UDC 004.056.5

DOI: 10.15587/1729-4061.2018.131471

NETWORK TRAFFIC FORECASTING BASED ON THE CANONICAL EXPANSION OF A RANDOM PROCESS

V. Savchenko
Doctor of Technical Sciences, Senior Researcher*
E-mail: savitan@ukr.net

O. Matsko
PhD, Associate Professor**

O. Vorobiov
Doctor of Technical Sciences, Professor**

Y. Kizyak
PhD
Scientific research laboratory**

L. Kriuchkova
Doctor of Technical Sciences, Associate Professor*

Y. Tikhonov
PhD, Associate professor*

A. Kotenko
PhD*

*Department of Information and cyber security
Information Security Institute
State University of Telecommunications
Solomianska str., 7, Kyiv, Ukraine, 03110
**Institute operational support and logistics
Ivan Chernyakhovsky National
Defense University of Ukraine
Povitroflotsky ave., 28, Kyiv, Ukraine, 03049

Досліджується проблема прогнозування мережевого трафіку у мережах TCP/IP на основі статистичних даних спостережень. Визначено, що існуючі протоколи (SNMP, RMON) не передбачають довготривалого прогнозування, яке є необхідним для модернізації мережі. Регресійні методи (AR, ARMA, ARIMA, SARIMA), які лежать в основі протоколів, використовують лише послідовність значень прогнозованого ряду, що унеможливлює довгострокове прогнозування. Зроблено висновок про відсутність універсального ефективного методу прогнозування часових послідовностей, якими описується трафік комп'ютерної мережі.

Розроблено модель прогнозу мережевого трафіку з урахуванням особливостей накопичення статистичних даних: наявності апріорних траєкторій, апостеріорного характеру прогнозування, скінченності дисперсії. Застосовується апарат канонічного розкладання випадкового процесу з урахуванням неоднорідності трафіку. Розроблено математичний апарат вирішення задачі екстраполяції реалізації, одержано вирази для оцінки похибки екстраполяції, вирази для відтворення апостеріорного випадкового процесу на основі моделювання. Враховуються похибки апріорних вимірювань, що дозволяє застосовувати зазначену модель у мережах при мінімумі діагностичних даних. Забезпечується точне визначення параметрів випадкового процесу у точках контролю та мінімум середнього квадрата похибки наближення у проміжках між цими точками.

Застосування запропонованої методики на основі канонічного подання випадкових процесів забезпечує вирішення задачі довгострокового прогнозування мережевого трафіку. Порівняльний аналіз методів прогнозування свідчить про наближення методу канонічного розкладання випадкового процесу до інтелектуальних методів прогнозування

Ключові слова: мережевий трафік, прогнозуючий контроль, випадковий процес, канонічне розкладання випадкового процесу

1. Introduction

A widespread emergence of TCP/IP networks requires effective methods to study them. Forecasting of characteristics of network traffic is one of the important tasks, both for operational management of network parameters in real time, as well as for the construction and optimization of telecommunication networks. Network providers should be able to forecast a number of requests (data volume) for future to optimize resources, manage loading adaptively, and regulate network parameters proactively. Thus, they will be able to improve service quality.

In addition, DDOS attacks, the main tool of which is a multiple increase in traffic, were and remain one of the most serious threats to computer networks [1]. This forces providers to monitor traffic very closely and forecast an amount of data that is being transmitted. In this case, parameters of a forecast can be: a number of requests over a certain time, a speed of requests arrival, a number of requests from a specific source (network) or a number of requests to a specific destination, time between requests, etc.

Statistic data on network traffic is a time series where specific values of time over a specified period (day, week, month, year) correspond to specific values of a traffic parameter.

At present, SNMP protocol *(Simple Network Management Protocol)* encloses computer network monitoring technologies, as it is the most common multi-agent network management protocol used to obtain information from network devices about their status, performance, and other characteristics. The protocol emerged in order to control routers at the Internet, which is part of TCP/IP stack. An extension of SNMP is RMON protocol *(Remote Network MONitoring)*, which controls a network remotely. In contrast to SNMP, which collects information on devices with a corresponding agent installed only, RMON collects information on traffic between network devices.

The mentioned protocols complete the task of forecasting over short intervals (seconds-minutes) successfully, but do not provide long-term forecasting, which is necessary for modernization of a network. Thus, the development of effective methods for a long-term forecasting of network traffic based on statistical data remains an actual scientific and applied problem.

2. Literature review and problem statement

Paper [2] outlines a general classification of methods for forecasting network traffic. The authors share a whole range of approaches to statistical, software and cognitive methods, methods of data-mining and machine learning. Despite a potentially promising prospect of artificial intelligence methods, statistical methods remain the most usable and effective to forecast traffic.

The forecasting theory of statistical time series includes hundreds of different methods based on a rather limited number of approaches and models. Thus, paper [3] considers the method of *forecasting by the last value* (extrapolation of the zero order). According to the method, we should take $X(t_0)$ as a forecasted value $\hat{X}(t_0+\theta)$

$$\hat{X}(t_0+\theta) = X(t_0). \qquad (1)$$

In this case, a forecasted value does not depend on forecasted time interval θ; only one point presents the prehistory, it is the last value of $X(t_0)$. The method does not take into account probabilistic characteristics at all. A forecasting error

$$e(t_0+\theta) = X(t_0+\theta) - \hat{X}(t_0+\theta)$$

in this case takes the form

$$e(t_0+\theta) = X(t_0+\theta) - X(t_0),$$

and its mean square at $m_x=0$,

$$\overline{e}^2(\theta) = M\left\{\left[\dot{X}(t+\theta) - \dot{X}(t)\right]^2\right\} = \\ = \sigma_x^2 - 2R_x(\theta) + \sigma_x^2 = 2\left[\sigma_x^2 - R(\theta)\right].$$

The simplicity of the method facilitated its wide application. However, this method is unsuitable for effective long-term forecasting of constantly changing network traffic parameters.

