



Institute of Information and
Computational Technologies

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

ADVANCED TECHNOLOGIES AND **COMPUTER SCIENCE**

2023
No1

www.atcs.iict.kz

Institute of Information and Computational Technologies

Advanced Technologies and computer science

№1

Almaty 2023

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

Institute of Information and Computational Technologies,

Advanced Technologies and computer science

This journal is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The publisher, the authors, and the editors are safe to assume that the advice and information in this journal are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published works and institutional affiliations.

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.

Editorial Team

Chief Editor

Kalimoldayev M.N., Academician of NAS RK, Doctor of Physics and Mathematics, Professor, DG of RSE "Institute of Information and Computational Technologies" SC MES RK (Kazakhstan)

Deputy chief editor: PhD, Mamyrbayev O.Zh (Kazakhstan)

Editorial team

- Amirgaliev Ye.N., Doctor of Technical Sciences, Professor, Kazakhstan
- Arslanov M.Z., Doctor of Physics and Mathematics, Professor, Kazakhstan
- Berdyshev A.S., Uzbekistan
- Biyashev R.G., Doctor of Technical Sciences, Professor, Kazakhstan
- Ischukova Ye.A., Candidate of Technical Sciences, Docent, Russia
- Krak, Ukraine
- Posypkin M.A., Doctor of Physics and Mathematics, Russia
- Khairova N.F., Doctor of Technical Sciences, Ukraine
- Keylan Alimhan, Japan (Tokyo Denki University)
- Marat Ahmet, Turkey
- Mohamed Othman, Малайзия (Universiti Putra Malaysia)
- Naohisa Otsuka, Japan (Tokyo Denki University)
- Ravil Muhamediev, Latvia
- Waldemar Wójcik, Poland

Contents

Splines interpolation analysis using maple package Aqila Karimi, Yalda Qani	4
Review of perspectives of programming olympiads in kazakhstan M.Syzdykov	10
Интеллектуалдық-ақпараттық технологиялар саласындағы инновациялық әзірлемелер және оларды бағдарламалық іске асыру негізінде бизнес процестерді болжау әдістеріндегі жаңа енгізілімдер А.М. Ахметова, А.С. Шаяхметова, Д.Ө. Жақан	16
Технологии организации хранения и запроса больших коллекций XML- документов А.А. Мухитова, А.С. Еримбетова, В. Баракнин	23
Survey of modern trends in computer security and artificial intelligence M. Syzdykov	29

UDC 519.6
IRSTI 27.41.77

SPLINES INTERPOLATION ANALYSIS USING MAPLE PACKAGE

Aqila Karimi, Yalda Qani

Faryab University, Maymana, Afghanistan

e-mail: yalda.qani@gmail.com

ORCID ID: <https://orcid.org/0000-0002-3030-5783>

Abstract Expressing relationships in data with functions is both useful and convenient. It allows us to estimate dependent variables at values of independent variables not given in the data, take derivatives, integrate, and even solve differential equations. This is done by interpolation in a precise way. Interpolation is a set of function points that passes through each data point. Linear polynomial interpolation, cubic spline, Lagrange and Newton are common interpolation methods. In interpolation problems, spline interpolation is often preferred over polynomial interpolation because it gives similar results even when using low-order polynomials, while avoiding the phenomenon common at higher orders. We used simple linear functions to introduce some basic concepts and issues related to spline interpolation. Then we derive an algorithm for fitting quadratic lines to the data. Finally, we present material on the cubic spline, which is the most common and useful version in engineering practice. Cubic spline is easy to display and calculate by Maple package. Because Maple is closed, the same math symbols used in classrooms can be used to enter data. In addition, the Maple package has many features, including converting outputs to MATLAB codes and LaTeX commands, where computational problems related to the subject are presented with precise and explicit answers.

Keywords: cubic spline, interpolation, Maple package, natural spline.

Introduction

Before examining spline interpolation, we first discuss about spline polynomials because in interpolation, the interpolation space is always involved and this space has foundations: for example, in Lagrange interpolation, which is a special type of polynomial interpolation, several Lagrange's terms form the basis of space, which is also the case in other types. Regarding spline interpolation, it can be said that this interpolation is a linear combination of spline polynomials such as $S_i(x)$, which is introduced below:

$$p(x) = \sum_{i=1}^n C_i S_i(x)$$

Once $S_i(x)$, is known, it suffices to obtain C_i . In this case, the interpolation polynomials are discovered. For this purpose, we first discuss the space consisting of spline polynomials [1].

A spline is a function which, together with its several derivatives, is continuous on $[x_0, x_n]$ and is such that on each separate subinterval $[x_i, x_{i+1}]$ it is some algebraic polynomial. The terminology was introduced by I. J. Schoenberg in 1946, although the idea was used earlier informally by several authors. Technically a 'spline' is a draughtsman's flexible implement which is used to draw 'smooth' curves through a series of points [2]. Splines were first introduced in 1946 by a person named I. J. Schoenberg [1]. The term "spline" originated from a thin, lexible strip, known as a spline, used by draftsmen to draw smooth curves over a set of points marked by pegs or nails. The data points at which two splines meet are called knots. The most commonly used splines are cubic splines, which produce very smooth connections over adjacent intervals [6]. Polynomial interpolation is an ancient practice, but the heavy industrial use of interpolation began with cubic splines in the 20th century [8]. The much broader application of splines to the areas of data fitting and computer-aided geometric design became evident with the widespread availability of computers in the 1960s [3]. Interpolation actually describes the problem of finding a curve (called an interpolation) that passes through a given set of real values f_0, f_1, \dots, f_n at real data points x_0, x_1, \dots, x_n , which are sometimes called abscissae or nodes [10]. Spline becomes very handy when considering functions that are used in applications requiring data interpolation

and/or smoothing. Most graphic user interface software makes use of the spline interpolation for data such as ASCII and image dataset. The interesting part of spline interpolation is its flexibility to fits low-degree polynomials to small subsets of the values [5].

Materials and methods

Because of these tasks, spline, cubic spline and B-spline interpolation were carefully calculated and analyzed using the following methods in MAPL.

Spline

We precisely defined the specified data points using the Maple Spline package. The result is a piecewise polynomial. We also specified the independent variable and then defined the spline. As you can see, we enter a list of data points in the first step to analyze the issue accurately.

> [[-5, 5], [0, 2], [4, 0]]

[[-5, 5], [0, 2], [4, 0]]

Now we specify the independent variable and then define the spline.

> s := CurveFitting[Spline]((1), x)

$$s := \begin{cases} 2 - \frac{49}{90}x + \frac{1}{60}x^2 + \frac{1}{900}x^3 & x < 0 \\ 2 - \frac{49}{90}x + \frac{1}{60}x^2 - \frac{1}{720}x^3 & \text{otherwise} \end{cases}$$

Now if we want to calculate a spline with specified end conditions, in that The Spline routine computes a degree d piecewise polynomial in variable v that approximates the points $\{(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)\}$. It should be said that if v is a numerical value, the polynomial value is returned at this point. The default value of d is 3. This can be changed using the degree=d option.

The main purpose of the Spline function is to calculate and return the piecewise formula of a spline interpolant. On numerical data, to perform fast spline interpolation, the Curve Fitting [Array Interpolation] command should be used instead. We can call the Spline routine in two ways. First, Spline (xydata, v, dgr, endpoints) accepts a list, array, or matrix, $[[x_0, y_0], [x_1, y_1], \dots, [x_n, y_n]]$ of data points.

Examples

> with(CurveFitting) :

> Spline([[0, 0], [1, 1], [2, 4], [3, 3]], v)

$$\begin{cases} \frac{4}{5}v^3 + \frac{1}{5}v & v < 1 \\ -2v^3 + \frac{42}{5}v^2 - \frac{41}{5}v + \frac{14}{5} & v < 2 \\ \frac{6}{5}v^3 - \frac{54}{5}v^2 + \frac{151}{5}v - \frac{114}{5} & \text{otherwise} \end{cases}$$

> Spline([[0, 0], [1, 5], [2, -1], [3, 0]], v, degree = 2, endpoints = 'periodic')

$$\begin{cases} \frac{16}{5}v^2 + \frac{24}{5}v & v < \frac{1}{2} \\ -\frac{44}{5}v^2 + \frac{84}{5}v - 3 & v < \frac{3}{2} \\ \frac{28}{5}v^2 - \frac{132}{5}v + \frac{147}{5} & v < \frac{5}{2} \\ \frac{16}{5}v^2 - \frac{72}{5}v + \frac{72}{5} & \text{otherwise} \end{cases}$$

Returns the linear system of equations used to solve the cubic spline interpolants.

Description

- The LinearSystem command retrieves the matrix and vector in the linear system of equations that were solved when computing the cubic spline interpolants.

- To calculate the cubic spline interpolants, the linear system of equations was solved and the matrix and vector were recovered using the LinearSystem command.
- A POLYINTERP structure is created using the CubicSpline command.
- The LinearSystem command only accepts interpolation structures created using theCubicSpline command, because the cubic spline interpolation method is the only method that has an associated linear system [https://www.maplesoft.com/].

Examples

> with(StudentNumericalAnalysis) :

> xy := $\left[[0, 1], \left[\frac{1}{2}, 1 \right], \left[1, \frac{11}{10} \right], \left[\frac{3}{2}, \frac{3}{4} \right], \left[2, \frac{7}{8} \right], \left[\frac{5}{2}, \frac{9}{10} \right], \left[3, \frac{11}{10} \right], \left[\frac{7}{2}, 1 \right] \right]$
 $xy := \left[[0, 1], \left[\frac{1}{2}, 1 \right], \left[1, \frac{11}{10} \right], \left[\frac{3}{2}, \frac{3}{4} \right], \left[2, \frac{7}{8} \right], \left[\frac{5}{2}, \frac{9}{10} \right], \left[3, \frac{11}{10} \right], \left[\frac{7}{2}, 1 \right] \right]$

> p1 := CubicSpline(xy, independentvar = 'x') :

> LinearSystem(p1)

$$\begin{pmatrix} 1. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. \\ 0.5000000000 & 2. & 0.5000000000 & 0. & 0. & 0. & 0. & 0. & 0. \\ 0. & 0.5000000000 & 2. & 0.5000000000 & 0. & 0. & 0. & 0. & 0. \\ 0. & 0. & 0.5000000000 & 2. & 0.5000000000 & 0. & 0. & 0. & 0. \\ 0. & 0. & 0. & 0.5000000000 & 2. & 0.5000000000 & 0. & 0. & 0. \\ 0. & 0. & 0. & 0. & 0.5000000000 & 2. & 0.5000000000 & 0. & 0. \\ 0. & 0. & 0. & 0. & 0. & 0.5000000000 & 2. & 0.5000000000 & 0. \\ 0. & 0. & 0. & 0. & 0. & 0. & 0.5000000000 & 2. & 0.5000000000 \\ 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 1. \end{pmatrix} \begin{pmatrix} 0. \\ 0.8291995878 \\ -2.116798351 \\ 2.237993817 \\ -1.135176915 \\ 1.102713844 \\ -1.175678461 \\ 0. \end{pmatrix}$$

Linear Splines

The simplest connection between two points is a straight line. First-order contours for a group of ordered data points can be defined as a set of linear functions [4]. With linear splines, straight lines (linear functions) are used for interpolation between the data points [6]. An arbitrary collection of n data points can be successfully interpolated by the linear spline. Linear splines, however, are not smooth. This flaw in linear splines is intended to be fixed by cubic splines. Degree 3 (cubic) polynomials are used in place of linear functions between the data points by a cubic spline [8].

The B-Spline curve

The B-Spline curve routine calculates a B-Spline curve based on control points. The B-Spline routine computes a piecewise function representing the kth order B-spline in the symbol v. The non-zero parts of this function are polynomials of degree $k - 1$. If the knots option is not provided, the uniform knot list $[0, 1, \dots, k]$ is used. The node list must contain exactly $k + 1$ elements. These elements must be in non-decreasing order. Otherwise, unexpected results may be produced. Nodes can have a number of up to $k - 1$. If the number of a node is m, the continuity in that node is $C(k - m - 1)$. This method returns a B-spline basis function.

Use the Curve Fitting[BSplineCurve] procedure to create a B-spline curve. This function is part of the Curve Fitting package and therefore it can be used in the form of BSpline(.) only after executing the command with (Curve Fitting). However, it can always be accessed via the long form of the command using Curve Fitting[BSpline] [https://www.maplesoft.com/].

