Institute of Information and Computational Technologies

ISSN : 2788-7677 (Online)
ISSN : 2788-7987 (Print)

ADVANCED TECHNOLOGIES AND COMPUTER SCIENCE

2021
№3

www.atcs.iict.kz

Institute of Information and Computational Technologies

Advanced Technologies and computer science

№3

Almaty 2021

ISSN: 2788-7677 (Online)
ISSN : 2788-7987 (Print)

Institute of Information and Computational Technologies,

Advanced Technologies and computer science

28 Shevchenko str., Almaty,Republic of Kazakhstan
7 (727) 272-37-11
atcs@iict.kz

About the Journal

Advance technologies and computer science is a bilingual scientific peer-reviewed, interdisciplinary, electronic journal of open access, including thematic areas:

- Section **"Applied mathematics, computer science and control theory"** includes papers describing modern problems in these areas.
- Section **"Information and telecommunication technologies"** also includes the following topics:
    - Data transmission systems and networks.
    - Internet technologies.
    - Cloud technologies.
    - Parallel computing.
    - Distributed computing.
    - Supercomputer and cluster systems.
    - Big data processing (Big-data).
    - Geographic Information Systems and Technologies.
- In the section **"Artificial intelligence technologies"** in addition to technology, there are works on topics:
    - Intelligent Management Systems.
    - Speech technology and computer linguistics.
    - Pattern Recognition and Image Processing.
    - Bioinformatics and biometric systems.
    - Human-machine interaction.
    - Machine learning.
    - Intelligent Robotic Systems.
- The section **"Information Security and Data Protection"** also covers topics:
    - Software and hardware information protection.
    - Mathematical methods for ensuring information security of complex systems.
- The section **"Modeling and optimization of complex systems and business processes"** may include:
    - Computational mathematics, numerical analysis and programming, mathematical logic.
    - Theory of Statistics.
    - Statistical Methods.

Contents

UDC 004.02

IRSTI 20.53.15

FUNCTIONAL HYPOTHESIS OF COMPLEXITY CLASSES

Mirzakhmet Syzdykov

al-Farabi Kazakh National University, Almaty, Kazakhstan
mspmail598@gmail.com
ORCID ID: https://orcid.org/0000-0002-8086-775X

Abstract. This work describes the hypothesis of the relation between the classes of complexity: for this purpose we define the functions over algorithms or state machines for which the equality holds true and, thus, the decision can be made towards polynomial reduction of the computational complexity of algorithms. The specific class of impractical or exponential measures of complexity against the polynomial ones is also discussed – for this case we divide these classes according to the discrete numbers which are known to the present time. We also present the approximate algorithm for the classical NP-complete problem like Traveling Salesman using the memory construction. The question of P and NP equality is important in decision-making algorithms which commonly decide inequality of these classes – we define the memory factor which is exponential and space consumption is non-deterministic. The memory consumption problem within the memorization principle or dynamic programming can be of varying nature giving us the decision to build the approximation methods like it's shown on the example of Traveling Salesman problem. We also give the notion of the past work in theory of complexity which, in our opinion, is of the same consideration in most cases when the functional part is omitted or even isn't taken into account. The model theorem with its proof of the equality of classes over congruent function is also given in the end of this article.

Keywords: computational complexity, algorithms, equality.

Intoduction

The problem of deciding whether the polynomial class of finite automata can accept the solution in the same time as finding the optimal one [1]:

$$O(A_P) = ? \; O(A_{NP}), \tag{1}$$

where A_P is the polynomial algorithm and A_{NP} is for general case when the algorithm can be of non-polynomial type and, thus, is impractical.

The "P versus NP" problem is an old questions discussed by several communications in [2, 3, 4]. The more philosophical outcome of the problem is presented in [5].

For the past time Karp's 21 NP-complete problems were observed as a bundle with common prooperties [6]. However, in this work we define the hypothesis over complexity classes using big-O notation.

For the practical algorithms we define the class of P-complete complexity as follows:

$$O(A_P) = \{1, N, N^2, .., N^{T.}\}, \tag{2}$$

where N is the input parameter of the *size* of the problem and T is a free parameter which is adequate to the computational environment which is presented more generally by Cook as a classical Turing tape automaton. As per our outcome, we use the definition of the algorithm as tuple:

$$A = <I, O, P> \tag{3}$$

where I is a set of input parameters, O is a set of output parameters and P stands for a procedural routine of the algorithm. Obviously the size of the problem to be computed is included in the set of inputs to the algorithm.

As our algorithm can be polynomial and, thus, complying to the computational device, the other side of NP-complexity is its classification as exponential like powerset and factorial or even Ackermann number [7]:

$$O(A_{NP}) = \{ 2^N, N! \sim N^N, \dots \} \quad (4).$$

Thus, the problem is either P- or NP-complete if:

$$A) = O(A_P)\text{: P-complete; } O(A) = O(A_{NP})\text{: NP-complete,} \quad (5)$$

where A is the general algorithm which is actually the union of the accepting algorithm in P and computing in NP:

$$A = \{ A_P, A_{NP} \} \quad (6).$$

Functional Hypothesis

The pre-historical turn of our hypothesis is first met in [8] for the subset construction [9] when the functor of counter state permits the equality:

$$F(A_P) = A_{NP}, \quad (7)$$

where **F(x)** is state functor.

We state that NP-complete for problem over accepting algorithm A_P and computing algorithm A_{NP} there exist function **f** and **g** so that the following equality holds true:

$$f(A_P) = g(A_{NP}) : O(A_P) \sim O(A_{NP}), \quad (8)$$

Thus, if these functions exist we may conclude that P ≠ NP or P = NP otherwise. The decision depends on the type of the functions where as of the set (2) and (4).

For the equation (8) we can define the limits over the size of the problem if it's NP-complete, when the input parameter N is almost infinite.

For practical reasons we will consider the classical NP-complete Traveling Salesman Problem (TSP) when there exist Bellman's function [10]:

$$u_i(t) = sup \{ u_{i-1}(t - t_0) + f_i(t_0) \}, t, t_0 \text{ in } R \quad (9)$$

Our approach uses the incremental method of cities inclusion and, thus, the function (9) for TSP will be as follows:

$$F(t, n) = sup \{ F(s, n - 1) + distance(s, t) \}, \quad (10)$$

where **s** and **t** are cities and **n** is the number of the passed cities to the current step of algorithm.

The equation (10) is also known as Nearest Neighbor Heuristics [11] and is polynomial in the number of steps required for the algorithm.

However this is not principal where to find exact solution, so there is an approximate polynomial algorithm for TSP which uses ant colony optimization [12] and exact which uses exponential memory [13].

Our construction uses also exponential time and is based upon the building of network in Figure 1. Each time we visit the mediate layer we memorize the path as a context for this state and push it back to the stack of the candidates with the value equal to the path in source graph between two nodes. This allows us to adjust the maximum size of the candidate stack and, thus, get an approximate solution within the predefined memory consumption which isn't allowed in algorithm Held-Karp's algorithm [13].

We state that memory consumption is vital in memorized computations of NP-complete problem, this case is partially described in [14] where the hardware architecture design is proposed for effective computations.

The memorized computations, thus, gives us the decision to find the solution over the limited set.

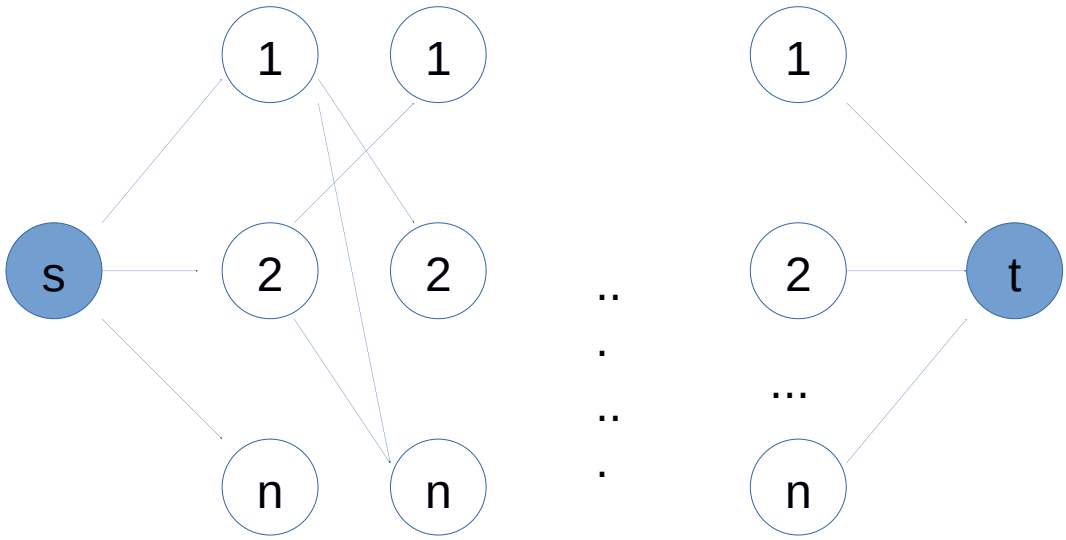

Figure 1 – TSP network for approximate solution

As to our hypothesis the following statement always holds true:

$$P \to f(x) \to NP, \tag{11}$$

where in (11) the P and NP are classes of complexity and f(x) is an arbitrary function according to the fact of congruence of the functor f*(x)* in counter state in the subset construction of the modified non-deterministic finite automaton, NFA, to the deterministic finite automaton, DFA, in [8].

This idea was first devised from the proof of existence of conversion between NFA and DFA made by Rabin and Scott in late 50's [9], who were also awarded the Turing Award for their proof of concept.

Thus, our main theorem based upon the fact that the NFA-counter in AND-construction algorithm is for the function f(x) in (11) and *P* and *NP* are incoming NFA and resulting DFA similarly, can be called as the theorem about congruence relations between P and NP-classes, where as the following inequality holds true, as the P-class is rather smaller than NP-class:

$$P \ll NP \tag{12}$$

The equation (12) is practically the common case when we state that *P* is in *NP*, however, due to decision-making fact we state that:

$$f(P) = NP \tag{13}$$

The equation (13) is according to the modified subset construction in our algorithm presented in [8]. We say that the P- and NP-classes are decidable if there exists another inverse function so that the following holds true:

$$f^{-1}(NP) = P \tag{14}$$

The above equation devised from the equation (13) gives the prominent fact about decidability of NP-classes. Thus, we have to conclude by induction that the equation (15) holds true if we decide that our NP-complete problem is of "over" NP-class, meaning that the it's too big of almost Ackermann numbers in function [7].

$$P = NP \tag{15}$$

The facts (14) and (15) can be proved due to the fact of the congruence of the function f(x) along the P-class and over to NP-class, seen in the main result like in equation (11).

Thus, according to our proofs of the decision against P- and NP-classes we can write our conclusion in the following form:

$$\begin{cases} P \neq NP : P \ll NP \\ P = NP : \exists F^{-1}(x) \end{cases} \tag{16}$$

In parallel computing there is the following hypothesis according to the following equation:

$$\lim_{N\to\infty} \frac{NP}{N} = P \quad (17)$$

where the *N* is the number of concurrent processes working on the parallel machine: the author supposes that with some type of generalization this could be even Turing machines or any other.

Model theorem

The problem of equality of P- and NP-classes of complexity is a Millennium Prize Theorem as proposed by Clay Mathematics Institute. Besides this fact, thisis question when dealing with decision-making in solving problems like, for example,sorting. We will show that it's solvable by back-reference matching in regular expression. Our further material is based upon the regular expression and automata theory.

Let's define the P- and NP-classes as follows:

$$P = \{ N: O(N) > o(N) \}$$

$$NP = \{ N! : O(N!) > o(N!) \}, \quad (18)$$

for example, if we deal with factorial classes of complexity.

Further we will state the model theorem around the congruence of P- and NP-classes, these theoremsborrowed from the original source which was developed by an author in late 2015 [8]. This theorem is model, further we show by proving by induction that P != NP basing upon existence of such model classes in finite automata theory.

In [8] the author deals with the problem of matching against two expressions on (non-)deterministicfinite automata for AND-operator. This is logical operator meaning the intersection of languages.