Work [4] considers another approach, it is *forecasting by mathematical expectation.* Here we use a mathematical expectation of m_x process as a future forecasted value $\hat{X}(t_0+\theta)$: $\hat{X}(t_0+\theta) = m_x$. A forecast error for this case takes the form $e(\theta)=X(t+\theta) - m_x$ and represents a deviation of the process from an average one at moment $t_0+\theta$. The mean error does not depend on forecast time and is equal to the process variance

$$\overline{e}^2 = M\left\{\left[X(t_0+\theta) - m_x\right]^2\right\} = \sigma_x^2.$$

The "last value" method is better for low forecasting time values θ. However, after θ2 at $\overline{e}^2(\theta) = \sigma_x^2$, the method of "mathematical expectation" yields greater accuracy. Finally, at θ→∞, the square of a forecasting error is half less of a last countdown on the average.

In forecasting of trends, we usually use regression and auto regression (AR – Auto Regressive). For a linear model, a forecast takes the form

$$y_{t+T} = b_0 + b_t(t+T),$$

where *T* is the depth of a forecast. People use polynomial, quadratic, exponential, logarithmic, hyperbolic and indicative dependences in addition to a linear one in practice. Work [5] proposes a more interesting application of the linear autoregressive model that we can apply in a quasilinear environment. The authors propose a multi-step recursive forecasting method that combines the traditional linear approach to direct strategy and an auto regression procedure that applies to each step of a forecast. Based on the results of the study, the authors conclude that the proposed method provides lower values of variance, although its value increases depending on lead time of forecasting. Thus, we can consider the approach as boundedly acceptable for a long-term forecasting of network flows.

A base of a large group of forecasting methods, in particular, methods described in work [6], is the application of *moving averages* (MA – *Moving Averages*). Here, we calculate each forecasted value from formula:

$$\tilde{x}_{t+1} = \frac{1}{N}\sum_{j=1}^{N} x_{t-j+1}, \qquad (2)$$

where *N* is the number of previous periods included to the moving average; x_j is the actual value at time *j*; $\tilde{x}_j$ is the forecasted value at moment *j*.

The moving average (2), in contrast to a simple average for the entire sample, contains data on a trend of data change, but it does not differ much from forecasting by mathematical expectation in accuracy.

Recently, mixed (ARMA) models have been used more often [7] for the description of time series. Their application domain is not limited to stationary processes. In addition, it is possible to reduce series with a specific homogeneous non-stationarity to stationary ones and to describe them by a modified form of ARMA model (Box-Jenkins model [8]). In this case, an auto regression model describes a stationary process, where a value of an indicator is a linear combination of a limited number of its previous values and a random component. For example, we can represent the process AR(*p*) in the following way

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + \\ + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} +,\ldots,+ b_q\varepsilon_{t-q}. \qquad (3)$$

In ARMA-model (3), only AR-part determines the stationarity of ARMA(p,q)-process. Therefore, disadvantages of the approach are the same as for the AR-process.

Paper [9] considers auto regressive *integrated* moving average models (ARIMA). Their base is the assumption that the data generation process is linear and they describe a stationary process. Attributes of such a process are: an order of autoregression (*p*), a necessary order of integration (*d*) and an order of the moving average in a model (*q*). Work [10] shows that such models are close to intelligent methods of forecasting in accuracy for short forecasting periods (1–5 steps). However, the use of current statistical values of statistics for a forecast only and ignoring the history of a process makes them ineffective in a long run.

Further improvement of considered ARMA and ARIMA models is SARIMA model [11], which takes into account a seasonal trend of SARIMA data $(p, d, q)\times(P,D,Q)S$, where p is a non-seasonal AR order, D is a seasonal coefficient, Q is a seasonal MA order, S is a seasonal period. Season consideration is one of the most effective ways of ensuring sufficient accuracy at significant time intervals. However, in this case, a forecasting process is more complicated due to the need to ensure the accuracy of determining a seasonal component.

In addition [10] to the methods that relate to artificial intelligence, *the method of group arguments consideration* (MGAC) [12] deserves an attention. The method solves all optimization questions by examining options among models based on training and verification data sequences. It does not use information on laws of information distribution. Several "partial" descriptions replace the full description of an object $\varphi=f_1(x_1, x_2, x_3,..., x_n)$

$$y_1=f_1(x_1, x_2),\ y_2=f_1(x_1, x_3),\ ...,\ y_m=f_1(x_{n-1}, x_n), \quad (4)$$

$$z_1=f_1(y_1, y_2),\ z_2=f_1(y_1, y_3),\ ...,\ z_p=f_1(y_{m-1}, y_m), \quad (5)$$

where $m=C_n^2,\ \ p=C_m^2.$

There are heuristic criteria used for boundary self-selection of the best variants in (4) and (5) successively: according to the coefficient of correlation, according to the criterion of variety of arguments, according to the criterion of conditionality of matrices and, most importantly, according to the criterion of the minimum standard error (ASE minimum). We should use ACE minimum successively several times for the selection of optimized variables. All other variables are auxiliary and have a purpose of reducing a volume of calculations. MGAC provides such selection of coefficients of partial equations, at which it is possible to achieve ACE minimum in the space of these coefficients. In this case, we determine coefficients of a complete equation by elimination of intermediate variables from "partial" equations. However, if a learning sequence is limited or short, some arguments and intermediate variables are harmful, which requires introduction of additional procedures for selection of equations and reduces accuracy of a forecast.

The analysis of the considered forecasting methods makes it possible to conclude that most of them do not take into account the dependence of a forecasted value on other already known additional factors. Specifically, auto regressive methods (AR, ARMA, ARIMA, SARIMA) do not use any auxiliary data except for a sequence of values of forecasted series. Such a forecast model is sufficiently effective for short intervals; however, it is ineffective for long-term forecasting.