Examples

> with(CurveFitting) :

> BSpline(2, u)

$$\begin{cases} 0 & u < 0 \\ u & 0 \leq u < 1 \\ 2 - u & 1 \leq u < 2 \\ 0 & 2 \leq u \end{cases}$$

> BSpline(2, u, knots = [0, a, 2])

$$\left\{ \begin{array}{ll} 0 & u < 0 \\ \frac{u}{a} & 0 < u < a \\ \frac{-u+a}{2-a} + 1 & a < u < 2 \\ 0 & 2 \leq u \end{array} \right.$$

Cubic spline

The concept of a spline originates from the drafting technique of using a thin, flexible strip (called a spline) to draw smooth curves through a set of points. In this technique, the draftsman places the paper on a wooden board and nails or pins into the paper (and the board) at the data points. A smooth cubic curve results from interweaving the tape between the pins. Hence, the name "cubic spline" has been adapted for polynomials of this type[3]. Cubic splines give useful alternatives to plain polynomials. Cubic splines are piecewise cubic functions s that are continuous and have continuous first and second derivatives [9]. The objective in cubic splines is to derive a third-order polynomial for each interval between knots [4]. Cubic splines may guarantee the continuity of the first and second derivatives at the nodes and are therefore more commonly used in practice [7]. In cubic splines, third-degree polynomials are used to interpolate over each interval between data points. Suppose there are $n + 1$ data points $(x_1, y_1), \dots, (x_{n+1}, y_{n+1})$ so that there are n intervals and thus n cubic polynomials. Each cubic polynomial is conveniently expressed in the form

$$S_i(x) = a_i(x - x_1)^3 + b_i(x - x_1)^2 + c_i(x - x_1) + d_i, \quad i = 1, 2, \dots, n$$

where a_i, b_i, c_i, d_i ($i = 1, 2, \dots, n$) are unknown constants to be determined [6]. The Cubic Spline command interpolates the given xy data points using the cubic spline method and stores all the computed information in a POLYINTERP structure. The POLYINTERP structure is then passed to various interpolation commands in the Student [Numerical Analysis] sub package, where information can be extracted from it and manipulated depending on the command. A residual term is not calculated by the Cubic Spline command, so it cannot be used in conjunction with the Remainder Term command or the Interpolant Remainder Term command. This method works numerically. That is, inputs that are not numeric are first evaluated to floating point numbers before continuing calculations [<https://www.maplesoft.com/>].

Examples

```
> with(Student[NumericalAnalysis]) :
> xy := [[0, 4.0], [0.5, 0], [1.0, -2.0], [1.5, 0], [2.0, 1.0], [2.5, 0], [3.0, -0.5]]
xy := [[0, 4.0], [0.5, 0], [1.0, -2.0], [1.5, 0], [2.0, 1.0], [2.5, 0], [3.0, -0.5]]
> p1 := CubicSpline(xy, independentvar = x) :
> expand(Interpolant(p1))

```

$$\left\{ \begin{array}{l} 4. - 8.48076923076923x + 1.92307692307692x^3 \\ -5.13461538461539x + 3.44230769230769 - 6.69230769230769x^2 + 6.38461538461538. \\ 21.2884615384615 - 58.6730769230769x + 46.8461538461538x^2 - 11.4615384615385x \\ 15.0576923076923x - 15.5769230769231 - 2.30769230769231x^2 - 0.538461538461538. \\ -64.8076923076923 + 88.9038461538461x - 39.2307692307692x^2 + 5.61538461538461. \\ -52.4423076923077x + 52.9807692307692 + 17.3076923076923x^2 - 1.92307692307692. \end{array} \right.$$

```
> Draw(p1)
```

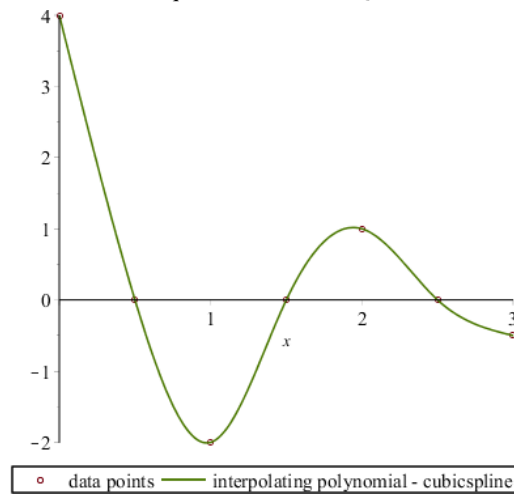


Figure 1 - Cubic spline interpolation natural boundary conditions.

```
> p2 := CubicSpline(xy, independentvar = x, boundaryconditions = clamped(0, 6)) :
> Draw(p2)
```

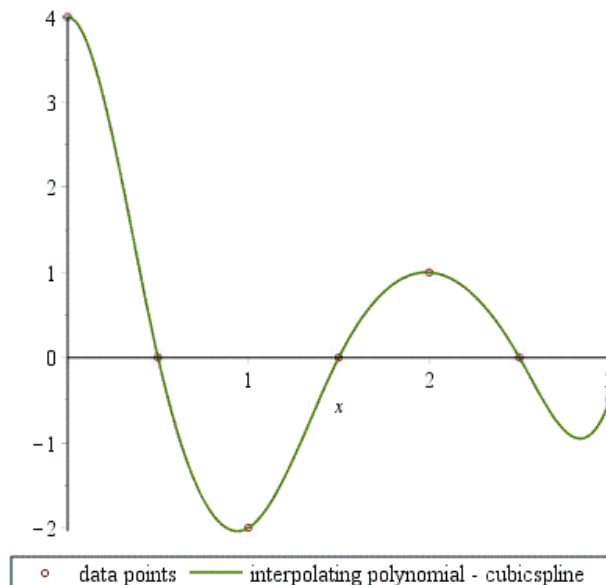


Figure 2- Cubic spline interpolation clamped boundary conditions.

Conclusion

Interpolation is a set of function points that passes through each data point, and expressing relationships in data with functions is both useful and convenient. This is done by interpolation in a precise way. In interpolation problems, spline interpolation is often preferred over polynomial interpolation because it gives similar results even when using low-order polynomials, while avoiding the phenomenon common at higher orders. The Maple package has good computing capabilities. We chose Maple package for accurate data analysis. From Maple package, you can use mathematical symbols used in classrooms to enter data. In addition, the Maple package has many features, including converting outputs to MATLAB codes and LaTeX commands, where computational problems related to the subject are presented with precise and explicit answers. Using the capabilities of the LinearSystem command, we recovered the matrix and vector in the linear system of equations that were solved when calculating the cubic spline interpolants. Also, to calculate the cubic spline interpolants, the linear system of equations was solved and the matrix and vector were recovered using the LinearSystem command. We also showed that a POLYINTERP structure is created using the CubicSpline command. The LinearSystem command only accepts interpolation structures created using the CubicSpline command, because the cubic spline interpolation method is the only method that has an associated linear system.

References

- [1] Allahviranloo T., Pedrycz W., Esfandiari A. Advances in Numerical Analysis Emphasizing Interval Data. CRC Press. 2022. 86.
- [2] Bose S. K. Numerical methods of mathematics implemented in Fortran. Springer Singapore. 2019. 197
- [3] Burden R. L., Faires J. D., Burden, A. M. Numerical analysis. Cengage learning. 2015. 143.
- [4] Chapra S. C., Canale R. P. Numerical methods for engineers, Seventh., 2015. 33.
- [5] Emetere, M. E. Numerical Methods in Environmental Data Analysis. Elsevier. 2022. 179.
- [6] Esfandiari R. S. Numerical methods for engineers and scientists using MATLAB®. Crc Press. 2017.
- 194
- [7] Li H. Introduction to Numerical Methods in Java. Numerical Methods Using Java: For Data Science, Analysis, and Engineering, 2022. 1-69. 249.
- [8] Sauer T. Numerical analysis. Addison-Wesley Publishing Company. 2011. 138.
- [9] Stewart D. Numerical Analysis: A Graduate Course. Springer. 2022. 258.
- [10] Faul A. C. A Concise introduction to numerical analysis. Chapman and Hall/CRC. 2018. 70.
- [11] <https://www.maplesoft.com/>

UDC 004.02
IRSTI 20.53.15

REVIEW OF PERSPECTIVES OF PROGRAMMING OLYMPIADS IN KAZAKHSTAN

Mirzakhmet Syzdykov

Satbayev University, Almaty, Kazakhstan
mspmail598@gmail.com

ORCID ID: <https://orcid.org/0000-0002-8086-775X>

Abstract. We present the research aimed at the current status of the programming Olympiads in which Kazakhstan is involved and holds a positive line, we propose various pros and cons for the training of the participants of these events. Since the participation of Kazakhstan in the International Olympiad in Informatics (IOI) for schoolers and International Collegiate Programming Contest (ICPC), which is upheld by Association of Computing Machinery (ACM), most of the better results were shown on IOI rather than ACM ICPC, where still participants from our country demonstrate stable tendency at gaining finals, however, still not receiving the prize pool consisting of the first twelve teams in the final rating. The problem which is to be solved by participants can be also classified as polynomial (P-complete) or non-polynomial (NP-complete), which, in turn, gives us the observation of what could be solved using rational solution. We demonstrate this fact on examples of two problems from ACM ICPC.

Keywords: Olympiad in Informatics, tendency, training, statistics.

Introduction

The problem which is to be solved can be classified as P or NP-complete. In fact, NP-complete problem can be truncated up to the fixed constant of the complexity factor of the size of the input data and, thus, forwarded further for being solved in polynomial time using brute force algorithm, which means the evaluation of the whole state space of input data within the composition of the problem statement. This problem can be also alternatively solved by using approximate algorithmic methods like Ant Colony Optimization (ACO) [1].

According to obtained statistics our teams show great and brilliant results in IOI, however, in ACM ICPC for students the prize pool still remains not obtained for the past twenty years of the participation of Kazakhstan in this event. On this occasion, we will try to justify what could be the possibility of this outcome and define pros and cons according to the experience of the author of this work as participant of ACM ICPC in 2004-2006.

Kearse et al. see the competition as a way of attracting the participant attitude towards Computer Science [2] – their research and demonstrate that it's one of the best educational practices for giving the interest to the science.

Patterson gives the statistics according to countries which are leading in ACM ICPC [3], as they're China, Russia and United States of America. Basing upon these results we will give short outcome why our finalists from Northern Eurasia Regional Contest (former Northeastern European Regional Contest – NEERC) cannot get the prize place in the final command score.

We give concise and clear comparison of P (polynomial) and NP (non-polynomial) complexity classes on the example of two problems [4, 5, 6].

We also show that solution is feasible within NP-complete problem if the number of elements in input data is very low (< 20) as per bit set definition so that $2^N < 10^9$, which is measured in performed as one second of processor operation cycle.

We will consider following problems to compare P and NP complexities:

- NP-complete: Problem "Box" [7];
- P-complete: Problem "Exploring Pyramids" [8].

The "P versus NP" theorem, which is still not proved, was formulated by Stephen Cook [5] and implies the relation between these two classes of complexity.

The code for both problems can be obtained from repository [9].

P versus NP in Programming Olympiads. The problem "Box" is NP-complete, however, as the

number of elements in input stream is less than 6, it can be solved in almost polynomial time:

```
const int N = 6;

...

bool check (int *P, int bitmask) {
    for (int i = 0; i < N; i++)
        for (int j = 0; j < N; j++)
            if (F[i][j][0] != -1)
                if (A[P[i]][F[i][j][0] ^ ((bitmask >> i) & 1)]
                    != A[P[j]][F[i][j][1] ^ ((bitmask >> j) & 1)])
                    return false;
    return true;
}

void solve() {
    int P[N];

    for (int i = 0; i < N; ++i) {
        P[i] = i;
    }

    do {
        for (int i = 0; i < (1L << N); ++i) {
            if (check(P, i)) {
                puts ("POSSIBLE");

                return;
            }
        }
    } while (next_permutation(P, P + N));

    puts ("IMPOSSIBLE");
}
```

Solution for this problem is P-complete, where $N = 6$ and $N! \ll 10^8$, when the complexity, however is $O(2^N * N! * N^3)$, which is also much lower than average time of running the program in few seconds on modern hardware.

