COMPUTATIONAL PROBLEMS OF ELECTRICAL ENGINEERING

Vol. 3, No. 2, 2013

MATHEMATICAL MODELING OF DYNAMICS OF WEBSITE QUALITY CHARACTERISTICS

Natalia Pasichnyk, Mykola Dyvak, Roman Pasichnyk

Ternopil National Economic University, Ternopil, Ukraine,

natalia.pasichnyk@gmail.com, mdy@tneu.edu.ua, roman.pasichnyk@gmail.com

Abstract. The article offers a set of dynamic models of predicting the website attendance. The numerical experiments were conducted to identify these models. It allowed the authors to structure an adequate model to predict overall and target attendance, basing on the analysis of the characteristics of business processes of an object represented by this website.

Keywords: website attendance, website traffic, discrete dynamic models, structure identification, prognosis properties

1. Introduction

Recently, the Internet and its components have become the separate objects of research. Despite the fact that this system is artificially created and its functionality is based on a respective logic, new functional requirements of its resources require its significant improvement and development due to the increasing number of users. One of the key components of this network is a website. Not long ago, website building was considered to be only creative work, for which it was impossible to use a formal approach. Recently, however, the interest in the development of formal methods for creating websites significantly increased.

Thus indicators to assess website quality were introduced; the systems, such as Google Analytics, which provide the dynamic assessment of these indicators were created. The issues of website analysis and quality control are examined in different works, especially those of A.M. Peleshchyshyn, V.M. Dubovyi, A.N. Moskvin, F. Khalil. As a next step of these studies the problem of analysis and computer-aided design of websites is considered in works [1-4]. Particularly, it is confirmed that their quality depends not only on the structure, but also on the content. However, the full formalism in the formulation of the task of creating high-quality website was not achieved. One way to form the task can be a formal model of evaluation of the quality characteristics of the resource, which would reflect the patterns of relation between the factors of influence and levels of achieved quality of the website in dynamics. This work is devoted to this topical task.

2. Characteristic of the factors that increase website attendance

Every website is created for users, and the increase in its attendance is one of the highest priority tasks of its owners. Of course, each website has its potential audience, but due to increased competition for users in the Web, they can expect to succeed with only the highest quality website. Site researchers comprise dozens of success factors [5]. However, the following key factors can be pointed out from this set: the page load time, ease of navigation and quality content. If the first factors are enabled insufficiently, it blocks the access to the next ones, and insufficient bringing into play at least one factor eliminates other good properties. The last factor from the set has the greatest variability, which mainly determines the essence of the entire website. Publications describing theoretically justified methods of the formation of site content, which would facilitate its attendance, are extremely rare.

The publications [6-8] state that the specialized thematic sites have an advantage in the competition for the consumer, because their content is more credible. The analysis of the types of specialized websites allows us to determine the following functions of the content: representation, reference, consulting, commercial. In this set the consultative function plays a key role. It can serve as a form of corporate text knowledge base, enrich the site users' specialized knowledge, serve as a basis for the site reference database and promote the informed choice of products, delivered by this website.

For being marketable, provided advisory information must be relevant. However, taking into account the different levels of awareness of the website users, it must also be complete, that is, it needs to be based on the conceptualization of the domain. These characteristics determine the use of ontologies to consult the advisory functions of websites. If the increase in overall attendance caused by the updated information in selected sections is desirable, easy navigation should be provided for the target audience of the website for these targeted sections. And to improve the navigation on the website, its structure should be understandable for the audience members interested in relevant topics. This structure is automatically determined by analyzing Web content as described in [1].

If the efficient attendance of the developing sections of the website should be provided, they must align with the key business process of the object presented through the website. The content of the subjects must meet the criteria of relevance and completeness. The efforts in forming the content of subjects according to these criteria can be organized basing on specific thesauri, which are formed automatically by the method described in [2].

Measures to improve attendance require certain costs. It is rational to evaluate their effectiveness before the implementation, including assessing by means of mathematical models. Website traffic forecasting is partially explored in the literature, for example, in [9, 10]. These works examine the users' navigation inside the website to identify the most frequent routes, as well as the choice of effective websites and webpages for advertising. At the same time, site developers are interested in the recommendations of a different sort, including the direction and pace of development of the site, that would ensure its maximum effectiveness.

