думку, самостійно добудовуючи речення. Такий результат є прямим наслідком того, що приклади у навчальному наборі містять лише речення, а не словосполучення.

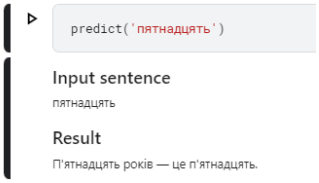


Рисунок 26 – «Творчість» NN

Однак, для перевірки слів і словосполучень в загальному застосовуються інші методи: перевірка за словником та перевірка простих правил (апостроф, чергування у-в, правопис пів-, напів- тощо).

Попередні результати тестування NN показують, що NN неправильно визначає суть поставленого завдання (виправлення помилок при збереженні початкового змісту речення / словосполучення), що може бути результатом відсутності окремих слів та словосполучень у навчальному наборі даних. На противагу цьому, NN виправляє речення паттернами з набору навчальних даних, що навпаки може свідчити про її перенавчання. Окрім того, причиною таких помилок може бути архітектура моделі – RoBERTa оригінально застосовується як енкодер для отримання контекстних представлень токенів. На рис. 27 NN виправила «чим…тим» на «що…то». Деякі

DOI 10.15588/1607-3274-2023-1-12

мовознавці вважають, що цей сполучник не є власне українським, а суржиком. Прості речення є зазвичай простим завданням для NN (рис. 28), але є складності (рис. 29).

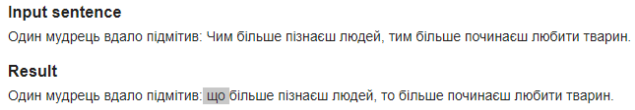


Рисунок 27 – Виправлення NN суржика

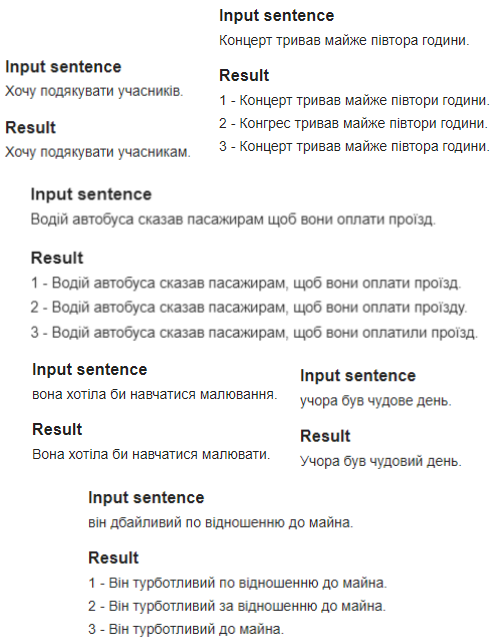


Рисунок 28 – Виправлення простих речень

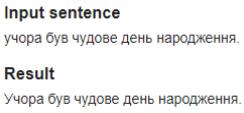


Рисунок 29 – Некоректне виправлення

Правильне речення є серед пропозицій моделі, однак проблемним є той факт, що більша ймовірність була призначена неправильному реченню (рис. 30). На рис. 31а NN правильно замінила словосполучення «за професії», однак слово «архітектор» замінене на найближче за значенням слово. Аналогічно, більшу ймовірність має неправильне речення (рис. 31б), але правильне виправлення є серед запропонованих.

У деяких реченнях mT5 пропонує більш правильні виправлення (рис. 32). На даних прикладах найкращу точність виправлення має МТ-модель mT5, однак для остаточної оцінки і порівняння NN необхідний достатньо великий тестовий корпус даних.

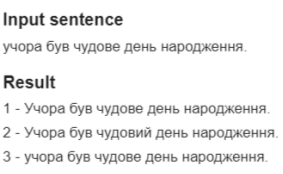
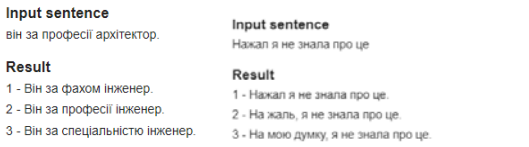


Рисунок 30 – Варіанти виправлення



а б

Рисунок 31 – Коректне виправлення

Рисунок 32 – Варіанти корекції різним NN-моделями

Із наведених прикладів видно, що mT5 краще вдається виправляти граматичні без зміни початкового змісту речення. Однак дана модель має вдвічі більше параметрів, ніж енкодер-декодер, заснований на Roberta (580 мільйонів проти 250), що вимагає значних обчислювальних ресурсів для навчання NN та отримання передбачення. Нейронна мережа M2M100 значно гірше виправляє граматичні помилки, ніж дві попередні моделі (рис. 33)

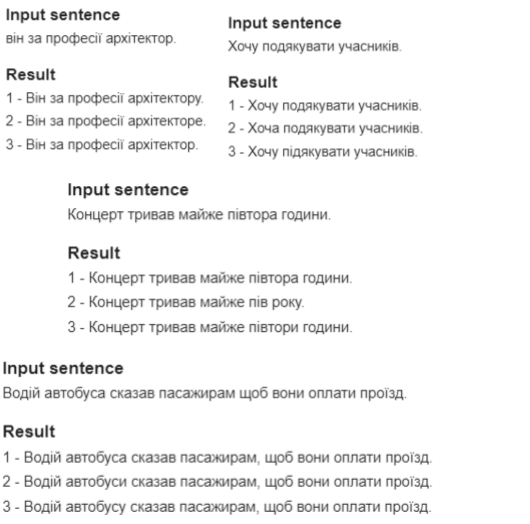


Рисунок 33 – Робота нейронної мережі M2M100

DOI 10.15588/1607-3274-2023-1-12

6 ОБГОВОРЕННЯ

Порівняння нейронних мереж за допомогою спеціальних метрик. Відповідно до [54], для оцінки GEC-якості не існує базових, уставлених метрик. Із метою оцінки таких систем в загальному використовують метрики МТ-якості, однак і вони мають певні недоліки.

BLEU (bilingual evaluation understudy) [55] – це алгоритм для оцінки МТ-якості. Оцінки розраховуються для окремих перекладених сегментів (зазвичай речень) шляхом порівняння їх з набором правильних перекладів хорошої якості. Ці бали потім усереднюються по всьому корпусу, щоб отримати оцінку загальної якості перекладу. Оцінка за BLEU завжди є числом від 0 до 1. Це значення вказує, наскільки текст-кандидат подібний до еталонних текстів, а значення, ближчі до 1, означають більш подібні тексти. Значення BLEU засноване на розрахунку кількості спільних N-грам у початковому реченні та його перекладі. Недоліком метрики BLEU є те, що вона заснована на порівняннях стрічок, і, відтак, ігнорується той факт, що деякі граматичні помилки можна виправити кількома способами. Окрім того, результати отримані результати GEC-якості не можна порівняти із результатами певних систем перекладу, оскільки у першому випадку більша кількість н-грамів у реченнях буде співпадати (або співпадати повністю, якщо речення не містить граматичних помилок і не було змінене).

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [56] – ще одна метрика для оцінювання МТ-якості. Вона заснована на використанні n-gram та орієнтована на використання статистичної та точної оцінки вихідного тексту. На відміну від метрики BLUE, дана метрика використовує функції співставлення синонімів разом із точною відповідністю слів. Метрика була розроблена, щоб вирішити проблеми, які мала BLUE, а також відтворити хорошу кореляцію з оцінкою експертів на рівні словосполучень або речень.

В результаті запуску метрики на рівні словосполучень кореляція з людським рішенням становила 0.964, тоді кореляція з BLUE становила 0,817 на тому ж наборі вхідних даних. На рівні речень максимальна кореляція з оцінкою експертів була 0,403 [56].

Як і в метриці BLUE, основна одиниця для оцінки – речення, алгоритм спочатку проводить вирівнювання тексту між двома реченням, рядком еталонного перекладу та рядком вхідного тексту для оцінювання. Дана метрика використовує декілька етапів встановлення відповідності між словами МТ й еталонного перекладу для зіставлення двох рядків:

1. Точне встановлення відповідності – визначаються рядки, що є ідентичними в еталонному і машинному перекладі.

2. Встановлення відповідності основ – проводиться стемінг (виділення основи слова) і визначаються слова з однаковим коренем в еталонному і машинному перекладі.

3. Встановлення відповідності синонімів – визначаються слова, що є синонімами відповідно до WordNet.

Розраховані метрики МТ-якості дають змогу лише частково порівняти моделі, оскільки більшість слів і словосполучень у початковому та виправленому реченні співпадають. Найкраще значення як BLEU (0,908) , так і METEOR (0,956) отримано для mT5, що співпадає із аналізом прикладів, у якому найбільш точні виправлення помилок без зміни початкового значення речення отримані саме для цієї нейронної мережі. M2M100 має більшу оцінку BLEU (0,847), ніж «Ukrainian Roberta» Encoder-Decoder (0,697), однак, суб'єктивно оцінюючи результати виправлення прикладів, M2M100 значно гірше справляється із цим завданням, ніж дві інші моделі. Для METEOR також M2M100 (0.925) має більшу оцінку, ніж «Ukrainian Roberta» Encoder-Decoder (0,876). Щодо розробки проекту можна зробити такі висновки:

– Для створення якісної ML-моделі для GEC у текстах тих мов, які є складними морфологічно, необхідна велика кількість паралельних або вручну розмічених даних. Ручна анотація даних вимагає багато зусиль професійних лінгвістів, що робить створення корпусів текстів, особливо морфологічно багатих мов, часо- та ресурсозатратним процесом.

– Вирішення завдання автоматичного виявлення та виправлення помилок в україномовних текстах вимагає подальших досліджень у зв'язку з невеликою кількістю робіт, що фокусуються на обробці саме української мови. Окрім того, відповідно до результатів дослідження Погорілого С. Д. і Крамова А. А. [27], методи, які застосовуються для обробки англійської мови, не можуть бути використані для української, оскільки остання є значно складнішою і морфологічно багатшою мовою.

– Поява у 2017 році нової архітектури глибинного навчання Transformer [3], що містить механізм уваги, дозволила значно спростити розробку мовних моделей, що тепер може відбуватися без застосування експертних знань домену, у даному випадку – лінгвістики. Більша частина досліджень застосовує модель Transformer для впровадження нейронного МТ тексту із помилками у його правильний варіант. Значним недоліком такого підходу для розробки моделей для різних мов є необхідність у великих корпусах анотованих або паралельних даних. Єдиний (на момент написання роботи) україномовний набір даних [15] містить усього 20 715 речень, такої кількості навчальних зразків не є достатньо для створення автоматичної інтелектуальної МТ-системи.

– Застосування перевірки правил та парсингу синтаксичних дерев при розробці GEC-системи в україномовних текстах може бути обумовлене такими двома факторами:

1) відсутністю достатньо великих анотованих / паралельних корпусів української мови для навчання

DOI 10.15588/1607-3274-2023-1-12

повністю автоматичних моделей, що засновані на алгоритмах глибинного навчання;

2) недостатньою спроможністю лінгвістичних моделей штучного інтелекту узагальнювати граматичні правила природньої мови [37].

– розробка якісної системи перевірки граматичної правильності речень в україномовних текстах вимагатиме поєднання ML-алгоритмів із кількома різними типами методів, зокрема і застосування експертних знань з комп'ютерної лінгвістики.

– Відповідно до результатів досліджень, для отримання найкращої GEC-точності за допомогою нейронної мережі, потрібно використовувати попередньо навчені MT-моделі, що підтримують українську мову.

– Найкраще значення як BLEU, так і METEOR було отримано для MT-моделі mT5, результати співпадають із аналізом власних прикладів, у якому найбільш точні виправлення помилок без зміни початкового значення речення були отримані саме для цієї нейронної мережі. M2M100 має більшу оцінку BLEU, ніж «Ukrainian Roberta» Encoder-Decoder, однак, суб'єктивно оцінюючи результати виправлення власних прикладів, M2M100 значно гірше справляється із цим завданням, ніж дві інші моделі. З метою економії обчислювальних ресурсів можливим також є застосування попередньо навченої нейронної мережі типу BERT, використовуючи її як у якості енкодера, так і декодера. Така нейронна мережа має вдвічі менше параметрів, ніж інші попередньо навчені MT-моделі, і показує задовільні GEC-результати.

ВИСНОВКИ

Результатом роботи є розроблена ML-модель для GEC в україномовних текстах. Ми також пропонуємо універсальну схему розробки GEC-системи для різних мов. Відповідно до отриманих результатів, нейронна мережа має здатність виправляти прості речення, написані українською, однак розробка повноцінної системи вимагатиме застосування перевірки орфографії за допомогою словників і перевірки правил, як простих, так і заснованих на результаті парсингу залежностей або інших ознак.

З-поміж трьох моделей, найкращі показники має попередньо навчена модель нейронного перекладу mT5. Найкраще значення як BLEU (0,908) , так і METEOR (0,956) отримано для mT5, що співпадає із аналізом прикладів, у якому найбільш точні виправлення помилок без зміни початкового значення речення отримані саме для цієї нейронної мережі. M2M100 має більшу оцінку BLEU (0,847), ніж "Ukrainian Roberta" Encoder-Decoder (0,697), однак, суб'єктивно оцінюючи результати виправлення прикладів, M2M100 значно гірше справляється із цим завданням, ніж дві інші моделі. Для METEOR також M2M100 (0,925) має більшу оцінку, ніж «Ukrainian Roberta» Encoder-Decoder (0,876).

З метою економії обчислювальних ресурсів можливим також є застосування попередньо навченої нейронної мережі типу BERT, використовуючи її як у якості енкодера, так і декодера. Така нейронна мережа має вдвічі менше параметрів, ніж інші попередньо навчені MT-моделі, і показує задовільні GEC-результати.

ПОДЯКИ

Роботу виконано в рамках держбюджетної теми «Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (ID:839 2017-05-15 09:20:01 (2459-315)). Дослідження провадилось в межах спільних наукових досліджень кафедри інформаційних систем та мереж НУ «Львівська політехніка» на тему «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, просторів даних та знань з метою прискорення процесів формування сучасного інформаційного суспільства». Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ІСМ НУ «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

ЛІТЕРАТУРА

1. Naghshnejad M. Recent Trends in the Use of Deep Learning Models for Grammar Error Handling / M. Naghshnejad, T. Joshi, V. N. Nair // ArXiv. – 2020. DOI: 10.48550/arXiv.2009.02358
2. Automated Grammatical Error Detection for Language Learners, Second Edition / [C. Leacock, M. Chodorow, M. Gamon, J. Tetreault] // Synthesis Lectures on Human Language Technologies. – 2014, Berlin : Springer. – 154 p. DOI: 10.1007/978-3-031-02153-4
3. Attention Is All You Need / [A. Vaswani, N. Shazeer, N. Parmar et al.] // ArXiv. – 2017. DOI: 10.48550/arXiv.1706.03762
4. Transformers: State-of-the-Art Natural Language Processing / [T. Wolf, L. Debut, V. Sanh et al.] // EMNLP 2020: Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, ALC Anthology, Oct. 2020 : proceedings. – P. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6
5. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / [J. Devlin, M.-W. Chang, K. Lee, K. Toutanova] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Minneapolis, Minnesota, ALC Anthology, June 2019 : proceedings. – P. 4171–4186. DOI: 10.18653/v1/N19-1423
6. Rozovskaya A. Grammar Error Correction in Morphologically Rich Languages: The Case of Russian / A. Rozovskaya, D. Roth // Transactions of the Association for Computational Linguistics. – 2019. – Vol. 7. – P. 1–17. DOI: 10.1162/tacl_a_00251
7. Bick E. DanProof: Pedagogical Spell and Grammar Checking for Danish / E. Bick // International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, Sep. 2015 : proceedings. – P. 55–62.
8. Design and construction of the Greek grammar checker / [P. Gakis, C. T. Panagiotakopoulos, K. N. Sgarbas et al.] //

DOI 10.15588/1607-3274-2023-1-12

Digital Scholarship in the Humanities. – 2017. – Vol. 32. – P. 554–576. DOI: 10.1093/llc/fqw025
9. Deksne D. A New Phase in the Development of a Grammar Checker for Latvian / D. Deksne // Human Language Technologies – The Baltic Perspective. – IOS Press, 2016. – P. 147–152. DOI: 10.3233/978-1-61499-701-6-147
10. Sorokin A. Spelling Correction for Morphologically Rich Language: a Case Study of Russian / A. Sorokin // 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain, Apr. 2017 : proceedings. – P. 45–53. DOI: 10.18653/v1/W17-1408.
11. Gill M. S. A Grammar Checking System for Punjabi / M. S. Gill, G. S. Lehal // Coling 2008: Companion volume: Demonstrations, Manchester, UK, Aug. 2008 : proceedings. – P. 149–152.
12. Go M. P. Developing an Unsupervised Grammar Checker for Filipino Using Hybrid N-grams as Grammar Rules / M. P. Go, A. Borra // PACLIC: 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers, Seoul, South Korea, Oct. 2016 : proceedings. – P. 105–113.
13. Shaalan K. F. Arabic GramCheck: a grammar checker for Arabic / K. F. Shaalan // Software: Practice and Experience. – 2005. – Vol. 35(7). – P. 643–665. DOI: 10.1002/spe.653
14. A Comprehensive Survey of Grammar Error Correction / [Y. Wang, Y. Wang, J. Liu, Z. Liu] // ArXiv. – 2020. DOI: 10.48550/arXiv.2005.06600
15. Syvokon O. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language / O. Syvokon, O. Nahorna // ArXiv. – 2021. DOI: 10.48550/arXiv.2103.16997
16. Lardinois F. Grammarly goes beyond grammar / F. Lardinois // TechCrunch. – 2019. – Access mode: https://techcrunch.com/2019/07/16/grammarly-goes-beyond-grammar/.
17. Lardinois F. Grammarly gets a tone detector to keep you out of email trouble / F. Lardinois // TechCrunch. – 2019. – Access mode: https://techcrunch.com/2019/09/24/grammarly-gets-a-tone-detector-to-keep-you-out-of-email-trouble/.
18. Grammarly Inc. About Us / Grammarly Inc. – Access mode: https://www.grammarly.com/about.
19. Grammarly Inc. Does Grammarly support languages other than English? / Grammarly Inc. – Access mode: https://support.grammarly.com/hc/en-us/articles/115000090971-Does-Grammarly-support-languages-other-than-English-.
20. LanguageTool. Languages / LanguageTool. – Access mode: https://dev.languagetool.org/languages.
21. LanguageTool. Error Rules for LanguageTool / LanguageTool Community. – Access mode: https://community.languagetool.org/rule/list?offset=0&max=10&lang=uk&filter=&categoryFilter=&_action_list=%D0%A4%D1%96%D0%BB%D1%8C%D1%82%D1%80.
22. LanguageTool. About / LanguageTool. – Access mode: https://languagetool.org/about.
23. Jayanthi S. NeuSpell: A Neural Spelling Correction Toolkit / S. Jayanthi, D. Pruthi, G. Neubig // ArXiv. – 2020. DOI: 10.48550/arXiv.2010.11085
24. Hunspell, Github. – 2021. – Access mode: https://github.com/hunspell/hunspell.
25. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages / M. Korobov // ArXiv. – 2015. DOI: 10.48550/arXiv.1503.07283
26. Tmienova N. System of Intellectual Ukrainian Language Processing / N. Tmienova, B. Sus // ITS: the XIX International Conference on Information Technologies and Security, Kyiv, Ukraine, Nov. 28, 2019 : proceedings. – P. 199–209.
27. Pogorilyy S. Method of noun phrase detection in Ukrainian texts / S. Pogorilyy, A. A. Kramov // ArXiv. – 2020. DOI: 10.48550/arXiv.2010.11548
28. Глибовець А. Алгоритм токенізації та стемінгу для текстів українською мовою / А. Глибовець, В. Точицький // Наукові записки НаУКМА. Комп'ютерні науки. – 2017. – Т. 198. – С. 4–8.
29. Hao S. A Research on Online Grammar Checker System Based on Neural Network Model / S. Hao, G. Hao // Journal of Physics. – 2020. – Vol. 1651. – P. 1–8. DOI: 10.1088/1742-6596/1651/1/012135
30. Батюк Т. М. Технологія соціалізації особистостей за спільними інтересами на основі методів машинного навчання та seo-технологій / Т. М. Батюк, В. А. Висоцька // Радіоелектроніка, інформатика, управління. – 2022. – № 2 (61). – С. 53–68. DOI: 10.15588/1607-3274-2022-2-6
31. The CoNLL-2014 Shared Task on Grammatical Error Correction / [H. T. Ng, S. M. Wu, T. Briscoe, et al.] // Conference on Computational Natural Language Learning: Shared Task, Baltimore, Maryland, Jun. 2014 : proceedings. – P. 1–14. DOI: 10.3115/v1/W14-1701
32. Chollampatt S. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction / S. Chollampatt, H. T. Ng // ArXiv. – 2018. DOI: 10.48550/arXiv.1801.08831
33. GECToR – Grammatical Error Correction: Tag, Not Rewrite / [K. Omelianchuk, V. Atrasevych, A. N. Chernodub, O. Skurzhanskyi] // ArXiv. – 2020. DOI: 10.48550/arXiv.2005.12592
34. The University of Illinois System in the CoNLL-2013 Shared Task / [A. Rozovskaya, K.-W. Chang, M. Sammons, D. Roth] // Conference on Computational Natural Language Learning: Shared Task, Sofia, Bulgaria, Aug. 2013 : proceedings. – P. 13–19.
35. The CoNLL-2013 Shared Task on Grammatical Error Correction /[H. T. Ng, S. M. Wu, Y. Wu et al.] // Conference on Computational Natural Language Learning: Shared Task, Sofia, Bulgaria, Aug. 2013 : proceedings. – P. 1–12.
36. A Simple Recipe for Multilingual Grammatical Error Correction / [S. Rothe, J. Mallinson, E. Malmi et al.] // ArXiv. – 2021. DOI: 10.48550/arXiv.2106.03830
37. Mita M. Do Grammatical Error Correction Models Realize Grammatical Generalization? / M. Mita, H. Yanaka // ArXiv. – 2021. DOI: 10.48550/arXiv.2106.03031
38. A Unified Strategy for Multilingual Grammatical Error Correction with Pre-trained Cross-Lingual Language Model / [X. Sun, T. Ge, S. Ma et al.] // ArXiv. – 2022. DOI: 10.48550/arXiv.2201.10707
39. Yasunaga M. LM-Critic: Language Models for Unsupervised Grammatical Error Correction / M. Yasunaga, J. Leskovec, and P. Liang // ArXiv. – 2021. DOI: 10.48550/arXiv.2109.06822
40. A Neural Grammatical Error Correction System Built on Better Pre-training and Sequential Transfer Learning / [Y. J. Choe, J. Ham, K. Park, Y. Yoon] // Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, Aug. 2019 : proceedings. – P. 213–227. DOI: 10.18653/v1/W19-4423

DOI 10.15588/1607-3274-2023-1-12

41. Wan Z. A Syntax-Guided Grammatical Error Correction Model with Dependency Tree Correction / Z. Wan, X. Wan // ArXiv. – 2021. DOI: 10.48550/arXiv.2111.03294
42. Neural Language Correction with Character-Based Attention / [Z. Xie, A. Avati, N. Arivazhagan et al.] // ArXiv. – 2016. DOI: 10.48550/arXiv.1603.09727
43. Parnow K. Grammatical Error Correction as GAN-like Sequence Labeling / K. Parnow, Z. Li, H. Zhao // ArXiv. – 2021. DOI: 10.48550/arXiv.2105.14209
44. Wang X. Research and Implementation of English Grammar Check and Error Correction Based on Deep Learning / X. Wang, W. Zhong // Scientific Programming. – 2022. – Vol. 2022. – Article ID 4082082. DOI: 10.1155/2022/4082082
45. Grammatical Error Correction in Low-Resource Scenarios / J. Náplava, M. Straka // W-NUT: 5th Workshop on Noisy User-generated Text, Hong Kong, China, Nov. 2019 : proceedings. – P. 346–356. DOI: 10.18653/v1/D19-5545.
46. Improving Grammatical Error Correction with Machine Translation Pairs / [W. Zhou, T. Ge, C. Mu et al.] // ArXiv. – 2020. DOI: 10.48550/arXiv.1911.02825
47. Raheja V. Adversarial Grammatical Error Correction / V. Raheja, D. Alikaniotis // EMNLP: Findings of the Association for Computational Linguistics, Online, Nov. 2020 : proceedings. – P. 3075–3087. DOI: 10.18653/v1/2020.findings-emnlp.275
48. Radchenko V. Ukrainian Roberta / V. Radchenko. Access mode: https://github.com/youscan/language-models.
49. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer / [L. Xue, N. Constant, A. Roberts et al.] // Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, Jun. 2021 : proceedings. – P. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.
50. Beyond English-Centric Multilingual Machine Translation / [A. Fan, S. Bhosale, H. Schwenk et al.]// ArXiv. – 2020. DOI: 10.48550/arXiv.2010.11125
51. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning / [Y. Tang, C. Tran, Xian Li et al.] // ArXiv. – 2020. DOI: 10.48550/arXiv.2008.00401
52. Platen von Patrick Leveraging Pre-trained Language Model Checkpoints for Encoder-Decoder Models / Platen von Patrick. – Access: https://huggingface.co/blog/warm-starting-encoder-decoder.
53. Rothe S. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks / S. Rothe, S. Narayan, A. Severyn // ArXiv. – 2019. DOI: 10.48550/arXiv.1907.12461
54. Ground truth for grammatical error correction metrics / [C. Napoles, K. Sakaguchi, M. Post, J. Tetreault] // 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 2015 : proceedings. – P. 588–593. DOI: 10.3115/v1/P15-2097
55. Bleu: a Method for Automatic Evaluation of Machine Translation / [K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu] // 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 2002 : proceedings. – P. 311–318. DOI: 10.3115/1073083.1073135.
56. Banerjee S. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments / S. Banerjee, A. Lavie // ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Michigan, Jun. 2005 : proceedings. – P. 65–72.
57. Platen von Patrick Encoder-Decoder models don't need costly pre-training to yield state-of-the-art results on seq2seq tasks / Platen von Patrick. – Access mode: https://twitter.com/patrickplaten/status/1325844244095971328.

