UDK 00.015.23

DATA ANALYSIS IN DISTRIBUTED INFORMATION SYSTEMS

S.S. Beknazarova
Professor of TUIT Doctor of technical science, associate of professor

Department of Audiovisual technologies, Tashkent University of information technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
E-mail: saida.beknazarova@gmail.com

S.S. Beknazarova
Doctor of technical Sciences, associate Professor, professor of the Department of Audiovisual technologies, faculty of Television technologies, Tashkent University of information technologies named after Muhammad al-Khwarizmi. Author of 189 research papers on audio, video, multimedia resources, and media education technologies.

Abstract. The article describes the methods of data analysis used in the design of distributed systems that provide a continuous flow of data without loss, inherent in various by nature properties. The main feature of a distributed system is a technique from performing such functions as receiving, processing, transmission of various data. In this we define that the data entered into the system has a quantitative property. Quantitative data-discrete and especially continuous data (text, image, audio, video) - gives way to a much wider spectrum of possibilities in terms of statistical analysis, due to the more perfect scales of measurement and the possibility of quantifying the differences between them.

Keywords: distributed system, lossless data processing, ensuring continuity of data flow, various types of information, data analysis, quality data.

Introduction. The main feature of a distributed system is a technique from performing such functions as receiving, processing, transmission of various data. In this we define that the data entered into the system has a quantitative property. Quantitative data-discrete and especially continuous data (text, image, audio, video) - gives way to a much wider spectrum of possibilities in terms of statistical analysis, due to the more perfect scales of measurement and the possibility of quantifying the differences between them.

Since case studies are often associated with a large amount of data at the level of abstraction, the researcher is the first to group them, having studied their internal structure, the question arises of presenting such data with a good visualization weight in a more compact and accessible form.

Let There Be n ta data of the projected system x1, x2,XP each of them x characterizes one quantitative sign. The issue of processing this data is solved. If the number of observations (n) is large enough (at least N≥ 50), then First They are brought and grouped in a certain order.

Variation Series-x1, x2, of the quantitative sign of X, located in the ascending (descending) order ...N value of XP independent observations.

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(i)} \leq x_{(i+1)} \leq \cdots \leq x_{(n)}$$

The element X(1) is called the I order statistics, the main order statistics are the minimum value of x(1)=min{x(2)} and the maximum value of x(1)=Max{x(2)}. The designation of the index in small brackets (i) is the designation (sign, symbol, symbol)of all accepted by the regulator of observations (regulated in descending order).

The difference between the largest and the smallest value of the symptom is called the range of vision of the variation range:

$$R = x_{(n)} - x_{(1)} = x_{max} - x_{min}$$

The irregular spread of the values of the sign under study of the range of Vision serves as an independent characteristic.

In large-scale Observations, they are grouped, with the aim of improving the presentation of empirical data (data visualizationurni) by generating a grouped line of data.

For the grouping of discrete quantitative data, the M1 of each X1 symptom is the Met periodicity. When the values are large enough, the grouped variation is carried out again (next) grouping of the row, turning it into an interval row.

Table 1. Grouping of discrete quantitative data

x_i the value of the token	x_i	$x_{(1)}$ $x_{(2)}$	…	$x_{(i)}$	…	$x_{(k)}$
m_i frequencies	m_i	m_1 m_2	…	m_i	…	m_k

The grouped discrete variation series represents the k value of the X1 symptom shown together with the corresponding M1 frequency or W1=mi/n (this frequency is called the empirical).

The grouped discrete variation range is expressed in the form of histograms or polygons in graphical form.

Polygon is a graphical representation of a discrete series of oscillations grouped in the form of a fracture that corresponds to all possible values of the sign on the abscissa axis, while on the ordinate axis the value of the frequencies m1 or W1=mi/n corresponds to the value of the relative frequencies.

To ensure a clear view, the scale on the arrows is selected voluntarily. Like histogram, Polygon discrete variable values make it possible to evaluate the frequency distribution, to determine the maximum (mod) and minimum encountered values of the symptom.