3. Formal model of the website attendance

The attendance of the content pages of the website promotes the higher rating in the relevant thematic domains. This leads to increasing the website traffic to other subjects in general, and to the target pages, that are directly aimed to fulfil website tasks in particular. One of the factors contributing to such attendance is the regular update of subjects with relevant site content. The effectiveness of this replenishment can be assessed using formal mathematical models, that will also predict the effectiveness of similar operations in the future.

We begin the construction of the attendance models from defining the resulting and input parameters. The state variables are chosen: general traffic x1 of all website pages (WS), and the traffic of resulting pages x2 at some subsite (SS), as a target information spreading tool. The intensity of filling the most popular website subjects with high-demand information are chosen as the management parameters of a model. This intensity is evaluated with the predicted traffic u_j of the mentioned subjects located at some low-level subsites. A conceptual diagram of website traffic is depicted in Figure 1.

Daily traffic is characterized with many occasional factors, which causes a problem of predicting it accurately. At the same time the website attendance is influenced not by individual extreme, but average metrics. Thus in our model we analyze average weekly traffic.

It is necessary to use the principle of external additions in order to determine optimal complexity of the model, as mentioned in [11]. It means that the parameters of the model are determined in a training subset, and the accuracy check is performed in the control subset. The following principles form the base for constructing the website attendance model: 1) traffic is inertial; 2) variations of the state variables should agree with variations of the system management parameters; 3) state parameters influence each other.

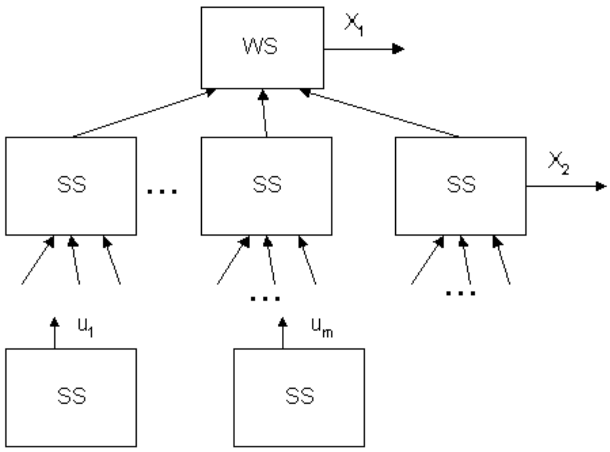


Fig. 1. Conceptual diagram of website traffic.

We add the lagged values of the predicted variable, which are summed with corresponding weighted traffic values for the topical subjects, in order to consider the traffic characteristics. The correspondence between the increase in the impact factor and the increase in the state variables is non-linear and can be modelled with power function with satisfactory accuracy. To coordinate the dynamics of the most visited subjects with resultant traffic we apply power function to the first one. The evolution of the website is considered through the introduction of new substructures at a time point T_j, whose additional traffic u_j significantly affects the overall attendance. We obtain the following number of model relationships as a result. We will include one lagged value of the state variable in the simplest model, and assume the influence of the development factors of new subjects to be a constant

$$x_{i,k} = g_{i,1} x_{i,k-1} + q_{i,1} \left(u_{1,k}\right)^{q_{i,2}} + \sum_{j=2}^{m} H(t-T_j)\left(g_{i,j} z^i_{j,k-1} + q_{i,2j-1}\left(u_{j,k}\right)^{q_{i,2j}}\right), \quad (1)$$

where

$$z^i_{j,k} = x_{i,k} - g_{i,1} x_{i,k-1} - q_{i,1}\left(u_{1,k}\right)^{q_{i,2}} - \sum_{l=2}^{j-1} H(t-T_l)\left(g_{i,l} z^i_{l,k-1} + q_{i,2l-1}\left(u_{l,k}\right)^{q_{i,2l}}\right) \quad (2)$$

is a forecast error of the $i-$ th outcome variable at a point of time k due to the influence of the j -th factor emerging at time T_j; H is the Heaviside function; g,q are model parameters to be identified.