For our model theorem let's define some assumptions:

Definition 1. The non-deterministic finite automata, NFA, is a tuple like:

$$NFA = <A, S, Start(S), Finish(S), S \times (A + \{epsilon\}) \times S>, \quad (19)$$

where A is a set of alphabet, S is a set of states, Start(S) and Finish(S) is a set of starting and final states,epsilon stands for an empty sign and S x (A + {epsilon}) x S is a set of transitions.

Thus starting from state from set Start(S) we match against the word on a finite tape (probably infinite) until reaching the state from set Finish(S) by the transitional reach on the transitions connecting states by the marks, some of which could be empty.

Later we will show the congruence of non-deterministic and deterministic automataover subset construction, thus proving the properties of DFA as to be atomic in some sense. Thus, the deterministic finite automata is tuple like:

Definition 2. Deterministic finite automata (DFA)

$$DFA[1] = <A, S, Starting, Finish(S), S \times A \times S>, \quad (20)$$

where A is a set of alphabet symbols,S is a set of states, Starting is a starting state from set S, Finish(S) is a sub-set of S defining the finishing points on automaton, S x A x S is a set of transitions (please note, without epsilon like in NFA).

Thus, besides like matching in NFA, matching in DFA is linear, starting from single source point and going through all the elementary transitions for a single alphabet symbol until reaching the finalstate marked by the Finish(S)-set.

In elementary steps of decision the matching by DFA is rather process of function devising than, like in NFA, full simulation:

F(Starting, x[1]) = S[2];
F(S[2], x[2]) = S[3];
…
F(S[n], x[n]) = S[n + 1] (iff in Finish(S), then automata accepts the word x[1]...x[n]).

Full simulation of NFA plays a vital role in making rounds to detect if the word lays in the final accepting state, same as in case of DFA.

Now as we have defined the main clusters of the model theorem, we have to show the congruenceof P- and NP-classes around our problem of matching against AND-operator as it was stated previously in [8].

For the languages L(r), where r is the regular expression, we define the new operation along with allothers like intersection of languages:

L(r[1]*r[2]) = {r[1] r[2]},

L(r[1] | r[2]) = L(r[1]) + L(r[2]),

L(r*) = L(epsilon) + L(r) + L(r*r...), L(epsilon) = {epsilon},

L(r[1] & r[2]) = L(r[1]) x L(r[2]) – the new operator by intersection of languages for regular expressions.

The building of NFA for L(r[1] & r[2]) is same in [8] as for pairing entities despite of that fact thatwe define the counter, which is a function in model theorem. For this purpose we define an extended NFA:

$$NFA\ Ext[1] = <A, S, Start(S), Finish(S), S \times A \times \{1, 2\} \times S>, \quad (21)$$

where A is a set of alphabet,S is a set of states, Start(S) and Finish(S) are starting and finishing states, S x A x {1, 2} x S is a set of transitions along the counters until the counter becomes active when simulating the matching of word on an infinite tape.

We can prove that along with subset construction [9] that our simulation leads to the construction ofan atomic DFA if we would simulate these as per subset construction algorithm.

Thus the extended DFA is a subset of an extended NFA within the AND-operator along the counterfunction as a regulator of simulating processes within a subset construction:

$$DFA\ Ext[1] = Subset\ Construction\ (NFA\ Ext[1], Counters = \{1, 2\}). \quad (22)$$

As we get to the NFA state during the subset construction it becomes active along the counter valuewhich is polar to the number of incoming edges to that point.

Let us reverse to our model theorem, while we have composited main entities for better evaluation of the theorem. Basically, the theorem is based upon that fact that there exist an algorithm (in our case this is a modified by a counters subset construction) so that there would exist an in-equality ofP- and NP-classes of complexity (P != NP). According to our subset construction there exist a counter-function along which the class of in-equality P standing for non-deterministic finite automaton NFA is raised to the class of NP standing for deterministic finite automaton DFA:

[NFA, P – Subset Construction(DFA, NP),

[NFA, P – Subset Construction(DFA, NP, f(x) = {1, 2}).

Thus according to the fact that there exists such f(x), so that:

Algorithm (modified subset construction):

$$f(NFA) = DFA, \quad (23)$$

where f(x) = {1, 2) is a counter-function in modified subset construction forintersection operator in regular expression.

As there exist such example this barely proves that:

$$f(P) = NP, \text{ and } P \mathrel{!}= NP. \quad (24)$$

In the next section we will show by induction that our model theorem states not only the barely fact decision of proving outcome, but rather the upper and lower-bounds for the complexities of P and NP.

Proof

Answering the Steven Cook's question stated by him in [1], we now can decide that function f(x) matter the set of entities along which the solution of the stated problem could be verified:

T = { f(x): over set of functions F }.

By the common occasion as we have proved the existence of congruent function f(x) presented by counters in subset construction algorithm for AND-operator [8], we now can afford the generalization of the model theorem by in-equality P != NP:

$$P - f(x) \rightarrow NP. \quad (25)$$

This is how it's used to be as outlined ahead. Thus, P- and NP-classes are congruent over an arbitrary function f(x), when class P is smaller (same as by definition) than class NP and, thus, limits to it over this function.

Conclusion

We have defined the set of polynomial and non-polynomial functions according to the computing complexity of the environment as they can differ from tape machines to non-deterministic finite automata and there could be the functions according to which the complexities as of big-O notation may equal and the problem, thus, can be solved using applicable or non-applicable or impractical approach.

By giving the proof to our model theorem we state the co-existence of the congruent functions for which the relation between classes can be devised.

The question is open for finding the functions satisfying the condition of equality of complexity classes.

Another question is the relation between memory and time for effective computations as the results can be dynamically computed and stored

References

[1] Cook, Stephen. The P versus NP problem. *The millennium prize problems* (2006). 87-104.

[2] Fortnow, Lance. The status of the P versus NP problem. *Communications of the ACM* 52.9. (2009). 78-86.

[3] Sipser, Michael. "The history and status of the P versus NP question." *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*. 1992.

[4] Cook, Stephen. The importance of the P versus NP question. *Journal of the ACM (JACM)* 50.1 (2003). 27-29.

[5] Aaronson, Scott, and P. Is. Is P versus NP formally independent?. *Bulletin of the EATCS* 81.109-136 (2003). 70.

[6] Filar, Jerzy A., Michael Haythorpe, Richard Taylor. "Linearly-growing reductions of Karp's 21 NP-complete problems." arXiv preprint arXiv:1902.10349 (2019).

[7] Ackermann Wilhelm. Zum hilbertschen aufbau der reellen zahlen. *Mathematische Annalen* 99.1 (1928). 118-133.

[8] Syzdykov Mirzakhmet. Algorithm to Generate DFA for AND-operator in Regular Expression. *International Journal of Computer Applications* 975 (2015). 8887.

[9] Rabin, Michael O., Dana Scott. Finite automata and their decision problems. *IBM journal of research and development* 3.2 (1959). 114-125.

[10] Bellman Richard. Dynamic programming. *Science* 153.3731 (1966): 34-37.

[11] Kizilateş, Gözde, Fidan Nuriyeva. On the nearest neighbor algorithms for the traveling salesman problem. *Advances in Computational Science, Engineering and Information Technology*. Springer, Heidelberg, 2013. 111-118.

[12] Dorigo, Marco, Gianni Di Caro. "Ant colony optimization: a new meta-heuristic." *Proceedings of the 1999 congress on evolutionary computation-CEC99* (Cat. No. 99TH8406). IEEE, 1999. 2.

[13] Held, Michael, Richard M. Karp. A dynamic programming approach to sequencing problems. *Journal of the Society for Industrial and Applied mathematics* 10.1 (1962). 196-210.

[14] Traversa, Fabio Lorenzo, et al. Memcomputing NP-complete problems in polynomial time using polynomial resources and collective states. *Science advances* 1.6 (2015).

UDC 551.590.2
IRSTI 37.23

GRAPH THEORY IN ANALYSIS OF PALEOCLIMATIC TIME SERIES

I.S. Knyazeva[1], N.G. Makarenko[2], A.A. Pak[3]
[1,2]The Central Astronomical Observatory of
the Russian Academy of Sciences at Pulkovo, Saint-Petersburg, Russia
[3]Institute of Information and Computational Technologies, Almaty, Kazakhstan
[1]iknyazeva@gmail.com, [2]ng-makar@mail.ru, [3]aa.pak83@gmail.com
[1]ORCID ID: https://orcid.org/0000-0003-0026-6047
[2]ORCID ID: https://orcid.org/0000-0001-9917-1421
[3]ORCID ID: https://orcid.org/0000-0002-8685-9355

Abstract. At the current work, we present the results of diagnostic of similarities of various paleo reconstructions of temperatures with the help of network approaches. The first step of data processing is transformations of correlational patterns of time series into the geometry of the corresponding network, which is an input for further processing steps by methods of algebraic topology. To detect the possible non-linear connections between climatic series and solar activity, we propose the approach based on network generation with the help of embedding time series into the feature space of the corresponding dimension. In conclusion, we give markovian chains for climatic reconstructions and Wolf numbers.

Keywords: climatic series, Wolf numbers, solar activity, algebraic topology

Introduction

The information about the climate of Earth at previous periods of time before the epoch with instrumental recordings is about the extraction data from auxiliary sources such as annual tree rings, glacial kerns, marine deposits, glacial varved clays, thermospeleological series. Various thermal reconstructions can be calculated from correlations of these proxy indicators with temperature of instrumental period. So, the result of such reconstructions of averaged temperature in North hemisphere is an inference about "hockey stick", namely the steep increase of averaged temperature in 20s century, which was after relatively stable state of temperature during 1000 years. These point of view is not common. It is not unlikely that the unique of current increase of temperature can be erroneous, due to the fact of understaged low-frequency trends in data [1].

There is an open question about the quality of the reconstruction. It is worth noting, that regression models of Sluckogo-Yula, which are the basis of the majority of reconstructions, actually, they are applicable only for stationary (or weakly stationary) random series, where randomness is endogenic property. That's poorly agreed with commonly approved climatic scenarios [2]. As the result of stated above, there are several main problems of inferences about the future scenarios of climate:

How to check the quality of reconstructions?

How to choose the best reconstruction?

What's the meaning of the "best" one of reconstructions?

How correctly compare reconstructions?

How correctly average reconstructions?

Recently we tried to address some of these questions. The main idea was about the checking, is it correct to use a differentiable model to generate reconstruction, e.g., Takens models [3]. The answer was negative: the Hölder regularity of paleoreconstructions turned out to be worse than that of the gauge instrumental series [4].

In this article, we use the methods of graph (network) theory to compare different reconstructions with each other and detect the sun-conditioned component in them. Networks are built directly from time series. The presence of such a component would make it possible to construct a criterion for choosing the best reconstruction within the framework in a correct way. Networked approaches have been used previously to visualize climate and meteorological data [5], [6].

We use quantitative network descriptors from time series, (1) to compare paleoreconstructions with each other and (2) to highlight a possible solar component in temperature reconstructions. For this, the annual series of Wolf numbers from 1090 to 2002 was used as a reconstruction of solar activity [7].

Considered 6 temperature reconstructions of the previous millennium, described in the work [8]. Let's list them: reconstructions by Briffa[9] and Esper [10], Jones [11] and Mann's reconstruction [12].

The last mentioned is considered the most comprehensive because it sourced paleoclimatic records from more than 100 different locations. It took into account data with different time resolutions and scales. The Crowley reconstruction [13] is based on a sample that uses only the data that is available at all times. Finally, Moberg's latest reconstruction [14] is based on high frequency (tree rings) and low frequency data, which are considered separately using wavelet decomposition. All of these reconstructions were investigated using network approaches.

2 Construction of complex networks from series

The main idea of constructing networks from series is to transform a time series into a complex network or graph [15], [16]. A graph is an ordered set of vertices (nodes) and edges (arcs). We describe below three ways to build networks from rows.

2.1 Loop networks

This method, proposed in [17], refers to the construction of complex networks from pseudo-periodic time series. The approach assumes that the time series is cyclical, and a separate cycle is taken as the base unit or node for building the network. A network of such rows is constructed as follows. The time series is divided into cycles, which can be of different lengths; each cycle is considered a network node. So, for the time series of Wolf numbers, the nodes are separate cycles. To construct edges between nodes, they are compared with each other in a suitable metric, for example, in Euclidean [16].