And for the problem "Exploring Pyramids" [8], we use case marks in the global array in order to save the time as per each case, when input is given in single file. We use the dynamic programming approach with memorization.

The complexity of this problem, thus, is also polynomial and is defined in big-O notation as $O(N^3)$.

The code for this problem is as follows:

```
const int N_MAX = 400;

typedef long long ll_t;

int S[N_MAX];
int P[N_MAX][N_MAX];
ll_t R[N_MAX][N_MAX];

...

ll_t F(int l, int r) {
```

```
if (l == r) return 1;

if (P[l][r] == CaseNumber) return R[l][r];

P[l][r] = CaseNumber, R[l][r] = 0;

for (int i = l + 1; i <= r; i++)
    if (S[i] == S[l])
        R[l][r] = (R[l][r] + F(l + 1, i - 1) * F(i, r)) % 1000000000;

return R[l][r];
}
```

Pros and Cons of Olympiad Training

Here we devise the positive and negative sides of the distant training of the participants of the Olympiad. It takes much more to probe through the different sides of the World Wide Web (WWW or W3) community which give the helping hand in solving particular problems from their volume sets. This approach recommended itself to be fast and simple when the user sends his solution through the automated system which judges this solution in one of the programming languages. The distant approach, however, practically shows worse results of preparation of Olympiad participants, rather, than attending special events upheld by sponsoring organizations.

Attending training courses is also not a good alternative to the variants of obtaining experience since the school of programming is formed by the authorities in this field who had a good skill expertise on the official event like, for instance, ACM ICPC or IOI.

Another question which is important to learn is a way of teaching students the Dynamic Programming (DP) [10]. In the statistics it's known that almost every student meets the difficulties of learning DP as well as other structures of the Computer Science in theory, which, in turn, is essential since the methods are to be studied along with practical exercises.

We conclude that the person showing good expertise in specific field of Informatics and Computer Science is able to demonstrate his skills on the event like Olympiad. However, another practice shows that Olympiad participants are trained from the early age – this shows good results, however, is more expensive in terms of time and management. We propose the easy way of getting to the point by solving problems which are published around W3. The stable working place would be also a better alternative for getting practical experience in the one of the programming languages, on which the solution is to be realized and solved, and the practical evaluation experience is obtained by giving the probe of realizing the classical algorithms and other solutions using one of the programming flavors. The most popular of which are those which aim for compiler rather than interpreter as the solving time is the primary goal on the Olympiad stage.

In the next section we'll continue the overview of the programming languages selection criteria.

Selection of Programming Language

The practice shows that most of the solutions are solved using C++ programming language, rather, than Java which is forwarded to be replacement being the second most popular realization programming language. This difference is justified by the fact that C/C++ utilizes convenient way of giving input and output to the problem solution, meanwhile, Java gives the additional task of implementing scanner, tokenizer or parser of the input data.

Together these two languages form the group of programming languages which derive the template library. Standard Template Library (STL) in C++ represents set of programming classes and data structures as well as modules to operate on basic level and utilities like sorting – above you can see the usage of “algorithm” module. The term “iterator” here is vital as STL bases itself

by overriding standard comparison and equivalence operators, thus, giving the order of the elements in a sorted collection like array, set (hash-set) or other user-defined data structure. It's quite enough to mention that it's achieved by overriding only "<" operator as the STL recognizes equivalence by using logical not "!"-operator.

Java programming language is also well suited as it's a virtual machine compiler and, thus, allows to avoid the invocation of programming elements which are using memory pointers which is unsafe in general.

C# programming language could be also a better alternative which supports the template library for data structures, however, it lacks I/O routines as well as many other programming languages.

Conclusion

We have defined the pros and cons of the elementary preparation of the participants for Informatics and Programming Olympiads showing that collaborative work gives better results when the financial sponsoring is limited.

Thus, we have devised that even NP-complete problems can be solved exactly with respect to the computational volume of the state space without using dynamic programming, as when the complexity fits into this volume.

We have also made an important conclusion of the programming language to be selected in order to solve problems more effectively in the limited time during the programming contest.

The pros and cons for development of Programming and Informatics Olympiads in Kazakhstan is also given and more is to be done as Kazakhstan still remains without prize place on the finals, despite the stable advances of our teams to ACM ICPC finals. Meanwhile, IOI participants show better results which is due to the non-limited rating scheme which differs as ACM problems are to be solved on the full set of tests and IOI problems are only scored for each test case.

Acknowledgements

The author expresses the gratitude to the former NEERC community for organizing contests in 2004-2006 as well as local sponsors of this contest like administration of Satbayev University during author's participation at these dates.

Funding

This work was fully supported by an educational grant of the Ministry of Education and Sciences from 2001 to 2006 during author's studying at Satbayev University.

References

- [1] Syzdykov M., Uzbekov M. Ant Colony Optimisation Applied to Non-Slicing Floorplanning. Olympiads in Informatics. 2015. 9.
- [2] Kearsse Iretta B., Charles R. Hardnett. Computer science olympiad: exploring computer science through competition. Proceedings of the 39th SIGCSE technical symposium on Computer science education. 2008.
- [3] Patterson David A. Reflections on a programming olympiad. Communications of the ACM 48.7. 2005. 15-16.
- [4] Fortnow Lance. The status of the P versus NP problem. Communications of the ACM 52.9. 2009. 78-86.
- [5] Cook Stephen. The importance of the P versus NP question. Journal of the ACM (JACM) 50.1. 2003. 27-29.
- [6] Aaronson Scott P. Is P versus NP formally independent? Bulletin of the EATCS 81.2003. 70. 109-136.
- [7] UVA. Problem Box. <https://vjudge.net/problem/UVA-1587> (accessed Jan 25, 2023).
- [8] UVA. Problem Exploring Pyramids. <https://vjudge.net/problem/UVA-1362> (accessed Jan 25, 2023).
- [9] Syzdykov M. ACM Problem Solutions. <https://github.com/mirzakhmets/ACM> (accessed Jan 25, 2023).
- [10] Forišek Michal. Towards a better way to teach dynamic programming. Olympiads in Informatics. 2015. 9. 45-55.

ОБЗОР ПЕРСПЕКТИВ ПРОВЕДЕНИЯ ОЛИМПИАД ПО ПРОГРАММИРОВАНИЮ В КАЗАХСТАНЕ

Мирзахмет Сыздыков

Казахский национальный исследовательский технический университет имени К.И.

Сатпаева Алматы, Казахстан

mspmail598@gmail.com

ORCID ID: <https://orcid.org/0000-0002-8086-775X>

Аннотация. Мы представляем исследование, направленное на текущее состояние олимпиад по программированию, в которых участвует Казахстан и придерживается положительной линии, мы предлагаем различные плюсы и минусы для подготовки участников этих мероприятий. С момента участия Казахстана в Международной олимпиаде по информатике (IOI) для школьников и Международном студенческом конкурсе по программированию (ICPC), который поддерживается Ассоциацией вычислительной техники (ACM), большинство лучших результатов были показаны на IOI, а не на ACM ICPC, где по-прежнему участники из нашей страны демонстрируют стабильную тенденцию в финале, однако, по-прежнему не получают призовой фонд, состоящий из первых двенадцати команд в итоговом рейтинге. Задача, которую предстоит решить участникам, также может быть классифицирована как полиномиальная (P-полная) или неполиномиальная (NP-полная), что, в свою очередь, дает нам представление о том, что можно было бы решить с помощью рационального решения. Мы демонстрируем этот факт на примерах двух задач из ACM ICPC.

Ключевые слова: олимпиада по информатике, тенденция, обучение, статистика.

ҚАЗАҚСТАНДА БАҒДАРЛАМАЛАУ БОЙЫНША ОЛИМПИАДАЛАРДЫ ӨТКІЗУ КЕЛЕШЕГІНЕ ШОЛУ

Мырзахмет Сыздыков

Қ. И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті Алматы, Қазақстан

mspmail598@gmail.com

ORCID ID: <https://orcid.org/0000-0002-8086-775X>

Аңдатпа. Біз Қазақстан қатысатын және оң бағытты ұстанатын бағдарламалау олимпиадаларының ағымдағы жай-күйіне бағытталған зерттеуді ұсынамыз. Осы іс-шараларға қатысушыларды дайындау үшін әртүрлі оң және теріс жақтарын ұсынамыз. Қазақстанның мектеп оқушыларына арналған Информатика бойынша халықаралық олимпиадаға (IOI) және есептеу

техникасы қауымдастығы (АСМ) қолдайтын бағдарламалау бойынша халықаралық студенттік конкурсқа (ICPC) қатысқан сәттен бастап, ең жақсы нәтижелердің көпшілігі АСМ ICPC-де емес, IOI-де көрсетілді, онда біздің еліміздің қатысушылары финалда тұрақты тенденцияны көрсетуде алайда, қорытынды рейтингтегі алғашқы он екі командадан тұратын жүлде қоры әлі де бұйырмай тұр. Қатысушылар шешетін тапсырманы көпмүшелік (P-толық) немесе көпмүшелік емес (NP-толық) деп жіктеуге болады, бұл өз кезегінде ұтымды шешім арқылы не шешуге болатыны туралы түсінік береді. Біз бұл фактіні АСМ ICPC-тен екі тапсырманың мысалы негізінде көрсетеміз.

Кілттік сөздер: информатика олимпиадасы, тренд, оқыту, статистика.

Сведения об авторе:

Анг.: Syzdykov Mirzakhmet - Satbayev University, Almaty, Kazakhstan

Каз.: Сыздықов Мырзахмет- Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан.

Рус.: Сыздыков Мырзахмет- Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан.

ӘҚЖ 004.89
МРНТИ 28.23.01

ИНТЕЛЛЕКТУАЛДЫҚ-АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР САЛАСЫНДАҒЫ ИННОВАЦИЯЛЫҚ ӘЗІРЛЕМЕЛЕР ЖӘНЕ ОЛАРДЫ БАҒДАРЛАМАЛЫҚ ІСКЕ АСЫРУ НЕГІЗІНДЕ БИЗНЕС ПРОЦЕСТЕРДІ БОЛЖАУ ӘДІСТЕРІНДЕГІ ЖАҢА ЕНГІЗІЛІМДЕР

А.М. Ахметова, А.С. Шаяхметова, Д.Ө. Жақан

Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан

E-mail: Dastan_zhakan@mail.ru

¹ORCID ID: <https://orcid.org/0000-0001-8360-5504>

²ORCID ID: <https://orcid.org/0000-0002-4072-3671>

Аңдатпа. Бұл мақалада интеллектуалдық-ақпараттық технологиялар саласындағы инновациялық әзірлемелер мен оларды бағдарламалық іске асыру негізінде бизнес процестерді болжау әдістері сөз болады. Сондай-ақ, оларды болжау әдістеріндегі жаңа енгізілімдер, нейрондық желілер жайы ашып қарастырылады. Кәсіпорындағы экономикалық талдау міндеттерін шешу кез-келген ұйымның дамуының әр кезеңінде бәсекеге қабілеттілігін арттыруды қамтамасыз ететін фактор болып табылады. Бұл мақалада кәсіпорын қызметін болжауда деректерді зияткерлік талдау әдістерін қолданудың негізгі ерекшеліктері ашылады, сонымен қатар оларды шешуде қолданылатын типтік есептер мен алгоритмдердің тізімі келтірілген.

Кілттік сөздер: бизнес процестер, интеллектуалдық-ақпараттық технология, нейрондық желілер, болжау, OLAP, Data Mining.

Кіріспе

Қазіргі экономика жағдайында кәсіпорындағы кез келген қызмет міндетті түрде барлық шаруашылық операциялардың, бизнес процестердің және басқа да қажетті ақпараттың ұсақ бөлшектерін құжаттаумен, тіркеумен және жазумен, басқаша айтқанда, ұйымдағы барлық ақпараттық ресурстарды тіркеумен бірге жүруі керек.

Оның ішінде шешім қабылдауды қолдау жүйесі жұмысын қамтамасыз ету, бизнес процестер жұмысындағы болашақ өзгерістерді жоспарлау және экономикалық қызмет көрсеткіштері үшін болжам жасау мақсатында ақпаратты терең талдауға жарамды түрге тиімді жинау, сақтау және келтіру мәселесі туындайды.