Accepted 19.09.2022.
Received 03.12.2022.

UDC 004.9

TECHNOLOGY FOR GRAMMATICAL ERRORS CORRECTION IN UKRAINIAN TEXT CONTENT BASED ON MACHINE LEARNING METHODS

Kholodna N. – PhD student of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

Vysotska V. – PhD, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. Most research in grammatical and stylistic error correction focuses on error correction in English-language textual content. Thanks to the availability of large data sets, a significant increase in the accuracy of English grammar correction has been achieved. Unfortunately, there are few studies on other languages. Systems for the English language are constantly developing and currently actively use machine learning methods: classification (sequence tagging) and machine translation. A large amount of parallel or manually labelled data is required to build a high-quality machine learning model for correcting grammatical/stylistic errors in the texts of those morphologically complex languages. Manual data annotation requires a lot of effort by professional linguists, which makes the creation of text corpora, especially in morphologically rich languages, mainly Ukrainian, a time- and resource-consuming process.

Objective of the study is to develop a technology for correcting errors in Ukrainian-language texts based on machine learning methods using a small set of annotated parallel data.

Method. For this study, machine learning algorithms were selected when developing a system for correcting errors in Ukrainian-language texts using an optimal pipeline, including pre-processing and selecting text content and generating features in small annotated data corpora. The neural network's use with a new architecture, a review of state-of-the-art methods, and a comparison of different pipeline stages will make it possible to determine such a combination of them, allowing a high-quality error correction model in Ukrainian-language texts.

Results. A machine learning model for error correction in Ukrainian-language texts has been developed. A universal scheme for creating an error correction system for different languages is proposed. According to the results, the neural network can correct simple sentences written in Ukrainian. However, creating a full-fledged system will require spell-checking using dictionaries and checking rules, both simple and based on the result of parsing dependencies or other features. The pre-trained neural translation

DOI 10.15588/1607-3274-2023-1-12

model mT5 has the best performance among the three models. To save computing resources, it is also possible to use a pre-trained BERT-type neural network as an encoder and a decoder. Such a neural network has half the number of parameters as other pre-trained machine translation models and shows satisfactory results in correcting grammatical and stylistic errors.

Conclusions. The created model shows excellent classification results on test data. The calculated machine translation quality metrics allow only a partial comparison of the models since most of the words and phrases in the original and corrected sentences are the same. The best value for both BLEU (0.908) and METEOR (0.956) is obtained for mT5, which is consistent with the case study in which the most accurate error corrections without changing the initial value of the sentence are obtained for such a neural network. The M2M100 has a higher BLEU score (0.847) than the "Ukrainian Roberta" Encoder-Decoder (0.697). However, subjectively evaluating the results of the correction of examples, the M2M100 does a much worse job than the other two models. For METEOR, M2M100 (0.925) also has a higher score than the "Ukrainian Roberta" Encoder-Decoder (0.876).

KEYWORDS: NLP, text pre-processing, error correction, grammatical error correction, machine learning, deep learning, text analysis, text classification, neural network.

REFERENCES

1. Naghshnejad M., Joshi T., Nair V. N. Recent Trends in the Use of Deep Learning Models for Grammar Error Handling, *ArXiv*, 2020. DOI: 10.48550/arXiv.2009.02358
2. Leacock C., Chodorow M., Gamon M., Tetreault J. Automated Grammatical Error Detection for Language Learners, Second Edition, *Synthesis Lectures on Human Language Technologies*, 2014. Berlin, Springer, 154 p. DOI: 10.1007/978-3-031-02153-4
3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need, *ArXiv*, 2017. DOI: 10.48550/arXiv.1706.03762
4. Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., Platen P. v., Ma C., Jernite Y., Plu J., Xu C., Scao T. L., Gugger S., Drame M., Lhoest Q., Rush A. Transformers: State-of-the-Art Natural Language Processing, *EMNLP 2020: Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, ALC Anthology, Oct. 2020 : proceedings*, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6
5. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Minneapolis.* Minnesota, ALC Anthology, June 2019 : proceedings, pp. 4171–4186. DOI: 10.18653/v1/N19-1423
6. Rozovskaya A., Roth D. Grammar Error Correction in Morphologically Rich Languages: The Case of Russian, *Transactions of the Association for Computational Linguistics*, 2019, Vol. 7, pp. 1–17. DOI: 10.1162/tacl_a_00251
7. Bick E. DanProof: Pedagogical Spell and Grammar Checking for Danish, *International Conference Recent Advances in Natural Language Processing.* Hissar, Bulgaria, Sep. 2015, proceedings, pp. 55–62.
8. Gakis P., Panagiotakopoulos C. T., Sgarbas K. N., Tsalidis C., Verykios V. S., Design and construction of the Greek grammar checker, *Digital Scholarship in the Humanities*, 2017, Vol. 32, pp. 554–576. DOI: 10.1093/llc/fqw025
9. Deksne D. A New Phase in the Development of a Grammar Checker for Latvian. Human Language Technologies, The Baltic Perspective, IOS Press, 2016, pp. 147–152. DOI: 10.3233/978-1-61499-701-6-147
10. Sorokin A. Spelling Correction for Morphologically Rich Language: a Case Study of Russian, *6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain, Apr. 2017, proceedings*, pp. 45–53. DOI: 10.18653/v1/W17-1408.
11. Gill M. S., Lehal G. S. A Grammar Checking System for Punjabi, *Coling 2008: Companion volume: Demonstrations, Manchester, UK, Aug. 2008 : proceedings*, pp. 149–152.
12. Go M. P., Borra A. Developing an Unsupervised Grammar Checker for Filipino Using Hybrid N-grams as Grammar Rules, *PACLIC: 30th Pacific Asia Conference on Language, Information and Computation, Oral Papers*. Seoul, South Korea, Oct. 2016 : proceedings, pp. 105–113.
13. Shaalan K. F. Arabic GramCheck: a grammar checker for Arabic, *Software: Practice and Experience*, 2005, Vol. 35(7), pp. 643–665. DOI: 10.1002/spe.653
14. Wang Y., Wang Y., Liu J., Liu Z. A Comprehensive Survey of Grammar Error Correction, *ArXiv*, 2020. DOI: 10.48550/arXiv.2005.06600
15. Syvokon O., Nahorna O. UA-GEC: Grammatical Error Correction and Fluency Corpus for the Ukrainian Language *ArXiv*, 2021. DOI: 10.48550/arXiv.2103.16997
16. Lardinois F. Grammarly goes beyond grammar, *TechCrunch*, 2019. Access mode: https://techcrunch.com/2019/07/16/grammarly-goes-beyond-grammar/.
17. Lardinois F. Grammarly gets a tone detector to keep you out of email trouble, *TechCrunch*, 2019. Access mode: https://techcrunch.com/2019/09/24/grammarly-gets-a-tone-detector-to-keep-you-out-of-email-trouble/.
18. Grammarly Inc. About Us / Grammarly Inc. Access mode: https://www.grammarly.com/about.
19. Grammarly Inc. Does Grammarly support languages other than English? / Grammarly Inc. Access mode: https://support.grammarly.com/hc/en-us/articles/115000090971-Does-Grammarly-support-languages-other-than-English-.
20. LanguageTool. Languages, LanguageTool. Access mode: https://dev.languagetool.org/languages.
21. LanguageTool. Error Rules for LanguageTool / LanguageTool Community. Access mode: https://community.languagetool.org/rule/list?offset=0&max=10&lang=uk&filter=&categoryFilter=&_action_list=%D0%A4%D1%96%D0%BB%D1%8C%D1%82%D1%80.
22. LanguageTool. About / LanguageTool. Access mode: https://languagetool.org/about.
23. Jayanthi S., Pruthi D., Neubig G. NeuSpell: A Neural Spelling Correction Toolkit, *ArXiv*, 2020. DOI: 10.48550/arXiv.2010.11085
24. Hunspell, Github, 2021, Access mode: https://github.com/hunspell/hunspell.
25. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages, *ArXiv*, 2015. DOI: 10.48550/arXiv.1503.07283
26. Tmienova N., Sus B. System of Intellectual Ukrainian Language Processing, *ITS: the XIX International Conference on Information Technologies and Security*, Kyiv, Ukraine, Nov. 28, 2019 : proceedings, pp. 199–209.

DOI 10.15588/1607-3274-2023-1-12

27. Pogorilyy S., Kramov A. A. Method of noun phrase detection in Ukrainian texts, *ArXiv*, 2020. DOI: 10.48550/arXiv.2010.11548
28. Hlybovets A., Tochytskyi V. Algorithm of tokenization and stemming for texts in the Ukrainian language, *Scientific notes of NaUKMA. Computer Science*, 2017, Vol. 198, pp. 4–8.
29. Hao S., Hao G. A Research on Online Grammar Checker System Based on Neural Network Model, *Journal of Physics*, 2020, Vol. 1651, pp. 1–8. DOI: 10.1088/1742-6596/1651/1/012135
30. Batiuk T. M., Vysotska V. Technology for Personalities Socialization by Common Interests Based on Machine Learning Methods And SEO-Technologies, *Radio Electronics, Computer Science, Control*, 2022, Vol. 2 (61), pp. 53–68. DOI: 10.15588/1607-3274-2022-2-6
31. Ng H. T., Wu S. M., Briscoe T., Hadiwinoto C., Susanto R. H., Bryant C. The CoNLL-2014 Shared Task on Grammatical Error Correction, *Conference on Computational Natural Language Learning: Shared Task.* Baltimore, Maryland, Jun. 2014, proceedings, pp. 1–14. DOI: 10.3115/v1/W14-1701
32. Chollampatt S., Ng H. T. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction, *ArXiv*, 2018. DOI: 10.48550/arXiv.1801.08831
33. Omelianchuk K., Atrasevych V., Chernodub A. N., Skurzhanskyi O. GECToR – Grammatical Error Correction: Tag, Not Rewrite, *ArXiv*, 2020. DOI: 10.48550/arXiv.2005.12592
34. Rozovskaya A., Chang K.-W., Sammons M., Roth D. The University of Illinois System in the CoNLL-2013 Shared Task, *Conference on Computational Natural Language Learning: Shared Task*. Sofia, Bulgaria, Aug. 2013, proceedings, pp. 13–19.
35. Ng H. T., Wu S. M., Wu Y., Hadiwinoto C., Tetreault J. The CoNLL-2013 Shared Task on Grammatical Error Correction, *Conference on Computational Natural Language Learning: Shared Task*. Sofia, Bulgaria, Aug. 2013, proceedings, pp. 1–12.
36. Rothe S., Mallinson J., Malmi E., Krause S., Severyn A. A Simple Recipe for Multilingual Grammatical Error Correction, *ArXiv*, 2021. DOI: 10.48550/arXiv.2106.03830
37. Mita M., Yanaka H. Do Grammatical Error Correction Models Realize Grammatical Generalization?, *ArXiv*, 2021. DOI: 10.48550/arXiv.2106.03031
38. Sun X., Ge T., Ma S., Li J., Wei F., and Wang H. A Unified Strategy for Multilingual Grammatical Error Correction with Pre-trained Cross-Lingual Language Model, *ArXiv*, 2022. DOI: 10.48550/arXiv.2201.10707
39. Yasunaga M., Leskovec J., and Liang P. LM-Critic: Language Models for Unsupervised Grammatical Error Correction, *ArXiv*, 2021. DOI: 10.48550/arXiv.2109.06822
40. Choe Y. J., Ham J., Park K., Yoon Y. A Neural Grammatical Error Correction System Built on Better Pre-training and Sequential Transfer Learning, *Workshop on Innovative Use of NLP for Building Educational Applications*, Florence. Italy, Aug. 2019 : proceedings, pp. 213–227. DOI: 10.18653/v1/W19-4423
41. Wan Z., Wan X. A Syntax-Guided Grammatical Error Correction Model with Dependency Tree Correction, *ArXiv*, 2021. DOI: 10.48550/arXiv.2111.03294
42. Xie Z., Avati A., Arivazhagan N., Jurafsky D., Ng A. Neural Language Correction with Character-Based Attention, ArXiv, 2016. DOI: 10.48550/arXiv.1603.09727
43. Parnow K., Li Z., Zhao H. Grammatical Error Correction as GAN-like Sequence Labeling, *ArXiv*, 2021. DOI: 10.48550/arXiv.2105.14209
44. Wang X., Zhong W. Research and Implementation of English Grammar Check and Error Correction Based on Deep Learning, *Scientific Programming*, 2022, Vol. 2022, Article ID 4082082. DOI: 10.1155/2022/4082082
45. Náplava J., Straka M. Grammatical Error Correction in Low-Resource Scenarios, *W-NUT: 5th Workshop on Noisy User-generated Text.* Hong Kong, China, Nov. 2019 : proceedings, pp. 346–356. DOI: 10.18653/v1/D19-5545.
46. Zhou W., Ge T., Mu C., Xu K., Wei F., Zhou M. Improving Grammatical Error Correction with Machine Translation Pairs, *ArXiv*, 2020. DOI: 10.48550/arXiv.1911.02825
47. Raheja V., Alikaniotis D. Adversarial Grammatical Error Correction, *EMNLP: Findings of the Association for Computational Linguistics, Online, Nov. 2020 : proceedings*, pp. 3075–3087. DOI: 10.18653/v1/2020.findings-emnlp.275
48. Radchenko V. Ukrainian Roberta. Access mode: https://github.com/youscan/language-models.
49. Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A., Raffel C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, Jun. 2021 : proceedings, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.
50. Fan A., Bhosale S., Schwenk H., Ma Z., El-Kishky A., Goyal S., Baines M., Celebi O., Wenzek G., Chaudhary V., Goyal N., Birch T., Liptchinsky V., Edunov S., Grave E., Auli M., Joulin A. Beyond English-Centric Multilingual Machine Translation, *ArXiv*, 2020. DOI: 10.48550/arXiv.2010.11125
51. Tang Y., Tran C., Li Xian, Chen P.-J., Goyal N., Chaudhary V., Gu J., Fan A. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, *ArXiv*, 2020. DOI: 10.48550/arXiv.2008.00401
52. Platen von Patrick Leveraging Pre-trained Language Model Checkpoints for Encoder-Decoder Models. Access: https://huggingface.co/blog/warm-starting-encoder-decoder.
53. Rothe S., Narayan S., Severyn A. Leveraging Pre-trained Checkpoints for Sequence Generation Tasks, *ArXiv*, 2019. DOI: 10.48550/arXiv.1907.12461
54. Napoles C., Sakaguchi K., Post M., Tetreault J. Ground truth for grammatical error correction metrics, *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.* Beijing, China, July 2015 : proceedings, pp. 588–593. DOI: 10.3115/v1/P15-2097
55. Papineni K., Roukos S., Ward T., and Zhu W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation, *40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 2002 : proceedings*, pp. 311–318. DOI: 10.3115/1073083.1073135.
56. Banerjee S., Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.* Michigan, Jun. 2005 : proceedings, pp. 65–72.
57. Platen von Patrick Encoder-Decoder models don't need costly pre-training to yield state-of-the-art results on seq2seq tasks. Access mode: https://twitter.com/patrickplaten/status/1325844244095971328.

DOI 10.15588/1607-3274-2023-1-12

УПРАВЛІННЯ У ТЕХНІЧНИХ СИСТЕМАХ

CONTROL IN TECHNICAL SYSTEMS

UDK 519.853: 658.52

METHOD OF ROUTING A GROUP OF MOBILE ROBOTS IN A FIXED NETWORK FOR SEARCHING THE MISSING OBJECTS IN A TECHNOLOGICAL DISASTER ZONE

Batsamut V. M. – Dr. Sc., Professor, The Deputy Head of the Scientific Research Center of the National Academy of the National Guard of Ukraine, Kharkiv, Ukraine.

Hodlevsky S. O. – The Researcher of the Scientific Research Center of the National Academy of the National Guard of Ukraine, Kharkiv, Ukraine.

ABSTRACT

Context. The relevance of the article is determined by the need for further development of models of collective behavior of systems with multi-agent structure construction endowed with intelligence that ensures synchronization of the joint efforts of various agents while achieving the goals set for the system. The method proposed in the article solves the problem of competition between different agents of a multi-agent system, which is important while performing search, rescue, and monitoring tasks in crisis areas of various origins.

Objective is to develop a method for determining the sufficient population of a multi-agent system and the optimal routes of movement of its individual elements in a stationary network for the most complete examination of a technological disaster zone (any given zone based on a certain transport network).

Method. We implemented the concept of a dynamic programming to search for all possible edge-simple longest paths connecting the directed subsets of vertices-sources and vertices-sinks in the structure of the model weighted directed graph. To this end, the modified Dijkstra method was applied. The modification comprises representing the weights of the arcs of the modeling directed graph with the negative values, which are further used in calculations according to the Dijkstra method. After finding the next edge-simple longest path, the arcs that make up it are fixed in the memory of the computer system (in the route plan) and removed from the graph structure, and the process is iteratively repeated. The search for paths takes place as long as the transitive closure between the vertices that are part of the specified subsets of source vertices and sink vertices is preserved. The developed method makes it possible to find such a set of traffic routes for the elements of the multi-agent system, which maximizes the area examined by them in a technological disaster zone (or the number of checked objects on the traffic routes) in one “wave” of the search and distributes the elements of a multi-agent system by routes that do not have common areas. A derivative of the application of the developed method is the determination of a sufficient population of a multi-agent system for effective search activities within the defined zone.

Results. 1) A method of routing a group of mobile robots in a stationary network for searching the missing objects in a technological disaster zone has been developed. 2) The working expression of the Dijkstra method for searching in the structure of a network object (in the structure of a model graph) for the longest paths has been formalized. 3) We have suggested a set of indicators for a comprehensive evaluation of route plans of a multi-agent system. 4) The method has been verified on test problems.

Conclusions. Theoretical studies and several experiments confirm the efficiency of the developed method. The solutions made using the developed method are accurate, which allows recommending it for practical use in determining in an automated mode route plans for multi-agent systems, as well as the required number of agents in such systems to perform the required amount of search tasks in a particular crisis area.

KEYWORDS: multi-agent system, group of mobile robots, routing, network object, weighted undirected (directed) graph, extreme paths, optimization criterion, method.

ABBREVIATIONS

MAS is a multi-agent system;

TDZ is a technological disaster zone.

NOMENCLATURE

G is a undirected weighted graph simulating the transport network in a technological disaster zone;

$\vec{G}$ is a directed weighted graph simulating the transport network in a technological disaster zone;

P is a set of graph vertices simulating turning points (intersections) of the transport network within a technological disaster zone;

A is a subset of vertices-sources such that $A \subset P$;

B is a subset of vertices-sinks such that $B \subset P$;

E is a set of edges (arcs) of the graph modeling paths (communications) inside a technological disaster zone;

$w(e_{ij})$ is a weighting coefficient of some edge (arc) e_{ij};

Direct is a general search direction;

DOI 10.15588/1607-3274-2023-1-13

k_{wvs} is a number of searching waves;

M_i is the longest edge-simple path between some vertices a_x and b_y from sets A and B respectively;

$\vec{G}'$, $\vec{G}''$,…, $\vec{G}^n$ is a sequence of substructures arising because of splitting the initial graph $\vec{G}$;

n is the number of found longest edge-simple paths between subsets A and B respectively;

L_i is the length of *i*-th edge-simple longest path;

i is the number of the edge-simple longest paths;

L_{total} is the total length of the transport network in the TDZ;

L_{PL} is a total length (weight) of the defined routes;

T_s is a total time of conducting search activities;

$\bar{v}_{agt}$ is the average speed of the agents moving along the defined routes;

T_{dir} is an established (directive) time for performing search activities;

K_{MAS} is the required number of MAS agents;

P_{dso} is a probability of detecting search objects according to a certain route plan of the MAS;

$\xrightarrow{T}$ is a transitive closure between the selected pair of vertices of the model graph $\vec{G}$.

INTRODUCTION

In the event of a large-scale technological disaster of a certain origin in any region of the country, which could be accompanied by the release of various types of poisonous or ionizing substances, a task of searching the missing objects and/or monitoring the operational (chemical, radiation) situation in a technological disaster zone (TDZ) may arise. In this case, the bodies for obtaining data on operational situation in such a zone can be both chemical or radiation reconnaissance groups, staffed with trained personnel and appropriate special equipment, protective gear and devices, and the groups of mobile robots equipped with appropriate devices for monitoring, measuring, and recording the values of individual parameters, as well as mixed groups.

The set of such bodies constitutes a multi-agent system (MAS), which in order to perform practical missions (achieving the goals) must have a common logic of behavior, as well as the logic of conduct of its individual agents [1, 2]. It should be noted here that the struggle to preserve the life and health of rescue and search units' personnel has recently outlined a global trend towards the use of exclusively robotic MAS in dangerous zones [3, 4]. A human in such a system endows it with its intelligence, performs control, logistical support of the system and is a consumer of the results of its functioning. The materials of this article are aimed at ensuring the functioning of robotic MAS.