Let $d(C_i, C_j)$ is a distance between two loops: $C_i = (x_1, x_2, ..., x_{n_i})$, $C_j = (y_1, y_2, ..., y_{n_j})$, and Δ is corresponding cutoff. Then the adjacency matrix $a(i, j)$ is defined as follows: if $d(C_i, C_j) < \Delta$, then $a(i, j) = 1$ and loops C_i, C_j considered to be connected. Contra versa they are disconnected, $a(i, j) = 0$. With the help of an adjacency matrix and appropriate algorithm one can produce a network [18].

2.2 Networks of embedded data

Another method [16], allows to build networks for arbitrary, not necessarily cyclic, time series using the technique of topological embeddings of scalar data into R^m. The choice of dimension and lag is a stand alone problem of chaotic dynamics and depends on the data [3]. After the data is embedded, each point, i.e. - vector of lagging coordinates, is considered as a separate node. The network is obtained by connecting a node with its k nearest neighbors, for some fixed k. In the case when two points are mutually close to each other, the connection is added only once, so that the network is built as a simple undirected graph. The average complexity of the network will be 2k. After that, by evaluating the correlation function, links caused by the close arrangement of nodes are excluded. The resulting network is an NxN connectivity matrix filled with zeros and ones. Various approaches can be used to visualize this matrix; general requirements for the construction of a graph are reduced to minimizing the number of intersections of the edges of the graph and the distance between the vertices [18]. Figure 1 shows an example of a time series representing one of the coordinates of the solution to the Ressler equation [19] and the corresponding network, for the following nesting parameters: dimension m =3, lag τ = 10.

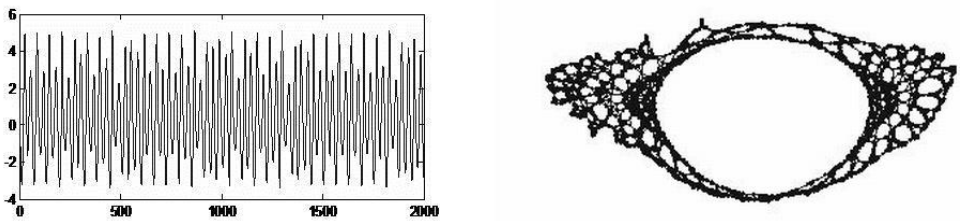

Figure 1 - Rössler time series and dive network (dim = 3, lag = 2)

The resulting network is close to a time series attractor [19] and reflects all the main features of the dynamical system underlying the time series. There are a number of simple approaches to describe graphs formally. One of them is based on the so-called "motives" that arise in the network and the frequency of their occurrence [20] . A motive of size n is called a subnet consisting of n nodes. In fact, each motif

represents one of the possible ways of connecting n points. For simplicity of calculations, motives of size 4 are usually considered, which contains 6 variants of motives (see Fig. 2)

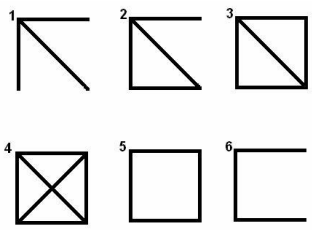

Figure 2- Variants of motives for a subnet with a number of nodes equal to 4.

It turns out that for time series generated by different dynamical systems, the distribution of motives is significantly different. Table 1 shows the frequencies of occurrence of motives from Fig. 2. for networks built by nesting three time series.

Table 1. Distribution of the occurrence of motives for three model series

Motif №	Lorenz	Noise	Resler
1	0,0736	0,02904	0,04229
2	0,25606	0,33112	0,29853
3	0,10145	0,09766	0,12309
4	0,01038	0,02226	0,01938
5	0,06893	0,01171	0,04808
6	0,4896	0,50822	0,46863

.

Note that more informative descriptors of the graph complexity are the spectrum of eigenvalues of the Laplace-Kirchhoff matrix [21].

2.3 Markovian networks

Consider a dynamic system (M,T) like mapping $T: M \to M$ defined on a compact subset $M \in R^d$. The random trajectory of this system is a sequence of points: $\left\{x, Tx, T^2x, ...\right\}$. To get a coarse-grained version of the dynamics, let's split M into a finite number of connected non-empty subsets: $\bigcup_{i-1}^{n} A_i \subseteq M$, $A_i \cap A_j = \varnothing$. Let's identify $i-$ state of a dynamic system with a cell A_i. We do not discuss the choice of a suitable partition here. It is known that the Markov model of dynamics is given by the matrix of transition probabilities: $\hat{P}_{ij} = \text{Prob}_{\mu}\{Tx \in A_j \,|\, x \in A_i\}$ [Froyland, 2001]. Here the probability is understood with respect to some natural ergodic measure μ, assuming that for almost all starting points $x \in M$ the trajectories have the same distribution. However, this measure is usually unknown. Therefore, instead of μ use the normalized Lebesgue measure:

$$P_{ij} = \frac{m\left(A_i \cap T^{-1}A_j\right)}{m\left(A_i\right)}$$

In other words, the probability of transition from the state i (or cell A_i) into state j (cell A_i) is determined by the relative proportion of points A_i who moved to A_j by affect of action T. The described ideas underlie the Markov networks proposed in the work [22].

Let's construct a histogram of samples for a time series so that each bin contains approximately the same number of points. In this case, the width of the bins will be of course different, and the bins themselves will play the role of partitioning the phase space, i.e. $\left\{A_i\right\}_{i=1}^{n}$. Let us assign a serial number to each bin,

and then encode each value of the time series in accordance with the belonging to a certain bin. As a result of such a procedure, instead of the original row, a new row consisting of bin numbers will be obtained. Then we calculate the frequency of all consecutive transitions from one number to another, we get a transition matrix of size $N \times N$, where N is number of bins.

In the event that the row is smooth, adjacent transitions will prevail and the matrix will be close to a diagonal form, but if, on the contrary, the structure is very chaotic, then the matrix will have many filled elements far from the main diagonal.

3. Numerical results

Below are the results of a network analysis of the monthly average Wolf numbers, the reconstruction of the annual values of a series of Wolf numbers (from 1090 to 2002) obtained by Yu.A. Nagovitsin in [7], and temperature reconstructions of the northern hemisphere over the last thousand years described in the introduction.

3.1 Comparison of solar activity cycles

The Wolf number series is one of the main indices of solar activity, and therefore we used it to identify possible solar-terrestrial connections. Wolf's Zurich series of monthly averages has been available since 1749 and contains 23 complete cycles. We have built cyclical networks for this series to identify similarities between different cycles. The division into nodes of the network was carried out in accordance with the known dates of the beginning and end of the cycles according to the list of solar activity cycles.

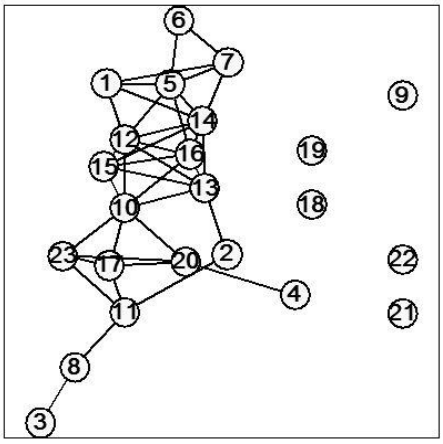

Figure 3- Relationships between cycles at the threshold $\Delta = 300$. Five cycles form isolated vertices.

Further, between each pair of cycles, the distance in the metric L_2 was calculated and a threshold $\Delta = 350$ was selected for which the adjacency matrix was constructed. The choice of this value was due to the fact that this is the minimum threshold at which each node has at least one connection with others. Figure 3 shows the minimum energy graph [18] for a threshold $\Delta = 300$ as an example. This analysis did not lead to any clusters associated with the shape of the cycles.

3.2 Analysis of the dynamics of solar activity and temperature reconstructions

To diagnose possible connections between temperature reconstructions and solar activity, we used embedded nets (see Section 2.2). For different lags, the networks constructed by the series of Wolf numbers demonstrated a stable topological structure, similar to the networks obtained for systems of deterministic chaos (see Figs. 1 and 4). Stability is understood here in the following sense of the similarity of the network built over the entire sample and its part. Figure 4 (top panel) clearly shows the similarity of the networks for the complete Wolf series and the network for the first ten cycles. In the bottom panel of Fig. 4, on the right, the net for the Ressler attractor for $m = 3, \tau = 2$, and on the left, the net for the Gaussian noise. The similarity of the networks for the Wolf numbers and the Ressler attractor confirms the conclusions of [19]. An analysis of the frequency of occurrence of motives also confirmed the connection that a number of monthly Wolf numbers belong to chaotic processes.

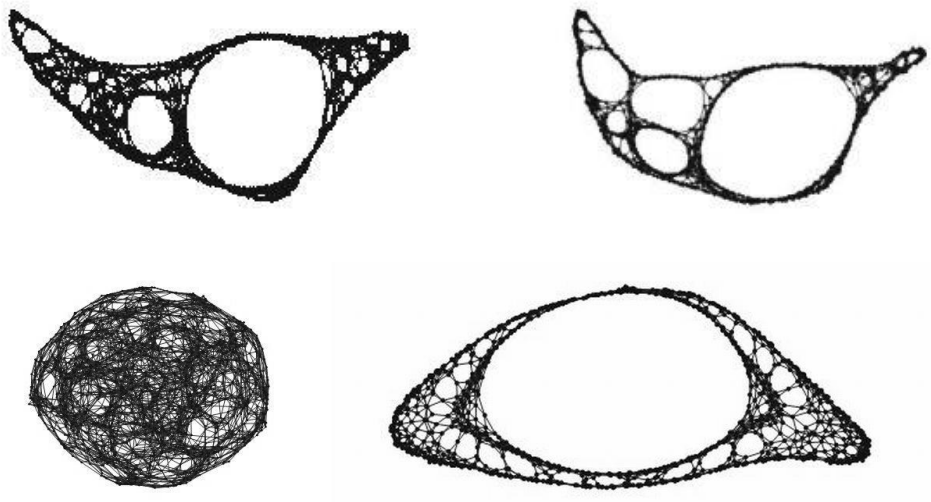

Figure 4- Networks for different tiers. Above - Wolf numbers, complete series (left), right first 10 cycles of Wolf series ($m=3,\tau=5$). Below, on the right - the network for the Rossler attractor ($m=3,\tau=2$), on the left - Gaussian noise ($m=3,\tau=2$)

To analyze climatic regimes, we also used dive nets for annual temperature reconstructions from 1000 to 1995 by various authors and for various diving parameters. In all cases, it turned out that networks are structures that are unstable in time and differ from each other. Here, as above, instability means the difference in the form of networks for different time intervals. Qualitatively, these networks do not resemble either a random series or networks of deterministic chaos.

Figure 5 shows examples of such networks for temperature reconstructions at immersion parameters $m=24,\tau=1$. We chose the dimension $m=24$ in connection with the indication of the possible presence of the Hale cycle in the climatic series [2]. It turned out that the networks of all temperature reconstructions differ to varying degrees from each other. Table 2 and Table 3 show the results of the numerical analysis of the graphs. It is interesting to note that the reconstructions of Briffa and Esper have a rather visual structure and similar parameters of the distribution of motifs and curvature. Perhaps this is due to the fact that they rely only on dendrodata. The rest of the reconstructions, despite the fact that they use similar multi-proxy data, differ significantly from each other. The Crowley and Moberg reconstructions stand out in particular. Thus, it can be assumed that the internal dynamics of the reconstructions is not sufficiently invariant with respect to the use of the initial data.

Figure 5- Types of networks for various climatic reconstructions. Left, net for Mann row, right, for Moberg row for $m=24,\tau=1$.

To analyze possible solar-terrestrial connections, networks for reconstruction of the annual Wolf numbers [7] were constructed with the same parameters as for temperature reconstructions. The results were quite unexpected. It turned out that year-on-year data is already poorly matched to networks for processes with a dynamic chaos regime. Moreover, visually, they are more similar to climatic reconstructions. It turned out that with an increase in the dimension of the phase space, the structure of the network changes and becomes closer to the structure of temperature reconstructions

Markov networks for temperature reconstructions also correspond to the picture obtained above: the correlation structure and internal dynamics of the variants are different. The degree of dominance of the main diagonal corresponds to the smoothness of the row; the more pronounced it is, the smoother the row.