Процестерді болжау қажеттілігі кез-келген кәсіпорынның, аймақтың, саланың немесе бөлімнің стратегиясын жасау кезінде туындайды. Сыртқы ортаның тез, айқын емес және күтпеген өзгерістеріне байланысты болжау мәселесі соңғы кездері ерекше күрделі болып жатыр. Болжау мамандары арасында экономика мен саясаттың көптеген факторларының болжамсыздығы мен сенімділігін ескере отырып, жалпы қабылданған математикалық әдістермен ғылыми негізделген болжамдар жасау мүмкін еместігі туралы мәлімдемелер пайда болды. Осы жерден болжаудың қолданыстағы әдістерін жетілдіру және жаңа әдістерін іздеу өзекті болып табылады.

Болжаудың жаңа әдістерін әзірлеудегі перспективалы бағыттардың бірі – нейрондық желілер әдісі. Нейрондық желілер өзін-өзі оқыту қабілеті және әртүрлі деректер арасында байланыстар мен үлгілерді орнату арқылы болжау және модельдеу мәселелерін шешуде тиімді болуы мүмкін.

Әдебиеттерге шолу

Жұмысымызда қарастырылып отырыған нейрондық желі технологиясы Р.Б. Сергиенконың [1] бұлыңғыр классификатор және В.В. Бухтоярованың [2] нейрондық

желілердің өзін-өзі реттейтін ұжымының әдісі сынды әзірлемелерін біріктіреді. Айтып өткен, Р.Б. Сергиенконың бұлыңғыр жіктеуішінің (классификатор) артықшылығы мынада, ол бір уақытта жіктеу алгоритмі және лингвистикалық түрде көрсетілген білімді алу алгоритмі ретінде әрекет етеді.

Нейрондық желілер болжау және модельдеу мәселелерін шешуде тиімді. Нейрондық желілер ұжымдарын қалыптастырудың жаңа әдісі (Бухтояров әдісі) бір нейрондық желінің тиімділігінен және нейрондық желілер ұжымдарын қалыптастырудың қолданыстағы әдістерінен (GASEN, Gabased 1, gabased 2) жоғары тиімділігімен ерекшеленеді.

Нейрондық желілер ұжымының жалпы шешімі жеке нейрондық желілер қабылдаған шешімдер негізінде жасалады, яғни жалпы шешім жеке нейрондық желілердің шешімдеріне тәуелді кейбір функция болып табылады. Жалпы шешімді қалыптастырудың эволюциялық тәсіліне әдістер кіреді: ықтималдық эволюциялық әдіс, генетикалық бағдарламалаудың гибридті алгоритмі [2]. Коэволюциялық алгоритм әр түрлі стандартты эволюциялық алгоритмдердің бәсекелестігі мен кооперациясы арқылы мәселені шешеді. Сонымен қатар, ең жақсы параметрлерді таңдау мәселесі шешіледі, коэволюциялық алгоритм шартсыз оңтайландыру есептеріндегі орташа параметрлері бар стандартты алгоритмге қарағанда тиімдірек және шартты оңтайландыру есептеріндегі ең жақсы параметрлері бар стандартты эволюциялық алгоритмге қарағанда тиімдірек.

Осы мамандардың әзірлемелерін біріктіре отырып, оларды бағдарламалық модульге енгізуге әрекет жасалынады. Әдістердің өзі, әзірленіп жатқан бағдарлама сияқты, әртүрлі салаларда қолданылуы мүмкін, бірақ болжау қажеттілігінің артуына байланысты дәл осы бағыт басымдыққа ие болды.

Әдістеме

Болжау кез-келген кәсіпорынның ең сұранысқа ие міндеттерінің бірі болып табылады. Бұл тауарларды сатуды болжау, сұраныстың өзгеруі, баға саясатының өзгеруі және т.б. айтып кете аламыз. Болашақта оқиғалардың дамуын ескере отырып, кәсіпорынның қызметін неғұрлым икемді түрде жоспарлауға, неғұрлым негізделген басқару шешімдерін қабылдауға болады. Көптеген кәсіпорындар өз қызметінің нәтижесінде дерекқорлар мен деректер қоймаларында сақталатын ақпараттың ауыр көлемін жинақтайды. Осы мәліметтер негізінде болжамды уақтылы жүзеге асыруға және негізделген шешімдер қабылдауға болады. Болжау мәселелерін шешу үшін статистикалық әдістерді де, нейрондық желілерді де қолдануға болады. Ол үшін уақыт қатарларының негізінде әр түрлі жеке әдістер мен болжау модельдерінің көп екендігі белгілі. Бұл әдістер нәтижелердің дәлдігі мен күрделілік дәрежесі бойынша бір-бірінен ерекшеленеді.

Нейрондық желілер болжау әдістерінің ішінде ерекше орын алады. Көптеген статистикалық әдістерден айырмашылығы, нейрондық желілер тәуелсіз айнымалылардың шексіз санын қолдануға мүмкіндігі бар. Бір рет құрылған құрылымды жаңа тәуелсіз міндеттерді шешу үшін қолдануға болады. Бірақ нейрондық желілерде болжау бірқатар кемшіліктерге ие. Әдетте, қолайлы модель жасау үшін көптеген бақылаулар жасап отыру қажет. Өйткені онда оқытудың қолайлы коэффициенттерін таңдауда қиындықтар туындап жатады.

Нәтижелер мен талқылау

Болжам жасау мақсатында ақпаратты терең талдауға жарамды түрге тиімді жинау, сақтау және келтіру мәселесі туындайтындығы белгілі. Осылайша, мәліметтер базасы кез-келген заманауи ұйымның өмірлік белсенділігінің ажырамас атрибутына айналады. Кәсіпорын дерекқоры әдетте белгілі бір нысанды немесе фактіні сипаттайтын жазбаларды сақтайтын кестелердің үлкен жиынтығынан тұрады (мысалы, тұтынушылар, жеткізушілер, тапсырыстар, сату көлемі туралы жазбалар). Жедел, сондай-ақ өзгермейтін

деректердің бүкіл ағыны неғұрлым құнды ақпаратты, атап айтқанда белгілі бір деректер арасындағы тенденциялар, заңдылықтар мен өзара тәуелділіктер туралы ақпаратты алу көзі бола алады.

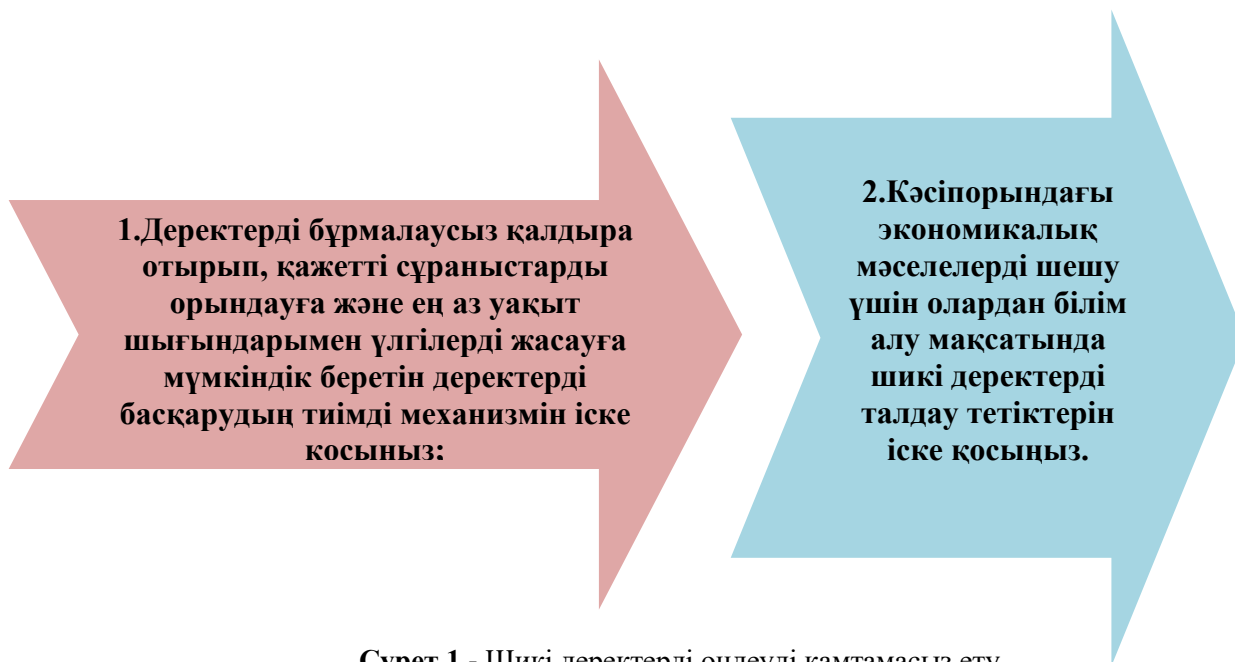
Мұндай ақпарат кейіннен кәсіпорындағы аналитикалық қызмет үшін қолданылады. Мысалы, сатып алу себетін талдау, тәуекелдер, стратегиялық және операциялық жоспарлау, белгілі бір экономикалық объектілер мен субъектілердің динамикалық тенденцияларын болжау немесе оқиғаны ретроспективті мәліметтер негізінде модельдеу [3].

Басқаша айтқанда, шикі деректерді терең талдауға жарамды түрге түрлендіру жүзеге асырылады. Шикі деректерді өңдеуді қамтамасыз ету үшін сізге қажет 2 нәрсе бар (кесте 1).

Жоғарыда айтылғандарға сәйкес, Кәсіпорында деректерді зияткерлік талдау құралдарын қолдайтын ақпараттық аналитикалық жүйе енгізілуде.

Нарықтың қазіргі жағдайында мұндай аналитикалық мәселелерді шешу үшін Data mining және OLAP технологиялары жиі қолданылады (ағылш. online analytical processing, нақты уақыттағы аналитикалық өңдеу) [4, 123].

Эдгар Коддтың анықтамасына сәйкес, OLAP деректерді өңдеу технологиясының негізгі міндеті көп өлшемді принцип бойынша құрылымдалған деректердің үлкен массивтерін талдау негізінде жинақталған ақпаратты дайындау болып табылады. OLAP технологиялары бизнес-аналитика (Business Intelligence, BI) бағдарламалық өнімдерінің құрамдас бөлігі болып табылады [5].



Сурет 1 - Шикі деректерді өңдеуді қамтамасыз ету

Нейронның математикалық моделі оның сипаттамаларын өзгертуге, толықтыруға және реттеуге мүмкіндік беретін әмбебап сызықтық емес элемент болып табылады. Мұндай желілердің басты ерекшелігі – олар нақты мысалдармен өзін-өзі оқытуға қабілетті. Нейрондық желілерді заңдылықтар жасырылған кіріс құжаттарының үлкен көлемі бар тапсырмаларда қолданған жөн.

Нейрондық желілерді қолдану көптеген қолданбалы экономикалық мәселелерді шешуде тиімді. Қаржы саласында нейрондық желілер клиенттердің мінез-құлқын болжау, алдағы мәміле тәуекелдерін бағалау, несие берушілердің ықтимал алаяқтық әрекеттерін болжау үшін қолданылады. Экономикалық қызметті жоспарлау кезінде нейрондық желілер сату көлемін анықтауға, өндірістік қуаттарды жүктеуге, жаңа өнімге сұранысты болжауға қатысуы мүмкін. Бизнес-аналитика және СІПР жүйелерінің жұмысында

нейрондық желілер бәсекелес фирмаларды салыстырмалы талдауға, филиалдардың жұмысын талдауға қатысады, даму тенденцияларын, деректердің үлкен көлеміндегі корреляцияларды анықтауда қолданылады [6, 455].

Қолданбалы есептерді шешкен кезде ең тиімді кірістерді (яғни айнымалыларды) анықтау үшін бір алгоритмді қолдану ұсынылады, содан кейін осы мәліметтер негізінде белгілі бір нәтижені болжау үшін басқа алгоритм қолданылады. Деректерді өндіру бір интеллектуалды құрылымға сүйене отырып, көптеген модельдерді құруға мүмкіндік береді, осылайша бір деректерді өндіру шешімінің бөлігі ретінде кластерлеу алгоритмін, шешім ағашының моделін, сондай-ақ деректердің әртүрлі көріністерін алу үшін жеңілдетілген Байес алгоритмінің моделін пайдалануға болады. Бір шешімде жеке тапсырмаларды орындау үшін бірнеше алгоритмдерді қолдануға болады. Мысалы, регрессия арқылы қаржылық болжамдарды алуға болады, ал нейрондық желі алгоритмі арқылы сату көлеміне әсер ететін факторларды талдауға болады.