Endowing the MAS with a certain behavior (intelligence) in matters of rational or optimal movement of its elements through the network structure while performing search (or monitoring) task, requires the development of certain routing methods.

In general, the problem of routing in any network object is to find some extreme paths (shortest or k-shortest, Hamiltonian) in its structure [5, 6, 7], less often trees [8] or minimal covering trees [9], etc. In this matter, the nature of the applied problem to be solved is decisive.

While search activities using a robotic MAS, the application of the above approaches to the routing of its agents is unacceptable, since:

1) Various conditions in a certain way determine the available points of input of MAS elements in the search and the number of such points, which in turn determines the general direction of the search within TDZ.

2) In order to achieve the search goals, it is necessary to check as many areas as possible (in the best case, all areas, although it is practically impossible to fulfill this requirement in one "pass" of the MAS in dense structures).

3) It is necessary that the same areas of the terrain are not revisited by different agents (implementation of the "checked and crossed out from the list" principle), which will increase the chances of the positive result of a search operation.

4) It is desirable to withstand the general direction of advancement of the entire MAS within the network object (a technological disaster zone).

The object of the study is to determine the set of optimal movement routes for the elements of a multi-agent system within the network object.

The subject of the study is the method of optimal routing of a group (swarm) of mobile robots within the structure of a fixed network for conducting search activities in a limited area.

The purpose of the study is to develop a method for determining the sufficient population of a multi-agent system and the optimal movement routes of its individual agents in a stationary network for the most complete examination of a technological disaster zone (any given zone based on a certain transport network).

1 PROBLEM STATEMENT

Routing problems are, for the most part, formalized and solved using graph theory models and methods [10]. That is why we will model the transport network in TDZ with some weighted undirected graph $G=(P,E)$, where $P=\{p_1,...,p_v\}$ – the set containing the vertices of the graph; $E=\{e_1,...,e_m\}$ – the set containing the edges of the graph; v – the number of graph vertices; m – the number of edges of the graph. At the same time, the set P is simulating the intersections of the transport network (points at which the agent decides about the direction of its further movement), and the set E – certain sections of paths connecting different intersections in the transport network of the TDZ. Each edge from the set E will have a certain weight coefficient $w(e_{ij})$, which will quantita-

DOI 10.15588/1607-3274-2023-1-13

tively characterize the edge e_{ij}. It could be a certain number of objects to be inspected located on this site; the length of the route measured in certain units; the amount of work to be done on the site, etc.

Considering the peculiarities of carrying out search activities in a certain area, the formulation of the task of determining traffic routes for the MAS will look as follows.

Assume that the undirected weighted graph $G=(P,E)$ has selected subset of vertices $A=\{a_x\}_{x=\overline{1,k}}$ (vertices-sources) and a subset of vertices $B=\{b_y\}_{y=\overline{1,z}}$ (vertices-sinks), where k and z are the numbers of such vertices, respectively, and $B\notin\varnothing$, $|A|<<|P|$, $|B|<<|P|$, $A\cap B=\varnothing$. Weighting coefficients $w(e_{ij})$ on the edges of the graph $G=(P,E)$ quantitatively characterize the corresponding area of the TDZ.

It is necessary to find all possible edge-simple longest paths $M_{i=\overline{1,n}}=(a_x, p_1, p_2,..., p_q, b_y)$, connecting vertices from subsets A and B, and the totality of which satisfies the following target function:

$$F=\sum_{i=1}^{n}L_i\rightarrow\max, \tag{1}$$

under the conditions:

$$T_s\le T_{dir}, \tag{2}$$

$$n=f\left[G(P,E),|A|,|B|,Direct\right], \tag{3}$$

$$e_{ij}\in M_1\cap e_{ij}\in M_2\cap...\cap e_{ij}\in M_n=\varnothing, \tag{4}$$

$$k_{wvs}=1. \tag{5}$$

Therefore, the solution of the problem (1)–(5) requires searching within the structure of the initial weighted graph $G=(P,E)$ of all extremal (edge-simple longest) paths between defined subsets of its vertices.

An edge-simple path is a path in which each edge (arc) of the graph occurs only once [10].

2 REVIEW OF THE LITERATURE

At present, the theoretical basis for the functioning of robotic vehicles is the theory of algorithms. The development of the theory of algorithms takes place in two directions: first, the expansion of the range of practical problems solved by the existing algorithms; secondly, the development and improvement of algorithms for solving new problems that arise during the creation and functioning of the MAS. Well-developed graph theory plays a leading role in routing of individual MAS agents.

A well-known and studied problem of graph theory having numerous practical applications is the problem of Hamiltonian paths, that is, whether there is a simple path in a graph in which each vertex of the graph occurs exactly once. Such a graph is called Hamiltonian. In the case when the graph does not contain a Hamiltonian path, in some applications it makes sense to search for a path of a maximum length in the graph. Finding such a path is known as the longest path problem. Like finding the Hamiltonian path, finding the longest path is also a difficult task.

The longest path problem is NP-complete on every class of graphs in which the Hamiltonian path problem is also NP-complete. Thus, in [11] the authors prove that even if a graph has a Hamiltonian path, the problem of finding a path of the length $n-n^{\varepsilon}$ for some $\varepsilon<1$ is a NP-complete, where n – is the number of vertices of the initial graph. The authors claim that there is no polynomial approximation algorithm with a constant factor for the longest path problem unless $P=NP$ [11]. Similar research results are also given in works [12, 13, 14, 15, 16].

It should be noted here that the Hamiltonian path problem is NP-complete on general graphs [17, 18] and remains NP-complete even when restricted to some small classes of graphs, such as splitting graphs [19], chordal bipartite graphs, strongly chordal graphs [20], directed path graphs [21], circular graphs [22], planar graphs [18] and grid graphs [23].

Polynomial solutions of this problem are known only for certain classes of graphs. Such algorithms were developed for graphs of intervals and presented in studies [24, 25, 26, 27], for doubly convex graphs – in a study [28], for graphs of arcs of circles – in a study [27] and graphs of comparability – in a study [29].

Unlike the Hamiltonian path problem, several polynomial complexity algorithms are known for the longest path problem that work with the structures of tree-type and some classes of graphs. A linear algorithm for finding the longest path on a tree-structure was proposed by Dijkstra in 1960, a formal description of which can be found in [30]. Later, based on the results of improving the Dijkstra algorithm for trees, the authors of [31] have solved the problem of finding the longest path for weighted trees and block graphs with a linear calculation time, and for cactuses with a polynomial calculation time – $O(n^2)$, where n – the number of vertices of the initial graph. Recently, polynomial algorithms were proposed that solve the problem of finding the longest path on bipartite graphs with computational complexity $O(n)$ [32], on Ptolemaic graphs with computational complexity $O(n^5)$ [33]. In [34], the authors presented their polynomial algorithm for interval graphs, which is based on the idea of dynamic programming and has a computational complexity of $O(n^4)$. In [35], the authors proposed a polynomial algorithm in which they also used the dynamic programming approach but applied lexicographic search in depth (the so-called LDFS graph routing) for co-comparable graphs. The computational complexity of such an algorithm is also limited to $O(n^4)$.

The analysis of the literature shows that the problem in the formal statement (1)–(5) was not posed or solved

DOI 10.15588/1607-3274-2023-1-13

by anyone. If the approach is not strict, then the task is to find in the structure of the model weighted graph $G=(P,E)$ every (or k, where $k>0$) of the longest edge-simple paths between the selected subsets of the vertices of the graph, which lie on opposite edges of it and the total sum of the weights of which (paths) maximizes the target function F, expression (1). At the same time, an additional and mandatory condition for the sought route plan is the requirement that there are no common edges in different paths (routes). Our article is dedicated to solving this problem.

3 MATERIALS AND METHODS

It is quite clear that the solutions of the problem (1)–(5) formulated above require not only estimates of the lengths of the critical paths, but also the critical paths themselves, that is a consecutive set of edges that make them (the paths). It should be noted that such identification capabilities, in contrast to the well-known in graph theory algorithms Floyd-Warshall [36, 37, 38], Shimbel [39], Danzig [5] and some others, which solve the problem of finding the lengths of extreme paths, the Dijkstra's algorithm provides.

In addition, to solve the problem (1)–(5), it is necessary to use the modified Dijkstra algorithm, since it (the problem) in this formulation belongs to the class of NP-complete and, therefore, it cannot be solved in polynomial time [41].

Modification of the initial structure. The problem of NP-completeness of the task of finding the longest paths (paths of the greatest total weight) in the structure of a network object is associated with the possible presence of cycles, which will lead to an unjustified increase in the total weight of the searched path (so-called "looping"), and therefore to the inability to adequately identify the path itself in the future. To solve such a "problem" when determining maximum (search for the longest paths), usually the initial undirected graph $G=(P,E)$ is represented in the form of an oriented graph $\vec{G}=\left(P,\vec{E}\right)$ without cycles. For this, the edges of the graph are directed along the general search direction, which is determined based on the nature of the practical problem to be solved (see Fig. 1). As a result of this transformation, the edges of the graph become arcs, that is, they get directionality.

Since a group of mobile robots during search activities moves from one side of the TDZ to the opposite (which can be imagined in the form of a "wave" that passes through this zone), then such a general direction also exists.

As can be seen from Fig. 1, there are no cycles in the structure of the oriented graph, which makes it possible to apply the modified Dijkstra algorithm to it to find the longest paths (paths of the greatest total weight).

Modification of initial data. The next problematic point that needs to be solved is the representation of the weight coefficients of the edges when solving the problem of maximizing the lengths of the routes by which the elements of the MAS advance (search for the longest paths). This problem appears because in Dijkstra's algorithm, at each iteration, the value of the accumulated total weight for some vertex is compared with the estimate of the total weight that this vertex can receive through another edge (arc) and the smallest of these values is chosen. The working expression of Dijkstra's algorithm in its classical form is as follows [5]:

$$d(x):=\min\{d(x),\ d(y)+w(y,x)\}. \qquad (6)$$

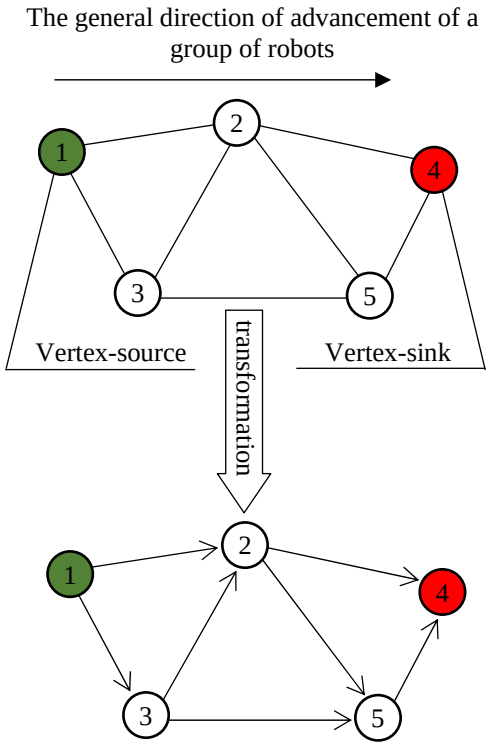


Figure 1 – Transformation of an unoriented graph $G=(P,E)$ into oriented $\vec{G}=\left(P,\vec{E}\right)$

Since the problem (1)–(5) is solved for the maximum of the function, the application of the classic Dijkstra algorithm according to the expression (6) will lead to incorrect results – the search for the shortest paths. To solve such a problem, we suggest representing the weighting coefficients of the arcs of the graph $\vec{G}=\left(P,\vec{E}\right)$ as numbers from the negative domain and applying the following modification of the working expression of the algorithm, namely:

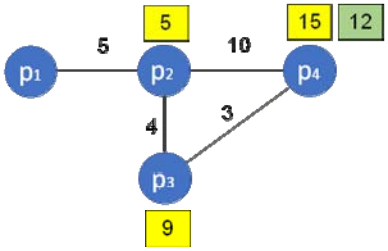
$$-d(x):=\min\{-d(x),\ -d(y)+(-w(y,x))\}. \qquad (7)$$

As a result of applying expression (7), the current negative score will be compared with an alternative also negative score, and if the alternative score is lower, then

DOI 10.15588/1607-3274-2023-1-13

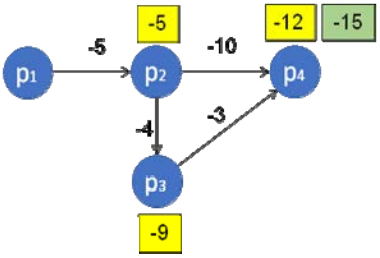
the current vertex will be assigned exactly that score. In other words, if in the structure of a weighted directed graph it is possible to increase the length of the current path due to the addition of a certain arc with a negative weight, then such an arc will be added, which will ensure that the longest path is found.

It should be noted here that estimates of path lengths will be presented in negative form, which will in no way interfere with their subsequent identification. In the future, the obtained estimates will be taken by the module, which will return them to their original physical meaning, Fig. 2.



$M_1=\{p_1, p_2, p_3, p_4\}=12$

a



$M_2=\{p_1, p_2, p_4\}=15$

b

Figure 2 – The result of applying classical and modified Dijkstra methods to find the extreme paths: a – finding the shortest path M_1 (classical method); b – finding the longest path M_2 (modified method)

As can be seen from Fig. 2, *b* that between the vertices p_1 and p_4 there are two alternative ways: $\{p_1, p_2, p_3, p_4\}$ and $\{p_1, p_2, p_4\}$. At the vertex p_4 you need to decide which of these paths is considered the longest. Two arcs are incidental to it: (p_3, p_4) and (p_2, p_4). Since the vertex p_3 has a label of –9 (the total weight of the path connecting this vertex to the source vertex p_1), the vertex p_4 through the arc (p_3, p_4) will receive a label of –12. Through an incident arc (p_2, p_4) having weight of –10, the vertex p_4 will receive a label of –15, because the vertex p_2 already has a label of –5. Therefore, the length of the path $\{p_1, p_2, p_3, p_4\}$ = –12, and the path $\{p_1, p_2, p_4\}$ = –15. Based on expression (7), the second of them is unambiguously chosen as the longest path. In the future, taking the received estimate of the length of the path by the module, we consider 15 units of the conditional length of the path.

On more branched structures, the process continues until all arcs are analyzed, and all vertices (including vertices-sinks) receive the corresponding conditional weight labels according to the expression (7). The identification of the extreme paths takes place on the reverse course, just as in the classic Dijkstra method.

Application of dynamic programming. The characteristics of edge-simple extreme paths are of particular importance in the context of search activities in TDZ, since it is undesirable for different agents of the MAS to move along paths that have common sections (arcs), even if the number of such sections is relatively small. In this sense, the movement of different MAS agents needs to be "separated" and redirected by different ways to increase the overall effectiveness of search activities (Fig. 3). The only place where the routes of different agents can intersect is the intersection (vertices in the structure of the initial model graph), which in no way hinders their simultaneous progress along the assigned routes and, in this part, does not reduce the overall effectiveness of search activities.

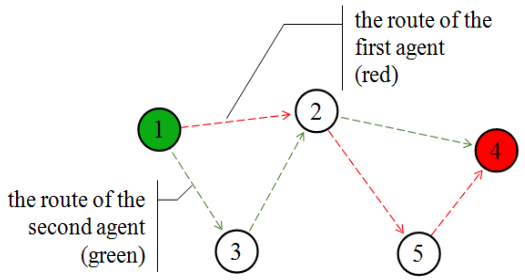


Figure 3 – Separated traffic routes of various MAS agents (option)

When solving the problem (1)–(5), to obtain a set of routes that do not have common sections (according to the additional condition (4) of the given problem) we propose to use the method of dynamic programming, which is to divide some general problem $\hat{Z}$ on a number of subproblems $Z', Z'',..., Z^n$, where n – the number of such subproblems, and to find such their corresponding solutions $R', R'',..., R^n$, that $R' \cup R'' \cup,..., \cup R^n$ will be the solution to a general problem $\hat{z}$. At the same time, it is considered if separate solutions $R', R'',..., R^n$, are rational (or optimal), then the solutions to the general problem $\hat{z}$ are also rational (or optimal).

Using dynamic programming approaches and solving the problem of finding all edge-simple longest paths, we propose to split the structure of the initial graph $\vec{G}$ into a sequence of substructures $\vec{G}'$, $\vec{G}''$,…, $\vec{G}^n$ and search for the appropriate junctions (routes) $M_1, M_2,..., M_n$ on each of them. The set of such routes will make up the general route plan for the MAS in a technological disaster zone.

Splitting the initial graph $\vec{G}$ into a set of substructures $\vec{G}', \vec{G}'',..., \vec{G}^n$ will be based on the iterative re-

DOI 10.15588/1607-3274-2023-1-13

moval from the current substructure of the arcs included in the composition of the extreme path found on this substructure.

The process of splitting the initial graph $\vec{G}$ will continue until there remains no path (transitive closure [42]) connecting the vertices from the sets A and B (see the statement of the problem), or until the fulfillment of another (additional) condition at the choice of the decision maker.

The developed method is structurally presented in Fig. 4.

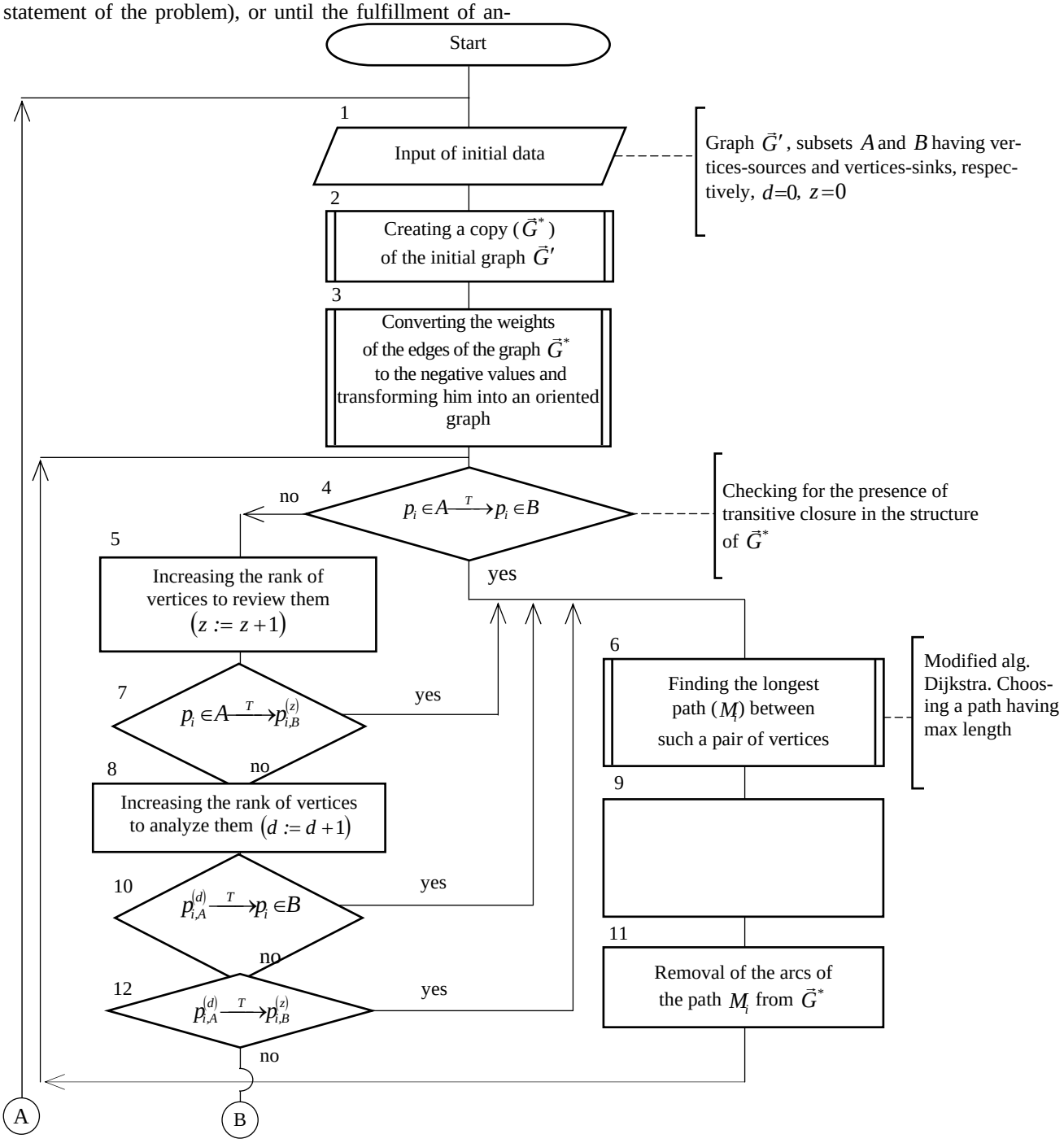


Figure 4 – Block diagram of the method of routing a group of mobile robots on a fixed network to search the missing objects in a technological disaster zone

DOI 10.15588/1607-3274-2023-1-13

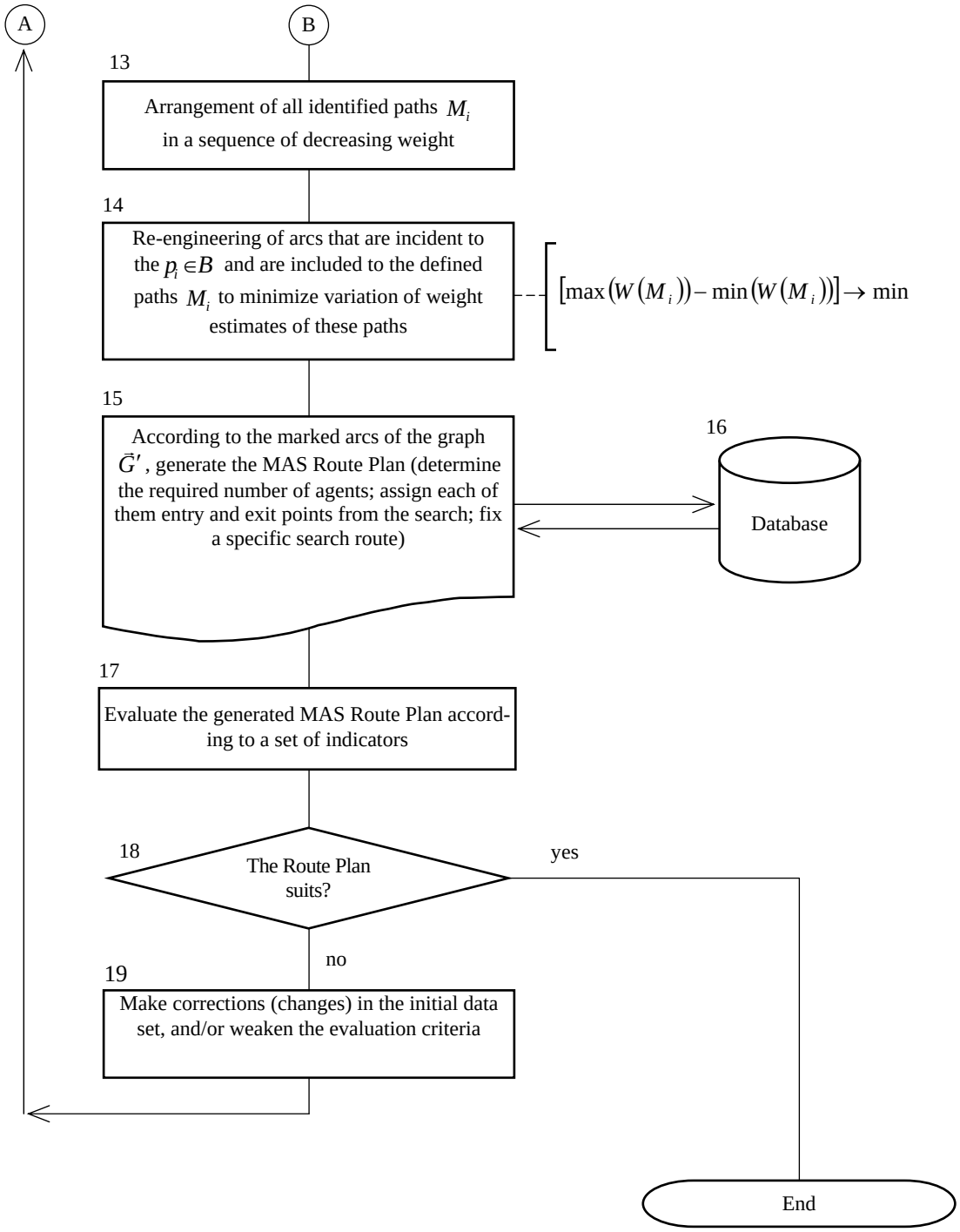


Figure 4, sheet 2

4 EXPERIMENTS

Let's assume that the TDZ transport network is modeled as weighted G' (Fig. 5).

