УДК 004.052

Gorbenko A. V.
Doctor of Technical Sciences, Professor, Department of Computer Systems and Networks National Aerospace University. N. E. Zhukovsky "HAI", Ukraine;
e-mail: A.Gorbenko @ csn.khai.edu
Ruban V. I.
graduate student of Computer Systems and Networks National Aerospace University. N. E. Zhukovsky "HAI", Ukraine;
e-mail: rubanvit@mail.ru

CONCEPTION OF THE MANAGEMENT THE COMPARATIVE ANALYSIS OF THE METHODS OF RECOVERY OF THE PASSED DATA OF TIME OF AVAILABILITY OF THE SERVER INFORMATION ECONOMIC AREA

Abstract. In the article presents results of research of primary statistical characteristics of time of processing and data transmission are given in systems of information economic area. On the basis of the received results the analysis of methods of recovery of the passed data is made. In work methods of data processing with admissions are analyzed and the most rational are chosen. Change of the law of distribution of selection depending on a way of elimination of the passed data is investigated. Forecasting of time of availability of the server of information economic area is executed. The goal of this paper is primary statistical characteristics research of data processing and transmission time in IES. Particularly, the server availability time (SAT) was studied within this work. This time consists of Client-to-Server connection time, Server data processing time, and Server answer transfer time. The values of service time were taken for research, that were received while studying of cloud computing performance. It was established that if data loss is not over 6% distribution law of initial model selection saves. If data loss is between 9% and 15% methods primary statistical characteristics of studied selection were determined without consideration of data missing. Missed data was filled with data distributed by the same law as for initial selection with admissions.

Keywords: time of availability of the server, statistical data, the distribution law, indicated of Hurst, method sliding average, admissions of data.

Formulas: 14; fig.: 3, tabl.: 3, bibl.: 6

JEL Classification:G 21, F 29, L 41.

Горбенко А. В.
д.т.н., професор кафедри Комп'ютерних систем та мереж Національного Аерокосмічного університету ім. Н.Е. Жуковского «ХАИ», Україна;
e-mail:A.Gorbenko @ csn.khai.edu;
Рубан В. І.
аспірант кафедри Комп'ютерних систем та мережі Національного аерокосмічного університету ім. Н.Е. Жуковского «ХАИ», Україна;
e-mail: rubanvit@mail.ru

ПОРІВНЯЛЬНИЙ АНАЛІЗ МЕТОДІВ ВІДНОВЛЕННЯ ПРОПУЩЕНИХ ДАНИХ ЧАСУ ДОСТУПНОСТІ СЕРВЕРА В ІНФОРМАЦІЙНО-ЕКОНОМІЧНИХ ПРОСТОРАХ

Анотація. У статті наведено результати дослідження первинних статистичних характеристик часу обробки і передачі даних в інформаційно-економічних просторах. На підставі отриманих результатів виконано аналіз методів відновлення пропущених даних. Проаналізовано прийоми обробки даних з пропускам та обрано найбільш раціональні. Досліджена зміна закону розподілу вибірки залежно від способу усунення пропущених даних. Виконано прогнозування часу доступності сервера економічних

просторів. Мета роботи полягала у дослідження первинних статистичних характеристик часу обробки і передачі даних в інформаційно-економічних просторах. Зокрема, в рамках даної роботи вивчалося час доступності сервера.

Ключові слова: інформаційно-економічний простір, час доступності сервера, статистичні дані, закон розподілу, показник Херста, метод ковзного середнього.