Neural networks algorithms and MGAC, as static models, cannot take into account the inertia of an object, but they work well on nonlinearity and multidimensionality. Another disadvantage of the algorithms is that we cannot apply them beyond limits of statistical data, that is, in order to implement them, in any case, a certain amount of preliminary information about a system behavior is required. Therefore, after consideration of a nature of a potential use of the method of forecasting network traffic as a process with a significant aftereffect, we can consider a use of the mentioned methods as problematic.

Upon an analysis, we can conclude that there is not any universal effective method for long-term forecasting of time sequences, which describes traffic of a computer network, at present. In one case, forecasting accuracy is suitable for short periods (seconds, minutes) only. In the other, complexity of calculations does not make possible to implement models on most network devices in real time. In addition, auto regressive methods do not use historical statistics that accumulate in the process of network operation.

Therefore, it is necessary to solve the problem of development of a methodology for long-term forecasting of time series taking into account the features specified above.

3. The aim and objectives of the study

The aim of present study is to substantiate the method for forecasting network traffic based on an approach that takes into account both a history of development of a process in general and an individual behavior of a trajectory of time series.

We set the following tasks to achieve the objective:

– development of a model to forecast network traffic;

– formation of parameters of an analytical description of an *a priori* random process and the methodology of forecasting *a posteriori* random process;

– modeling and evaluation of effectiveness of the results obtained.

4. Development of a model for forecasting network traffic

The study of volumes of network traffic indicates the existence of daily, weekly, and annual cycles (Fig. 1), which makes it possible to use this information along with operational statistic data on traffic.

A feature of long-term forecasting of network traffic is a possibility of taking into account previous trends based on existing statistics on a load of servers and network equipment and existence of "individual" trajectories of a system's performance.

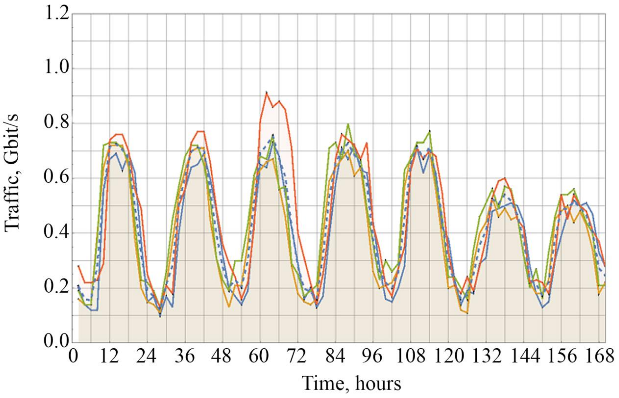


Fig. 1. Typical dynamics of network traffic within a week

To specify a task for each of elementary events, which we understand as averaged observational data during a certain period (a week) $\omega\in\Omega(t_k,\theta)$, it is necessary to put some quantitative sign of a state $e_\omega(s), t_k\le s\le t_k+\theta$ at unambiguous correspondence. $e_\omega(s)$ will be implementation of some random process of a change in parameters of traffic in time. Thus, the forecasting model will be the random process $E(t)$ on the interval $t_k\le s\le t_k+\theta$.

Upon determining a certain region of acceptability E_0, a network administration task will be to avoid intersection of any implementation of a random process $E(t)$ and limits of a region of acceptability E_0 a in time interval $t_k \le s \le t_k + \theta$:

$$e_\omega(s) \in E_0,\ \ t_k \le s \le t_k + \theta. \qquad (6)$$

Paper [13] shows that there is a single-valued connection between a random process $E(t)$ and a random function $T(e)$. That makes it possible to use a probabilistic description of $T(e)$ function instead of an initial $E(t)$ random process for models of traffic forecasting. Thus, for the task of determination of a traffic behavior, one can consider a certain scalar random function $T_1(e)$, which determines random time before the first intersection of $E(t)$ random process of an arbitrary surface given by the value of restrictions of e vector. The principle of determination of time T for a specific trajectory on an example of implementation of a scalar process $E(t)$ with a region of acceptability $E_0=[a, b]$ (Fig. 2).

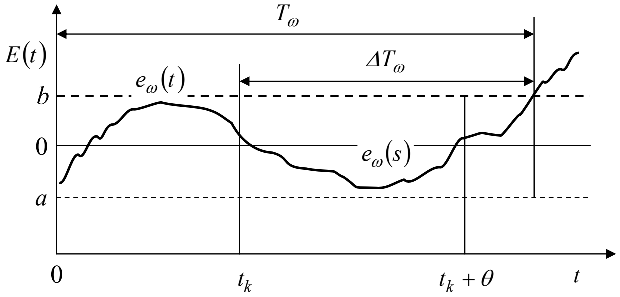


Fig. 2. Forecasting the traffic

The presented statement of the problem is somewhat idealized, since it does not include measurement errors associated with various fluctuations, but we can use it for further solution of the problem of individual forecasting. Real processes of a network traffic behavior are extremely varied and complex enough, since they depend on a large number of factors and conditions. However, we can identify some common features for this type of random processes: unsteadiness of a process, presence of a significant aftereffect, multidimensionality, and existence of a connection between separate implementations of a random process.

To determine appropriate time for the beginning of implementation of measures aimed at regulation of traffic (neutralization of a DDOS attack), it is necessary to solve the problem of individual forecasting of its time trajectory. In this case, we can use data on observations on parameters of traffic that form a vector of parameters $X=(X_1, X_2,..., X_H)$ as data inputs. There is a region of acceptability S_0 defined for values of this vector. Execution of the condition $X \in S_0$ in it means that a network is operating in its normal mode. An occasional process $X(t)$ emerges at a change of a value of a parameter vector in time, it describes evolution of network parameters in time. Let us also define that $X(t)$ process is determined statistically on the time axis at $t \ge t_1$, and the moment t_1 corresponds to the moment of the beginning of observations at a given time interval (day, week, month, etc.).

Let us assume that for a specific observation (specific trajectory) ω, a control moment of observation $t_k \ge t_1$ is defined, then information on ω parameter is given by implementation interval $x_\omega(t) \in S_0$, $t_1 \le t \le t_k$ of a random process $X(t)$, which can be obtained from the data on traffic monitoring.

