ITHEA

International Journal INFORMATION MODELS & ANALYSES

2013 Volume 2 Number 1

International Journal
INFORMATION MODELS & ANALYSES
Volume 2 / 2013, Number 1

IJ IMA is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ IMA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals are given on www.ithea.org .
The camera-ready copy of the paper should be received by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

General Sponsor of IJ IMA is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION MODELS AND ANALYSES" Vol.2, Number 1, 2013

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia,
and Federation of the Scientific - Engineering Unions /FNTS/ (Bulgaria).

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: **Ina Markova**

Printed in Bulgaria

ISSN 1314-6416 (printed) **ISSN 1314-6424 (CD)** **ISSN 1314-6432 (Online)**

A JOINT GLOBAL AND LOCAL TONE MAPPING ALGORITHM FOR DISPLAYING WIDE DYNAMIC RANGE IMAGES

Alain Horé, Chika A. Ofili, and Orly Yadid-Pecht, Fellow, IEEE

***Abstract**: In this paper, we introduce an efficient improved tone mapping algorithm that can be used for displaying wide dynamic range (WDR) images on conventional display devices that are mainly of low dynamic range (LDR). That algorithm, which enhances the exponent-based tone mapping algorithm of Ofili et al., uses both local and global image information for improving the contrast and increasing the brightness of images. It also directly operates on the Bayer domain instead of the luminance/chrominance domain that is used by the vast majority of tone mapping algorithms. Experimental results performed on different WDR images show that we are able to get images that are more pleasant visually when compared to nine other tone mapping algorithms. These observations are also confirmed numerically through the use of the TMQI, an objective image quality measure.*

***Keywords**: Wide dynamic range images, tone mapping, color filter array, contrast enhancement.*

***ACM Classification Keywords**: A.0 General Literature - Conference Proceedings; I.4.3 Image Processing and Computer Vision – Enhancement*

Introduction

In imaging, the dynamic range describes the luminance ratio between the brightest part and the darkest part of a scene [1]. Natural sceneries often have a wide dynamic range (WDR) ratio that exceeds 100,000:1 [2]. Through the adaptation mechanism in human visual system (HVS), we are capable of simultaneously detecting different parts of a scene with high dynamic range. In fact, the HVS is capable of perceiving scenes over five orders of magnitude and can gradually adapt to scenes with dynamic range of over nine orders of magnitude. However, the conventional imaging display devices (such as LCD monitors, mobile phones) have a limited dynamic range (LDR), and they can reproduce dynamic ranges around two or three orders of magnitude. Due to recent technological improvements, modern image sensors can capture WDR images that accurately describe real world sceneries [3]. However, when these captured WDR images are displayed on standard display devices, they appear to be over-exposed in well-lit scenes or under-exposed in dark scenes. This leads to the loss of image details. A tone mapping algorithm is used to adapt the captured wide dynamic range scenes to the low dynamic range devices while maintaining the details from the original scenery. In Figure 1, it is shown how our proposed tone mapping algorithm can be used to enhance the quality and colors of an image.

There are two main classes of tone mapping algorithms: global techniques and local techniques. Global techniques are also called tone reproduction curves (TRC). They manipulate each individual pixel without considering its neighborhood [2, 4-8]. In fact, a single function is used for all the pixels, which outputs the same value for the same input intensity. Mathematically, a tone reproduction curve can be defined as follows:

$$y(p) = \psi\left(x(p)\right) \tag{1}$$

Figure 1. Effect of tone mapping on the quality of images.

where *x* is the original WDR image, *y* the final low dynamic range (LDR) tone mapped image, and *p* is a pixel. $\psi : \Gamma \to \Omega$ is a function that maps wide dynamic range intensities of the set Γ to the set of low dynamic range intensities Ω. The mapping function ψ can be for example a gamma function, a logarithmic function, a power function, or any continuous mathematical function. Common definitions for Γ and Ω are $\Gamma = [0, 2^{16} - 1]$ and $\Omega = [0, 2^8 - 1]$. Although TRC-based techniques are computationally efficient since they are simple to carry out, they are prone to issues like halos which appear as false black structures around pixels at the border between bright and dark areas [2]. Also, they are generally not enough efficient to enhance contrast and to preserve details because they do not consider the contrast characteristics of the local region [2].

Local tone mapping algorithms, also called tone reproduction operators (TRO), are spatial location dependent, and varying transformations are applied to each pixel depending on its surrounding [3]. Thus, they use different tonal curves in different regions of an image [2, 9 - 15]. They improve the local contrast and details of an image. However, they are computationally expensive, and are subject to artifacts such as false colors and false contours. Mathematically, a tone reproduction curve can be defined as follows:

$$y(p) = \mathrm{Z}\left(x(p), \zeta_p\right) \tag{2}$$

where *x* is the original WDR image, *y* the final low dynamic range (LDR) tone mapped image, *p* is a pixel, and ζ_p is a neighborhood of pixel *p*. $\mathrm{Z} : \Gamma \times \mathrm{N} \to \Omega$ is the mapping function that takes an input WDR intensity value from Γ and a neighborhood from N, and computes the final LDR intensity as an element of Ω. Neighborhoods can be expressed in the form of matrices as commonly found in image processing.

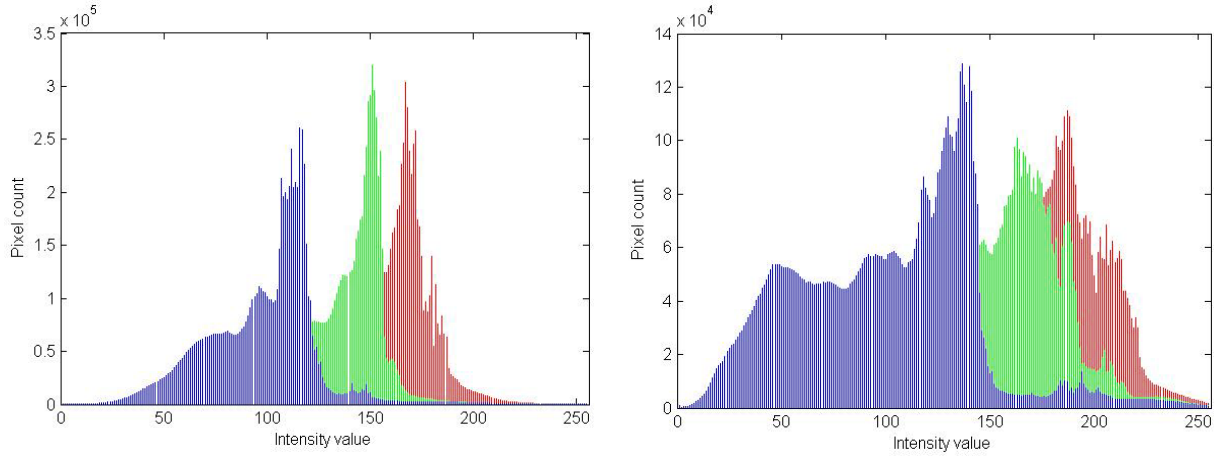
In [8], Ofili *et al*. have proposed a combined global and local tone mapping algorithm for displaying WDR images. In the rest of the paper, this algorithm which was developed in our lab, Integrated Sensors and Intelligent Systems lab (ISIS), will simply be referred to as our original tone mapping algorithm. It is based on an exponential function, and uses a low-pass Gaussian filter for getting the local information needed for performing tone mapping while preserving details. However, it is prone to false colors, and the local contrast is not always good as can be

seen in Figure 2 where we show two images from our original exponent-based tone mapping algorithm and the modified algorithm presented further in this paper. As we can see in the figure, the image from the modified algorithm contains more details and contrast than the image obtained from our original exponent-based tone mapping algorithm. We can also notice blur and false colors (reddish appearance) in the image from our original tone mapping algorithm. The histograms of the red, green and blue bands of the images of Figure 2 are shown in Figure 3. As we can notice, the histogram is more spread with the modified tone mapping algorithm in comparison to our original exponent-based tone mapping algorithm, which means better contrast.

The rest of the paper is as follows: first, we describe our original exponent-based tone mapping algorithm, and then we introduce our modified tone mapping algorithm which enables to get images with better contrast and reduced false colors. After that, we present some experimental results, and we end the paper with the concluding remarks.

(a) (b)

Figure 2. Tone mapped images. (a) Our original exponent-based tone mapping algorithm. (b) The modified exponent-based tone mapping algorithm presented in this paper.



(a) (b)

Figure 3. Histograms of the red, green and blue bands for the images shown in of Figure 2. (a) Our original exponent-based tone mapping algorithm. (b) The modified tone mapping algorithm presented in this paper.

Original exponent-based tone mapping algorithm

The tone mapping algorithm of Ofili *et al*. [8] is defined by:

$$\begin{cases} y(p) = x_{\max} \times \dfrac{1 - e^{-\frac{x(p)}{x_0(p)}}}{1 - e^{-\frac{x_{\max}}{x_0(p)}}} \\ x_0(p) = \kappa \times \mu_x + (x * h)(p) \end{cases} \quad (3)$$

where x is the original WDR image, y the final LDR image, p a pixel, x_{max} the maximum value for the display device (for example, for an 8-bit display device, $x_{max}=2^8-1=255$). x_0 is the adaptation factor and it is computed as the sum of a global component, $\kappa \times \mu_x$, and a local component $x * h$. In the global component part, κ is a parameter that plays a major role in brightening an image. When it increases, the tone mapped image becomes darker, while it becomes brighter when κ decreases. μ_x is the average intensity of WDR image x. Regarding the local component that is used to extract the local information, * denotes the convolution operation, and h is a low-pass filter. In this paper, a 2D Gaussian filter is used for h.

Modified tone mapping algorithm

As reported in the introduction, our original exponent-based tone mapping algorithm does not perform well regarding the contrast of an image. In fact, this is due to the local component term $x * h$ in the adaptation factor x_0. For getting more contrast, we propose to add a multiplicative constant 0.5 to that local component, and thus the modified tone mapping equation is given by:

$$\begin{cases} y(p) = x_{\max} \times \dfrac{1 - e^{-\frac{x(p)}{x_0(p)}}}{1 - e^{-\frac{x_{\max}}{x_0(p)}}} \\ x_0(p) = \kappa \times \mu_x + \dfrac{(x * h)(p)}{2} \end{cases} \quad (4)$$

Like in the case of our original tone mapping algorithm, the modified tone mapping equation is applied to images extracted from a Bayer color filter array, an idea originally developed by Meylan *et al*. [16]. Thus, we do not perform tone mapping on a luminance band and then combine the result with chrominance information as is done by different authors. In Figure 4, we show the traditional tone mapping workflow, and we also show the model developed by Meylan *et al*. and used in this paper.

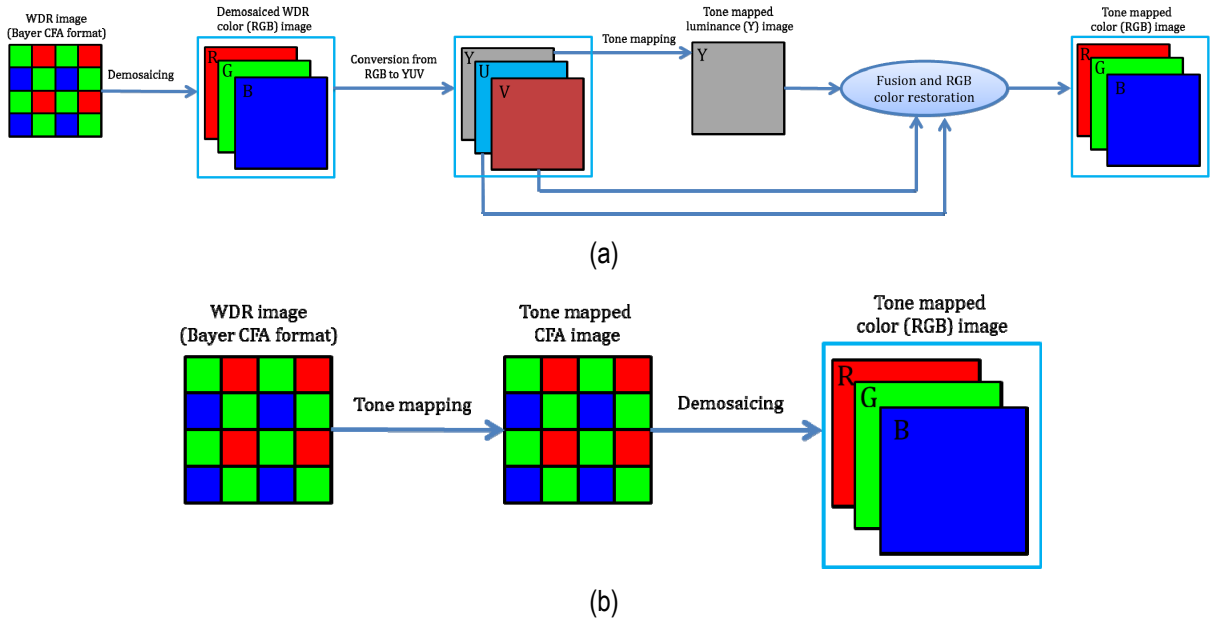


Figure 4. Image processing workflow. (a) Traditional model. (b) Model used in our tone mapping system and originally proposed by Meylan *et al*. [16]

Experimental results

For evaluation purposes, the modified exponential-based tone mapping algorithm is applied to three WDR images. The resulting tone mapped images are compared with nine other tone mapping operators: our original exponent-based tone mapping algorithm [8], bilateral filtering algorithm of Durand *et al*. [17], logarithmic mapping of Drago *et al*. [18], gradient tone mapping algorithm of Fattal *et al*. [11], display adaptive tone mapping algorithm of Mantiuk *et al*. [19], retinal display algorithm of Meylan *et al*. [16], Ashikmin's spatially varying operator [20], tone reproduction algorithm of Reinhard *et al*. [21], and Pattanaik *et al*.'s multiscale observer model [22]. Figures 5 to 22 show three WDR images tested as well as the resulting tone mapped images obtained from the different tone mapping algorithms. All image results were obtained from software implementations of the different tone mapping algorithms. As can be noticed in the figures, the modified exponent-based tone mapping algorithm gives images with good contrast and brightness as well as enhanced details. This confirms the good effect of modifying the local contrast component in the adaptation factor. For objective comparison, a tone mapping quality index proposed by Yeganeh *et al*. [23] is used, which assesses the effectiveness of the different tone mapping operators. The objective assessment algorithm produces three image quality scores that are used in evaluating the image quality of a tone mapped image: the structural similarity (S) between the tone mapped image and the original WDR image, the naturalness (N) of the tone mapped image, and the overall image quality measure which the authors call "tone mapping quality index" (TMQI). TMQI is computed as a non-linear combination of both the structural similarity score (S) and the naturalness score (N). It generally ranges from 0 to 1, where 1 is the highest in terms of image quality. The objective test results (TMQI) for the different tone mapping operators implemented are shown in Table 1. As can be seen, the TMQI values from the modified exponent-based tone mapping system are in the high score range that can be attained using this objective test. We have obtained the highest values for the images tested in comparison to the nine other tone mapping algorithms, which highlights the good performance of our algorithm. Moreover, we should note that although the tone mapping system by Meylan *et al*. [16] is also applied on the Bayer format of the colored WDR image, lower TMQI scores were obtained for all the images in comparison to our tone mapping system.

Table 1 TMQI scores for the various tone mapping operators.

Tone mapping operators	Image 1	Image 2	Image 3
Modified exponent-based tone mapping algorithm (this work)	**0.9626**	**0.8582**	**0.8463**
Original exponent-based tone mapping algorithm [4]	0.9411	0.8057	0.7730
Fattal *et al*. [5]	0.8675	0.7272	0.7371
Durand *et al*. [6]	0.9414	0.7474	0.7578
Drago *et al*. [7]	0.8932	0.8049	0.7888
Meylan *et al*. [3]	0.8628	0.8344	0.7793
Mantiuk *et al*. [8]	0.9159	0.7983	0.8200
Pattanaik *et al*. [9]	0.8295	0.7425	0.6903
Reinhard *et al*. [10]	0.9552	0.8228	0.8030
Ashikhmin [4]	0.7934	0.5357	0.7039

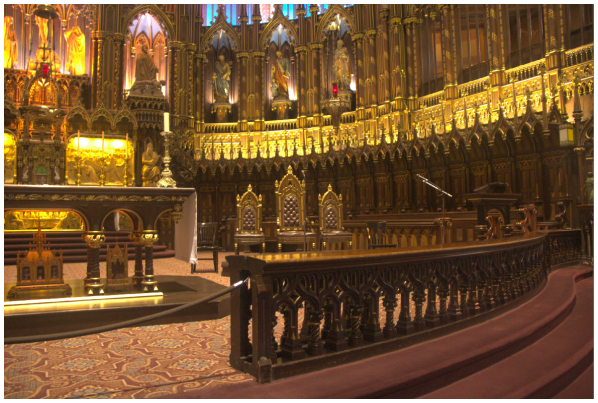
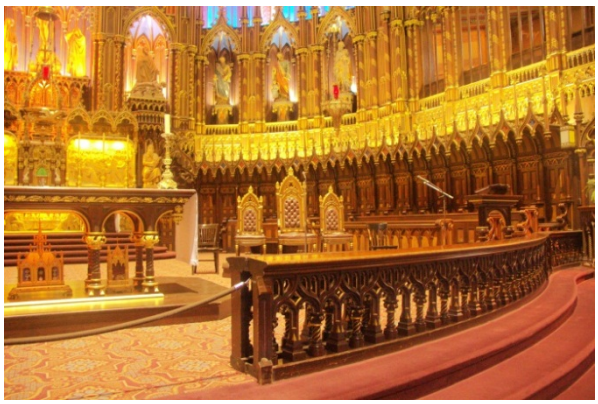


Figure 5 : Original WDR image



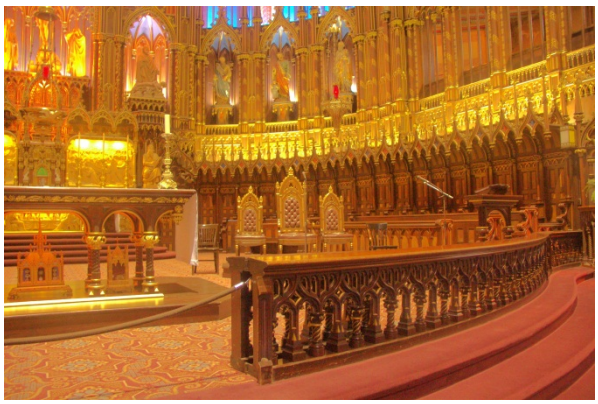


Figure 6 . Tone mapped images. (Left) Our modified tone mapping algorithm. (Right) Our original algorithm.

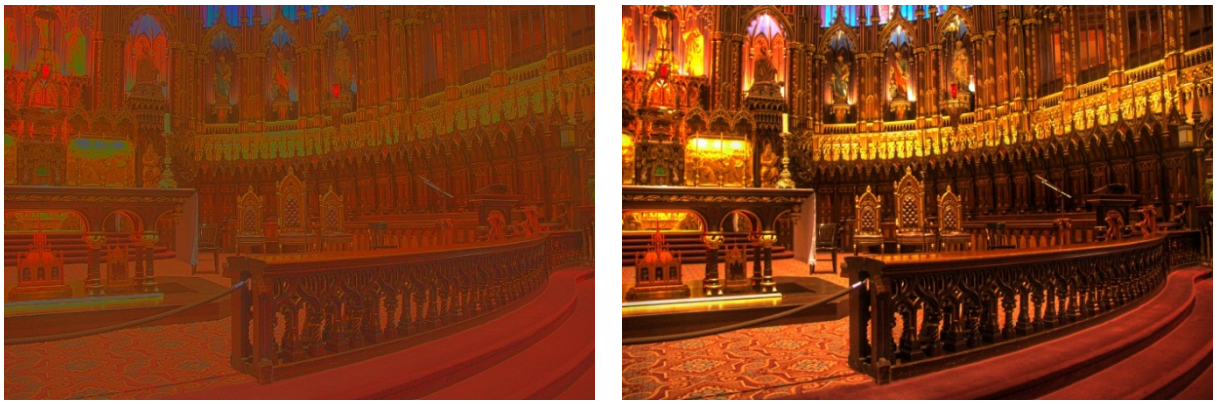


Figure 7. Tone mapped images. (Left) Ashikhmin's algorithm. (Right) Fattal et al.'s algorithm.

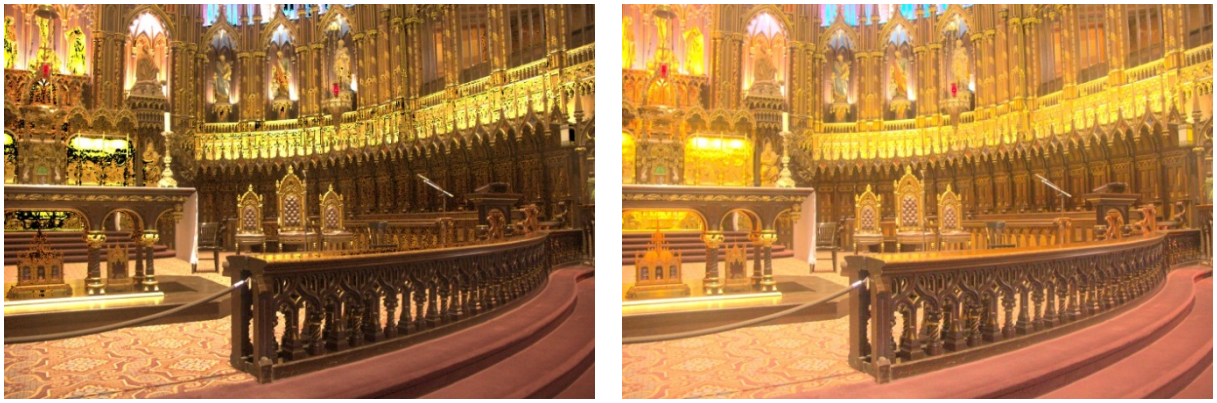


Figure 8. Tone mapped images. (Left) Durand et al.'s algorithm. (Right) Drago et al.'s algorithm.

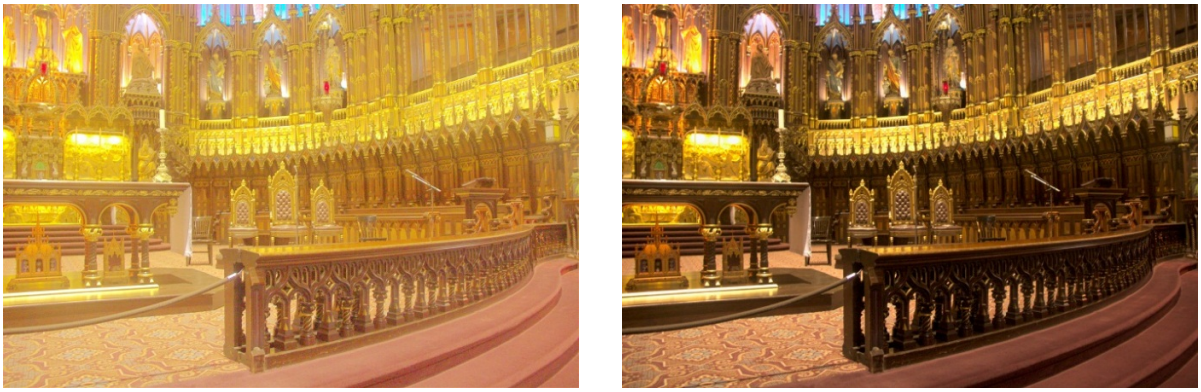


Figure 9. Tone mapped images. (Left) Meylan et al.'s algorithm. (Right) Mantiuk et al.'s algorithm

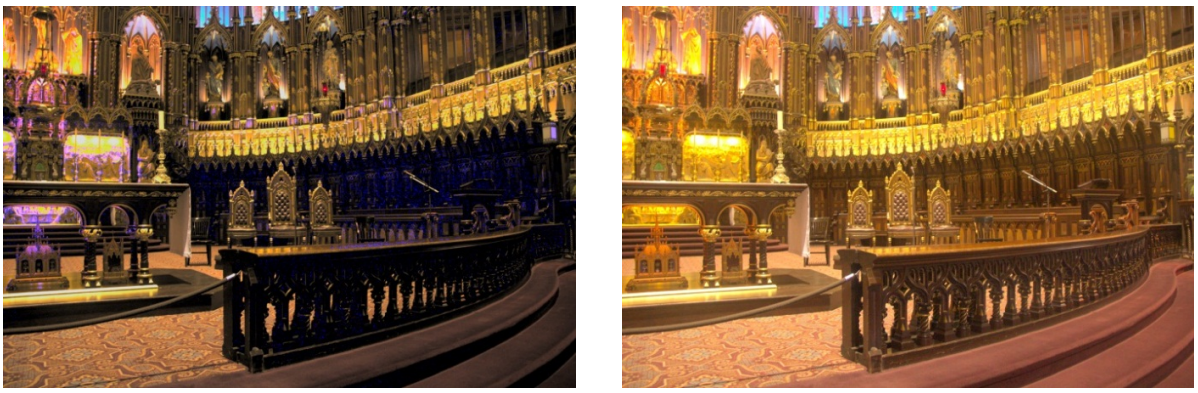


Figure 10. Tone mapped images. (Left) Pattanaik et al.'s algorithm. (Right) Reinhard et al.'s algorithm.

Figure 11 : Original WDR image.

Figure 12 : Tone mapped images. (Left) Our modified tone mapping algorithm. (Right) Our original algorithm.

Figure 13. Tone mapped images. (Left) Ashikhmin's algorithm. (Right) Fattal et al.'s algorithm.

Figure 14. Tone mapped images. (Left) Durand et al.'s algorithm. (Right) Drago et al.'s algorithm.

Figure 15. Tone mapped images. (Left) Meylan et al.'s algorithm. (Right) Mantiuk et al.'s algorithm.

Figure 16. Tone mapped images. (Left) Pattanaik et al.'s algorithm. (Right) Reinhard et al.'s algorithm.

Figure 17. Original WDR image

Figure 18. Tone mapped images. (Left) Our modified tone mapping algorithm. (Right) Our original algorithm.

Figure 19. Tone mapped images. (Left) Ashikhmin's algorithm. (Right) Fattal et al.'s algorithm.

Figure 20. Tone mapped images. (Left) Durand et al.'s algorithm. (Right) Drago et al.'s algorithm.

Figure 21. Tone mapped images. (Left) Meylan et al.'s algorithm. (Right) Mantiuk et al.'s algorithm.

Figure 22. Tone mapped images. (Left) Pattanaik et al.'s algorithm. (Right) Reinhard et al.'s algorithm.

Conclusion

In this paper, we have presented a modified and enhanced exponent-based tone mapping algorithm that exploits both global and local image information for producing low dynamic range images with high brightness and good contrast. Experimental tests performing with different WDR images have shown that high objective quality measure values are attainable using our algorithm, and it is also able to produce visually pleasant images. Due to its simplified mathematical model, our algorithm can be a good candidate for implementation on system-on-a-chip.

Bibliography

[1] O. Yadid-Pecht and R. Etienne-Cummings, “CMOS imagers: from phototransduction to image processing”, Kluwer Academic Publisher, 2004.

[2] J. Duan, W. Dong, R. Mu, G. Qiu, and M. Chen, “Local contrast stretch based tone mapping for high dynamic range images”, IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision, pp. 26-32, 2011.

[3] Y. Dattner and O. Yadid-Pecht, “High and low light CMOS imager employing wide dynamic range expansion and low noise readout”, IEEE Sensors Journal, vol. 12, no. 6, pp. 2172-2179, 2012.

[4] J. Tumblin and H. Rushmeier, “Tone reproduction for realistic images”, IEEE Computer Graphics and Applications, vol. 13, no. 6, pp. 42-48, 1993.

[5] G. W. Larson, H. Rushmeier, and C. Piatko, “A visibility matching tone reproduction operator for high dynamic range scenes”, IEEE Transactions on Visualization and Computer Graphics, vol. 3, no. 4, pp. 291 – 306, 1997.

[6] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg. “A model of visual adaptation for realistic image synthesis”, ACM SIGGRAPH, pp. 249-258, 1996.

[7] J. Duan, G. Qiu, and G.m D. Finlayson. “Learning to display high dynamic range images”, Pattern Recognition, vol. 40, no. 10, pp. 2641-2655, 2007.

[8] C. A. Ofili, S. Glozman, and O. Yadid-Pecht, “An in-depth analysis and image quality assessment of exponent-based tone mapping algorithm”, International Journal Information Models and Analysis, vol. 1, no. 3, pp. 236-250, 2012.

[9] D. Lischinski, Z. Farbman, M. Uyttendaele, R. Szeliski. “Interactive local adjustment of tonal values”. ACM Transactions on Graphics. vol. 22, no. 3, pp.646–653, 2006.

[10] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images”, ACM SIGGRAPH, pp. 267-276, 2002.

[11] R. Fattal, M. Werman, and D. Lischinski, “Gradient domain high dynamic range compression”, ACM Transactions on Graphics, vol. 21, no. 3, pp. 249-256, 2002.