Белгілі бір есепті шешу үшін алгоритмдерді таңдауды жеңілдету үшін келесі кестеде әр алгоритм әдетте қолданылатын есептердің түрлері келтірілген (кесте 2).

Кесте 2. Есептердің түрлері

Тапсырмалардың мысалдары	Сәйкес алгоритмдер
<p>Дискретті атрибутты болжау</p> <ul style="list-style-type: none"> - Клиенттерді әлеуетті сатып алушылар тізімінен жақсы және жаман үміткерлер ретінде белгілеу. - Келесі алты ай ішінде сервердің істен шығу ықтималдығын есептеу. - Пациенттердің ауруларын дамыту нұсқаларының жіктелуі және байланысты факторларды зерттеу. 	<ul style="list-style-type: none"> - Шешім қабылдау ағашының (жіктеу ағашы немесе регрессия ағашы) алгоритмі - Жеңілдетілген Байес алгоритмі - Кластерлеу алгоритмі - Нейрондық желі алгоритмі
<p>Үздіксіз атрибутты болжау</p> <ul style="list-style-type: none"> - Келесі жылға сату болжамы. Өткен жылдар мен маусымдық тенденцияларды ескере отырып, сайтқа кірушілер санының болжамы. - Демографияны ескере отырып, тәуекелді бағалауды қалыптастыру. 	<ul style="list-style-type: none"> - Шешім қабылдау ағашының алгоритмі - Уақыт қатарларының алгоритмі - Сызықтық регрессия алгоритмі
<p>Реттілікті болжау</p> <ul style="list-style-type: none"> - Компанияның веб-сайты бойынша қозғалыс бағытын талдау. - Сервердің істен шығуына әкелетін факторларды талдау. Жалпы іс-әрекеттер бойынша ұсыныстарды тұжырымдау мақсатында емханаға бару кезіндегі іс-қимылдар реттілігін қадағалау және талдау 	<ul style="list-style-type: none"> - Тізбекті кластерлеу алгоритмі

Белгілі бір қолданбалы тапсырманы шешетін бір жүйеде бірнеше интеллектуалды және ақпараттық технологияларды бөлісудің екі әдісін ажыратуға болады. Бірінші әдіс – тапсырманың жалпы шешімін жасау үшін бірнеше интеллектуалды технологияларды бөлісу, ол жасанды нейрондық желілер ұжымы осындай бірлестіктің мысалы бола алады. Мұндай бірлестіктерде мәселені жалпы шешудің жолын таңдау туралы мәселе маңызды болып табылады.

Бір жүйеде ақпараттық технологияны біріктірудің екінші тәсілі – басқаларын автоматты түрде жобалау үшін кейбір технологияларды қолдану [2]. Мұндай бірлестікке деген қажеттілік интеллектуалдық-ақпараттық технология дизайнының тиімділігін

арттыру қажеттілігімен де, ИАТ негізіндегі жүйелерді жобалау процесін автоматтандыруға деген ұмтылыспен де негізделуі мүмкін. Соңғы дәлел, әсіресе, техникалық қызметтің көптеген салаларының қазіргі жағдайына тән уақытша және адами ресурстарға шектеулердің қатаңдығы мен ақпараттық технологияны қолданатын жүйелердің өсіп келе жатқан ауқымына байланысты өте маңызды болып көрінеді. Оларды біріктірудің бір мысалы – нейрондық желілерді жобалау және оқыту үшін эволюциялық алгоритмдерді қолдану. Интеллектуалдық-ақпараттық технологиямен бөлісудің екі әдісін бір жүйеде біріктіру технологияларды генерациялау процестерін күшейту және автоматтандыру арқылы оны жобалау үшін қажетті қымбат ресурстарға қойылатын талаптарды азайту кезінде осындай жүйенің көмегімен алынған шешімдердің сапасын едәуір арттыруға мүмкіндік береді. Сайып келгенде, осындай бірлестіктерінің деректерін зияткерлік талдау үшін пайдалану адам қызметінің көптеген салаларында қолданбалы мәселелерді шешудің тиімділігін едәуір арттыра алады. Интеграцияланған ядроны бағдарламалық іске асырудың нәтижесі бағдарламалық компонент (компьютер бағдарламасы) болып табылады. Оған қойылатын талаптар, оның ішінде тұтастай алғанда аналитикалық модульге қойылатын талаптар анықталады.

Функционалдық талаптардың 3 түрін көрсете аламыз:

1. Аналитикалық модульдің интеграцияланған ядросының бағдарламалық жасақтамасы деректерді өңдеуге және деректерді талдаудың келесі типтік мәселелерін шешудің нәтижелерін ұсынуға мүмкіндік беруі керек:

а) модельдеу. Математикалық модельдерді құруды, тиімділікті талдауды және қолдануды қамтиды.

ә) болжау. Болжамды модельдерді құруды, тиімділікті талдауды және қолдануды қамтиды.

б) жіктеу. Ол жүктелген деректермен сипатталған объектілердің белгілі бір сыныптарға жататындығын анықтауды және осы процедураны жүзеге асырудың тиімділігін бағалауды қамтиды.

2. Бағдарлама пайдаланушы интерфейсі арқылы да, оны алу арқылы да талдау үшін басқару параметрлерін анықтау және деректерді жүктеу мүмкіндігін қамтамасыз етуі керек конфигурация файлдары мен деректер файлдарын түсіндіру.

3. Бағдарлама міндеттерді шешу үшін ақпараттық технологияны пайдалану тиімділігіне статистикалық зерттеулер жүргізу мүмкіндігін қамтамасыз етуі тиіс.

4. Бағдарлама жұмыс нәтижелерін (құрастырылған модельдер, болжамдар, алгоритмдік ядроның сынақ нәтижелері) пайдаланушы интерфейсіне де, берілген құрылымы бар файлдарға да түсіруді қамтамасыз етуі керек.

Қорытынды

Қорытындылай келе, біздің жұмысымызда интеллектуалдық-ақпараттық технологиялар саласындағы инновациялық әзірлемелер және оларды бағдарламалық іске асыру негізінде бизнес процестерді болжау әдістеріндегі жаңа енгізілімдер жайы сөз болды. Нәтижесінде, бизнес процестерді болжауда деректерді интеллектуалды талдау әдісін қолданудың, соның ішінде нейрондық желілерді қолданудың артықшылықтары аталып өтті.

Осылайша, кәсіпорын қызметін талдаудағы болжау мәселелерін шешу бұрын анықталмаған тәуелділіктерді анықтауға, сатып алу сұранысының даму векторларын бағалауға, клиенттердің мінез-құлқын талдауға мүмкіндік береді. Бұл ретте аналитикалық платформаларды пайдалану қабылданатын басқару шешімдерінің тиімділігін арттыруға мүмкіндік береді, шығындарды, сату көлемін оңтайландыруға, нарықтық жағдайды бағалауға ықпал етеді, бұл барынша пайда алуға мүмкіндік береді. Сонымен қатар, алынған және жүйеленген деректерді бақылау жүйесін ұйымдастыру үшін пайдалануға болады, бұл тұтастай алғанда кәсіпорынның тиімділігін арттыруға көмектеседі.

Әдебиеттер тізімі:

[1] Сергиенко Р. Б. Метод формирования нечеткого классификатора самонастраивающимися коэволюционными алгоритмами. Искусственный интеллект и принятие решений. 2010. 3. 98-106.

[2] Бухтояров В. В. Разработка комплексного эволюционного подхода для проектирования коллективов нейронных сетей для задач моделирования. Труды Международных научно-практических конференций AIS'10/CAD- 2010. С. 271-279.

[3] Сопов Е.А. СППР на базе интеллектуальных информационных технологий Data Mining для задач бюджетного управления. Труды Международных конференций «Вычислительные и информационные технологии в науке, технике и образовании».

[4] Курейчик В.В., Сороколетов П.В., Щеглов С.Н. Анализ современного состояния автоматизированных систем приобретения и представления знаний. Известия ЮФУ. Технические науки. 2008. 9 (86). 120-125.

[5] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. 2004. 336.

[6] Гладков Л.А., Гладкова Н.В. Проблемы и перспективы развития технологий интеллектуального анализа и извлечения данных. Конгресс по интеллектуальным системам и информационным технологиям «AIS-IT'09». Труды конгресса. 2009. 1. 454-458.

References

[1] Sergiyenko R. B. Metod formirovaniya nechetkogo klassifikatora samonastravayushchimisiya koevolutsionnymi algoritmami. Iskusstvennyy intellekt i prinyatiye resheniy. 2010. 3. 98-106.

[2] Bukhtoyarov V. V. Razrabotka kompleksnogo evolyutsionnogo podkhoda dlya proyektirovaniya kollektivov neyronnykh setey dlya zadach modelirovaniya. Trudy Mezhdunarodnykh nauchno-prakticheskikh konferentsiy AIS'10/CAD- 2010. S. 271-279.

[3] Sopov Ye.A. SPPR na baze intellektual'nykh informatsionnykh tekhnologiy Data Mining dlya zadach byudzhethnogo upravleniya. Trudy Mezhdunarodnykh konferentsiy «Vychislitel'nyye i informatsionnyye tekhnologii v nauke, tekhnike i obrazovanii».

[4] Kureychik V.V., Sorokoletov P.V., Shcheglov S.N. Analiz sovremennogo sostoyaniya avtomatizirovannykh sistem priobreteniya i predstavleniya znaniy. Izvestiya YUFU. Tekhnicheskkiye nauki. 2008. 9 (86). 120-125.

[5] Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. Metody i modeli analiza dannykh: OLAP i Data Mining. 2004. 336.

[6] Gladkov L.A., Gladkova N.V. Problemy i perspektivy razvitiya tekhnologiy intellektual'nogo analiza i izvlecheniya dannykh. Kongress po intellektual'nym sistemam i informatsionnym tekhnologiyam «AIS-IT'09». Trudy kongressa. 2009. 1. 454-458.

INNOVATIVE DEVELOPMENTS IN THE FIELD OF INTELLECTUAL AND INFORMATION TECHNOLOGIES AND INNOVATIONS IN METHODS OF FORECASTING BUSINESS PROCESSES BASED ON THEIR SOFTWARE IMPLEMENTATION

A.M. Akhmetova, A. S. Shayakhmetova, D. O. Zhakan
Al-Farabi Kazakh National University, Almaty, Kazakhstan

Abstract. This article will focus on innovative developments in the field of intellectual and information technologies and methods of forecasting business processes based on their software implementation. Innovations in their forecasting methods and neural networks are also revealed. Solving the problems of economic analysis at the enterprise is a factor that ensures the competitiveness of any organization at every stage of its development. This article reveals the main features of the application of data mining methods in forecasting the activities of an enterprise, and also provides a list of typical tasks and algorithms used in their solution.

Keywords: business processes, intellectual and information technologies, non-neural networks, forecasting, OLAP, Data Mining.

«Интеллектуалдық-ақпараттық технологиялар саласындағы инновациялық әзірлемелер және оларды бағдарламалық іске асыру негізінде бизнес процестерді болжау әдістеріндегі жаңа енгізілімдер
А.М. Ахметова, А.С. Шаяхметова, Д.Ө. Жакан

ИННОВАЦИОННЫЕ РАЗРАБОТКИ В ОБЛАСТИ ИНТЕЛЛЕКТУАЛЬНО-ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И НОВОВВЕДЕНИЯ В МЕТОДАХ ПРОГНОЗИРОВАНИЯ БИЗНЕС-ПРОЦЕССОВ НА ОСНОВЕ ИХ ПРОГРАММНОЙ РЕАЛИЗАЦИИ

А. М. Ахметова, А. С. Шаяхметова, Д. О. Жакан
Казахский национальный университет им. Аль-Фараби, Алматы, Казахстан

Аннотация. В данной статье речь пойдет об инновационных разработках в области интеллектуально-информационных технологий и методах прогнозирования бизнес-процессов на основе их программной реализации. Также раскрываются нововведения в методах их прогнозирования, нейронных сетях. Решение задач экономического анализа на предприятии является фактором, обеспечивающим повышение конкурентоспособности любой организации на каждом этапе ее развития. В данной статье раскрываются основные особенности применения методов интеллектуального анализа данных в прогнозировании деятельности предприятия, а также приводится перечень типовых задач и алгоритмов, используемых при их решении.