Формул: 14; рис.: 3, табл.: 3, бібл.: 6

Горбенко А. В.
д.т.н., профессор кафедры Компьютерных систем и сетей Национального аэрокосмического университета им. Н.Е. Жуковского «ХАИ», Украина;
e-mail:A.Gorbenko @ csn.khai.edu;
Рубан В. И.
аспирант кафедры Компьютерных систем и сетей Национального аэрокосмического университета им. Н.Е. Жуковского «ХАИ»;
e-mail: rubanvit@mail.ru

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ВОССТАНОВЛЕНИЯ ПРОПУЩЕННЫХ ДАННЫХ ВРЕМЕНИ ДОСТУПНОСТИ В ИНФОРМАЦИОННО-ЭКОНОМИЧЕСКИХ ПРОСТРАНСТВАХ

Анотация. В статье приведены результаты исследования первичных статистических характеристик времени обработки и передачи данных в информационно-экономических пространствах**.** На основании полученных результатов выполнен анализ методов восстановления пропущенных данных. В работе проанализированы приемы обработки данных с пропускам и выбраны наиболее рациональные. Исследовано изменение закона распределения выборки в зависимости от способа устранения пропущенных данных. Выполнено прогнозирование времени доступности сервера в информационно-экономических пространствах. Цель работы состояла в исследование первичных статистических характеристик времени обработки и передачи данных в информационно-экономических пространствах. В частности, в рамках данной работы изучалось время доступности сервера.

Ключевые слова: Информационно-экономические пространства, время доступности сервера, статистические данные, закон распределения, показатель Херста, метод скользящего среднег,опропуски данных.

Формул: 14; рис.: 3, табл.: 3, библ.: 6

Introduction. The processes of information economics formation are followed by the broad implementation of Information and communication technology that gives opportunity for different organizations to present their products and services in convenient format, to analyze rivals activity, market situations and consumer needs online, the grows of economic activity which is reached by placing of all economic activity types in different Information and Economic Spaces (IES) and by network forms of cooperation [6, p. 80-90].

At present business becomes electronic, it means that any commercial activity among partners (purchases and sales of products and services, operations with securities at the stock market, contracts formation and execution, etc) is executed with the help of electronic documents exchange in information space. Cloud Computing is one of the most progressive directions of information technologies development. Availability, fault tolerance, profitability, simplicity, flexibility and scalability are the advantages of cloud computing.

Research analysis and problem definition. Papers [2, 3] are dedicated to timing characteristics research of cloud computing and Web services. In these papers such parameters were studied: statistical parameters of connection time between computer and remote cloud provider, data processing time and server response time which includes data processing time and remote server connection time. During data transmission connection break and data loss are possible to occur. Influence of these events on statistical characteristics of transmitted data was not studied.

The goal of this paper is primary statistical characteristics research of data processing and transmission time in IES. Particularly, the server availability time (SAT) was studied within this work. This time consists of Client-to-Server connection time, Server data processing time, and Server answer transfer time. The values of service time were taken for research, that were received while studying of cloud computing performance [3]. The research of IES performance and reliability is actual because of user needs to have continuous access to them without data loss.

Research results. Statistical data used in paper was received as follows [3]. The Client and Server applications were developed. The Server application was installed on the remote Cloud Computing Azure computer. The Client application was gathering parameters of connection time and remote server data processing time every minute during 24 hours. In common, 1440 values were received, the initial selection based on these values was established, with 0% of lost data. Due to the fact that data transmission in Client-Server system is inevitably followed by data loss for different reasons, the research of data missing influence on primary statistical processing results was made.

The following approach was used for the analysis of missed data influence on the results of statistical characteristics determination. 0, 1, 3, 6, 9, 12, 15% of data was removed from initial selection. Data for removing was selected randomly. The conclusions about lost data influence on the quantity of received statistical results of selections properties were made based on comparisons between received statistical characteristics for each selection.

According to recommendations from paper [5] the following methods of data processing with losing were used:

1. The "Gluing" method - primary statistical characteristics of studied selection were determined without consideration of data missing. With such approach the selection size decreases, and therefore freedom degrees number also decreases when checking of the formulated statistical hypotheses.

2. The "Zero" method - primary statistical characteristics of studied selection were determined with consideration of filling admissions with zeros. In such case freedom degrees number does not decrease, but estimations shift of received characteristics appears comparing to initial selection.