The grouped cumulative discrete variation range represents the value of the xi symptom displayed along with the corresponding cumulative mih frequencies or wih= mih/n frequencies.

Table 2. Discrete variation range

x_i the value of the token	$x_{(1)}$	$x_{(2)}$	…	$x_{(k)}$
collected m_{ih} frequencies	$m_{1h} = m_1$	$m_{2h} = m_1 + m_2$	…	$m_{kh} = n = \sum_{i=1}^{k} m_i$

The graphical representation of the grouped cumulative discrete series of variations is described in the form of cumulates.

Cumulative curve line (cumulative line graph) (cumulative line graph), or ogiva (ogive) - a graphical representation of the grouped cumulative discrete line of variation in the form of columns, when viewing it, all the values that can be a sign on the axis of the abscissa are distinguished, while on the ordinate axis, the frequencies or the relative frequencies that are accumulated to this value are distinguished. Cumulants indicate the number (or percentage)of all objects of cumulants that do not exceed the value of the given signs.

Example: In the distributed system, 30 randomly selected pieces of data are received, and in the processing of 10 pieces of data (text, image, audio, video), the presence of errors is required to compile a discrete line of data variation according to the data given in the results table of the numerical analysis and obtain different graphical images of the data series – histogram, Polygon, cumulation and 30 randomly selected pieces of information received in the projected distributed system.

Table 3. Selected pieces of information received

5	1	3	4	5	6
1	3	4	6	8	1
4	6	3	3	2	2
2	6	3	3	7	2
0	0	4	1	2	4

Solution:

In order to formulate (build) a series of variations, it is necessary to put the value in the order of multiplying. It is easy to do this with the help of the sorting A – ya function in MS Excel. The number of error availability is given in Table 2 of the variation range obtained by order.

30 randomly selected pieces of information received in the projected distributed system.

Table 4. Projected distributed system

x_i	0	0	1	1	1	1	2	2	2	2	2	3	3	3	3
	3	3	4	4	4	4	4	5	5	6	6	6	6	7	8

It is possible to draw (construct) a grouped row by easily calculating how many times each value matches according to the resulting data series.

If the number of values is too much, you can compile the grouped data series using MS Excel's frequency function at once, without compiling the data series. This makes it possible to calculate the frequency of values of a given mass of data that fall into their given intervals and receive the given values.

To do this, after determining the possible values of the sign, it is necessary to divide the sphere (Gray) consisting of Army yachts (the number of bunda yachts will be more than one unit of the number of values), call the frequency statistical function, divide the array of data and the array of values of the sign in the corresponding windows of the function Table 5.

Table 5. Values of the sign in the corresponding windows

5 1 3 4 5 6 1 3 4 5 8 1 4 6 3 3 2 2 2 6 3 3 7 2 0 0 4 1 2 4	0 1 2 3 4 5 6 7 8	2 4 5 6 5 2 4 1 1 0

In this case, the data obtained in a non-large array of data can be easily checked according to Table 5.

In the same way it is possible to construct (construct) a grouped line of data analysis in MS Excel using the – histogram module.

Thus, the obtained grouped series of data error frequency and relative frequencies of error and cumulative series of data error are presented in Table 6.

Grouped series of variation in the number of errors in the processing of 30 randomly selected data

Table 6. Cumulative series of data error

xi the values of the symptom – the number of existence of errors	0	1	2	3	4	5	6	7	8
mi frequency-the number of data in which the error is the same as the number of available	2	4	5	6	5	2	4	1	1
relative frequency wi=mi/n	$\frac{2}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{6}{30}$	$\frac{5}{30}$	$\frac{2}{30}$	$\frac{4}{30}$	$\frac{1}{30}$	$\frac{1}{30}$
the frequency of occurrence collected - the number of errors to be present at least xi-mih number of data being	2	6	11	17	22	24	28	29	30
wih=mih/n relative accumulated frequency	$\frac{2}{30}$	$\frac{6}{30}$	$\frac{11}{30}$	$\frac{17}{30}$	$\frac{22}{30}$	$\frac{24}{30}$	$\frac{28}{30}$	$\frac{29}{30}$	1