4. Conclusions

The results obtained can be summarized as follows:

1. The structure of cyclic networks for the monthly average Wolf numbers does not contain any distinguished clusters, which could be expected in the case of the existence of one or several “typical” cycles with analogs close to them, in the sense of the metric used.

2. Immersion networks for monthly average Wolf numbers are qualitatively and in terms of the distribution of motives similar to dynamic chaos (like the Ressler attractor). The nets for the full row and its half are similar. Perhaps this indicates the existence of a certain scale-invariant measure.

3. Networks for paleoclimatic series are qualitatively different from both random networks and dynamic chaos networks. They differ significantly from each other, indicating a lack of coordinated dynamics of reconstruction. Mutual differences decrease with an increase in the dimension of investment, which is possibly due to the presence of long correlations (24 years, 60 years or more). Markov networks support these findings.

4. The structure of the networks for the reconstruction of the annual Wolf numbers substantially depends on the dimension of the immersion. At low dimensions, networks are close to dynamic chaos; at large, they are similar to the networks of climatic series. Perhaps this is due to the absence of significant long-term relationships between cycles in the analyzed time interval.

References

[1] Christiansen B. “Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?” *J. Clim.*, (2011). 24(3). 674–692, doi: 10.1175/2010JCLI3646.1.

[2] Monin A. S., D. M. Sonechkin, *Observational Climate Fluctuations*. Moscow: Nauka, 2005.

[3] Ott E., Sauer T., Yorke J. A., Coping with chaos. Analysis of chaotic data and the exploitation of chaotic systems. (1994).

[4] Makarenko N. G., Karimova L. M., Helama S., Eronen M. “Evaluating direct and indirect evidence of climatic change by Hölder regularity and order pattern in time series,” *Geomagn. Aeron.*, (2013). 53. 8. 922–926, Dec., doi: 10.1134/S0016793213080148.

[5] Tsonis A. A., Swanson K. L., Roebber P. J. What do networks have to do with climate? Bull. Am. Meteorol. Soc. 2006. 87(5). 585–595, doi: 10.1175/BAMS-87-5-585.

[6] Paluš M., Hartman D., Hlinka J., Vejmelka M. Discerning connectivity from dynamics in climate networks. Nonlinear Process. Geophys. 2011. 18(5). 751–763, doi: 10.5194/npg-18-751-2011.

[7] Nagovitsin Y. A. Nonlinear mathematical model of the solar cycle process and possibilities for reconstruction of past activity. Pisma v AZh. 1997. 23(11).

[8] Ogurtsov M.G., Junger H., Lindholm M., Helama S., Dergachev V. A. Quasisecular cyclicity in the climate of the earth’s nothern hemisphere and its possible correlation with solar activity variations. Solar-Terrestrial Phys. 2007. 10(123). 10–15.

[9] Briffa K. R. Annual variability in the Holocene: Interpreting the message of ancient trees. Quat. Sci. Rev. 2000. 19. 87–105.

[10] Esper J., Cook E. Schweingruber F. H. Low-frequency signals in long tree-ring chronologies and the reconstruction of past temperature variability. *Science (80-.).* 2002. 295. 2250–2253.

[11] Jones P. D., Briffa K. R., Barnett T. P. Tett F. B. High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with General Circulation Model control-run temperatures. *The Holocene*. 1998. 8(4). 455–471, doi: 10.1191/095968398667194956.

[12] Mann M. E., Bradley M. K., Raymond S. Hughes. Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophys. Res. Lett.* 1999. 26. 759–762.

[13] Crowley T. J. Lowery T. S. How warm was the Medieval warm period?. *Ambio. 2000*. 29. 51–54.

[14] Moberg A., Sonechkin D. M., Holmgren K., Datsenko M. H., Karlén W. Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data *Nature*. 2005. 433(7026). 613–617, doi: 10.1038/nature03265.

[15] Gao Z. K., Small M., Kurths J. Complex network analysis of time series. *EPL. 2016.* 116(5). 50001, doi: 10.1209/0295-5075/116/50001.

[16] Small M., Zhang, J. Xu X. Transforming time series into complex networks. in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering.* 2009. 5(2). 2078–2089, doi: 10.1007/978-3-642-02469-6_84.

[17] Zhang J., Small M. Complex network from pseudoperiodic time series: Topology versus dynamics. *Phys. Rev. Lett. 2006.* 96 (23). 238701, doi: 10.1103/PhysRevLett.96.238701.

[18] Battista D., Eades G., Tamassia R., Tollis I.G., Graph Drawing: Algorithms for the Visualization

of Graphs. Prentice Hall. 1999.

[19] Letellier C., Aguirre L. A., Maquet J., and R. Gilmore, Evidence for low dimensional chaos in sunspot cycles. *Astron. Astrophys. 2006.* 449(1). 379–387. doi: 10.1051/0004-6361:20053947.

[20] Xu X. K., Zhang J., Li P., Small M. Changing motif distributions in complex networks by manipulating rich-club connections. *Phys. A Stat. Mech. its Appl. 2011*. 390(23–24). 4621–4626, doi: 10.1016/j.physa.2011.06.069.

[21] Alekseev V. V., Makarenko N. G., Knyazeva I. S. Graph Dynamics of Solar Active Regions: Morse–Smale Complexes and Multiscale Graphs of Magnetograms. *Astron. Lett. 2020.* 46(7). 488–500, doi: 10.1134/S1063773720070014.

[22] Shirazi A. H., Jafari G. R., Davoudi J., Peinke J., Rahimi M. R. Tabar, Sahimi M. Mapping stochastic processes onto complex networks *J. Stat. Mech. Theory Exp.* 2009. 9(7). P07046, doi: 10.1088/1742-5468/2009/07/P07046.

UDC 629.764.7
IRSTI 89.01.94

THE INFORMATION AND ANALYTICAL SYSTEM OF INNOVATIVE TECHNOLOGY FOR CREATING A PROMISING LAUNCH VEHICLE WITH IMPROVED ENVIRONMENTAL CHARACTERISTICS

[1]Alipbayev K., [2]Trushlyakov V.I., [3]Yermoldina G., [3]Bapyshev A., [3]Suimenbayev N.
[1]Head of the Institute of Telecommunications and Space Engineering, Almaty University of Power Engineering and Telecommunications
[2] Omsk State Technical University, Russia
[3] Institute of information and computational technologies, Kazakhstan
e-mail: gulerm@mail.ru
[1]ORCID ID: https://orcid.org/0000-0003-1652-9470
[2]ORCID ID: https://orcid.org/0000-0002-8444-6880
[3]ORCID ID: https://orcid.org/0000-0002-0276-6670
ORCID ID: https://orcid.org/0000-0003-1634-6738

Abstract. The article is devoted to the problem of ensuring the ecological safety of the operation of rocket and space complexes. Within the framework of the concept of the controlled descent of the spent first stages of the launch vehicle, a fundamentally new approach to the problem of ensuring the environmental safety of land fall areas is proposed. This approach ensures fire and explosion safety and a decrease in the size of the fall areas as well as the extraction of unused fuel residues in the tanks and pipings of the worked-off stage for performing maneuvers to descend into a given region.

The object of research is the processes, as well as information flows circulating in the analytical information systems of the fall areas, the cosmodrome and the design bureau to develop recommendations for innovative technology for creating launch vehicles with improved environmental characteristics.

The objective of research is to analyze the state of the problem of technogenic impact of liquid-fueled launch vehicles and preparation of initial data for the development of an upgraded information and analytical system.

The presented original concept consists in a combined method of reducing the anthropogenic impact of the separated parts of launch vehicles by installing an additional system. This system provides a controlled descent of the worked-off stages to a given aiming point, and optimizing the fall areas based on the criteria of resistance to the technogenic impact of space rocket launches through the development of information and analytical system. A methodology has been developed for the selection of these sites, based on the creation of an information and analytical system for the area of the fall, which will become an element of the existing system of environmental monitoring of the Baikonur cosmodrome.

Key words: information and analytical system, database, launch vehicle, software package.

1. Introduction

The development of modern space vehicles, including advanced launch vehicles with liquid-propellant rocket engines (LPRE), in accordance with the accepted recommendations of such organizations as the UN technical Subcommittee on the peaceful uses of outer space (COPUOS), the inter-Agency space debris coordination Committee (IADC) [1] should be accompanied by a radical change in approaches to the process of creating and operating launch vehicles.

Currently, trends in reducing the man-made impact on the environment of launch vehicles with main liquid rocket engines (LPRE) to prevent explosions in the orbits of separating parts (SP) of launch vehicles are given in Support to the IADC Space Debris Mitigation Guidelines [2], and domestic works [3, 4].

However, there is a great example of a comprehensive solution to the problem. This refers to the SpaceX project, when the first stage of the launch vehicle after performing a maneuver in the atmosphere lands strictly in the designated place on the Earth's surface or in the ocean. In addition to SpaceX, a number of innovative projects (Blue Origin, Sierra Nevada Corporation, etc.) are

being built in the United States using a similar model [5, 6]). In [7], the possible flight patterns of the separating part of the first stage equipped with a rocket-dynamic rescue system are analyzed using the example of a medium-class oxygen-kerosene two-stage launch vehicle, which is similar in its energy and mass characteristics to the Falcon-9 launch vehicle (LV), compiled according to the results of ballistic design.

In turn, AIS is considered as an analytical information system for theoretical and experimental research and information support for bench tests of an Autonomous onboard descent system (AODS).

Main tasks in the development of AODS for the first stage:

- theoretical and experimental studies of evaporation processes in various schemes of heat supply to tanks;
- simulation of various descent schemes of SP 1, including the use of aerodynamic braking by the SP 1 body, etc.
- selection of AODS design parameters that increase not only environmental safety (ES) in the fall areas, but also increase other tactical and technical characteristics of the LV (payload mass, expansion of inclination ranges, reduction of the cost of launching the LV) , etc.

For the purposes of this research, the information and analytical system (IAS) is considered as a system that provides a procedure for analyzing and making recommendations on innovative technology for creating advanced launch vehicles with improved environmental characteristics.

The main factors of the technogenic impact of launches of LV with sustainer LPRE on the ecosystem of the allotted territories of Kazakhstan are reflected in a number of works [9, 10]. The development of promising launch vehicles with sustainer liquid jet engines, in accordance with the accepted recommendations [2, 8], provides for a significant reduction in the technogenic impact of the LV operation on the environment, including:

– preventing the clogging of near-earth space by upper WS with cruising LPRE, which are large-sized explosive space debris;

– a dramatic reduction in the number and areas of FA on the Earth's surface for the separating parts of the launch vehicle, which are fire hazardous and toxic objects, leading to chemical contamination of soil and soils with residues of liquid toxic PC such as: asymmetric dimethylhydrazine, nitric acid, kerosene

2 Analysis of existing information and analytical systems of the Baikonur complex

The IAS of the cosmodrome environmental monitoring system (CEMS) is designed for collecting, processing, analyzing, and summarizing environmental information when solving a set of tasks of scientific and technical support for environmental monitoring in terms of providing them with structured information in the decision-making system. It serves as the basis for industrial environmental monitoring, which is carried out within the framework of the departmental environmental monitoring system of the Baikonur cosmodrome (CEMS), linked to the Unified state system for monitoring the environment and natural resources of the Republic of Kazakhstan. It should be noted that State environmental monitoring is carried out by specially authorized state bodies of the Republic of Kazakhstan, in accordance with the Environmental code of the Republic of Kazakhstan. State environmental monitoring on the territory of the Baikonur complex is carried out in agreement with the authorized bodies of the Russian Side in accordance with the ownership of the objects.

Currently, the applied aspects of the IAS of the environmental monitoring system of the Baikonur complex are quite well developed. They are reflected in a number of main works on the functioning of the monitoring system, during the regular operation of rocket and space complexes and support facilities of the cosmodrome, as well as in possible emergency situations [9, 12].