[12] J. Tumblin and G. Turk, “LCIS: A boundary hierarchy for detail preserving contrast reduction”, ACM SIGGRAPH, pp. 83-90, 1999.

[13] J. Duan, M. Bressan, C. Dance, and G. Qiu, “Tone-mapping high dynamic range images by novel histogram adjustment”, Pattern Recognition, vol. 43, no. 5, pp. 1847-1862, 2010.

[14] D. J. Jobson, Z. Rahman, and G. A. Woodell, “A multiscale retinex for dridging the gap between color images and the human observation of scenes”, IEEE Transactions on Image processing, vol. 6, no. 7, pp. 965-976, 1997.

[15] J. W. Lee, R. H. Park, and S. Chang, “Tone mapping using color correction function and image decomposition in high dynamic range imaging”, IEEE Transactions on Consumer Electronics, vol. 56, no. 4, pp. 2772-2780, 2010.

[16] L. Meylan, D. Alleysson, and S. Süsstrunk, "A Model of retinal local adaptation for the tone mapping of color filter array

images", Journal of the Optical Society of America A, vol. 24, no. 9, pp. 2807-2816, 2007.

[17] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images", ACM Transactions on Graphics, vol. 21, no. 3, pp. 257-266, 2002.

[18] F. Drago, K. Myszkowski, N. Chiba, and T. Annen, "Adaptive logarithmic mapping for displaying high contrast scenes", Computer Graphics Forum, vol. 22, no. 3, pp. 419-426, 2003.

[19] R. Mantiuk, L. Kerofsky, and S. Daly, "Display adaptive tone mapping", ACM Transactions on Graphics, vol. 27, no. 3, pp. 193-202, 2008.

[20] M. Ashikhmin, "A Tone mapping algorithm for high constrast images", Eurographics Workshop on Rendering, pp. 145-155, 2002.

[21] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Display adaptive tone mapping", ACM Transactions on Graphics, vol. 21, no. 3, pp. 267-276, 2002.

[22] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg, "Time-dependent visual adaptation for fast realistic image display", ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques, pp. 47–54, 2000.

[23] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images", IEEE Transactions on Image Processing, vol. 22, no. 2, pp. 657-667, 2013.

Authors' Information

Alain Horé *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, T2N 1N4, Canada; e-mail: ahore@ucalgary.ca.*

Chika Antoinette Ofili *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, T2N 1N4, Canada; e-mail: cofili@ucalgary.ca.*

Orly Yadid-Pecht *is with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, T2N 1N4, Canada; e-mail: orly.yadid.pecht@ucalgary.ca.*

COMPONENT MODELING: ON CONNECTIONS OF DETAILED PETRI MODEL AND COMPONENT MODEL OF PARALLEL DISTRIBUTED SYSTEM

Elena Lukyanova

Abstract: Connections of detailed Petri *N* model and its component *CN*-net on the level of their structural and dynamic properties are investigated.

Keywords: *component Petri net, components-places, components-transitions, languages of component net with components-places and component net with components-transitions.*

ACM Classification Keywords: *Petri net, language of Petri net, homomorphism, epimorphism, model, model verification.*

Introduction

One of the stages of process of research of the dynamic systems of any complication is a construction of qualitative model of the investigated system. Realization of researches of model makes it possible to set a number of specific properties of model with subsequent interpretation of the results in relation to the target system. Petri nets [Kotov, 1984] are convenient means of modeling (detailed modeling) of various parallel distributed systems. In a detailed modeling of real systems and objects we can receive a larger network, which makes the analysis of detailed models impracticable. When taking component Petri net (CN -net) as a model of parallel distributed system [Lukyanova 1, 2012, Lukyanova, 2011], we get the possibility to work with the model that is much smaller than the original Petri detailed model. It is important that CN -model does not lose its conformity for the description of the initial investigated system.

This paper describes currently obtained results on the links between the detailed Petri model and component Petri model of investigated parallel distributed system. These are the results of the relationship of structural properties of CN - and detailed Petri models, of the language of detailed Petri model and the language of CN - model with only components-transitions. In the paper, we determine the language of component Petri net containing only components-places, its relations with the language of detailed model and characteristics of input languages.

Component modeling of parallel distributed systems

On the initial stage of component modeling, analysis of the original complex system for allocating its constituent simpler objects – groups of identical and single-type processes (processes that are of the same type differ only in the number of identical parallel processes). It allows forming allocated groups of identical and single-type processes as blocks of composite model components at the phase of model construction. Thus, in the modeling process, we obtain a detailed model of the original system, which has identical and single-type processes placed in the appropriate blocks – composite components (components-places C_p and components-transitions C_t) [Lukyanova 2, 2012]. Among the constituent components there may be identical and single-type composite components [Lukyanova 3, 2012]. Petri net constructed is called component Petri net.

Definition 1. Component Petri net (CN -net) is a directed graph, described by the ordering quinary $CN = (P, T, F, W, M_0)$, where P is a finite set of places consisting of subsets P_1 and P_2 (P_1 is a finite set of component-places, P_2 – a finite set of places that are left after the separation of component-places); T – final set of transitions, consisting of subsets T_1 and T_2 (respectively, the set of components-transitions and a set of transitions that are left after the separation of the component transitions); $F \subseteq P \times T \bigcup T \times P$ – the incidence relation between places and transitions; $W : F \to N \setminus \{0\}$ – the multiplicity function of arcs; M_0 – the initial marking of net.

The ratio of the incidence F and multiplicity function of arcs W determine the function of the incidence I, defining the rule $I : (P \times T \bigcup (T \times P) \to N)$. Incidence function defines that the elements of one set of arcs cannot be connected, and describes the sets of input and output elements.

Component-place Cp designs some single-type processes of the detailed Petri model of the investigated system, which begins and ends with the place (places). Component-place is triple $Cp = (N, X, Y)$, where N is Petri net, $X \subseteq P$, $Y \subseteq P$ – the sets of its initial and final places correspondingly, not having respectively the input and output arcs, and $X \bigcap Y = \varnothing$. Component Cp, as a structural element of CN -net, is a place that has input and output arcs, and as in the regular Petri nets is a condition that determines the possibility of an event – transition firing in CN -net.

Component-transition Ct is an area of net of the detailed model, designing some of the single-type process, beginning and ending with transition (transitions).

Component transition Ct is triple $Ct = (N, U, V)$ where N is Petri net, $U \subseteq T$, $V \subseteq T$ – respective sets of its initial and final transitions that do not have the respective input and output arcs, and $U \bigcap V = \varnothing$. Component Ct as part of CN -net has input and output arcs, and is an event and transition of CN -net.

Remark 1. It is natural to denote corresponding places and transitions in identical and single-type composite components with the same symbols, and corresponding places and transitions with the same symbols in identical parallel processes.

As an example of CN -model, Fig. 1 shows a small CN -net. This CN -net includes:

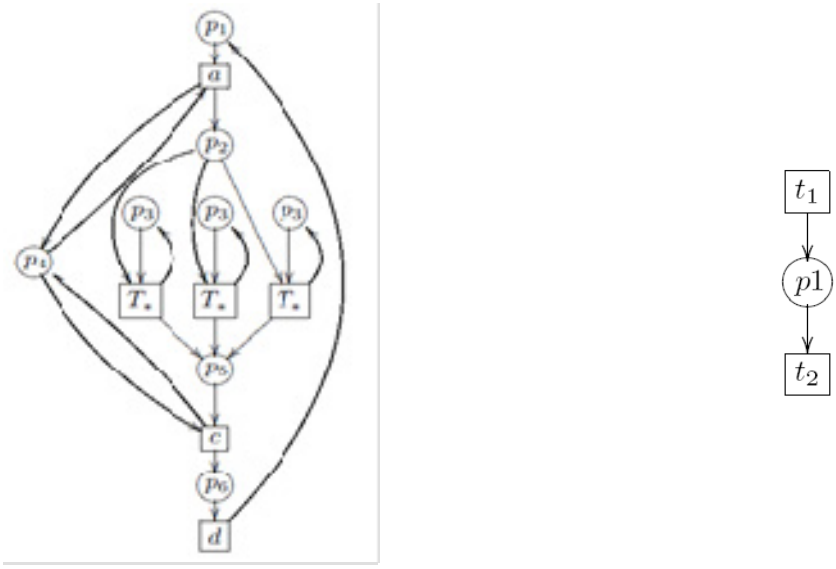
1). Three identical parallel processes. According to [Lukyanova 3, 2012, Lukyanova 4, 2012], in these processes, relevant places and transitions, according to Remark 1, are indicated by the same markings (Fig. 1, a);

2). Composite components – components-transitions *T** (Fig. 1, b).

At that, CN -net of Fig. 1 is itself one of composite components-places P_8^*, P_{10}^*, P_{14}^* P_{16}^*, P_{17}^*, P_{20}^* – component-place P_{16}^* in CN -net [Lukyanova 3, 2012], as shown at Fig. 2.

CN-model analysis to determine the structural properties

The effectiveness of analysis of CN -models to determine the structural characteristics of a detailed Petri model of parallel distributed system has been stated in [Lukyanova, 2011, Lukyanova 3, 2012]. The proposed analysis is based on the analysis of the component net (CN -net), in which the component parts are considered only as places and transitions, and on the analysis of one representative from each group of single-type components. From this group of single-type components, "minimal" representative is selected – an integral component with the least number of identical parallel processes. So here is the final theorem [Lukyanova 3, 2012].



a) b)

Figure 1. a) CN -net; b) component-transition T^* in CN -net.

Figure 2. CN -net, in which P_8^*, P_{10}^*, P_{14}^* P_{16}^*, P_{17}^*, P_{20}^* are components-places.

Theorem 1. Detailed Petri model of the investigated system does not have structural property if corresponding CN -net does not have this property.

Detailed Petri model of the investigated system has structural property if corresponding CN -net has this property subject to the fulfillment of following properties of the structural elements of the CN -net:

a) for each group of single-type components-transitions, the component-transition with the least number of identical parallel processes is live;

b) for each group of single-type components-places, a combined system of inhomogeneous linear Diophantine equations (SILDE) corresponds to a component-location with the least number of identical parallel processes.

Languages of *CN*-model

CN -net research involves the analysis not only of structural characteristics, but also of a series of dynamic characteristics. Dynamic behavior of the modeled system is described in terms of the net functioning. Functioning of the net is formally described by the set of the sequence of firing and set of achievable net marking.

While determining language, generated by CN -net, one should consider:

- presence of two different types of composite components in the component Petri net: components-places Cp and components-transitions Ct;
- two-aspects approach is applicable to the operation of the component Petri nets [Lukyanova, 2011, Lukyanova 2, 2012]: on the one hand, the composite components of the CN -net serve immediately; on the other hand, the study of the functioning of composite component is an integral part of the investigation of the properties of the CN -net model.

This means that we need a separate consideration of the nature of the sets of possible sequences of events or of sets of reachable markings in the composite components themselves, in component net only with components-transitions, in component net only with components-places, and in component net with both components-places and components-transitions.

In the study of CN -language net, containing only the components-transitions C_t [Lukyanova 4, 2012], operation of the net is described in terms of sequences of firing of transitions [Kapitonova, 1988, Kotov, 1984].

Given X and Y respectively finite alphabets of detailed Petri model and CN -model only with components-transitions of the investigated system, then X^* is a set of all words in alphabet X, $Y^* = (X \bigcup \{T_1^*, T_2^*, ..., T_n^*\})^*$ – a set of all words in the alphabet $Y = X \bigcup \{T_1^*, T_2^*, ..., T_n^*\}$, where T_k^* ($k = 1,2,..., n$) are names of the various components-transitions C_{t_k} $(k = 1,2,..., n)$ in the CN -net.

Definition 2. Let's call the language $L_t(N)$ of a detailed Petri model N, in which composite components – components-transitions can be separated, its free language, for which marking for the identical and single-type components-transitions C_t .is the same, according to the Remark 1.

Definition 3. Language $L_t(CN)$ of CN -net, containing only components-transitions, is the set of firing sequences of CN -net, which is a subset of the set of all words in alphabet Y, resulting from expansion of the alphabet X of original detailed model N with names of components-transitions T_k^*. At that, in Y^* the words are allowed, not containing those characters from X, which are used in marking of transitions in components-transitions in the net N.

It is stated in [Lukyanova 4, 2012] that:

- languages of the composite components, differing only in number of identical parallel processes and modeling single-type processes are congruent;
- transition of words in alphabet X of detailed Petri model N of the investigated parallel distributed system in set of words of alphabet Y of its -model, containing only the components-transitions, is determined up to epimorphism preserving concatenation of words;

– the language of -net, containing only components-transitions, is surjective homomorphic mapping of the language of original detailed Petri model N of investigated parallel distributed system.

Let's examine component Petri net, which contains only components-places as an integral components. In this component net, some areas of detailed Petri model are accumulated in special places of CN -net – components-places Cp . Place in a Petri net is a condition for determining the possibility of an event – transition firing. Therefore, in the case of CN -net, containing only the components-places, we will examine a set of reachable markings in nets N and CN for the study of its language. We can present possible changes in the net marking, that result from transition firing, in the form of a graph marking – a directed graph whose set of nodes is formed by the set of reachable net marking. Such a graph, describing the dynamics of the net functioning, reflects the distribution of tokens – change of the conditions for the possibility of an event in the net. Then the operation of the net can be described in terms of reachable marking [Kapitonova, 1988, Kotov, 1984]. To do this, let's mark the nodes of the graph of reachable markings with symbols so to have valid character string (set of words) in some alphabet T , obtained by writing out symbols of nodes along the paths in the graph of reachable markings starting at the initial marking. We have set $T*$ of all words in alphabet T , made up of the symbols of the nodes of net reachable markings, a subset of which is the set of possible sequences of nodes of net reachable markings. Then each step of the composite component also generates a symbol from the set of names that corresponds to the node of the graph of reachable markings of this component.

Given net N contains m places, then marking of net N is described by m -dimensional vector which coordinates correspond to places, ordered according to the numeration of places in the net. In the net N , components-places are separated, suppose the total number of places that were allocated to the components is Cp_i $(i = 1,2,...,n)$, k . Then marking of CN -net will be configured by l -dimensional vector ($l = m - k$), which coordinates correspond to places, ordered according to the numeration of places in the CN -net. In this case, the numeration of places, adopted in the net N , is preserved in CN -net.

Let A and B , respectively, are finite alphabets of detailed model N and CN -model only with components-places. Then A – set of names of m -dimensional vectors corresponding to the nodes of the graph of reachable markings of net N , B – set of names of l -dimensional vectors corresponding to the nodes of the graph of reachable markings of net CN . Set $A*$ – set of all words in the alphabet A , $B* = (\phi(A) \bigcup \{P_1^*, P_2^*, ..., P_n^*\})^*$ – set of all words in the alphabet $B = \phi(A) \bigcup \{P_1^*, P_2^*, ..., P_n^*\}$, where P_i^* $(i = 1,2,...,n)$ are the names of the different nodes in the graph of reachable markings of CN -net, that correspond to the names of nodes of components-places Cp_i $(i = 1,2,...,n)$. Mapping ϕ transits names of *m*-dimensional vectors in the names of l -dimensional vectors.

Let a – name of m -dimensional vector $(a_1, a_2, ..., a_m)$ from the alphabet A , then $\phi(a) = a'$ – name of l -dimensional vector $(a_1', a_2', ..., a_l')$ from the alphabet B . As a result of combining of areas of the net N , which are separated in the components Cp in the CN -net, in places – each element a_j' $(j = 1,2,...,l)$ is a mapping at least of one element a_s $(s = 1,2,...,m)$. Here comes a theorem.

Theorem 2. The mapping ϕ is surjective mapping, which transits alphabet A in alphabet B .

Definition 4. The language $L_p(N)$ of detailed Petri model N, in which composite components – components-places Cp can be separated, is its free language that is defined in terms of the set of reachable marking in the net. Places and transitions in the identical and single-type components-places are marked according to Remark 1.

Definition 5. The language $L_p(CN)$ of CN-net, containing only components-places, is the set of sequences received by singling out symbols of nodes along the paths in the graph of reachable markings of CN-net, starting in initial marking and leading to each reachable marking in the net. Language $L_p(CN)$ is a subset of the set of all words in alphabet B. Alphabet B consists of the mapping of the alphabet A and the names of the nodes corresponding to component-places Cp_i.

Consider the transition ς of symbol sequences of the graph of reachable markings of net N into symbol sequences of the graph of reachable markings of net CN, containing only components-places. Thus, we consider the transition, transforming the words in the language $L_p(N)$ in the words in the language $L_p(CN)$ of CN-net, containing only components-places.

Suppose that a word $q \in A^*$ has a form

$$q = (a_1, a_2, ..., a_m)(b_1, b_2, ..., b_m)(p_1', p_2', ..., p_m')(p_1'', p_2'', ..., p_m'')(c_1, c_2, ..., c_m)(d_1, d_2, ..., d_m) = \\ = abp'p''cd .$$

Symbols a, b, c, d denote the names of the nodes of the graph of reachable markings of detailed model N, which are not nodes of the graph of markings of any component-place Cp_i. Symbols p', p'' are the names of the nodes of graph of marking of the component-place Cp_i. So in the word under consideration $abp'p''cd$, the names of nodes of the graph of marking of one component Cp_i take part.

Let's make following notation in the word q: $ab = q_1$, $p'p'' = \overline{q}$, $cd = q_2$. Then the record of the original word *q* we get as $q = q_1 \overline{q} q_2$.

At transition ς, the mapping of word $q \in A^*$ is a word $\varsigma(q) = h \in B^*$: $h = \phi(q_1) P^* \phi(q_2)$, where $\phi(q_1) = \phi(a)\phi(b)$ and $\phi(q_2) = \phi(c)\phi(d)$. Thus the effect of the transition ς on the word *q* is determined by words mapping involved in the concatenation of the word q:

$$\varsigma(q) = \varsigma(q_1 \overline{q} q_2) = \varsigma(q_1)\varsigma(\overline{q})\varsigma(q_2) = \phi(a)\phi(b) P^* \phi(c)\phi(d) = a'b' P^* c'd'.$$

Finally, for any word $q = abp'p''cd$ from A^*, we get its mapping $h = \varsigma(q) = a'b' P^* c'd'$ – a word from B^*.

Transition ς is completely determined by the values on the letters in the alphabet A so, that each symbol $y \in B$ is a mapping at least of one symbol $x \in A$. The conclusions on transition ς are as follows:

1. $\varsigma(xy) = \varsigma(x)\varsigma(y)$ holds for all x and y over A;
2. $\varsigma(e) = e$, where e is empty word;

3. $\varsigma(x) = \phi(x)$ for words x of any length from the names of the nodes of reachable markings of network N, that are not the names of the nodes of graph of reachable markings of any constituent component Cp_i;

4. $\varsigma(x) = P_i^*$ for all words of any length *x* from the names of the nodes of graph of reachable markings of composite component Cp_i.

Then for $L_p(CN)$ – language of CN -net, containing only the components-places, we receive:

$$L_p(CN) = \varsigma(L(N)) = \{y / y = \varsigma(x), \ \text{где} \ x \in L_p(N)\}.$$

Resulting from what was said above we have theorems.

Theorem 3. Language of CN -net, containing only the components-places, is surjective homomorphic mapping of the language of the original detailed Petri model N of the investigated parallel distributed system.

Theorem 4. Epimorphism ς generates epimorphism ϕ.

Properties of the languages of *CN*-model

Important problems of Petri nets: a problem of membership, a problem of emptiness, finiteness problem – are solved on the language level. Let's examine the corresponding problems for the above languages. The problem of membership is connected with a check of membership of any word p to language L, in the problem of emptiness we need to find out whether the set L is empty, in the problem of finiteness one need to find out whether L is a finite set.

Theorem 5. The problem of membership is solvable for languages $L_t(N)$, $L_p(N)$, $L_t(CN)$, $L_p(CN)$.

The proof is based on the fact that the procedure for checking if p is an element of one of the mentioned languages, will end in a finite number of steps. In the case of languages $L_t(N)$ and $L_t(CN)$, for any word p of corresponding net we must determine, that p is a sequence of firing of transitions of this net. It's enough:

1) to verify that at the initial marking M_0, transition will fire, which symbol is the first in the word p;

2) to change the marking M_0 into immediately following after M_0 marking M_1;

3) to check the possibility of transition firing at M_1, transition symbol is standing second in the word p, etc.

The word p has a finite number of symbols, therefore, the process of successive inspections will be completed in a finite number of steps. In the case of languages $L_p(N)$ and $L_p(CN)$, we play similarly, stating that the word p is a sequence of changing of net markings along the paths in the graph of reachable markings, starting at the initial marking and occurring as a result of firing of its transitions.

Theorem 6. Reachability problem in the given Petri net marking is reducible to the problem of language membership to corresponding component net.

The proof is based on finding a membership of some marking of net N to the set $R(N)$ of its reachable markings. For this, it is enough for its corresponding component net to figure out whether mapping of the word, corresponding to this marking, is an element of language CN.

Theorem 7. Emptiness problem is soluble for languages $L_t(N)$, $L_p(N)$, $L_t(CN)$ and $L_p(CN)$.

The proof is based on the verification of the following fact, whether at least one transition of this net fires at the initial marking.

Corollary 1. Language $L_t(N)$ ($L_p(N)$) is empty if and only if the language $L_t(CN)$ ($L_p(CN)$) is empty.

Conclusion

The results obtained in the study of connection between structural and dynamic properties of detailed Petri model and component model of parallel distributed system allow analyzing of detailed Petri model with the help of its CN-model. It is effective because CN-models meet modern requirements for models of large systems and complex real tasks to be manageable and easy to analyze. In this paper, we determine the language of component Petri net, containing only components-places, and it is stated that the language of CN-net, containing only the components-places, is a surjective homomorphic mapping of the language of the original detailed Petri model of investigated parallel distributed system; at the same time, the mapping of the alphabet A of net N in the alphabet B of CN-net (in terms of the set of reachable markings in the net) is an epimorphism, which is generated by transition of language $L_p(N)$ to the language $L_p(CN)$. In the languages under consideration $L_t(N)$, $L_p(N)$, $L_t(CN)$ and $L_p(CN)$, the problems of membership, emptiness, and finiteness are examined.

Bibliography

[Kotov, 1984] V. E. Kotov. Petri nets / V. E. Kotov – Moscow: Nauka, 1984. – 160 p. (in Russian).

[Lukyanova 1, 2012] Lukyanova E. A. On component modeling of systems with concurrency / E. A. Lukyanova // Naukovi Zapysky of NaUKMA. Computer Science. – 2012. – T. 138. – P. 47–52. (in Ukrainian).

[Lukyanova, 2011] Lukyanova E. A. On component analysis of parallel distributed systems / E. A. Lukyanova // TVIM – 2011. – № 2. – P. 71–81. (in Russian).

[Lukyanova 2, 2012] Lukyanova E. A. The structural elements of a Petri net component / E. A. Lukyanova // Problemy programuvannya – 2012. – № 2-3. – P. 25–32. (in Russian).

[Lukyanova 3, 2012] Lukyanova E. A. The study of the structural elements of the single-type CN-network during the component modelling and analysis of complex systems with concurrency / E. A. Lukyanova, A. V. Dereza // Cybernetics and Systems Analysis, 2012, – № 6. – P. 20–29. (in Russian).

[Lukyanova 4, 2012] . Lukyanova E. A. On the Relationship between the Language of CN-Model with Component Junctions, and the Language of Detailed Petri Model of Parallel Distributed System/ E. A. Lukyanova // Visnyk of Kyiv Univ. im. Tarasa Shevchenka – 2012. **–** in press (in Ukrainian).

[Kapitonova, 1988] Kapitonova Y. V. The mathematical theory of design of computer systems / Y. Kapitonova, A. A. Litichevsky – Moscow: Nauka, 1988. – 294 p. (in Russian).

Authors' Information

Elena Lukyanova– *PhD, doctoral candidate of Kyiv Univ. im. Tarasa Shevchenka, Faculty of Cybernetics, Kyiv, Ukraine; Vorovsky St., Bl. 60, Flat 239, Simferopol, Crimea, 95053, Ukraine;*
e-mail: lukyanovaea@mail.ru
Major Fields of Scientific Research: Modelling, Simulation, Formal verification.

MODEL FOR ASTRONOMICAL DATING OF THE *CHRONICLE* OF HYDATIUS

Jordan Tabov

Abstract. *This article presents a 'soft' model for astronomical dating of the seven eclipses mentioned in the Chronicle of Hydatius. The information about them is used in the construction of the two main components of the model: 1) "Template" ("image" of the initial "septet" of eclipses, described by Hydatius), and 2) "Distance".*

We assume that some of the data in the Chronicle may be incorrect. This means that the Template is considered as a fuzzy image of those seven real eclipses described by Hydatius. We assume that the inaccuracy of the data is small, i.e. that the parameters (dates, days of the week, etc.) in the Template do not differ much from the corresponding parameters of the initial (real) septet of eclipses, but are in some sense "close" to them.

For a more precise definition of this "closeness" in the paper is proposed a formula for the "distance" from the Template to any septet of real (happened in the past) eclipses. It allows us to calculate and compare the distances from the Template to all septets of real eclipses and to choose several "closest" to the Template (at the shortest distance to it) septets of real eclipses.

These septets are the target of the procedure of dating in the proposed model; they should be subject to further analysis and individual comparisons to determine which one of them is the most likely prototype of the Template, and thus its most probable dating.

Keywords: *Chronicle* of Hydatius, astronomical dating, eclipses, soft model, fuzzy information.

ACM Classification Keywords: I.5 PATTERN RECOGNITION; I.5.1 Models

1 Hydatius, his Chronicle and the eclipses in it

According to [New Advent Catholic Encyclopedia, 2013], Hydatius lived after the middle of the V century. From 427 to about 468 he was a bishop of Lemica. His Chronicle (continuation of the Chronicle of St. Jerome) has reached us in a single almost full copy. Its first part, describing events from 379 to 427, is based on records by writers preceding him, while the second part contains his descriptions of events from 427 to 468, of which he was a contemporary.

The Chronicle of Hydatius has been published and commented many times. Two of its editions are considered as main: 1) in Migne, PL LI, 873-890 [Idatii, 1861, pp. 873-890], and LXXIV, 701-750 [Idatii 1879, pp. 701-750]; and 2) in "Mon. Ger. Hist.: Auct. Antiq. ", XI (ed. Mommsen), 13-36 [Hydatii, 1894, pp. 1-36]. Among the recent editions we should mention: [Vilar, 2004], [Muhlberger, 2006] and [Cardoso, 1995].

In the Chronicle there are brief notes about the events, ordered chronologically by year of the reign of the successive Roman emperors, starting with Theodosius I. The years are numbered and the numbers start from 1 for each of the emperors. Occasionally, year numbers appear, according to the "Era" of Eusebius and other calendars. For the eclipses, there are given the day and the month, and sometimes also the day of the week, in which they occurred.

By H-1619, H-1634, H-1845, H-PL and H-MGH we denote five editions of the Chronicle of Hydatius – respectively of 1619 [Idatii, 1619], 1634 [Idatii, 1634], 1845 [Idatii, 1845], in the series *Patrologia Latina* – [Idatii, 1861], [Idatii, 1879] and *Monumenta Germaniae Historiae* [Hydatii, 1894].

Fig. 1 represents a fragment of the beginning of the Chronicle in the edition H-1619. After the title with the name of the 39th Roman Emperor Theodosius the Chronicle starts with the first year – indicated by 1 – of his reign: "1 Theodosius …". In the right margin somebody has added by hand the number 379 – it denotes the AD year. It is important to note, that the number 379 is not present in the original text of the chronicle.


IDATII EPISCOPI
CHRONICON.

ROMANORVM XXXIX.
Theodoſius per Gratia-
num in conſortium regni
adſumptus, cum ipſo, &
Valentiniano iuniore re-
gnat annis XVII.
1 Theodoſius natione Hiſpanus, de
prouincia Gallæcia, ciuitate Cauca, à

Figure 1. Fragment from the beginning of the Chronicle in the edition H-1619. After the title (containing the name of the 39th Roman Emperor Theodosius) the Chronicle starts with the first year – indicated by 1 – of Theodosius' reign: "1 Theodosius …". In the right margin somebody has added by hand the number 379 – it denotes the year of AD.

Another fragment from the beginning of the Chronicle in the edition H-1619 is shown in **Fig. 2.** The text is related to the second year – indicated by 2 – of the reign of Theodosius. Here, in the right margin, the year of AD 380 is written by hand, and the year CCXC is printed according to a reckoning "by Olympiads", indicating in this way the beginning of the 290th four year cycle.

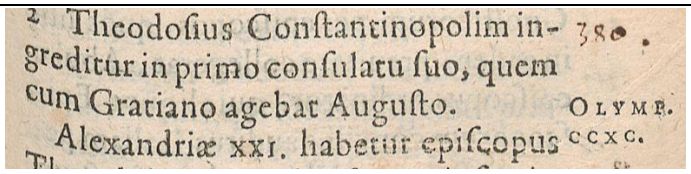

2 Theodoſius Conſtantinopolim in-
greditur in primo conſulatu ſuo, quem
cum Gratiano agebat Auguſto.
Alexandriæ XXI. habetur epiſcopus

Figure 2. Fragment of the first page of the Chronicle in the edition H-1619 with a text about the second year of the reign of Theodosius. In the right margin the year of AD 380 is written by hand, and the year CCXC is printed, according to a reckoning "by Olympiads", indicating in this way the beginning of the 290th four-year cycle.