In order to choose the right direction for further research, we complicate the previous model with the help of individual elements, analyzing the values of errors. At the first phase we supplement the model described in (1) – (2) with the mutual influence of state variables

$$x_{i,k} = g_{i,1}x_{i,k-1} + g_{i,2}x_{i*,k} + q_{i,1}\left(u_{1,k}\right)^{q_{i,2}} + \sum_{j=2}^{m} H(t-T_j)\left(g_{i,2j-1}z^{i}_{j,k-1} + g_{i,2j}z^{i*}_{jk-1} + q_{i,2j-1}\left(u_{j,k}\right)^{q_{i,2j}}\right), \quad (3)$$

where

$$i^* = \begin{cases} 2 & \text{when} \quad i=1, \\ 1 & \text{when} \quad i=2, \end{cases}$$

$$z^{i}_{j,k} = x_{i,k} - g_{i,1}x_{i,k-1} - g_{i,2}x_{i2,k-1} - q_{i,1}\left(u_{1,k}\right)^{q_{i,2}} - \sum_{l=2}^{j-1} H(t-T_l)\times \left(g_{i,2l-1}z^{i}_{l,k-1} + g_{i,2l}z^{i*}_{l,k-1} + q_{i,2l-1}\left(u_{l,k}\right)^{q_{i,2l}}\right) \quad (4)$$

At the next step, model (1)–(2) is supplemented with including the two previous values of the state variables, getting the relation

$$x_{i,k} = g_{i,1}x_{i,k-1} + g_{i,2}x_{i,k-2} + q_{i,1}\left(u_{1,k}\right)^{q_{i,2}} + \sum_{j=2}^{m} H(t-T_j)\times \left(g_{i,2j-1}z^{i}_{j,k-1} + g_{i,2j}z^{i}_{j,k-2} + q_{i,2j-1}\left(u_{j,k}\right)^{q_{i,2j}}\right) \quad (5)$$

where

$$z^{i}_{j,k} = x_{i,k} - g_{i,1}x_{i,k-1} - g_{i,2}x_{i,k-2} - q_{i,1}\left(u_{1,k}\right)^{q_{i,2}} - \sum_{l=2}^{j-1} H(t-T_l)\times \left(g_{i,2l-1}z^{i}_{l,k-1} + g_{i,2l}z^{i}_{l,k-2} + q_{i,2l-1}\left(u_{l,k}\right)^{q_{i,2l}}\right). \quad (6)$$

At the final stage the model (1)–(2) is supplemented by taking into account the intensity variability of the influence of control parameters according to the exponential law, because the intensity of the influence factor on state variables declines with time

$$x_{i,k} = g_{i,1}x_{i,k-1} + q_{i,1}e^{-k\cdot q_{i,2}}\left(u_{1,k}\right)^{q_{i,3}} + \sum_{j=2}^{m} H(t-T_j)\times \left(g_{i,2j-1}z^{i}_{j,k-1} + q_{i,3j-2}e^{(k-T_j)\cdot q_{i,3j-1}}\left(u_{j,k}\right)^{q_{i,3j}}\right) \quad (7)$$

where

$$z^{i}_{j,k} = x_{i,k} - g_{i,1}x_{i,k-1} - q_{i,1}e^{-k\cdot q_{i,1}}\left(u_{1,k}\right)^{q_{i,3}} - \sum_{l=2}^{j-1} H(t-T_l)\left(g_{i,2l-1}z^{i}_{l,k-1} + q_{i,3l-2}e^{-(l-T_l)\cdot q_{i,3l-1}}\left(u_{l,k}\right)^{q_{i,3l}}\right). \quad (8)$$

4. Identification of dynamic model

It is required to collect appropriate statistical material for identifying the model. This task is completed with the Google Analytics tool, and particularly with its component, a Content Drill Down option. In this case we choose weekly attendance characteristics of the relevant subjects. Since the model must be predictable, the control parameters are formed as modeled characteristics. These models are built on the basis of traffic statistics of the most active sections and the analysis of the processes accompanying them. Analysing state changes in the processes accompanying significant changes in dynamics of parameters u_j, we create the model of their dynamics as a discrete function

$$v_j(s_{j,l}) = \hat{u}_{j,l} \quad l = \overline{1, p_j}, \quad (9)$$

where $\{s_{j,l}\}_{l=1}^{p_j}$ is the set of identified states of a parameter u_j,

p_j is the quantity of identified states of the parameter u_j,

$\hat{u}_{j,l}$ is average evaluation of the parameter u_j in its state $s_{j,l}$.