In this case, we can formulate the problem of individual forecasting a traffic performance as the problem of determination of a posterior (conditional) law of time distribution of the output of a process $X(t)$ outside a region of acceptability S_0 relative to implementation $x_\omega(t)$, that is, as a task on determining probability

$$P^{ps}(s) = P\left\{X(s) \in S_0 / x_\omega(t)\right\},\ \ t_1 \le t \le t_k, s \ge t_k. \qquad (7)$$

Expression (7) is a conditional probability that a specific trajectory ω is guaranteed to fall into the domain $s>t_k$, if, by the moment of time t_k, inclusively, the implementation $x_\omega(t)$, $t_1 \le t \le t_k$ determined its state. That is, we solve the problem of individual forecasting of a trajectory of network traffic.

The mentioned features determine approaches to solving a problem on individual forecasting. The main of them are: analytical solution and statistical modeling. Application of classical methods of extrapolation of random processes involves a point estimation of a future value of implementation at some point in time $s>t_k$ only. This does not solve the complete problem of forecasting of a behavior of a trajectory (7).

Methods, which include the following steps, are more suitable to solve such a problem:

1. Obtaining an analytical description of the investigated random process $X(t)$ suitable for further modeling.
2. Development of an algorithm that takes into account a value of a specific implementation $x_\omega(t)$, $t_1 \le t \le t_k$ and construction of *a posteriori* random process $X^{ps}(s)$, $s>t_k$ on its basis.
3. Modeling of a set of implementations of *a posterior* process $X^{ps}(s)$, $s>t_k$ and obtaining of characteristics of such a process on its basis.

To obtain a universal and convenient method to solve the problem, it is expedient to use an approach based on the representation of the examined process by the following model

$$X(t) = m(t) + \sum_\nu V_\nu \phi_\nu(t), \qquad (8)$$

where $m(t)$ is the mathematical expectation of a process; $\varphi_v(t)$ are the non-random (coordinate) time functions; V_v are random, non-correlated coefficients ($M\left[V_\nu\right]=0,\ M\left[V_\nu, V_\mu\right]=0,$ $v \ne \mu$).

This representation, proposed in work [13], makes it possible to apply it to any real random process, which is useful in view of a large variety of behavioral variants of random network traffic processes. Papers [14, 15] propose a description of random processes based on some canonical representation. Expression (8) describes its general meaning. In this case, it is possible to highlight main properties of such representation, which make the approach acceptable to describe a network traffic performance.

5. Formation of parameters of a scalar *a priori* random process

Let us assume that we have a scalar random process $X(t)$ given by a random sequence $X(t_i)=X(i)$, $i=\overline{1,I}$ of eigenvalues in discrete observation series t_i. At the same time, its canonical representation takes the form

$$X(i) = m(i) + \sum_{v=1}^{i} V_v \phi_v(i),\ \ i=\overline{1,I}, \qquad (9)$$

where V_v is a random coefficient with the following characteristics $M[V_\nu]=0,\ M[V_\nu,V_\mu]=0,\ v\neq\mu;\ \phi_\nu(i)$ is a non-random coordinate function, $\varphi_v(v)=1$, $\varphi_v(i)=0$ at $v>i$.

We can write expressions for variance and correlation functions as

$$D(i)=\sum_{v=1}^{i}D_v\phi_v^2(i),\quad i=\overline{1,I}; \qquad (10)$$

$$D(i,j)=\sum_{v=1}^{\inf(i,j)}D_v\phi_v(i)\phi_v(j),\quad i,j=\overline{1,I}. \qquad (11)$$

Work [15] states that determination of canonical representation's elements by the following recurrence relations provide optimal properties of a canonical representation:

$$V_1=\overset{\circ}{X}(1),\quad V_i=\overset{\circ}{X}(i)-\sum_{v=1}^{i-1}V_v\phi_v(i),\quad i=\overline{2,I}; \qquad (12)$$

$$D_1=D(1),\quad D_i=D(i)-\sum_{v=1}^{i-1}D_v\phi_v^2(i),\quad i=\overline{2,I}; \qquad (13)$$

$$\phi_v(i)=\frac{1}{D_v}M\left[V_v\overset{\circ}{X}(i)\right],\quad v=\overline{1,I},\quad i=\overline{v,I}. \qquad (14)$$

Thus, expression (9) provides a solution to the problem on modeling a scalar process with dependent components. As follows from expression (14), the only restriction imposed on the investigated random process is finiteness of variance of a random process. This is performed usually for real processes of study into network traffic, which ensures universality of the method of canonical representation of random processes. At the same time, expression (9) defines a random process at points of observation t_i precisely and provides a minimum of standard approximation error in the intervals between these points. Consequently, the canonical representation of random processes can provide a solution to the problem of forecasting a network traffic (7).

6. Forecasting of a network traffic based on the canonical representation of a linear scalar *a posteriori* process

Let us determine an *a priori* random process $X(t)$ as a canonical representation (9) on a discrete number of points $t_i,\ i=\overline{1,I}.$ to solve the problem on obtaining an analytical description of *a posteriori* random process at the apparatus of canonical representation. Let us assume that at some moments $t_\mu,\ \mu=\overline{1,k},\ k<I$, which coincide with moments t_i in determination of a process for $i\leq k$, as a result of control we obtain values of $x(\mu),\ \mu=\overline{1,k}$ of a segment of specific implementation of the process $X(t)$. Than we need to obtain a description of *a posteriori* analytical process $X^{ps}(t)$, which arises from *a priori* $X(t)$ on the basis of observational data.

We can obtain the description as follows. Let us consider a value $x(1)$ of the implementation of the process obtained as a result of control. Representation (9) is correct for this value, which takes the form at $\mu=1$

$$x(1)=m(1)+v_1. \qquad (15)$$

Thus, expression (15) specifies a value v_1 of a random coefficient V_1, which corresponds to the result of the first observation.

