DEVELOPMENT OF THE ALGORITHM FOR CLOUD SERVICE PERFORMANCE PREDICTION

Babkin O.V.[1], Varlamov A.A.[2] (Russian Federation), Gorshunov R.A.[3] (Slovakia), Dos E.V.[4] (REPUBLIC OF BELARUS), Kropachev A.V.[5], Zuev D.O.[6] (United States of America) Email: Varlamov551@scientifictext.ru

[1]Babkin Oleg Vyacheslavovich - Strategy Consultant,
IBM;
[2]Varlamov Aleksandr Aleksandrovich – Technical Director
SHARXDC LLC,
MOSCOW;
[3]Gorshunov Roman Aleksandrovich - Solution Architect,
AT&T, BRATISLAVA, SLOVAKIA;
[4]Dos Evgenii Vladimirovich - Lead DevOps Architect,
EPAM, MINSK, REPUBLIC OF BELARUS;
[5]Kropachev Artemii Vasilyevich - Principal Architect,
LI9 TECHNOLOGY SOLUTIONS, NORTH CAROLINA;
[6]Zuev Denis Olegovich - Independent Consultant,
NEW JERSEY,
UNITED STATES OF AMERICA

Abstract: *main stages of data center service performance prediction were discussed, specifically data monitoring and gathering, calculation and prediction of key indexes and performance index prediction. It was proposed to build data center service performance prediction algorithm based on analysis of service transactions index, service resource occupancy index and service performance index. Prediction of the indexes is based on chaotic time series analysis that was used to estimate service transactions index time series trend, radar chart method to calculate service resource occupancy index value and weighted average method to calculate service performance index. For performance prediction is proposed to use fuzzy judgment matrix with service transactions index and service resource occupancy index as input values. Next stages include definition of fuzzy closeness degree and estimation the best matching value of the indexes at the predicted moment by similarity matching algorithm. It was taken into consideration that service transactions index is usually represented by nonlinear time series and thus the index time series parameters have to be predicted by chaos theory and for the calculation of this index can be used estimation procedure of Lyapunov exponent value. Radar chart demonstrates service resource occupancy index estimation of shared storage, mobile storage, memory, computational capability and network bandwidth. It was noticed that for calculation of service performance index values' dataset it is necessary to find nearness degree of service transactions index and service resource occupancy index it is proposed to estimate first membership degree as a parameter of membership function. Therefore, prediction technique was based on the fuzzy nearness category that use input values of service transactions index and service resource occupancy index dynamic changes which have to be considered as a real time process.*

Keywords: *data center, service transactions index, service resource occupancy index, service performance index. fuzzy judgment matrix, Lyapunov exponent, radar chart.*

РАЗРАБОТКА АЛГОРИТМА ПРОГНОЗИРОВАНИЯ ПРОИЗВОДИТЕЛЬНОСТИ ОБЛАЧНЫХ СЕРВИСОВ

Бабкин О.В.[1], Варламов А.А.[2] (Российская Федерация), Горшунов Р.А.[3] (Словакия), Дос Е.В.[4] (Республика Беларусь), Кропачев А.В.[5], Зуев Д.О.[6] (Соединенные Штаты Америки)

[1]Бабкин Олег Вячеславович - стратегический консультант,
IBM;
[2]Варламов Александр Александрович – технический директор,
ООО "Шаркс Датацентр",
г. Москва;
[3]Горшунов Роман Александрович - архитектор решений,
AT&T, г. Братислава, Словакия;
[4]Дос Евгений Владимирович - ведущий DevOps архитектор,
EPAM, г. Минск, Республика Беларусь;
[5]Кропачев Артемий Васильевич - главный ИТ архитектор,
Li9 Technology Solutions, г. Северная Каролина;
[6]Зуев Денис Олегович – независимый международный эксперт, г. Нью Джерси,