The ideology of development of ecological safety of operation of space vehicles, as determined by the sponsors, is to develop recommendations for improving space-rocket equipment and technologies with improved environmental performance based on in-depth and comprehensive analysis of the environmental effects of space activities.

Based on the conducted research, it can be concluded that the existing IAS of the Baikonur complex do not fully reflect the issues of developing recommendations for improving and developing innovative technology for creating launch vehicles with improved environmental characteristics and require modernization.

The concept is based on three important postulates that take into account the features of land-based spaceports, as well as the requirements of the indigenous population to ensure quality living conditions.

Postulate 1. The life cycle of the WS should not end, as it is currently implemented in the logic of operation of almost all Russian launch vehicles launched from the Baikonur cosmodrome - by achieving the specified motion parameters, shutting down the sustainer LPRE. The operating stage of the WS, similar to the spacecraft, should still be implemeneted, which provides for its transfer to the disposal orbit after the end of the active operation period. At this stage, the WS must ensure that the technogenic impact on the environment in the area of its intended fall is minimized.

Postulate 2. At the present stage of the study, it is not expected to return the WS to the cosmodrome with its soft landing and subsequent reuse, similar to the first WS of the Falcon-9 launch vehicle.

Postulate 3. Perfect fall WS with practically dry fuel tanks and fuel lines with minimum deviation from the projected aiming point WS in the R-neighborhood from the optimum energy point of incidence of the WS.

IAS of the Baikonur cosmodrome is the basis for environmental monitoring of space activities on the territory of the Republic of Kazakhstan. The innovative nature of the creation of rocket units consists in a comprehensive solution to the problem: the development of rocket units based on gasification, the development of AODS, the evaluation of the results using the IASlv.

The interaction between IAScd and IASlv, like any information exchange between complex technical systems, is iterative in nature, which can be divided into several stages and levels, both as each IAS is ready and the current tasks solved by each IAS.

1) At the current level, the primary task is to create an IASlv and form a database for each FA of the most acceptable points of WS fall from the condition $\min \{C_i^{\text{v}}[\vec{R}_i(x_i, y_i)]\}$.

2) The information obtained is necessary for conducting research within the framework of IASlv, IASESLV for the following purposes:

a) synthesis of various control programs for the LV movement on the active section of the launch trajectory without taking into account restrictions on the areas of the WS fall (calculation $\vec{R}_{fra}(x, y)$);

б) distance estimation $\Delta\vec{R}$between $\vec{R}_{\text{fra}}(x, y)$ and recommended points of falling assets obtained in IASfa of conditions $\min\{C_i^{\text{v}}[\vec{R}_i(x_i, y_i)]\}$;

в) development of proposals for changing the design parameters of the WS for the implementation of maneuvers on the descent path.

Implementation of the presented concept of enhancing the ecological safety of launch vehicles with sustainer LPRE will reduce the environmental load on the environment in the FA of the Baikonur cosmodrome due to a dramatic reduction in the FA area (controlled descent of the WS), a significant reduction in the likelihood of vegetation fires (due to the almost complete extraction of liquid fuel residues), selection of the safest (from an ecological point of view) points of the fall of the environment in the territory of the designated FA.

3 Development of the initial data structure of the modernized IAS-M

The formation of the structure of the initial data is based on the ontology of information interaction of the IAS of various levels developed within the framework of the project (Figure 1).

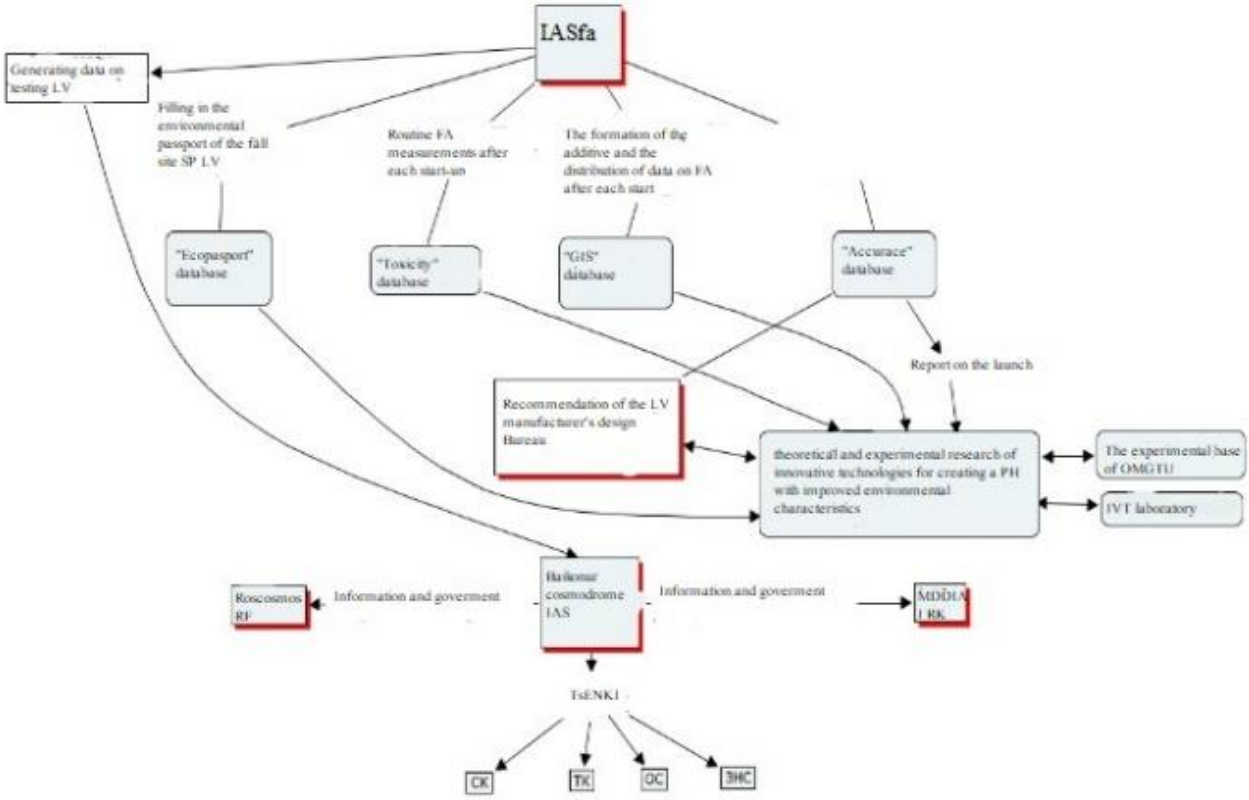

Figure 1 - IAS information interaction ontology

3.1 Development of the IAS-M functional structureThe functional structure of the IAS-M includes:

IASkd, - the existing IAS of the cosmodrome

IASfa- additionally created IAS of the area of the WS fall

IASlv - the existing IAS for the development of LV

IAS_{lv}^{es}- additionally created IAS for the formation of recommendations for the creation of a launch vehicle with improved environmental characteristics.

IASlv is designed to generate data when developing promising launch vehicles in the area of WS landing. Functionally, IASlv solves the following tasks:

а) according to the data received about the upcoming launch of the LV from the IASlv (the initial point of aiming the fall of the WS in the selected RP$\vec{R}_{pr}^{int}(x_i, y_i)$, optimal aiming point $\vec{R}_{fra}(x, y)$, at which the mass of the payload injected into a given orbit is maximum, the partition of the FA area into N sections with areas Si (i = 1, …N), so that $\sum_{i=1}^{N} S_i = S_\Sigma$;

б) in the selected N areas, N possible predicted coordinates of the points of WS fall are selected;

в) distances are estimated $\Delta\vec{R}_i = \vec{R}_{fra}(x, y) - \vec{R}_{pr}(x_i, y_i)$ to assess the possibility of the WS maneuver by shifting the point of WS fall by these values and are transmitted to the IASlv;

г) based on the passport of this RP is calculated for each predicted point of incidence $\vec{R}_i(x_i, y_i)$ environmental damage $E_i[\vec{R}_i(x_i, y_i)]$ from falling into this i-th section and, accordingly, the cost of restoration work $C_i^{\text{v}}[\vec{R}_i(x_i, y_i)]$;

г) the obtained information is transmitted to the IASrn for calculating the programs for controlling the movement of the launch vehicle in the active section of the injection trajectory and programs for controlling the movement of the WS in the section of descent to the selected point, which is determined from the analysis of the data array $\{C_i^{\text{v}}[\vec{R}_i(x_i, y_i)]\}$, оцScores of the ballistic capabilities of the AODS for maneuvering to change the coordinates of the point of incidence by $\Delta\vec{R}_i = \vec{R}_{fra}(x, y) - \vec{R}_{\text{прр}}(x_i, y_i)$.

As follows from the given Figure 2, it is possible to ensure the fall of the WS into areas with significantly different landscape conditions. In this case, it is assumed that an ABS is installed on the WS, which provides control of the OWSS movement on the descent trajectory. As a result of this control, the accuracy of the WS fall is similar to the accuracy of the SP1LV "Falcon-9" landing when landing on a cosmodrome or a floating barge.

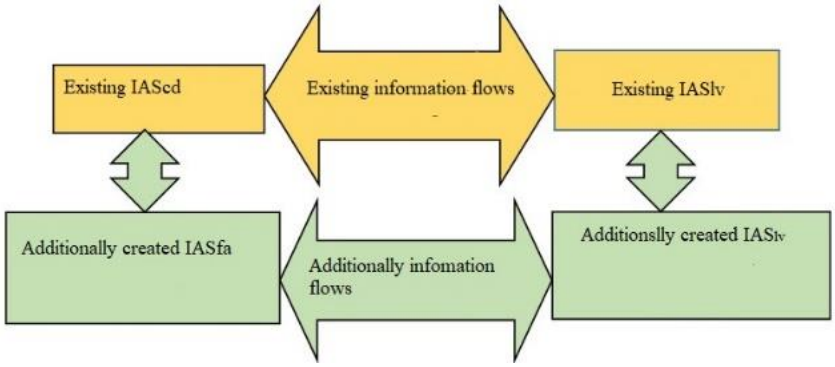

Figure 2 - General schematic diagram of information flows between IAScd, IASfa, IASlv, IAS_{lv}^{es}

In addition to the information received from the IASfa, which is necessary to improve the environmental safety of the LV, the IASfa is carrying out work.

3.2 Architecture of the IAS-M software package

The purpose of the software package is to ensure the management of the IAS, focused on achieving the goals of the project: researching an innovative technology for creating a launch vehicle with improved environmental characteristics.

The architecture of the software package (Figure 3) supports working with a distributed database. The system is divided into several subsystems ("nodes"), each "node" has its own database (DB). Synchronization of "nodes" with the central database is ensured through replications, which allows the use of the DBMS in conditions of insufficient bandwidth of communication channels.

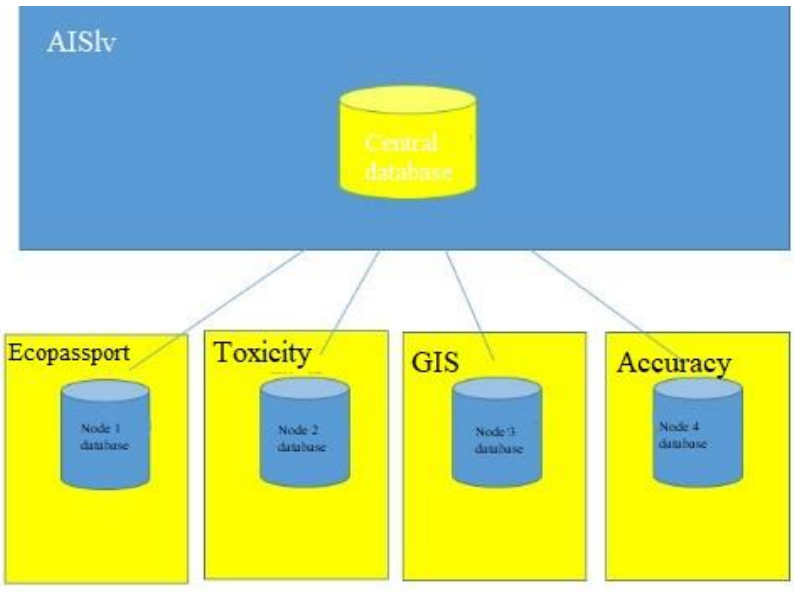

Figure 3 - Architecture of the software package

3.3 Initial data for IAS-M development

1) "Ecopassport»:

The subsystem concentrates the initial data containing data on the WS FA, which is necessary to ensure the functioning of the IAS-M during the regular testing and operation of the LV and includes:

- General information about the company responsible for the operation of the fall area;
- General information about the area where the SP LV falls and adjacent territories;
- Characteristics of natural and climatic conditions on the territory of the fall areas;
- Information about anthropogenic sources of pollution in FA and adjacent territories. Characteristics of the background ecological state of the NE;
- Characteristics of sources of FA pollution in rocket and space activities;
- Stability of soils;
- Impact of the SPLV in FA and adjacent territories;
- Levels of pollution of objects of the natural environment on the territory of FA and adjacent territories of RFC and their toxic derivatives: atmospheric air, biotic objects and levels of their pollution. Assessment of the environmental situation in FA and adjacent territories;
- Information about the amount of payments for pollution of the NE of FA, environmental protection measures
- Regulatory and reference data.