Coefficients $V_i,\ i=\overline{1,I}$ of the canonical representation (9) are not independent although they are uncorrelated to each other. Therefore, specification of V_1 value leads generally to a change in distribution density of the remaining coefficients $V_i,\ i=\overline{2,I}.$ To obtain an analytical description of *a posteriori* process that is valid within the correlation theory, let us assume that coefficients of the initial representation (9) are pairwise independent:

$$f_2(v_i,v_j)=f_1(v_i)f_1(v_j),\quad i=\overline{1,I-1},\quad j=\overline{i+1,I}. \qquad (16)$$

We use this assumption and substitute the value V_1 obtained in (15) in formula (9). We obtain an expression for *a posteriori* random process, which passes through the point $x(1)$ at moment $i=1$:

$$X^{(1)}(i)=m(i)+\left(x(1)-m(1)\right)\phi_1(i)+\sum_{v=2}^{i}V_v\phi_v(i),\quad i=\overline{1,I}. \qquad (17)$$

A mathematical expectation of such a process will take the form

$$m^{(1)}(i)=m(i)+\left(x(1)-m(1)\right)\phi_1(i),\quad i=\overline{1,I}. \qquad (18)$$

We can take an advantage of this and find

$$X^{(1)}(i)=m^{(1)}(i)+\sum_{v=2}^{i}V_v\phi_v(i),\quad i=\overline{1,I}. \qquad (19)$$

If we obtain the following value $x(2)$ of the same process implementation under a regular control, then representation (19) is correct for this value, where

$$x(2)=m^{(1)}(2)+v_2.$$

We repeat the operation for the case $\mu=1$ and obtain

$$m^{(2)}(i)=m^{(1)}(i)+\left(x(2)-m^{(1)}(2)\right)\phi_2(i),\quad i=\overline{1,I}; \qquad (20)$$

$$X^{(2)}(i)=m^{(2)}(i)+\sum_{v=3}^{i}V_v\phi_v(i),\quad i=\overline{1,I}. \qquad (21)$$

Taking into account the recurrence of expressions for mathematical expectation of *a posteriori* random process, a general expression for an arbitrary number $k<I$ of control points takes the form:

$$m^{(0)}(i)=m(i),\quad i=\overline{1,I},$$

$$m^{(k)}(i)=m^{(k-1)}(i)+\left(x(k)-m^{(k-1)}(i)\right)\phi_k(i),\quad i=\overline{1,I}; \qquad (22)$$

$$X^{(k)}(i)=m^{(k)}(i)+\sum_{v=k+1}^{i}V_v\phi_v(i),\quad i=\overline{1,I}. \qquad (23)$$

Thus, expressions (22) and (23) completely describe a linear *a posteriori* process, where expression (22) is a mathematical expectation of this process at t_i points.

Expression (22), as a conditional mathematical expectation, is an extrapolation operator, it is optimal for the minimum criterion of a standard error in the class of linear operators. At the same time, a standard error of extrapolation will be

$$\sigma_k(i)=\sqrt{M\left[\left\{m^{(k)}(i)-X(i)\right\}^2\right]},\quad i=\overline{k+1,I}, \qquad (24)$$

where the expression $\left\{m^{(k)}(i)-X(i)\right\}$, $i=\overline{k+1,I}$ is an absolute error of extrapolation at k known values of control.

The operator of mathematical expectation $M[\bullet]$ in (24) means averaging over all possible realizations of an output random process (9). It is advisable to perform the averaging in two stages: initially over the ensemble of implementations of *a posteriori* process by a fixed known part of implementation, and then - over all possible realizations at $i\le k$:

$$\sigma_k^2(i)=M_k\left\{M_i\left\{m^{(k)}(i)-X(i)/x(\mu),\mu=\overline{1,k}\right\}\right\}=$$

$$=M_k\left\{\sum_{v=k+1}^{i}D_v\phi_v^2(i)\right\}=D^{(k)}(i),\ i=\overline{k+1,I}. \qquad (25)$$

Thus, a standard extrapolation error is equal to the variance of *a posteriori* process.

We compare expressions (10) and (25) and establish that $D^{(k)}(i)\le D(i)$, $k\le i\le I$, since there are no first k members from an expression (10) in (25). In this case, a degree of reduction of *a posteriori* variance serves as a measure of efficiency of the solution to the extrapolation problem relative to the *a priori* one.

Thus, expression (22) makes it possible to resolve the problem of extrapolation of implementation optimally, expression (25) – to evaluate an error of extrapolation, and expression (23) – to reproduce *a posteriori* random process in general based on modeling. The indicated linear analytical model of *a posteriori* random process based on the canonical representation can solve the problem of long-term forecasting of network traffic.

7. Modeling and discussion of results of application of the method for forecasting network traffic

We studied the efficiency of forecasting on the basis of an apparatus of the canonical expansion of time series of network traffic through modeling. We used the statistics of total traffic of an enterprise's local network node based on observations collected using WireShark 2.2.7 software package by DNS, FTP, FDDI, HTTP, ISQ, IRV6, IPH, IRC, MAPI, MOUNT, NECIBIOS, NFS, NNTP, POP, PPR, TCR, TELLNET and X25 protocols as the source data. The period of observation was a calendar week. We averaged the obtained statistical data averaged for two-hour intervals during a day (Fig. 1) for convenience of processing and visibility of the identified trends.

We applied the forecasting methodology (9) to (23) to the process. Fig. 1 shows its implementation. We took the points of time series, which correspond to a separate trajectory of observation 1 (Fig. 1 – a blue curve), as initial values of observations. We selected the mentioned curve as a control one, and the initial values of a time series – as the initial data of observations, which correspond to t=2, 4, 6, 8 hours of observations.

Fig. 3, *a* shows work of the forecasting algorithm at t=2 hours. As we can see, the knowledge of only one value of implementation makes it possible to recreate the process (the effect of a curve of mathematical expectation) in general. However, specific values of forecasted traffic are very different from the actual ones (a control trajectory). A physical content of the above is that if we know average parameters of network traffic and an entry point to a forecast, we can predict future performance of a system with sufficient accuracy. That is, a device "selects" a necessary trajectory of a system behavior itself depending on an entry point and an average trajectory. The accuracy of the result, as we indicated, lies within the variance of a random process.