Let's give the edges of the graph a general direction from the vertex p_5 towards the opposite edge of the graph. Thus, subset A will contain only one vertex p_5. We assume the subset B consists of vertices p_3 and p_{10}. Let's represent the weights of the edges as negative values.

Step I. Let's find the longest path from the vertex p_5 to any vertex from the subset B. The current weight estimates obtained due to the modified Dijkstra method and the identified paths themselves are presented in Fig. 6.

DOI 10.15588/1607-3274-2023-1-13

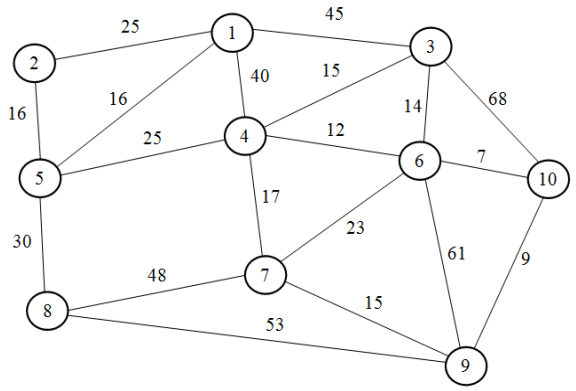


Figure 5 – Weighted graph $\vec{G}'$ simulating the transport network in a technological disaster zone (the weights of edges are conditional lengths of the corresponding communications)

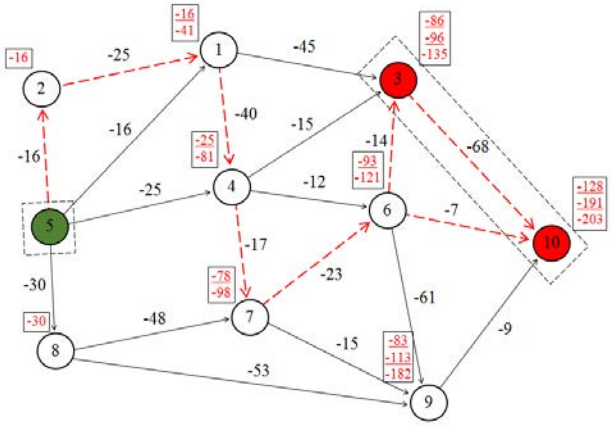


Figure 6 – Initial weighted directed graph $\vec{G}'$

We obtained two paths (routes) to the vertices from the subset B, namely: $M_1 = \{p_5, p_2, p_1, p_4, p_7, p_6, p_3, p_{10}\}$ and $M_2 = \{p_5, p_2, p_1, p_4, p_7, p_6, p_{10}\}$. The absolute weight of route M_1 is greater than that of route M_2, because (|–203|>|–135|). Therefore, the route M_1 is included in the MAS route plan.

Step II. Subsequently, all arcs that make up the route M_2 are removed from the structure $\vec{G}'$. Along with them, all hanging vertices arising from the removal of arcs are removed. As a result, we get the following structure $\vec{G}''$ on which the next longest paths are to be searched (see Fig. 7).

After applying the modified Dijkstra method to the structure of $\vec{G}''$, two paths (routes) to the vertices from the subset B are obtained, namely: $M_3 = \{p_5, p_1, p_3\}$ and $M_4 = \{p_5, p_4, p_6, p_9, p_{10}\}$. The absolute weight of route M_4 is greater than that of route M_3, because $|-107| > |-61|$. Therefore, route M_4 is included in the MAS route plan.

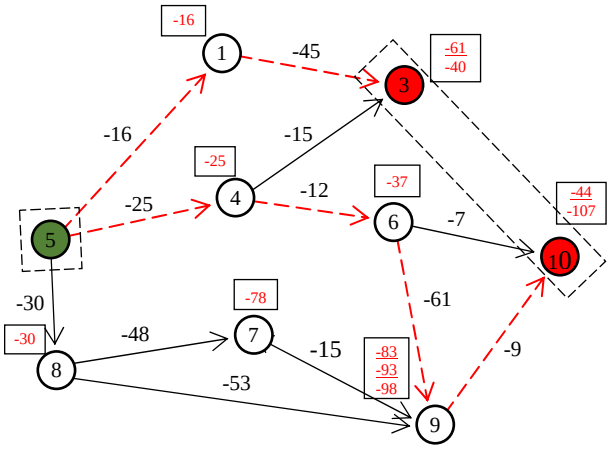


Figure 7 – Weighted graph $\vec{G}''$

Step III. All arcs that make up the route M_4 are to be removed from structure $\vec{G}''$. All hanging vertices are also to be removed. We obtain the following structure $\vec{G}'''$, on which the longest path to any vertex from subset B is also being searched (see Fig. 8). Such a route is the only route $M_5 = \{p_5, p_1, p_3\}$ and its weight is |–61|. Route M_5 is to be included into the MAS route plan.

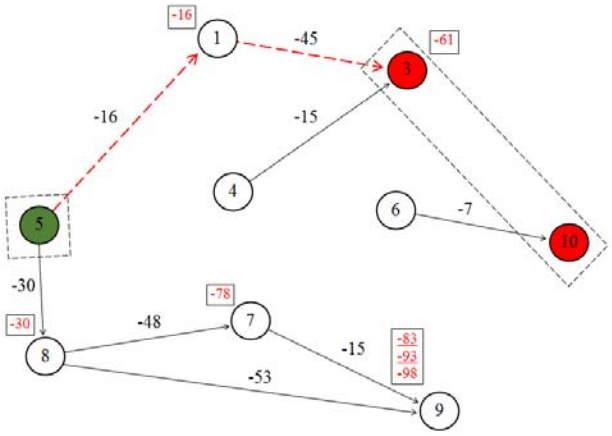


Figure 8 – Weighted graph $\vec{G}'''$

Step IV. All arcs that make up the route M_5 from the structure of $\vec{G}'''$ are to be deleted. Hanging vertices are also removed. As a result, we get the following structure of $\vec{G}''''$ (see Fig. 9). We can see from the figure that in this structure there are no paths connecting vertex p_5 to any vertex from subset B. So, based on the results of this step, the problem (1)–(5) has been basically solved. On the initial structure, all edge-simple longest paths between the defined sets of vertices in the structure of the initial network object $\vec{G}'$ have been determined. We will name such paths as first-level paths because they directly connect vertices from subsets A and B.

DOI 10.15588/1607-3274-2023-1-13

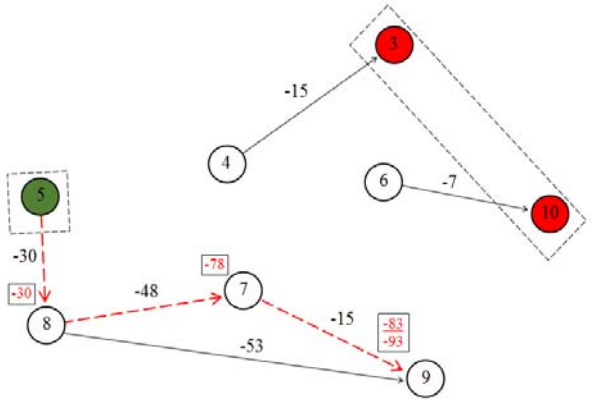


Figure 9 – Weighted graph $\vec{G}''''$

However, vertex p_5 has another unused incident arc (p_5, p_8). Let's find the maximum path to any vertex (in this case, to one that does not belong to subset B). We will consider such paths as paths of the second level: they connect the initial vertex-source with some vertex that is adjacent to some vertex from the subset B. Such paths can also be included into the MAS route plan, thereby increasing its efficiency.

In structure $\vec{G}''''$, such a route is the only route $M_6 = \{p_5, p_8, p_7, p_9\}$ with total weight |–93|. Route M_6 is also included into the MAS route plan.

So, as a result of successive splitting of the initial graph $\vec{G}'$, namely: $\vec{G}' \to \vec{G}'' \to \vec{G}''' \to \vec{G}''''$, four edge-simple longest paths (those that do not have mutual arcs) were found in its structure, namely: $M_1 = \{p_5, p_2, p_1, p_4, p_7, p_6, p_3, p_{10}\}$; $M_4 = \{p_5, p_4, p_6, p_9, p_{10}\}$; $M_5 = \{p_5, p_1, p_3\}$; $M_6 = \{p_5, p_8, p_7, p_9\}$. The paths and their absolute weights are presented in the figure (Fig. 10).

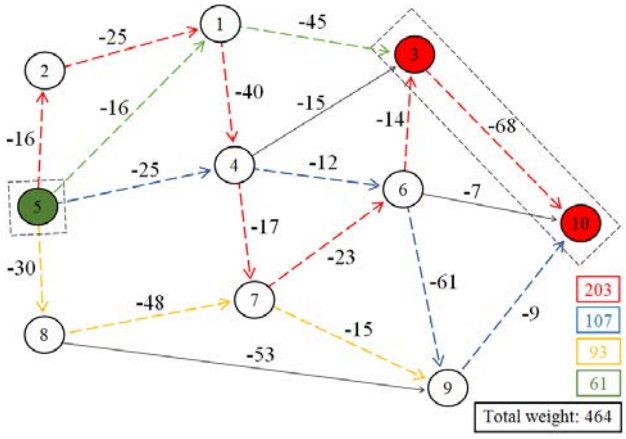


Figure 10 – The set of searched edge-simple longest paths in the structure of the initial network object $\vec{G}'$ (marked with different colors)

5 RESULTS

The scheme of routes presented in Fig. 10, contains the following information:

– to carry out search activities within the TDZ, whose transport network is modeled by graph $\vec{G}'$, it is sufficient to have four agents as part of the MAS.

– for each MAS agent, its starting point (boundary) of entering the search and the final point (boundary) of exiting the search are defined.

– each MAS agent within the TDZ is assigned a specific search (movement) route.

– routes of movement of various MAS agents do not cross sections (arcs).

– in total, the traffic routes of the MAS are optimized both in terms of the direction of traffic and the length of the routes.

The assessments of the developed MAS route plan, based on the totality of the proposed indicators, are presented in Table 1.

Table 1 – Assessments of the developed MAS route plan according to a set of indicators

№	Indicator (parameter)	Indicator value
1	Required number of MAS agents, K_{MAS}, (units): $K_{MAS} \cong n$, where n – the number of defined (selected) search routes	4
2	MAS route plan, PL_{MAS}: $PL_{MAS} = \{M_1, M_2, ..., M_n\}$, where n – the number of defined (selected) search routes	$M_1 = \{p_5, p_2, p_1, p_4, p_7, p_6, p_3, p_{10}\}$; $M_4 = \{p_5, p_4, p_6, p_9, p_{10}\}$; $M_5 = \{p_5, p_1, p_3\}$; $M_6 = \{p_5, p_8, p_7, p_9\}$

DOI 10.15588/1607-3274-2023-1-13

Table 1 – Assessments of the developed MAS route plan according to a set of indicators (continuation)

№	Indicator (parameter)	Indicator value
3	Route length, L_i, (km)	$L_1 = 203$; $L_4 = 107$; $L_5 = 61$; $L_6 = 93$
4	Total length (weight) of routes, L_{PL}, (km): $$L_{PL} = \sum_{i=1}^{n} L_i ,$$ where n – the number of defined (selected) search routes	464
5	Assessment of the search time in TDZ, T_s, (hrs.): $$T_s = \frac{\max\{L_i\}}{\bar{v}_{agt}},\ i = 1...n,$$ where $\bar{v}_{agt}$ – the average speed of the agents moving along the defined routes, km/h.	If: $\bar{v}_{agt} = 5$ km/h. $T_s \approx 40{,}6$ hours
6	Probability of finding search objects in case of implementation of the MAS route plan, P_{dso}: $$P_{dso} = \frac{L_{PL}}{L_{total}},$$ where L_{total} – the total length of the transport network in the TDZ	If: L_{total} =539 km P_{dso} =0,86
7	Probability of finding search objects within the prescribed time for conducting search activities, $P_{dso}(T_s \le T_{dir})$: $$P_{dso}(T_s \le T_{dir}) = 1 - e^{\left(-\frac{T_{dir}}{T_s}\right)},$$ where T_{dir} – the prescribed time	If: T_{dir} =30 hours $P_{dso}(T_s \le T_{dir})$ =0,52

6 DISCUSSION

The set of approaches proposed in the article, together with the modified Dijkstra's method, allowed to develop a method that searches in the structure of any network object, between subsets of its vertices, all existing edge-simple longest paths, which allows to organize the routing of MAS agents in a certain area by paths that do not intersect their individual sections.

At the same time, it is worth noting that the structure of the paths themselves, as well as their number, will depend on the defined (prescribed) points of entry into the search and points of withdrawal from the search, which allows, when planning search activities, to quickly consider several options for future actions and choose the most appropriate one (to order the options by the degree of their advantage).

In addition, having carefully studied the scheme of routes presented in Figure 10, it is possible to conclude that it is expedient to reengineer the routes to reduce the variation of values by their absolute weight (see Fig. 4, Block 14). The idea of reengineering is as follows. If there is an arc (edge) connecting them between any pair of vertices from a subset B, and this arc (edge) is included in some extreme path, and this path has the maximum (higher) weight score among other paths, then this arc (edge) should be included in the path with the minimum (lower) weight. It should be noted that based on the results of using the proposed method, the minimally required number of MAS agents to conduct search activities in an area with a certain transport network could be found.

It is obvious that in the real structures, and hence in the structures which are denser and having more vertices, the number of iterations to split the initial structure and, accordingly, to find the edge-simple longest path, can be close to the very dimension of the initial structure itself. Therefore, the computational complexity of the combinatorial algorithm that implements the developed method will be determined by the computational complexity of its "basic element" – the Dijkstra's algorithm, which is estimated as $O(n^2)$ [40], as well as the computational complexity of the algorithm that determines the presence of transitive closure between vertices from subsets A and B, which is evaluated as $O(n^3)$ [41]. Considering the strengths of subsets A and B, the total computational complexity of the combinatorial algorithm could be estimated as $O\left(|A| \cdot |B| \cdot \left(n^3 + n^2\right)\right)$, where $|A|$ and $|B|$ – the strengths of the corresponding subsets; n – the number of vertices of the model graph G.

The obtained polynomial estimate of computational complexity is quite acceptable for the use of such an algorithm in a real-time framework.

CONCLUSIONS

DOI 10.15588/1607-3274-2023-1-13

The article solves the actual scientific and applied problem of finding all possible edge-simple longest paths between defined subsets of vertices of an initial undirected graph.

The scientific novelty of the developed method is as follows:

1) the representation of the weights of the edges of the initial model graph by the negative values, which allows finding the longest paths between a specified pair of vertices using the classical Dijkstra method;

2) the application of the dynamic programming method, which makes it possible to find the longest paths $M_1, M_2,...,M_n$ in the set of obtained substructures $\vec{G}', \vec{G}'',..., \vec{G}^n$ of the initial model graph $\vec{G}$, which will constitute the complete combination of the edge-simple longest paths.

Since the basic element of the developed method is the Dijkstra's method, which belongs to the accurate class, it can be assumed that the developed method is also accurate.

The practical value of the method is that its application significantly simplifies the process of developing effective route plans for the elements of the MAS in a particular transport network. The projection of general theoretical and methodological statements and conclusions made during the study on the problem of routing of the MAS elements in a certain area, allows to make the search activities effective and quickly develop several options for search actions. The developed method will have a practical use if it is implemented based on a geoinformation system.

A promising direction for further research is the development of routing methods of the MAS performing search activities in a non-stationary network, as well as the development of a set of indicators and criteria for prompt decision-making regarding the optimal (rational) bypassing by agents of the obstacles that suddenly appear on the route path.

ACKNOWLEDGEMENTS

This article highlights one of the results obtained by the authors in 2020–2021 while implementing the research project (state registration number 0120U002173) at the Research Center of the National Academy of the National Guard of Ukraine. The authors are grateful to their colleagues for their support during the research and active participation in the discussion of the results. All authors declare that they have neither financial support nor obligations.

REFERENCES

1. Konovalenko O., Brusentsev V. Multi-agent management and decision support systems, *Bulletin of the National Technical University "KhPI". Ser.: Mechanical engineering and CAD*, 2019, Vol. 1, pp. 18–27. DOI: 10.20998/2079-0775.2019.1.03
2. Kravari K., Bassiliades N. A Survey of Agent Platforms, *Journal of Artificial Societies and Social Simulation*, 2015, Vol. 18, № 1, pp. 1–18. DOI: 10.18564/jasss.2661
3. Schurr N., Schurr N., Marecki J. et al. The Future of Disaster Response: Humans Working with Multiagent Teams using DEFACTO, *Conference: AI Technologies for Homeland Security : Papers from the 2005 AAAI Spring Symposium, Technical Report SS-05-01*. Stanford, California, USA, March 21–23, 2005 : proceedings, Menlo Park, California : The AAAI Press, 2005, pp. 9–16.
4. Ayanian N. DART: Diversity-enhanced Autonomy in Robot Teams, *The International Journal of Robotics Research.* – 2018, pp. 1–8. DOI: 10.1177/0278364919839137.
5. Minieka E. Optimization algorithms on networks and graphs. Moscow, Mir, 1981, 323 p.
6. Held M., Karp R. The Travelling-Salesman Problem and Minimum Spanning Trees: Part II, *Math. Programming*, 1971, Vol. 1, pp. 6–25. DOI:10.1007/BF01584070
7. Bellman R. On a Routing Problem, *Quarterly of Applied Mathematics*, 1958, Vol. 16 (1), pp. 87–90. DOI: 10.1090/qam/102435
8. Hadjiconstantinou E., Christofides N., Mingozzi A. A new exact algorithm for the vehicle routing problem based on q-paths and k-shortest paths relaxations, *Annals of Operations Research*, 1995, Vol. 61, pp. 21–43. DOI:10.1007/BF02098280
9. Christofides N., Mingozzi A., Toth P. Exact algorithm for the Vehicle Routing Problem Based on Spanning Tree and Shortest Paths Relaxations, *Mathematical Programming*, 1981, Vol. 20, No. 1, pp. 255–282. DOI:10.1007/BF01589353
10. Christofides N. Theory of graphs. Algorithmic approach. Moscow, Mir, 1978, 432 p.
11. Karger D. Motwani R., Ramkumar G.D.S. On approximating the longest path in a graph, *Algorithmica*, 1997, Vol. 18, pp. 82–98. DOI: 10.1007/BF02523689
12. Feder T., Motwani R. Finding large cycles in Hamiltonian graphs, *16th annual ACM-SIAM Symp. on Discrete Algorithms (SODA), Vancouver, 23–25 January 2005 : proceedings.* Philadelphia, United States, Society for Industrial and Applied Mathematics, 2005, pp. 166–175. DOI: 10.5555/1070432
13. Gabow H. N. Finding paths and cycles of super polylogarithmic length, *36th annual ACM Symp. on Theory of Computing (STOC), Chicago, 13–16 June 2004 : proceedings.* New York, United States : Association for Computing Machinery, 2004, pp. 407–416. DOI: 10.1145/1007352.1007418
14. Gabow H. N., Nie S. Finding long paths, cycles, and circuits, *19th annual International Symp. on Algorithms and Computation (ISAAC), Gold Coast, Australia, December 2008 : proceedings.* Berlin, Springer-Verlag, 2008, pp. 752–763. DOI: 10.1007/978-3-540-92182-0_66
15. Vishwanathan S. An approximation algorithm for finding a long path in Hamiltonian graphs, *11th annual ACM-SIAM Symp. on Discrete Algorithms (SODA), San Francisco, 9–11 January 2000 : proceedings.* Philadelphia, United States : Society for Industrial and Applied Mathematics, 2000, pp. 680–685. DOI: 10.5555/338219
16. Zhang Z., Li H. Algorithms for long paths in graphs, *Theoretical Computer Science*, 2007. Vol. 377, Issue 1–3. pp. 25–34. DOI: 10.1016/j.tcs.2007.02.012
17. Garey M. R., Johnson D. S. Computers and Intractability: A Guide to the Theory of NP-completeness. New York, W.H. Freeman, 1979, 340 p.
18. Garey M. R., Johnson D. S., Tarjan R. E. The planar Hamiltonian circuit problem is NP-complete, *SIAM*