2) «Toxicity».

The subsystem concentrates the initial data containing data on the hygienic normalization of RFC and products of their chemical transformation in objects:

- Scale of contamination of FA RFC and products of their transformations;
- Characteristics of physical and chemical properties (PCP) and biological effects of liquid rocket propellants and their transformation products;
- Maximum permissible concentrations of rocket fuel components and some products of their chemical transformation in environmental objects;
- Results of in-house determination of rocket fuel components and products of their chemical transformation in environmental objects for each launch of the LV;
- Characteristics of physical and chemical properties and biological effects of liquid rocket fuels and their transformation products;
- Methods and technologies for detoxification of soils contaminated with toxic RFC;
- The initial concentration of RFC in the experimental sites and the results of quantitative chemical analysis;
- Results of identification of RFC and products of their transformation in soil samples before and after detoxification using the method of chromatography-mass spectrometry;
- Hygienic studies of FA and areas of influence;
- Regulatory and reference data.

3) «GIS».

The subsystem concentrates the source data containing GIS data for the WS FA:

- Attribute data on economic activity in a certain area of impact of separating parts of launch vehicles and in adjacent territories;
- Road network and terrain conditions;
- Power grid of the locality in FA;
- Cartographic materials: placing households in FA maps and FA maps FA soil, etc.
- Records of data on the number and coordinates of the fall of the WS in FA;
- SR (launch date, type of LV, SC, total launch mass of the LV);
- SR withdrawal routes on the territory of Kazakhstan;
- Regulatory and reference data.

4) «Accuracy».

В подсистеме концентрируются исходные данные, containing design parameters for ballistic descent WS results WS landing in FA for the evaluation of decisions made according to the requirements of the guaranteed drop at the given coordinates of the target point:

- Calculated data of the WS ballistic descent;
- Meteorological data for the launch day of the SR;
- Telemetry and external trajectory measurements on the active part of the trajectory (APT) of the first stage of the SR;
- Possible scenarios for the development of the environmental situation in the SRV areas, depending on the deviation from the specified coordinates of the aiming points;
- Regulatory and reference data.

5) « Modeling».

The subsystem concentrates the standard mathematical apparatus necessary for making recommendations for improving the LV.

4 Conclusion

1. The analysis of existing IAS is carried out. An ontology of IAS information interaction has been developed: IAScd, IASfa, IASlv, IAS_{lv}^{es} to assess the technogenic impact of the launch of the LV on the selected fall area, integrated into the General IAS of the Baikonur cosmodrome.
2. The functional structure and architecture of the software package were developed, and the initial data for the development of IAS-m were prepared.

5 Acknowledgements

The research was supported by grant no. AP09258759 "Development of a model of an information-forecasting system for determining the areas of launch and fall of ultra-light launch vehicles, taking into account the requirements of ecological safety".

References

[1] IADC space debris mitigation guidelines. www.iadc-online.org.: 25.05.2018.

[2] IADC-04-06, Rev 5.5 May 2014 Support to the IADC Space Debris Mitigation Guidelines. http://www.iadc-online.org/Documents/IADC-04(https://news.mail.ru/incident/30087239/?frommail=1.

[3] Space Debris Mitigation Guidelines of the Committee on the Peaceful Uses of Outer Space. United Nations Office for Outer Space Affairs/ Vienna, 2010. www.unoosa.org. 30.04.2018.

[4] Shatrov, ya. T. Ensuring environmental safety of rocket and space activities: in 3 books. - /Shatrov Ya. T. - Korolev, Moscow region: TsNIIMash publishing house. 2009 (In Russian).

[5] Patent no. RU2290352C2 Method for neutralization of fuel and oxidizer in the separating stages of a launch vehicle and a device for its implementation. http://www.findpatent.ru/patent/229/2290352.html: 15.06.2018 (In Russian).

[6] Titov B. A., Rychkov S. A. reducing the size of the fall areas of spent blocks of a Soyuz-type launch vehicle with deliberate partitioning of their design. Bulletin of SSAU, 2007, no. 1, pp. 90-97 (In Russian).

[7] Pat. no. 0002572014 "Method for reducing the fall areas of spent rocket blocks of the first stage of the launch vehicle when they are connected in parallel" from 27.12.2015 (In Russian).

[8] Trushlyakov V., Zharikov L., Lempert D. Development of solid fuel gas generating compositions to ensure non explosiveness of spent orbital stages of liquid rocket of space launch vehicles. Eurasian Chemico-Technological Journal. 2017. 19(1). 63-70.

[9] Environmental problems and risks of rocket and space technology impacts on the environment: a reference guide. under the General editorship of Adushkin V. V., Kozlov S. I., Petrov A.V. - M: publishing House "Ankil", 2000. 640. (In Russian).

[10] Krassnov V.M., Suymenbayev B.T. Sputtered rocket fuel in the atmosphere, the size of the cloud and the trajectory of its flight. abstracts of reports of geophysical research. 2001. 3. (In Russian).

[11] Suymenbayev B. T., Krasnov V. M. Some aspects of the consequences of accidents of the proton launch vehicle over the territory of Kazakhstan. Ecological systems and devices. 2001. 2. (In Russian).

[12] Zubatov J., Tauasarov A., Kozlovsky V., Alekseeva D., Bisarieva S., Pozdnyakov A., Gusarova N. Environmental safety activities of the cosmodrome "Baikonur". under the editorship of doctor of technical Sciences, academician MANEB J. Gubatova. Almaty (2011). (In Russian).

UDC 004.02
IRSTI 20.53.15

LOCAL SEARCH IN NON-DETERMINISTIC FINITE AUTOMATA WITH EXTENSIONS

Mirzakhmet Syzdykov
al-Farabi Kazakh National University, Almaty, Kazakhstan
mspmail598@gmail.com
ORCID ID: https://orcid.org/0000-0002-8086-775X

Abstract. In this work we present the theoretical approach over solving the back-reference problem in regular expression matching within the almost polynomial time using local search within the memory, while within the growth of capturing groups we obtain the exponential results: for this purpose we develop the modified matching algorithm operating on non-deterministic finite automata within the modified search algorithm and presence of the specific method also over extended regular expressions. This is made due to the algorithm which can be adjusted for approximate searching allowing us to imply extended operators and features of modern regular expressions like intersection, subtraction and complement, as well as back-references. The review of past work on this issues is also done: to the present time there is no discrete algorithm in systems like automata for local search. Thus, we obtain the new result of matching the pattern locally while the simulating algorithm works as usual. The obtained result also refers to the membership problem with local bound which can be set in the main algorithm presented in this article.

Keywords: Back-reference, Finite Automata, Local Search, Extended regular expressions.

Introduction

As of the regular expression languages nature and the proposed solution of extending it with a repeater like back-reference, in the modern era this feature is used in analyzing specific sequences in the medicine [1].

Thus, the back-reference problem is considered widely as NP-complete and the backtracking algorithm is used for the matching problem [2].

However, due to the memoization [2] the backtracking process can be optimized — as we use the same approach in giving the memoization of the matching process for back-references and corresponding capturing groups.

The solution also covers the regular expression features like look-ahead and extended operators.

Typically we define the non-deterministic finite automaton (NFA) as a tuple of alphabet A, states S, list of transitions T, starting state s0 and list of finishing or accepting states:

$$<A, S, T, s_0, F> \quad (1)$$

In this work we will present the modified NFA for organizing the local search on back-reference, positive or negative look-ahead and look-behind and extended set of operators like intersection, complement and subtraction.

Related past work review

In [1] the importance and application of back-reference operator is given.

In [2] the memoization principle is covered which we use in simulating the NFA within the capturing groups and back-references on the counter state – this method is presented further in this article.

The work [3] deals with the backtracking matchers as they are commonly used in matching regular expression with back-references as an easy alternative to the approach presented in this paper, which uses the memoization of the state context and counter simulation for the state referred as ending in NFA for back-reference operator in regular expression.

In [4] the subset construction of converting NFA to deterministic finite automaton (DFA) is given.

In [5] the construction of NFA for the regular expression is given.

In [6] the past work for converting regular expression to finite automata is presented.

In [7] the definition of transitional NFA is given with respect to the context of captured groups to be

optimized towards memory consumption.

In [8] the NFA constructions are presented for extended operators in regular expression like intersection, subtraction and complement which are to be used in local search during matching process as well as a back-reference operator.

The modern work [9] describes the logical constructions to model the features like positive or negative look-ahead and look-behind which is met in Perl and other regular expression flavors.

In [10] the algorithm for converting regular expression directly into deterministic finite automaton is presented. It's important to us as the example of the past work.

In [11] the effective polynomial algorithm is given for extended regular expressions. We will show further that our automaton is as well quadratic in memory consumption as in complexity.

In [12] the theory of complexity is presented – this is a seminal work of theory of complexity when the optimization problem is to be non-reliable if we deal with solution set explosion along the algorithm to check the correct solution in this set.

We will state further that according to the NP-completeness of back-reference problem [13], using the counter heuristics we achieve polynomial results in time and memory.

The work [14] considers the state explosion problem when converting NFA to DFA using techniques to avoid it. We consider modified or extended NFA in this work which can be built by any existing algorithm and mainly is from linear up to quadratic number of states in it.

The work [15] is important for us in the sense of exact polynomial complexity, the algorithm for Traveling Salesman Problem (TSP) given in the article is exponential in time and can be considered as not optimal in the sense of modern computer architecture.

Methodology

Before proceeding to the NFA, we can mention that it can be transformed to DFA according to the algorithm by subset construction [4]. We can define the Thompson NFA [5] for our algorithm as it's not a conceptual decision and can be changed even to the derivative constructions by Berry-Sethi [10] or other algorithms [6]. The use of NFA instead of DFA for local search isn't principal as it will be shown further.

The preliminary and the key idea is to define the context of captured groups and back-references in regular expression as following set of positions in the input string for each of the states in NFA:

$$Context(S) = \{ (L_i, R_i), (B_j, E_j) \}, i = 1..N, j = 1..M, \quad (2)$$

where the N is the number of capturing groups, M is the number of back-references and the values in the set L and R are the positions of matched captures for the corresponding back-reference in the searched text – B and E.

To use the memoization principle we will extend the particular polynomial matching algorithm or even family of such algorithms [5] with the context principle where each context is dependent on the input string position is stored in each reached state during the simulation, while the modern solution can solve in general more effectively as to the memory consumption [7].

Thus, we obtain the polynomial results for our algorithm according to our results, when its complexity depends on the state of context in the state of simulated NFA as the stored set (2) is saved in memory – thus giving the possibility to exclude repeated intervals in the input during the pattern matching process.

In this work for simulation purposes of local search we present the NFA with counter edges, where the counter is decreased by one at each step of the matching process, when the position in input is increased by one, and is included in the set of active states when the configuration complies to the present input, as well, as initialized to the counter value of the matched sub-string and is included for candidate when the counter value equals zero.

The counter for the set of states S in automaton can be defined as follows:

$$Counter(S) = \{ (S_i, C_i) \}, C_i \text{ in } N+, \quad (3)$$

where Si is the state of non-deterministic automaton and Ci is the corresponding counter value.