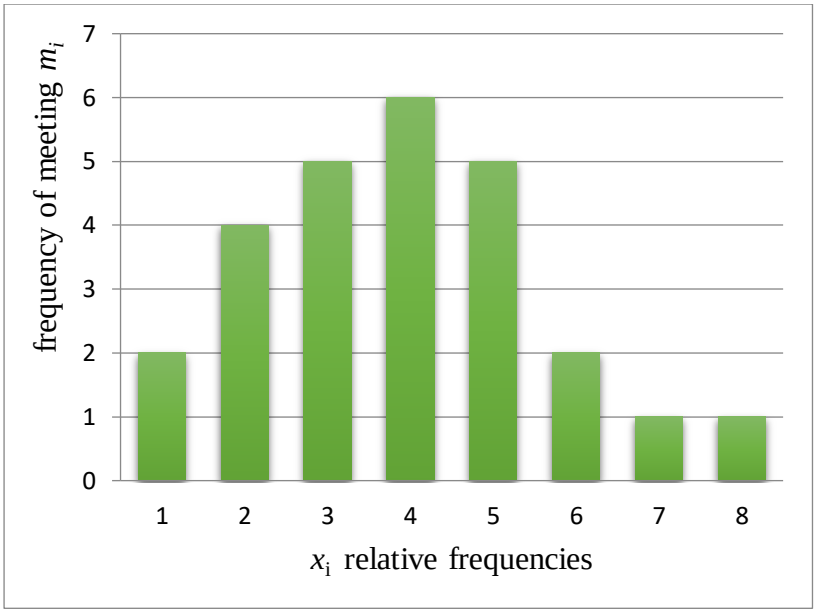

Figure 1. Histogram of the frequency of the number of errors available

On the basis of the data in Table 3, we can build all the required graphs of the frequencies and relative frequencies at which there will be errors – histogram, Polygon and cumulation.

Taqsimlash mod by histogram and Polygon – the number of errors that can be present at the maximum frequency: here these are the errors that can be present in 3 units (three in 6 units of dataraydi), and the minimum value of the symptom is 7 and the error that can be present in 8 units is only three in one raceraydi.

The polygon is built on the same points as the histogram, only in this we get a different view of the graph –not a column diagram, but a point-to-line chart.

For the construction of the cluster, we use the frequencies collected in Table 3 (Figure 3).

The cumulative graph allows you to find the number of objects with sign values that do not exceed the given. For example, 3-table and 3-as can be seen from the picture, 24 data have the number of existence of errors of no more than 5 (from 0 to 5 soles).