When prognosing the sequence of states $S_{j,K} = \{s_{j,0},...,s_{j,K}\}$ of the paramer u_j, we receive the discrete set of its predicted values

$$U_j = \{v_j(s_{j,0}),...,v_j(s_{j,K})\}. \quad (10)$$

Substituting the set of predicted values instead of observed ones into the notation (1)–(2), we obtain the predictive dynamic model of the following form:

$$x_{i,k} = g_{i,1}x_{i,k-1} + q_{i,1}\left(v_1(s_{1,k})\right)^{q_{i,2}} + \sum_{j=2}^{m} H(t-T_j)\left(g_{i,j}z^{i}_{jk-1} + q_{i,2j-1}\left(v_j(s_{j,k})\right)^{q_{i,2j}}\right), \quad (11)$$

$$z^{i}_{j,k} = x_{i,k} - g_{i,1}x_{i,k-1} - q_{i,1}\left(v_1(s_{1,k})\right)^{q_{i,2}} - \sum_{l=2}^{j-1} H(t-T_l)\left(g_{i,l}z^{i}_{l,k-1} + q_{i,2l-1}\left(v_l(s_{l,k})\right)^{q_{i,2l}}\right). \quad (12)$$

Similar transformations are applied to model relations (3) – (8), summarizing the simulation results in minimized form as follows

$$x_{i,k} = f_{i,k}^n(\vec{g},\vec{q},x_{i,k-1},x_{i,k-2},x_{i2,k},W),\ n=\overline{1,4}, \quad (13)$$

where n is a serial number of the proposed model, $W=\bigcup_j U_j$ is a matrix of predicted values of the traffic of the developed subjects.

Let us generalize this relation by entering a vector of state variables $\vec{x}_k$ for time discrete k, a set of vectors of state variables X_k for all time discretes prior to $k+1\pi$ discrete and a set of vectors of developed subjects attendances for the entire analysed time interval. Then the representation (13) is simplified again

$$\vec{x}_k = F_k^n(\vec{g},\vec{q},X_k,W),\ n=\overline{1,4} \quad (14)$$

To identify these models the Method of Least Squares is used. Let the values of state variables in time discretes with numbers being the elements of sets $K_A=\left\{k_1^A,...k_{A_1}^A\right\}$ and $K_B=\left\{k_1^B,...k_{B_1}^B\right\}$ create relevant training and control samples. Then the criteria function of the Method of Least Squares has the form

$$V(\vec{g},\vec{q},\vec{X}_k \mid k\in K_A) = \sum_{k\in K_A}\left[\vec{x}_k - F_k^n\left(\vec{g},\vec{q},X_k,W\right)\right]^2 \quad (15)$$

Absolute and relative errors of the constructed model are estimated as follows:

$$E_k(n,K_A) = \left|\vec{x}_k - F_k^n\left(\vec{g},\vec{q},X_k,W\right)\right|, k\in K_A\bigcup K_B \quad (16)$$

$$\varepsilon_k(n,K_A) = E_k\left(n,K_A\right)/\Delta(n,K_A), k\in K_A\bigcup K_B, \quad (17)$$

where $\Delta(n,K_A)=\max_k E_k\left(n,K_A\right)$.

The main criterion for the accuracy of the model is considered to be its biggest relative error.

5. Numerical experiments

Daily attendance is characterized by a large number of random factors, making its accurate prediction difficult. At the same time, the characteristics of the site are affected not by individual extreme measures, but by average ones. Therefore, in our model we analyze average weekly attendance. The attendance model was identified with the use of data collected for the site of the Faculty of computer information technologies (FCIT) of Ternopil National Economic University (TNEU) (http://tanet.tneu.org) during a period of 16 weeks. During this period one control factor was revealed and one additional factor employed. Thus the period of analysis is divided into two sub-periods lasting 8 weeks each.

Fig. 1 shows the simulation results of the total attendance for two periods by the first control parameter. In terms of the minimum absolute error, the simple inertial model taking into account the impact of changes in the intensity of the control parameter was selected for the first time period. From the analysis of model behavior for the two periods, it is evident that previously built model should be extended.