Ключевые слова: бизнес-процессы, интеллектуально-информационные технологии, нейронные сети, прогнозирование, OLAP, Data Mining.

Авторлар жайында мәлімет:

Қаз: Ахметова Ардақ Мергенбайқызы – Әл-Фараби атындағы Қазақ ұлттық университетінің доценті, PhD ardak_66@mail.ru

Рус: Ахметова Ардак Мергенбаевна – PhD, доцент Казахского национального университета им. аль-Фараби, ardak_66@mail.ru

Англ: Akhmetova Ardak Mergenbaevna – PhD, Associate Professor of Al-Farabi Kazakh National University, ardak_66@mail.ru

Қаз: Шаяхметова Асем Серикбаевна – Әл-Фараби атындағы Қазақ ұлттық университетінің доценті, PhD asemshayakhmetova@mail.ru

Рус: Шаяхметова Асем Серикбаевна – PhD, доцент Казахского национального университета им. аль-Фараби, asemshayakhmetova@mail.ru

Англ: Shayakhmetova Asem Serikbayevna – PhD, Associate Professor of Al-Farabi Kazakh National University, asemshayakhmetova@mail.ru

Қаз: Жақан Дастан Укенұлы – Әл-Фараби атындағы Қазақ ұлттық университетінің магистранты

Рус: Жакан Дастан Укенович-магистрант Казахского национального университета имени Аль-Фараби

Англ: Zhakan Dastan Ukenovich-Master's student of Al-Farabi Kazakh National University

УДК 004.62;65
МРНТИ 20.23.17, 20.23.21, 20.53.19

ТЕХНОЛОГИИ ОРГАНИЗАЦИИ ХРАНЕНИЯ И ЗАПРОСА БОЛЬШИХ КОЛЛЕКЦИЙ XML-ДОКУМЕНТОВ

А.А. Мухитова^{1,2}, А.С. Еримбетова^{2,3}, В. Баракнин⁴

¹Казахский национальный университет имени аль-Фараби, Алматы, Казахстан

²Институт информационных и вычислительных технологий МОН РК, Алматы, Казахстан

³Satbayev University, Алматы, Казахстан

⁴Федеральный Исследовательский Центр Информационных и Вычислительных технологий, РАН, 630090, Новосибирск, Россия

¹ORCID ID: <https://orcid.org/0000-0002-4081-7694>, mukhitova.aigul@gmail.com

²ORCID ID: <https://orcid.org/0000-0002-2013-1513>, aigerian@mail.ru

⁴ORCID ID: <https://orcid.org/0000-0003-3299-0507>

Аннотация. В настоящее время веб-приложения, использующие XML репозитории (банковское дело, бухгалтерский учет, управление персоналом, резервирование авиакомпаний, мониторинг и прогнозирование погоды, электронное управление и электронная коммерция) обрабатывают огромное количество информации временного характера и нуждаются в полной истории изменения данных и схем для эффективного и прозрачного управления информационной системой. Существующие XML СУБД и XML-инструменты не имеют поддержки данной функции, в связи с чем возникает сложность построения истории изменений, сложность возврата к предыдущему состоянию, возможность потери данных. В следствие чего возникает проблема эффективного управления хранением временных данных при работе с большим количеством небольших XML-документов и множества дополнительных элементов/атрибутов.

Ключевые слова: схема кодирования XML, отображение XML-данных, база данных NoSQL, обработка и оптимизация XML-запросов.

Введение

XML был разработан в качестве стандартного языка и формата де-факто в промышленности для хранения и обмена данными между разнородными системами. Иерархическая природа XML поставила множество исследовательских задач перед сообществом баз данных [1]. Одна фундаментальная проблема заключается в эффективном хранении и запросе таких данных с древовидной структурой.

Чтобы запросить большую коллекцию XML-документов, аналитики данных обычно создают специальные программы для поиска во всех документах определенных тегов или местоположений, что обычно приводит к значительному количеству файловых операций ввода-вывода, что не показывает своей эффективности. Преобразование XML-документов в отношения и, следовательно, преобразование XML-запросов в SQL-запросы по этим отношениям является общепризнанным механизмом, но все еще существуют несоответствия между XML-структурированными данными и данными отношений [2-3]. Кроме того, такие методы не могут обеспечить удовлетворительную производительность и обычно не масштабируются по мере роста количества документов. Другим направлением усовершенствования является адаптация собственных баз данных XML, таких как BaseX, Sedna и XDB. Хотя эти базы данных имеют значительные преимущества с точки зрения удобства использования, но они не используют сложные возможности хранения и обработки запросов.

Системы запросов к коллекциям XML-документов, рассредоточенных по P2P-сетям и децентрализованные системы интеграции данных для XML (например, Piazza [4]) могут повысить производительность обработки XML-запросов за счет метода фрагментации. К сожалению, эти методы все же имеют некоторые недостатки. Во-первых, трудно осмысленно распределить XML-данные по узлам данных, поскольку распределенные файловые системы обычно не поддерживают формат XML. Во-вторых, поскольку

обработка XML-данных не является простой задачей, разработчикам приходится либо использовать сторонние библиотеки или инструменты, такие как Avro и Mahout, либо писать свои собственные интерфейсы. В-третьих, извлечение XML-данных из этих систем способом, подобным XQuery, невозможно без дополнительного уровня управления данными поверх распределенной файловой системы.

Существует множество языков и систем обработки данных, разработанных для масштабируемых архитектур, таких как MapReduce и Dryad для облачных вычислений. В частности, Pig, Hive, Jaql и Score можно использовать для хранения XML-документов и запросов к ним. Однако, как показал результат проведенного нами эксперимента, многие системы не оптимизированы, в результате чего возникают проблемы производительности, включая высокие требования к хранилищу, длительное время загрузки и длительное время выполнения запросов.

Способы хранения и запроса XML-документов

XML в реляционных базах данных

Предпринимаются значительные усилия по хранению и запросу XML-документов с использованием реляционных баз данных. Ключевой проблемой является сопоставление вложенных элементов с плоскими таблицами [5]. В целом существующие методы можно классифицировать как отображение нормализации и кодирование узлов.

Преыдушие результаты показали, что отображение нормализации способно обеспечить хорошую эффективность запросов. XML-схема и нормализация данных являются двумя краеугольными камнями для обеспечения такой производительности. С другой стороны, отображение с кодировкой узлов обеспечивает более общее решение - все элементы и атрибуты XML представлены числовыми кодами местоположений узлов в XML-дереве, например Dewey Order [6], Ordpath [7]. Иерархические отношения и порядок документа неявно фиксируются структурным соединением кодировок элементов. При сопоставлении с кодировкой узлов таблицы данных всегда плотные, а элементы из одного XML-документа хранятся последовательно. Из-за отсутствия информации о схеме обработки запросов с сопоставлениями кодирования узлов, как правило, такие разработки очень затратны.

Во временной настройке данные XML могут развиваться по времени транзакции и/или допустимого времени. Таким образом, они могут иметь время транзакции, достоверное время или временный формат. Когда данные XML различных временных форматов могут сосуществовать в одном и том же хранилище XML, именуемый многопользовательским репозиторием XML.

Собственные базы данных XML

Системы баз данных Native XML (NXD) создаются с нуля для хранения XML-документов и запросов к ним: Berkeley DB XML, Apache Xindice, eXist, dbXMKL, Sedna, BaseX, xDB и OrientX. Эти системы изначально хранят XML-данные с использованием трех основных технологий хранения, а именно методов на основе текста, моделей и схем [8].

<i>База данных</i>	<i>Используемая технология хранения XML-документов</i>
Berkeley DB XML	база данных типа «ключ-значение», хранит XML-документы в их родном формате, основанном на текстовом подходе к хранению
Xindice, eXist, dbXML	хранит XML-документы в коллекциях как логические единицы хранения, используя подход к хранению на основе моделей
Sedna	использует блоки/наборы данных в качестве логических единиц хранения и хранит XML-документы, используя подход кластерного хранения на основе схемы

BaseX	использует табличное представление древовидных структур XML для хранения XML-документов
EMC Documentum xDB	хранит XML-документы в интегрированной объектно-ориентированной базе данных
OrientX	упорядочивает XML-документы в соответствии с их схемой и сохраняет их в наборах данных как логические единицы хранения на основе кластерного подхода к хранению с использованием этой схемы

Собственные базы данных XML имеют лучшую производительность по сравнению с подходами «XML на реляционных базах данных». Однако эти системы имеют два потенциальных ограничения. Во-первых, размер как данных, так и индекса, потребляемых собственной базой данных XML, намного больше, чем в собственных базах данных XML [9]. Во-вторых, собственные системы баз данных XML не используют сложные возможности хранения и запросов, уже предоставляемые существующими системами баз данных [10].

XML в распределенных системах хранения

Развитие однорангового (P2P) обмена информацией проложило путь к поддержке приложений управления XML-данными в распределенной среде. В некоторых более ранних работах изучалось совместное использование данных P2P в неструктурированных сетях, включая Piazza [11] и PeerDB [12]. Обе они являются одноранговыми системами управления данными, которые позволяют совместно использовать разнородные данные распределенным и масштабируемым образом. Основная проблема этих систем заключается в том, что запросы могут передаваться всем одноранговым узлам, что значительно увеличивает сетевой трафик.

Разработанная платформа KadoP опирается на известную технологию распределенных хэш-таблиц (DHT) для поддержки сложных запросов к общим XML-данным [13]. Авторы [14] разработали методы для оценки XPath-запросов к XML-дереву, которое разделено по горизонтали и вертикали и распределено по нескольким сайтам. Однако этот подход имеет дело только с логическими запросами XPath и запросами XPath с выбором данных.

XML в облачных системах хранения

Многие системы, языки и модели данных были разработаны для обработки массивных наборов данных поверх распределенных файловых систем (например, HDFS) и в рамках MapReduce. Типичные примеры включают Jaql, Pig, Hive, Score, HadoopDB, HadoopXML и ChuQL. Pig — это динамически типизированный язык запросов с гибкой вложенной реляционной моделью данных и удобным синтаксисом, подобным скрипту, для разработки потоков данных, которые оцениваются с помощью MapReduce. Hive использует гибкую модель данных и MapReduce с синтаксисом, основанным на SQL. Score от Microsoft имеет те же функции, что и Hive, за исключением того, что он использует Dryad в качестве параллельной среды выполнения. HadoopDB является расширением Hive и запрашивается с помощью языка запросов, подобного SQL. Вместо хранения данных в распределенной файловой системе Hadoop он управляет несколькими экземплярами Postgres для хранения данных. Он теряет универсальность Hadoop, который поддерживает произвольные модели данных, но улучшает уровень хранения Hadoop, приближая части запроса к данным. HadoopXML — это система, которая одновременно обрабатывает множество запросов шаблонов веток для большого объема XML-данных с помощью Hadoop. ChuQL — это расширение MapReduce для XQuery для обработки XML с помощью Hadoop.

Были разработаны некоторые новые платформы, предлагающие более общие операторы, включая Hyracks и Nephelē. Hyracks предоставляет разделенную параллельную модель для выполнения вычислений с интенсивным использованием данных в кластерах без совместного использования. Nephelē — это универсальная система для обработки данных в веб-масштабе. Абстракцией программирования для написания

задач являются контракты параллелизации (Parallelization Contracts - PACTs), состоящие из контрактов ввода/вывода. Камачо-Родригес и др. [15] представили подход PAXQuery для неявного распараллеливания XQuery посредством преобразования алгебраического плана XQuery в параллельный план PACT.

Другое направление — отображение XML-документов в различные базы данных NoSQL. Все хранилища NoSQL созданы на основе декларативной обработки запросов с использованием XQuery и используют стратегии отображения модели данных, таких как ключ/значение, столбцы и ориентированные на документы, в модель данных XDM.

Камачо-Родригес и др. [16] описали механизм распределенных запросов для управления большими объемами XML-документов поверх Amazon Cloud. Представлены три стратегии индексирования и реализовано подмножество XPath и XQuery. В исследовании, проведенном Wen-Syan et al. [17], облачное информационное устройство под названием Xbase предлагается специально для крупномасштабных и сложных медицинских приложений. Xbase применяет гибридный подход к распределенным репозиториям XML-документов на основе RDBMS и Hadoop HDFS, что обеспечивает более высокую производительность обработки запросов, а также снижает затраты на проектирование системы.