An Olympiad is a period of four years associated with the Olympic Games of the Ancient Greeks. During the Hellenistic period it was used as a calendar epoch. It was generally agreed that the first Olympic games had happened during the summer of 776 BC [Bickerman, 1980, p. 75]. Hence, by this reckoning, the first Olympiad lasted from the summer of 776 BC to that of 772 BC [Olympiad, 2013].

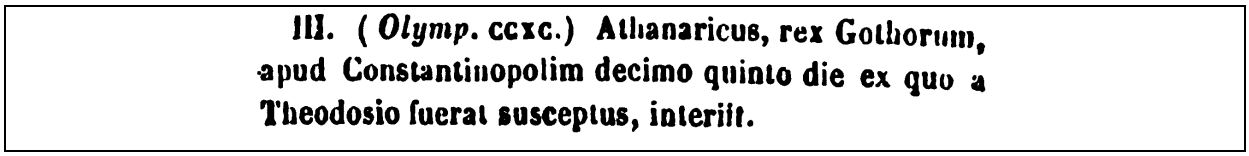

III. (*Olymp.* CCXC.) Athanaricus, rex Gothorum, apud Constantinopolim decimo quinto die ex quo a Theodosio fuerat susceptus, interiit.

Figure 3. Fragment of the Chronicle in H-PL with a text about the third year of the reign of Theodosius. The notation *Olymp. CCXC* is in brackets.

According to the text in H-PL (**Fig. 3**), the beginning of the CCXC (290th) four-year cycle is not in the second, but in the third (here the number 3 is written in Latin digits) year of the reign of Theodosius. The dating *Olymp. CCXC* is in brackets, which means, that it is not a part of the original text.

Also, in the third year of Theodosius is the beginning of the CCXC (290th) four-year cycle "by Olympiads" in H-MGH (**Fig. 4**). There are no brackets here and therefore the reader could be misled. The supposed respective AD year is given as a hint in the right margin.


OLYMP. CCLXXXX.

6. 'III. *Aithanaricus rex Gothorum aput Constantinopolim XV die, ex quo a Theo-* 381
'*dosio fuerat susceptus, interiit.*'

Figure 4. Fragment of the Chronicle in H-MGH with a text about the third year of the reign of Theodosius.

Another chronological guide is put **(Fig. 5**) in the margin of the fragment of the text of the Chronicle in [H-1619, p. 16]: for the 29th year of the reign of Arcadius and Honorius no occurrences are mentioned, but in the margin by hand is put the respective year AD. For the 30th year in the margin are printed:

- the year according to the "Era of Abraham": II.CCCCXL, i.e. 2440, and
- the year by Olympiads: Olymp. CCCI, i.e. [the beginning of] 301st Olympiad.

By hand is added the AD year: 424, and the comment that, according to the Chronicle of Marcelline, the year is 423.

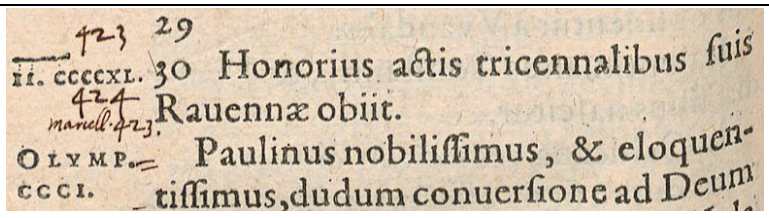
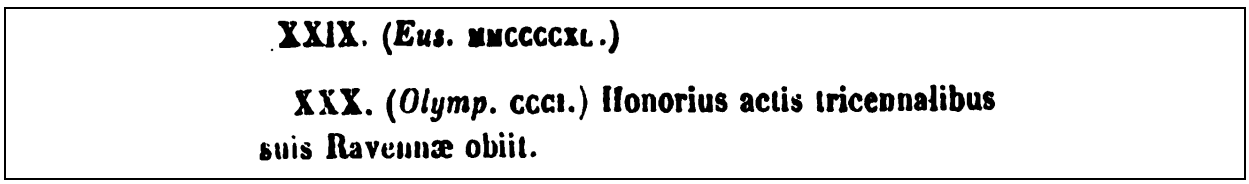

423 29
II. CCCCXL. 30 Honorius actis tricennalibus suis
424
Marcell. 423. Rauennæ obiit.
OLYMP. Paulinus nobiliſſimus, & eloquen-
CCCI. tiſſimus, dudum conuerſione ad Deum

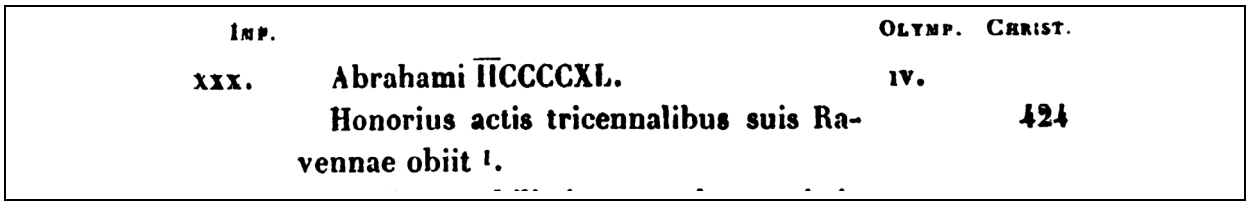
Figure 5. Fragment of the Chronicle in H-1619 with a text about the 29th and 30th year of the reign of Arcadius and Honorius.


XXIX. (*Eus.* MMCCCCXL.)

XXX. (*Olymp.* CCCI.) Honorius actis tricennalibus suis Rauennæ obiit.

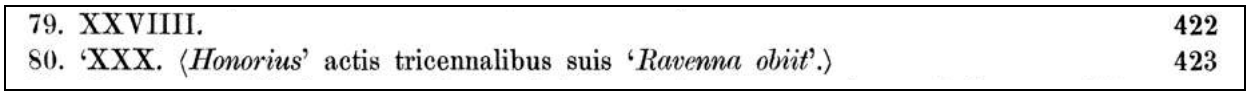
Figure 6. Two fragments of the Chronicle in H-PL with a text about the 29th and 30th year of the reign of T Arcadius and Honorius.

In H-PL 2440th year by Abraham started in the 29th year of Arcadius and Honorius (**Fig. 6**), in H-1845 – in the 30th (**Fig. 7**). In H-MGH the guiding mentioning of the 2440th year by Abraham is missing (**Fig. 8**).



Inp.		Olymp.	Christ.
xxx.	Abrahami IICCCCXL.	iv.	
	Honorius actis tricennalibus suis Ravennae obiit [1].		424

Figure 7. Fragment of the Chronicle in H-1845 with a text about the 29th and 30th year of the reign of Arcadius and Honorius.

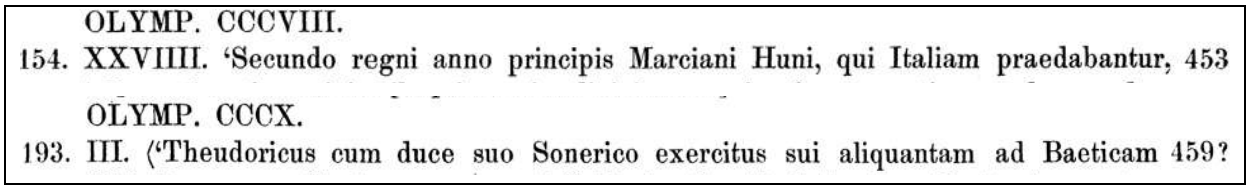


79.	XXVIIII.	422
80.	'XXX. (*Honorius*' actis tricennalibus suis '*Ravenna obiit*'.)	423

Figure 8. Fragment of the Chronicle in H-MGH with a text about the 29th and 30th year of the reign of Arcadius and Honorius.

In the various editions of the Chronicle the text of the manuscript is "supplemented" by different pieces of information from other sources, whereas the publisher's intent was to help the readers to relate the story told by the author to the most important events of the times. Often a detailed "chronological net" is given for the years in various calendars.

However, the additional information sometimes is in contradiction with the data within the chronicle and with the astronomical dating of the eclipses mentioned there. For example, according to the construction in H-MGH it turns out that the first years of the 308th and 310th Olympiad should correspond to AD 453 and 459. This is impossible because the interval between them would not be 459-453 = 6 years (**Fig. 9**), but 8 years. The Editor of H-MGH (Momsen) has noticed this and has put a question mark (**Fig. 9**).



	OLYMP. CCCVIII.	
154.	XXVIIII. 'Secundo regni anno principis Marciani Huni, qui Italiam praedabantur,	453
	OLYMP. CCCX.	
193.	III. ('Theudoricus cum duce suo Sonerico exercitus sui aliquantam ad Baeticam	459?

Figure 9. Two fragments from the Chronicle of Hydatius in H-MGH; according to the Editor, the first year of the 308th Olympiad and the first year of the 310th Olympiad correspond to AD 453 and AD 459, respectively.

All these examples show that some of the details in the Chronicle – e.g. the intervals of time between eclipses – perhaps are not exact, maybe due to misinterpretation or for other reasons. But which of them exactly? The answer to this question must be consistent with the analysis of the whole "astronomical picture" in the Chronicle, and it is natural to expect it to be part of the goals set by this model and its investigation. Either way, it is appropriate to focus only on the text of the Chronicle and on the data contained in it.

2 Template of the eclipses in the Chronicle of Hydatius

Seven eclipses are described in the chronicle of Hydatius: five solar and two lunar. In the "List of Ginzel" [Ginzel, 1899] they bear numbers 59-65; here they will be numbered consecutively from 1 to 7. The respective texts in the Chronicle (as in H-PL) are:

Eclipse **H1.** "Solis facta defectio tertio idus Novembris feria secunda."

Eclipse **H2.** "Solis facta defectio die decimo quarto kal. Augusti, qui fuit quinta feria."

Eclipse **H3.** "Solis facta defectio die nono kal. Januarias, qui fuit tertia feria."

Eclipse **H4.** "Quinto kal. Octobris a parte Orientis luna fuscatur."

Eclipse **H5.** "Quinto idus Junias die, quarta feria, ab hora quarta in horam sextam, ad speciem lunae quintae vel sextae, sol de lumine orbis sui minoratus apparuit."

Eclipse **H6.** "In provincia Gallaecia prodigiorum videntur signa diversa. Aera D, VI nonas Martias pullorum cantu, ab occasu solis luna in sanguinem plena convertitur. Idem dies sexta feria fuit."

Eclipse **H7.** "Decimo tertio kalend. Augusti die, secunda feria, in speciem lunae quintae sol de lumine suo ab hora tertia in horam sextam cernitur minoratus."

From these texts we obtain the following parameters of the eclipses (date = month and day, day of the week, time and phase):

H1 Solar eclipse on November 11, Monday.

H2 Solar eclipse on July 19, Thursday.

H3 Solar eclipse on December 24, Tuesday.

H4 Lunar eclipse on September 27. It was seen in the East (Eastern parts of the Empire) and was not seen in the West.

H5 Solar eclipse on June 9, Wednesday. Time – from the 4th hour till the 6th hour. Phase – about 0.7 – 0.8 (like 5- or 6- day moon).

H6 Lunar eclipse on March 2, Friday.

H7 Solar eclipse on July 20, Monday. Time – from the 3^{d} hour till the 6th hour. Phase – about 0.7 – 0.8 (like 5-day moon).

- **Comments for the parameters – why fuzzy?**

Now the day and the night together have 24 hours and the length of the hours is constant (does not change). In the summer, the day lasts longer than 12 hours, and the night is shorter; in the winter is the opposite.

But in the past it was not so.

The day was divided into two halves, light and dark, rather than in two twelve-hour periods; the light half was divided in 12 hours, as well as the dark one [Stephenson, 1997, p. 381]. Thus, in winter, an hour would be longer at night than during the day. Their length remained constant throughout the day, but changed from day to day; it could be said that the hours were at the same time "seasonal".

The days begun with the sunrise or (as e.g. in the Jewish calendar) with the sunset [Stephenson, 1997, p. 381]. This was so probably because people measured time with sundials. The variable length of the hours is embarrassing for accurate astronomical calculations, so together with the improvement of the methods and the instruments for measuring time, the astronomers begun using fixed-length hours, which gradually entered in the

daily life. The starting point for counting the hours moved at midnight or in the middle of the day, and this led to the emergence of other beginnings of the day and night.

But if the beginning of the day in one calendar is shifted with respect to the beginning of the day in another calendar, the transition from the one to the other is ambiguing, because the day in one includes parts of two days of the other. This is true also for different versions of the Julian Calendar.

May be namely the different beginnings of the days have caused the formal "shifts" in one day, which we sometimes meet in the old documents. As an example consider the following record from Braunschweig for a total solar eclipse, which occurred (according to the modern standard astronomical Julian calendar) on June 16 1406:

"1406. In this year there was an eclipse of the Sun so that the Sun stopped shining (vorgingk or schyn) before the Prime of the day (i.e. the Office held c. 6 a.m.) on St. Vitus' day (Jun 15); it was so dark that people could not recognize one another." (*Bothonis Chronicon Brunsvicensis picturatum*: [Stephenson, 1997, p. 405]). From this it becomes clear that if a historical source of information gives for an eclipse the date June 15, in the modern astronomical literature its date could be June 16.

Other similar examples are considered in [Stephenson, 1997, p. 406-407].They illustrate a part of the difficulties, which we meet when trying to determine the exact dates of the eclipses, mentioned or described in old chronicles and documents. The possibility that some dates mentioned in the Chronicle of Hydatius are shifted in one or two days with respect to the modern astronomical Julian calendar is taken into account in the construction of the model presented here.

- **Template Description:**

  1. Parameters of the eclipses mentioned in the Chronicle of Hydatius:

  Parameters of the eclipse H1 according to the Chronicle

  1-1. Solar eclipse. Date of the eclipse - November 11.

  1-2. Day of the week on which the eclipse occurred - Monday.

  Parameters of the eclipse H2 according to the Chronicle

  2-1. Solar eclipse. Date of the eclipse – July 19.

  2-2. Day of the week on which the eclipse occurred - Thursday.

  Parameters of the eclipse H3 according to the Chronicle

  3-1. Solar eclipse. Date of the eclipse - December 24.

  3-2. Day of the week on which the eclipse occurred - Tuesday.

  Parameters of the eclipse H4 according to the Chronicle

  4-1. Lunar Eclipse. Date of the eclipse - September 27.

  4-2. Seen in the Eastern part of the Empire (Constantinople), and

  4-3. Not seen in the Western parts of the empire (Spain).

  Parameters of the eclipse H5 according to the Chronicle

  5-1. Solar eclipse. Date of the eclipse - June 9.

  5-2. Day of the week on which the eclipse occurred - Wednesday.

  Parameters of the eclipse H6 according to the Chronicle

  6-1. Lunar Eclipse. Date of eclipse - March 2.

6-2. Day of the week on which the eclipse occurred - Friday.

Parameters of the eclipse H7 according to the Chronicle

7-1. Solar eclipse. Date of the eclipse - July 20.

7-2. Day of the week on which the eclipse occurred - Monday.

2. Intervals between the eclipses

Eclipses **H1** and **H2** occurred during the reign of Emperors Arcadius and Honorius: respectively in the IX year of reign, and in the XXIV. Thus, the interval between these two eclipses is approximately 16 years.

Eclipses **H3** and **H4** occurred during the reign of Emperor Theodosius: during the XXIII year of reign, and during the XXVIII year. Thus, the interval between these two eclipses is (about) 5 years.

Eclipses **H6** and **H7** occurred during the reign of Emperor Norton: respectively in the I and the II year. Thus, the interval between these two eclipses is approximately 1 year.

The reigns of Emperors bearing consecutive numbers in the List of the Roman Emperors sometimes follow directly one after another with a small interval of days or months between them, but sometimes the interval is longer. "Overlapping" of the reigns also could not be excluded. So the other three intervals between successive eclipses between **H2** and **H3**, **H4** and **H5**, and H6 and H& may vary more widely. Formally the data in the Chronicle give the following very rough approximations:

- Between **H2** and **H3** - about 29 years;
- Between **H4** and **H5** - about 7 years;
- Between **H5** and **H6** - about 5 years.

3. Calendar and other astronomical data:

In the year of the eclipse **H5** Easter was on March 28 (**Fig. 10**).

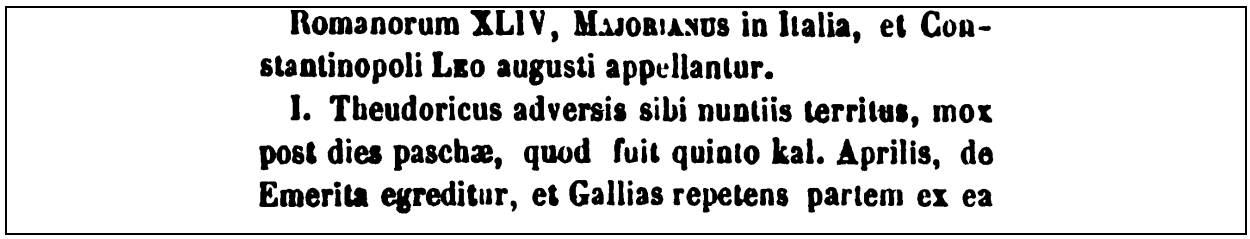

Romanorum XLIV, Majorianus in Italia, et Constantinopoli Leo augusti appellantur.
I. Theudoricus adversis sibi nuntiis territus, mox post dies paschæ, quod fuit quinto kal. Aprilis, de Emerita egreditur, et Gallias repetens partem ex ea

Figure 10. The text (in X-PL) for the first year of the reign of Emperor Majorianus, in which it is mentioned that Easter was on the fifth day of the April calends (March 28); the note about the eclipse **H5** is several lines after it. [Stephenson 1997, p. 406-407].

3 Distances from the Template to septets of real eclipses

Let $G_E = \{E_1, E_2, \ldots, E_7\}$ be a set of seven eclipses which occurred in the past.

How much does this set "differ" from the set of eclipses G_H = {**H1**, **H2**, …, **H7**} – differ in the astronomical parameters described above for the the set G_H?

To answer this question, we suggest a "metric" that models "closeness" of (and the "distance" between) individual eclipses and groups of eclipses respectively to **H1, H2, ..., H7** and $\mathbf{G_H}$. It is important for the application of the "template" described above in dating ancient eclipse and in particular for the evaluation of the "closeness" of $\mathbf{G_E}$ to $\mathbf{G_H}$.

Let **E** be any eclipse.

We start with defining rules for giving "scores" for the "closeness" of **E** respectively to each one of the eclipses **H1 - H7**.

Let ***m*** be a fixed positive number.

Scores for evaluation of the closeness of E to the eclipse H1:

The total score $\mathbf{e_1}$ for the closeness of an eclipse **E** to the eclipse **H1** is the sum of the scores for **1-1** and **1-2**:

1-1 The date of the eclipse H1 is November 11th. If the date of E is

- November 11 => score for **1-1:** ***m*** points;
- November 10 or 12 => score for **1-1:** 0.9***m*** points;
- November 9 or 13=> score for **1-1:** 0.5***m*** points;
- Another day => score for **1-1:** 0 points.

1-2 If the score for **1-1** is 0 points, the score for **1-2** is also 0 points; if the score for **1-1** is different from 0, the score for **1-2** is determined by the following rules. Taking into account, that the day of the week on which the eclipse **H1** occurred is Monday, if the day of the week on which occurred **E** is

- Monday => score for **1-2:** 0.5***m*** points;
- Tuesday or Sunday => score for **1-2:** 0.4***m*** points;
- Wednesday or Saturday => score for **1-2:** 0.3***m*** points;
- Another day => score for **1-2:** 0 points.

The rules for the determination of the scores for the closeness to the other four solar eclipses - **H2, H3, H5** and **H7** – are omitted, because they are completely analogous to that for the case of **H1**; different are only the dates and the corresponding days of the week.

The rules for the determination of the scores for the closeness of a lunar eclipse **E** to the lunar eclipses **H4** and **H6** are different:

Scores for evaluation of the closeness of E to the eclipse H4:

The total score $\mathbf{e_4}$ for the closeness of a lunar eclipse **E** to the lunar eclipse **H4** is the sum of the scores for **4-1, 4-2** and **4-3**:

4-1 The date of the eclipse **H4** is September 27. If the date of **E** is

- September 27 => score for **4-1:** ***m*** points;
- September 26 or 28 => score for **4-1:** 0.9***m*** points;
- September 25 or 29 => score for **4-1**: 0.5***m*** points;
- Another day => score for **4-1**: 0 points.

4-2 If ***E*** was seen in the Eastern part of the Empire (Jerusalem),

the score for **4-2** is 0.5***m*** points, otherwise 0 points.

4-3 If ***E*** was not seen in the Western part of the Empire,

the score for **4-3** is ***m*** points, otherwise 0 points.

If the score for **4-1** is 0 points, the scores for **4-3** and **4-2** are also 0 points.

Scores for evaluation of the closeness of E to the eclipse H6:

The total score $\mathbf{e_6}$ for the closeness of a lunar eclipse **E** to the lunar eclipse **H5** is the sum of the scores for **6-1** and **6-2**:

6-1 The date of the eclipse **H6** is March 2. If the date of E is

- March 2 => score for **6-1:** ***m*** points;
- March 1 or 3 => score for **4-1:** 0.9***m*** points;
- February 28/29 or March 4 => score for **6-1:** 0.5***m*** points;
- Another day => score for **6-1:** 0 points.

If the score for **6-1** is 0 points, the score for **6-2** is also 0 points; if the score for **6-1** is different from 0, the score for **6-2** is determined by the following rules. Taking into account, that the day of the week on which the eclipse **H6** occurred is Friday, if the day of the week on which occurred **E** is

- Friday => score for **6-2:** 0.5***m*** points;
- Wednesday or Friday => score for **6-2:** 0.4***m*** points;
- Tuesday or Saturday => score for **6-2:** 0.3***m*** points;
- Another day => score for **6-2:** 0 points.

Scores for evaluation of the lengths of time intervals between the eclipses $E_1, E_2, \dots, E_7$

Let ***n*** be a fixed positive number.

The intervals are in years and are equal to the differences between the years (in the Julian calendar) in which the respective eclipses occurred.

Denote by f_i the score for the closeness of the interval between E_i and E_{i+1} to the interval between H_i and H_{i+1}.

If the interval between E_1 and E_2 is:

16 years => $f_1 = n$ points, 15 or 17 years => $f_1 = 0{,}9n$ points, 14 or 18 years => $f_1 = 0{,}4n$ points, in other cases $f_1 = 0$ points.

The rules for calculation of the scores f_3 and f_6 are similar.

If the interval between E_2 and E_3 is:

29 years => $f_1 = 0.2n$ points, 28 or 30 years => $f_1 = 0{,}1n$ points, in the other cases $f_1 = 0$ points.

The rules for calculation of the scores f_4 and f_5 are similar.

The uncertain length of the intervals between the successive eclipses H_2 and H_3, H_4 and H_5, and H_5 and H_6 create additional difficulties for adequate evaluation of the "closeness" of G_E to G_H. More significant deviations of the interval between E_i and E_{i+1} from the interval between H_i and H_{i+1} in more than two cases should be subject of a special attention.

Scores for Easter in the year of E_5

Let **p** be a fixed positive number; by g denote the score for Easter in the year of E_5 .

If in the year of E_5 Easter was on

- March 28 => $g = p$;
- March 27 or 29 => $g = 0{,}9p$;
- March 26 or 30 => $g = 0{,}5p$;
- Another day => $g = 0$.

4 Closeness of a set G_E of 7 eclipses to the septet G_H

Let $G_E = \{E_1, E_2, \ldots, E_7\}$ be a set of seven eclipses. We define a "distance" of G_E to the septet $G_H = \{H1, H2, \ldots, H7\}$ ("of Hydatius") by the astronomical parameters described above for the group G_H.

We define the "distance" ***d*** from G_E to G_H in the following way:

$$d = 11{,}5\,m + 3{,}6\,n + p - (e_1 + e_2 + \ldots + e_7 + f_1 + f_2 + \ldots + f_6 + g).$$

It is easy to check that in case of coincidence of the respective parameters for the eclipses of G_E and G_H the distance ***d*** is equal to 0. The less is ***d***, the "closer" is the septet G_E to G_H.

5 Searching for the closest (to G_H) septet of eclipses G_E

The suggested formula for calculation of the distance from G_E to G_H is an essential part of our model for astronomical dating. It is natural to combine it with different methods for determination of the closest to G_H "septets" of eclipses in a given historical period – for example, in the time interval from AD 300 to AD 600.

A brief description of a possible approach how to "search" for suitable "septets" at shortest distance from G_H is given below.

1) We reduce the list of all eclipses of the period (assuming that this is the interval from AD 300 to AD 600) to its part **L,** containing only the eclipses visible from the Mediterranean region.

2) We select from this list **L** seven sets of eclipses G^1, G^2, … , G^7*:* the set G^1 contains only those eclipses of **L**, whose date is "around the date of **H1**", i.e. about November 11, and more precisely, in the framework of the proposed Template, on the days from 9 to 13 November inclusive. Similarly, we select the other sets G^2, G^3,…, G^7.

3) We determine the number of the sets among G^1, G^2,… ,G^7, having at least one element (eclipse) in the interval that starts from 300 and has a length of 100, i.e. in the interval (300, 400).

4) If the number determined in 3) is seven, we consider all possible septets of eclipses $G_E = \{E_1, E_2, \ldots, E_7\}$ in the same interval (300, 400), where E_i is from G^i for i = 1,2, …, 7. For each such septet we calculate its distance ***d*** to G_H and choose several such septets with smallest ***d***.

5) We replace successively the interval (300, 400) by the intervals (301, 401), (302, 402), etc. and to each of them apply 3) and 4).

6) If the number of 3) is 6, we apply 4) and 5) for sextets (sets of six) of eclipses G_E, replacing by 0's the scores of the "missing eclipse" (of one of the sets G^1, G^2, … , G^7) in the formula for ***d***.

The determined in this way septets and sextets of eclipses are "candidates" for sets of eclipses that are closest to G_H and therefore are "astronomically probable" datings of G_H in the interval from AD 300 to AD 600.

6 Distance from the traditional dating of the eclipses in the *Chronicle* to the Template G_H

The seven eclipses of the *Chronicle* of Hydatius are contained in the famous "List of Ginzel" [Ginzel, 1899] of all "antient" eclipses mentioned and described in the major historical documents. There they are numbered successively from 59 to 65 and are dated as follows:

№ 59: **402-Nov-11** (November 11, AD 402)

№ 60: **418-Jul-19**

№ 61: **447-Dec-23**

№ 62: **451-Sep-26**

№ 63: **458-May-28**

№ 64: **462-Mar-2**

№ 65: **464-Jul-20**

At what extent does this dating satisfy the astronomical information about the seven eclipses in the Chronicle of Hydatius?

Or, in other words, how far (at what distance) is this dating from $\mathbf{G_H}$?

To give an appropriate answer to the last question we will calculate the distance ***d*** from the septet of eclipses № 59, № 60, № 61, № 62, № 63, № 64, № 65 to $\mathbf{G_H}$.

We put

$\mathbf{E_1 = 402\text{-}Nov\text{-}11\ ;\ E_2 = 418\text{-}Jul\text{-}19\ ;\ E_3 = 447\text{-}Dec\text{-}23\ ;\ E_4 = 451\text{-}Sep\text{-}26\ ;}$

$\mathbf{E_5 = 458\text{-}May\text{-}28;\ E_6 = 462\text{-}Mar\text{-}2\ ;\ E_7 = 464\text{-}Jul\text{-}20.}$

First we calculate the the scores $e_1, e_2, \ldots, e_7, f_1, f_2, \ldots, f_6$ and ***g:***

The total score e_1 of the closeness of an eclipse of **E** to **H1** equals the sum of the scores for **1-1** and **1-2**.

1-1 The date of the eclipse **H1** is November 11. Since the date of $\mathbf{E_1}$ is also November 11, the score for **1-1** is ***m*** points;

1-2 The score for 1-1 is different from 0; the day of the week on which the eclipse **H1** occurred is Monday, and the day of the week on which occurred $\mathbf{E_1}$ is Tuesday; therefore the score for **1-2** is 0.4***m*** points.

Hence $e_1 = m + 0{,}4m = 1.4m$.