DOI 10.15588/1607-3274-2023-1-13

J. Computing, 1976. Vol. 5, pp. 704–714. DOI: 10.1137/0205049
19. Golumbic M. C. Algorithmic Graph Theory and Perfect Graphs, *Vol. 57 : Annals of Discrete Mathematics. 2nd Edition*. Amsterdam, North-Holland Publishing Co., 2004, 592 p. ISBN: 9780080526966
20. Müller H. Hamiltonian circuits in chordal bipartite graphs, *Discrete Math.*, 1996, Vol. 156, pp. 291–298. DOI: 10.1016/0012-365X(95)00057-4
21. Narasimhan G. A note on the Hamiltonian circuit problem on directed path graphs, *Information Processing Letters*, 1989, Vol. 32, pp. 167–170. DOI: 10.1016/0020-0190(89)90038-0
22. Damaschke P. The Hamiltonian circuit problem for circle graphs is NP-complete, *Information Processing Letters*. – 1989, Vol. 32, pp. 1–2. DOI: 10.1016/0020-0190(89)90059-8
23. Itai A., Papadimitriou C. H., Szwarcfiter J. L. Hamiltonian paths in grid graphs, *SIAM J. Computing*, 1982, Vol. 11, pp. 676–686. DOI: 10.1137/0211056
24. Arikati S. R., Pandu Rangan C. Linear algorithm for optimal path cover problem on interval graphs, *Information Processing Letters*, 1990, Vol. 35, pp. 149–153. DOI: 10.1016/0020-0190(90)90064-5
25. Bertossi A. A. Finding Hamiltonian circuits in proper interval graphs, *Information Processing Letters*, 1983, Vol. 17, pp. 97–101. DOI: 10.1016/0020-0190(83)90078-9
26. Chang M. S., Peng S. L., Liaw J. L. Deferred-query: An efficient approach for some problems on interval graphs, *Networks*, 1999, Vol. 34, Issue 1, pp. 1–10. DOI: 10.1002/(sici)1097-0037(199908)34:1<1::aid-net1>3.0.co
27. Damaschke P. Paths in interval graphs and circular arc graphs, *Discrete Math.*, 1993, Vol. 112, pp. 49–64. DOI: 10.1016/0012-365X(93)90223-G
28. Asdre K., Nikolopoulos S. D. The 1-fixed-endpoint path cover problem is polynomial on interval graphs, *Algorithmica*, 2009, Vol. 58, pp. 679–710. DOI: 10.1007/s00453-009-9292-5
29. Damaschke P., Deogun J. S., Kratsch D. et al. Finding Hamiltonian paths in cocomparability graphs using the bump number algorithm, *Order*, 1992, Vol. 8, pp. 383–391. DOI: 10.1007/bf00571188
30. Bulterman R., Sommen F. van der, Zwaan G. et al. On computing a longest path in a tee, *Information Processing Letters*, 2002, Vol. 81, pp. 93–96. DOI: 10.1016/S0020-0190(01)00198-3
31. Uehara R., Uno Y. Efficient algorithms for the longest path problem, *15th annual International Symp. on Algorithms and Computation (ISAAC), Hong Kong*, December 2004 : *proceedings.* Berlin, Springer-Verlag, 2004, pp. 871–883. DOI:10.1007/978-3-540-30551-4_74
32. Uehara R., Valiente G. Linear structure of bipartite permutation graphs and the longest path problem, *Information Processing Letters*, 2007, Vol. 103, pp. 71–77. DOI: 10.1016/j.ipl.2007.02.010
33. Takahara Y., Teramoto S., Uehara R. Longest path problems on ptolemaic graphs, *IEICE Trans. Inf. and Syst.*, 2008, Vol. 91-D, pp. 170–177. DOI: 10.1093/ietisy/e91-d.2.170
34. Ioannidou K., Mertzios G., Nikolopoulos S. The Longest Path Problem has a Polynomial Solution on Interval Graphs, *Algorithmica*, 2011, Vol. 61, pp. 320–341. DOI: 10.1007/s00453-010-9411-3
35. Mertzios G. B., Corneil D. G. A Simple Polynomial Algorithm for the Longest Path Problem on Cocomparability Graphs, *SIAM Journal on Discrete Mathematics*, 2012, Vol. 26, Issue 3, pp. 940–963. DOI: 10.1137/100793529
36. Floyd R. Algorithm 97: Shortest Path, *Communications of the ACM*, 1961, Vol. 5 (6), pp. 344–348. DOI: 10.1145/367766.368168
37. Hougardy S. The Floyd-Warshall algorithm on graphs with negative cycles, *Information Processing Letters*, 2010, Vol. 110 (8–9), pp. 279–281. DOI: 10.1016/j.ipe.2010.02.001
38. Warshall S. Algorithm on Boolean matrices, *Journal of the ACM*, 1962, Vol. 9 (1), pp. 11–12. DOI: 10.1145/321105.321107
39. Shimbel A. Structural parameters of communication networks, *Bulletin of Mathematical Biophysics*, 1953, Vol. 15 (4), pp. 501–507. DOI: 10.1007/BF02476438
40. Dijkstra E. W. A note on two problems in connexion with graphs, *Numerische Mathematik*, 1959, Vol. 1, Issue 1, pp. 269–271. DOI:10.1007/BF01386390
41. Lipskyy V. Combinatory for programmers, Moscow, Mir Press., 1988, 213 p.
42. Batsamut V., Manzura S., Kosiak O. et al. Fast Algorithm for Calculating Transitive Closures of Binary Relations in the Structure of a Network Object, *International Journal of Computing*, 2021, Vol. 20(4), pp. 560–566. DOI: 10.47839/ijc.20.4.2444

Received 22.11.2022.
Accepted 03.02.2023.

УДК 519.853: 658.52

МЕТОД МАРШРУТИЗАЦІЇ ГРУПИ МОБІЛЬНИХ РОБОТІВ НА СТАЦІОНАРНІЙ МЕРЕЖІ ДЛЯ ВИКОНАННЯ ЗАВДАНЬ ПОШУКУ ЗНИКЛИХ ОБ'ЄКТІВ В ЗОНІ ТЕХНОГЕННОЇ АВАРІЇ

Бацамут В. М. – д-р військ. наук, професор, заступник начальника науково-дослідного центру службово-бойової діяльності Національної гвардії України Національної академії Національної гвардії України, Харків, Україна.
Годлевський С. О. – науковий співробітник науково-дослідного центру службово-бойової діяльності Національної гвардії України Національної академії Національної гвардії України, Харків, Україна.

АНОТАЦІЯ

Актуальність. Актуальність статті обумовлюється потребою у подальшому розвитку моделей колективної поведінки систем із мультиагентною побудовою структури, у наділенні таких систем інтелектом, який забезпечує синхронізацію спільних зусиль різних агентів у ході досягнення поставлених перед системою цілей. Запропонований у статті метод усуває проблему конкуренції між різними агентами мультиагентної системи, що є важливим у ході виконання пошукових, рятувальних, моніторингових завдань у кризових районах різного характеру походження.

DOI 10.15588/1607-3274-2023-1-13

Мета роботи полягає у розробленні методу визначення достатньої чисельності мультиагентної системи та оптимальних маршрутів руху її окремих елементів на стаціонарній мережі для максимально повного обстеження зони техногенної аварії (будь-якої заданої зони, в основі якої лежить певна транспортна мережа).

Метод. Застосовано ідею динамічного програмування для пошуку в структурі модельного зваженого орієнтованого графа всіх можливих реберно-простих найдовших шляхів, що з'єднують директивно визначені підмножини вершин-істоків та вершин-стоків. З цією метою застосовано модифікований метод Дейкстри. Модифікація полягає у представленні ваг дуг моделюючого орієнтованого графа значеннями з від'ємної області з подальшою роботою метода Дейкстри з цими значеннями. Після відшукування чергового реберно-простого найдовшого шляху, дуги, що його складають, фіксуються у пам'яті обчислювальної системи (у маршрутному плані) та видаляються зі структури графа і процес ітераційно повторюється. Пошук шляхів відбувається доти, поки зберігається транзитивне замкнення між вершинами, що входять до складу визначених підмножин вершин-істоків та вершин-стоків. Розроблений метод дозволяє знайти таку сукупність маршрутів руху для елементів мультиагентної системи, яка максимізує обстежену ними площу в зоні техногенної аварії (або кількість перевірених об'єктів на маршрутах руху) за одну "хвилю" пошуку, та розподіляє елементи мультиагентної системи маршрутами, що не мають спільних ділянок. Похідною застосування розробленого методу є визначення достатньої чисельності мультиагентної системи для ефективного проведення пошукових заходів у межах визначеної зони.

Результати. 1) Розроблено метод маршрутизації групи мобільних роботів на стаціонарній мережі для виконання завдань пошуку зниклих об'єктів в зоні техногенної аварії; 2) Формалізовано робочий вираз методу Дейкстри для пошуку в структурі мережевого об'єкту (в структурі модельного графа) шляхів найбільшої довжини; 3) Запропонована сукупність показників для комплексного оцінювання маршрутних планів мультиагентної системи; 4) Виконано верифікацію методу на тестових задачах.

Висновки. Проведені теоретичні дослідження та низка експериментів підтверджують працездатність розробленого методу. Рішення, що виробляються із використанням розробленого методу, є точними, що дозволяє рекомендувати його до практичного використання при визначенні в автоматизованому режимі маршрутних планів для мультиагентних систем, а також потрібної кількості агентів в таких системах для виконання необхідного обсягу пошукових завдань у певному кризовому районі.

КЛЮЧОВІ СЛОВА: мультиагентна система, група мобільних роботів, маршрутизація, мережевий об'єкт, зважений неорієнтований (орієнтований) граф, екстремальні шляхи, критерій оптимізації, метод.

ЛІТЕРАТУРА

1. Коноваленко О. Мультиагентні системи управління та підтримки прийняття рішень / О. Коноваленко, В. Брусенцев // Bulletin of the National Technical University "KhPI". Ser.: Mechanical engineering and CAD. – 2019. – № 1. – С. 18–27. DOI: 10.20998/2079-0775.2019.1.03
2. Kravari K. A Survey of Agent Platforms / K. Kravari, N. Bassiliades // Journal of Artificial Societies and Social Simulation. – 2015. – Vol. 18, № 1. – P. 1–18. DOI: 10.18564/jasss.2661
3. The Future of Disaster Response: Humans Working with Multiagent Teams using DEFACTO / [N. Schurr, N. Schurr, J. Marecki et al.] // Conference: AI Technologies for Homeland Security : Papers from the 2005 AAAI Spring Symposium, Technical Report SS-05-01, Stanford, California, USA, March 21–23, 2005 : proceedings. – Menlo Park, California : The AAAI Press, 2005. P. 9–16.
4. Ayanian N. DART: Diversity-enhanced Autonomy in Robot Teams / N. Ayanian // The International Journal of Robotics Research. – 2018. – P. 1–8. DOI: 10.1177/0278364919839137
5. Майника Э. Алгоритмы оптимизации на сетях и графах / Э. Майника. – Москва : Мир, 1981. – 323 с.
6. Held M. The Travelling-Salesman Problem and Minimum Spanning Trees: Part II / M. Held, R. Karp // Math. Programming. – 1971. – Vol. 1 – P. 6–25. DOI:10.1007/BF01584070
7. Bellman R. On a Routing Problem / R. Bellman // Quarterly of Applied Mathematics. – 1958. – Vol. 16 (1). – P. 87–90. DOI: 10.1090/qam/102435
8. Hadjiconstantinou E. A new exact algorithm for the vehicle routing problem based on q-paths and k-shortest paths relaxations / E. Hadjiconstantinou, N. Christofides, A. Mingozzi // Annals of Operations Research. – 1995. – Vol. 61. – P. 21–43. DOI:10.1007/BF02098280
9. Christofides N. Exact algorithm for the Vehicle Routing Problem Based on Spanning Tree and Shortest Paths Relaxations / N. Christofides, A. Mingozzi, P. Toth // Mathematical Programming. – 1981. – Vol. 20, No. 1. – P. 255–282. DOI:10.1007/BF01589353
10. Кристофидес Н. Теория графов. Алгоритмический подход / Н. Кристофидес. – Москва : Мир, 1978. – 432 с.
11. Karger D. On approximating the longest path in a graph / D. Karger, R. Motwani, G.D.S. Ramkumar // Algorithmica. – 1997. – Vol. 18. – P. 82–98. DOI: 10.1007/BF02523689
12. Feder T. Finding large cycles in Hamiltonian graphs / T. Feder R. Motwani // 16th annual ACM-SIAM Symp. on Discrete Algorithms (SODA), Vancouver, 23–25 January 2005 : proceedings. – Philadelphia, United States : Society for Industrial and Applied Mathematics, 2005. – P. 166–175. DOI: 10.5555/1070432
13. Gabow H. N. Finding paths and cycles of super polylogarithmic length / H. N. Gabow // 36th annual ACM Symp. on Theory of Computing (STOC), Chicago, 13–16 June 2004 : proceedings. – New York, United States : Association for Computing Machinery, 2004, P. 407–416. DOI: 10.1145/1007352.1007418
14. Gabow H.N. Finding long paths, cycles, and circuits / H. N. Gabow, S. Nie // 19th annual International Symp. on Algorithms and Computation (ISAAC), Gold Coast, Australia, December 2008 : proceedings. – Berlin : Springer-Verlag, 2008. – P. 752–763. DOI: 10.1007/978-3-540-92182-0_66
15. Vishwanathan S. An approximation algorithm for finding a long path in Hamiltonian graphs / S. Vishwanathan // 11th annual ACM-SIAM Symp. on Discrete Algorithms (SODA), San Francisco, 9–11 January 2000 : proceedings. – Philadelphia, United States : Society for Industrial and Applied Mathematics, 2000. – P. 680–685. DOI: 10.5555/338219
16. Zhang Z. Algorithms for long paths in graphs / Z. Zhang, H. Li // Theoretical Computer Science. – 2007. – Vol. 377, Issue 1–3. – P. 25–34. DOI: 10.1016/j.tcs.2007.02.012

DOI 10.15588/1607-3274-2023-1-13

17. Garey M. R. Computers and Intractability: A Guide to the Theory of NP-completeness / M. R. Garey, D. S. Johnson. – New York : W.H. Freeman, 1979. – 340 p.
18. Garey M. R. The planar Hamiltonian circuit problem is NP-complete / M. R. Garey, D. S. Johnson, R. E. Tarjan // SIAM J. Computing. – 1976. – Vol. 5. – P. 704–714. DOI: 10.1137/0205049
19. Golumbic M. C. Algorithmic Graph Theory and Perfect Graphs / M. C. Golumbic. – Vol. 57 : Annals of Discrete Mathematics. 2nd Edition. – Amsterdam : North-Holland Publishing Co., 2004. – 592 p. ISBN: 9780080526966
20. Müller H. Hamiltonian circuits in chordal bipartite graphs / H. Müller // Discrete Math. – 1996. – Vol. 156. – P. 291–298. DOI: 10.1016/0012-365X(95)00057-4
21. Narasimhan G. A note on the Hamiltonian circuit problem on directed path graphs / G. Narasimhan // Information Processing Letters. – 1989. – Vol. 32. – P. 167–170. DOI: 10.1016/0020-0190(89)90038-0
22. Damaschke P. The Hamiltonian circuit problem for circle graphs is NP-complete / P. Damaschke // Information Processing Letters. – 1989. – Vol. 32. – P. 1–2. DOI: 10.1016/0020-0190(89)90059-8
23. Itai A. Hamiltonian paths in grid graphs / A. Itai, C. H. Papadimitriou, J. L. Szwarcfiter // SIAM J. Computing. – 1982. – Vol. 11. – P. 676–686. DOI: 10.1137/0211056
24. Arikati S. R. Linear algorithm for optimal path cover problem on interval graphs / S. R. Arikati, Rangan C. Pandu. // Information Processing Letters. – 1990. – Vol. 35. – P. 149–153. DOI: 10.1016/0020-0190(90)90064-5
25. Bertossi A. A. Finding Hamiltonian circuits in proper interval graphs / A. A. Bertossi // Information Processing Letters. – 1983. – Vol. 17. – P. 97–101. DOI: 10.1016/0020-0190(83)90078-9
26. Chang M. S. Deferred-query: An efficient approach for some problems on interval graphs / M. S. Chang, S. L. Peng, J. L. Liaw // Networks. – 1999. – Vol. 34, Issue 1. – P. 1–10. DOI: 10.1002/(sici)1097-0037(199908)34:1<1::aid-net1>3.0.co
27. Damaschke P. Paths in interval graphs and circular arc graphs / P. Damaschke // Discrete Math. – 1993. – Vol. 112. – P. 49–64. DOI: 10.1016/0012-365X(93)90223-G
28. Asdre K. The 1-fixed-endpoint path cover problem is polynomial on interval graphs / K. Asdre, S. D. Nikolopoulos // Algorithmica. – 2009. – Vol. 58. – P. 679–710. DOI: 10.1007/s00453-009-9292-5
29. Finding Hamiltonian paths in cocomparability graphs using the bump number algorithm / [P. Damaschke, J. S. Deogun, D. Kratsch et al.] // Order. – 1992. – Vol. 8. – P. 383–391. DOI: 10.1007/bf00571188
30. On computing a longest path in a tee / [R. Bulterman, F. van der Sommen, G. Zwaan et al.] // Information Processing Letters. – 2002. – Vol. 81. – P. 93–96. DOI: 10.1016/S0020-0190(01)00198-3
31. Uehara R. Efficient algorithms for the longest path problem / R. Uehara, Y. Uno // 15th annual International Symp. on Algorithms and Computation (ISAAC), Hong Kong, December 2004 : proceedings. – Berlin : Springer-Verlag, 2004 – P. 871–883. DOI:10.1007/978-3-540-30551-4_74
32. Uehara R. Linear structure of bipartite permutation graphs and the longest path problem / R. Uehara, G. Valiente // Information Processing Letters. – 2007. – Vol. 103. – P. 71–77. DOI: 10.1016/j.ipl.2007.02.010
33. Takahara Y. Longest path problems on ptolemaic graphs / Y. Takahara, S. Teramoto, R. Uehara // IEICE Trans. Inf. and Syst. – 2008. – Vol. 91-D. – P. 170–177. DOI: 10.1093/ietisy/e91-d.2.170
34. Ioannidou K. The Longest Path Problem has a Polynomial Solution on Interval Graphs / K. Ioannidou, G. Mertzios, S. Nikolopoulos // Algorithmica. – 2011. – Vol. 61. – P. 320–341. DOI: 10.1007/s00453-010-9411-3
35. Mertzios G. B. A Simple Polynomial Algorithm for the Longest Path Problem on Cocomparability Graphs / G. B. Mertzios , D. G. Corneil // SIAM Journal on Discrete Mathematics. – 2012. – Vol. 26, Issue 3, – P. 940–963. DOI: 10.1137/100793529
36. Floyd R. Algorithm 97: Shortest Path / R. Floyd // Communications of the ACM. – 1961 – Vol. 5 (6). – P. 344–348. DOI: 10.1145/367766.368168
37. Hougardy S. The Floyd-Warshall algorithm on graphs with negative cycles / S. Hougardy // Information Processing Letters. – 2010. – Vol. 110 (8–9). – P. 279–281. DOI: 10.1016/j.ipe.2010.02.001
38. Warshall S. Algorithm on Boolean matrices / S. Warshall // Journal of the ACM. – 1962. – Vol. 9 (1). – P. 11–12. DOI: 10.1145/321105.321107
39. Shimbel A. Structural parameters of communication networks / A. Shimbel // Bulletin of Mathematical Biophysics. – 1953. – Vol. 15 (4). – P. 501–507. DOI: 10.1007/BF02476438
40. Dijkstra E. W. A note on two problems in connexion with graphs / E. W. Dijkstra // Numerische Mathematik. – 1959. – Vol. 1, Issue 1. – P. 269–271. DOI:10.1007/BF01386390.
41. Lipskyy V. Combinatory for programmers / V. Lipskyy. Moscow : Mir Press., 1988. – 213 p.
42. Fast Algorithm for Calculating Transitive Closures of Binary Relations in the Structure of a Network Object / [V. Batsamut, S. Manzura, O. Kosiak et al.] // International Journal of Computing. – 2021. – Vol. 20(4). – P. 560–566. DOI: 10.47839/ijc.20.4.2444

DOI 10.15588/1607-3274-2023-1-13

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2023. № 1
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2023. № 1

UDC 004.451.36:681.5

ANALYSIS OF RISK TERMINAL FLOWS IN TECHNOGENIC SYSTEMS ARISING IN THE PROCESS OF THREAT IMPACT

Sabat V. I. – PhD, Associate Professor, Associate Professor of the Department of Information Multimedia Technologies of Ukrainian Academy of Printing, Lviv, Ukraine.

Sikora L. S. – Dr. Sc., Professor, Full Member of the Engineering Academy of Ukraine, Professor of the Department of Automated Control Systems of the Institute of Computer Sciences and Information Technologies, Lviv, Ukraine.

Durnyak B. V. – Dr. Sc., Professor, Honoured Worker of Science and Technology of Ukraine, Rector of Ukrainian Academy of Printing, Lviv, Ukraine.

Povkhan I. F. – Dr. Sc., Professor, Dean of the Faculty of Information Technologies, Uzhhorod National University, Uzhhorod, Ukraine.

Polishchuk V. V. – Dr. Sc., Associate Professor, Professor of the Department of Systems Software, Uzhhorod National University, Uzhgorod, Ukraine.

ABSTRACT

Context. The analysis of the risk terminal flows in technogenic systems is carried out, which arise in the process of the impact of informational and cognitive threats in the automated document management system as part of the hierarchical production system.

The object of the research is the process of functioning of complex systems with a hierarchical structure, in which automated document management systems with a high level of data flow protection for decision-making are used to provide the information quality control of technological processes.

The subjects of the research are the methods and means of constructing an information protection system to ensure the reliable functioning of automated document management systems and making targeted decisions in hierarchical structures with minimal risk of exposure to external threats and attacks.

Objective is to develop a complex model for assessing the risk of the document management system failure as part of a hierarchical production system under the active threats.

Method. For the first time, the cause-and-effect diagram of the event formation with the active action of threat factors and attacks is substantiated and developed, the interpretation of risk in a technogenic system is defined, and the risk in the space of states is presented as a change in the trajectory in the system transitions to the limit operation mode. For the first time, a category diagram of the structure of risk generation under the threat factors and a system-category diagram of interaction in the system *risk* ↔ *emergency-active nature* is constructed, a system-category scheme of risk formation under the active threat factors is suggested. For the first time, a cognitive diagram for assessing losses in the event of a risk situation arising from incorrect actions of the personnel is substantiated.

Results. As a result of the research, a system-category diagram of the impact of a set of threats on the system functioning mode and process is constructed, a method is developed for calculating the level of system strategic security of energy-active hierarchical systems in the process of attacks and threats, and a complex model for assessing the risk of a system functioning failure under active threats is suggested.

Conclusions. Under the action of active obstacles, cognitive and system factors at the operational and strategic levels of the control hierarchy, due to wrong decisions and informational disorientation, emergency situations and risks of the system function loss and its target-orientation arise. The analysis of a set of risks and the suggested category diagram of the risk generation structure under the impact of threat factors form the basis of the development of the probability structure of the risk concept based on the *attack* ↔ *consequence* model, as well as the construction of a system-category diagram of the interaction in the game *active factor* ↔ *accident risk*. This, in turn, makes it possible to construct a system-category scheme for the formation of risk terminal flows in technogenic systems that arise in the process of threat impact. A complex model for assessing the risk of system failure under threats can be used to construct protection systems for any hierarchical control structures of technogenic systems.

KEYWORDS: technogenic systems, threats, vulnerabilities, risk assessment, decision making, control of hierarchical systems.

NOMENCLATURE

α_{risk} is an assessment of the risk level;

α_d is an acceptable risk level;

KIA_i is a crisis information agent;

F_i are attack factors;

SV_i is a technogenic system (TS);

Π_i are resource flows;

$\{x_i\}$ is a vector of possible alternatives on the set X of the target space partitioning;

ω_i are external negative factors from the set Ω;

f is a function of the relationship between decision x and consequence y;

$StratU$ is a control strategy;

R_i are control system resources;

$V(x_i, \omega)$ is a selected alternative for the space partitioning of the system states;

T_m is an interval of the threat;

D_i is an event in the security system, situation;

τ_i is a time of the emergency event;

C_i is a consequence of an emergency event;

$Sit\Pi S$ is a situation in the space of states;

DOI 10.15588/1607-3274-2023-1-14

$\text{Resurs}(TO)$ is a resource of the technological object under the impact of external threat factors;

R_d is a permissible resource value under normal conditions;

$Opt(U/S_i)$ is an optimal control strategy;

V_i are losses upon successful completion of the attack;

$A(f_i)$ is an activity of the factor action on the aggregate structure and control system;

P_i is a probability of the event occurrence;

N_{di}^{f} is a consequence of the influencing factor f_i;

V_{ni}^{f} is a significance of the consequences after the effect of the influencing factor f_i;

$\text{U}(W)$ is a utility function;

Sh_W is a scale of the utility function with the values interval I;

K_i is a mechanisms and requirements for information confidentiality;

$\{Z_i\}$ are types of threats;

$\{Va_i\}$ are loss of information authenticity;

$r(\Pi\alpha_{risk})$ is a flow of cognitive risks;

$\Pi\mu(\alpha_{risk})$ are errors of the personnel;

$\alpha_{risk\,opt}$ is an optimal risk value, which does not exceed the permissible value.