Fig. 2 shows the simulation results of the total attendance for two periods with two control parameters. The maximum relative error of the model was 17.24%. General view of the constructed model is presented with the following equation

$$\begin{aligned} x_{1,k} &= g_{1,1}x_{1,k-1} + q_{1,1}e^{-k\cdot q_{1,2}}\left(v_1(s_{1,k})\right)^{q_{1,3}} + \\ &+ H(t-T_2)\times \\ &\times\left(g_{1,2}z_{2,k-1}^1 + q_{1,4}e^{(k-T_2)\cdot q_{1,5}}\left(v_2(s_{2,k})\right)^{q_{i,6}}\right)\times \\ &\times(1\pm\varepsilon_1) \end{aligned}, \quad (18)$$

where

$$z_{2,k}^1 = x_{1,k} - g_{1,1}x_{1,k-1} - q_{1,1}e^{-k\cdot q_{1,2}}\left(v_1(s_{1,k})\right)^{q_{1,2}},$$

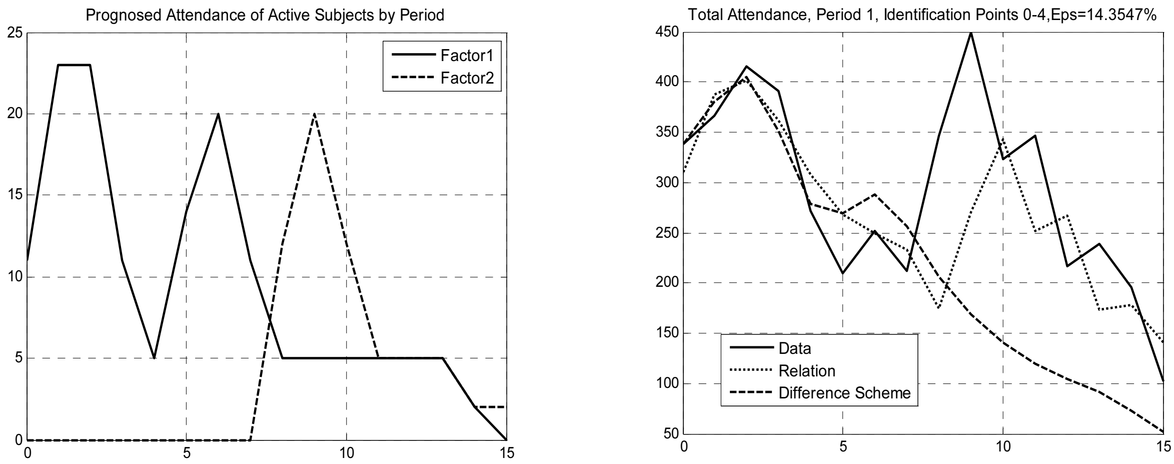
$$\begin{aligned} &g_{1,1}=0.7166,\ g_{1,2}=0.1946, \mathrm{T}_2=7,\ \varepsilon_1=17.24\%, \\ &q_{1,1}=7.0164,\ q_{1,2}=0.0489, q_{1,3}=0.9820, \\ &q_{1,4}=44.9988,\ q_{1,5}=0.0,\ q_{1,6}=0.5350. \end{aligned}$$

When identifying the model of resulting attendance, the impact of the total attendance had to be taken into account in order to achieve satisfactory accuracy. The same model gave good results for two periods of analysis (accuracy was 14.58%) without additional refinements. Simulation results are shown in Fig. 3, and the structure of the identified model is presented in the following ratio

$$\begin{aligned} x_{2,k} &= (19) \\ &= \left(g_{2,1}x_{2,k-1} + g_{2,2}x_{1,k} + q_{2,1}e^{-k\cdot q_{2,2}}\left(v_1(s_{1,k})\right)^{q_{2,3}}\right)\times \\ &\times\left(1\pm\varepsilon_2\right) \end{aligned}$$