Шенк и др. [18] разработали постоянное хранилище под названием Xomoda, которое отображает XML-данные в хранилище с широкими столбцами. В Xomoda XML-документы разбиты на узлы, и эти узлы индексируются двумя разными индексами. Xomoda имеет некоторые общие черты с XML2HBase, но отличается двумя важными аспектами. Во-первых, Xomoda фокусируется только на запросах осей XML, и авторы не обсуждали другие операции с данными XML. Во-вторых, Xomoda хранит всю информацию (включая индекс пути, имя, тип, значение и количество одноуровневых узлов) каждого узла в XML-дереве в HBase, что в свою очередь создает огромную таблицу данных и проблему с производительностью.

Анализ и выводы

Все вышеизложенные подходы были предложены с целью хранения и извлечения разнородных данных на основе XML. Основная стратегия запросов, используемая для реляционных решений, заключается в извлечении табличных данных с использованием традиционного декларативного языка запросов: XPath или XQuery. Но их использование для обеспечения возможностей улучшения скорости обработки и масштабирования является малоэффективной.

На этом фоне большие преимущества в распределенной и параллельной обработке имеют базы данных NoSQL. Они широко используются для крупномасштабного хранения больших данных и высокопроизводительных вычислений.

Облачные системы предназначены для обеспечения достаточного объема памяти и компьютерных ресурсов для управления и совместного использования больших наборов разнородных данных. Это позволяет предотвратить неоправданную избыточность данных в сети на основе сервис-ориентированной архитектуры и соответствующих протоколов. Интеграция данных поддерживает операции с композицией и временным (виртуальным) представлением данных, хранящихся у различных владельцев. Данные остаются под контролем владельца и извлекаются по требованию клиентов для интегрированного доступа, возможно на коммерческой или свободной основе.

Заключение

В дальнейшем, в проводимом нами исследовании по внедрению адаптивного графического редактора XML-записей в распределенной информационной системе, предполагается добавление в систему версионирования документов, и, как следствие, добавление в приложение редактирования гетерогенных XML-документов поддержки версионирования. Также планируется добавление поддержки темпоральных (временных) документов, с разными версиями одного и того же элемента в документе, а также мульти-

схемных документов.

Список литературы

- [1] Nassiri H., Machkour M., Hachimi M. Integrating xml and relational data. Proc.Comput. Sci. 2017. 110 422–427.
- [2] Brahmia, Z., Hamrouni, H., Bouaziz, R. XML data manipulation in conventional and temporal XML databases: A survey. Computer Science Review Volume 36, 1 May 2020, №100231.
- [3] A.V. Lyamin, E.N. Cherepovskaya. “Xml-relational mapping using production rule system.” Intelligent Systems Conference, IntelliSys 2017, IEEE, 2017, 422–429.
- [4] A.Y. Halevy, Z.G. Ives, J. Madhavan, P. Mork, D. Suciu, I. Tatarinov. “The Piazza peer data management system.” IEEE Trans. Knowl. Data Eng. 16 (7) (2004), 787–798.
- [5] L.J. Chen, P.A. Bernstein, P. Carlin, D. Filipovic, M. Rys, N. Shamgunov, J.F. Terwilliger, M. Todic, S. Tomasevic, D. Tomic. “Mapping xml to a wide sparse table.” IEEE Trans. Knowl. Data Eng. 26 (6) (2014), 1400–1414.
- [6] I. Tatarinov, S.D. Viglas, K. Beyer, J. Shanmugasundaram, E. Shekita, C. Zhang. “Storing and querying ordered xml using a relational database system.” Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, ACM, 2002, 204–215.
- [7] P. O’Neil, E. O’Neil, S. Pal, I. Cseri, G. Schaller, N. Westbury. “Ordpaths: insert-friendly xml node labels.” Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, ACM, 2004, 903–908.
- [8] M. Marjani, F. Nasaruddin, A. Gani, S. Shamshirband. “Measuring transaction performance based on storage approaches of native xml database.” Measurement 114 (2018), 91–101.
- [9] S. Balamurugan, A. Ayyasamy. “Performance evaluation of native xml database and xml enabled database.” Int. J. Adv. Res. Comput. Sci. Softw. Eng. 7 (5) (2017).
- [10] A.M. Saba, E. Shahab, H. Abdolrahimpour, M. Hakimi, A. Moazzam. “A comparative analysis of xml documents, xml enabled databases and native xml databases.” Preprint, arXiv:1707.08259, 2017.
- [11] A.Y. Halevy, Z.G. Ives, J. Madhavan, P. Mork, D. Suciu, I. Tatarinov. “The Piazza peer data management system.” IEEE Trans. Knowl. Data Eng. 16 (7) (2004), 787–798.
- [12] W.S. Ng, B.C. Ooi, K.-L. Tan, A. Zhou. “Peerdb: a p2p-based system for distributed data sharing.” 19th International Conference on Data Engineering, 2003, Proceedings, IEEE, 2003, 633–644.
- [13] S. Abiteboul, I. Manolescu, N. Polyzotis, N. Preda, C. Sun. “Xml processing in dht networks.” 2008 IEEE 24th International Conference on Data Engineering, IEEE, 2008, 606–615.
- [14] G. Cong, W. Fan, A. Kementsietsidis, J. Li, X. Liu. “Partial evaluation for distributed xpath query processing and beyond.” ACM Trans. Database Syst. 37 (4) (2012), 32–43.
- [15] J. Camacho-Rodríguez, D. Colazzo, I. Manolescu Paxquery. “Efficient parallel processing of complex xquery.” IEEE Trans. Knowl. Data Eng. 27 (7) (2015), 1977–1991.
- [16] J. Camacho-Rodríguez, D. Colazzo, I. Manolescu. “Building large xml stores in the Amazon cloud.” 2012 IEEE 28th International Conference on Data Engineering Workshops, ICDEW, IEEE, 2012, 151–158.
- [17] W.-S. Li, J. Yan, Y. Yan, J. Zhang. “Xbase: cloud-enabled information appliance for healthcare.” Proceedings of the 13th International Conference on Extending Database Technology, EDBT, ACM, 2010, 675–680.
- [18] A. Šenk, M. Valenta, W. Benn. “Distributed evaluation of xpath axes queries over large xml documents stored in mapreduce clusters.” 2014 25th International Workshop on Database and Expert Systems Applications, IEEE, 2014, pp. 253–257.

TECHNOLOGIES FOR STORING AND QUERYING LARGE COLLECTIONS OF XML DOCUMENTS

A.Mukhitova^{1,2}, A. Yerimbetova^{2,3}, V. Barakhnin⁴

¹al-Farabi Kazakh National University, Almaty, Kazakhstan

²Institute of Information and Computational Technologies, Almaty, Kazakhstan

³Satbayev University, Almaty, Kazakhstan

⁴Federal Research Center for Information and Computational Technologies, Ac., 630090 Novosibirsk, Russia

¹ORCID ID: <https://orcid.org/0000-0002-4081-7694>, mukhitova.aigul@gmail.com

²ORCID ID: <https://orcid.org/0000-0002-2013-1513>, aigerian@mail.ru

⁴ORCID ID: <https://orcid.org/0000-0003-3299-0507>

Abstract. Currently, web applications using XML repositories (banking, accounting, personnel

management, airline reservations, weather monitoring and forecasting, electronic control and e-commerce) process a huge amount of temporary information and need a complete history of data and schema changes for effective and transparent management of the information system. Existing XML DBMS and XML tools do not support this function, which makes it difficult to build a history of changes, difficult to return to the previous state, and the possibility of data loss. As a result, there is a problem of effective management of temporary data storage when working with a large number of small XML documents and many additional elements /attributes.

Keywords: XML encoding scheme, XML data mapping, NoSQL database, XML query processing and optimization

XML ҚҰЖАТТАРДЫҢ ҮЛКЕН КОЛЛЕКЦИЯСЫНА СҰРАНЫСТАР МЕН САҚТАУДЫ ҰЙЫМДАСТЫРУ ТЕХНОЛОГИЯЛАРЫ

А.Мухитова^{1,2}, А.Еримбетова^{2,3}, В. Баракнин⁴

¹Ақпараттық және есептеуіш технологиялар институты ҒК БҒМ, Алматы, Қазақстан

²Әл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы, Қазақстан

³Satbayev University, Алматы, Қазақстан

⁴Ақпараттық және Есептеу Технологиялары бойынша Федералды Зерттеу Орталығы, РФА,
630090, Новосибирск, Ресей

Аңдатпа. Қазіргі уақытта XML репозиторийлерін қолданатын веб-қосымшалар (банк ісі, бухгалтерлік есеп, адам ресурстарын басқару, авиакомпанияларды брондау, ауа райын бақылау және болжау, электрондық үкімет және электрондық коммерция) үлкен көлемдегі уақытша ақпаратты өңдейді және ақпараттық жүйені тиімді және ашық басқару мақсатында деректер мен схема өзгерістерінің толық тарихын қажет етеді. Қолданыстағы XML ДҚБЖ және XML құралдары бұл функцияны қолдамайды, бұл өзгерістер тарихын құруды қиындатады, алдыңғы күйге оралуды қиындатады және деректердің жоғалу мүмкіндігін тудырады. Нәтижесінде, уақытша мәліметтерді сақтауды тиімді басқару мәселесі көптеген шағын XML құжаттарымен және көптеген қосымша элементтермен/атрибуторлармен жұмыс істегенде туындайды.

Кілттік сөздер: XML кодтау сызбасы, XML деректерін салыстыру, NoSQL деректер қоры, XML сұраныстарын өңдеу және оңтайландыру.

Сведения об авторах

Қаз: Мухитова Айгуль - Әл-Фараби атындағы Қазақ ұлттық университетінің докторанты, mukhitova.aigul@gmail.com

Рус: Мухитова Айгуль - докторант Казахского национального университета им. аль-Фараби, mukhitova.aigul@gmail.com

Англ: Mukhitova Aigul – a doctoral student at Al-Farabi Kazakh National University, mukhitova.aigul@gmail.com

Қаз: Еримбетова Айгерим Сембековна – Ақпараттық және есептеуіш технологиялар институтының аға қызметкері, Satbayev University профессоры, PhD, т.ғ.к., қауымд. профессор, aigerian@mail.ru

Рус: Еримбетова Айгерим Сембековна – старший научный сотрудник Института информационных и вычислительных технологий, профессор Satbayev University, PhD, к.т.н., ассоц. профессор, aigerian@mail.ru

Англ: Yerimbetova Aigerim Sembekovna – Senior Researcher of Institute of Information and Computing Technologies, Professor of Satbayev University, PhD, Candidate of Technical Sciences, Assoc. Professor, aigerian@mail.ru

Қаз: Владимир Баракнин - Ақпараттық және Есептеу Технологиялары бойынша Федералды Зерттеу Орталығының профессоры, т. г. д. РФА, 630090, Новосибирск, Ресей

Рус: Владимир Баракнин – д.т.н., профессор Федерального Исследовательского Центра Информационных и Вычислительных технологий, РАН, 630090, Новосибирск, Россия

Англ: Vladimir Barakhnin - Professor, Doctor of Engineering Sciences of Federal Research Center for Information and Computational Technologies, Ac., 630090 Novosibirsk, Russia.

UDC 004.02
IRSTI 20.53.15

SURVEY OF MODERN TRENDS IN COMPUTER SECURITY AND ARTIFICIAL INTELLIGENCE

Mirzakhmet Syzdykov

Satbayev University, Almaty, Kazakhstan

mshpmail598@gmail.com

ORCID ID: <https://orcid.org/0000-0002-8086-775X>

Abstract. In this article we present the modern trends like Explainable Machine Learning (EML) versus classical Machine Learning (ML), the computer security is to be evaluated from the point of view of the hashing and authentication schemes in Hyper-Text Transfer Protocol (HTTP), most of the emphasis of this work is made towards hashing in representing the identity of data sources and making possible to use these hashes as crypto currencies in block-chaining technology, the survey also gives the definition of the basic hashing algorithm in the assumption that $P \neq NP$, where P is the polynomial class of complexity and NP is a non-polynomial class, since this fact was proved before as equality between P and NP , this is still remains challenging to prove that hashing function can be dual for either input or output, thus giving the possibility of compromising the computer security system and obtain the prior hash in hashing queue in block-chaining and, probably, partial chunks of the hashed data.