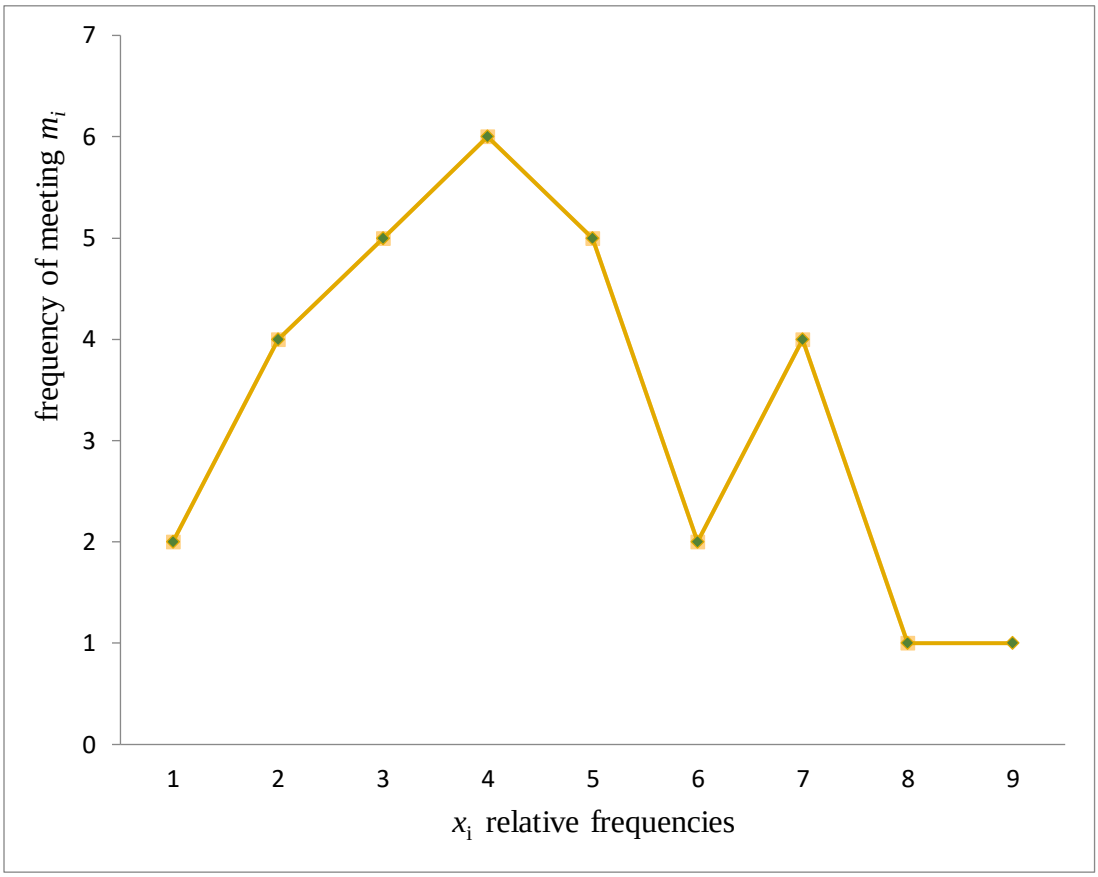

*Figure 2.*The polygon of the frequency of the error number of existence

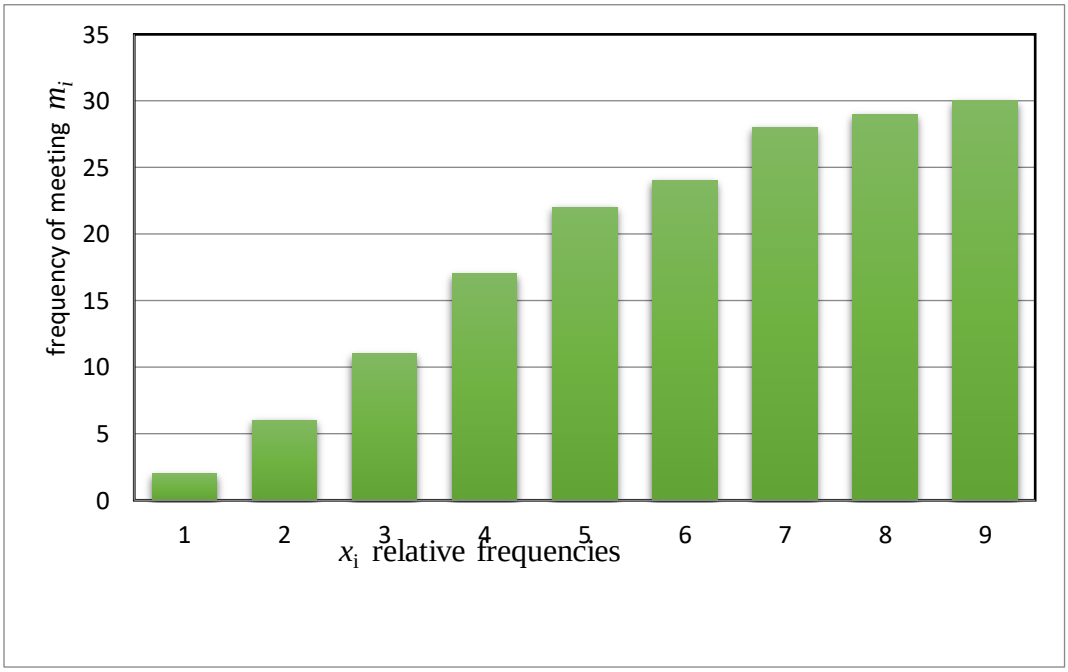

Figure 3. Error cumulation of the number frequency of existence

The graph of the histogram, Polygon and Cumulus of relative frequencies looks exactly the same, only on the ordinate Arrows are located relative frequencies, respectively, and they indicate their share, and not their number. For example, we see only a histogram of relative frequencies with drying (Figure 4).