Similarly

$e_2 = m + 0{,}4m = 1{,}4m$

$e_3 = 0{,}9m + 0{,}5m = 1{,}4m$

$e_4 = 0{,}9m + 0{,}5m = 1{,}4m$

$e_5 = 0$

$e_6 = m + 0{,}5m = 1{,}5m$

$e_7 = m + 0 = m$,

and therefore

$$e_1 + e_2 + e_3 + e_4 + e_5 + e_6 + e_7 = 8{,}1m.$$

Further

$f_1 = n$

$f_2 = 0{,}2n$

$f_3 = 0{,}9n$

$f_4 = 0{,}2n$

$f_5 = 0{,}1n$

$f_6 = 0{,}9n$

$f_2 = 0{,}2n,$

hence

$$f_1 + f_2 + f_3 + f_4 + f_5 + f_6 = 3{,}3n.$$

Finally, since in 458 Easter was on April 20, while according to the Chronicle in the year of the eclipse **H5** Easter was on March 28,

$$\boldsymbol{g} = 0.$$

Now for the distance $\boldsymbol{d}$ from $\mathbf{G_E}$ to $\mathbf{G_H}$ we obtain

$$\boldsymbol{d} = 11{,}5\boldsymbol{m} + 3{,}6\boldsymbol{n} + \boldsymbol{p} - (\mathbf{e}_1 + \mathbf{e}_2 + \ldots + \mathbf{e}_7 + \boldsymbol{f}_1 + \boldsymbol{f}_2 + \ldots + \boldsymbol{f}_6 + \boldsymbol{g})$$
$$= 11{,}5\boldsymbol{m} + 3{,}6\boldsymbol{n} + \boldsymbol{p} - (8{,}1\boldsymbol{m} + 3{,}3\boldsymbol{n} + 0)$$
$$= 3{,}4\boldsymbol{m} + 0{,}3\boldsymbol{n} + \boldsymbol{p}.$$

For $\boldsymbol{m} = \boldsymbol{n} = \boldsymbol{p} = 20$ we have $\boldsymbol{d} = 94$. Because

$$\boldsymbol{d}_{max} = 11{,}5\boldsymbol{m} + 3{,}6\boldsymbol{n} + \boldsymbol{p} = 230 + 72 + 20 = 322,$$

then $\boldsymbol{d} = 94$ equals about 28.3% of $\boldsymbol{d}_{max}$.

7 The problem H5

The zero score $\mathbf{e_5} = 0$ is impressive. Could this score be a result of incorrect dating of the eclipse **H5** (**№ 63** in the List of Ginzel)?

This question is interesting, since an analysis of the eclipses in the second half of the V century suggests the hypothesis that from astronomical point of view there is another preferable eclipse for the role of **H5**: the eclipse **476-Jun-07**.

Let us put $\mathbf{E^1_5}$ **= 476-Jun-07** and consider the septet $\mathbf{G^1_E} = \{\mathbf{E_1, E_2, E_3, E_4, E^1_5, E_6, E_7}\}$ instead of $\mathbf{G_E} = \{\mathbf{E_1, E_2, E_3, E_4, E_5, E_6, E_7}\}$. Most of the scores involved in the formula for the distance $\boldsymbol{d}^1$ between $\mathbf{G^1_E}$ and $\mathbf{G_H}$ coincide with the respective scores already calculated for the distance $\boldsymbol{d}$ between $\mathbf{G_E}$ and $\mathbf{G_H}$. Different are only

$\mathbf{e^1_5} = 0{,}5\boldsymbol{m} + 0{,}3\boldsymbol{m} = 0{,}8\boldsymbol{m}$

$\boldsymbol{f}^1_4 = 0$

$\boldsymbol{f}^1_5 = 0$

$\boldsymbol{g}^1 = \boldsymbol{p}.$

Substituting in the formula we get

$$\boldsymbol{d}^1 = 2{,}6\boldsymbol{m} + 0{,}6\boldsymbol{n}.$$

The choice $\boldsymbol{m} = \boldsymbol{n} = \boldsymbol{p} = 20$ leads to $\boldsymbol{d}^1 = 64$.

The distance $\boldsymbol{d}^1$ is significantly less than the distance $\boldsymbol{d}$; from the point of view of the defined above distance the septet of eclipses $\mathbf{G^1_E}$ (which includes the eclipse **476-Jun-07** instead of **458-May-28**) is substantially closer to $\mathbf{G_H}$, i.e. to the basic parameters of the seven eclipses of the *Chronicle*, than $\mathbf{G_E}$.

This numerical result reflects an important substantive fact: that the date of Easter in the year of the eclipse $\mathbf{E^1_5}$ **= 476-Jun-07**, coincides with the date March 28, mentioned in the *Chronicle* as the date of Easter in the year of the eclipse **H5**. It provides a strong argument in favor of the hypothesis that the $\mathbf{E^1_5}$ is a better candidate than $\mathbf{E_5}$ for the role of **H5**, in the sense that the $\mathbf{G^1_E} = \{\mathbf{E_1, E_2, E_3, E_4, E^1_5, E_6, E_7}\}$ better than $\mathbf{G_E}$ meets the main parameters of the eclipses of the *Chronicle*.

The coincidence of the year of an eclipse "around June 9" (i.e. in the range from June 7 to 11) with a year of "Easter on March 28" is a very rare event. Therefore, if the other six eclipses of the *Chronicle* Hidatsiy are dated correctly (in $\mathbf{G_E}$), the probability that **H5** is a description of the later eclipse $\mathbf{E^1_5}$ **= 476-Jun-07** is rather high.

Is such a thing possible?

At a first glance it seems that the answer is clearly negative. A reason for it is the chronological sequence of the events in the Chronicle.

But it could happen that in the extant copy of the *Chronicle* the chronological order is not correct: that it has been changed, may be occasionally, for instance, if 2-3 leaves of the manuscript of the Chronicle, which have been in the end, were moved in the middle. Having in mind the corrupted form of the preserved manuscript of the *Chronicle*, such a possibility is not excluded.

But evidently a "time shift" of $\mathbf{E_5}$ in about 18 years would also cause a similar shift (with respect to the contemporary historical events) of the rule of Emperor Majoranus.

Thus it becomes clear that to answer the question whether the year of the eclipse **H5** is 476, additional research and analysis of the sources concerning the epoch of Emperor Majoranus are necessary.

8 Concluding remarks and acknowledgments

The datings of ancient eclipses are used in history; they are usually searched for in the range of 30-40 years, and rarely - if the eclipses are mentioned in ancient Shumerian, Babylonian and Egyptian sources – in intervals of several centuries.

The narrow range facilitates the search, and, which is more important, it reduces at a great extent the number of "solution" to reasonable figures.

For building the chronology of ancient history on the basis of exact sciences - in which astronomy should play a leading role - it is necessary to investigate the dating of eclipses in a large range, covering several hundreds to several thousands of years. In this case, the number of the possible "purely astronomical solutions" for a great number of particular historical eclipses could be great.

From this perspective, the task of searching for astronomical dating of a "group of eclipses," which occurred "close" to each other (in the time), has the advantage that the resulting number of "solutions" is much smaller.

Dating of "group of eclipses" has been carried out by N. Morozov in the 20-ies of the XX century, for the so called "Triad of Thucydides" [Morozov, 1928]. Half a century later, now with the help of computer, dating of the same triad was carried out by A. Fomenko [Fomenko, 1990, pp. 91-95].

It should be noted that the approaches of Morozov and Fomenko are "harder" than the one proposed here; it treats the information from the work of Thucydides as "entirely accurate."

I offer thanks to V. Umlenski, M. Nikiforov, V. Vatchkova, S. Velev, P. Petrov, K. Markov and A. Vasileva, who helped to shape the ideas about the "soft modeling" of the problem of the dating "the septet of Hydatius" and for the presentation of these ideas in the present paper.

Bibliography

[Bickerman, 1980] E.J.Bickerman. Chronology of the Ancient World (Aspects of Greek & Roman Life) (2nd sub ed.). Cornell University Press, Ithaca, NY, 1980.

[Cardoso, 1995] J.Cardoso. Crónica by Idatius, Bishop of Chaves. Livraria Minho, Braga, 1995.

[Fomenko, 1990] A.T.Fomenko. Methods of Statistical Analysis of Historical Texts and Applications to Chronology. Moscow University,1990. (In Russian)

[Ginzel, 1899] F.K.Ginzel. Spezieller Kanon der Sonnen- und Mondfinsternisse fur das Landregebiet der klassische Altertumswissenschaften und ben Zeitraum von 900 vor Chr. 600 bis nach Chr. von F.K.Ginzel Standigem Mitgliede des Konigl. Astronomische Recheninstitutes. Mayer & Muller, Berlin, 1899.

[Hydatii, 1894] Hydatii Lemici. Continuatio chronicorum Hieronymianorum. Monumenta Germaniae Historiae: Auct. Antiq., XI (ed. Mommsen). Apud Weidmannos, Berolini, MDCCCXCIV.

[Idatii, 1619] Idatii Episcopi Chronicon, et Fasti Consvlares. Lutetiae Parisiorum. Ex Officina Nivelliana. Apud Sebastianvm Cramoisy, via Iacobaea, sub Ciconiis, M.DC.XIX.

[Idatii, 1634] Idatii Episcopi Chronica. In: Prudencio de Sandoval. Historias de Idacio Obispo que escrivio poco antes que España se perdiese: De Isidoro Obispo de Badahoz, De Sebastiano Obispo de Salamanca , De Sampiro Obispo de Astorga, De Pelagro, Obispo de Quiedo. Impresso en Pamplona, por Nicolas de Assiayn Impressor del Reyno de Navarra, M.DC.XXXIV.

[Idatii, 1845] J.M.Garzon. Idatii Episcopi Chronicon. M. Hayez, Reg. Acad. Typographus, Bruxellis, 1845.

[Idatii, 1861] Idatii Episcopi Chronicon. Patrologia Latina *v. 51*, 1861.

[Morozov, 1928] N.Morozov. Christ. Book 4. Part III. Chapter III. Gosudarstvennoe Izdatelstvo, 1928. (In Russian)

[Muhlberger, 2006] S. Muhlberger. The fifth-century chroniclers : Prosper, Hydatius, and the Gallic Chronicler of 452. Francis Cairns, Cambridge, 2006.

[New Advent Catholic Encyclopedia, 2013] Hydatius of Lemica. New Advent Catholic Encyclopedia. Visited March 11 2013. http://www.newadvent.org/cathen/07592a.htm

[Newton, 1974] R.R.Newton. Two uses of ancient astronomy. Phil. Trans. R. Soc. Land. A. 276, 99-110 (1974).

[Olympiad, 2013] Olympiad. From Wikipedia, the free encyclopedia. Visited March 11 2013. [http://en.wikipedia.org/wiki/Olympiad].

[Stephenson, 1997] F.R.Stephenson, Historical Eclipses and Earth's Rotation, Cambridge University Press, 1997.

[Vilar, 2004] X.B.Vilar. Idacio Lémico chronica (376-469). CRP 06. Xunta de Galicia, Santiago de Compostela, 2004.

Authors' Information

Jordan Tabov *– Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Acad. G.Bonchev Str. Block 8, 1113 Sofia, Bulgaria; e-mail: tabov@math.bas.bg*

Major Fields of Scientific Research: Applications of mathematics and informatics in the humanities, Didactics of mathematics and informatics

CONNECTIVITY CONTROL IN AD HOC SYSTEMS: A GRAPH GRAMMAR APPROACH

Alexander Mikov, Alexander Borisov

Abstract*: We discuss the problem of connectivity within large-scale dynamic distributed information systems. Ad hoc system can adapt themselves through resource management or reconfiguration to achieve specific goals, such as functions, performance, energy budget and reliability. One of the mostly important goals is to keep a possibility of routing for any two nodes of the distributed system. Structure of the system can be described by a time-graph. In some moments of the time the structure of a network changes: new nodes can be included into the network and old nodes can be deleted. Thus the connectivity property of the time-graph is a logical function of time. A cause of disconnection may be technical (hardware failures, low energy, server overloading etc) or organizational (information secure, regular breaks or random interruptions). To improve the quality of service we have to control the distributed information system structure. The first step of a self-management (autonomic) system design is to describe permissible structures and forbidden structures. We propose to use the well-known method of graph grammars for this purpose. A finite set of graph grammar rules defines an infinite (but countable) set of permissible structures. An inference process allows us to get some permissible graph after doing of a finite sequence of steps. In this work we solve the problem of a set of graph grammar rules description for such property of graphs as connectivity. A detailed description of the rule set is used for rewriting of any graph (connected or disconnected) to a connected graph. Also we discuss a second step of a self-aware system design: including of the graph grammar into a feedback control cycle. An autonomic system has the property of self-awareness, i.e. the system contains its model and manages itself using this model. During a life-time cycle the current model can be compared with the permissible set described by graph grammar. If the current model isn't belong to the permissible set then the distributed system turns a connectivity renewal process on. Some problems of graph grammar algorithms complexity are discussed.*

Keywords*: time-graph, grammar, autonomic system, ad hoc system.*

ACM Classification Keywords*: H. Information Systems: H.3 Information Storage and Retrieval: H.3.4 Systems and Software – Distributed systems. C. Computer Systems Organization: C.2 Computer-Communication Networks: C.2.4 Distributed Systems – Distributed applications*.

Introduction

Autonomic computing is assumed solving problems (tasks) in self-managed computer environments. Such computer environment is based on local or global computer network, which functions almost without interference of the person – administrator of the network. Autonomic system is self-awareness, i.e. it contains its model and manages itself using this model. Frequently it is ad hoc network consisting in mobile nodes and which is created undertime (rather quickly) for a short time, when emergency situation is happened and so on. In this case self-management is very important. It can include self-generation, self-supporting in active functioned period and self-conservation at the completion of the network.

In this work some aspects of autonomic network structure modeling for the purpose of building software, which implements the model, for supporting self-managing algorithms, are under study.

Mobile networks graphs

At the generation of the network (time $t = 0$) specified links (relations) structure between computational nodes, which is defined by solving task (for example, placement of nodes and signal distance wireless connection) appears. It is natural that such structure can be represented by the graph $G(0) = G_0$. The graph $G(t)$ is changing in the process of functioning. In different times some nodes leave the system, but other nodes are included in it. Moving of nodes leads to increasing/decreasing distance between nodes and power of signal and respectively appearing new edge in graph $G(t)$ or removing existing edge [1]. Gradual reducing of battery charge also leads to decreasing of amount of edge in respective graph.

Commonly it is desirable (often necessary), that graph $G(t)$ has such property as permanent connectivity. Any pair of vertexes must be connected with at least one simple chain. The length of such chain is also important, because the quantity of hops has direct impact on latency of information transmission, i.e. qualitative characteristic of computer network. A routing task consists in searching of the shortest paths. Due to changing of graph $G(t)$ the routing task must be solved again after each change of the graph. Of course it can't be solved completely at the periods of disconnected graph [2].

So two conclusions can be made:

- not all graphs are acceptable and acceptable graphs have different qualitative characteristics;
- network must possess knowledge of its graph.

Networks, which belong to studied class, as a rule, are decentralized, that's why the question about placement of “knowledge” about graph is urgent. On the local level it is knowledge about neighbors (one-hop). Acquisition of local knowledge require the least cost, but it is difficult make a global decision (for example, routing) basing on local knowledge. The next level is knowledge about neighbors of neighbors (two hop), which is more difficult to receive and it is less reliable, because it quickly goes out of date. In general, the more knowledge is global the less it is reliable.

Graph acceptable is defined not only by connectivity, but also other characteristics of solving by the network task.

Label Ω as set of graphs. Graph $G(t)$ takes on a value from this set, $G(t) \in \Omega$, making in this set a trajectory. Based on the above the set Ω can be separated on two subsets: Ω_p – permissible and Ω_p – forbidden graphs of the networks [3].

Mobile networks grammars

The natural approach to defining sets of graphs with specified properties is graph grammars or graph rewriting systems. Grammar can be used for checking belonging of graph to the set $G(t) \in \Omega_p$, i.e. for solving the task of recognition.

Some approaches to graph rewriting systems exist. One of them is algebraic approach which based on the category theory. At the same time algebraic approach is divided to some approaches. Double-pushout approach (DPO) and single-pushout approach (SPO) are the most important.

Boolean algebra and matrix theory based approaches also exist and it is called matrix graph grammars.

Except of algebraic approaches, determine graph rewriting systems, which based on logic and database theory, can be separated out.

In this work algebraic approach – single-pushout approach (hereinafter referred to as SPO) – is studied.

In this approach rule p: ($L \rightarrow_r R$) consists of a rule name p and of an injective partial morphism r, called the rule morphism. Graphs L and R are called left-hand and right-hand side of the rule, respectively. Graph grammar GG is a pair $GG = \langle (p: r)_{p \in P}, S \rangle$, where $(p: r)_{p \in P}$ is a family of rule morphisms indexed by rule name, and S is the start graph of the grammar [4].

In practice classic graph grammar theory is not enough often. One of the additional mechanisms is application conditions. Application conditions allow setting the requiring context for application of the grammar rule. Application conditions separated to two classes: positive and negative application conditions. Positive constraints require existence of the elements (edges or vertexes) and vice versa, negative forbid the existence of some edges or vertexes [5].

Now describe graphical layout for representing application conditions. All constraints are distinguished by dotted border. If the area inside the dotted border is crossed by the line, it means that it is negative application condition else positive [6].

The next set of directives was developed to describe graph grammar in autonomic computer system software and also transform (edit) grammar:

- Directive CreateRule <identifier>.
- Directive

  Set <*rule identifier*> (*left* | *right*) *graph* <*graph identifier*>

  sets left or right graph in the rule depending on selected parameters (*left* or *right*).
- Directive *AddRule* <*rule identifier*> adds the rule to the grammar.
- Directive *RemoveRule* <*rule identifier*> removes rule.
- Directive

  AddConnectingRule <*rule identifier*>
  (*if* (*exist* | *absent*) *arc* <*vertex identifier* > (*in* | *out*) <*vertex identifier*>
  then make arc <*vertex identifier*> (*in* | *out*) <*vertex identifier*>)

  adds connecting (merging) rule in the rule.
- Directive AddConnectingRule <vertex identifier> (<vertex identifier> link <vertex identifier>) adds rule.
- Directive GetListRules shows the list of all rules (their identifiers) of the grammar.
- Directive

  Show (<*rule identifier*> | *startGraph*)

  shows either information about rule with specified identifier or start graph of the graph grammar.
- Directive RemoveConnectingRule <rule identifier> <order number of connecting rule> removes connecting rule.
- Directive SetStartGraph <graph identifier> sets the start graph.
- Directive SetName <identifier> defines name for the grammar.
- Directive Save saves grammar in the file.
- Directive Load <identifier> loads grammar from file (<identifier> sets filename without extension).
- Directive Derivate <depth of the derivation tree> builds derivation of the graph grammar with specified depth of the derivation tree to avoid circularity.

Representing grammars in software-based network models

The performance of the algorithms of self-management and program logic in the first place depend on representing basic data structures. So a program operates with graph grammars, i.e. data with complex structure, the problem of their representing became top priority task.

To implement graphs list representation was chosen. It has some advantages in comparison with matrix representations. First of all, it is rather quick way to remove and paste vertexes that are the one of the most basic actions for getting graph grammar derivation. The second advantage is memory saving, because matrixes are often sparse. After object-oriented decomposition we get three classes (*Graph*, *Node*, *Arc*) for representing such data structure as graph. By his nature graph grammars are more complex objects. It follows even from this fact that graphs are included in graph grammars. Three classes (*GraphGrammar*, *Rule*, *CRule*) can be assigned for representation these structures.

The basic class is GraphGrammar. It has some properties and methods:

```
class GraphGrammar
  {
    public string name = "";
    public Graph startGraph;
    Set <string> T, N;
    public List <Rule> rules = new List <Rule>();
    public List <Graph> derivation = new List <Graph>();
                    // methods
    }
```

Property *startGraph* contains link on start graph; *T* and *N* are alphabets of terminal and nonterminal labels respectively; field *rules* is a list of object of class *Rule*, which is described below. Graphs will be added in list derivation during the process of graph grammar derivation. It will be necessary in future for checking results of work of grammar.

Now describe class *Rule*:

```
class Rule
  {
    public string name = "";
    public Graph left;
    public Graph right;
    public List <CRule> crules = new List <CRule>();
  }
```

Left and right side graphs can be assigned as well as in mathematical representation of graph grammar. Also class *Rule* contains the list of merging rules (*connecting rules*), which are defined by class *CRule*.

```
class CRule
  {
    public string SourceId1;
    public int Dir1;
    public string TargetId1;
    public int State;
    public string SourceId2;
```

```
    public int Dir2;
    public string TargetId2;
    public string LinkSource;
    public string LinkTarget;
}
```

Every object of this class is a merging rule and its properties can be interpreted like this:

"*If State* (1-*exist*, 0-*absent*) *edge SourceId*1 *dir*1 (1-*from*, 0-*in*) *TargetId*1
*then add edge SourceId*2 *dir*2 (1-*from*, 0-*in*) *TargetId*2".

Here *TargetId*1 and *TargetId*2 are labels of vertexes which exist on right side of the rule, but *SourceId*1 and *SourceId*2 are labels, which absent on the right side of rule.

A merging rule can be defined by another way. For this purpose the properties *LinkSource* and *LinkTarget* exist. If some labels are bound with them then program will be try to connect pasted part with the rest graph by them; *LinkSource* is a vertex label in the pattern graph, and *LinkTarget* is a vertex label in the pasted graph. Therefore, if before changing the edge is connected with the vertex with label *LinkSource* then this edge must be also in vertex with the label *LinkTarget*. This way is less flexible, but more handy.

Class *GraphGrammar* also contains some methods including methods for graph grammar derivation, saving in file and loading from file and so on.

Procedures of work with grammars for the algorithms of self-management

One of the tasks of the self-management of the autonomic network is to check belonging current network graph $G(t)$ to the set of acceptable graphs $G(t) \in \Omega_p$. Since, the set of acceptable graphs Ω_p is specified by graph grammar, the solving of this task is based on algorithm of parsing.

To implement derivation of the graph grammar, three base algorithms are required: Subgraph isomorphism, replacing a found subgraph by a substitutional graph, an algorithm of graph grammar derivation.

```
flag = false
k = 0
ISOMORPH(∅)
if flag then return ( G_X ≈ G_Y, correspondence i ↔ f_i )
    else ( G_X ¬≈ G_Y )
procedure ISOMORPH(S)
    k = k+1
    if S = V_Y then flag = true
    for v ∈ V_Y \ S while not flag do
        if MATCH then { f_k = v; ISOMORPH(S ∪ {v}) }
    k = k+1
return
procedure MATCH
  [give out true, if vertex v ∈ V_y \ S can be matched with vertex k ∈ V_X]
return
```

Replacing algorithm

Let we have links on vertexes and edges, which are needed to replace. And let we have a graph, which should substitute. Then:

1. Save information about the edge relations of the removing part of the graph to the rest graph (or copy graph) and then remove all corresponding vertexes and edges.
2. Paste graph from the right side of the rule into the source graph.
3. Apply merging rules.

Merging rules can be applied as follows

1. Find in the copied (unchanged) graph the vertex with identifier *SourceId*1. This vertex mustn't be among the vertexes which are subject to remove. If such vertexes don't exist then stop the process.
2. Then find the vertex with identifier *TargetId*1 among removed vertexes (for this target also use copied graph). If such vertex wasn't found then the process go back to step 1.
3. Depending on *dir*1 and *State* check existence or absence of the edge between these vertexes.
4. If condition wasn't satisfied then go to step 2. Otherwise go to step 5.
5. Find in the copied (unchanged) graph the vertex with identifier *SourceId*2. This vertex mustn't be among the vertexes which are subject to remove. If such vertexes don't exist then stop the process.
6. Then find the vertex with identifier *TargetId*2 among the vertexes, which were pasted. If any vertex wasn't founded then the process go back to step 5.
7. Add an edge and go back to step 6.

In the case of a merging rule is specified by labels of the vertexes which are needed to be linked between each other, the merging rules application algorithm will be next:

1. Find the vertex with the label *LinkSource* in the unchanged graph. This vertex must belong to the set of vertexes which will be removed.
2. Find the vertex with label *LinkTarget* in the changed graph, and this vertex must be among those, which were added.
3. If any edge enters in the vertex with label *LinkSource* then add this edge to the finite graph, replacing vertex with the label *LinkSource* by vertex with the label *LinkTarget*.
4. Execute the process for the outgoing edges in much the same way.

As a result, after executing of the algorithm, the graph, which will became a part of procedure of the graph grammar derivation, included in the algorithm of recognition of belonging the autonomic network graph to the set of acceptable graphs, will be created.

Solving graph connectivity problem by graph grammars

As stated above, one of the problems which can be solved by graph grammars is the problem of graph connectivity. Graph grammar, making arbitrary undirected disconnect graph connected by minimum number of edges, is presented next. The source graph being used as a start graph of this graph grammar. All vertexes in the source graph must have identifier "*v*".

The first rule (figure 1) chooses vertex and make it connected (i.e. set it identifier "*c*"). This rule used only on the first step of the derivation. It can be removed if before the application of the graph grammar the identifier "*c*" will be set for a random vertex of the source graph.

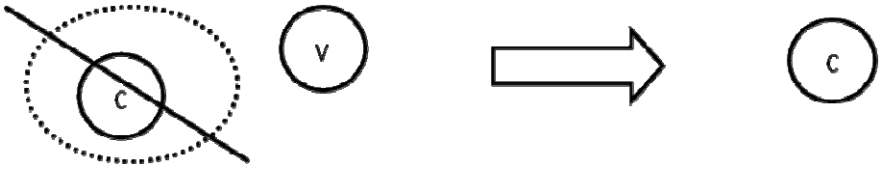


Figure 1: Starting rule of the graph grammar

The second rule (figure 2) labels adjacent vertex as connected. And when all except one vertexes of the connectivity component are labeled the third rule (figure 3) is applied. The identifier of the last vertex is set "*e*".

Figure 2: Rule for labeling connected vertexes

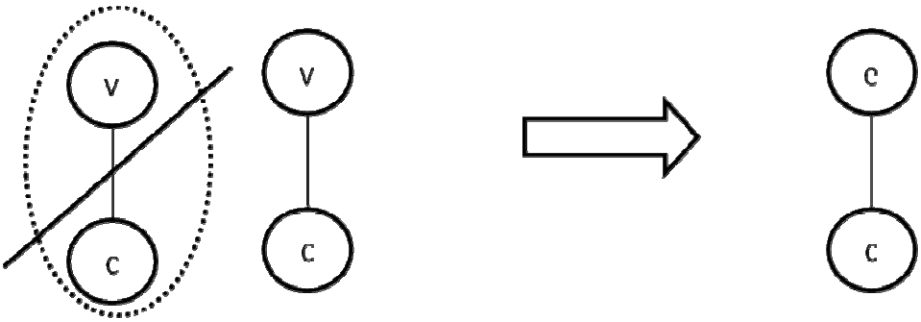


Figure 3: Rule for checking last vertex of the connectivity component

When all vertexes of the connectivity component are passed the fourth rule (figure 4) makes the edge from vertex with identifier "*e*" to any isolated vertex. And the fifth rule (figure 5) makes such edge from any vertex with identifier "*c*" (it's necessary to give all possible connected graphs). Using identifier "*t*" is caused by the necessary to avoid ambiguity in describing of merging rules. This identifier is temporary and replaced by the sixth rule (figure 6).

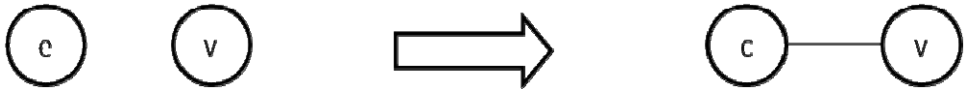


Figure 4: Adding edge to isolated vertex (from last vertex)

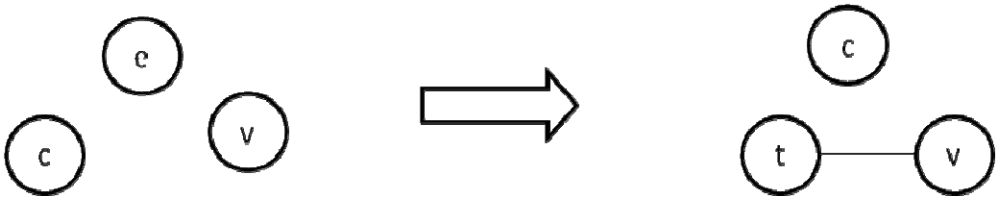


Figure 5: Adding edge to isolated vertex (from any vertex)

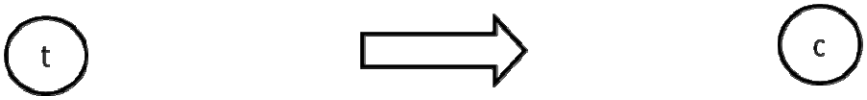


Figure 6: Renaming the vertex

This graph grammar can't give correct results in the case of empty (fully disconnected) graphs. That's why the seventh rule (figure 7) was designed. This rule will apply only if there is now any vertex with identifier "*c*", which connected with any other vertex, in the graph.