3. The "Average" method - primary statistical characteristics of studied selection were determined with consideration of filling admissions with average values. In such case it is also possible to get estimations shift of received characteristics comparing to initial selection.

4. The "Random" method primary statistical characteristics of studied selection were determined with consideration of filling admissions with quasi-normal numbers distributed by normal law when average value and mean square deviation coincides with with similar characteristics of an initial data set with admissions.

5. In addition to these method, the following "Distributions" approach was used. Missed data was filled with data distributed by the same law as for initial selection with admissions.

The Statgraphics V.15 system was used for determination of distribution law type. The results of this procedure are shown in table 1.

Table 1

*Change of selection distribution law depending of missed data filling approach**

Missed data filling approach	Part of missed data (%)						
	0	1	3	6	9	12	15
Gluing	T1	T1	T1	T1	T1	T1	T1
Average	T1	T1	T1	T1	T2	T3	T3
Random	T1	T1	T1	T1	T1	T2	T2
Distributions	T1	T1	T1	T1	T1	T1	T1

* T1 – distribution of maximum value, T2 – loglogistic distribution,
T3 – Laplace's distribution.

Received distribution laws look as follows.

T1 - function of the density distribution of maximum value

$$f(x) = \frac{1}{\beta}\exp\left\{-\frac{x-\alpha}{\beta} - \exp\left(\frac{x-a}{\beta}\right)\right\} \quad (1)$$

Distribution parameters α and β are related to mathematical expectation m and to dispersion with s^2 equations:

$$m = \alpha + \beta\Gamma^{-2}, \qquad s^2 = \frac{\beta^2\pi^2}{6} \quad (2)$$

T2 – loglogistic distribution with density:

$$f(x) = \frac{1}{\sigma x}\frac{\exp(z)}{[1+\exp(z)]^2}, \quad (3)$$

where:

$$z = \frac{\ln(x) - \mu}{\sigma}.$$

In such case parameters of position μ and scale σ are related to mathematical expectation m and to dispersion with s^2 conditions:

$$\mu = \exp(\mu)\Gamma(1+\sigma)\Gamma(1-\sigma) \quad (4)$$

$$s^2 = \exp(2\mu)\left[\Gamma(1+2\sigma)\Gamma(1-2\sigma) - \Gamma^2(1+\sigma)\Gamma^2(1-\sigma)\right] \quad (5)$$

where $\Gamma(.)$ – Euler's gamma-function.

T3 – Laplace's distribution with density:

$$f(x) = \frac{\lambda}{2}e^{-\lambda|x=\mu|}. \quad (6)$$

Parameter of this distribution μ is equal to mathematical expectation, dispersion σ^2 is related to parameter λ with equation:

$$\sigma^2 = \frac{2}{\lambda^2}. \quad (7)$$

The conclusion from table 1 is if data loss is not over 6% all studied methods of filling missed data save distribution law of initial model selection. If data loss is between 9% and 15% only "Gluing" and "Distributions" methods save distribution law of initial model selection. When restoring missing data it is important not only to save distribution law of initial selection, but also to achieve the fact that initial model selection (selection without missing) would not be significantly (in statistical sense) different from selections with artificially filled missed data.

Non-parametric criteria of selection coincidence hypothesis checking were used to check the influence of missed data filling approach on received distributions parameters. The checking results and the list of used criteria are presented in table 2. Model selection was used as standard sample in all studied cases. It was established that symbol (*) means absence of difference between compared selections by all criteria. The list of used criteria and their symbols are provided in the note to table 2. All computations were performed using AtteStat program.

Table 2

*Influence of missed timing data filling approach on saving of distribution law appearance and parameters**

Missed data filling approach	Part of missed data (%)					
	1	3	6	9	12	15
Zero	*	K	V, M, VW, K	V, M, VW, Z, K	V, M, VW, S, Z, K	V, M, VW, S, Z, K
Gluing	*	*	*	*	*	*
Average	*	*	*	Z, A	VW, Z, A	VW, Z, A *
Random	*	*	*	*	*A, S	*S, K
Distributions	*	*	*	*	*	*

* V – Vilkokson's Criterion, M – Mann-Whitney's Criterion, VW – Van der Waerden's Criterion, S – Kritery Sevidzha, Z – Ziegel-Tyyuki's Criterion, A – Ansari-Bredli's Criterion, K – Klotts's Criterion.