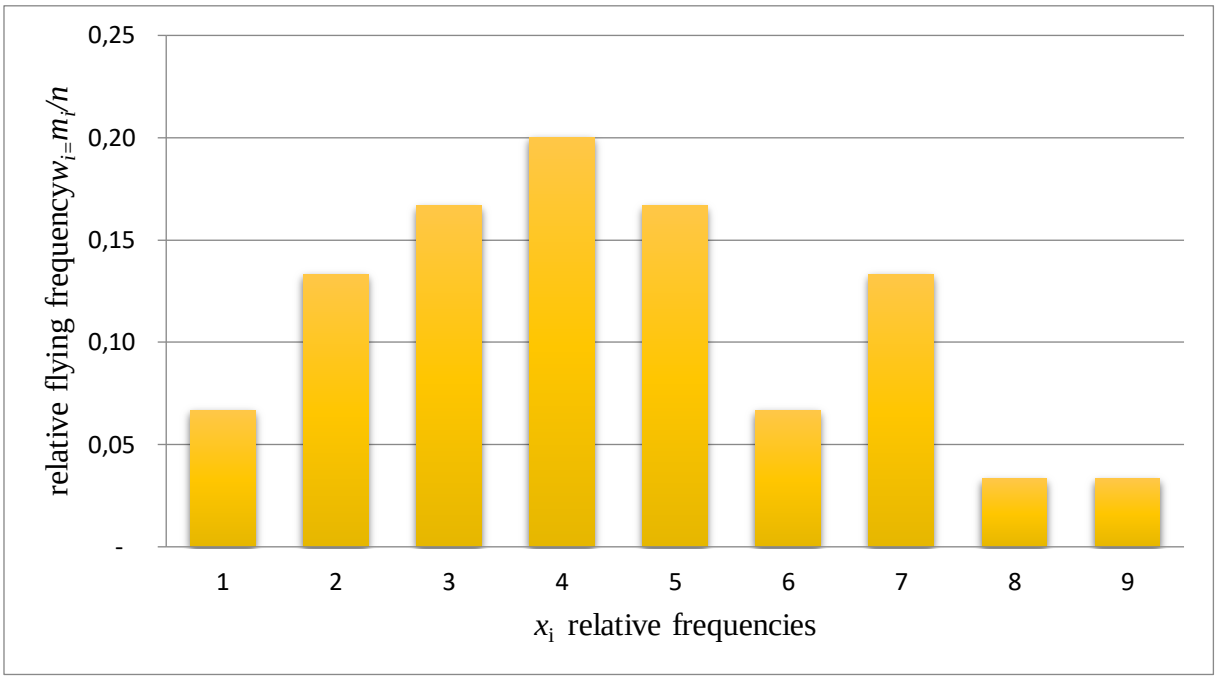

Figure 4. Histogram of the relative frequency of the number of existence of errors

The graph of relative frequencies shows the percentage of objects with the corresponding variable value. For example, as can be seen from the Figure 4 and Table 3, the fashion for three errors has a relative frequency of meeting, which is equal to 0,2. This means that 20 percent of all joint objects (selected data) (30 to 6 of them) have exactly 3 errors.