Thus, the local search even for extended constructions [8] can be achieved by implying the counters in (3) and simulating the NFA within the active zero value when the state can be used further in the non-local search before it will be pushed to the structure like stack which is typically used in the simulation of NFA along the input string. The back-references and extended operators [8] are to be matched locally with respect to the counter and state composition (3) when it's pushed to the stack for processing with the newly acquired counter value and, when, it equals to zero within the process of matching simulation, it becomes active or popped out from the same stack and forwarded to the matching procedure.

We obtain the following algorithm of simulating the local search with delay using the counter values (3):

Algorithm-1. The simulation of NFA with the local delay using counter.

```
define stack T;

define set of states S;

define the input string P;

T := {};

S := { Starting state of automaton };

for i in 1..|P| loop

for s in S loop

        if matched(P[i..j]): T.push ((s, |P[i..j]|));

        S.pop (s);

end loop

for t in T loop

        t.C := t.C – 1;

        if t.C = 0 then

                T.pop (t);

                S := S + { t.S };

        end if

end loop

end loop
```

The above algorithm procedure works in at most quadratic time and memory as it's bound to all the sub-strings from the input.

The Algorithm-1 can be extended for the back-references in “matched” function when the context (2) is used to compare the captured group and the current input for equivalence after which the state is to be considered.

Finally we state that the local search works on the extended NFA within the sub-program Algorithm-1 as follows:

$$<A, S, T, s_0, F, Context(S), Counter(S)> \quad (4)$$

The automaton (4) can be reduced only to counters if the back-references aren’t present.

Thus, we obtain the principally new results for simulating NFA for extended operators like intersection subtraction and complement [8] as well as for a back-reference within the space of counter states and local search using structures in (2) and (3). These structures are to be used for simulating NFA for effective memoization of the captured groups and back-references (2) and for pop-operation delay, according to the matched string during local search using the structure (3).

As back-reference problem is considered to be NP-complete – we solve it in polynomial time using the extra memory consumption using memoization principle and delay processing.

The class of NP is to be considered exponential or even factorial. The backtracking algorithm for matching without complexity optimization gives almost factorial time, while in our method the memoization and counter state space is used which is limited to the quadratic of input string as all the sub-strings are to be considered for matching.

In [8] using the algorithm from [4] the construction of deterministic finite automata (DFA) is presented. However, due to the state explosion problem these DFA cannot be universal and are bound to special cases where the explosion occurs. In this work we give the definition of local search with structures (2) and (3) which can be used in NFA which is linear in size to the regular expression with features like extended operators (intersection, complement and subtraction) and back-references.

The features like pre-matching [9] can be also modeled in NFA instead of building the DFA with re-written structure using &-operator [8].

In general the local search is also effective as other algorithms [11] when computed according to the specific cases.

So as to NP-completeness as we use quadratic memory in pattern matching routine the following statement of the complexity classes holds true:

$$P = NP: M \text{ in } P^2, \quad (5)$$

where M is the memory factor.

Conclusion

Despite the fact that our idea is based upon the memory consumption and the complexity reduction to the polynomial order, there are open questions concerning the determinism and the configuration of reached states, which can change the behaviour of the matching process. This question remains open, however, we can be sure that local search with counter heuristics solves the problem of in-memory matching.

The other point of view is the equality of P and NP classes with the use of memory (5) in computations for the NP-complete back-reference problem.

References

[1] Hassan, Mohammed Ahmed Ezzelregal, and Mohamed Elsayed Hasan. "Finding a Tandem Repeats Motifs in the Completed Genomes of Human Coronavirus (hku1) Which is Identified as a Hotspot Region for the Viruses Recombination by Using Regular Expression Language." (2019).

[2] van der Merwe, Brink, et al. "Memoized Regular Expressions." *International Conference on*

Implementation and Application of Automata. Springer, Cham, 2021.

[3] Berglund, Martin, Frank Drewes, and Brink Van Der Merwe. "Analyzing catastrophic backtracking behavior in practical regular expression matching." *arXiv preprint* arXiv:1405.5599 (2014).

[4] Rabin, Michael O., and Dana Scott. "Finite automata and their decision problems." *IBM journal of research and development* 3.2 (1959): 114-125.

[5] Thompson, Ken. "Programming techniques: Regular expression search algorithm." *Communications of the ACM* 11.6 (1968): 419-422.

[6] Champarnaud, J-M., J-L. Ponty, and Djelloul Ziadi. "From regular expressions to finite automata." *International journal of computer mathematics* 72.4 (1999): 415-431.

[7] Laurikari, Ville. "NFAs with tagged transitions, their conversion to deterministic automata and application to regular expressions." *Proceedings Seventh International Symposium on String Processing and Information Retrieval*. SPIRE 2000. IEEE, 2000.

[8] Syzdykov, Mirzakhmet. "Deterministic automata for extended regular expressions." *Open Computer Science* 7.1 (2017): 24-28

[9] Berglund, Martin, Brink van der Merwe, and Steyn van Litsenborgh. "Regular Expressions with Lookahead." *Journal of Universal Computer Science* 27.4 (2021): 324-340.

[10] Berry, Gerard, and Ravi Sethi. "From regular expressions to deterministic automata." *Theoretical computer science* 48 (1986): 117-126.

[11] Roşu, Grigore. "An effective algorithm for the membership problem for extended regular expressions." *International Conference on Foundations of Software Science and Computational Structures.* Springer, Berlin, Heidelberg, 2007.

[12] Cook, Stephen. "The P versus NP problem." *The millennium prize problems* (2006): 87-104.

[13] Alevoor, Praveen, Pratik Sarda, and Kalpesh Kapoor. "On Decidability and Matching Issues for Regex Languages." *Proceedings of International Conference on Advances in Computing*. Springer, New Delhi, 2013.

[14] Yu, Xiaodong, Bill Lin, and Michela Becchi. "Revisiting state blow-up: Automatically building augmented-fa while preserving functional equivalence." *IEEE Journal on Selected Areas in Communications* 32.10 (2014): 1822-1833.

[15] Held, Michael, and Richard M. Karp. "A dynamic programming approach to sequencing problems." *Journal of the Society for Industrial and Applied mathematics* 10.1 (1962): 196-210.

UDC 004.8, 004.9
IRSTI 28.23.29, 28.23.37

CALCULATION OF FILTRATION CHARACTERISTICS OF WELLS URANIUM DEPOSITS

Yunicheva N.R., Kuchin Y., Mukhamedieva E.L.
Institute of Information and Computational Technologies KN MES RK, Almaty
e-mails: naduni@mail.ru , ykuchin@mail.ru, muhamedijeva@gmail.com
ORCID ID: https://orcid.org/0000-0001-6351-3450,
ORCID ID: https://orcid.org/0000-0002-5271-9071

Abstract. Possession of information about the filtration characteristics of the host rocks allows you to plan the volume of ore production. The existing methodology for calculating the filtration properties of wells for uranium extraction by the method of underground borehole leaching is based on a system of rules that take into account only one geophysical parameter (apparent resistance-CS). The method gives a relatively low accuracy of calculating the filtration coefficient. In addition, it is not applicable in the case of defects in the recording of CS or distortion of values under the action of acid, which is widely used in uranium mining. At the same time, other geophysical parameters can be used to calculate the filtration coefficient, the use of which can increase the accuracy of the calculation by 10-20% when using machine learning methods.

The paper assumes the study of the applicability of machine learning methods for predicting the filtration properties of the host rocks of wells for uranium extraction by the PSV (underground well leaching) method and a numerical assessment of the advantages of such an approach. A machine learning method will be developed to predict the filtration properties of rocks based on inaccurately marked electrical logging data.

Keywords: Machine learning, geophysical studies of wells, filtration coefficient, inaccurate data.

1. Introduction

All over the world, uranium is the main resource for the operation of nuclear power plants. Deposits of uranium ores are not evenly distributed around the globe. Today, only 28 countries of the world extract valuable raw materials in their bowels. The main world reserves of uranium in the world are located in 8 countries. We will tell you a little more about the countries with the largest uranium reserves.

Figure 1 shows the countries with the largest uranium reserves according to [1].

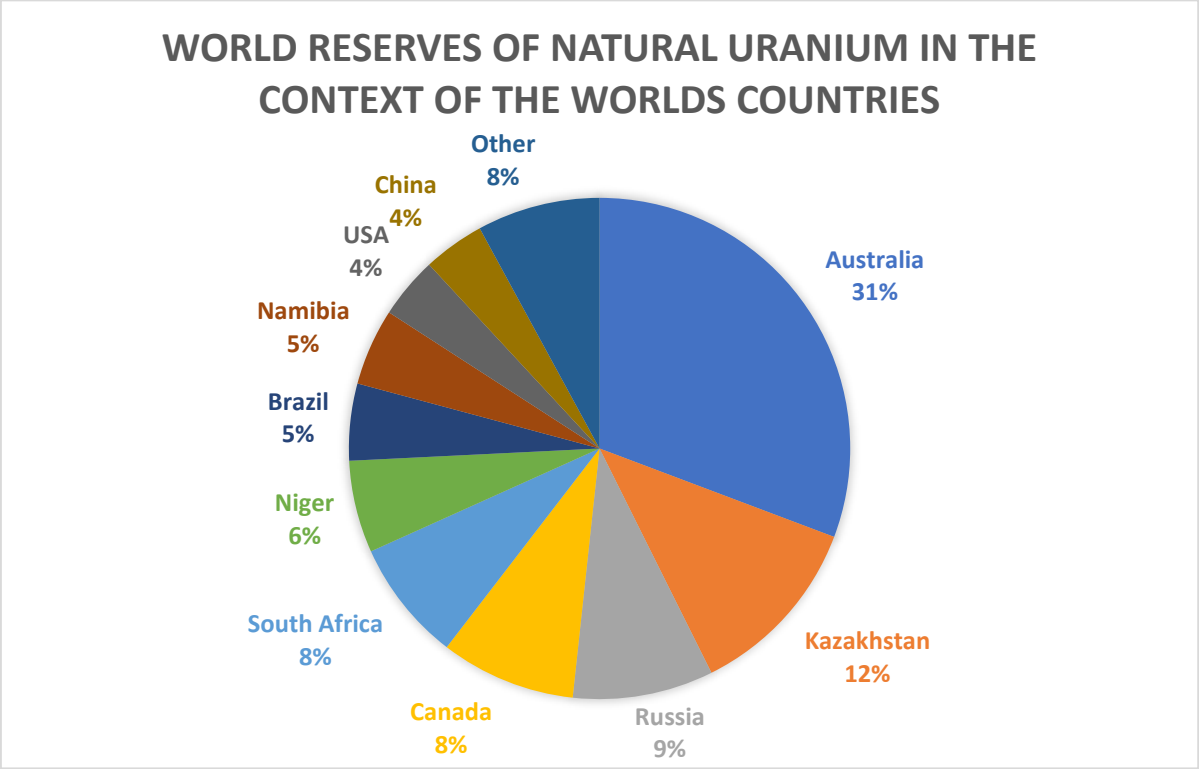

Figure 1 - World reserves of natural uranium in the context of the worlds countries

Analysis of information related to the leadership of countries in uranium reserves does not allow us to conclude that these same countries are leaders in uranium mining.

In 2018, the world's largest uranium miners produced 86% of the world's uranium mined, according to the World Nuclear Association. The main uranium mining companies are mining corporations from Kazakhstan, Canada, Australia: they account for two-thirds of the world's production.

Kazakhstan accounts for 39% of global uranium production. Production is carried out using the underground well leaching method (UWL). In this method, uranium is extracted through a network of pumping and injection wells, through which the leaching solution is circulated. For uranium production planning, the filtration properties of the host rocks are an important characteristic. Knowledge of the filtration properties in the form of a filtration coefficient allows planning the volume of ore extraction. The currently used method for calculating the filtration properties of wells for uranium production by the in-situ leaching method is based on a system of rules that takes into account only one geophysical parameter (apparent resistance). However, this method is not applicable in the case of a defect in the recording of apparent resistance or distortion of values under the influence of acid, which is widely used in uranium mining. However, other geophysical parameters may also be used to calculate the filtration coefficient, the use of which may improve the accuracy of the calculation. Multiple parameters can be taken into account with actual measurements using machine learning models (MLM).