where

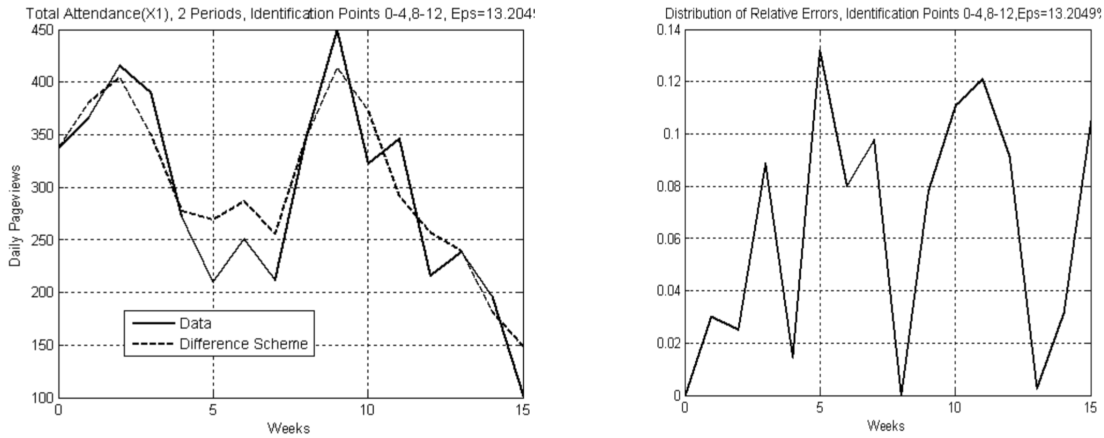
$$\begin{aligned} &g_{2,1}=0.2,\ g_{2,2}=0.25, \\ &q_{2,1}=5.0,\ q_{2,2}=0.0,\ q_{2,3}=0.06, \\ &\mathrm{T}_2=7, \varepsilon_2=13.28\% \end{aligned}$$



a) profile of control parameters for two periods

b) simulation of the total attendance for one factor

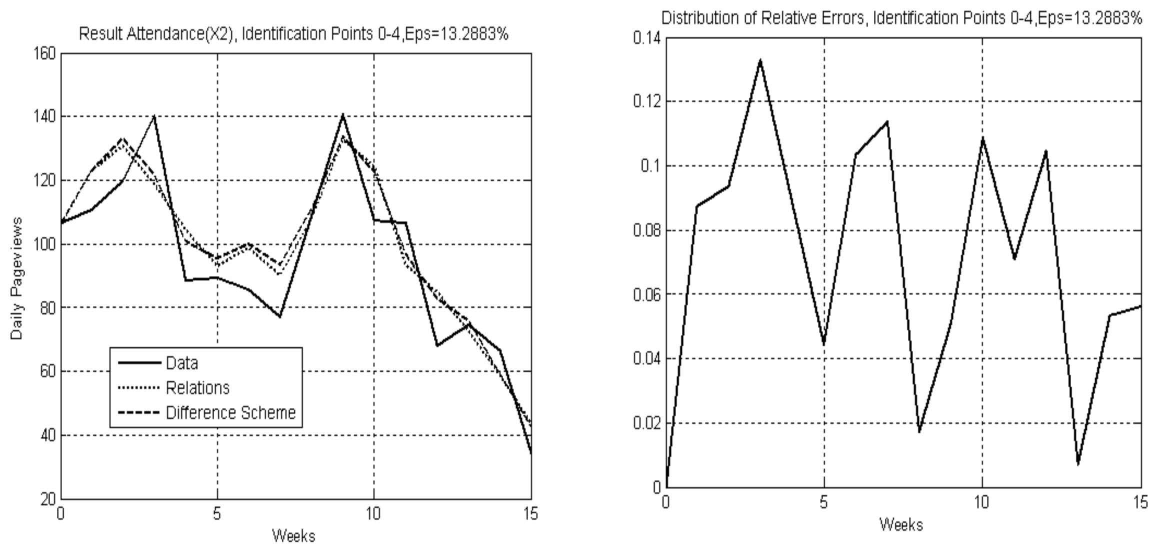
Fig.1. Total attendance simulation results for two periods with the first control factor.



a) profile of control parameters for two periods

b) simulation of the total attendance for one factor

Fig. 2. Total attendance simulation results for two periods with the first control factor.



a) model of the resulting attendance

b) distribution of errors for the resulting attendance model

Fig. 3. Total attendance simulation results for two periods with both control factors.

6. Conclusion

As a result of the linguistic approach and the theory of systems, the set of nonautonomous discrete dynamical models prognosing the website attendance was obtained for the first time. The impact factors in the models are described by the business processes of an object represented by this website, which makes it possible to evaluate the quality of site content. The method of identifying these models is introduced. The effectiveness of the proposed method is confirmed by solving the practical problems of the predicted traffic for the website of an educational department.