It is very relevant when processing large-scale information, especially when conducting modern scientific research, before the researcher stands a serious issue regarding the correct grouping of initial data. If the data is of discrete importance, as we have seen, problems do not arise – it is necessary to calculate only the mi frequency of each xi symptom. If the symptom under study is of continuous importance (more common in practice), the choice of the optimal number of intervals for grouping the symptom is not considered a trivial matter.

In order to group continuous random sizes, the entire range of vision of the symptom R=x(n)-x(1) is divided by the number of K intervals.

Grouped intervals (continuous) are listed in order on the value of that symptom in the series of variation (ai < x < bi) is said to be intervals, bunda I-the corresponding frequencies of the number of observations (mi) or (mi/n) are shown together with the relative frequencies i = 1, 2, ...k.

Table 7. Cumulative series of data error

$a_i \div b_i$ ranges of sign value	$a_1 \div b_1$	$a_2 \div b_2$	…	$a_i \div b_i$	…	$a_k \div b_k$
m_i frequency	m_1	m_2	…	m_i	…	m_k

Histogram and cumulation (ogiva), which have been studied in detail by US, is an excellent tool for visualizing data, allowing us to get a primary idea of the structure of the data. Such graphs (Figure 5) are formed both into continuous data and discrete data, taking into account that only continuous data can completely fill the field of its own values, which can be, if it accepts any values.

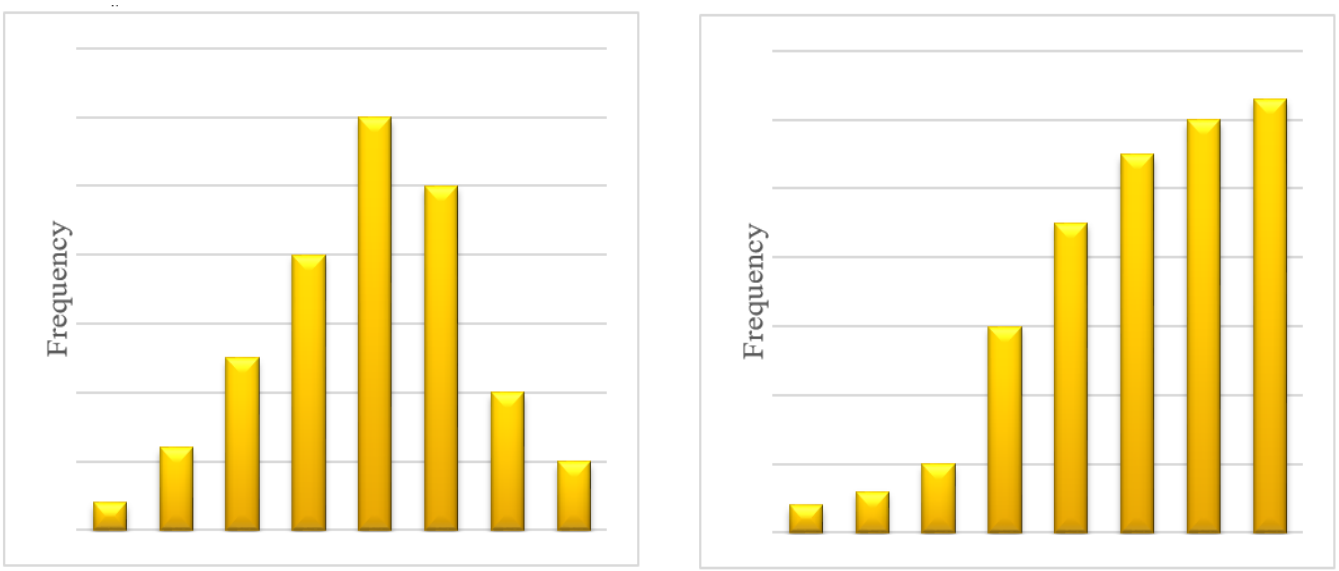

Figure 5. Histogram of the Interval row (in the chap) and cumulusatasi (on the right) sample (46-page)

Therefore, in histogram and cumulation, the columns must be touching each other (adjacent), the sign value must not have plots on which it can be located (that is, the histogram and cumulation should not have «holes» on which the value of the variable under study can be located, as in Figure 6-on the absciss arrows).