Соединенные Штаты Америки

Аннотация: *проведен анализ основных этапов прогнозирования эффективности обслуживания центров обработки данных, в частности мониторинга и сбора данных, расчета и прогнозирования ключевых аспектов, и прогнозирование коэффициента производительности. Было предложено построить алгоритм прогнозирования эффективности обслуживания центра обработки данных на основе анализа коэффициента транзакции, коэффициента использования машинных ресурсов и коэффициента производительности сервиса. Прогнозирование коэффициентов основано на анализе временных рядов, который использовался для оценки временных рядов коэффициента транзакций, метода радар-диаграммы для расчета значения коэффициента использования машинных ресурсов и средневзвешенного метода оценки для расчета коэффициента производительности сервиса. Для прогнозирования производительности предлагается использовать матрицу нечетких суждений с коэффициентом транзакций и коэффициентом занятости ресурса службы в качестве входных значений. Следующие этапы включают определение степени нечеткой близости и алгоритма соответствия подобия. Было указано, что коэффициент служебных операций обычно представлен нелинейными временными рядами, и, следовательно, параметры временного ряда коэффициента должны быть предсказаны теорией хаоса, а значит для расчета этого коэффициента может быть использована процедура расчета экспоненты Ляпунова. Радарная диаграмма демонстрирует оценку коэффициента использования машинных ресурсов для общего хранилища данных, мобильных хранилищ, памяти, вычислительных возможностей и пропускной способности сети. Метод прогнозирования основывался на категории нечетких приближений, которые используют входные значения коэффициента транзакций и динамические изменения коэффициента использования машинных ресурсов, которые должны рассматриваться в рамках процесса, который анализируется в режиме реального времени.*

Ключевые слова: *центр обработки данных, коэффициента транзакции, коэффициента использования машинных ресурсов, коэффициента производительности сервиса, матрица нечетких суждений, экспонента Ляпунова, радарная диаграмма.*

1. Introduction

Nowadays requirements to cloud platform data center services performance have significantly grown. Thereby it's important to develop effective and multipurpose algorithm of estimation of key aspects that refers to the stability of the network infrastructure work. Efficient strategy should be based on analysis of the whole dataset of gathered information of monitoring platform and to be able to predict indexes of data center performance at any moment of time with high accuracy.

Assigned task could be solved by mathematical methods of chaotic analysis and fuzzy logic but adaption of them stands nontrivial task. In order to identify the main aspects of the problem, an analysis of recent studies and publications was done. It was analyzed aspects of data center service performance that are mentioned to be key ones [1, 2], specifically service transactions index, service resource occupancy index and service performance index. To solve problem of prediction of those were studied works devoted to chaotic analysis [2-4], radar chart method [1, 5] and weighted average method [6]. Also within the bounds this study were analyzed fundamental mathematical materials [7-9] related to fuzzy logic in order to use it at cloud platform data center services performance analysis and prediction. Systematic analysis shows possibility to develop effective technique based on monitoring and gathering of information for estimation and accurate prediction of key aspects that refers to the data center service performance.

2. Service indexes prediction and calculation procedure

Data center service performance prediction procedure [1, 2] usually includes following stages (Figure 1):

- data center indicators' data monitoring and gathering;
- calculation and prediction of key indexes of data center infrastructure work;
- data center service performance index prediction.