References

[1] N. Pasichnyk and M. Dyvak, "Method and algorithm for building structure of a web-site content, based on ontological approach," *Wisnyk Donetskogo NTU, seria "Informatyka, kibernetyka ta kompjuterni nauky",* vol. 15, pp. 184–189, 2012. (Ukrainian).

[2] N. Pasichnyk, V. Spilchuk, R. Shevchuk, and I. Spivak, "Formalized analysis of the web-site structure," in *Proc. International Conference on Modern Problems of Radio Engineering Telecommunications and Computer Science (TCSET)*, p. 391, Lviv-Slavsko, Ukraine, 2012.

[3] N. Pasichnyk, "Method forming an ontological content, based on analysis of information at specialized web-sites". *Wisnyk HNU: Inzenerija*, vol. 5, pp. 241–244, 2012. (Ukrainian).

[4] N. Pasichnyk, A. Melnyk, and N. Dobrovolska, "Management the website attendance based on the projected traffic," in *Proc. International conference "The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)"*, p. 277, Polyana-Svalyava, Ukraine, 2013. (Ukrainian)

[5] P. J. Meyers, "25-point Website Usability Check-list," *http://www.usereffect.com/topic/25-point-website-usability-checklist*

[6] A. Wostrow, "Why do thematic-oriented sites are gainful," *http:// www. seoded. ru /webmaster /sozdanie - saita/ tematicheskie -sajty. html* (Russian)

[7] S. Frank, "Elements of a successful, profitable website," *http:// entrepreneurs. about. com/ cs/ marketing / a / uc072403 .htm*

[8] T.Michael, "Niche sites make money," *http:// www. nichefinding. com/ niche-sites-make-money*

[9] A. Gorbunov, "Markov models for website traffic". In *IMAT-2007*, p. 65-73, Sverdlovsk, Russia: Ural University, 2007.

[10] F. Khalil, "Combining. web data mining techniques for web page access prediction," *http:/ eprints. usq. edu. au/ 4341/*, p.197, Quinsland, 2008.

[11] H. Madala and A. Ivakhnenko, *Inductive learning algorithms for complex system modelling*. Boca Raton, FL, USA: CRC Press, 1994.

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ДИНАМІКИ ХАРАКТЕРИСТИК ЯКОСТІ ВЕБ-САЙТУ

Наталя Пасічник, Микола Дивак, Роман Пасічник

Проаналізовано множину динамічних моделей прогнозуваня відвідуваності веб-сайтів. Ідентифіковано ці моделі на підставі численних експериментів. Це дало змогу структурувати адекватні моделі для прогнозування загальної та результуючої відвідуваності, грунтуючись на аналізі характеристик бізнес-процесів об'єкта, який цей веб-сайт представляє.

Natalia Pasichnyk – Master's Degree in Economic Cybernetics, Ternopil National Economic University, Ukraine (2010). Currently PhD student of the Computer Science Department, Ternopil National Economic University, Ukraine. Areas of research interests: modeling of dynamic systems with ucertanity.

Mykola Dyvak – Professor (2004), Dr. of Technical Science in Mathematical Modeling and Computing Methods awarded by Lviv Polytechnic National University, Lviv, Ukraine (2003). Post Ph.D. study in Intelligent Control, Tsinghua University, Beijing, China (1994). Ph.D. in Control of Technical System awarded by Moscow Power Energy University, Moscow, Russia (1992).

Master's Degree in Radio Engineering awarded by Lviv Polytechnic National University, Lviv, Ukraine (1986). Currently Dean of the Faculty of Computer Information Technologies and Head of the Computer Science Department at Ternopil National Economic University, Ukraine. Responsibilities include management and lecturing to students in Software Engineering. Areas of research interests: mathematical modeling of static and dynamic system, computing methods, parallel and distributed calculations, interval and reliable calculations, interval data analysis, application of the methods to environmental monitoring and distributed system.

Roman Pasichnyk – Associate professor (1997), Ph.D in Physics and Mathematics in Computational Mathematics awarded by Kazan University, Russia (1989). Master's Degree in Applied Mathematics awarded by Lviv National University, Ukraine (1979). Currently Associate professor of Economic Cybernetics and Informatics Department at Ternopil National Economic University, Ukraine. Areas of research interests: modeling of processes in complex systems.