INTRODUCTION

The functioning of any technogenic system in the environment is vulnerable to a variety of interdependent negative factors, independent disturbances and threats that can lead to crisis situations, accidents and catastrophes, if preventive measures to assess the accident risk are not developed and decision-making procedures are not established for control under active informational and cognitive threats. Usually, such procedures and means of counteracting negative factors are provided at the beginning of the design of the protection system and are described in the form of provisions in the security policy of any organization with a hierarchical control system. However, there are negative factors and threats that may arise in the process of the system functioning, which also require the use of a probability approach to determine the risks associated with the action of active threats. Therefore, an important task of the research in the protection system is and will always be the analysis of risk flows, which together form a complex model for assessing the control failure risk for technogenic systems.

If one proceeds from the system concept of assessing the situation in local, district, regional infrastructures, it can be concluded that they were not ready for aggressive attacks on their structure and control process, because they were based on the concept of terminal stability. The action of informational and infrastructural attacks of an aggressive type led to their collapse and the emergence of emergency situations with a high level of accident risk.

The object of the research is the process of functioning of complex systems with a hierarchical structure, in which automated document management systems with a high level of data flow protection for decision-making are used to provide the information quality control of technological processes.

The subjects of the research are the methods and means of constructing an information protection system to ensure the reliable functioning of automated document management systems and making targeted decisions in hierarchical structures with minimal risk of exposure to external threats and attacks.

The goal of this work is to develop a complex model for assessing the risk of the document management system failure as part of a hierarchical production system under active threats.

1 PROBLEM STATEMENT

To achieve the goal of the scientific research, it is necessary to solve the following tasks:

– for the first time, to develop a cause-and-effect diagram of the event formation during the active action of threat factors and attacks, to form the interpretation of risk in a technogenic system in the space of states as a change in the trajectory in the system transitions to the limit functioning mode;

– for the first time, to construct a category diagram of the structure of risk generation under the threat factors and a system-category diagram of interaction in the system *risk* ↔ *emergency-active nature*;

– for the first time, to develop a system-category scheme of risk formation in the conditions of active threat factors and a cognitive diagram for assessing losses in the event of a risk situation arising from incorrect actions of personnel;

– to test and verify the suggested complex model for assessing the risk of the document management system failure as part of a hierarchical production system for the example of risk assessment of printing productions, as well as to offer a system-category interaction diagram in the game *active factor* ↔ *accident risk*.

Problem setting is formulated as follows. Let one have some research object SV_i in the input, which is assessed by many indicators depending on the control strategy $StratU$ and security policy. The output may deviate from the control target due to the set of threats $\{Z_i\}$ and the attack factors F_i, which these threats use. As a result of the construction of category diagrams of risk formation in the space of system states under the threat factors in the time interval T_m, it is possible to assess the amount of losses in the risk situations and in the event of incorrect actions of the personnel, the level of which can, for example, be represented by using the models of utility functions for risk assessment $\alpha_{risk} = \{\alpha_{r1}(F_1),\ldots,\alpha_{rn}(F_n)\}$ from the interval [0; 1]. Moreover, the indicators $\alpha_{ri}(F_i)$

DOI 10.15588/1607-3274-2023-1-14

can represent the whole system of criteria and models, on the basis of which one aggregated assessment $\alpha_{risk\,opt}$ on the scale of the utility function Sh_W is derived for each factor of threats and attacks F_i, which is equated to the acceptable risk level α_d.

In addition to quantitative assessments, the reasoning of experts analysing the object is used for the research object. For this, on the basis of experience and knowledge about the research object SV_i, a group of experts (or an expert) analyses it, draws conclusions and assigns one linguistic assessment to each indicator α_{risk}, from the set $r(\Pi\alpha_{risk}) = \{F_i; Z_i; KIA_i; Va_i; \mu(\Pi\alpha_{risk})\}$.

Thus, for a complex assessment of the risk level for a technogenic system, it is necessary to conduct an analysis of the attack factors, multiple threats, loss of authenticity of the control information and possible errors of the personnel, which lead to risk situations and deviations of the system from target orientation. On the basis of the presented input data, for the research object SV_i, it is necessary to derive the initial aggregate assessment of the risk level in the process of the threats impact on the technogenic system $\alpha_{risk} \in [0;1]$. Analysing the value of the assessment α_{risk} and equating it to the permissible values α_d established at the design stage and prescribed in the security policy SV_i, it is possible to adjust the level of its protection by implementing optimal countermeasures.

2 REVIEW OF THE LITERATURE

Analysis of the problem of the emergency and risk situations occurrence, under active threats and attacks, shows the importance of constructing models of attack penetration channels and methods of action on system nodes. In the work of A. M. Shurygin [1], the use of statistical methods for forecasting risk situations and risks assessing is substantiated; the basic concepts, essence, objectives and methods of information protection and the organization of printing productions are revealed in the educational manuals of Vietnamese scientists [2, 3]; in the publications dedicated to the printing industry, the production of securities [4, 5], the peculiarities of the control of printing production and the protection of printing products are considered; scientific publications are devoted to various thorough means of information protection: by Schneier Bruce [6] on applied cryptography, by Mykhailo Sikorsky, and others [7] on analysis of malicious software; the work of L. H. Koval and other scientists [8] is on the analysis of biometric identification methods. With the help of the above-mentioned works, the methods and means of printing production protection are developed, also using the normative legal framework of the Laws of Ukraine [9], as well as the methods of software and hardware protection of network technologies [10], mobile communication devices [11].

In the work [12], modern concepts of information protection against information attacks and system threats are substantiated; the authority control in information security systems, in particular, new approaches to the protection of printing companies, is discussed in the monograph [13].

In [14], information technologies for controlling complex hierarchical systems under threats and information attacks are studied. The concept of risk assessment is formed on the basis of determining the probability and frequency of threats and vulnerabilities for the company assets. The scientific article [15] describes an expert model for assessing risks and security incidents of airport network and information systems, based on intellectual analysis of knowledge using the apparatus of fuzzy sets. Nevertheless, the model is not able to assess the system impact on the process of the functioning system control. The document [16] analyses the concept of risk and safety of subway passengers in cases of malicious technogenic incidents. As a result, the importance of protecting passengers in terms of increasing safety and avoiding dangerous conditions is proven, using the example of the Athens metro system. These studies reveal the essence of hierarchical systems and their vulnerability to technogenic disasters under the external attacks and internal threats impact.

Modern developed methods of analysis of general industrial control systems of hierarchical technogenic structures are presented in the works of foreign scientists [17, 18], and they are given in [19, 20] specifically for complex systems.

The paper [21] presents the application of an object-oriented Bayesian network for scenario risk assessment. A model of probability coverage of key factors influencing accidents in fragmented structures is developed. In the studies [22], a model-based methodology for hybrid control of risk assessment of reliability, availability, maintainability, and safety for critically important systems is proposed. As a result, a method of cyber security risk analysis for industrial control systems is created. Agrawal et al. [23] define the ontology to represent the ISO/IEC 27,005, 2018 standards, with the aim of providing a step-by-step understanding of the meaning of security concepts and their relationships. For example, cyber security ontologies are developed by Arbanas & Čubrilo [24], who are able to construct 52 security ontologies. Researchers such as Blanco et al. [25] considered 31 security ontologies. Both studies group the security ontologies into three categories: general, specific, and theoretical.

Two popular risk assessment methods approved for the nuclear sphere use probability risk assessment [26, 27], and others use dynamic Bayesian networks [28, 29].

In the works of J. Rabcan and others [30, 31], the problem of developing a new algorithm with the application of a fuzzy classifier for signal classification is proposed. The results can be used to automate the process of constructing recognition models using precedents.

Nevertheless, a thorough systematic analysis of terminal risks in technogenic systems under the active threats of resource, information and cognitive types based on category models of influence channels on the control process has not been carried out to date.

DOI 10.15588/1607-3274-2023-1-14

In view of all the above-mentioned facts, it has been decided to carry out an innovative study on the development of a complex model for assessing the risk of the document management system failure as part of a hierarchical production system under the active threats impact.

3 MATERIALS AND METHODS

To analyse risks in technogenic systems and construct schemes and methods for their minimization and control, it is necessary to apply a risk analysis methodology based on the following components:

– the source of risk factors (activity, power, channels);

– the scenario of active actions and the factors impact on the system functioning process (structural, resource, information failure);

– the analysis of the results of the factors effects on the system;

– the structure of the intrusion zone of the active system, which allows the action of an intelligent threat agent through unidentified channels;

– active agents that form the means of the function and structure collapse of the technogenic hierarchical system at the physical level and errors in projects.

The source of risk is related to the consequences of active actions through the scenario – the chain of events of risk implementation in the system under certain conditions, which leads to negative consequences and accidents (Fig. 1).

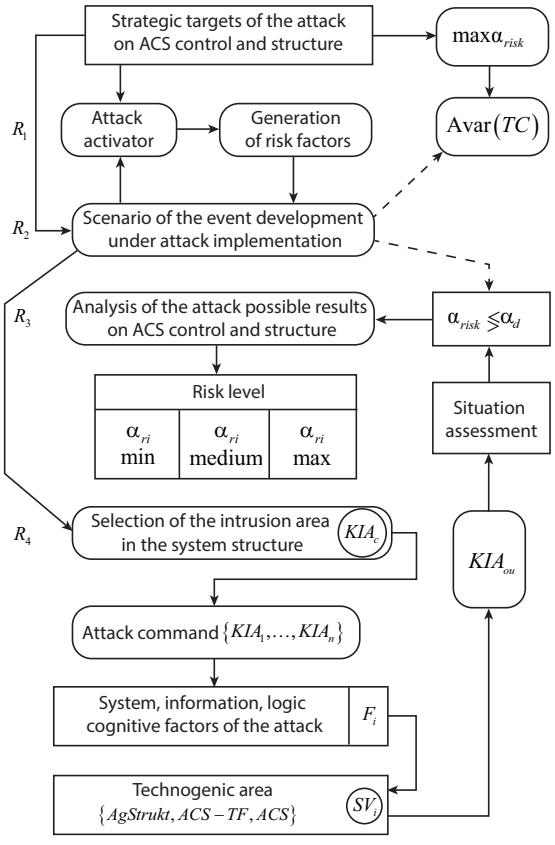


Figure 1 – Informational and cognitive map of the system intrusion process

Chains, paths and directions of connections are actually scenarios of the development of a dangerous situation from the point of view of different positions. They describe the scenarios of events that can happen to the system under the action of active factors generated by the source of risks – an active agent, an attacking system, a hidden internal crisis agent, errors of managerial personnel when making decisions.

With the action of active obstacles, cognitive and system factors at the operational and strategic levels of the control hierarchy, due to incorrect decisions and information disorientation, emergency situations and risks of loss of system functions and its target-orientation arise (Fig. 2).

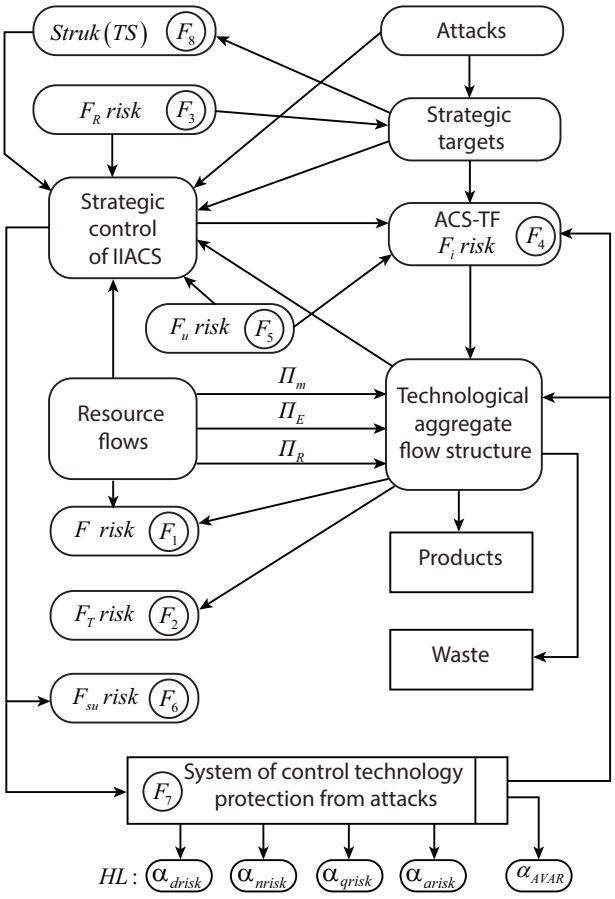


Figure 2 – Category diagram of the formation of hybrid complexes of active risks under the attack factors flow

Here are the schemes for forming a set of risks:

1. Physical risks – physical loss of resources during the system operation, collapse of structures, blocks, aggregates, nodes;

2. Technogenic risks – failure of systems, networks, computers, communications, DoS attacks and threatening destructive actions, power supply failures;

3. Position and cognitive risks – non-compliance with the criteria, regarding the position and abilities of the person, which leads to wrong decisions;

DOI 10.15588/1607-3274-2023-1-14

4. Information risks – loss of data, unauthorized access to ports, terminals, cryptosystems, attacks on databases and information security systems;

5. Management risks – access to decision-making systems, analysis, control, forgery of powers, disorientation of personnel;

6. General organizational risks – conflicts, personnel risks, erroneous setting of targets, inadequacy of situations perception in the control system, inadequacy of strategic targets, sabotage;

7. Risks of violation of the system security level – associated with attacks on the existing levels of the control system, selection and processing of data and decisions;

8. System risks – associated with possible errors when selecting a system concept (targets, structure, dynamics, data processing, control strategies) – by control and operational, project staff.

The level of risk (permissible, sufficient, limit, warning, emergency) is the basis of the classification of both system functioning modes and the assessment of the reliability of the functioning of aggregates, blocks, and control processes in complex integrated hierarchical systems.

The risk of control failure will be analysed in conditions of stochastic uncertainty of the situation under the action of active factors in the space of system states and the target of the control system.

Let one have $\{x_i\}$ – a vector of possible alternatives on an admissible set X of the space partitioning of targets (modes, states).

A rational selection is made according to the consequences that the control action leads to under the factors impact $\omega_i \in \Omega$ at the moment t_i.

The connection between the decision f regarding x and the consequence y is determined by:

$$\exists Strat(u/f): y = f(x, \omega, t_i),$$

$$f: (X \times \Omega \times T) \to Y \to \langle f_{onm} \in F \rangle,$$

where $\{f(x,\omega)/t_i\}$ – is the function that characterizes (costs-spending); $f: (X \times \Omega) \to Y$ – is the model of the decision-making process; $y = f\left(x_i\,|_{i=1}^{n}, \omega\right)$ – defines the process of calculating the consequences of the stochastic factor on x_i.

The two-stage decision-making process, under the influencing factor, has the following structure related to the cause-and-effect representation: [1]

1. The decision is made for the first move according to the alternative $x_i \in X$, then random factors are implemented $\omega_i \in \Omega$ ($AF_i \to \tau_i \to D_i \to Sit\Pi S$).

2. The decision is made $y = Y(x_i, \omega_i)$, which corrects x_i ($\exists StratU, U: x \to y$).

Implementation costs x_i will be $f_1(x_i)$, and for the implementation $y = f_2(x, y, \omega)$, thus, if:

$$\exists W\left(\text{Resurs}(TO) > R_d\right) \Rightarrow \exists Strat(U/f_1, f_2):$$

$$: \left\langle f(x_i, \omega) = \begin{cases} f_1(x_i), \\ f_2\left(x^*, y(x, \omega), \omega\right) \end{cases} \right\rangle \to \langle Opt(U/S_i) \rangle.$$

From the above, it is possible to construct loss functions according to the risk level of control selection on the terminal decision-making cycle for active target-oriented control. At the same time, the level of losses under decision-making risk may depend on the selection of strategies:

1. $F_1(x) = f\left(x_i, \omega^*\right)$ – are normalized losses of resources, products;

2. $F_2(x) = \max f(x, \omega)$ – are the highest losses of the system type;

3. $F_3(x) = \overset{\omega_i \in \Omega}{\text{M}}\, f(x_i, \omega) = f(x, \omega) P(d\omega)$ – is the function of the probable average risk under the action of active threats flows;

4. $F_4(x) = P\left\{ f(x_i, \omega) \le f(x, \omega) \int P(d\omega) \right\}:$ $: (\omega \mid f(x, \omega) \le c)$ – is the probability of losses under active threats that create an emergency situation.

The risk assessment is determined according to the situation and the activation of the threats, i.e.:

$\alpha_{risk}(\omega_i \mid x_i) = f(x_i, \omega) = F(x)$ – is determined by the selected alternative on the space partitioning of the system state in the form of a category diagram (Fig. 3).

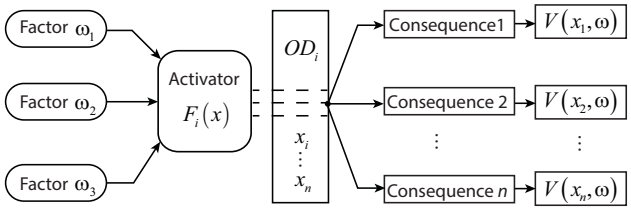


Figure 3 – Category diagram of the structure of risk generation under threat factors

According to the above category diagram, the probability structure of the concept of risk is justified based on the model $\langle attack \to consequence \rangle$:

$$\forall t_i \in T_m, \left\langle \begin{array}{l} \text{if } P_i(t_i) > 0, \text{ then} \\ Risk = \bigcup_{i=1}^{n} (P_i, C_i) \Rightarrow \max(C_r / V_i) \end{array} \right\rangle,$$

where P_i – is the event probability; C_i – is its consequence; V_i – are resource losses at C_r; T_m – is the interval of the treat action.

DOI 10.15588/1607-3274-2023-1-14

Then the general risk form for a certain type of active influencing factor is determined according to the formula:

$$\forall t_i \in T_m, V_m^f > 0, P_i > 0 : \begin{cases} Risk = \bigcup_{i=1}^{n} \left(P_i, N_{di}^f, V_{ni}^f \right) \to \min \alpha_v, \\ \text{at } A\left(f_i \right) \to 0 \end{cases},$$

Then the structure of the interaction system $\langle structure \leftrightarrow active\ factors \rangle$ can be presented in the form of a category diagram of threats impact in Fig. 4, 5, 6.

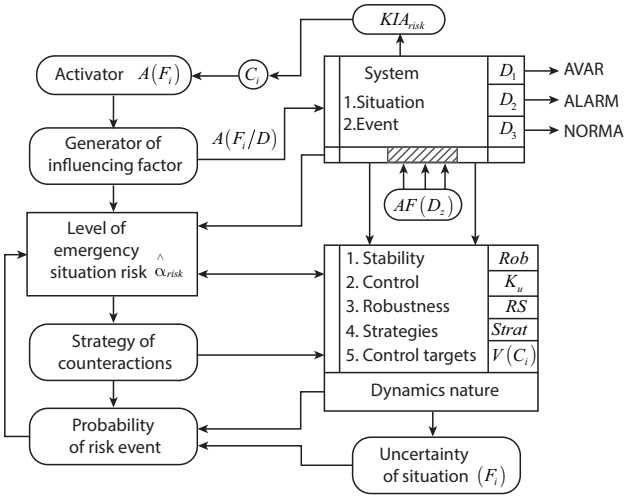


Figure 4 – System-category diagram of interaction in the game $\langle active\ factor \leftrightarrow accident\ risk \rangle$

According to the above analysis of events in the system, which is affected by both control actions and active threats, a set of utility functions of the action of a complex of factors $\left\{ F_k \middle|_{k=1,m} \right\}$ is constructed from the selected loss minimization strategies.

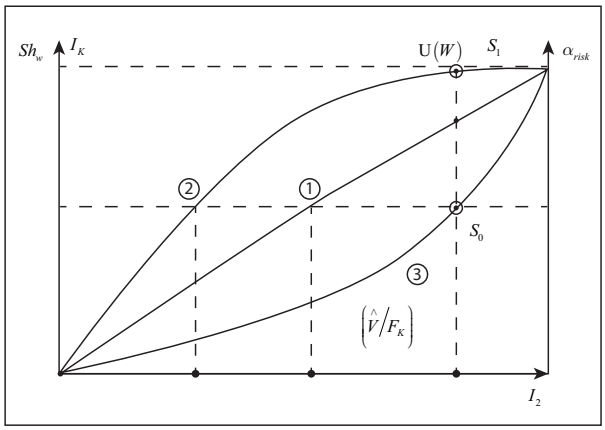


Figure 5 – Graph of the utility function when making decisions

Symbols in Fig. 5: Sh_W – is the scale of the utility function with the value interval I_k; $\mathrm{U}(W)$ – is the Neumann-Morgenstein utility function when forming the decision selection on the set of alternatives; $\left(\hat{W} = \sum_{i=1}^{n} W_i P_i \right)$ – is the expected benefit of the event with the probability of the consequences of the individual selection of the behaviour strategy (target-oriented) in relation to the system, situation $\left(U(W) = P(U \mid S) - (1-p) U(S) \right)$ – i.e. between maximum and minimum (S_1, S_0).

Graphs of the utility function $\mathrm{U}(W)$ when making decisions of the PMD-KIA (the person making the decision), with the expected benefit $\left(\hat{V}/F_{Ki} \right)$, determines the losses in case of incorrect actions of the operator with different types of behaviour:

1. PMD_1 –indifferent to risk (cognitively resistant);
2. PMD – not prone to risk (mentally unstable);
3. PMD_3 – prone to risk when making management decisions.

According to the determination of the usefulness level from control actions, under threats, a system-category scheme of the formation of a risk situation in the technogenic hierarchical structure (*TS*) is constructed (Fig. 6).

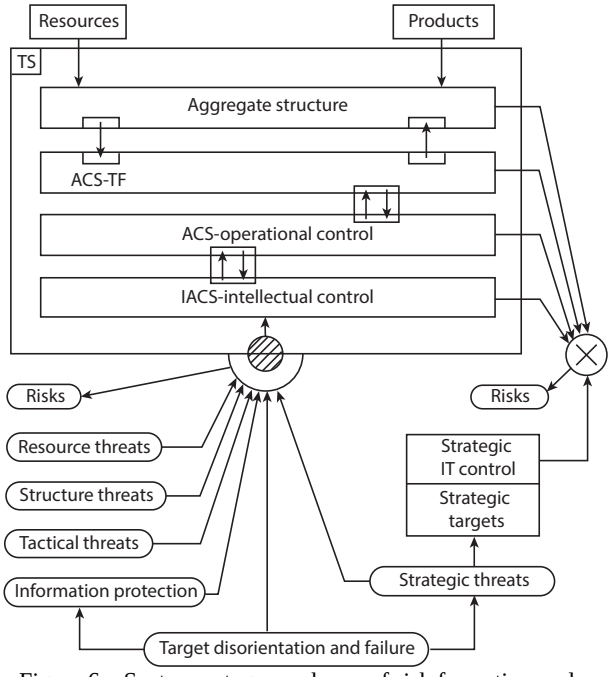


Figure 6 – System-category scheme of risk formation under active threats to the functioning of technogenic hierarchical structure objects

4 EXPERIMENTS

Informational and cognitive threats will be analysed and studied in the automated document management system as part of the hierarchical production system.