An increase in the number of observations to t=4, 6, 8 hours (Fig. 3, *b*) increases reliability of further forecasting and we can speak about meeting condition (7) at t=8 hours.

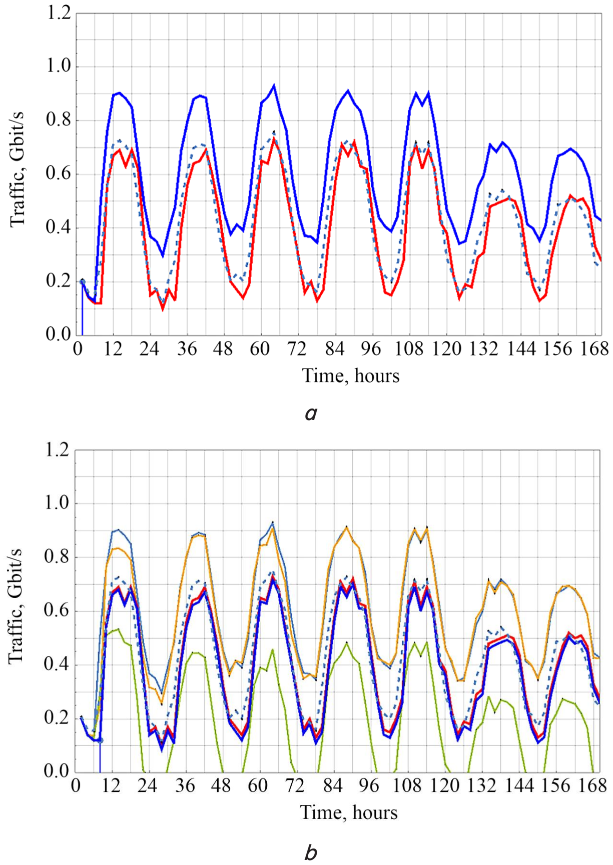
$$P^{ps}(s)=P\left\{X(s)\in S_0/x_\omega(t)\right\}\ge 0{,}95,\quad i=\overline{k+1,I}$$

at

$$0\le t\le 8, s\ge 8.$$

That is, probability of an error in selection of a correct trajectory depends on the number of observed initial data.

Interesting is the moment of studying the method on a subject of forecasting the anomalies of network traffic. The red trajectory has a characteristic "spike" for a period of 60...72 hours in Fig. 1. Application of the proposed method with the initial data of the specified trajectory according to data on the first 2 hours (Fig. 4, *a*) with a general background of the process (Fig. 1) reproduces an anomaly correctly generally, although with significant deviations from the control curve. An increase in the number of observations to t=8 hours leads to the minimum deviations from the trajectory (Fig. 4, *b*).

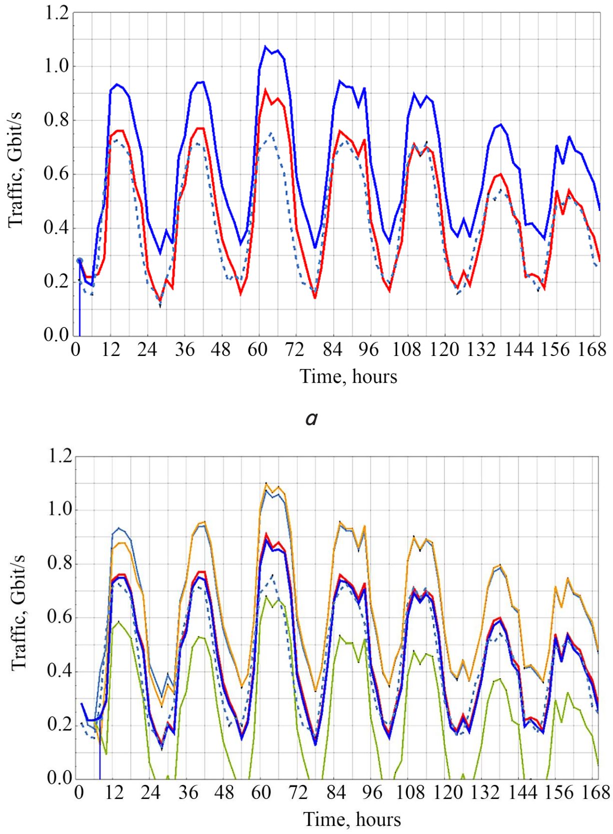


a

b

Fig. 3. Forecast of network traffic (— forecast; — comparable trajectory; --- an average value): *a* – t=2 h; *b* – t=8

At the same time, it is logical to assume that in this case, forecasting accuracy will depend too much on the behavior of a trajectory, which leads to abnormal traffic, as well as a frequency of observed anomalies. Therefore, this situation is unlikely to be repeated on any interval since it is not known in advance when the abnormal release will actually occur.



a

b

Fig. 4. Forecast of traffic anomalies (— forecast; — comparable trajectory; --- average value): *a* — *t*=2 h; *b* — *t*=8 hours

It is necessary to perform statistical consideration (accumulation) of abnormal behavior and implementation of the method based on short-term forecasting to implement possibility of forecasting of anomalies. As for AR and MA methods, in this case, only short-term intervals (seconds, minutes) will be effective, which is not very effective from a practical point of view.

We performed modeling with a use of the statistical test method to compare the accuracy of forecasting by the method of canonical decomposition of a random process with previously considered methods. We took an average absolute percentage error MAPE – Mean Average Percentage as the criterion for evaluation and calculated it in terms of expression

$$MAPE = \frac{100\,\%}{n}\sum_{i=1}^{n}\left|\frac{X_i - \vec{X}_i}{X_i}\right|,$$

where X_i is the current value; $\vec{X}_i$ is the predicted value of a variable at moment i; n is timeliness of a forecast (steps).