The height of the column corresponds to the mi frequency - this intervalalga the number of observations dropped or mi/n relative frequency – the percentage of observations. Intervals should not intersect and, as a rule, have the same width.

References

[1] Lighthill, M.J. A theory of traffic on long crowded roads / M.J. Lighthill, G.B. Whitham. – 1955. – Vol. 229. Pp. 317-345.

[2] В. И., Швецов Математическое моделирование транспортных потоков / Швецов В. И. – 2003.

[3] Payne, H. J. Models of freeway traffic and control / H. J. Payne. – 1971. – Pp. 51-56.

[4] Kerner, B. S. Deterministic spontaneous appearance of traffic jams in slightly inhomogeneous traffic flow / B. S. Kerner, P. Konhaeuser, M. Schilke // Phys. Rev. – P. 1995.

[5] Prigogine, I. A Boltzmann-like approach for traffic flow / I. Prigogine. – 1961.

[6] Newell, G. F. Nonlinear effects in the dynamics of car following / G. F. Newell. – 1961.

[7] Cremer, M. A fast simulation model for traffic flow on the basis of boolean operations /M. Cremer, J. Ludwig. – 1986. – Pp. 297-303.

[8] Nagel, K. A cellular automaton model for freeway traffic / K. Nagel, M. Schreckenberg //Physique I France. – 1992.

[9] А.Н., Котов. Моделирование дорожного движения на многополосной магистрали при помощи двумерного вероятностного клеточного автомата с тремя состояниями / Котов А.Н. – 2008.

[10] Automated Stationary Obstacle Avoidance When Navigating a Marine Craft (Conference Paper) Sedova, N., Sedov, V., Bazhenov, R., Karavka, A., Beknazarova, S. SIBIRCON 2019 - International Multi-Conference on Engineering, Computer and Information Sciences, Proceedings October 2019, Nomer stati 8958145, Pages 491-495 2019 International Multi-Conference on Engineering, Computer and Information Sciences, SIBIRCON 2019; Novosibirsk; Russian Federation; 21 October 2019 do 27 October 2019; Nomer kategoriiCFP1911E-ART; Kod 156894.

[11] USING SOCIAL NETWORK COMMUNITIES AS A TOOL FOR ORGANIZING IT EDUCATION D. Luchaninov, R. Bazhenov, T. Gorbunova, S. Beknazarova, L. Putkina, A. Vasilenko INTED2019 Proceedings Pages: 4104-4109 Publication year: 2019 ISBN: 978-84-09-08619-1ISSN: 2340-1079 doi: 10.21125/inted.2019.1029Conference name: 13th International Technology, Education and Development Conference Dates: 11-13 March, 2019 Location: Valen-cia, Spain.

АНАЛИЗ ДАННЫХ В РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

С. С. БЕКНАЗАРОВА
в.и.о.профессора Ташкентского университета информационных технологий имени Мухаммеда Аль-Хорезми, доктор технических наук, доцент

Кафедра аудиовизуальных технологий, Ташкентский университет информационных технологий имени Мухаммеда Аль-Хорезми, Ташкент, Узбекистан
E-mail: saida.beknazarova@gmail.com

Аннотация. В статье описываются методы анализа данных, используемые при проектировании распределенных систем, обеспечивающих непрерывный поток данных без потерь, присущих различным по своей природе свойствам. Главной особенностью распределенной системы является техника от выполнения таких функций, как прием, обработка, передача различных данных. При этом мы определяем, что данные, вводимые в систему, обладают количественным свойством. Количественные данные-дискретные и особенно непрерывные данные (текст, изображение, аудио, видео) - уступают место гораздо более широкому спектру возможностей с точки зрения статистического анализа, благодаря более совершенным шкалам измерения и возможности количественной оценки различий между ними.

Ключевые слова: распределенная система, обработка данных без потерь, обеспечение непрерывности потока данных, различные виды информации, анализ данных, количественные данные.