The security system is primarily focused on identifying threats and, accordingly, risks of losses for the organization. Such threats include threats to confidentiality, integrity, availability, accountability, authenticity, and reliability of information.

DOI 10.15588/1607-3274-2023-1-14

When assessing the above threats, it is important to directly analyse the security issues with system assets, as this can affect possible threats and, as a result, the selection of protective measures. Usually, such an assessment contains a specific approach to each organization in particular, and also requires the involvement of specialists not only in the field of information protection, but also directly from the organization's specialists, who can assess the system assets, analyse the consequences of damage to these assets and means for restoring the working mode of the system after possible incidents. An information security incident is any unforeseen or undesirable event that can disrupt operations or information security. Information security incidents are: loss of services, equipment or devices; system failures or overloads; user errors; non-compliance with policies or guidelines; violation of physical protection measures; uncontrolled system changes; software failures and failures of technical means; violation of access rules.

Let one consider each of the threats in more detail, analyse the possible consequences of successfully implementing attacks on the system and the necessary countermeasures to reduce them [2–4], and construct a category diagram of the threats actions to the system and control processes (Fig. 7, 8).

Threats to the data privacy about the entire system usually use information access attacks. Threats to privacy can be countered by appropriate privacy and identification services. The methods of such attacks on the information system (*IS*) are usually reduced to three unauthorized actions: spying, eavesdropping and interception.

The mechanisms for ensuring the confidentiality of information in the form of files are presented in Table 1.

Table 1 – File confidentiality mechanisms and requirements for them

Mechanisms for ensuring confidentiality	K_1	Physical security control
	K_2	Access control to the computer files
	K_3	File encryption
File confidentiality requirements	K_4	Identification and authentication
	K_5	Correct setting of the computer system
	K_6	Correct key control when using encryption

Spying (ZA_1) is carried out by accessing unauthorized information and viewing it. If these are paper documents, first of all, it is necessary to ensure their physical protection, preventing third parties from accessing confidential information (security, locks, safes, surveillance and alarm systems, etc.) [12]. If it is an electronic document, then in addition to physical protection and identification of authorized users, it is necessary to implement their authentication methods. At the same time, the following mechanisms of document files confidentiality and requirements for them should be taken into account [13]:

Eavesdropping (ZA_2) can be carried out, for example, by connecting to a line and listening to a telephone conversation. Electromagnetic radiation from sources of information dissemination in *IS* can also be used.

Let one study the threats from information attacks.

If the information is transmitted through an internal or external network, it is possible to intercept it.

In the case of interception, it is important to take into account the fact that an attacker can carry out such an attack, but in order to prevent its further successful development, it is necessary to implement countermeasures to encrypt such information. At the same time, it is necessary to introduce reliable encryption technologies for information flows, or all communication traffic (Fig. 7) [6].

Loss of privacy can lead to the following negative consequences:

– loss of public trust or lowering of the organization's image in the society (VR_1);

– liability before the law, including liability for violation of legislation in the field of data protection (VR_2);

– negative influence on the organization policy (VR_3);

– creating a threat to the safety of the organization's personnel (VR_4);

– financial losses (VR_5).

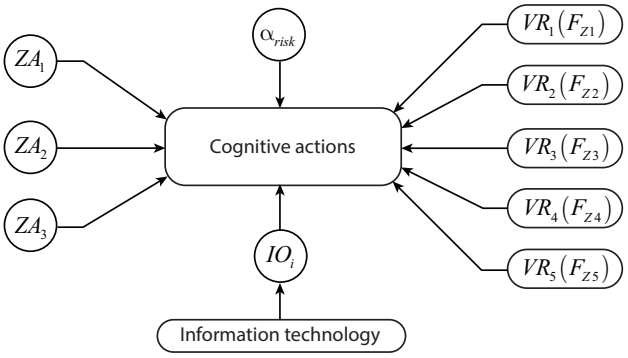


Figure 7 – Cognitive diagram for assessing losses in the risk situation

Symbols in Fig. 7: $\{ZA_i\}$ – are types of threats affecting the system; α_{risk} – is the risk level; IO_i – is a type of the information operation; $\{VR_i\}$ – are losses under the factors impact $\{F_i\}$.

Let one consider informational threats to the data integrity as the basis for errors in the formation of management decisions.

The threat to the information integrity leads to the implementation of attacks on its modification, so all possible negative events that can lead to such incidents should also be analysed (Fig. 8). Such events include the following:

– physical access (ZB_1) to the places of storage of information carriers – it threatens the integrity of the information located on such carriers);

DOI 10.15588/1607-3274-2023-1-14

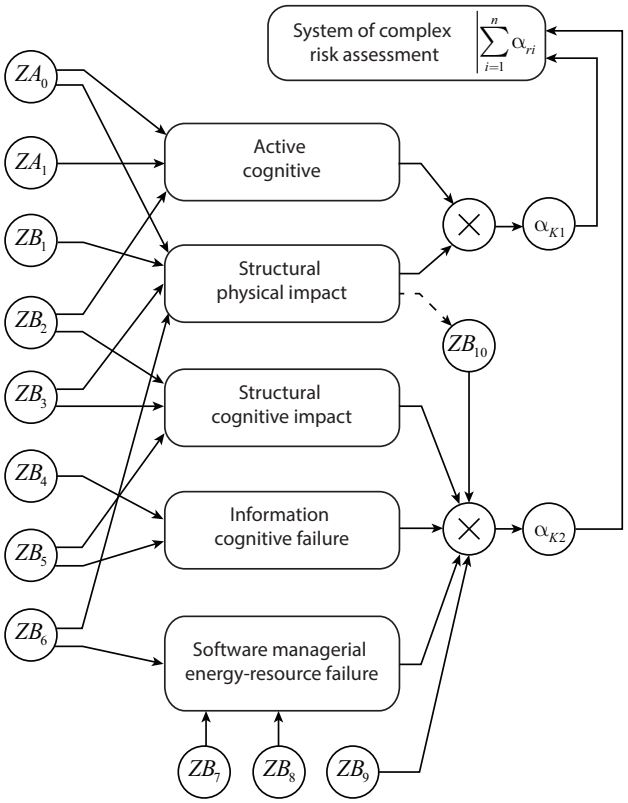


Figure 8 – System-category diagram of the impact of a set of threats on the system

– maintenance error (ZB_2) – it occurs if it is carried out irregularly or without compliance with all security policy procedures, then the integrity of the relevant information is at risk;

– malicious code (ZB_3) – it can lead to a violation of integrity, for example, if changes are made to data or files by an unauthorized person who gained access using malicious code, or such changes are made by the code itself [7];

– spoofing a legitimate user ID (ZB_4) – it can be used to bypass authentication and related services and security functions. As a result, integrity problems may arise every time when the information is accessed and modified under the guise of a legitimate user [8];

– sending messages along a wrong or changed route (ZB_5) – it can lead to a violation of integrity, for example, when messages are changed and then transmitted to the original addressee;

– software failure (ZB_6) – it may violate the integrity of data and information that is processed with the help of such software;

– failures in power sources (ZB_7) – (electrical supply and ventilation) – it can cause integrity problems, if such violations are the cause of other malfunctions. For example, power outages can lead to hardware failure, technical malfunctions, or data storage problems. Such problems include technical malfunctions;

– unauthorized access to computers, data, services and applications (ZB_8) – it can become a threat to the information integrity if their unauthorized change is possible [13];

– use of unauthorized programs and data (ZB_9) – it creates a threat to the information integrity in the storage device and during its processing in the system, if these programs and data are used to illegally change information or contain malicious code;

– unauthorized access to the place of storage of information carriers (ZB_{10}) – it may endanger the integrity of this information, because in this case, unauthorized changes to the information recorded on this information carrier are possible.

The integrity service monitors the reliability of information. With the proper level of organization, it gives users confidence that the information is correct and no one has changed it. The integrity service must work together with the identification service to perform a reliable verification of the authenticity of the person, his authenticity to the admission level. Therefore, the integrity service is a “shield” against modification attacks [18].

Loss of integrity can lead to the following:

– making wrong decisions (NR);

– failure in the organization’s commercial operations (NK_o);

– loss of public trust or lowering of the public image of the organization (VD_r) that performs social activities;

– financial losses (VF) from crisis situations and emergencies;

– liability before the law, including liability for violations of legislation in the field of data protection (ZV_Z).

Hybrid threats to the availability of an automated hierarchical control system from attacking agents usually include events that allow attackers to carry out denial-of-service attacks. Among security specialists, such attacks are also called DoS (Denial-of-Service) attacks. The following threats are considered that can lead to the specified attacks [14]:

– destructive actions (ZRD_1) – destructive attacks, which can also be called vandalism;

– physical access (ZRD_2) – to storage locations of information carriers – it threatens the readiness for functioning of storage facilities;

– equipment malfunction (ZRD_3) – connection and failure of communication services;

– maintenance error (ZDp) – it often occurs if maintenance is carried out irregularly or with errors;

– malicious code (ZKd) – it can be used to bypass the authentication and related services and security functions. As a result, this can lead to a loss of accessibility. For example, if data or files are destroyed by a person

DOI 10.15588/1607-3274-2023-1-14

who gained unauthorized access using malicious code, or the code itself erases files [7];

– spoofing a legitimate user ID (ZIK) – it can be used to bypass the authentication and all related services and security functions. As a result, accessibility problems may arise every time when impersonating a legitimate user makes it possible to delete or destroy information [5];

– incorrect routing (ZMi) or change of message routing [10];

– abuse of resources (ZR), which leads to failures of the network mode.

– natural disasters (ZKS) – impact on the structure and energy supply;

– software failures (ZZp) – it can lead to the unavailability of data and information that is processed with the help of these programs;

– disruptions in supply $(ZNRp)$ – it can lead to availability problems if these disruptions are the cause of other malfunctions. For example, power outages can cause hardware failure, technical malfunctions, or data storage problems. Therefore, it is advisable to provide workplaces with uninterrupted power supply units [12];

– technical malfunctions (ZNI) of nodes, blocks, system structures;

– theft (ZRs) of spare sets for communication and control systems, which leads to an accident;

– traffic overload $(ZNpt)$ – it will reduce system reliability;

– transmission errors (ZNp) – effects of interference on data transmission channels and systems;

– unauthorized access to computers (ZND), data, services and applications – it can become a threat to the information accessibility, if unauthorized destruction of this information is possible;

– use of unauthorized programs and data $(ZNpd)$ – it creates a threat to the information accessibility in the storage device and during processing in the system, if programs and data are used to destroy information or if they contain malicious code;

– unauthorized access to storage locations of information carriers $(ZNdn)$ – it can lead to a risk of information accessibility, since in this case unauthorized destruction of information recorded on these carriers is possible [14].

In accordance with the presented conditions, a diagram of risk formation in the technogenic system is constructed (Fig. 9).

The information accessibility service supports its readiness for work, allows access to computer systems, data stored in these systems, and programs. This service provides the information transfer between two endpoints or computer systems. It is mainly about the information

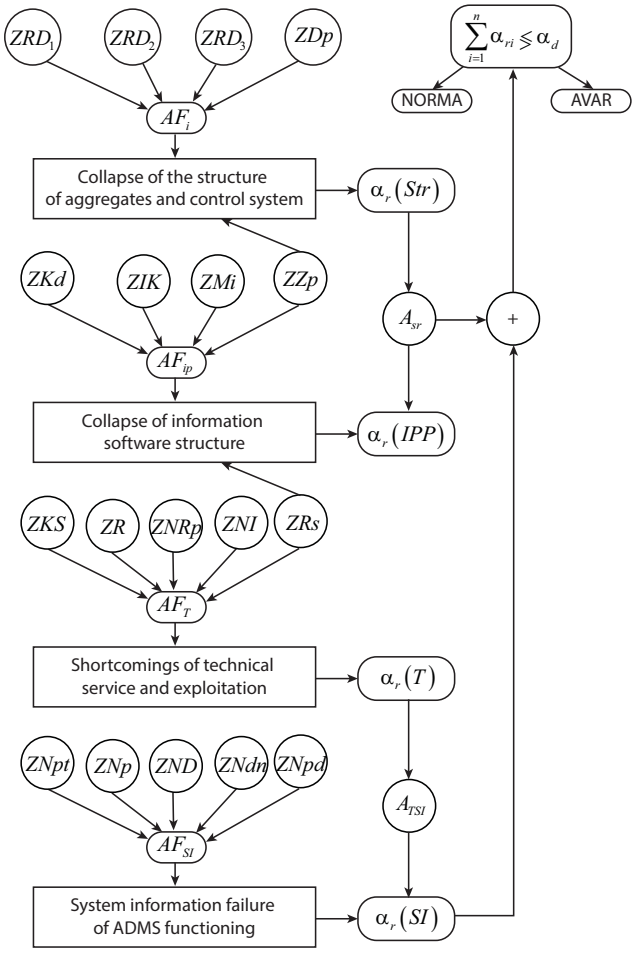


Figure 9 – Diagram of formation of the system accident risk under a complex of threats

presented in electronic form (but also suitable for ordinary documents).

Loss of access to data in ADMS and ACS can lead to the following consequences and the action of threat factors:

– making wrong decisions (F_{nr});

– inability to perform important assigned tasks (F_{nc});

– loss of public trust or lowering of the public image of the organization (F_{vd});

– financial losses (F_{vf});

– liability before the law, including liability for violations of legislation in the field of data protection and nonfulfilment of contracts within the established terms (F_{vz});

– significant costs for restoring (F_{vv}) the system structure, communication channels, and software.

Let one consider the threats of accountability for the hierarchy levels of the cyber technogenic system in the implementation of targeted control tasks.

When protecting accountability, any threat that may lead to the performance of actions that are not characteris-

DOI 10.15588/1607-3274-2023-1-14

tic of this object or entity should be taken into account: collective use of accounts; lack of possibility of operational control of actions; imitation of a legitimate user (masquerade); software failure; unauthorized access to the computer, data, services and applications; unsatisfactory authentication. Such threats typically use disclaimer attacks [12, 14].

The loss of accountability, under the factors impact of the system functioning process failure can lead to the following informational and cognitive consequences:

– manipulation of the system by users (ZF_m);

– deception of personnel at the levels of the system hierarchy (disinformation) (ZF);

– industrial spying (possibility of attacks) (ZF_p);

– uncontrolled actions leading to emergency situations (ZF_d);

– false accusations of incorrect decisions of individuals (ZF_z);

– liability before the law, including liability for violation of legislation in the field of data protection (ZF_v).

Let one consider threats to authenticity in the event of data failure and system disorientation.

Trust in authenticity can be undermined by any threat that causes a person, system, or process to doubt that an object is who it claims to be. Examples of the occurrence of such a situation are the change of data without proper control, the origin of unverified or unsupported data (Fig. 10). [14]

Loss of authenticity can lead to the following consequences in the ADMS system:

– deception of the lower levels $(Va_{1.1})$ and disorientation of the upper levels of the control hierarchy $(Va_{1.2})$;

– use of reliable processes with unreliable data (which can lead to a misleading result) (Va_2);

– manipulation of the organization from the outside – structure $(Va_{3.1})$ and targets $(Va_{3.2})$;

– industrial spying regarding target documents (Va_4);

– false accusations that lead to conflicts in the system (Va_5);

– liability before the law, including liability for violation of national legislation in the field of data protection (Va_6).

According to these factors, category diagrams of the formation of hybrid attacks on the information authenticity are developed for each control system (Fig. 10).

Threats of management data inaccuracies are not necessary for assessing the situation and making control decisions.

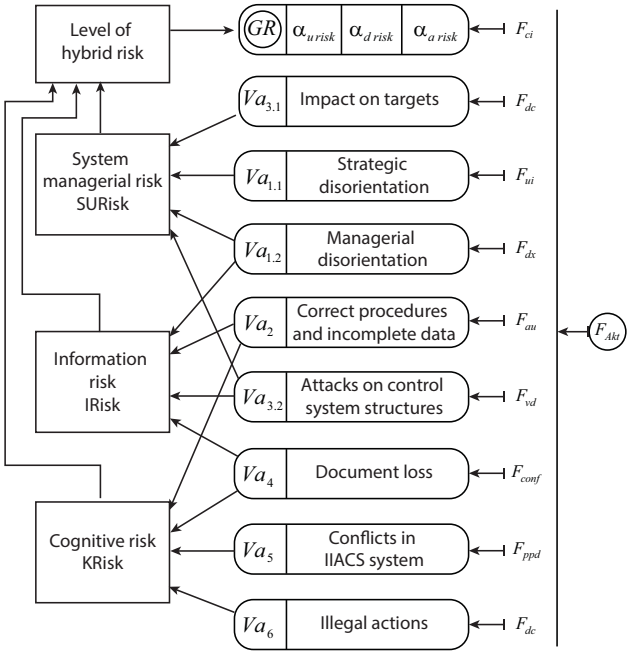


Figure 10 – Category diagrams of the authenticity loss of the control process under internal and external cognitive threats

Any threat that can lead to inconsistent behaviour of systems or processes leads to a decrease in reliability. Examples of such threats are illogical functioning of the system and unreliable suppliers. Decreasing credibility leads to poor customer service and loss of trust.

The loss of reliability can lead to the following consequences in the process of control decision-making under the threat factors:

– deception of the personnel, which leads to a conflict;

– loss of market share due to disorganization;

– decrease in motivation in the work of the organization's personnel, which leads to the emergence of risk situations during control;

– unreliability of suppliers;

– decrease in customer confidence in the service system;

– liability before the law, including liability for violations of legislation in the field of data protection.

Any attack is implemented through the performance of certain actions that disrupt the performance of the protection system and the automated control system as a whole. For a successful attack, an attacker needs to identify a weak spot in the chain of the protection system and, due to the threats and vulnerabilities of the system, make an illegal intervention in its work. At the same time, the main attention of attackers is aimed at security services, which are focused on countering attacks. Therefore, for the reliable functioning of the security system, an important task is to establish the operation of all its security services and analyse the formation of the characteristics of influencing factors and risk components (Fig. 11).

DOI 10.15588/1607-3274-2023-1-14

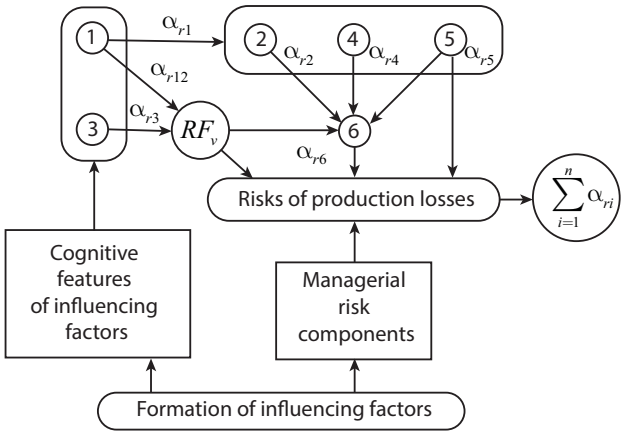


Figure 11 – A complex model for assessing the system failure risk under active threats

5 RESULTS

The analysis of literary sources and the results of research on the behaviour and knowledge level of operational personnel in energy (KMDandA, ACS) at thermal power plants (coal) was carried out on the basis of a system analysis of cognitive engineering psychology, methods of the theory of knowledge. This became the basis for the development of a table of crisis skills, which provide the possibility of selection and formation of operational anti-crisis management teams in the conditions of active threats and attacks of a complex type (Table 2).

Table 2 – Factors and skills

№	Requirements to the activity	Factor	Coefficient α_r
1.	Information processing of images	FI_V	0.1–0.5
2.	Operational actions	FI_{od}	0.1–0.95
3.	To form images of situations	FI_{syt}	0.1–0.5
4.	Factor of target-oriented actions	FCS_u	0.05–0.95
5.	Factor of action tactics generation	FG_{td}	0.1–0.35
6.	Factor of sensory information perception	FSS_i	0.1–0.25
7.	Factor of skills to implement strategy	FR_{str}	0.1–0.9
8.	Ability to master knowledge	$KFIZ_1$	0.1–0.5
9.	Ability to construct models of objects and event scenarios	$KFIZ_2$	0.05–0.3
10.	System target-orientation when the mode is broken	KIZ_3	0.1–0.95
11.	Formation of images of terminal situations	KIZ_4	0.05–0.25
12.	Analysis of the dynamics of events in the system, control modes	KIZ_5	0.05–0.5
13.	Forecast of the consequences of a person's managerial actions	KIZ_6	0.05–0.95
14.	Genetic features of a person's thinking	IKK_9	0.01–0.3
15.	Motivational and will-power ability to make decisions	IKK_M	0.5–0.95
16.	Cognitive stress resistance	SKI	0.1–0.95
17.	The level of system and professional knowledge of the operator	RSP_zI	0.5–0.9
18.	Ability to apply knowledge in crisis situations (creativity)	$FKSit$	0.1–0.95

Tables of this type ensure the construction of effective tests for the selection of personnel for teams of operational and strategic control levels, which are capable of resisting active threats of a high-risk level.

According to Table 2, a category diagram of influencing factors on the formation of possible cognitive risks – staff errors is constructed (Fig. 12).

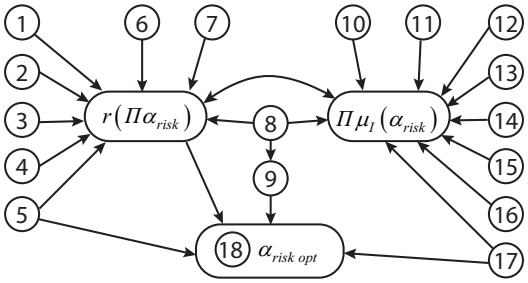


Figure 12 – Category diagram of cognitive risk assessment

6 DISCUSSION

The developed complex model of risk assessment increases the accuracy and objectivity of assessing risk situations, because on the one hand it uses quantitative assessments of the research object (based on data on possible losses in the event of incidents) according to various models and criteria, and on the other hand it uses the experience, knowledge and competencies of experts in the subject area.

The general concept of this approach can be applied not only to companies of the printing industry, but also to any technogenic structures with a hierarchical control structure. Quantitative assessment of losses is presented in the model input obtained according to the analysis of possible threats and vulnerabilities, for example, organizational assets, cognitive characteristics of factors affecting the technogenic structure, taking into account risk coefficients, characteristics of various management components of risk and linguistic considerations of experts-specialists in the field of security and control and decision-making systems. The risks of production losses are formed in the terminal cycle of production and can also be changed under the active threats in the process of production and control. It is necessary to take into account all the threats and vulnerabilities of such technogenic hierarchical systems, and only then it is possible to determine a complex indicator of the system failure risk under active threats. The weight of each constituent indicator of risk can be determined on the basis of the probability approach of the occurrence of one or another event depending on the level of threat to a certain asset, or on the basis of statistical surveys of experts in a specific subject area of research. As a rule, probability approaches coincide with practical statistical results, but unforeseen force majeure circumstances may arise, which, for example, we are currently experiencing during military operations, which destroy all scientific forecasts and conclusions of experts. Assessment and risk analysis of technogenic hierarchical structures was carried out in peacetime, during 2020–2021, for which there are relevant Implementa-

DOI 10.15588/1607-3274-2023-1-14

tion Acts, in particular for the printing production. Based on the received complex assessment of the system malfunction risk, recommendations can be provided for improving the protection system for a certain production and its separate structures, which increases the reliability of the security system.