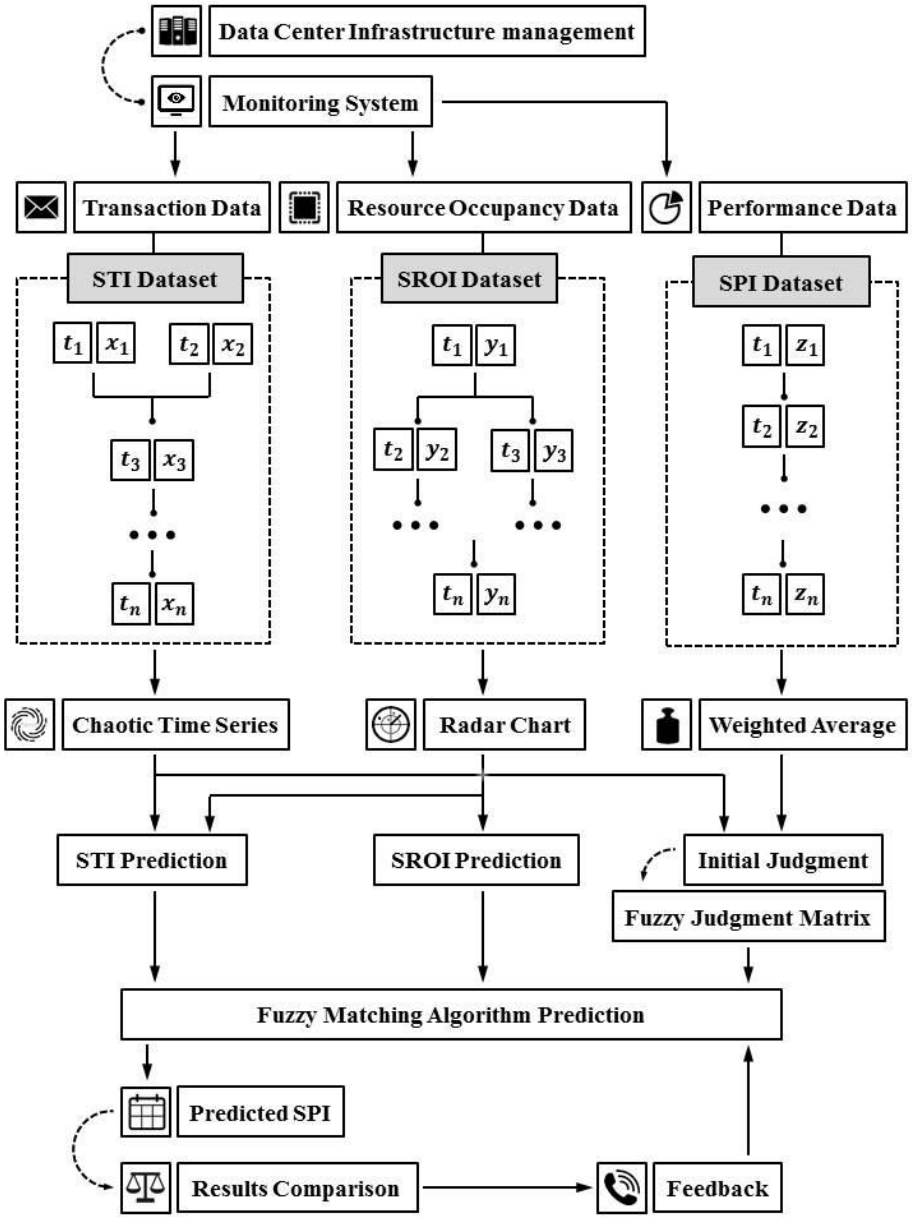

Fig. 1. Data center service performance prediction scheme

Indicators' data contains recorded by virtual machine (VM) monitoring plugins information about transaction logs, utilization level of physical resources (shared storage, computational capability, network bandwidth, etc.) and response time of each monitoring spot which refers to the system performance. Analysis of gathered data allows defining key indexes of data center infrastructure work efficiency (Figure 1):

- service transactions index (STI);
- service resource occupancy index (SROI);
- service performance index (SPI).

STI value refers to the number of data center's transactions that require service to process. This index indicates service's loads at each moment and should be recorded as a time series x_i: $[x_1, x_2, \dots, x_n]$ which corresponds to the time chart t_i: $[t_1, t_2, \dots, t_n]$. As it shown at Figure 1 usually STI time series have to be modeled as nonlinear sequence. Thus STI trends can be predicted by nonlinear time series forecasting methods based on artificial neural networks (ANNs) platform. In other hand SROI value refers to data center servers' physical resources allocated to the service at each moment and SPI value refers to the data center service's response time at each moment. It should be mentioned that SPI directly reflects service performance while this index is the comprehensive result of the key monitoring points' analysis.

Prediction of key indexes procedure includes a variety of methods or algorithms that can be used. Within the bounds of this study it is proposed to use (Figure 1):

- chaotic time series analysis to estimate STI time series trend [2-4];

- radar chart method to calculate SROI value [1, 5];
- weighted average method to calculate SPI value [6].

Performance prediction of the modern data center service work process should be based on fuzzy judgment matrix. It uses STI and SROI values (Figure 1) according to the definition of fuzzy closeness degree, and estimate the best matching value of STI and SROI at the predicted moment by similarity matching algorithm. Thereby, SPI which corresponds to the obtained value represents prediction result data center service performance which is be compared with value that was obtained experimentally.

3. STI time series prediction algorithm

It was mentioned above that modern data center service based on cloud paradigm is usually has to be represented by nonlinear system. It could be added that STI time series are would be nonlinear time series on cloud platform. Thereby, STI time series parameters have to be predicted by chaos theory.