The presence of symbol in table cell points to criterion that discarded hypothesis about selections coincidence. So, only "Gluing" and "Distributions" methods of restoring missing data are recommended for practical usage.

Statistical row properties were studied for the reasonable choice of forecasting method. At the first stage fractal row properties were defined, specifically Hurst index H.

In paper [4] following procedure of Hurst index computation is proposed. Relation between Hurst index H and data row statistical characteristics is defined with the formula:

$$R / S = \left(\frac{\pi}{2} N \right)^H , \quad (8)$$

where:

S – mean square deviation of observations timing row,

N – number of observations.

The Hurst index H is defined with the formula:

$$H = \frac{\lg(R / S)}{\lg(\pi N / 2)} . \quad (9)$$

Formula (10) is used to compute mean square deviation of observations row S:

$$S = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \overline{X} \right)} . \quad (10)$$

where: – arithmetic mean of studied observations timing row for N timing periods.

The range of accumulated deviation is the most important element of Hurst index computation formula. In general, it is evaluated as follows:

$R = \max_{1 \le u \le N} Z_u - \min_{1 \le u \le N}$ (11)

where: Z_u is accumulated deviation of row elements from mean value:

$$Z_u = \sum_{i=1}^{u} \left(x_i - \overline{X} \right) . \quad (12)$$

In paper [4] it is recommended to change left part of formula (8) using following correction if the observations number N is less than 250:

$R / S_T = R / S \times 0.998752 + 1.051037$ (13)

Table 3

Hurst index evaluation for SAT

All selection		Last hundred	
Arithmetic mean X	2007,678	Arithmetic mean X	1958,158
Standard deviation S	255,464	Standard deviation S	225,5333
Range R	57494,667	Range R	3698,297
Normalized Range R/S	225,060	Normalized Range R/S	16,39801
Hurst index Ht	0,709	Hurst index Ht	0,570525

As it is shown in paper [4], in case of Hurst index (Ht) is between 0,326 and 0,674 the model of changing row values is Wiener process. The physical analogue of this process is Brownian motion around average value of observations row.

Evaluation results of index H that is determined on all data array ("all selection") and on last hundred observations ("last hundred") are given in table 3.

The obtained data was served as explanation for the choice of sliding average as forecasting method.

The quality of forecasting results was evaluated with retrospective forecast method for six last values using formula:

$$\varepsilon = \sum_{i=m-k}^{m} \frac{|\hat{x}_i - x_i|}{x_i} \qquad (14)$$

where: ε – value of average relative forecast error,

m – capacity of data array for which average relative forecast error was determined

x_i – actual value,

$\hat{x}_i$ – calculated value.

Table 4

Quality estimation of SAT forecast

Selection type	Sliding average order		
	3	5	7
«All selection»	0,072	0,049	0,065
«Last hundred»	0,028	0,035	0,041

The chart of studied process is shown on fig. 1. The x-axis represents number of experiments (server calls), the y-axis represents server availability time in milliseconds.

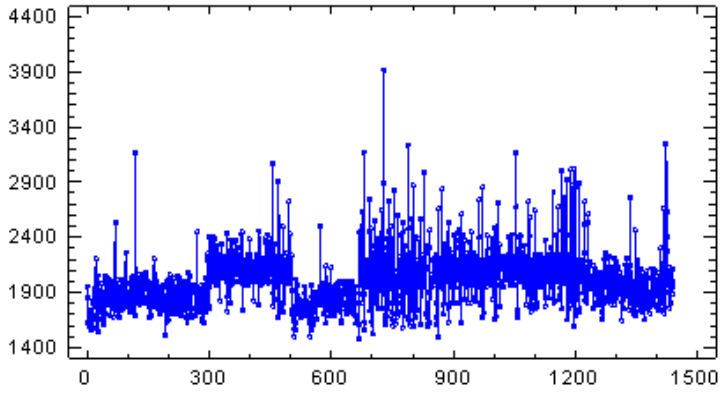

Fig. 1. Server availability time

Autocorrelation function of SAT value changing process is shown on fig. 2. According to this figure it is obvious that statistical dependence between sequential SAT observations is very weak that indirectly confirms received value of H.