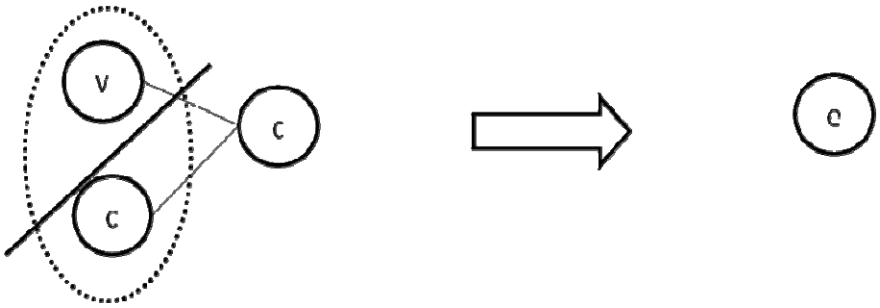


Figure 7: The rule for empty graphs

The leaves of the derivation tree of this graph grammar will be connected graphs built on basis of the source graph (graph grammar start graph) by adding minimum number of edges as it possible (connected component number minus one). The various cases of component connecting are possible, from case, when all components are connected with one (central component), to sequential connecting of the components. Disadvantage of this grammar is the large size of the derivation tree containing a priori unnecessary, redundant branches. In practical using the first and sixth rules can be removed, that improves performance of this grammar.

Adaptation of a self-managed computer structure

Providing the permanent correspondence between variable structure *G*(*t)* and solving by the system task is only a part of the self-management problem. The mentioned structure changes arise from the inner reasons: nodes breakdown, appearing of the new nodes, changes in the routes and in the connectivity because of nodes moving or battery charge decreasing/recovery [7]. The autonomic computer network must have such property as adaptation. This means that change of the outer conditions, change of the problem class, solving by the network, can lead to change of the acceptable structure class and accordingly to change of the graph grammar [8].

The architecture of such system includes two control loops: inner and outer.

The target of the inner loop functioning is permanent (periodic) control of the correspondence between system structure and graph grammar *GG*. Since grammar specifies the regularity of the network structures class to solve some general problem, the setting of the network structure for specific task and providing scalability are included in this target. Depending on the scale of the problem (it is difficult) the network structure of the different scales (amount of nodes and edges) can be generated by grammars. Thus in this case grammars can be used not only for the checking the structure, but also for generating it (generative graph grammar).

The target of the outer loop is transforming the graph grammar *GG*, setting it for the problem class. For this purpose operations under graph grammars, by which algorithm, getting the general task description (from wide class) as parameter and giving out the graph grammar of the autonomic computer system structure on the exit, are introduced.

Conclusion

Touched in the article topic is on the border of such fields of research as the theory of computational processes, the graph (and more complex structures) theory, the theory of automatic management. In the last time the theory of mobile processes are developed by R. Milner [9] and his followers [10] with using the category theory, conception of the bigraph and bisimulation.

The next advancement as regards analytical aspect is connected with the enhancement graph grammars to more complex formal systems, describing adequately not only structural, but also behavioral aspects. As for implementation grammars in the models of the autonomic systems it is required development of the parallel and distributional algorithms of the derivation.

Acknowledgement

The paper is published with support by Russian Foundation for Basic Research (project number 12-07-00763-a).

Bibliography

[1] O. Andrei, H. Kirchner. A Higher-Order Graph Calculus for Autonomic Computing. In: Graph Theory, Computational Intelligence and Thought, P. 15-26. Ed. M. Lipshteyn, V. E. Levit, R. M. Mcconnell. Springer, Heidelberg, 2009.

[2] P. Bottoni, F. De Rosa, K. Hoffman, M. Mecella. Applying algebraic approaches for modeling workflows and their transformations in mobile networks. In: Mobile Information Systems, Vol. 2, Issue 1, P. 51-76, IOS Press Amsterdam, The Netherlands, January 2006.

[3] M. Saksena, O. Wibling, B. Jonsson. Graph Grammar Modeling and Verification of Ad Hoc Routing Protocols. In: TACAS'08/ETAPS'08 Proceedings of the Theory and practice of software, 14th international conference on Tools and algorithms for the construction and analysis of systems, P. 18-32. Ed. C. R. Ramakrishnan, J. Rehof. Springer, Heidelberg, 2008.

[4] H. Ehrig, K. Ehrig, U. Prange, G. Taentzer. Fundamentals of algebraic graph transformations, Springer, 2006

[5] L. Lambers, H. Ehrig, F. Orejas. Conflict detection for graph transformation with negative application conditions. In: ICGT'06 Proceedings of the Third international conference on Graph Transformations, P. 61-76. Ed. A. Corradini, H. Ehrig, U. Montanari, L. Ribeiro, G. Rozenberg. Springer, Heidelberg, 2006.

[6] A. Habel, R. Heckel, G. Taentzer. Graph grammars with negative application conditions. In: Fundamenta Informaticae – Special issue on graph transformations, Vol. 26, Issue 3-4, P. 287-313. Ed. A. Skowron, G. Engels, H. Ehrig, G. Rozenberg. IOS Press Amsterdam, The Netherlands, June 1996.

[7] M. C. Huebscher, J. A. McCann. A survey of Autonomic Computing — degrees, models and applications. In: ACM Computing Surveys (CSUR), Vol. 40, Issue 3, P. 1-28. ACM New York, NY, USA, August 2008.

[8] I. B. Rodriguez, K. Drira, C. Chassot, M. Jimaiel. A rule-driven approach for architectural self adaptation in collaborative activities using graph grammars. In: International Journal of Autonomic Computing, Vol. 1, Issue 3, P. 226-245. Inderscience Publishers, Geneva, Switzerland, May 2010.

[9] R. Milner. Pure bigraphs: structure and dynamics. In: Information and Computation, Vol. 204, Issue 1, P. 60-122. Academic Press, Inc. Duluth, MN, USA, January 2006.

[10] F. Bonchi, F. Gadducci, B. Koenig. Synthesising CSS bisimulation using graph rewriting. In: Information and Computation, Vol. 207, Issue 1, P. 14-40. Academic Press, Inc. Duluth, MN, USA, January 2009.

Authors' Information

Alexander Mikov *– ACM Member, professor, head of the computing technologies chair, P.O. Box: Kuban State University, 149, Stavropolskaya str., Krasnodar, 350040, Russia; e-mail: alexander_mikov@mail.ru.*

Major Fields of Scientific Research: Distributed information systems, Simulation systems and languages

Alexander Borisov *– student of the Kuban State University, P.O. Box: Kuban State University, 149, Stavropolskaya str., Krasnodar, 350040, Russia; e-mail: nillerprog@gmail.com.*

Major Fields of Scientific Research: Distributed information systems, Simulation systems and languages

CITATION-PAPER RANK DISTRIBUTIONS AND ASSOCIATED SCIENTOMETRIC INDICATORS – A SURVEY

Vladimir Atanassov, Ekaterina Detcheva

Abstract: *This paper is devoted to studying the interrelations between most widely used scientometric indicators (in particular, Hirsch's h-, Egghe's g- and Zhang's e- indexes) for several more or less realistic citation-paper rank distributions. The analysis is provided for both continuous and discrete representations and is illustrated further on with examples for simultaneous time evolution (during a scientific career) of these indicators, computed by using real life scientometric data. The aim of the study is to illuminate specific properties of the indicators, the pros and cons of their use in various situations (citation-paper rank distributions) and (hopefully) to contribute for a fair and better scientific assessment.*

Keywords*: citation-paper rank distributions, scientometric indicators, h-index, g-index, e-index, approximate relations, data analysis*

ACM Classification Keywords*: H. Information Systems, H.2. Database Management, H.2.8. Database applications, subject: Scientific databases; I. Computing methodologies, I.6 Simulation and Modeling, I.6.4. Model Validation and Analysis*

Introduction

During the past decade the citation-based assessment of scientific activity has been essentially refined by considering details contained in *citation-paper rank distributions* and by suggesting various *scientometric indexes*. The first and most popular of them – the **Hirsch index** ([Hirsch, 2005], [Hirsch, 2007]) – has been introduced for a simple citation-paper rank distribution resulting from an extremely simplified model of a publication-citation process. Being a compromise between productivity and impact, this index ensures the opportunity for scientific assessment by a single number – a dream for many who are involved in various aspects of managing science. Although welcomed by most scientists, Hirsch's index has been criticized for underestimating the score of the most cited papers. The ***g*-index** [Egghe, 2006], constructed from informetric point of view for a Lotka (papers *vs.* citations) or Zipf (citations *vs.* paper ranks) distributions has been suggested as an alternative at least for two reasons. The first one is better accounting for the most cited papers, while the second (and in our opinion, more important) one is, that *g* as a true integral characteristic of the distribution is less subjected to statistical variability. On its turn the *g*-index has been criticized for an effect (we refer to it as *saturation* of *g*) which takes place when the total number of citations exceeds the square of the total number of publications. This criticism has lead to the appearance of the **e-index** [Zhang, 2009], also an integral characteristic that accounts for the excess of citations ignored in Hirsch's index estimation and at the same time free from this drawback. This, however, could not stop the explosion of improvements and nowadays we have several tens of indexes and numbers for scientific activity assessment [Schreiber, 2010], featuring its various aspects, like citation-paper rank distribution details (*e.g.* [Bornmann *et al*, 2010],[Cabrerizo et al, 2010]), accounting for the number of authors [Schreiber, 2008], the effect of self-citations and scientific fields specifics (*cf.* [Schreiber, 2007], [Iglesias and Pecharroman, 2007], [Alonso et

al, 2009], [Ferrara and Romero, 2012]). As an example one could point out that Harzing's *Publish or Perish* tool, based on Google Scholar database [A.-W. Harzing, 2012] estimates about 16 indicators and indexes for individual scientist's evaluation. However, these indicators (representing information squeezed from the citation-paper rank distribution) have their common origin and hence are mutually related.

The aim of this paper is to study the interrelations between most widely used scientometric indicators as Hirsch's *h*-, Egghe's *g*- and Zhang's *e*- indexes for several model continuous and discrete citation-paper rank distributions. The results obtained might be helpful to realize the pros and cons of the use of these indicators in various situations of scientific assessment. In particular, we address problems as, to what extent the indexes are robust (i.e. distribution independent), how many citations of the most cited papers are ignored by the *h*-index, which index – *g* or *e* – performs better in different cases, as well as at what conditions a saturation of *g* occurs.

The paper is organized as follows: in the first section we define the citation-paper rank distributions in both discrete and continuous representations. The second section introduces the scientometric indexes in the form they appear in the original papers ([Hirsch, 2005], [Egghe, 2006], [Zhang, 2009]). Next two sections consider the relations among the scientometric indicators for various discrete and continuous citation-paper rank distributions. Some of the theoretical conclusions are illustrated with examples for time evolution of scientometric indexes during real scientific careers.

Citation-paper rank distributions

Citation-paper rank distribution is defined as the sequence $\{I_c(I_p); I_p = 1, N_p\} : I_c(I_p) \geq I_c(I_p + 1); I_p = 1, N_p - 1$ of citations $I_c(I_p)$ to the paper I_p, where the set of N_p papers has been arranged in descending order to the number of citations gained, *i.e.* most cited placed first. We emphasize that the native, real life distributions (see Fig. 1) are discrete and consist of nonnegative integers. In this study, however, we consider also (as approaches to reality) continuous versions $C(P)$ of the discrete distributions $I_c(I_p)$, as well as discrete model distributions that consist of nonnegative real numbers. The first approach is justified when a large amount of data is analyzed (Fig. 2), while the second one appears in a natural way when approximating integer data with real-valued functions and *vice versa*.

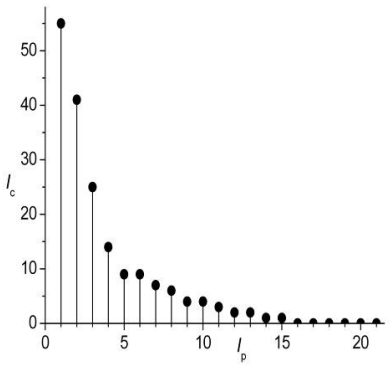


Figure 1. Real life citation-paper rank distribution example

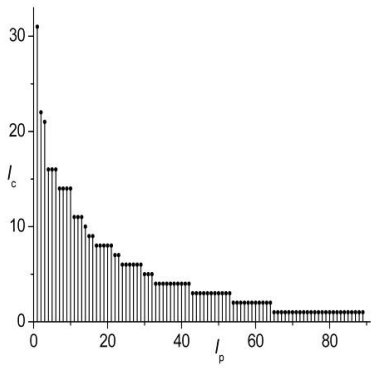


Figure 2. Citation-paper rank distribution example for a large amount of data

In this study, everywhere except explicitly stated, the continuous distributions will be considered as defined on a finite interval of papers $P \in (0, P_m], P_m \geq 1$ and varying between the maximal citation count $C_m = C(0) \geq 0$ and zero (some examples are shown on Fig.3). We analyze a class of distributions such that d^2C / dP^2 does not

change its sign in the interval under consideration. Integration in the continuous case is performed (with one exception) with lower bound equal to zero – one could imagine it as summing the area of 'stripes' (0,1], (1,2],… for the first, second *etc.* papers respectively.

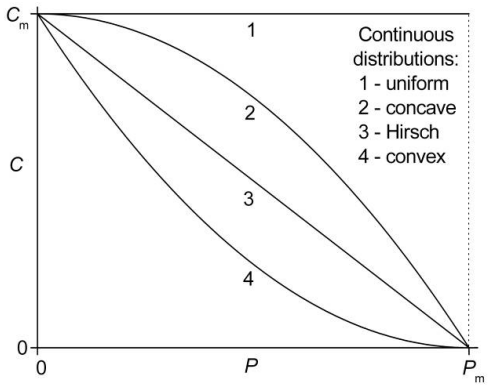


Figure 3. Continuous citation-paper rank distributions (1 – uniform, 2 – concave, 3 – negative slope linear (Hirsch) and 4 – convex). C_m and P_m are citation count of the most cited paper and number of papers, respectively.

Scientometric indicators

The scientometric indicators are considered to be a convenient measure to assess and compare scientific activity, *e.g.* in situations where the use of citation-paper rank distribution is not possible, or reduction of scientometric information is necessary due to time and effort considerations. Scientometric indicators are closely associated with citation-rank distributions. The most widely used are as follows:

- total number of papers N_p (discrete case) or P_m (continuous case);
- total number of citations $N_c = N_c(N_p) \equiv \sum_{Ip=1}^{Np} I_c(I_p)$ or $N_c = N_c(P_m) \equiv \int_0^{Pm} C(P)dP$ for the discrete or continuous case, respectively; it should be noted that (in order to have a distribution) N_c must remain finite, even when considering distributions with *infinite* number of papers, *i.e.* N_p (or P_m) $\to \infty$.;
- average number of citations per paper (N_c / N_p) and average number of citations per year;
- scientometric *indexes*: h -index, g -index, e -index.

Further on we recall the definitions of the scientometric indexes in the way they appear in the original papers and comment some general, more or less distribution-independent properties and relations.

- ***h*-index:**

A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have $\le h$ citations each [Hirsch, 2005], *i.e.* $h : \{I_c(I_p) \ge h$ for $I_p \le h$ and $I_c(I_p) \le h$ for $I_p > h\}$ (discrete case) and $C(h) = h$ (continuous case, Fig. 4). Obviously, h cannot exceed neither N_p (or P_m), nor $I_c(1)$ (or C_m), for discrete (or continuous) distributions. This index has been constructed assuming approximately linear negative-slope citation-paper rank distribution and it equals twice the harmonic mean of impact and productivity. Hirsh's index is the most popular among all other indexes (and the oldest one). The criticism against its use (apart

from the general criticism against scientometrics itself) is due to the fact that usually h ignores a large amount of citations to the first h most cited papers.

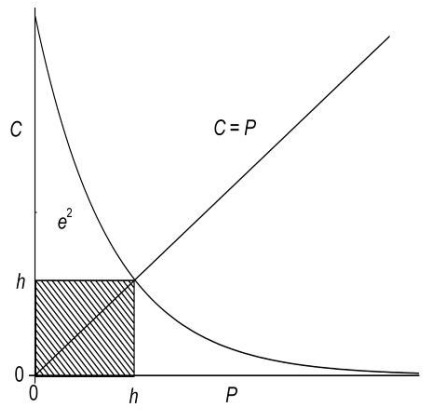


Figure 4. Illustrating Hirsh's h and Zhang's e definitions ([Hirsch, 2005], [Zhang, 2009])

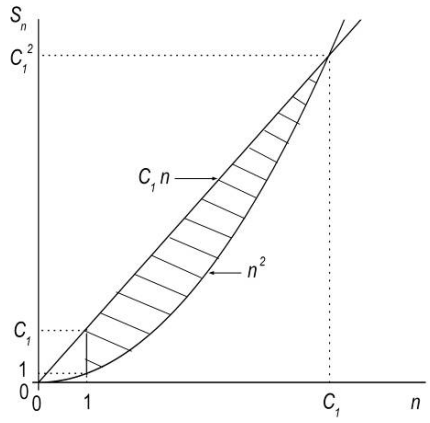


Figure 5. On the existence and uniqueness of g

$$(S_n = \sum_{i=1}^{n} C_i, C_1 \geq C_2 \geq \ldots \geq C_n \geq 0, n \geq 1)$$

- ***g*-index:**

The g -index is introduced as an improvement of the h -index to measure the global citation performance of a set of articles. If this set is ranked in decreasing order of the number of citations that they received, the g -index is the (unique) largest number g such that the top g articles received (together) at least g^2 citations [Egghe, 2006]. Let $N_c(G)$ is the number of citations gained by the first G most cited papers, then $g = \max\{G : N_c(G) \geq G^2\}$. The inequality $g \geq h$ follows immediately from the definition of g ; L. Egghe has also proved its existence and uniqueness for arbitrary citation-paper rank distribution (see also Fig. 5). The g -index is considered to represent the most cited papers better than h does. However, since g is associated with papers in the set, it cannot exceed N_p (or P_m) and remains constant ($g = N_p$ or P_m) if $N_c \geq N_p^2$ or, for a continuous distribution, $N_c \geq P_m^2$. This drawback (illustrated on Fig. 6) has been discussed in [Zhang, 2009], where a possible solution to the problem in the form of introducing virtual papers of zero citation count has been found unacceptable. An additional problem of g is the following: a scientist's saturated (i.e. limited by the number of papers) g -index could be increased by simply publishing (until the saturation level is exceeded) additional papers of mediocre quality, that probably will not be cited at all.

- ***e*-index:**

The *e*-index accounts for the excess citations (represented by e^2) in addition to the h^2 citations of the *h*-core papers [Zhang, 2009]. It is defined as $e^2 = N_c(h) - h^2$ and is free from the constraints on h and g (Fig. 4). The following inequalities take place independently on the citation-paper rank distribution:

$$I_c(g) \leq h \leq g \leq \min(I_c(1), N_p) \text{ or } C(g) \leq h \leq g \leq \min(C_m, P_m), \quad h^2 + e^2 \leq N_c(g),$$

as well as

$$e^2 \geq -\frac{1}{2}h^2(dC/dP)_{P=h} \text{ for a } \textit{convex} \text{ distribution } (d^2C/dP^2 \geq 0),$$

$$e^2 \leq -\frac{1}{2}h^2(dC/dP)_{P=h} \text{ for a } \textit{concave} \text{ distribution } (d^2C/dP^2 \leq 0).$$

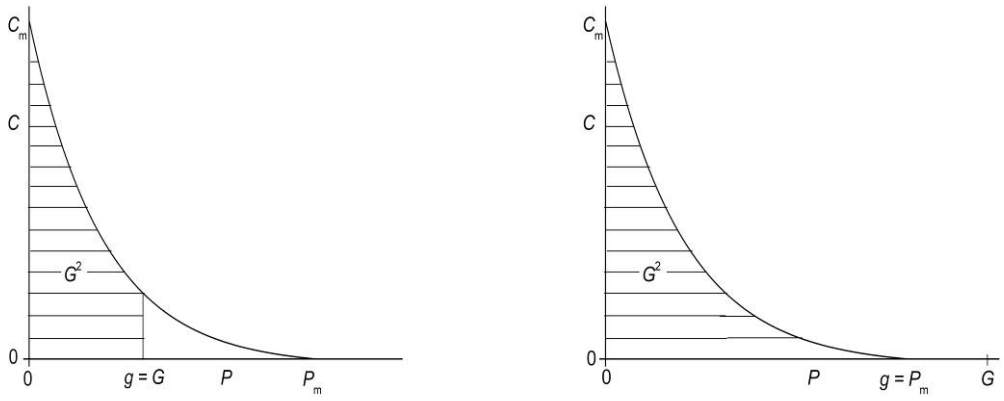


Figure 6. Illustrating Egghe's g index definition:

(left – far from saturation, $g^2 = G^2 \leq N_c \leq P_m^2$; right – saturated, $G^2 \geq N_c \geq P_m^2 = g^2$)

Relations between scientometric indicators (continuous case)

Continuous citation-paper rank distributions are considered as an approach to the real life discrete integer-valued ones. Their advantages include the opportunity to *analytically* compute scientometric indexes and to derive explicitly more or less exact relationships between them. At the same time continuous distributions keep most of the properties and peculiarities of the discrete ones that make them suitable for miscellaneous model studies. Further on we consider two groups of continuous distributions: finite sized ($P_m < \infty$ – uniform, linear negative-slope (or Hirsch), three-parameter polynomial and three-parameter positive exponent power-law distributions) and infinite size ($P_m \to \infty$ – exponential and Pareto distributions).

- **Uniform distribution**

This simple but not quite realistic distribution (Fig. 3, curve 1) reveals clearly the constraints on the h- and g-indexes imposed by the finite number of papers P_m. It is defined as $C(P) = C_m$ for $0 < P \leq P_m$ and

$$N_c = C_m P_m,\ h = \min(C_m, P_m),\ g = h,\ e^2 = (C_m - h)h \tag{1}$$

In this special case h and g coincide and the number of h-core citations is $e^2 + h^2 \geq g^2$. The three indexes h, g and e account for number of citations that represent $r_h = h^2 / N_c$, $r_g = N_c(g) / N_c$ and $r_{h-core} = (h^2 + e^2) / N_c$ parts of all citations N_c, as follows:

$$r_h = \min(s, 1/s),\ r_g = r_{h-core} = \min(1, s),\ s = C_m / P_m. \tag{2}$$

- **Linear negative-slope distribution**

This distribution (Fig. 3, curve 3) has been obtained in [Hirsch, 2005] by assuming constant publication rate (number of papers per year) and constant citation productivity (number of citations per paper per year. It is defined as $C(P) = C_m - sP$ for $0 < P \le P_m$, where the (constant) slope is $s = C_m / P_m$. We have

$$N_c = \frac{1}{2} C_m P_m,\ h = \left(\frac{1}{C_m} + \frac{1}{P_m} \right)^{-1},\ g = \min\left(\left(\frac{1}{C_m} + \frac{1}{2P_m} \right)^{-1}, P_m \right),\ e^2 = \frac{1}{2}(C_m - h)h. \tag{3}$$

The following relations take place:

$$N_c = \frac{(1+s)^2}{2s} h^2, \tag{4}$$

$$g / h = \begin{cases} 2(1+s)/(2+s), & 0 < s \le 2 \\ (1+s)/s, & s \ge 2 \end{cases} \tag{5}$$

$$e^2 = \frac{1}{2} s h^2. \tag{6}$$

Now the ratio g / h reaches its maximum $\frac{3}{2}$ for $s = 2$, where the total number of citations $N_c = g^2 = P_m^2$. Further on, for the relative citation count associated with the indexes one obtains

$$r_h = 2s/(1+s)^2,\ r_{h-core} = s(s+2)/(s+1)^2,\ r_g = \begin{cases} 8s/(2+s)^2, 0 < s \le 2 \\ 1, \quad s \ge 2 \end{cases} \tag{7}$$

Hence h^2 contains no more than 50 percent of all citations (a minimum of $r_h = 0.5$ at $s = 1$). We note that for $s \ge 2$ the g -index remains constant (saturated on a level $g = P_m$) and accounts for <u>all</u> citations.

- **Three-parameter polynomial distribution**

Let us denote $x = P / P_m, y = C / C_m$, then we could consider

$$y(x) = 1 - (1 + \frac{1}{2}\rho)x + \frac{1}{2}\rho x^2 \text{ for } 0 \le x \le 1, 0 \le y \le 1 \tag{8}$$

as a three-parameter polynomial distribution (Fig. 3, curves 2-4), where the third parameter is the constant second derivative $d^2y / dx^2 = \rho$. It covers the linear negative distribution ($\rho = 0$, Fig. 3 - curve 3) as well as convex ($0 < \rho \le 2$) and concave ($-2 \le \rho < 0$) distributions. Both limiting cases ($\rho = 2$, $y = (1-x)^2$) and ($\rho = -2$, $y = 1 - x^2$) are displayed on Fig. 3 (curves 2 and 4, respectively). Now the slope depends on x :

$$dy / dx = -\left(1 + \frac{1}{2}\rho\right) + \rho x,\ (dy/dx)_{x=0} = -1 - \frac{1}{2}\rho,\ (dy/dx)_{x=1} = -1 + \frac{1}{2}\rho. \tag{9}$$

The scientometric indicators are listed as follows:

$$N_c = \frac{1}{2} C_m P_m \left(1 - \frac{1}{6}\rho\right), \tag{10}$$

$$h / P_m = 4 / \left[2\frac{1+s}{s} + \rho + \sqrt{\left(2\frac{1+s}{s} + \rho\right)^2 - 8\rho} \right], \tag{11}$$

$$g = \min(G, P_m),\ G / P_m = 2 / \left[\frac{1}{s} + \frac{1}{2} + \frac{1}{4}\rho + \sqrt{\left(\frac{1}{s} + \frac{1}{2} + \frac{1}{4}\rho \right)^2 - \frac{2}{3}\rho} \right], \quad (12)$$

$$e^2 / h^2 = \frac{1}{3}\left\{ \frac{1}{2} s\left(1 + \frac{1}{2}\rho\right) - 1 + \sqrt{\left[s\left(1 - \frac{1}{2}\rho\right) + 1 \right]^2 + 2s\rho} \right\}, \quad (13)$$

where $s = C_m / P_m$. Egghe's g is saturated (*i.e.* $G \geq P_m$, which corresponds to $N_c \geq P_m^2$) for $s \geq 12 / (6 - \rho)$. The latter inequality implies that for the limiting case of concave distribution ($\rho = -2$) saturation occurs for $s \geq 3 / 2$, while for the limiting convex distribution ($\rho = 2$) saturation takes place for $s \geq 3$. Fig. 7 illustrates the behavior of g / h and e / h for $\rho = 0, \pm 2$. Obviously, Egghe's index g better accounts for the excess of citations in the h-core papers for s below the saturation point (between 1.5 and 3, depending on the second derivative ρ). Above this limit, however, g / h rapidly decreases, while e / h keeps on growing. Table 1 gives a notion of how the scientometric indicators and citation partition for this distribution looks like (since $s = 1$ no saturation of g occurs). We note the robust behavior of e and the large amount of citations accounted by g .

Table 1. Indicators and citation partition for three-parameter polynomial distribution with $C_m = 100$, $P_m = 100$.

	N_c	h	g	e	r_h	r_g	r_{h-core}
$\rho = 2$	3333	38	55	33	0.44	0.91	0.76
$\rho = 0$	5000	50	66	35	0.50	0.89	0.75
$\rho = -2$	6667	61	79	40	0.57	0.94	0.81

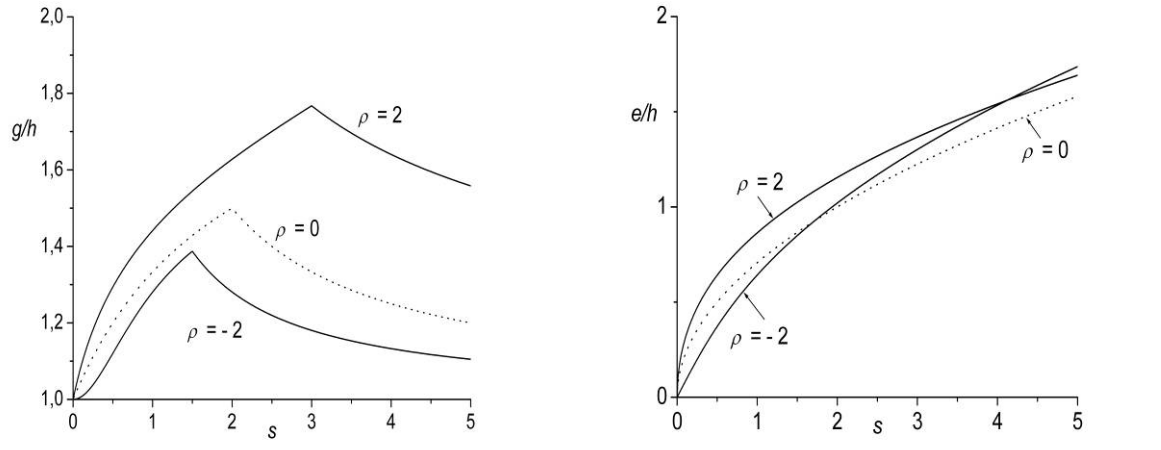


Figure 7. Index ratios g / h (left) and e / h (right) dependence on $s = C_m / P_m$, for the limiting cases of convex ($\rho = 2$), and concave ($\rho = -2$) three parameter polynomial distribution, as well as for the negative slope linear one ($\rho = 0$, dot line)

- **Three-parameter (positive exponent) power-law distribution**

This distribution is defined as $y = 1 - x^{\alpha}$, where $x = P / P_m, y = C / C_m$, $0 < x \le 1, 0 \le y \le 1$ and $\alpha > 0$. It is *convex* for $0 < \alpha < 1$ and *concave* for $\alpha > 1$. Its slope is $dy / dx = -\alpha x^{\alpha-1}$, $(dy / dx)_{x=0} = -\infty, 1$ or 0 for $0 < \alpha < 1, \alpha = 1$ or $\alpha > 1$ respectively and $(dy / dx)_{x=1} = -\alpha$. The total number of citations is $N_c = C_m P_m \alpha / (\alpha + 1)$ and the scientometric indexes are obtained as (unique) solutions to the equations:

$$(h / P_m)^{\alpha} + \frac{1}{s}(h / P_m) - 1 = 0 \tag{14}$$

$$(G / P_m)^{\alpha} + \frac{\alpha + 1}{s}(G / P_m) - (\alpha + 1) = 0, \tag{15}$$

$$\frac{e^2}{h^2} = \frac{\alpha}{\alpha + 1}\left(\frac{C_m}{h} - 1\right), \tag{16}$$

where, as usual, $s = C_m / P_m$ and $g = \min(G, P_m)$. The saturation of g occurs for $s \ge 1 + (1 / \alpha)$. The cases $\alpha = 1$ and $\alpha = 2$ reproduce the linear (negative slope) and the $\rho = 2$ concave distributions considered previously in the paper. Equations (14-16) can be explicitly solved for a *convex* distribution with $\alpha = 1 / 2$ (Fig. 8):

$$\frac{h}{P_m} = \frac{4}{\left(1 + \sqrt{1 + (4 / s)}\right)^2}, \quad \frac{G}{P_m} = \frac{9}{\left(1 + \sqrt{1 + (9 / s)}\right)^2}, \quad \frac{e^2}{h^2} = \frac{1}{6}\left(1 + \sqrt{1 + \frac{4}{s}}\right). \tag{17}$$

For this particular case we have $g = G$ for $s \le 3$ and $g = P_m$ above this limit.