For reconstruction of STI time series should be used delay embedding theorem (Takens' theorem). Let us suppose that time series x_i: $[x_1, x_2, \dots, x_n]$ which corresponds to the time t_i: $[t_1, t_2, \dots, t_n]$ have power system dimension d and thus the system must be considered form d-dimensional state vector $x_i(t)$ that evolves according to an unknown but continuous and deterministic dynamic. For simplified form of Takens' theorem [1, 7-9] adapted to the time series prediction it could be said that observable result F_x is a smooth function of x_i dataset. $F_x(t)$ have to be supplemented by observations made within certain time lag τ multiplied by values $k = 1 \dots m$:

$$F_x(t,k): [F_x(t), F_x(t-\tau), F_x(t-2\tau), \dots F_x(t-k\cdot\tau), \dots F_x(t-m\cdot\tau)]. \qquad (1)$$

It's obvious that for increasing number of lags m will lead motion in the lagged space to become more predictable, and for $m \to \infty$ system will tend to become deterministic and equivalent to original state space. Takens' theorem [1] demonstrates that lagged vectors become deterministic at a finite dimension of $m \geq 2d + 1$. Thereby STI time series prediction's target function $F_x(t_i)$ of m-dimensional phase space with N phase points could be defined in every point in space phase as:

$$\begin{cases} F_x(t_i) = [x(t_i), x(t_i+\tau), x(t_i+2\cdot\tau), \dots, x(t_i+\tau\cdot(m-1))] \\ m \geq 2d+1; i = 1,2 \dots N; N = n-\tau\cdot(m-1) \end{cases}. \qquad (2)$$

It has to be noticed that $m \geq 2d + 1$ is not a necessary but sufficient condition of determination of system dynamic.

STI time series' calculation could be done not only by qualitative analysis but also by quantitative algorithm. It's based on calculating some chaotic quantities. Most effective way is to estimate Lyapunov exponent value. Lyapunov exponent of a dynamical system is a quantity that characterizes the rate of separation of infinitesimally close trajectories []. Two trajectories in phase space with initial separation δZ_0 diverge as:

$$|\delta Z(t)| \approx e^{\lambda t}|\delta Z_0| => \lambda = \lim_{t\to\infty}\left(\lim_{\delta Z_0\to 0}\left(\frac{ln(|\delta Z(t)|/|\delta Z_0|)}{t}\right)\right). \qquad (3)$$

where λ is the Lyapunov exponent and $\delta Z_0 \to 0$ criteria ensures the validity of the linear approximation at each moment of time. Thereby, biggest obtained value of Lyapunov exponent (MLE: maximal Lyapunov exponent) is a parameter which could be used for estimation whether a system is a chaotic one ($\lambda > 0$) or not ($\lambda \leq 0$). It should be noticed that initial separation vector usually contain some component in the direction associated with the MLE, and thus effect of the other exponents can be neglected.

For analysis of STI time series proposed mathematical model could be slightly simplified. Let us suppose that we need to predict x_{n+k} for dataset of x_i: $[x_1, x_2, \dots, x_n]$. We have to choose a point X_i for prediction center in a phase space of the system. X_i is defined as:

$$X_i: [x_n - \tau\cdot(m-1), x_{n+1} - \tau\cdot(m-1), \dots, x_{n+k} - \tau\cdot(m-1)] \qquad (4)$$

The next step is to define nearest point $X_j \epsilon \{X_1, X_2, \dots, X_{i-1}\}$. While distance between X_i and X_j is d, then d could be defined as $d = |X_i - X_j|$. Therefore MLE could be estimated by comparison of $|X_i - X_{i+1}|$ and $|X_j - X_{j+1}|$ differences.

$$|X_i - X_{i+1}| = e^{\lambda_1}|X_j - X_{j+1}| => \lambda_1 = ln\left(\frac{|X_i - X_{i+1}|}{|X_j - X_{j+1}|}\right). \qquad (5)$$

While λ_1 is obtained MLE for time series x_i: $[x_1, x_2, \dots, x_n]$ it predicts x_{n+1}. To predict x_{n+k} it should be done k-step prediction.

4. SROI and SPI value prediction algorithm

As it was mentioned above SROI value refers to data center service physical resources utilization level. Physical resources are distributed on different servers and VMs so estimation of SROI value is nontrivial task. Most efficient method of SROI analysis is development of radar char, a graphical method of displaying multivariate data more than two quantitative variables [1, 5].

The radar chart area R(t) for SROI evaluation and prediction can be gotten as follows:

$$R(t) = \frac{sin(2\pi/3)}{2} \cdot \sum_{i,j}[y_i \times y_j] \qquad (6)$$

At Figure 2 is demonstrated radar chart that can be used for SROI analysis for five resources.

- shared storage;
- mobile storage;
- memory (RAM and cash-memory);
- computational capability (CPU);
- network bandwidth.