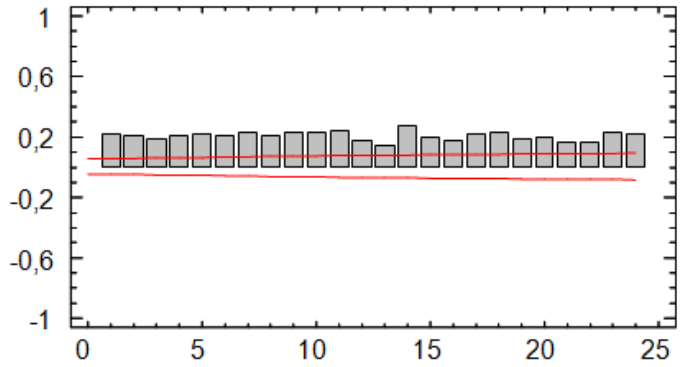

Fig. 2. Autocorrelation function of random SAT process

The chart of autocorrelation function of SAT process first differences was received to check its stationarity (fig. 3). According to this chart SAT process can be considered stationary.

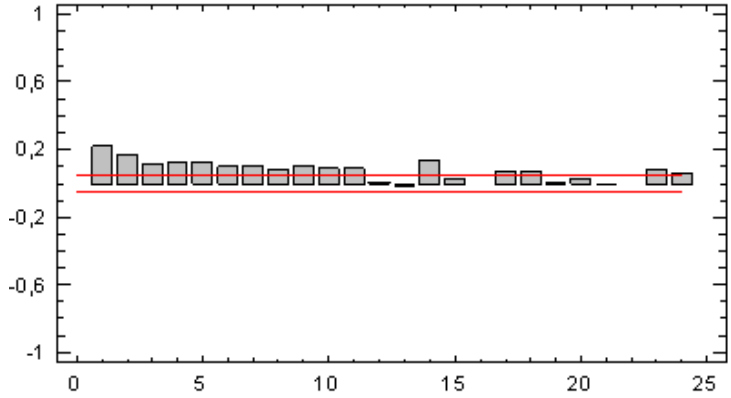

Fig. 3. Autocorrelation function of first differences of SAT random process

Conclusion. It was established that if data loss is not over 6% distribution law of initial model selection saves. If data loss is between 9% and 15% only "Gluing" and "Distributions" methods save distribution law of initial model selection.

Fractal properties of SAT row were studied using Hurst index in this paper. The method of SAT row values forecasting was chosen on this basis. The quality of forecasting results was evaluated using retrospective forecast method for six last values. Results showed that sliding average method with step equal to three should be used for SAT forecasting.

Література

1. Концепция Cloud Computing [Текст] / В. И. Лымарь, М. И. Макарова, Е. В. Зубкова, Т. В. Кортева // Конкурентоспособность территорий: материалы XV Всерос. экон. форума молодых ученых с междунар. участием в рамках III Евразийского экономического форума молодежи «Диалог цивилизаций – «ПУТЬ НАВСТРЕЧУ» (Екатеринбург, 17–18 мая 2012 г.). В 9 ч. Ч. 8. Направления : 11. Исследования менеджмента, маркетинга и логистики ; 19. Информационные процессы инновационного бизнеса / [отв. за вып. М. В. Федоров, Э. В. Пешина]. – Екатеринбург : Изд-во Урал. гос. экон. ун-та, 2012. – С. 255–256.