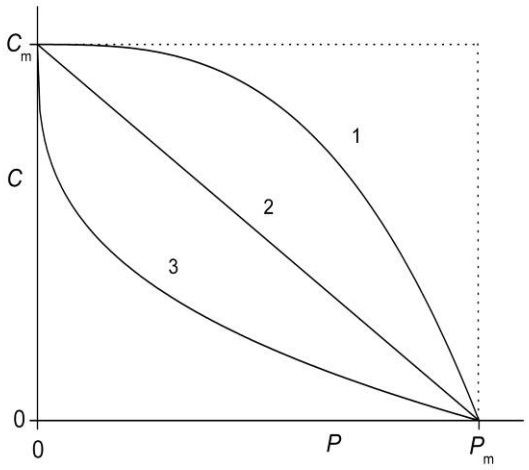


Figure 8. Three-parameter (positive exponent) power-law distribution (1 – concave with $\alpha = 2$, 2 – negative slope linear, 3 – convex with $\alpha = 1 / 2$)

Further on we consider two examples of continuous citation-paper rank distributions with infinite number of papers ($P_m \to \infty$). There are several peculiarities associated with these distributions. The obvious one is absence of g-index saturation for arbitrary large maximal number of citations C_m. This leads to an important inequality, $e^2 + h^2 \le g^2$ and consequently, $e \le g$. Another peculiarity is the fact that, paradoxically, sometimes

the maximal number of citations appears to be greater than their total number ($C_m > N_c$). However, due to some of their properties (of interest for deeper studies, *cf.* [Egghe, 2005]) we give results that might be compared with those of other distributions.

- **Exponential distribution**

is defined as $C(P) = C_m \exp(-\beta P)$, for $0 \le P < \infty$ and $\beta > 0$. Let us introduce $\underline{C} = \beta C$, $\underline{C}_m = \beta C_m$, $\underline{P} = \beta P$, $\underline{h} = \beta h$, $\underline{g} = \beta g$ and $\underline{e} = \beta e$, then we have $\underline{C}(\underline{P}) = \underline{C}_m \exp(-\underline{P})$. The total number of citations is $N_c = C_m / \beta = \underline{C}_m / \beta^2$ and the re-scaled scientometric indexes are obtained as solutions to:

$$\underline{h} \exp(\underline{h}) = \underline{C}_m,\ \underline{g}^2 / [1 - \exp(-\underline{g})] = \underline{C}_m,\ \underline{e}^2 = \underline{C}_m - \underline{h}(\underline{h} + 1). \quad (18)$$

As it could be seen from (Fig. 9), $\underline{g}$ and $\underline{e}$ lay close to each other; both of them are much greater than $\underline{h}$ and hence, represent better the effect of most cited papers than $\underline{h}$ does. Note that $N_c < C_m$ for $\beta > 1$.

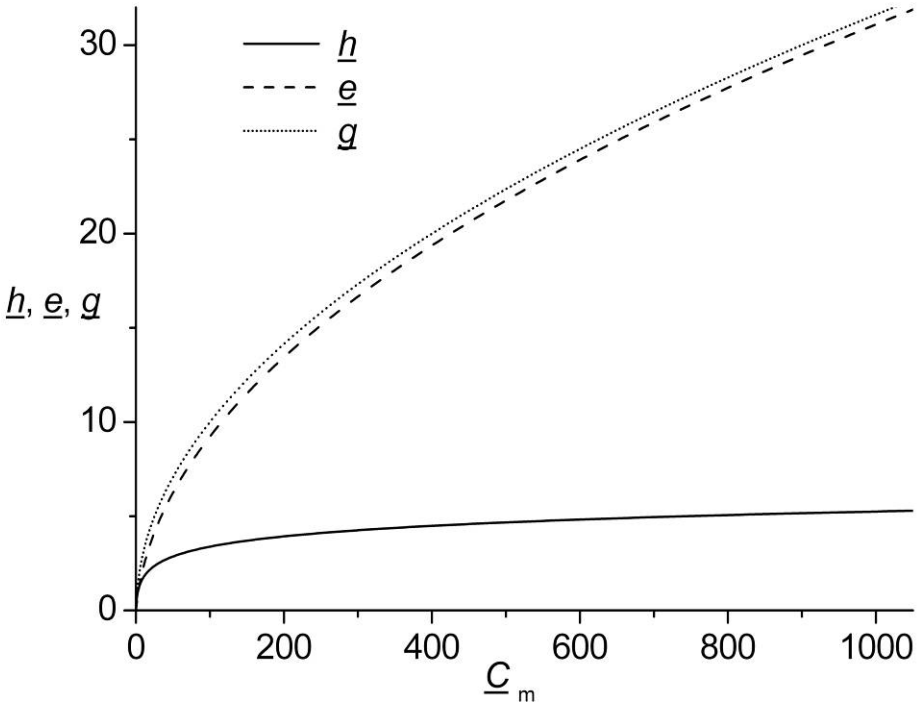


Figure 9. Re-scaled indexes $\underline{h} = \beta h$, $\underline{g} = \beta g$, $\underline{e} = \beta e$ versus re-scaled maximum citation count $\underline{C}_m = \beta C_m$.

- **Pareto distribution**

The convex distribution $C(P) = C_m P^{-\alpha}, 1 \le P < \infty, \alpha > 1$ (introduced by *Alfredo Pareto* (1848-1923) for other purposes) is one of the most commonly used in scientific fields as informetrics, scientometrics and other 'metrics'. It is *scale-free* and has *the product property* [Egghe, 2005]. The total citation count is $N_c = C_m / (\alpha - 1)$, i.e. $N_c < C_m$ for $\alpha > 2$. The scientometric indexes h and e are

$$h = C_m^{1/(\alpha+1)}, \tag{19}$$

$$e^2 = (h^{1+\alpha} - \alpha h^2)/(\alpha - 1), \tag{20}$$

while g is obtained as a solution to

$$g^2 - N_c(1 - g^{1-\alpha}) = 0. \tag{21}$$

Let $\alpha = 3$, then we have $N_c = \frac{1}{2}C_m$, $h = (C_m)^{1/4}$ and

$$g = \frac{1}{2}\sqrt{C_m\left(1 + \sqrt{1 - 8/C_m}\right)} \text{ for } C_m \geq 8,\ e^2 = \frac{1}{2}\sqrt{C_m}(\sqrt{C_m} - 3) \text{ for } C_m \geq 9. \tag{22}$$

For $C_m \to \infty$ e asymptotically approaches g:

$$g^2 / e^2 = 1 + 3C_m^{-1/2} + 7C_m^{-1} - 6C_m^{-3/2} + O(C_m^{-2}),\ C_m >> 1. \tag{23}$$

There is some concern about use of this distribution in analyzing ranked scientometric data (*cf.* [Atanassov and Detcheva, 2012]), mainly associated with the fact that in most real life cases Pareto exponent α is close to (and many times less than) unity. We also note that $\alpha = 1$ corresponds to Lotka's exponent 2 and fractal dimension 1 [Egghe, 2005].

Relations between scientometric indicators (discrete case)

At a first glance these distributions should better describe the real life citation-paper rank histograms. This is probably true, but one should bear in mind that although the *argument* is a positive integer, the *function* (i.e. the distribution) itself is (generally) a positive real number. This inconvenience is usually overcome by considering the nearest integer part of the result *(cf. e.g.* [Clauset *et al*, 2009]). Therefore, in most cases further on we derive relationships that are in this sense only approximately true (denoted here with '$\approx$').

Although all of the continuous distributions addressed in the previous section have their discrete representations, we restrict ourselves with considering two discrete Pareto distributions, resulting from the continuous one, however, better suited for scientometric data analysis.

- **Zeta distribution**

This distribution is defined for all positive integers:

$$I_c(I_p) \approx I_{cm} I^{-\alpha},\ I_p = 1, 2, \ldots,\ \alpha > 1. \tag{24}$$

The total number of citations is

$$N_c \approx I_{cm}\zeta(\alpha), \tag{25}$$

where $\zeta(\alpha)$ is the Riemann zeta function. By definition, for h we have

$$I_c(h+1) - 1 \leq h \leq I_c(h). \tag{26}$$

or

$$1 \leq \frac{I_{cm}}{h^{\alpha+1}} \leq \left(1 + \frac{1}{h}\right)^{\alpha}. \tag{27}$$

Hence (for $h > 2\alpha$)

$$I_{cm} \approx h^{\alpha+1},\ h \approx (I_{cm})^{\frac{1}{\alpha+1}}, \tag{28}$$

(*cf.* [Egghe and Rousseau, 2006]). Under these assumptions, we can estimate h by knowing I_{cm} and *vice versa*. Since the distribution is defined for an infinite sized set of positive integers we have no saturation effects for the g -index; the latter is obtained as a solution to the equation:

$$g^2 / S(\alpha, g) \approx h^{\alpha+1} \approx I_{cm}, \tag{29}$$

where

$$S(\alpha, N) = \sum_{I=1}^{N} I^{-\alpha} \tag{30}$$

is the *incomplete* Riemann zeta function (Figs. 10 and 11). Further on, for e^2 one obtains

$$e^2 \approx h^2 \left[S(\alpha, h) h^{\alpha-1} - 1 \right]. \tag{31}$$

The dependence of Hirsh's h, Egghe's g and Zhang's e on the maximal citation count I_{cm} for a zeta distribution with power exponent $\alpha = 1.1$ is demonstrated on Fig. 12.

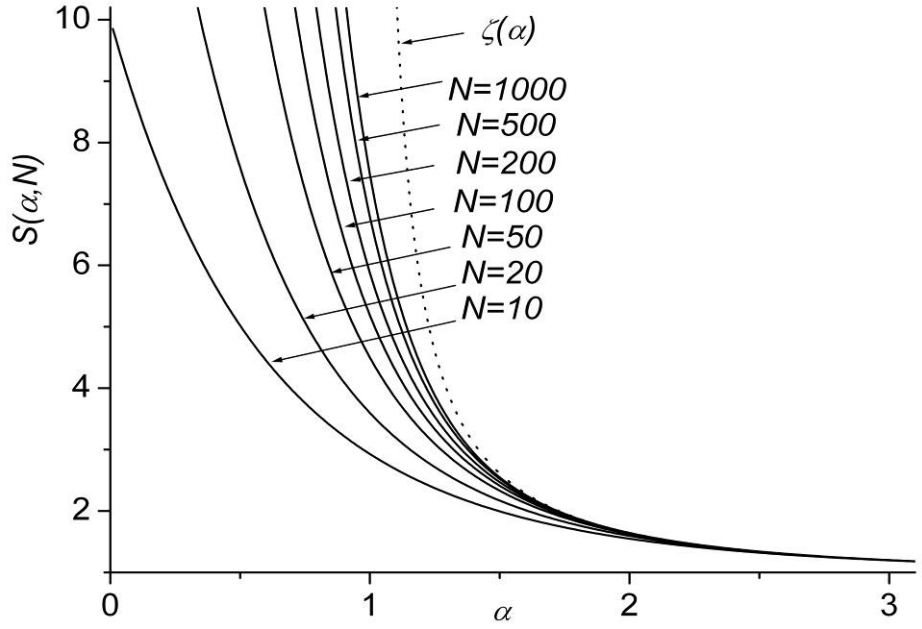


Figure 10. The incomplete zeta-function $S(\alpha, N) = \sum_{I=1}^{N} I^{-\alpha}$ *versus* power exponent α for various N.

Similarly to the continuous case, this *Pareto*-type distribution (defined on *infinite* set of positive integers) has severe problems when the power exponent α approaches, or falls down below unity (*cf.* [Atanassov and Detcheva, 2012]). A possible solution to the problem is the use of

- **Zipf distribution**

named after (the famous law of) *George Kingsley Zipf* (1902-1950), defined for a *finite* set of positive integers:

$$I_c(I_p) \approx I_{cm} I_p^{-\alpha}, \quad I_p = 1, 2, \ldots, N_p, \quad \alpha \geq 0, \tag{32}$$

where N_p is the total number of papers and the total number of citations is

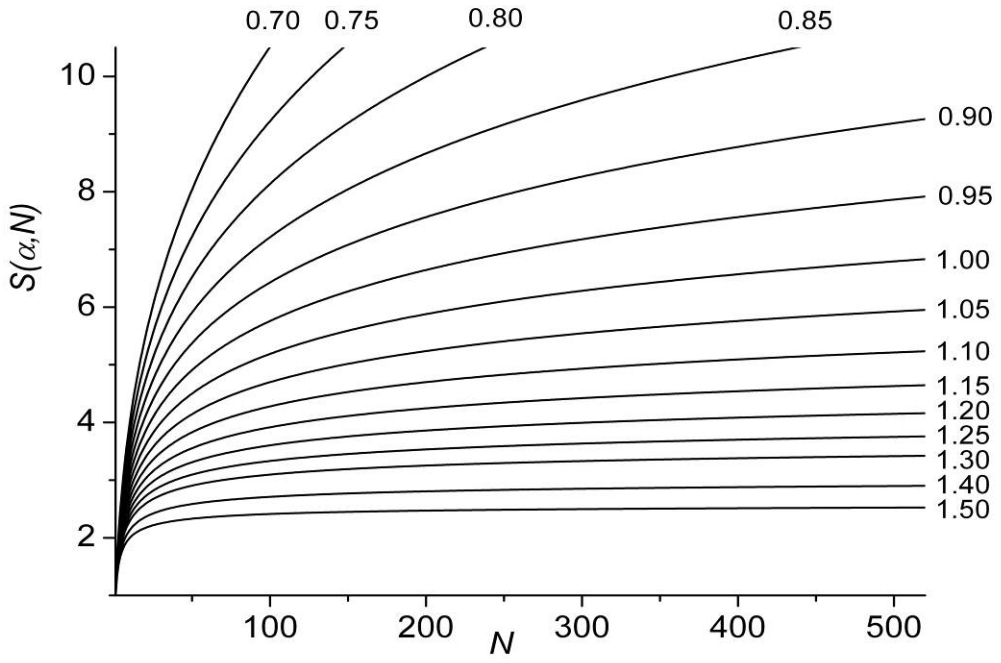


Figure 11. The incomplete zeta-function $S(\alpha, N) = \sum_{I=1}^{N} I^{-\alpha}$ *versus* N for various power exponents α.

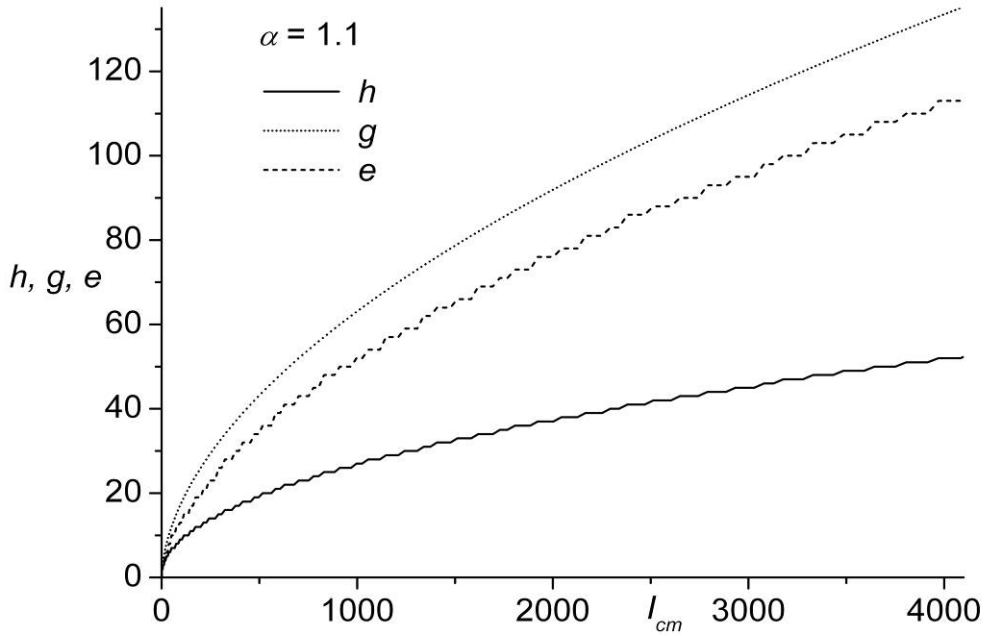


Figure 12. Dependence of h, g and e on maximal citation count I_{cm} for a zeta distribution with $\alpha = 1.1$.

$$N_c \approx I_{cm} S(\alpha, N_p)\,. \tag{33}$$

Now we have saturation effects for both h:

$$h = \min(H, N_p)\,,\ H \approx I_{cm}^{\frac{1}{\alpha+1}}\ , \tag{34}$$

and g:

$$g = \min(G, N_p)\ , G^2 \,/\, S(\alpha, G) \approx I_{cm}\,, \tag{35}$$

while the expression for e^2 (Eq. 31) remains unchanged. This distribution fits quite well many of the citation-paper rank histograms ([Atanassov and Detcheva, 2012], [Atanassov, 2012]); however, one must pay for this with introducing a second parameter, namely the total number of papers N_p. The good news is that, for large enough N and α greater than (and not very close to) unity, $S(\alpha, N)$ is a slowly varying function of N (Fig. 11). Note that in the limit $N \to \infty$ we have $S(\alpha, N) \to \zeta(\alpha)$; for α close to unity this could mean $N > 10^6$ and more. Bearing in mind that the total number of scientific sources that appeared in the whole world history nowadays hardly exceeds several tens of millions, one should be cautious when using zeta distributions, in particular, with lower power exponents. In addition, bearing in mind our *nearest integer* convention one may ask himself what happens when $I_c(I_p) < 1/2$. The answer is (*cf.* [Atanassov and Detcheva, 2012]), that this zeta distribution is indistinguishable from a Zipf 's one with $N_p \approx (2I_{cm})^{1/\alpha}$. In most cases this limit does not exceed 10^3-10^4.

- **Kronecker-type discrete distribution**

might be considered as limiting case of Zipf or zeta distributions with power exponents $\alpha \to \infty$. It is of interest for analyzing cases where a single paper has been highly cited compared to all others; this situation is far not as exotic as it seems (*cf.* next section and the examples in. [Zhang, 2009]). It is defined as $I_c(I_p) = I_{cm}$ for $I_p = 1$ and zero elsewhere. The scientometric indicators are easily obtained to be $N_c = I_{cm}, h = 1$, $g = \min(1, G) = 1$, where $G = \sqrt{N_c} = \sqrt{I_{cm}}$ and $e = \sqrt{G^2 - 1} \approx G - \left(2G\right)^{-1}$ for $I_{cm} >> 1$. We see that a severe loss of citations occurs in the determination of h and g, while e accounts for all of them, as suggested in [Zhang, 2009].

Time evolution examples

One of the advantages in using scientometric indexes is the opportunity to represent in a clear and concise way the scientific activity of a scholar during long periods of his/her academic career. We have chosen two examples to illustrate the main results of the analysis in the previous sections. Both of them are characterized with a Zipf – like citations-paper rank distributions, however, with different power exponents.

The first example (Fig. 13) represents time evolution of the scientometric indexes h, g, e for a twenty-year period of scientific activity in the field of *photonics* (more details can be found in [Atanassov, 2012]). The power exponent α varies gently between 0.8 and 1.0. Due to the high productivity and hence, large number of published papers no saturation effects on g can be observed and this index accounts for the excess of citations in the h-core papers significantly better than e does. This behavior approximately follows the relations between the scientometric indexes for a zeta-distribution with $\alpha = 1.1$ illustrated in Fig. 12.

The second example (Fig. 14) illustrates the effect of saturation on g. This is a case study for a situation where

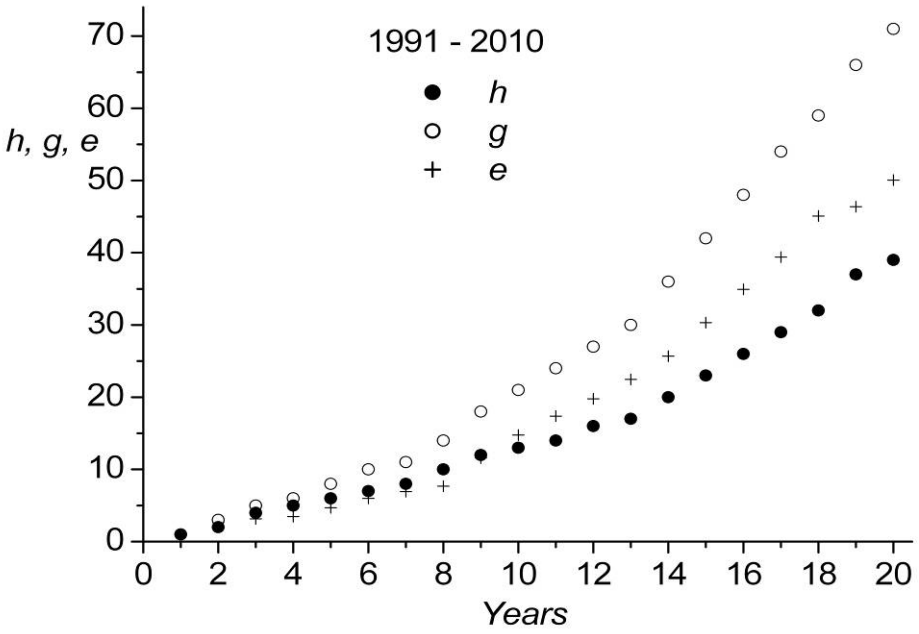


Figure 13. Time evolution of scientometric indexes (case study 1): no saturation of g is observed.

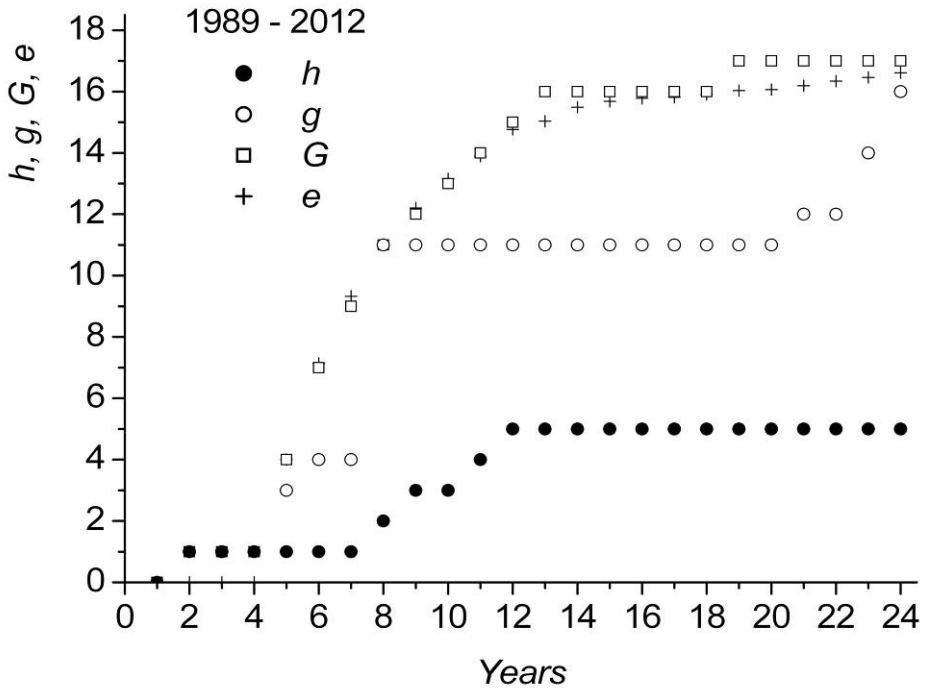


Figure 14. Time evolution of scientometric indicators (case study 2): significant saturation of g takes place.

a scientist has one highly cited paper while the rest of his/her papers is far not so popular. In this real life case g is limited by the number of publications N_p and is noticeably smaller than G, computed by solving Egghe's

relation $G^2 = N_c(G)$. Thus G is approximately equal to Zhang's e – a situation typical for the Kronecker-type distribution considered previously. Similar examples have encouraged Zhang to introduce his e -index. Note that the increase in g observed for $Years$ varying between 20 and 24 is due to papers with negligible citation count.

Summary and conclusions

In this paper we have obtained relations between scientometric indicators like total number of citations, Hirsch's, Egghe's and Zhang's indexes for various model citation-paper rank distributions in continuous and discrete representations. The theoretical considerations have been illustrated with two examples for time evolution of these indicators during real scientific careers.

Our main conclusions are summarized as follows:

- the Hirsch index, compromising between productivity and impact at the same time ignores a considerable amount of citations to the (highly cited) papers; this effect is stronger for distributions of convex type (in particular, the continuous exponential or Pareto distributions) and/or where the ratio of maximum citation count (number of citations gained by the most cited paper) and number of publications significantly deviates from unity;
- a quite general drawback of the *h*-index is its relatively strong dependence on the distribution shape that in the real life could result in statistical instability;
- Egghe's *g*-index , as a true integral characteristic seems to be statistically stable; up to a certain limit it accounts best for the highly cited papers, compared with the other two indexes; however, above this limit (where the total number of citations exceeds the square of the total number of papers) the *g*-index reaches its maximum value (equal to the total number of papers); in such regime of *saturation* the index accounts for all of the citations; it would grow only if the total number of papers (even of zero citation count) is increased; therefore, our conclusion is, that the *g*-index performs best below, and has severe problems above the saturation limit;
- a saturation of the *g*-index takes place for a finite number of papers only; the paper count where saturation occurs for convex distributions is greater than that for concave ones;
- Zhang's *e*-index appears to be quite robust with respect to the distribution shape; however, it accounts for less citations than Egghe's *g*-index (below saturation) does;

Since the *g*-saturation occurrence might be easily overlooked, care is needed when computing and comparing this index. In conclusion, we believe that such model studies could prove useful in analyzing various aspects of scientific activity assessment, in particular, the self-citations effect on the scientometric indexes which seems to be an appropriate topic for future studies.