As expected, we observed the most significant deviations (up to 65 %) at forecasting of traffic by an average value. Errors are charges for simplicity and reliability of the method at this approach. In general, it is possible to use such a method for rough estimation of network parameters only.

For the study of the canonical decomposition method in comparison with other methods of forecasting, we used capabilities of Wolfram Mathematica 11.0 package. In particular, we developed an appropriate software tool for its implementation based on the initial data of Fig. 1 and methods (9) to (23). In addition, we used standard capabilities of the package to determine parameters of ARIMA (6, 1, 0) models with further determination of values of an absolute error of MAPE forecast. Auto regression models and models of moving averages MA, as well as an average value forecasting, also yield sufficiently large forecast errors of 35–50 %. That is explained by the non-stationary nature of a process and the presence of a seasonal component. We obtain somewhat higher accuracy (MAPE=20–35 %) when applying ARIMA models (with weak seasonality) and SARIMA (in the case of significant seasonal effects).

We developed 3 modules of a program, which differ in criteria for selection of appropriate models – "regularity", "minimal shift" and "absolute immunity to noise", to study the method of group consideration of arguments. This approach provides accuracy at the level of 12–15, 10–12 and 8–10 %, respectively, depending on selection criteria.

A comparative analysis (Table 1) of forecasting methods by MAPE indicator makes it possible to conclude that the method of canonical decomposition of a random process is close to intelligent methods, in particular to the method of group consideration of arguments.

This feature exists not only due to precise characteristics of the methods, but also due to the general paradigm. As with intelligent methods (artificial neural networks, the method of group consideration of arguments), there is an idea of recognition and "selection" of the most expedient trajectory used in the method of the canonical decomposition of a random process. At the same time, the possibility of application of the method at intervals not covered by statistics or under conditions of insufficient amount of output data, provides for certain advantages over intellectual methods.

Table 1

Comparative analysis of forecasting methods

Forecasting method	Average deviation from control implementation (MAPE), %
By an average value (mathematical expectation)	35...65
By a moving average (MA, ARMA, ARIMA,...)	20...50
Method of group consideration of arguments	8...15
Canonical implementation of a random process	6...12

Thus, the modeling results confirm the adequacy of the forecasting model of a process of change of network traffic, which is based on the idea of forecasting an individual trajectory of a random process. The model takes into account errors of *a priori* measurements, defines a random process at control points accurately and provides a minimum standard error of approximation in intervals between these points.

It is possible to use the proposed approach as a static model, while it can take into account inertia of an object, work out nonlinearities and multidimensionality. Another advantage of the proposed methodology is that we can apply it beyond the limits of statistical data based on the consideration of previous information on the behavior of a system only.

The disadvantage of the approach is the dependence of forecasting results on the amount of statistical information. This circumstance limits its use for forecasting network anomalies. It is increasingly "more difficult" to forecast a future abnormal value of traffic with significant amounts of method statistics, because calculations of variances and correlation functions are performed on a complete statistical sample. It will be increasingly difficult to recognize changes in seasonal trends having large volumes of statistical data of the method. In addition, using a complete sample requires significant computing resources, which is not always possible for stand-alone or small-sized networks without centralized management.

The way out of such situation can be:

1) parallel application of auto regressive methods or methods based on artificial intelligence for forecasting of anomalies;

2) artificial reduction of a sample size (a number of observations) by cutting off old sequences when the new ones appear.

The indicated aspects of the application of the methodology should become a subject for further research in the specified direction.

8. Conclusions

1. Long-term forecasting of network traffic parameters is one of the necessary elements of analysis and optimization of telecommunication networks with the aim of more rational use of resources, network load management and protection against DDOS attacks. The basis of such forecasting can be results of observations on the behavior of network traffic during a day, a week, a month, a year. Forecasting of traffic parameters by classical auto regressive models (AR, ARMA, etc.) is impossible due to the presence of a seasonal component in the process. Requirements for reliability of the accumulated statistics increase significantly at application of more sophisticated models (such as SARIMA). For processes with dependent components, we can resolve the problem of finding of unknown future values on the basis of the canonical representation of a random process model. The only limitation imposed on a random process under investigation is finiteness of its variance.

2. An analytical description of an *a priori* random process based on its canonical representation determines a process at the control points exactly and provides a minimum standard error of approximation in the intervals between these points. The method of forecasting *a posteriori* random process uses an extrapolation operator, which is optimal for the criterion of minimum standard error. Its base is a conditional mathematical expectation determined by the assumption of a pairwise independence of random coefficients of the initial representation of an *a priori* process.

3. The modeling results confirm the overall effectiveness of the approach to forecasting traffic parameters based on the canonical decomposition of a random process. At the same time, due to the accuracy of forecasting, the proposed method is close to methods of artificial intelligence, although it requires much less computing power. In order to implement a possibility of forecasting the anomalies of network traffic, we need to apply statistical consideration (accumulation) of abnormal performance with further involvement of short-term forecasting methods. In the proposed general model of the forecasting of time series, errors of *a priori* measurements affect a value of the mathematical expectation of *a posteriori* process only. Therefore, we can define an error of an individual forecast arising from measurement errors of measurements of values of controlled implementation as a difference between mathematical expectations of real and ideal *a posteriori* process.

Implementation of the mentioned approach by providers and telecommunication companies will make planning of distribution of network resources more flexible. It will improve the quality of customer service and make it possible to take measures to counter cyber-attacks. The direction of further research in this area may be a wide range of issues for improvement of the method to ensure possibility of forecasting on an interval outside the available statistics, under conditions of high noise of source data or its partial lack.