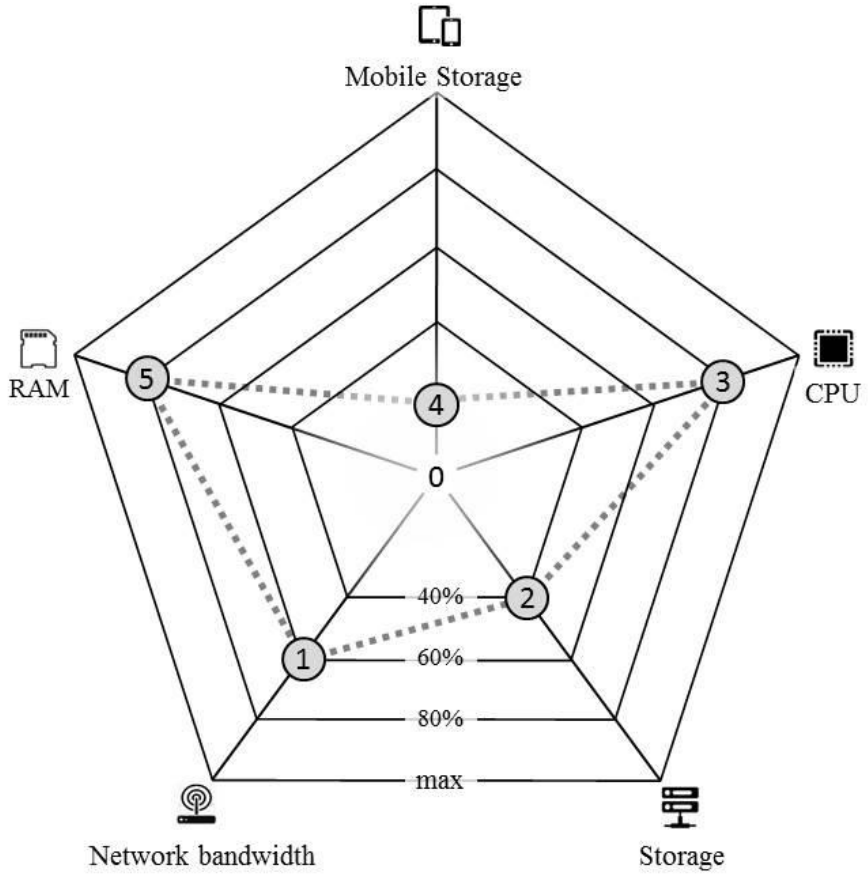

Fig. 2. Data center service resources occupancy radar chart

There are several methods of effective SPI prediction but all of them based on estimation of monitoring points response time dataset T_i: $[T_1 \ldots T_n]$. Thereby, basic equation for SPI at any moment of time could be defined as:

$$P_i = \frac{1}{n}\sum_{i=1}^{n} T_i \qquad (7)$$

Prediction technique is based on the fuzzy nearness category that use input values of STI and SROI values dynamic changes (as a real time process). A fuzzy matching algorithm estimates the nearness degree of STI and SROI of the prediction time. The nearness level of STI values' dataset (X_n) and SROI values' dataset (Y_n) for the n time moments to be predicted should be estimated to X_i and Y_i that a closest ones to X_n and Y_n, respectively. To calculate SPI values' dataset as a set of predicted performance values at predicted n time moments X_i and Y_i values have to be used (Figure 3).