The developed model of a complex assessment of the system failure risk under active threats does not give a quantitative assessment of the risk, as it determines weak points for the system functioning. In particular, the amount of cognitive risk is difficult to quantify until an incident occurs – a staff or manager error. Therefore, the study substantiated a cognitive diagram for assessing losses in the event of a risk situation with incorrect actions of personnel. The graph of the utility function in decision-making of a decision-maker with expected gains and losses, in case of incorrect actions with different types of behaviour, which affects the cognitive component of the risk level, is presented. In the process of testing with the help of a category diagram *active factor* ↔ *accident risk*, it is possible to reduce the level of cognitive risk, thanks to the increase in the competencies and skills of the personnel.

Obtaining a quantitative risk assessment for the research object has a number of advantages, namely: it combines quantitative (reliable) assessments with the experience, knowledge, and competencies of experts in the subject area; it is based on the definition of the decision-making utility function; the managerial personnel can be trained and decision-making levels can be adjusted according to the method proposed in the research. The suggested category diagrams and a complex risk assessment model can be used not only in the process of designing protection systems, but also in the process of system operation to solve those problems where there is no data for training and where it is necessary to periodically monitor the security system.

The disadvantages of this approach include the fact that the obtained coefficients of cognitive risk depend on the division of the interval [0,1], and their values depend on the competencies of experts in certain subject areas of research. The training of managerial personnel in the security system also depends on this.

CONCLUSIONS

The paper solves the scientific and applied task of developing a complex model for assessing the risk of system failure under active threats, which, on the one hand, uses quantitative assessments of the object, and on the other hand, takes into account the experience, knowledge and competencies of experts in the relevant subject area.

The scientific novelty of the conducted research is as follows:

– for the first time, a cause-and-effect diagram of the event formation during the active action of threat factors and attacks has been developed;

– the interpretation of risk in a technogenic system in the space of states as a change in the trajectory in the system transitions to the limit functioning mode has been improved;

– for the first time, a category diagram of the structure of risk generation under the impact of threat factors and a system-category diagram of interaction in the system *risk* ↔ *emergency-active nature* have been constructed;

– for the first time, a system-category scheme of risk formation in the conditions of active threat factors and a cognitive diagram for assessing losses in the event of a risk situation arising from incorrect actions of personnel have been developed;

– the proposed complex model for assessing the risk of the document management system failure as part of a hierarchical production system for the example of risk assessment of printing productions has been tested and verified, and in addition, a system-category interaction diagram in the game *active factor* ↔ *accident risk* has been suggested.

The practical significance of the obtained results is that the proposed model of complex risk assessment has been tested in the document management system as part of the hierarchical system of printing production and can be used in various technogenic hierarchical systems when solving managerial decision-making tasks, designing and improving protection systems.

The further research of the problem can be seen in the development of software for assessing the risk of system functioning under active threats to technogenic hierarchical structures.

ACKNOWLEDGEMENTS

The work has been performed within the framework of the state-budget research themes of Ukrainian Academy of Printing: “Development of information technologies for the protection of electronic resources in automated document management systems”; Uzhhorod National University: “Methods and means of software engineering for implementing big data analytics processes based on information and technical platforms of electronic science”.

The authors would also like to express their gratitude their partners Aquaflex Plus LLC, Prime-Pak LLC, Sofi Company LLC for the data provided for the implementation of the experimental part of the study.

REFERENCES

1. Shurygin A. M. Applied stochastics: robustness, estimation, forecast. Moscow, Finance and statistics, 2000, 224 p.
2. Kavun S. V., Nosov V. V., Manzhai O. V. Information security. Tutorial. Kharkiv, PH. KhNEU, 2008, 352 p.
3. Veretilnyk T. I., Mysnyk L. D., Mysnyk B. V., Kapitan R. B. Organization of publishing and printing activities: Tutorial Cherkasy. Cherkasy, State Technology University, 2020, 157 p. [Electronic resource] https://er.chdtu.edu.ua/bitstream/ChSTU/3380/1/ORGANIZATION OF%20POLIGRAPHIC%20ACTIVITY.pdf
4. Kovaleva V. V., Samarin Yu. N. Selection of management system for a printing company, *CompuArt. Journal for printers and publishers*, 2007, No. 11, pp. 61–64.

DOI 10.15588/1607-3274-2023-1-14

5. Honcharov S. V. Financial security of the securities market of Ukraine. Poltava, Poltava State Agrarian Academy, 2019, pp. 40–42.
6. Schneier Bruce. Applied cryptography. Protocols, algorithms, source texts in C language. 2nd edition. Moscow, Triumf, 2002, 816 p.
7. Michael S., Andrew H. Practical Malware Analysis: The Hands – On Guide to Dissecting Malicious Software; translated from English. Chernikov S., St. Petersburg, 2018, 786 p.
8. Koval L. H., Zlepko S. M., Novitskyi H. M., Krekoten E. H. Methods and technologies of biometric identification according to the results of literary sources, *Scientific notes of TNU named after V.I. Vernadskyi*. Vinnytsia, VNTU, 2019, Vol. 30 (69), Part 1, No. 2, pp. 104–112. [Electronic resource] https://www.tech.vernadskyjournals.in.ua/journals/2019/2_2019/part_1/19.pdf.
9. Law of Ukraine "On electronic digital signature", *Bulletin of the Verkhovna Rada*, 2003, No. 36, P. 276.
10. Schneider B. Secrets and Lies: Digital Security in a Networked World. New-York, WCP, 2002, 368 p.
11. Senkivskyi V. M., Petyak Y. F., Kozak R. O., Lytovchenko O. V. Information technology for effective data protection of publishing systems on mobile devices. Lviv, UAP, 2020, 272 p.
12. Bobalo Y. Ya., Horbaty I. V., Bondarev A. P. Information security. Lviv, Lviv Polytechnic University, 2019, 580 p.
13. Durnyak B. V., Sabat V. I., Shvedova L. E. Authority control in information protection systems. Lviv, UAP, 2016, 148 p.
14. Sabat V. Sikora L., Durnyak B., Lysa N., Fedevych O. Information technologies of active control of complex hierarchical systems under threats and information attacks, *The 3rd International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS-2022)*. Khmelnytskyi, Ukraine, May 25–27, 2022. https://ceur-ws.org/Vol-3156/paper23.pdf
15. Kelemen M., Polishchuk V., Gavurová B., Andoga R., Szabo S., Yang W., Christodoulakis J., Gera M., Kozuba J., Kaľavský P., Antoško M. Educational Model for Evaluation of Airport NIS Security for Safe and Sustainable Air Transport. Sustainability, 2020, 12, 6352. https://doi.org/10.3390/su12166352.
16. Milioti Christina, Kepaptsoglou Konstantinos, Deloukas Alexandros, Apostolopoulou Efthymia Valuation of man-made incident risk perception in public transport: The case of the Athens metro, *International Journal of Transportation Science and Technology*, 2022, Vol. 11, pp. 578–588. https://doi.org/10.1016/j.ijtst.2021.07.003.
17. Sicard F., Zamai É., Flaus J. M. An approach based on behavioral models and critical states distance notion for improving cybersecurity of industrial control systems, *Reliab Eng Syst Saf*, 2019, Vol. 188, pp. 584–603. 10.1016/J.RESS.2019.03.020
18. Cormier A., Ng C. Integrating cybersecurity in hazard and risk analyses, *J Loss Prev Process Ind*, 2020, Vol. 64. Article 104044, 10.1016/j.jlp.2020.104044
19. Schmittner C., Gruber T., Puschner P., Schoitsch E. Security application of Failure Mode and Effect Analysis (FMEA), *Computer safety, reliability, and security*. Springer International Publishing, Cham, 2014, pp. 310–325.
20. Vessels L., Heffner K., Johnson D. Cybersecurity risk assessment for space systems, *2019 IEEE Space Comput Conf. (SCC)*, 2019, pp . 11–19. 10.1109/SpaceComp.2019.00006
21. Domeh Vindex, Obeng Francis, Khan Faisal, Bose Neil, Sanli Elizabeth Risk analysis of man overboard scenario in a small fishing vessel, *Ocean Engineering*, 2021, Vol. 229, Article 108979. https://doi.org/10.1016/j.oceaneng.2021.108979.
22. Alanen Jarmo, Linnosmaa Joonas, Malm Timo, Papakonstantinou Nikolaos, Ahonen Toni, Heikkilä Eetu, Tiusanen Risto Hybrid ontology for safety, security, and dependability risk assessments and Security Threat Analysis (STA) method for industrial control systems, *Reliability Engineering & System Safety*, 2022, Vol. 220, Article 108270. https://doi.org/10.1016/j.ress.2021.108270.
23. Agrawal V. A. Comparative study on information security risk analysis methods, *J Comput (Taipei)*, 2017, pp. 57–67. 10.17706/jcp.12.1.57-67
24. Arbanas K., Čubrilo M. Ontology in information security, *J Inf Org Sci*, 2015, Vol. 39, pp. 107–136.
25. Blanco C. Lasheras J. , Fernández-Medina E. , Valencia-García R. , Toval A. Basis for an integrated security ontology according to a systematic review of existing proposals, *Comput Stand Interfaces*, 2011, Vol. 33, pp. 372–388.
26. Zhou T., Modarres M., Droguett E. L. Multi-unit nuclear power plant probabilistic risk assessment: a comprehensive survey, *Reliab Eng Syst Saf*, 2021, Vol. 213. Article 107782. 10.1016/J.RESS.2021.107782
27. Modarres M., Zhou T., Massoud M. Advances in multi-unit nuclear power plant probabilistic risk assessment, *Reliab Eng Syst Saf*, 2017, Vol. 157, pp. 87–100. 10.1016/J.RESS.2016.08.005
28. Kim J., Shah A.U.A., Kang H.G. Dynamic risk assessment with bayesian network and clustering analysis, *Reliab Eng Syst Saf*, 2020, Vol. 201, Article 106959, 10.1016/J.RESS.2020.106959
29. DeJesus Segarra J., Bensi M., Modarres M. A bayesian network approach for modeling dependent seismic failures in a nuclear power plant probabilistic risk assessment, *Reliab Eng Syst Saf*, 2021, Vol. 213, Article 107678. 10.1016/J.RESS.2021.107678
30. Rabcan J., Levashenko V., Zaitseva E., Kvassay M., Subbotin S. Application of Fuzzy Decision Tree for Signal Classification, *IEEE Transactions on Industrial*, 2019, No. 15(10), pp. 5425–5434. https://doi.org/10.1109/TII.2019.2904845
31. Rabcan J., Levashenko V., Zaitseva E., Kvassay M., Subbotin S. Non-destructive diagnostic of aircraft engine blades by Fuzzy Decision Tree, *Engineering Structures*, 2019, No. 197, P. 109396. https://doi.org/10.1016/j.engstruct.2019.109396

Received 06.01.2023.
Accepted 11.02.2023.

DOI 10.15588/1607-3274-2023-1-14

УДК 004.451.36:681.5

АНАЛІЗ ТЕРМІНАЛЬНИХ ПОТОКІВ РИЗИКІВ У ТЕХНОГЕННИХ СИСТЕМАХ, ЯКІ ВИНИКАЮТЬ В ПРОЦЕСІ ВПЛИВУ ЗАГРОЗ

Сабат В. І. – канд. техн. наук, доцент, доцент кафедри інформаційних мультимедійних технологій Української академії друкарства, Львів, Україна.

Сікора Л. С. – д-р техн. наук, професор, дійсний член Інженерної Академії України, професор кафедри автоматизованих систем управління Інституту комп'ютерних наук та інформаційних технологій, Львів, Україна.

Дурняк Б. В. – д-р. техн. наук, професор, заслужений діяч науки і техніки України, ректор Української академії друкарства, Львів, Україна.

Повхан І.Ф. – д-р техн. наук, професор, декан факультету інформаційних технологій ДВНЗ «Ужгородський національний університет», м. Ужгород, Україна.

Поліщук В.В. – д-р техн. наук, доцент, професор кафедри програмного забезпечення систем ДВНЗ «Ужгородський національний університет», Ужгород, Україна.

АНОТАЦІЯ

Актуальність. Проведено аналіз термінальних потоків ризиків в техногенних системах, які виникають в процесі впливу інформаційних і когнітивних загроз в автоматизованій системі управління та документообігу в складі ієрархічної системи виробництва.

Об'єктом дослідження є процес функціонування складних систем з ієрархічною структурою, в яких для інформаційного забезпечення якісного управління технологічними процесами використовуються автоматизовані системи документообігу з високим рівнем захисту потоків даних для прийняття рішень.

Предметом дослідження є методи та засоби побудови системи захисту інформації для забезпечення надійного функціонування автоматизованих систем документообігу та прийняття цільових рішень в ієрархічних структурах з мінімальним ризиком впливу зовнішніх загроз і атак.

Метою даної роботи є розроблення комплексної моделі оцінки ризику збою системи управління та документообігу в складі ієрархічної системи виробництва при дії активних загроз.

Метод. Вперше обгрунтовано і розроблено причинно-наслідкову діаграму формування події при активній дії факторів загроз і атак, визначено трактування ризику в техногенній системі та представлено ризик у просторі станів як зміну траєкторії при переході системи в граничний режим функціонування. Вперше побудовано категорну діаграму структури породження ризиків при дії факторів загроз та системно-категорну діаграму взаємодії в системі *ризик* ↔ *аварійно-активний характер*, запропоновано системно-категорну схему формування ризиків в умовах дії активних факторів загроз. Вперше обгрунтовано когнітивну діаграму для оцінки втрат при виникненні ризикової ситуації при некоректних діях персоналу.

Результати. В результаті досліджень побудовано системно-категорну діаграму впливу комплексу загроз на режим і процес функціонування системи, розроблено метод обчислення рівня системної стратегічної безпеки енергоактивних ієрархічних систем в процесі дії атак і загроз та запропоновано комплексну модель оцінки ризику збою функціонування системи при дії активних загроз.

Висновки. При дії активних завад, когнітивних і системних факторів на оперативному та стратегічному рівнях ієрархії управління із-за неправильних рішень та інформаційної дезорієнтації виникають аварійні ситуації та ризики втрати функцій системи і її цілеорієнтованості. Аналіз комплексу ризиків і запропонована категорна діаграма структури породження ризиків при дії факторів загроз, лягли в основу розроблення ймовірніснісної структури поняття ризику на підставі моделі *атака* ↔ *наслідок*, а також побудови системно-категорної діаграми взаємодії в грі *активний фактор* ↔ *ризик аварії*. Це, в свою чергу, дало можливість побудови системно-категорної схеми формування термінальних потоків ризиків в техногенних системах, які виникають в процесі впливу загроз. Комплексна модель оцінки ризику збою системи при дії загроз може бути використана для побудови систем захисту для будь-яких ієрархічних структур управління техногенними системами.

КЛЮЧОВІ СЛОВА: техногенні системи, загрози, вразливості, оцінка ризику, прийняття рішень, управління ієрархічними системами.

ЛІТЕРАТУРА

1. Shurygin A. M. Applied stochastics: robustness, estimation, forecast / A. M. Shurygin. – M. : Finance and statistics, 2000. – 224 p.
2. Kavun S. V. Information security. Tutorial / S. V. Kavun, V. V. Nosov, O. V. Manzhai. – Kharkiv : PH. KhNEU, 2008. – 352 p.
3. Organization of publishing and printing activities: Tutorial Cherkasy / [T. I. Veretilnyk, L. D. Mysnyk, B. V. Mysnyk, R. B. Kapitan]. – Cherkasy : State Technology University, 2020. – 157 p. [Electronic resource] https://er.chdtu.edu.ua/bitstream/ChSTU/3380/1/ORGANIZATION OF%20POLIGRAPHIC%20ACTIVITY.pdf
4. Kovaleva V. V. Selection of management system for a printing company / V. V. Kovaleva, Yu. N. Samarin // Computer-Art. Journal for printers and publishers. – 2007. – No. 11. – P. 61–64.
5. Honcharov S. V. Financial security of the securities market of Ukraine / S. V. Honcharov. – Poltava : Poltava State Agrarian Academy, 2019. – P. 40–42.
6. Schneier Bruce. Applied cryptography. Protocols, algorithms, source texts in C language. 2nd edition. – M. : Triumf, 2002. – 816 p.
7. Michael S. Practical Malware Analysis: The Hands – On Guide to Dissecting Malicious Software / S. Michael, H. Andrew ; translated from English. Chernikov S. – St. Petersburg, 2018. – 786 p.
8. Koval L. H. Methods and technologies of biometric identification according to the results of literary sources / [L. H. Koval, S. M. Zlepko, H. M. Novitskyi, E. H. Krekoten] // Scientific notes of TNU named after V. I. Vernadskyi.

DOI 10.15588/1607-3274-2023-1-14

– Vinnytsia : VNTU, 2019. – Vol. 30 (69) – Part 1. – No. 2. – P. 104–112. [Electronic resource] https://www.tech.vernadskyjournals.in.ua/journals/2019/2_2019/part_1/19.pdf.
9. Law of Ukraine "On electronic digital signature" // Bulletin of the Verkhovna Rada, 2003. – No. 36. – P. 276.
10. Schneider B. Secrets and Lies: Digital Security in a Networked World / B. Schneider. – New-York : WCP, 2002. – 368 p.
11. Information technology for effective data protection of publishing systems on mobile devices / [V. M. Senkivskyi, Y. F. Petyak, R. O. Kozak, O. V. Lytovchenko]. – Lviv : UAP, 2020. – 272 p.
12. Bobalo Y. Ya. Information security / Y. Ya. Bobalo, I. V. Horbaty, A. P. Bondarev/ – Lviv : Lviv Polytechnic University, 2019. – 580 p.
13. Durnyak B. V. Authority control in information protection systems / B. V. Durnyak, V. I. Sabat, L. E. Shvedova. – Lviv : UAP, 2016. – 148 p.
14. Information technologies of active control of complex hierarchical systems under threats and information attacks / [V. Sabat, L. Sikora, B. Durnyak et al.] // The 3rd International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS-2022) Khmelnytskyi, Ukraine, May 25–27, 2022. https://ceur-ws.org/Vol-3156/paper23.pdf
15. Kelemen, M. Educational Model for Evaluation of Airport NIS Security for Safe and Sustainable Air Transport. Sustainability / [M. Kelemen, V. Polishchuk, B. Gavurová et al.]. – 2020. – 12. – 6352. https://doi.org/10.3390/su12166352.
16. Valuation of man-made incident risk perception in public transport: The case of the Athens metro / [Christina Milioti, Konstantinos Kepaptsoglou, Alexandros Deloukas, Efthymia Apostolopoulou] // International Journal of Transportation Science and Technology. – 2022. – Vol. 11. – P. 578–588. https://doi.org/10.1016/j.ijtst.2021.07.003.
17. Sicard F. An approach based on behavioral models and critical states distance notion for improving cybersecurity of industrial control systems / F. Sicard, É. Zamai, J. M. Flaus // Reliab Eng Syst Saf, 2019. – Vol. 188. – P. 584–603. 10.1016/J.RESS.2019.03.020
18. Cormier A. Integrating cybersecurity in hazard and risk analyses / A. Cormier, C. Ng. // J Loss Prev Process Ind, 2020. – Vol. 64. – Article 104044, 10.1016/j.jlp.2020.104044
19. Security application of Failure Mode and Effect Analysis (FMEA) / [C. Schmittner, T. Gruber, P. Puschner, E. Schoitsch] // Computer safety, reliability, and security. – Springer International Publishing, Cham, 2014. – P. 310–325.
20. Vessels L. Cybersecurity risk assessment for space systems / K. Heffner, D. Johnson // 2019 IEEE Space Comput Conf. (SCC), 2019. – P. 11–19. 10.1109/SpaceComp.2019.00006
21. Domeh Vindex. Risk analysis of man overboard scenario in a small fishing vessel / [Vindex Domeh, Francis Obeng, Faisal Khan et al.] // Ocean Engineering. – 2021. – Vol. 229. – Article 108979. https://doi.org/10.1016/j.oceaneng.2021.108979.
22. Hybrid ontology for safety, security, and dependability risk assessments and Security Threat Analysis (STA) method for industrial control systems / [Jarmo Alanen, Joonas Linnosmaa, Timo Malm et al.] // Reliability Engineering & System Safety. – 2022. – Vol. 220. – Article 108270. https://doi.org/10.1016/j.ress.2021.108270.
23. Agrawal V. A. Comparative study on information security risk analysis methods / V. A. Agrawal // J Comput (Taipei). – 2017. – P. 57–67. 10.17706/jcp.12.1.57-67
24. Arbanas K. Ontology in information security / K. Arbanas, M. Čubrilo // J Inf Org Sci, 2015. – Vol. 39. – P. 107–136.
25. Basis for an integrated security ontology according to a systematic review of existing proposals / [C. Blanco, J. Lasheras, E. Fernández-Medina et al.] // Comput Stand Interfaces. – 2011. – Vol. 33 – P. 372–388.
26. Zhou T. Multi-unit nuclear power plant probabilistic risk assessment: a comprehensive survey / T. Zhou, M. Modarres, E. L. Droguett // Reliab Eng Syst Saf. – 2021. – Vol. 213. – Article 107782. 10.1016/J.RESS.2021.107782
27. Modarres M. Advances in multi-unit nuclear power plant probabilistic risk assessment / M. Modarres, T. Zhou, M. Massoud // Reliab Eng Syst Saf. – 2017. –Vol. 157. – P. 87–100. 10.1016/J.RESS.2016.08.005
28. Kim J. Dynamic risk assessment with bayesian network and clustering analysis / J. Kim, A.U.A. Shah, H.G. Kang // Reliab Eng Syst Saf. – 2020. – Vol. 201. – Article 106959, 10.1016/J.RESS.2020.106959
29. DeJesus Segarra J. A bayesian network approach for modeling dependent seismic failures in a nuclear power plant probabilistic risk assessment / J. DeJesus Segarra, M. Bensi, M. Modarres // Reliab Eng Syst Saf. – 2021. – Vol. 213 – Article 107678. 10.1016/J.RESS.2021.107678
30. Application of Fuzzy Decision Tree for Signal Classification // [J. Rabcan, V. Levashenko, E. Zaitseva et al.] // IEEE Transactions on Industrial. – 2019. – No. 15(10). – P. 5425–5434. https://doi.org/10.1109/TII.2019.2904845
31. Non-destructive diagnostic of aircraft engine blades by Fuzzy Decision Tree // [Rabcan J., Levashenko V., Zaitseva E. et al.] // Engineering Structures. – 2019. – No. 197. – P. 109396. https://doi.org/10.1016/j.engstruct.2019.109396

DOI 10.15588/1607-3274-2023-1-14

Наукове видання

Радіоелектроніка,

інформатика,

управління

№ 1/2023

Науковий журнал

Головний редактор – д-р техн. наук С. О. Субботін

Заст. головного редактора – д-р техн. наук Д. М. Піза

Комп’ютерне моделювання та верстання — С. В. Зуб
Редактор англійських текстів — С. О. Субботін

Оригінал-макет підготовлено у редакційно-видавничому відділі НУ «Запорізька політехніка»

Свідоцтво про державну реєстрацію
КВ № 24220-14060 ПР від 19.11.2019.

Підписано до друку 17.02.2023. Формат 60×84/8.
Папір офс. Різогр. друк. Ум. друк. арк. 19,76.
Тираж 300 прим. Зам. № 41.

69063, м. Запоріжжя, НУ «Запорізька політехніка», друкарня, вул. Жуковського, 64

Свідоцтво суб’єкта видавничої справи
ДК № 6952 від 22.10.2019.