Acknowledgments

This paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Alonso et al, 2009] S. Alonso, F. J. Cabrerizo, E. Herrera-Viedma and F. Herrera. h-index: a review focused in its variants, computation and standardization for different scientific fields, Journal of Informetrics 3 273-289 (2009)

[Atanassov and Detcheva, 2012] V. Atanassov, E. Detcheva. Theoretical analysis of empirical relationships for Pareto-distributed scientometric data, Int. J. Information Models and Analyses 1(3) 271-282 (2012)

[Atanassov, 2012] V. Atanassov. Time evolution of scitation-paper rank distributions and its implications for scientometric models, presented at "Evaluating Science: Modern Scientometric Methods" Conference Sofia May 21-22, 2012,COST Action "Physics of Competition and Conflicts, https://sites.google.com/site/scientometrics2012sofia/presentations (2012)

[Bornmann et al, 2010] L. Bornmann, R. Mutz and H.-D. Daniel. The h-index research output measurement: two approaches to enhance its accuracy, Journal of Informetrics 4 407-414 (2010)

[Cabrerizo et al, 2010] F. J. Cabrerizo, S. Alonso, E. Herrera-Viedma and F. Herrera. q2 –index: quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core, Journal of Informetrics 4 23-28 (2010)

[Clauset et al, 2009] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data, J. SIAM Review 51 (4) 661-703 (2009)

[Egghe and Rousseau, 2006] L. Egghe and R. Rousseau, An informetric model for the Hirsch-index, Scientometrics 69 (1) 121-129 (2006)

[Egghe, 2005] L. Egghe. Power Laws in the Information Production Process: Lotkaian Informetrics, Elsevier (2005)

[Egghe, 2006] L. Egghe. Theory and practice of the g-index, Scientometrics 69 (1) 131-152 (2006)

[Ferrara and Romero, 2012] E. Ferrara and A. Romero. A scientific impact measure to discount self-citations, submitted to J. Am. Soc. Information Sci and Technology (2012)

[Harzing, 2012] A.-W. Harzing. Publish or Perish v. 3.6.449, http://www.harzing.com (2012)

[Hirsch, 2005] J.E. Hirsch. An index to quantify an individual's scientific research output, Proc. Nat. Acad. Sci. 102 (46) 16569-16572 (2005)

[Hirsch, 2007] J.E. Hirsch. Does the h index have predictive power?, Proc. Nat. Acad. Sci. 104 (49) 19193-19198 (2007)

[Iglesias and Pecharroman, 2007] J. E. Iglesias and C. Percharroman. Scaling the h-index for different ISI fields, Scientometrics 73 (3) 303-320 (2007)

[Schreiber, 2007] M. Schreiber. Self-citation corrections for the Hirsch index, Europhysics Lett. 78 (3) 30002 (2007)

[Schreiber, 2008] M. Schreiber. A modification of the h-index: the h(m)-index accounts for multi-authored manuscripts, Journal of Informetrics 2 (3) 211-216 (2008)

[Schreiber, 2010] M. Schreiber. Twenty Hirsch index variants and other indicators giving more or less preference to highly cited papers, arXiv:1005.5227v1 [physics.soc-ph] (2010)

[Zhang, 2009] C.-T. Zhang. The e-index, complementing the h-index for excess citations, PloS ONE 4(5) e5429 (2009)

Authors' Information

Vladimir Atanassov *– Institute of Electronics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria; e-mail: v.atanassov@abv.bg*

Major Fields of Scientific Research: Plasma Physics and Gas Discharges, Radars and Ocean Waves, Nonlinearity and Chaos, Scientometrics

Ekaterina Detcheva *– Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,1113 Sofia, Bulgaria; e-mail: detcheva@math.bas.bg*

Major Fields of Scientific Research: Web-based applications, Image processing, analysis and classification, Knowledge representation, Business applications, Applications in Medicine and Biology, Applications in Psychology and Special Education, Computer Algebra.

SUB-OPTIMAL NONPARAMETRIC HYPOTHESES DISCRIMINATING WITH GUARANTEED DECISION

Fedor Tsitovich, Ivan Tsitovich

Abstract: *We study the problem of testing composite hypotheses versus composite alternatives when there is a slight deviation between the model and the real distribution. The used approach, which we called sub-optimal testing, implies an extension of the initial model and a modification of a sequential statistically significant test for the new model. The sub-optimal test is proposed and a non-asymptotic border for the loss function is obtained. Also we investigate correlation between the sub-optimal test and the sequential probability ratio test for the initial model.*

Keywords: *statistics, robustness, sequential analysis.*

ACM Classification Keywords*: G.3 PROBABILITY AND STATISTICS - Nonparametric statistics*

Introduction

The sequential probability ratio test (SPRT) was developed by Abraham Wald [10] under the influence of Neyman and Pearson's 1933 result. Further the test was modified for composite hypotheses testing using Bayesian approach. But on practise this approach could not provide required probability of an error decision. This imperfection was rectified by using tests those guaranteed a required significance level for all distributions from the alternative hypotheses. Such tests were considered, for example, in [1], [3], and [2]. An asymptotically optimal sequential test for nonparametric composite hypotheses having controlled observations and assuming an indifference zone was obtained in [4].

In this paper we provide another modification of the SPRT which rectify the following. The SPRT relies on the assumption that the real distribution f exactly matches with one of the distributions g_i those determine the simple hypotheses $\mathcal{H}_i^s$. But this condition is not often met on practise. Due to avoid this problem we suppose to use neighborhoods of the initial distribution those constrain new composite hypotheses. The way how to extend the initial model (type of the neighborhoods of the distributions from the initial hypotheses) is a complex problem which should be solve based on an experiment nature. In this paper we provides the test for neighborhoods those can be applied in situation when sample data contain outliers.

In some case distribution tails also should be considered. In [6]–[8] there are results obtained for exponential and heavy-tailed distributions.

The next reason why the composite hypotheses should be considered is disturbance of independency. Often observations are considered independent due to simplification of the real situation, i.e it is just an approximation. The optimal strategy from [4] requires an estimation of the dependance parameters. If the dependance is weak (we call this case as Problem 1) this becomes very difficult task on practise, so that strategy can not be used as is. A consideration of the new composite hypotheses can help avoiding incorrect decisions made because of a dependency of observations.

If a dependence is more significant (Problem 2) the sequential test should include a stage of a consistent estimation of the dependency parameters. Based on the result of this stage the observations may be transformed in such way that the new observations are considered as independent. But by a discrepancy in the estimation the new observations are not actually independent, instead they should be considered as weakly dependent, so this situation can be reduced to the described above. More details could be find in [9].

The obtain test for the composite hypotheses is a robust against mentioned above deviations between the model and the real situation. Out approach is applicable if the composite hypothesis are "small" in some sense, so the

asymptotically optimal test from [4] can not be used as is, because it will extremely increase a sample size for making a consistent estimate of the real distribution. Also, there is no need to estimate the distribution from the extended composite hypotheses. We just need to find what the closest to real distribution from the initial model is and accept that hypothesis. So the main objective of the sub-optimal approach is to provide the robust test of choosing the proper hypothesis form the initial model. Also, the obtained sup-optimal procedure converges to the asymptotically optimal sequential test when the neighborhood size converges to 0.

Setting of the problem

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space and $x_1, x_2, \ldots$ be identical distributed random variables with values in a subset $X \subset \mathbf{R}$. Let $f(x)$ be their common density with respect to some nondegenerate measure μ. The data $x_1, x_2, \ldots$ generate the statistical filter $\Big\{\mathcal{F}_n\Big\}$, $\mathcal{F}_n = \sigma(x_1, \ldots, x_n)$. Let $g_i(x)$ be densities with respect to μ, those denote simple hypotheses

$$\mathcal{H}_i^s : f = g_i(x),\ i = 1, \ldots, m. \tag{1}$$

If $x_1, x_2, \ldots$ are independent and a sample can contain outliers then we are going to modify the initial simple hypotheses into composite in the following way. Let us define the neighborhoods

$$\mathcal{O}_{g_i} := \Big\{g : g = g_i(x)(1 + h(x))\Big\}, \tag{2}$$

where functions $h(x)$ are satisfy the following conditions:

$$1) \quad \sup_{x \in X} |h(x)| \leq \varepsilon < 1, \tag{3}$$

$$2) \quad \int_X g(x)\, d\mu(x) = 1. \tag{4}$$

The first condition indicates that the neighborhoods are small in some sense. The second condition just means that each function from $\mathcal{O}_{g_i}$ is a density. Let $\mathcal{P}_i$ be a set of measures with densities from $\mathcal{O}_{g_i}$. Those neighborhoods are used as the new extended composite hypotheses:

$$\mathcal{H}_i : \mathsf{P} \in \mathcal{P}_i \tag{5}$$

It is shown in [5] the way of neighborhoods using when the sample may contains outliers.

If $x_1, x_2, \ldots$ are dependent let $f_{n+1}(x|x_1, \ldots, x_n)$ be the conditional distribution of x_{n+1} given $\mathcal{F}_n$. Let $\mathcal{P}_i$ be the set of measures on $(x_1, x_2, \ldots)$, that satisfy the following conditions:

$$\forall\ \mathsf{P} \in \mathcal{P}_i, \quad \mathsf{E}_{\mathsf{P}} f_{n+1}(x|x_1, \ldots, x_n) = g_i(x),\ |f_{n+1}(x_{n+1}|x_1, \ldots, x_n) - g_i(x_{n+1})| \leq \varepsilon \quad \mathsf{P}\text{ a.s.}$$

Defined above hypotheses can be applied in cases of:

1. Weak dependance.
2. Strong mixing condition for densities.

It was shown in [9] that described above model for dependant sample can be reduced to the model (1) with neighborhoods (2).

A determination of the neighborhood type is a complex problem. It should be defined according to the experiment characters. One of them is neighborhoods those can be applied in a situation when sample data contain outliers. The next reason, why we consider those neighborhoods, is caused by the fact that often statisticians use limiting

theorems for setting hypotheses testing problems. In this case, the model distribution is just an approximation of the real distribution and the approximation accuracy is estimated according to the rate of convergence known for the used limiting theorem for setting of the problem.

The condition (3) is natural only for a compact set X. If, for example, $X = \mathbf{R}$ then it is necessary take into account distribution's tails and the condition (3) may be too strong from practical point of view. Instead of (3) we impose the condition that densities on distribution tails are bounded from above by the known functions $t_i^-(x)$ and $t_i^+(x)$, i.e. densities $\widetilde{g}_i(x)$ from modified $\mathcal{O}_{g_i}$ should satisfy the following conditions:

$$|\widetilde{g}_i(x) - g_i(x)| \leq \varepsilon\, g_i(x),\ a_i^- \leq x \leq a_i^+; \tag{6}$$

$$\widetilde{g}_i(x) \leq t_i^-(a_i^- - x),\ x < a_i^-; \tag{7}$$

$$\widetilde{g}_i(x) \leq t_i^+(x - a_i^+),\ x > a_i^+; \tag{8}$$

and the condition (4) is substituted by

$$G_i := \int_{a_i^-}^{a_i^+} g_i(x)d\mu(x) < 1, \tag{9}$$

$$\inf_{x\in A_i} g_i(x) \geq g_i^0 > 0, \tag{10}$$

where $A_i := [a_i^-; a_i^+]$ is the segment where the main part of probability is concentrated.

A sequential test d consists of a stopping time τ and a $\mathcal{F}_\tau$-measurable decision rule δ, $\delta = r$ means that H_r, $r = 0, \ldots, m$, is accepted.

Definition 1. *We call a strategy d admissible if it satisfies the following conditions:*

$$\forall\, i \neq j,\ \sup_{\mathsf{P}\in\mathcal{P}_j} \mathsf{P}(\delta = i) \leq \alpha,\ 0 < \alpha < 1. \tag{11}$$

The conditions (11) means that the test is α level significant for each distribution from $\mathcal{P} := \cup_{i=1}^m \mathcal{P}_i$. The class of such strategies is denoted by $\mathcal{D}(\alpha)$.

As a loss function we use a sample size, this brings to the following definition of a risk function.

Definition 2. *The risk function of $d = \langle\tau, \delta\rangle$ is* $R_{\mathcal{H}_i}(d) := \sup_{\mathsf{P}\in\mathcal{P}_i} \mathsf{E}_\mathsf{P}\tau$.

We take this risk function because we do not estimate the probability low P and the strategy d needs to be good for any low from $\mathcal{P}_i$ if the hypothesis $\mathcal{H}_i$ is true.

In this paper we will analyze how extension on the initial model impacts the risk function. Define the main term of the risk function as

$$J_{\mathcal{H}_i}(d) = \lim_{\alpha\to 0} \frac{R_{\mathcal{H}_i}(d)}{|\ln\alpha|}.$$

Definition 3. *A strategy $d^* \in \mathcal{D}(\alpha)$ is called sub-optimal for the hypotheses (5) discriminating if*

$$\lim_{\varepsilon\to 0} J_{\mathcal{H}_i}(d^*) = \lim_{\varepsilon\to 0} \inf_{d\in\mathcal{D}(\alpha)} J_{\mathcal{H}_i}(d).$$

Sub-optimal strategy d_0 description

For a simplicity of notations, we suppose that $m = 2$, i.e. we test $\mathcal{H}_1$ versus alternative $\mathcal{H}_2$. For $\mathsf{P} \in \mathcal{P}$ Define $A(\mathsf{P})$ as the alternative hypotheses, i.e. $A(\mathsf{P}) := \mathcal{P}_2$ if $\mathsf{P} \in \mathcal{P}_1$ and $A(\mathsf{P}) := \mathcal{P}_1$ if $\mathsf{P} \in \mathcal{P}_2$. Let $\mathsf{I}(f, g)$ be the Kullback–Leibler information number, i.e.

$$\mathsf{I}(f, g) := \mathsf{E}_f z_{f,g}(x) := \int_X z_{f,g}(x) f(x) d\mu$$

where $z_{f,g}(x) := \ln \frac{f(x)}{g(x)},\ x \in X.$

Let us define statistics $L_i(n)$ that will be the base for out stopping rule:

$$l_{g_i}(g;n) := \sum_{k=1}^{n} z_{g_i,g}(x_k),\ L_i(n) := \inf_{g \in A(g_i)} l_{g_i}(g;n). \tag{12}$$

Then the stopping time τ_0 is

$$\tau_0 := \min\{n :\ \max_{i=1,2} L_i(n) \geq -\ln \alpha\}$$

and the decision rule δ_0 is defined by the following $\delta_0 = i$ if $L_i(\tau) \geq -\ln \alpha$. This definition is correct because if $L_1(n) > 0$ then $L_2(n) < 0$ and conversely.

If X is compact then

$$L_1(n) = \sum_{i=1}^{n} \ln \frac{g_1(x_i)}{g_2(x_i)} - n\ln(1+\varepsilon) = l_{g_1}(g_2;n) - n\ln(1+\varepsilon)$$

and $L_2(n) = -l_{g_1}(g_2;n) - n\ln(1+\varepsilon)$.

We can see that the statistics L_i are similar to the statistics used in stopping rule of the SPRT of simple hypotheses $\mathcal{H}_i^s$, but for each observation a new term in $L_i(n)$ is less on $\ln(1+\varepsilon)$ than the corresponding term of the Wald's statistic because of uncertainness in the probability low definition.

For the unbounded X and defined by (6)–(9) composite hypotheses we get more complicated formulas for the statistics $L_i(n)$:

$$L_1(n) = \sum_{i=1}^{n} \ln \frac{g_1^*(x_i)}{\tilde{g}_2^*(x_i)},\ L_2(n) = \sum_{i=1}^{n} \ln \frac{g_2^*(x_i)}{\tilde{g}_1^*(x_i)}$$

where

$$\tilde{g}_i^*(x) = \begin{cases} t_i^-(x - a_i^-), & \text{if } x < a_i^-; \\ g_i(x)(1+\varepsilon), & \text{if } a_i^- \leq x \leq a_i^+; \\ t_i^+(x - a_i^+), & \text{if } x > a_i^+ \end{cases} \quad g_i^*(x) = \begin{cases} t_i^{*-}(x - a_i^-), & \text{if } x < a_i^-; \\ g_i(x)(1+c_i), & \text{if } a_i^- \leq x \leq a_i^+; \\ t_i^{*+}(x - a_i^+), & \text{if } x > a_i^+. \end{cases} \tag{13}$$

Here functions t_i^{*-} and t_i^{*+} satisfy to (7) – (8), c_i is obtained from the equation

$$\int_{-\infty}^{+\infty} g_i^*(x) d\mu(x) = 1$$

should satisfy to the condition $|c_i| \leq \varepsilon$.

Results

The lower bound for an admissible strategy is obtained in the following

Theorem 1. *If $d \in \mathcal{D}(\alpha)$ then*

$$R_{\mathcal{H}_1}(d) \geq \frac{(1-2\alpha)(|\ln \alpha| + \ln(1-\alpha))}{\inf\limits_{p_i(x) \in \mathcal{G}_i} \inf\limits_{p(x) \in A(g_i)} \mathsf{I}(p_i, p)}.$$

The test d_0 defined above is admissible according to the following result.

Theorem 2. $d_0 \in \mathcal{D}(\alpha)$.

For the test d_0 we derived the non-asymptotic upper bound for the risk function.

Theorem 3. *Assume the conditions (3) and (4), define*

$$I^-(g_1,g_2) := (1-\varepsilon)\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^+ - (1+\varepsilon)\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^-.$$

If $\mathsf{E}_{g_1}|\ln\frac{g_1(x)}{g_2(x)}|^{1+b} \le C_1 < \infty$ *for* b *such that* $1 > b > 0$*, then the risk function of the test* d_0 *is bounded from above as followed*

1.
   - *if* $0 < b < \frac{1}{2}$ *then* $R_{\mathcal{H}_1}(d_0) \le \dfrac{|\ln\alpha| + K_1|\ln\alpha|^{1-b} + K_2|\ln\alpha|^{1-2b} + K_3}{I^-(g_1,g_2) - \ln(1+\varepsilon)}$,
   - *if* $b = \frac{1}{2}$ *then* $R_{\mathcal{H}_1}(d_0) \le \dfrac{|\ln\alpha| + K_1|\ln\alpha|^{\frac{1}{2}} + K_2'|\ln|\ln\alpha|| + K_3'}{I^-(g_1,g_2) - \ln(1+\varepsilon)}$,
   - *if* $\frac{1}{2} < b < 1$ *then* $R_{\mathcal{H}_1}(d_0) \le \dfrac{|\ln\alpha| + K_1|\ln\alpha|^{1-b} + K_3}{I^-(g_1,g_2) - \ln(1+\varepsilon)}$.
2. *If* $\mathsf{E}_{g_1}\left|\ln\frac{g_1(x)}{g_2(x)}\right|^2 \le C_1 < \infty$ *then* $R_{\mathcal{H}_1}(d_0) \le \dfrac{|\ln\alpha| + K_4}{I^-(g_1,g_2) - \ln(1+\varepsilon)}$ *where the constant* K_4 *does not depend on* α *and* g_1*-distribution from* $\mathcal{P}_1$.
3. *If* $\inf_{x\in X} g_i(x) =: G_1^- > 0, \sup_{x\in X} g_i(x) =: G_i^+ < \infty$ *then* $R_{\mathcal{H}_1}(d_0) \le \dfrac{|\ln\alpha| + K_4}{I^-(g_1,g_2) - \ln(1+\varepsilon)}$ *and* $K_4 = \frac{G_1^+}{G_2^-}$.

In this formulas

$$K_1 := \frac{(1+\varepsilon)}{b(1-b)(I^-(g_1,g_2) - \ln(1+\varepsilon))},\ K_2 := \frac{(1+\varepsilon)(1-b)C_2}{b(1-2b)(I^-(g_1,g_2) - \ln(1+\varepsilon))},$$

$$K_2' := \frac{(1+\varepsilon)C_2}{I^-(g_1,g_2) - \ln(1+\varepsilon)},$$

$$K_3 := \frac{(1+\varepsilon)}{I^-(g_1,g_2) - \ln(1+\varepsilon)}\left[\left(u_0 + \frac{1}{bu_0^b}\right)(u_0 + C_2u_0^{1-b}) - \frac{u_0^{1-b}}{b(1-b)^2} - \frac{C_2u_0^{1-2b}}{b(1-2b)}\right],$$

$$K_3' := \frac{(1+\varepsilon)}{I^-(g_1,g_2) - \ln(1+\varepsilon)}\left[\left(u_0 + \frac{2}{\sqrt{u_0}}\right)(u_0 + C_2\sqrt{u_0}) - 8\sqrt{u_0} - \frac{C_2\ln u_0}{2}\right],$$

$$u_0 := C_1^{\frac{1}{1+b}},\ a_1 = \mathsf{E}_f\nu_{u_0},\ C_2 := \frac{(1+\varepsilon)\,a_1}{b(1-b)u_0}.$$

Theorem 4. *On propositions of the theorem 3 the strategy* d_0 *is sub-optimal, i.e.*

$$\lim_{\varepsilon\to 0} J_{\mathcal{H}_i}(d_0) = \frac{1}{\mathsf{I}(g_i,g_j)} = \lim_{\varepsilon\to 0}\inf_{d\in\mathcal{D}(\alpha)} J_{\mathcal{H}_i}(d),$$

and

$$J_{\mathcal{H}_i}(d_0) \le \frac{1}{\mathsf{I}(g_i, g_j)} + \frac{1 + \mathsf{I}(g_i, g_j)}{\mathsf{I}(g_i, g_j)^2}\,\varepsilon + o(\varepsilon).$$

Define for $p_1 \in \mathcal{G}_1$ and $p_2 \in \mathcal{G}_2$

$$I_1(p_1) := \int_{-\infty}^{+\infty} \ln\left(\frac{g_1^*(x)}{\widetilde{g}_2^*(x)}\right) p_1(x)d\mu(x),\ I_1(p_2) := \int_{-\infty}^{+\infty} \ln\left(\frac{g_2^*(x)}{\widetilde{g}_1^*(x)}\right) p_2(x)d\mu(x).$$

Theorem 5. *If desctrubitions from $\mathcal{P}$ satisfy to (6)–(8) and $\mathsf{E}_{p_1}|\ln\frac{g_1^*(x_i)}{\widetilde{g}_2^*(x_i)}|^{1+b} \le C_1 < \infty$ uniformly for all $p_1 \in \mathcal{G}_1$, then*

$$R_{\mathcal{H}_1}(d_0) \le \frac{|\ln\alpha| + K_1|\ln\alpha|^{1-b} + K_2|\ln\alpha|^{1-2b} + K_3}{I_1(p_1)},$$

where the constants K_1, K_2, and K_3 do not depend on α and distribution $p_1 \in \mathcal{P}_1$.
If $\mathsf{E}_{p_1}|\ln\frac{g_1^*(x_i)}{\widetilde{g}_2^*(x_i)}|^2 \le C_1 < \infty$ uniformly for all $p_1 \in \mathcal{G}_1$, then

$$R_{\mathcal{H}_1}(d_0) \le \frac{|\ln\alpha| + K_4}{I_1(p_1)} \tag{14}$$

where the constant K_4 does not depend on α and distribution $p_1 \in \mathcal{P}_1$.
If $\sup_{x\in X}\ln\frac{g_1^*(x_i)}{\widetilde{g}_2^*(x_i)} \le K_5$ where the constant K_5 does not depend on α and distribution p_1, then

$$R_{\mathcal{H}_1}(d_0) \le \frac{|\ln\alpha| + K_5}{I_1(p_1)}. \tag{15}$$

In contrast with Theorem 3, the upper bounds obtained in Theorem 5 may be not close to the lower bound in Theorem 1 even if ε and $1 - G_i$ are very small.

Some numerical results

In this section we present some simulation results for the suboptimal tests described above. Here we consider the case when X is a compact. Let X be the segment $[0; 1]$. Let densities g_1 and g_2 be such that

$$g_1(x) = 1,\ \text{if } x \in [0; 1],\ g_2(x) = \begin{cases} 0,2, & \text{if } x \in [0; 0,5]\,; \\ 1,8, & \text{if } x \in (0,5;\ 1]\,. \end{cases}$$

The neighborhoods of the hypotheses $g_i(x)$ are

$$\mathcal{G}_1 = \Big\{\widetilde{g}_1(x)\,|\,\forall\, x \in [0; 1], \quad |\widetilde{g}_1(x) - 1| \le \varepsilon\Big\},$$

$$\mathcal{G}_2 = \Big\{\widetilde{g}_2(x)\,|\,\forall\, x \in [0; 0,5]\,,\ |\widetilde{g}_2(x) - a| \le \varepsilon, \forall\, x \in [0,5; 1]\,,\ |\widetilde{g}_2(x) - (2 - a)| \le \varepsilon\Big\}.$$

Let $z_1, z_2, \ldots$ be a sequence of uniformly distributed random numbers on $[0, 1]$. A sample $x_1, x_2, \ldots$ is calculated based on $z_1, z_2, \ldots$ according to the formula

$$x_i := \begin{cases} z_i(1 + \varepsilon), & \text{if } z_i \in [0;\ 0{,}5], \\ 1 - (1 - z_i)(1 - \varepsilon), & \text{if } z_i \in (0{,}5;\ 1]. \end{cases}$$

The distribution function of x_i satisfies to the condition (3).

Define the following notations: R_{SPRT} is the expected sample size of SPRT; R is the expected sample size of the suboptimal test; P_{SPRT} is the error probability of SPRT; P is the error probability of the suboptimal test.

Table. *Numerical results based on 10000 simulations*

α	ε	R_{SPRT}	R	P_{SPRT}	P
0.01	0.05	13.11	16.73	0.0056	0.0034
0.001	0.05	18.12	20.14	0.001	0.0001
0.01	0.1	14.4	18.38	0.01	0.0029
0.001	0.1	20.16	25.93	**0.0016**	0.0001
0.01	0.15	15.86	26.51	**0.0178**	0.0023
0.001	0.15	22.52	36.74	**0.0032**	0.0001

The simulation results mentioned in the table above shows that probability errors made by the SPRT test can increase the significance level α in the case of the uncertainty (2) (driven by ε). That can not be explain by statistical error because P_{SPRT} does not get into the confidence interval of level 0.995. Let us note that the error fraction increases if α decreases, i.e. if $\varepsilon = 0.15$ and $\alpha = 0.01$ then P_{SPRT} is 1.7 times greater than the significance level α. If $\varepsilon = 0.15$ and $\alpha = 0.001$ then P_{SPRT} is 3.2 times greater than the significance level α. On the contrary, the suboptimal test provides the required error probability, because the additional term $\ln(1+\varepsilon)$ compensates the uncertainty in the model.

Conclusion

If there is deviation in the hypotheses discrimination problem then the initial model should be extended due to reflect the known a priori information about possible deviation. This approach leads us to considering nonparametric sets of probability distributions those are neighborhood of the initial distributions. A statistically significant test for the obtained composite hypotheses becomes robust for the initial problem.

In the case of a compact codomain of observations X it is possible to use as the neighborhoods all densities those relative error against the known densities are uniformly bounded from above (see (3)). In the case of an unbounded X it is necessary to consider the distributions tails decrease rate because it may significant impact the decision rule.

The special approach called sup-optimal was introduced. It allow get a robust stoppling rule with a risk function close to the risk function of the optimal tests.

An influence of dependence sample elements is essential and impact the error probability. If the test is applicable only for independent observations, but indeed a sample is dependant, then error probability can be greater then the promised value α. Developing a statistic of a robust test we should consider a priory information on dependence.

In case of the Problem 2 ε becomes a random value and if we are going to make a robust test with level of significance α we can get ε_0 such that $\mathrm{P}(\varepsilon < \varepsilon_0) < \alpha/2$ and construct $d_0 \in \mathcal{D}(\alpha/2)$ assuming $\varepsilon = \varepsilon_0$. A sample size required for robust desction d_0 increases when ε decreases, but number of observations required for estimation of the dependence parameters decrease when ε decrease, therefore there is a problem of finding the optimal value of parameter ε_0 in the case of Problem 2.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[1] H. Chernoff. Sequential Analysis and Optimal Design. Philadelphia: SIAM. 1972.

[2] V. Dragalin, A. Novikov. Adaptive Sequential Tests for Composite Hypotheses. Statistics and Control of Random Processes. Frontiers in Pure and Applied Math. Moscow: TVP Publishers. V. 4. 1995. P. 12–23.

[3] J. Kiefer, J. Sacks. Asymptotically Optimal Sequential Inference and Design. Ann. Math. Statist. 1963. V. 34, no. 3. P. 705–750.

[4] M. Malyutov, I. Tsitovich. Asymptotically Optimal Sequential Hypothesis Testing. Problems of Information Transmission. 2000. V. 36, no. 4. P. 370–382.

[5] F. Tsitovich. Some sub-optimal strategies of hypotheses discriminating. Proceedings of the conference Information Technologies and Systems ITS'07. Moscow: IITP. 2007. P. 110–115. (in Russian)

[6] F. Tsitovich. Sub-optimal sequential tests for hypotheses of distributions with exponential tails. Proceedings of the conference Information Technologies and Systems ITS'09. Moscow: IITP. 2009. P. 416–422. (in Russian)

[7] F. Tsitovich. Sub-optimal sequentioal tests and observation errors. Proceedings of the conference Information Technologies and Systems ITS'10. Moscow: IITP. 2010. P. 252–257. (in Russian)

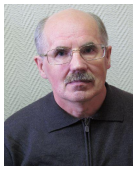
[8] F. Tsitovich. Properties of Suboptimal Sequesntial Tests of Testing Nonparametric Hypotheses with Exponentially Decreasing Tails. Journal of Communication Technology and Electronics. 2011. V. 56, no. 6. P. 748–757.

[9] I. Tsitovich. Sub-optimal Nonparametric Hypotheses Discriminating from Small Dependent Observations. Pliska. Studia mathematica Bulgaria. 2009. V. 19. P. 283–292.

[10] A. Wald. Sequential Analysis. New York, Wiley. 1947.