2. Benchmarking Dependability of a System Web Application. [Text] / Yuhui Chen, Alexander Romanovsky, Anatoliy Gorbenko, Vyacheslav Kharchenko. // 14th IEEE Int. Conf. on Engineering of Complex Computer Systems. – ICECCS'2009: conference proceedings. – Potsdam (Germany), 2009. – P. 146–153.

3. Рубан, В. И. Экспериментальное исследование производительности Cloud Computing [Текст] / В. И. Рубан, А. В. Горбенко // Системы управления, навигации и связи. – 2012. – Вип. 4 (1). – С. 189–191.

4. Найман, Э. Расчёт показателя Херста с целью выявления трендовости (персистентности) финансовых рынков и макроэкономических индикаторов [Текст] / Э. Найман // Економіст. – 2009. – № 10. – С. 25–29.
5. Литтл, Р. Дж. А. Статистический анализ данных с пропусками [Текст] / Р. Дж. А. Литтл, Д. Б. Рубин ; пер. с англ. А. М. Никифорова. – М. : Финансы и статистика, 1991. – 334 с.
6. Тардаскина, Т. Н., Электронная коммерция [Текст] / Т. Н. Тардаскина, Е. Н. Стрельчук, Ю. В. Терешко. – Одесса : ОНАС им. А. С. Попова, 2011. – 128 с.
7. Балабанов, И. Т. Электронная коммерция [Текст] / И. Т. Балабанов. – Санкт–Петербург : Питер, 2001. – 336 с.
8. Прикладной статистический анализ данных. Теория. Компьютерная обработка. Области применения [Текст] / С. В. Алексахин, А. В. Балдин, А. Б. Николаев, В. Ю. Строганов. – М. : Приор, 2002. – 688 с.
9. Gorbenko, A. Using Inherent Service Redundancy and Diversity to Ensure Web Services Dependability. In Methods, Models and Tools for Fault Tolerance [Text] / A. Gorbenko, V. Kharchenko, A. Romanovsky. – LNCS 5454. – 2009. – P. 324–341
10. Боровиков, В. STATISTICA. Искусство анализа данных на компьютере [Текст] / В. Боровиков. – 2–е изд. – Санкт–Петербург : Питер. – 2003. – 688 с. + CD : ил. – (Для профессионалов).
Стаття надійшла до редакції 27.01.2015

References

1. Lymar, V. I., Makarov, M. I., Zubkova, Ye. V., & Korteva, T. (2012). *Concept Cloud Computing*.
2. Chen, Y., Romanovsky, A., Gorbenko, A. V., & Kharchenko, V. S. (2009). *Benchmarking Dependability of a System Web Application*.
3. Ruban, V. I., & Gorbenko, A. V. (2012). Experimental study of the performance Cloud Computing. *Control systems, navigation and communication, 4,* 189–191.
4. Nayman, E. (2009). Hurst exponent calculation to identify trended (persistence) of financial markets and macroeconomic indicators. *Ekonomist*, *10*, 25–29.
5. Little, RJ. A., Rubin, D. B., & Nikiforov, A. M. (1991). *Statistical analysis of gaps data*. Moscow: Finance and Statistics.
6. Tardaskina, T. N, Strelchuk, Ye. N., & Tereshko, Yu. V. (2011). *Electronic commerce*. Odessa.
7. Balabanov, I. T. (2001). *Electronic commerce*. Sankt-Peterburg.
8. Aleksakhina, S. V., Baldina, A. V., Nikolaeva, A. B., & Stroganova, V. Yu. (2002). *Applied statistical analysis. Theory. Computer processing. Areas of application*.
9. Gorbenko, A., Kharchenko, V., & Romanovsky, A. (2009). *Using Inherent Service Redundancy and Diversity to Ensure Web Services Dependability. In Methods, Models and Tools for Fault Tolerance. LNCS 5454.*
10. Borovikov, V. (2003). *Statistica. Art data analysis on the computer*. Sankt-Peterburg: Piter.
Received 27.01.2015