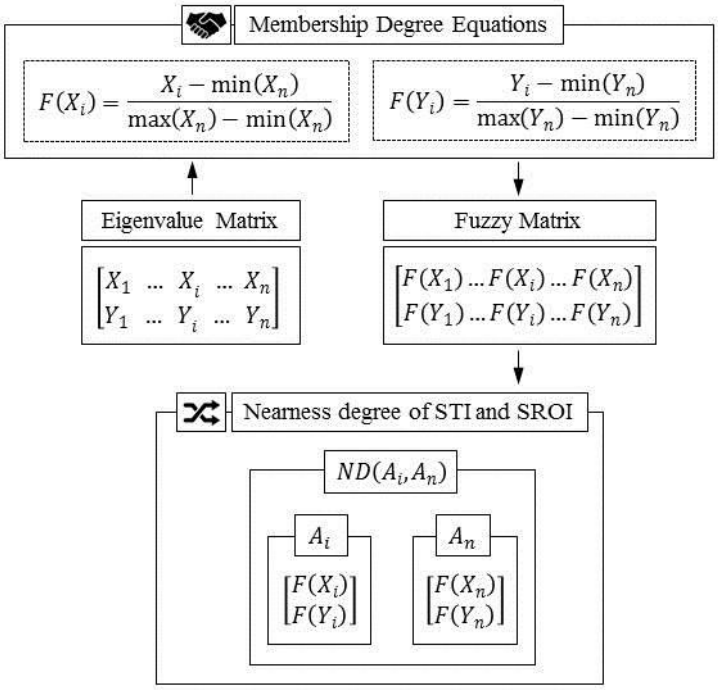

Fig. 3. Evaluation of nearness degree of STI and SROI

Estimation of SPI values' dataset is impossible without getting nearness degree of STI and SROI which is based on calculating of membership degree (Figure 3). Membership degree is a value of membership function $F \in [0; 100\%]$ that refers to the correlations between element and some characteristic [1, 9]. Calculation of membership function is based on eigenvalue matrix of X_i and Y_i datasets:

$$\begin{cases} F(X_i) = \frac{X_i - \min(X_n)}{\max(X_n) - \min(X_n)} \\ F(Y_i) = \frac{Y_i - \min(Y_n)}{\max(Y_n) - \min(Y_n)} \end{cases}. \qquad (8)$$

It allows obtaining fuzzy matrix of $F(X_i)$ and $F(Y_i)$ datasets (Figure 3). Together with $F(X_n)$ and $F(Y_n)$ datasets it should be used to obtain nearness degree:

$$ND(A_i, A_n) = \Big(\big(F(X_i) \wedge F(X_N)\big) \vee \big(F(X_i) \wedge F(X_N)\big)\Big) \wedge \Big(\big((1 - F(X_i)) \wedge (1 - F(X_N))\big) \vee \big((1 - F(Y_i)) \wedge (1 - F(Y_N))\big)\Big), \qquad (9)$$

where A_i represents the matrix in moment i (estimated moment of time), and A_n represents the matrix in moment n (predicted moment of time).

5. Conclusions

Main stages of data center service performance prediction, such as indicators' data monitoring and gathering, calculation and prediction of key indexes of data center infrastructure work and performance index prediction were discussed. It was proposed to build data center service performance prediction algorithm based on analysis of service transactions index, service resource occupancy index and service performance index. Prediction of the indexes was based on chaotic time series analysis that was used to estimate service transactions index time series trend, radar chart method to calculate service resource occupancy index value and weighted average method to calculate service performance index.

For performance prediction was proposed to use fuzzy judgment matrix with service transactions index and service resource occupancy index as input values. Next stages include definition of fuzzy closeness degree and estimation the best matching value of the indexes at the predicted moment by similarity matching algorithm. It was taken into consideration that service transactions index is usually represented by nonlinear time series. It was noticed that the index time series parameters have to be predicted by chaos theory and thereby for the calculation of this index was used estimation procedure of Lyapunov exponent value. Radar chart that was used for service resource occupancy index estimation was built for five main resources of cloud platform service: shared storage, mobile storage, memory, computational capability and network bandwidth. For calculation of

service performance index values' dataset it is necessary to find nearness degree of service transactions index and service resource occupancy index it was proposed to estimate first membership degree. Therefore, prediction technique was based on the fuzzy nearness category that use input values of service transactions index and service resource occupancy index dynamic changes which was considered as a real time process.

References / Список литературы

1. *Wu C. & Guo J.*, 2015. Software Monitoring in Data Centers. Handbook on Data Centers. 1209-1253.
2. *Newcombe L.*, 2014. Data Center Financial Analysis, ROI and TCO. Data Center Handbook. 103-137.
3. *Román-Flores H. & Ayala V.*, 2018. Chaos on Set-Valued Dynamics and Control Sets. Chaos Theory.
4. *Tang R., Fong S. & Dey N.*, 2018. Metaheuristics and Chaos Theory. Chaos Theory.
5. *Hongliang L., Anxin L., Bin Z., Tiefu Z. & Xin Z.*, 2008. A Fuzzy Comprehensive Evaluation Method of Maintenance Quality Based on Improved Radar Chart, 2008 ISECS International Colloquium on Computing, Communication, Control, and Management.
6. *Shi J., Liu Y. & Zhou W.*, 2011. The domain decomposition method based on weighted average. 2011 IEEE International Conference on Computer Science and Automation Engineering.
7. *Harris J.*, 2000. An introduction to fuzzy logic applications. Dordrecht: Kluwer Academic.
8. *Anderson M.*, 2015. Fuzzy logic. Parkdale, OR: Black Opal Books.
9. *Dimitrov V. & Korotkich V.*, 2011. Fuzzy logic: A framework for the new millennium. Heidelberg: Physica.