MLM are divided into two groups [2, 3, 4]:

– *Unsupervised Learning (UL)* [5].

– *Supervised Learning (SL)* [6].

UL solve the clustering problem when a multitude of previously unmarked objects are grouped by an automatic procedure based on the properties of these objects [7, 8].

SL is used when a potentially infinite number of objects distinguish final groups (classes) of some marked objects. The algorithm uses this initial classification as a model and marks out the unmarked objects and assigns them to a suitable group (class) based on the properties of these objects. Although such algorithms are known, it is not trivial to select or develop a classifier for a particular practical task. What is required is a set of computational experiments, significant

preliminary work on data preparation, etc.

For many years, the focus of machine learning researchers has been on petroleum geophysics, where the main task is to isolate reservoirs. In uranium geophysics, the standard set of studies contains far fewer methods, while at the same time it is necessary to determine the properties of many fairly small alternations. It makes the application of MLMs in uranium deposits more difficult than in oil geophysics.

The task of determining filtration properties from log data refers to supervised learning or learning by example. The mathematical problem of example-based learning should be regarded as an optimization task which can be solved by searching for the minimum value of the cost function $J(\theta)$ in all available examples, defined as the sum of squares of the difference between the "predicted" value and the real value of y in a numerious examples *m*. Hypothesis $h_\theta(x)$ is selected, which, with a certain set of parameters $\theta_i \in \Theta$ provides a minimum value $J(\theta)$:

$$J(\theta) = min\frac{1}{2m}\sum_{i=1}^{m}(h_\theta\left(x^{(i)}\right) - y^{(i)})^2 \quad (1.1)$$

where *m* – numerious examples, h_θ – function of the hypothesis, which can be linear ($h_\theta = \theta_0 + \theta_1 x$) or non-linear (example, $h_\theta = \theta_0 + \theta_1 x + \theta_2 x^2$) with a different set of parameters $\theta_i \in \Theta$.

In order to find the optimal function $h_\theta(x)$ a gradient descent algorithm is used, the essence of which is the sequential change of θ_0, θ_1 parameters using an expression

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1) \quad (1.2)$$

where α – learning parameter, and $\frac{\partial}{\partial\theta_j}J(\theta_0,\theta_1)$ is a derivative of the cost function of θ_j. The sign := means an assignment as opposed to an equal sign (=) in algebraic expressions.

In this process, the algorithm steps are performed in such a way that both parameters are first changed simultaneously based on expression 1.2, and only then new values are assigned to them. In other words, the algorithm sequence of one algorithm step for the case of two parameters, expressed in pseudo-code, will be the next:

$temp\theta_0 := \theta_0 - \alpha\frac{\partial}{\partial\theta_0}J(\theta_0,\theta_1)$;

$temp\theta_1 := \theta_1 - \alpha\frac{\partial}{\partial\theta_1}J(\theta_0,\theta_1)$;

$\theta_0 := temp\,\theta_0$;

$\theta_1 := temp\theta_1$.

With regard to differentiation

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_1^{(i)}$$

Several algorithms are useful for conducting the experiments, for example, ElasticNet (linear regression with L1 and L2 regularisations), gradient boosting (an ensemble of weak solution trees aggregated in a meta-busting model), and a neural network with a hidden layer.

ElasticNet [9] implements a model that combines two types of regulators: linear and square. Expression of the cost function for ElasticNet

$$J(\theta) = \min \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \quad + \alpha * l1 * \| \theta \| + 0.5 * \alpha * (1 - l1) * \theta^2$$

In this case, if l1=0, he regularisation is only carried out using the square value of the parameters (l2), if l1=1, only the linear regulator (l1) is used, in other cases, when 0<l1<1 both types are combined. In this case

$$\alpha * l1 * \| \theta \| + 0.5 * \alpha * (1 - l1) * \theta^2$$

can be expressed as a sum

a * l1 + b * l2 ,

where alpha = a + b и l1 = a / (a + b).

The essence of gradient boosting [10] s that, after calculating the optimum regression coefficient values and obtaining the hypothesis function $h_\theta(x)$ c using some algorithm (a), the error is calculated and a new function $h_{b\theta}(x)$ is selected, possibly using another algorithm (b), so that it minimizes the error of the previous one.

$$h_\theta(x^{(i)}) + h_{b\theta}(x^{(i)}) - y^{(i)} \to \min$$

In other words, it is about minimizing the function

$$J_b = \sum_{i=1}^{m} L(y^{(i)}, h_\theta(x^{(i)}) + h_{b\theta}(x^{(i)}))$$

where L – error function that takes into account the results of algorithms a and b. If $J_b(\theta)$ is still large, the third algorithm (c) is selected, and etc. Often as algorithms (a), (b), (c) and etc., relatively shallow solution trees are used. The value of the error function gradient is used to find the minimum of function $J_b(\theta)$

$$L(y^{(i)}, h_\theta(x^{(i)}))_{i=1}^{m}$$

Given that the minimization of the function $J_b(h_{b\theta}(x^{(i)}))_{i=1}^{m}$ is achieved in the direction of the antigradient of the error function, algorithm (b) is configured so that the target values are not $(y^{(i)})_{i=1}^{m}$, but the antigradient ($-L'(y^{(i)}, h_\theta(x^{(i)}))_{i=1}^{m}$), i.e. the $(x^{(i)}, -L'(y^{(i)}, h_\theta(x^{(i)}))$ pairs are used instead of $(x^{(i)}, y^{(i)})$ pairs in training algorithm (b).

Multilayer artificial neural networks (multilayer perceptron) are one of the most popular methods of supervised learning, especially for multiple classes. The following cost function is used to set the θ neural network weight:

$$J(\Theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(h_\Theta(x^{(i)}))_k + (1 - y_k^{(i)}) \log(1 - h_\Theta(x^{(i)}))_k] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{S_l}\sum_{j=1}^{S_{l+1}}(\Theta_{ji}^{l})^2 ,$$

where L – number of neural network layers; S_l – number of neurons in layer l; K – number of classes (equal to the number of neurons in the output layer); Θ – weight matrix, and the hypothesis function is often a sigmoidal (logistic) function $h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$,

To minimize the error (learning) function of the multilayer ANN, the Back Propagation Error (BPE) algorithm [11] and its modifications are used to speed up the learning process.

In the execution of the project, it is planned to determine the algorithm and its parameters, which will give the best result in predicting the filtration coefficient, by performing computational experiments using interval inaccurate data. The complexity of this process lies in its high variability. For example, if the researcher has large data volumes and significant computing power at his disposal, deep learning methods that solve complex machine learning tasks with high

precision can be used. On the other hand, if the number of examples is less than the number of properties, it is difficult to use support vector machines (SVM) because they are prone to retraining in this case. Thus, by selecting a certain number of methods to solve a task and changing their parameters (e.g. regularisation coefficient, number of layers of neural networks, etc.), it is necessary to assess the results of their work using one or more indicators, e.g. percentage (fraction) of correctly classified examples (accuracy). For the evaluation of 1st and 2nd type errors, the following are used: precision, recall, and the aggregates are F1 Score, Precision-Recall, and ROC. The following are used to assess the quality of regression dependence: r2_score, mean_absolute_error (absolute difference of target & predicted values), mean_squared_error (squared difference of target & predicted values), median_absolute_error (robust to outliers).

There are currently no systems in place to intelligently interpret well logging data from uranium deposits for the purpose of calculating filtration characteristics. The method developed as a result of the project will make it possible to determine the value of the filtration coefficient with greater accuracy. Preliminary experiments suggest that the use of MLM will increase the accuracy of the calculation by 10-20%.

2. Description of computational experiments and results

As the initial data, the results of hydrogeological studies of wells performed at the exploration stage were used as the only source of actual data on the filtration properties of rocks.

Computational experiments include data preprocessing, the formation of training and test data sets, and the calculation of the parameters of a classification or regression model. Data preprocessing has special features in this case, since interval samples obtained during electrical logging of wells serve as data sources. The formation of a training and test data set also has special features, since they must be formed taking into account the downhole data splitting and data splitting into folds.

To find the best regression parameters and regularization coefficients, a grid-search was performed using the parameters of three algorithms with the recording of the results of testing cross-validation methods.

- A set of models has been formed based on the gdir-search principle.
- Testing was carried out, as a result of which model quality indicators were obtained, including the standard error and determination coefficient, after which the models were sorted by these indicators on a test sample and the top of the best models with optimal meta-parameter settings were selected.
- To train the model, a dataset was formed containing data on the Budennovskoye field.
- Splitting into folds for cross-validation was performed in two ways:

The usual stratified splitting into folds, which ensures that the number of objects with different target values (the value of the filtration coefficient) is proportional to the total sample in the test and training sample

Splitting into folds by wells — with this approach, each individual well was separated into the test sample in order to exclude the possibility of retraining the model for specific physical and filtration properties of individual wells, since a simple stratified splitting into folds cannot provide the calculation of indicative metrics of the assessment quality in such cases.

- The search of regressors and their parameters took place according to the grid search principle, in which the significant parameters of the classifier are selected and their search takes place in a given interval and with a given step, after which the model is selected that has optimal quality indicators on test samples.

Table 1 shows the regressors that showed the best results on different sets of input data.

Table 1. Results of regressors on different sets of input data

	Перрессор	Correlation train	Correlation test	MSE train	MSE test
КС	MLP Regressor - one hidden layer with 201 neurons, 600 max iterations	0,494	0,613	32,389	35,872
КС+Код породы	MLP Regressor - one hidden layer with 201 neurons, 850 max iterations	0,502	0,488	30,968	38,384
КС+ПС	MLP Regressor - one hidden layer with 201 neurons, 600 max iterations	0,596	0,535	25,785	30,472
КС+ПС+ Код породы	MLP Regressor - one hidden layer with 251 neurons, 100 max iterations	0,698	0,570	21,244	30,363
КС	Расчет по принятой методике		**0,261**		**48,32**

Conclusion

Thus, the applied approach allowed us to evaluate the filtration properties of rocks better than the existing method. It is revealed that the correlation with the actual data is higher and less than the standard error. The expansion of the dataset due to logging data for a larger number of wells will increase the accuracy of the model.

Acknowledgements

This research has been funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant No. AP09562260 "Development of a method for calculating the filtration characteristics of wells of uranium deposits based on machine learning algorithms working with inaccurately marked interval data".)

References:

[1] URL: https://eenergy.media/2018/11/06/zapasy-urana-u-kakih-stran-oni-samye-bolshie (13.07.2021).

[2] Ayodele T. O. Types of machine learning algorithms. New advances in machine learning. 2010. 3. 19-48.

[3] Ibrahim H. A. H. et al. Taxonomy of machine learning algorithms to classify real time interactive applications. *International Journal of Computer Networks and Wireless Communications*. 2012. 2(1). 69-73.

[4] Muhamedyev R. Machine learning methods: An overview. Computer Modelling & New Technologies. 2015. 19(6). 14-29.

[5] Hastie T., Tibshirani R., Friedman J. Unsupervised learning. The elements of statistical learning. Springer, New York, NY, 2009. 485-585.

[6] Kotsiantis S. B., Zaharakis I., Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007. 160(1). 3-24.

[7] Jain A. K., Murty M. N., Flynn P. J. Data clustering: a review. ACM computing surveys (CSUR). 1999. 31(3). 264-323.

[8] Ashour B. W., Ying W., Colin F. Review of Clustering Algorithms. Non-Standard Parameter Adaptation for Exploratory Data Analysis. Studies in Computational Intelligence. 2009. 249. 7-28.

[9] ElasticNet. [Electronic resource]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html.

[10] Friedman J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*. 2001. 1189-1232.

[11] Werbos P. J. New tools for prediction and analysis in the behavioral sciences. Ph. D. dissertation, Harvard University. 1974. 38.

Responsible for the release: PhD, Shayakhmetova A.S.

Merkebayev A.

Deputy chief editor: PhD, Mamyrbayev O.Zh

Signed in print 03.03.2021
Edition of 50 copies. Format 60x84 1/16. Paper type.
Order No. 4.

Publication of the Institute of Information and Computational Technologies

28 Shevchenko str., Almaty, Republic of Kazakhstan
7 (727) 272-37-11
atcs@iict.kz