Keywords: hashing, algorithm, block-chaining, computer security, artificial intelligence.

Introduction

We present the description of HTTP-protocol authentication schemes like “basic” and “digest” based upon encoding and hashing validation of the user identity.

Even since, the proof of equivalence of complexity classes it doesn't mean that there could exist the inverse function for hash as this question remains open – we give the strict proof that even since assumption of equality of P and NP -classes ($P = NP$), this is still impossible to create the inverse function. This proof is based upon definition of the universal hashing algorithmic steps which are dependent on the input. At any time and at each point this becomes evident that data to be obtained from the inverse steps cannot lay within the hash size which should as small as possible with respect to the average size of input data which are hashed on the one-direction function like hash probing.

We also give the notion to the rapidly growing interest of variety of communities for Machine Learning. We give the definition of algorithm of Explainable Machine Learning in order to define the extensible role of the ML in the field of algorithmic approach rather than resorting to the classical approach of neural networks.

HTTP basic authentication

Authentication schemes like basic and digest were presented in the official standard [1]. The basic authentication scheme prerequisites are discussed in [2]. As we operate on HTTP it's necessary to study the initial version of the protocol [3] – to the present day the version “2.0” is in active usage by many software servers like Apache or nginx, we name our project as “Alumni”, version 1.0. The extension to the digest authentication protocol is presented in [4]. The security issues were well studied by author in [5]: these issues commonly are of programmatic case like buffer overflow conditions, when stack is rewritten in the limited space and the return address is thus modified giving the possibility of compromising the system and running arbitrary shell code [5].

According to the server statistics in World Wide Web (WWW) the most popular web servers are IIS, Apache and nginx.

The authentication scheme of the basic authentication is provoked by the server response if “Authorization: -field is missing and authorization token is followed encoded using base 64 encoding algorithm, the encoded string contains both login and password separated by “:” symbol.

In turn the server sends “WWW-Authenticate: basic[/digest], realm='...’” response if the authentication values are missing and authorization is required for the requested source. In this response

the field “realm” corresponds to the arbitrary identification of secured source.

The Socket Security Layer (SSL) is used in the modern web software and, thus, avoids the evaluation of data by reading data directly on Transfer Control Protocol (TCP) level.

The basic architecture of the web server is based upon the queues and, thus, represents the classical example of the system of mass-servicing.

The basic authorization can be replaced with the usage of hash token after validating the user data.

Hashing as a Modern Trend

The hashing is the term which is about to obtain the short part describing the whole part of the input data, most of these functions are to be one-directional thus giving the protection against the inverse determination of the hashed data.

The example of algorithm to be studied for the case of security level is presented in [6] and is named as a Whirlpool algorithm.

In other meaning the hashing functions give the time complexity and performance of $O(1)$ for the hash trees which are widely used in the modern Database Management Systems (DBMS) like Oracle, etc.

The block-chaining based upon the hash algorithm which describes the crypto currency is a task of obtaining the new hash at each transaction with the help of encryption algorithm – we divide these algorithms as past and present, which was elaborated by the author of this work. The classical example of hashing in block-chain is to hash the prior value to the values like additional attributes and, thus, giving the new value for the next crypto currency transaction. We elaborate and give the simple algorithm to obtain these hash values:

1. Generate random hash and encrypt using one-way algorithm with arbitrary key;
2. Decrypt the hash value;
3. Add amount of crypto currency as a big number to the hash;
4. Encrypt the next value;
5. Yield the value from step 4.

From the above algorithm, it follows that we don't need the classical approach of mining in order to obtain the hash values at each step as this process by itself compromises the security risks of data collection by third parties and, thus, is vulnerable to the hashing attacks.

The last statement is prevalent since the MD5-hash was compromised in a distributed network in year, circa, 2003.

Our algorithm is secure as the key-values evaluation relays on the SSL connection, where the encryption key can be secured and stored in the local storage – this technology is widely used by the modern browsing software like Google Chrome and Microsoft Edge.

The basic hashing algorithm is defined as the number of steps for the data-window of the fixed size to be processed against the larger amount of data:

1. Generate window length less than input;
2. Set current position to zero;
3. For each input symbol module current position increment, set the value in the window array to the bit-mask operation like exclusive-OR.
4. Repeat the step 3 until the end of input.

As we have stated before, the hashing function cannot be inversed – we prove it by using the above example: since the window is less than input it's obvious that the whole part of data cannot be obtained even if the algorithm is known to the performer of this task.

Explainable Machine Learning

Modern trends of Machine Learning like the TensorFlow and imaginary detection is briefly discussed in [7]. The existing algorithms are described in [8] along with a neural network method. The use of Machine Learning in the present day within the practical experience is presented in [9] – this

research states the question of the evaluation of Machine Learning in different types of application and its impact on the global stage. The modern usage of ML in wireless networks as the application is presented in [10].

After the overview of the research on the latest publications for Machine Learning and its application with pros and cons, we present the algorithmically practical solution for ML, known as Explainable Machine Learning, or, simply, EML.

EML relays on input data and can be represented as the matrix of sorted elements by the category and priority at each row and column, the steps of building this matrix and evaluation of the query are defined as follows:

1. Sort the input data matrix with given priority of each factor;
2. Query the result for each data in the sorted matrix for the short-coming range of lower and upper row;
3. For the number values do the floor or ceil of the value and compare it same way as the word entities in the input matrix.

Neural networks which are based upon sigmoid function can be extended for EML as the algorithm for arbitrary function with respect to the term various probability type which is well-known in modeling theory of mass-servicing queues.

The algorithm for arbitrary function $f(x)$ is as follows:

1. Define the function $f(x)$ and its range L and R ;
2. Compute the minimum and maximum values of function $f(x)$ on this range using differential calculus;
3. Train the neural network by applying function $f(x)$ divided by the minimum and maximum, with subtraction of minimum.

The steps above give the novel example of the arbitrary learning without usage of sigmoid-function for neural network training.

Conclusion

Thus, we have presented the novel block-chaining algorithm which is secure and fast and doesn't require expensive mining operation to be performed on the specific devices which utilize Nvidia chipset bundle for this purpose.

The trend for computer security is also given with respect to the hash evaluation after authorizing user using standard methods or methods which involve security token or "cookie" safety.

We have also represented the modern results on the Machine Learning, which is named in this article as Explainable Machine Learning and, thus, is more relevant and doesn't require the neural network to be trained as the EML-decision tree can be expanded by online query also in the linear time as well as the query performance time $O(n)$, where n is the number of factors in the data matrix of the input.

Thus, the modern trends are to be seen from other point of view with respect to the prior results on algorithmic stage like the prior experience of the authors of these methods.

Acknowledgements

The author expresses gratitude to the creators of SharpDevelop programming environment for creating a valuable and free package for effective and professional software development in C# on .NET 2.0 platform.

Funding

This work was partially supported by an educational grant of the Ministry of Education and Sciences of the Republic of Kazakhstan during author's studying at the Satbayev University from 2001 to 2006.

References

- [1] Franks John, et al. HTTP authentication: Basic and digest access authentication. 1999. 2617.

- [2] Reschke Julian. The 'Basic' HTTP authentication scheme. 2015. 7617.
- [3] Fielding Roy, Julian Reschke. Hypertext transfer protocol (http/1.1): Authentication. 2014. 7235.
- [4] Franks John, et al. An extension to HTTP: digest access authentication. 1997. 2069.
- [5] Сыздыков Мирзахмет. Win32-buffer/heap overrun conditions или секреты написания эксплоитов. Security Lab.ru.
https://www.securitylab.ru/analytics/216279.php?el_id=216279&VOTE_ID=101&view_result=Y. (accessed Jan 25, 2023)
- [6] Barreto P. S., Vincent Rijmen. The Whirlpool hashing function. First open NESSIE Workshop, Leuven, Belgium. 2000. 13
- [7] Jordan Michael I., Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. Science 2015. 349.6245. 255-260.
- [8] Mahesh Batta. Machine learning algorithms-a review. International Journal of Science and Research (IJSR). 2020. 9. 381-386.
- [9] Athey Susan. The impact of machine learning on economics. The economics of artificial intelligence: An agenda. University of Chicago Press. 2018. 507-547.
- [10] Chen Mingzhe, et al. Artificial neural networks-based machine learning for wireless networks: A tutorial. IEEE Communications Surveys & Tutorials 21.4. 2019. 3039-3071.

ОБЗОР СОВРЕМЕННЫХ ТЕНДЕНЦИЙ В ОБЛАСТИ КОМПЬЮТЕРНОЙ БЕЗОПАСНОСТИ И ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Мырзахмет Сыздыков

Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан
mspmail598@gmail.com

ORCID ID: <https://orcid.org/0000-0002-8086-775X>

Аннотация. В этой статье мы представляем современные тенденции, такие как объяснимое машинное обучение (EML) по сравнению с классическим машинным обучением (ML), компьютерная безопасность должна оцениваться с точки зрения схем хэширования и аутентификации в протоколе передачи гипертекста (HTTP), основной акцент в этой работе сделан на хэширования для представления идентичности источников данных и обеспечения возможности использования этих хэшей в качестве криптовалют в технологии блокчейн-цепочек, в исследовании также дается определение базового алгоритма хэширования в предположении, что P не равно NP, где P - полиномиальный класс сложности, а NP - неполиномиальный класс, поскольку этот факт был доказан ранее как равенство между P и NP, все еще остается сложной задачей доказать, что функция хэширования может быть двойной как для ввода, так и для вывода, что дает возможность поставить под угрозу систему компьютерной безопасности и получить предыдущий хэш в очереди хэширования в блочной цепочке и, возможно, частичные фрагменты хэшированных данных.

Ключевые слова: хэширование, алгоритм, блочная цепочка, компьютерная безопасность, искусственный интеллект.

КОМПЬЮТЕРЛІК ҚАУІПСІЗДІК ПЕН ЖАСАНДЫ ИНТЕЛЛЕКТТІҢ ЗАМАНАУИ ТЕНДЕНЦИЯЛАРЫНА ШОЛУ

Мырзахмет Сыздыков

Қ. И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан
mspmail598@gmail.com

ORCID ID: <https://orcid.org/0000-0002-8086-775X>

Аңдатпа. Бұл мақалада біз классикалық машиналық оқытумен (ML) салыстырғанда түсіндірме машиналық оқыту (EML) сияқты заманауи тенденцияларды ұсынамыз, компьютерлік қауіпсіздікті гипермәтінді беру протоколындағы (HTTP) хэштеу және аутентификация сызбалары тұрғысынан бағалау керек. Бұл жұмыстың негізгі бағыты деректер көздерінің сәйкестігін көрсету және қамтамасыз ету үшін хэштеу болып табылады. Бұл хэштерді блокчейн тізбегі

технологиясында криптовалюта ретінде пайдалану мүмкіндіктері, зерттеу сонымен қатар P NP - ге тең емес деген болжаммен негізгі хэш алгоритмін анықтайды. Мұндағы p-көпмүшелік күрделілік класы, ал NP-көпмүшелік емес класс, өйткені бұл факт бұрын P мен NP арасындағы теңдік ретінде дәлелденген, хэштеу функциясы енгізу үшін де, шығару үшін де қосарлы болуы мүмкін екенін дәлелдеу әлі де қиын міндет болып табылады. Бұл компьютерлік қауіпсіздік жүйесіне қауіп төндіруге және алдыңғысын алуға мүмкіндік береді блокчейндегі хэштеу кезегіндегі хэш және мүмкін болатын хэштелген деректердің ішінара фрагменттері болады.

Кілттік сөздер: хэштеу, алгоритм, блокчейн, компьютерлік қауіпсіздік, жасанды интеллект.

Сведения об авторе:

Анғ.: Syzdykov Mirzakhmet - Satbayev University, Almaty, Kazakhstan

Қаз.: Сыздықов Мырзахмет- Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті, Алматы, Қазақстан.

Рус.: Сыздықов Мырзахмет- Казахский национальный исследовательский технический университет имени К.И. Сатпаева, Алматы, Казахстан.

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

Deputy chief editor: PhD, Mamyrbayev O.Zh

The editorial board of the journal " Advanced technologies and computer science " is not responsible for the content of published articles. The content of the articles belongs entirely to the authors and is posted in the journal solely under their responsibility.

Signed in print 03.03.2020
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