References

1. Khalimonenko A., Kupreev O., Ilganaev K. DDoS attacks in Q4 2017 // Securelist-2018. URL: https://securelist.com/ddos-attacks-in-q4-2017/83729
2. Prasad K. M., Reddy A. R. M., Rao K. V. DoS and DDoS Attacks: Defense, Detection and Traceback Mechanisms – a Survey // Global Journal of Computer Science and Technology: E Network, Web & Security. 2014. Vol. 14, Issue 7. P. 14–32.
3. Kandananond K. A Comparison of Various Forecasting Methods for Autocorrelated Time Series // International Journal of Engineering Business Management. 2012. Vol. 4. P. 4. doi: 10.5772/51088
4. Montgomery D., Jennings C. L., Kulahci M. Introduction to Time Series Analysis and Forecasting. 2nd ed. John Wiley & Sons, 2015. 671 p.
5. Taieb S. B., Hyndman R. Boosting Multi-Step Autoregressive Forecasts // Proceedings of the 31st International Conference on Machine Learning, PMLR. 2014. Vol. 32, Issue 1. P. 109–117.
6. Dias G. F., Kapetanios G. Estimation and Forecasting in Vector Autoregressive Moving Average Models for Rich Datasets // CREATES Research Paper No. 2014-37. 2014. 102 p.
7. Neusser K. Autoregressive Moving-Average Models // Springer Texts in Business and Economics. 2016. P. 25–44. doi: 10.1007/978-3-319-32862-1_2
8. Time Series Analysis: Forecasting and Control / Box G., Jenkins G. M. et. al. 5th ed. Wiley, 2015. 712 p.
9. Salamanca Céspedes J. E., Rodríguez Y. G., López Sarmiento D. A. Development of an Univariate Method for Predicting Traffic Behaviour in Wireless Networks Through Statistical Models // International Journal of Engineering and Technology (IJET). 2015. Vol. 7, Issue 1. P. 27–36.

10. Elwasify A. I. A Combined Model between Artificial Neural Networks and ARIMA Models // International Journal of Recent Research in Commerce Economics and Management (IJRRCEM). 2015. Vol. 2, Issue 2. P. 134–140.
11. Hidayatulloh I., Bustoni S. A. SARIMA-Egarch Model to Reduce Heteroscedasticity Effects in Network Traffic Forecasting // Journal of Theoretical and Applied Information Technology. 2017. Vol. 95, Issue 3. P. 554–560.
12. Confidence matching in group decision-making / Bang D., Aitchison L., Moran R., Herce Castanon S., Rafiee B., Mahmoodi A. et. al. // Nature Human Behaviour. 2017. Vol. 1, Issue 6. P. 0117. doi: 10.1038/s41562-017-0117
13. Atamanyuk I., Kondratenko Y. P., Sirenko N. N. Forecasting Economic Indices of Agricultural Enterprises Based on Vector Polynomial Canonical Expansion of Random Sequences // Proceedings volume of 5th International Workshop on Information Technologies in Economic Research (ITER) in ICTERI. Kyiv, 2016. P. 458–468.
14. Pugachev V. S. Probability Theory and Mathematical Statistics for Engineers. URL: https://legyrinez.firebaseapp.com/aa122/probability-theory-and-mathematical-statistics-for-engineers-by-v-s-pugachev-b00jez1dwq.pdf
15. Atamanyuk I. P., Kondratenko Yu. P. Information technology of polynomial forecast control of trouble-free operation of technical systems // System Research and Information Technologies. 2013. Issue 1. P. 43–52.

UDC 519.711; 681.5; 621.382

DOI: 10.15587/1729-4061.2018.132052

DEVELOPMENT OF METHODS, MODELS, AND MEANS FOR THE AUTHOR ATTRIBUTION OF A TEXT

I. Khomytska
Assistant
Department of Applied Linguistics *

V. Teslyuk
Doctor of Technical Sciences, Professor
Department of Automated Control Systems*
E-mail: vasyl.m.teslyuk@lpnu.ua

A. Holovatyy
PhD, Associate Professor
Department of Information Technologies
Ukrainian National Forestry University
Henerala Chuprynky str., 103,
Lviv, Ukraine, 79057

O. Morushko
PhD, Associate Professor
Department of Social Communication and Information Activity*

*Lviv Polytechnic National University
Bandery str., 12, Lviv, Ukraine, 79013

Рівень точності авторської атрибуції текста не є достатньо високий на лексичному та синтаксичному рівнях мови, бо ці рівні не є строго організованими системами. У даному дослідженні авторська атрибуція текста грунтується на диференціації фоностатистичних структур стилів.

Розроблено систему диференціації фоностатистичних структур стилів, яка відрізняється від існуючих вибраним рівнем мови – фонологічним. На цьому рівні мови можна отримати результати з більшою точністю. Окрім того, побудована система грунтується на модульному принципі, що дає змогу швидко модифікувати розроблений програмний продукт.

Розроблено методи та моделі, які грунтуються на теорії математичної статистики і дають змогу підвищити точність диференціації фоностатистичних структур стилів. Побудовано метод комплексного аналізу фоностатистичних структур стилів, багатофакторний метод визначення ступенів дії факторів стилю, підстилю та авторської манери викладу. Побудовано статистичну модель стилевої диференціації за методом ранжування та статистичну модель визначення загальної стилевої маркованості досліджуваного текста. Розроблено програмну систему диференціації текстів.

Критерієм диференціації текстів є середні частоти груп приголосних фонем. В процесі реалізації системи використана мова програмування java, що забезпечує платформо-незалежність програмного продукту.

Наведено результати застосування розроблених методів, моделей та програмних засобів, які підтверджують, що авторська атрибуція текста на фонологічному рівні є ефективнішою.

Розроблені методи, моделі та засоби авторської атрибуції текста можна використати при встановленні відсотку творчого внеску кожного із співавторів наукових праць

Ключові слова: середні частоти груп приголосних фонем, стилева, підстилева та авторська диференціація текстів, програмна система, метод, фонема, фонологічний рівень

1. Introduction

Modern information technologies (IT) are widely used in various fields of science and technology. One of such areas is applied linguistics [1, 2] where IT has been applied to the author's attribution by using content analysis [3], for the attribution of texts in legal proceedings [4, 5], and for a linguistic analysis of the text commercial content [6]. IT is employed in the semantic analysis of Ukrainian texts [7] and for carrying out scientific research related to programs