Authors' Information

Fedor Tsitovich *- Associate professor; National research university Higher school of economics; 20 Myasnitskaya Ulitsa, Moscow, 101000, Russia; e-mail: ftsitovich@gmail.com*
Major Fields of Scientific Research: Mathematical statistics and its applications



Ivan Tsitovich *- Associate professor; National research university Higher school of economics; Chief Scientific Researcher, Institute for Information Transmission Problems, RAS, Bol'shoi Karetnyi per., 19, Moscow, 127994, Russia; e-mail: cito@iitp.ru*
Major Fields of Scientific Research: Mathematical statistics and its applications, Teletraffic theory

LINGUISTIC AND PROGRAM TOOLS FOR DEBUGGING AND TESTING OF SIMULATION MODELS OF COMPUTER NETWORKS

Elena Zamyatina, Alexander Mikov, Roman Mikheev

***Abstract**: This paper focuses on the problem of validation and verification of computer network simulation models. Authors propose to use special linguistic and program tools of CAD system TriadNS in this case. First of all it should be noted that TRiadNS is a computer system which was developed for computer network design. Simulation is the main method for investigation of designed computer networks. But it is very important to have a credible simulation result. It is necessary for target users to have sufficient confidence that results generated by a simulation run reflect real world operation to a large degree. Authors observe the specifications of the simulation model in TriadNS.Net, consider the program tools for simulation model analysis (information procedures and conditions of simulation) and propose to use them for simulation model validation and verification, debugging and testing. Besides, the authors suggest program tools including the intellectual agents and ontology for localization of mistakes determined during verification and validation processes. Moreover the authors show how the specific features of hierarchical simulation models in TRIADNS make the process of testing and debugging of simulation models flexible.*

***Keywords**: simulation model, debugging, testing, computer networks, validation, verification.*

***ACM Classification Keywords**: I.6 SIMULATION AND MODELING: I.6.8 Types of Simulation – Distributed. I.2 ARTIFICIAL INTELLIGENCE: I.2.5 Programming Languages and Software – Expert system tools and techniques.*

Introduction

The role of computer networks is rather high nowadays. Computer networks are widely used in distributed processing of information. The evidence of this is the widespread of corporate information systems, Grid-technologies, cloud computing. One more example – social networking without which many people do not realize their life.

Widespread computer networks impose demands for speed and reliable information transmission, for efficient processing of information. For this reason, it becomes necessary to study the traffic, new efficient protocols, new algorithms (for example, routing algorithms), to study new devices and control algorithms for these devices, to investigate new types of computer networks and sufficient principles of it functioning.

Analytical methods are not always possible to apply for the study of computer networks because of the complexity of this object of research. The field experiments do not give the opportunity to explore all aspects of the designed network. So the researchers have to use the methods and software tools for simulation. More precisely, it is efficient to apply the linguistic and program tools of network simulation. There is a large number of such software [Salmon, 2011].

The primary purpose of this paper is to discuss one of the approaches of scientific information quality enhancing. It is well known that qualifying standards are rather high for scientific information received during some scientific experiment. The validity of scientific information means a high degree of conformity between scientific results and

a problem to be solved. It is very important to apply the proper method, to find a corresponding approach, to create a respective mathematic model. And it is actual for simulation and simulation model too. [Соколов, 2005]. The problems of validity are coupled with problems of debugging and testing.

Some authors [Bagrodia, 1999] picked up such problems of network simulation model validity as *simulator validation, protocol validation* («does the simulation model of a given network protocol faithfully replicate details of the protocol based on its specification or implementation», is, for example, TCP model provided by NS library correct with respect to actual TCP implementations), *system validation* (the simulation of physical resources using in order to identify bottlenecks, such as processing delays and overheads), *scenario validation* (the degree of sensitivity of results to minor (or major) variations in the critical parameters of the assumed scenario) and so on.

Authors suggest linguistic and program tools for simulation model validation, more precisely, debugging and testing. These tools and an expert component for mistake localization are considered below. Authors suggest using a linguistic constructions "information procedures" and "conditions of simulation" for debugging and testing. First of all let us consider what is meant by the terms "validation" and "verification".

Verification and Validation

The simulation model is a representation or abstraction of something such as entity, or a system. So a model can't be perfect because it is an abstraction. But we intend to build a credible simulation model and to receive the credible results of simulation. Both the modelers and users of simulation model are interested in reliable simulation results because it is very important to accept the right decision.

Verification and validation (V&V) are considered usually as a single process of M&S (Modeling and Simulation) but it is not accurate reasoning because each of them purposes on the different aim.

The aim of verification is to be sure that simulation model implementation is proper, to determine whether it corresponds to conceptual description and specification. Simulation model verification "is substantiating that the model is transformed from one form into another, as intended, with sufficient accuracy. Model verification deals with building the model right [Balci, 1998].

The aim of validation is to determine the degree to which a model is an accurate representation of a real system from the perspective of intended use of the model. Model validation "is substantiating that the model, within its domain of applicability, behaves with satisfactory accuracy consistent with the M&S objectives. Model validation deals with building the right model [Balci, 1998] .

In addition we must say about model debugging and testing. Model debugging supposes the detection of mistakes, its localization and elimination. Simulation model testing "is ascertaining whether inaccuracies or errors exist in the model [Balci, 1998].

We can mention one more definition coupled with verification and validation. It is accreditation – a statement of M&S sponsors that simulation results are intended for use. Accreditation is "the official certification that a model or simulation is acceptable for use for a specific purpose" [Balci, 2002].

Related works

V&V is well known problem. It is discussed in many papers. So one may become acquainted with these problems thanks to numerous publications of [Balci, 1998, 2002; Law, 2004; Sargent, 2005, 2007] and the other authors.

These papers present different paradigms of verification and validation, define different stages, give recommendations and consider methods but most of them don't show how to do it.

Structural and Operational validity in TRIADNS

Let us consider the simulation model validation more precisely. It is well known that V&V must be fulfilled on different levels (input data, simulation model elements, simulation model subsystems and its interconnections).

The testing of simulation model adequacy and accuracy includes it's structure testing (one must determine whether the structure of the simulation model, a list of objects and their interconnections correspond to investigator's intentions, let us name this type of validity as the structural one), primitive functions testing, behavior testing (one must examine whether the simulation model functionality corresponds to investigator's concepts. Let us name this type of validity as an operational one). Moreover, simulation model validation has to be executed at each stage of simulation model design. It is necessary to return to the previous stage if the process of simulation model validation shows some errors.

Structural simulation model validity may be fulfilled by functions and procedures of Triad-model structure layer. These functions and procedures may test the topological specification of a simulation model in particular. Operational validation is defined as the process of confirming that simulation results closely approximate real world results. The operational validity may be carried out by the information procedures. Information procedures allow to determine whether any value of simulation model variables at some concrete moment of simulation time is equal to the specific one in right (valid) model and so on. (We shell discuss the information procedures possibilities more precisely below). If it is not so (the values are not equal) then the cognitive agents will define the type of mistake using specific ontology and localize it following some rules which use knowledge about simulation model structure and behavior. Let us consider the specification of simulation model in Triad. We must remind that CAD TRIADNS [Замятина, 2012] is intended for computer network design and analyses and therefore it has some specifications.

Simulation Model Representation in TRIADNS

Program model in Triad.Net is represented by several objects functioning according to some scenario and interacting with one another by sending messages. Program model [Mikov, 1995] is μ={*STR*, *ROUT*, *MES*} and it consists of three layers, where *STR* is a layer of structures, *ROUT* – a layer of routines and *MES* – a layer of messages appropriately. The layer of structure is dedicated to describe objects and their interconnections, but the layer of routines presents their behavior. Each object can send a message to another object. So, each object has the input and output poles (P_{in} – input poles are used to send the messages, P_{out} – output poles serve to receive the messages). One level of the structure is presented by graph *P* = {*U*, *V*, *W*}. P-graph is named as graph with poles. A set of nodes *V* presents a set of programming objects, *W* – a set of connections between them, *U* – a set of external poles. The internal poles are used for information exchange within the same structure level; in contrast, the set of external poles serves to send messages to the objects situated on higher or underlying levels of description. Special statement ***out*** <*message*> ***through*** <*name of pole*> is used to send the messages.

One can describe the structure of a system to be simulated using such a linguistic construction:

structure <*name of structure*> ***def*** (<*a list of generic parameters*>)
(<*a list of input and output parameters*>)
<*a list of variables description*> <*statements*>) ***endstr***

Special algorithms (named "routine") define the behavior of an object. It is associated with particular node of graph P = {U, V, W}. Each routine is specified by a set of events (E-set), the linearly ordered set of time moments (T-set), and a set of states {Q-set}. State is specified by the local variable values. Local variables are defined in routine. The state is changed if an event occurs only. One event schedules another event. Routine (as an object) has input and output poles (Prin and Prout). An input pole serves to receive messages, output – to send them.

One can pick out input event e_{in}. All the input poles are processed by an input event, an output poles – by the other (usual) event.

Routine <имя> (<*a list of generic parameters*>) (<*a list of input and output formal parameters*>)
initial <*a sequence of a statements*> ***endi***
event <*a sequence of a statements*> ***ende***
event <*a name of an event*> <*a sequence of statements*> ***ende*** …
event<*a name of an event*><*a sequence of a statements*> ***ende endrout***

There are two possibilities to build a simulation model of a computer network: via text or graphical editor.

Let us present a simulation model in Triad (one can pick it out at fig.1.). Here is a fragment of computer network.

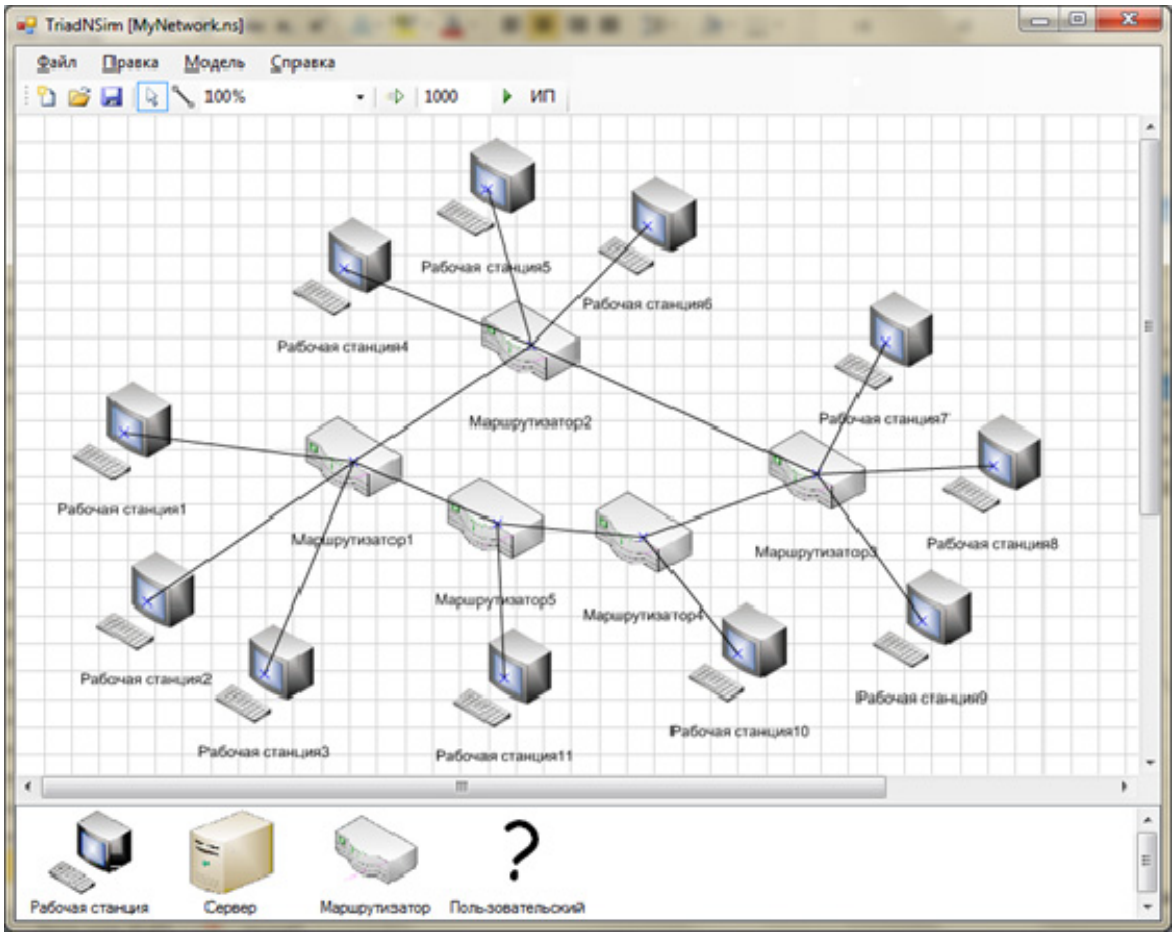


Figure 1. A fragment of computer network

Computer system consists of some workstations and routers and provides message sending and receiving. The structure of simulation model may be described by graph constant ***dcycler,*** three of these graph constants are connected with 3 nodes associated with work stations. Eight nodes represent workstations Hst [i]. These nodes are added in cycle. Each router is intended to fulfill the same algorithm. So designer defines the same routine Router (appropriate program module is saved in data base) for each router (node Rout [i] of structure) and the same routine Host to each workstation (node Hst [i]).

A program model in Triad.Net isn't static. Triad language includes the special type of variables – type "model". There are several operations with the variable type "model". The operations are defined for the model in general and for each layer. For example, one may add or delete a node, add or delete an edge (arc), poles, union or

intersection of graphs. Routine layer permits to add or delete any event, layer of messages – to add or delete types or selectors. Besides, one or another routine can be assigned to the node in structure layer (using some rules). As a result the behavior of the object associated with this node would be changed.

```
Type Router,Host; integer i;
M:=dcycle(Rout[5]<Pol>[5]); M:=M+node (Hst[11]<Pol>);
for i:=1 by 1 to 5 do
    M.Rout[i]=>Router; M:=M+edge(Rout[i].Pol[1] — Hst[i]);
endf
for  i:=1 by 1 to 3 do M:=M+edge(Rout[i].Pol[2] - Hst[2*I -1]); endf;
for i:=0 by 1 to11 do M.Hst[i]=>Host; endf;
```

Algorithm of investigation

The objects of simulation model are managed by the special algorithm during the simulation run. Let us name it as “simulation algorithm” (CAD system Triad has distributed version and corresponding algorithm for distributed objects of simulation model too) [Миков, 2009]. CAD system Triad includes analyses subsystem implementing the algorithm of investigation – special algorithm for data (the results of simulation run) collection and processing.

The analysis subsystem includes special objects of two types**:** *information procedures* and *conditions of simulation*. Information procedures are “connected” to nodes or, more precisely, to routines, which describe the behavior of particular nodes during simulation experiment. Information procedures inspect the execution process and play a role of monitors of test desk.

Conditions of simulation are special linguistic constructions defining the algorithm of investigation because the corresponding linguistic construction includes a list of information procedures which are necessary for investigator.

The algorithm of investigation is detached from the simulation model. Hence it is possible to change the algorithm of investigation if investigator would be interested in the other specifications of simulation model. For this one need to change the conditions of simulation. But the simulation model remains invariant. We may remind that it is not possible in some simulation systems.

One can describe the information procedure as so:

information procedure<*name*>*(*<*a list of generic parameters*>*)(*<*input and output formal parameters*>*)*
initial <*a sequence of statements*> ***endi***
<*a sequence of statements*>
processing <*a sequence of statements*>*…* ***endinf***

It is possible to examine the value of local variables, the event occurrence and the value of messages which were sent or received. A part of linguistic construction ‘processing’ defines the final processing of data being collected during simulation run (mean, variance and so on).

Let us present the linguistic construction ***conditions of simulation***:

Conditions of simulation<*name*>(<a list of generic parameters>)(<input and output formal parameters>)
initial <a sequence of statements> ***endi***
<a list of information procedures> <a sequence of statements>
processing <a sequence of statements>…***endcond***

The linguistic construction ***conditions of simulation*** describes the algorithm of investigation which defines not only the list of information procedures but the final processing of some information procedure and checks if conditions of simulation correspond to the end of simulation.

simulate <a list of an elements of models, being inspected>

on conditions of simulation <name> (a list of actual generic parameters>)

[<a list of input and output actual parameters>]

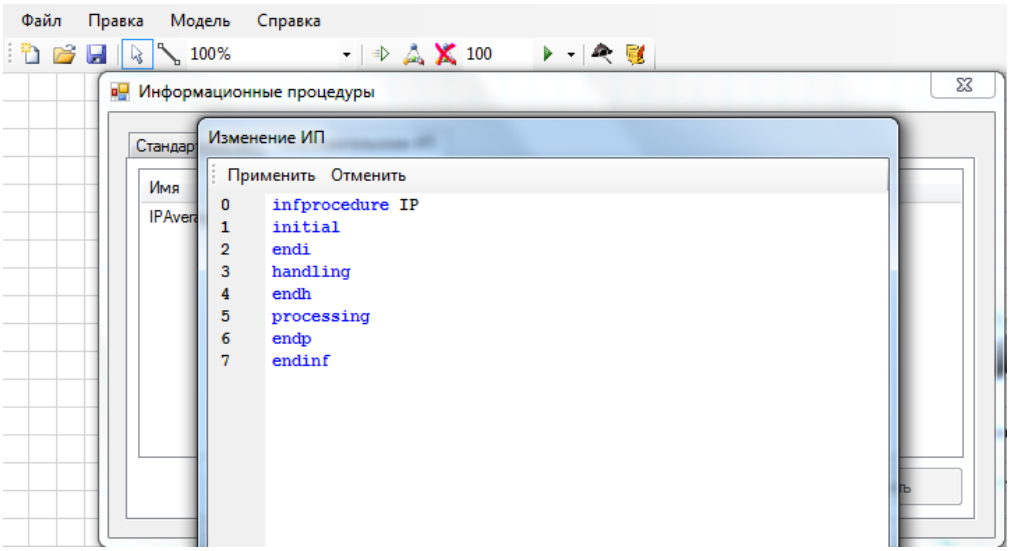


Figure 2. The form for information procedure

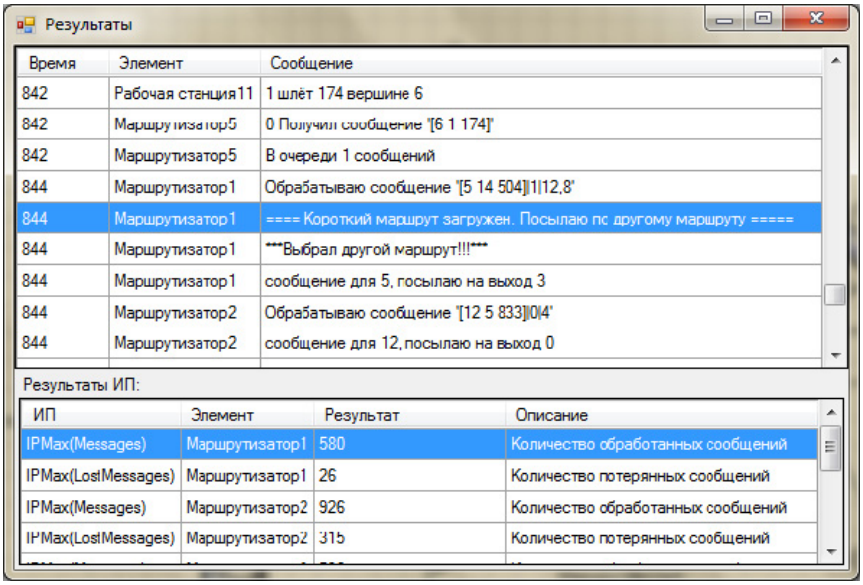


Figure 3. The results of simulation

Simulation run is initialized after simulation statement processing. One can pay an attention to the fact that the several models may be simulated under the same *conditions of simulation* simultaneously.

The subsystem of visualization represents the results of simulation. One can see the representation of the results of simulation run at fig.3.

Information procedures for simulation model validation

First of all let us discuss error detection in TRIADNS. Primarily we shall consider the types of errors which may be recognized by the information procedures. There are the following types of errors:

- incorrect temporary delays;
- wrong messages transfer;
- semantic incorrectness of signal conversion;
- semantic incorrectness of data exchange;
- invalid management of simulation model functioning;
- semantic incorrectness of changing of states of simulation model;
- forbidden simulation model states.

It is advisable to determine the correctness of the following values for simulation model being represented above: time slice between sending of message from one workstation and receiving of message by another one (this time slice must be less than some limiting value). Moreover it is important to be sure that the value of received message is valid and it is received by the correct input of workstation.

Information procedures are convenient not only for simulation model analyses but for simulation model debugging and validation too. CAD system TriadNS (and CAD system Triad too) has a set of standard information procedures for temporary delays recognition.

The information procedures using for debugging and testing are defined so: $DB = \{P, I, A\}$, where P is an algorithm representing information procedure action, I – a set of input parameters of this information procedures, where $I = Ip \cup Ie \cup Iv$, Ip – a set of input parameters and the type of this parameter is pole, Iv – a set of input parameters and the types of these parameters coincide with the types of local variables, Ie – a set of parameters with type *event*. Some information procedures may monitor one element of simulation model.

These procedures serve as a basis of knowledge based debugger. Information procedure *more_interval* (*t*) [*e*1, *e*2], for example, is able to fix time slice between two events *e*1 (maybe it is the event of message sending) and *e*2 (the event of message termination).

Invalid temporary delays may be recognized by information procedures *all_events* (*e*) (this procedure defines the particular beginning of event and its termination), *all_changes* (*var*) (procedure defines each instant of the time when indicated variable is changed) and so on.

Investigator may use information procedure *schedule_event (e)* (the names of all registered events which preceded the indicated one); *inspect_change* (*e*) (the values of all variables which were changed before the event *e* occurrence, event *e* is an actual argument of procedure).

So we named some standard information procedures. But an investigator may use linguistic constructions (information procedure and conditions of debugging) for specific occasions of debugging and so he will share the debugger functionality.

Let us consider the example of information procedure for the detection of the sequence of events arrival:

```
information procedure event_sequence (in ref event  E1, E2, E3; out Boolean arrived)
initial interlock (E2,E3); Arrived := false;
  case of  e1: available(e2);
           e2: available(E3):
           e3:ARRIVED:=true;
```

endc

endinf

So investigator may detect the arrival of the sequence of events $E1 \rightarrow E2 \rightarrow E3$. The statement ***interlock*** provides input parameter blocking (event $E1$ in this case). It means that information procedure doesn't watch parameters being marked in interlock statement. The statement ***available*** allows beginning the marked parameter monitoring again.

Another example concerns the problem of forbidden states of simulation model detection. Only such linguistic and programming tool as information procedure may detect this error because only information procedure may determine the value of local variables of different routines at the same moment of simulation time during simulation experiment and may compare these values. Simulation model editor and information procedures for data collection and model monitoring are presented at fig.4.

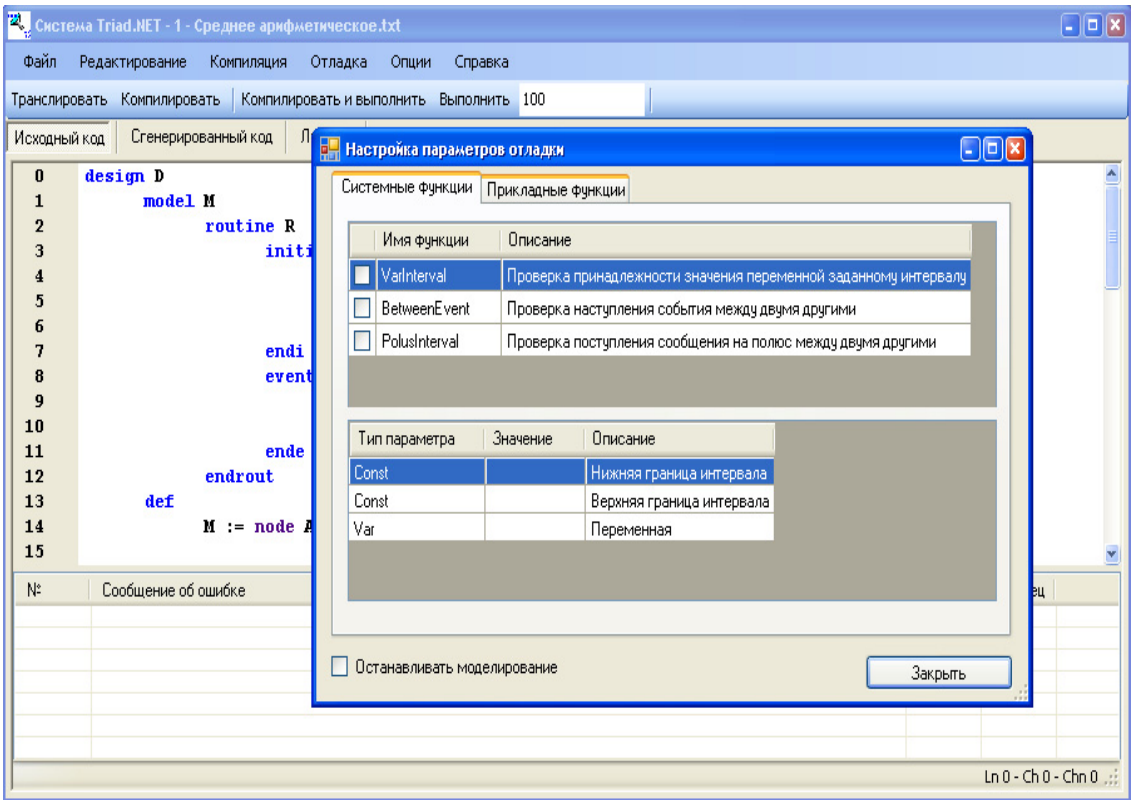


Figure 4. Simulation model editor and information procedures for algorithm of investigation and debugging

Knowledge based debugger

Thus we have discussed the problem of error debugging and testing and consider the linguistic and program tools in Triad intending to solve these problems. We briefly presented information procedures, its linguistic construction and examples of applying these tools for simulation model monitoring during the simulation run. Special linguistic construction *conditions of debugging* contains a list of the information procedures which are intended for simulation model monitoring and error detection. The construction *conditions of debugging* are a part of simulate statement. Debugger starts its work when simulation model begins to fulfill the *simulate* statement.

Let us consider such a situation: we want to be sure in data transfer correctness from the workstation Hst[1].Pol[1] (Pol[1] – it is an output polus) to workstation Hst[10].Pol[1] (fig.5). The example of simulate statement one can see below:

simulate M ***on*** Checker (M.Hst[1].Pol[1], M.Hst[10].Pol[1])

Corresponding linguistic construction **conditions of debugging** is represented below:

conditions of debugging Checker (*input real* a, b; ***output boolean*** answer)
(… Check_Value(1.0,1.0) (***input*** ia, ib; ***output*** ianswer); …).

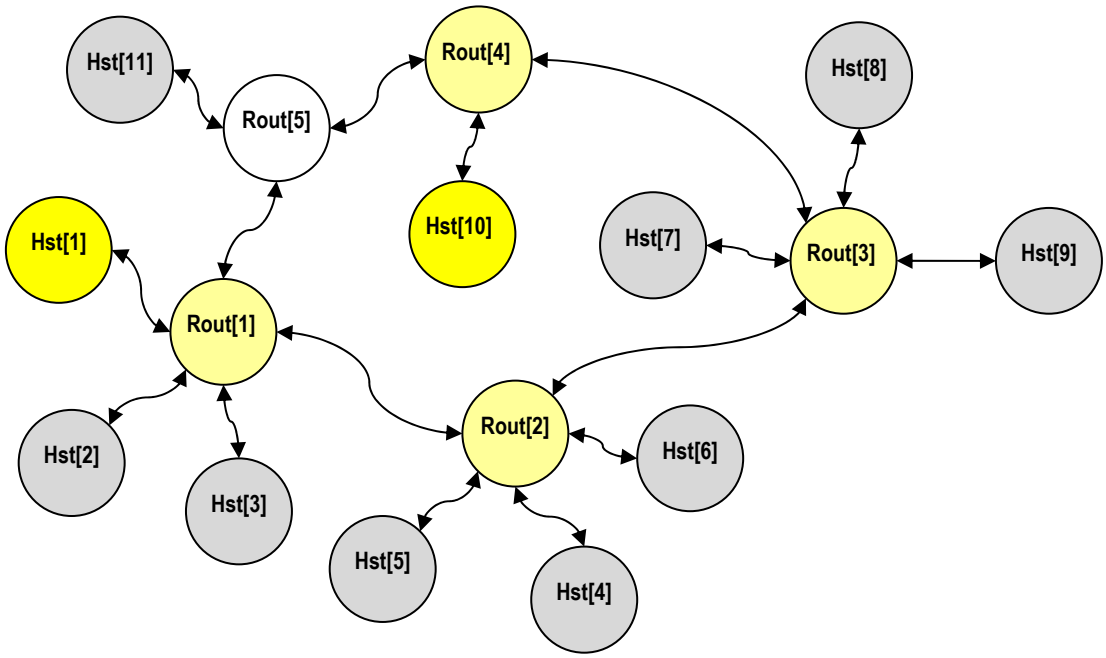


Figure 5. Graph representation of the simulation model structure

The simulation run has to be terminated because of error detection. It is necessary to localize error and neutralize it. Error localization and neutralization is knowledge based.

Information procedures are convenient not only for simulation model analyses but for simulation model debugging and validation too.

Let us return to the example (fig. 5). We shall suppose that the routine associated with node Hst[1] sends a signal to routine associated with HST[10]. Let debugger detects invalid signal received at input of this routine. So this invalid signal may be the result of invalid processing in routines Rout[1], Rout[2], Rout[3], Rout[4]. Moreover it is advisable to check the sequence of events and temporary delays validity.

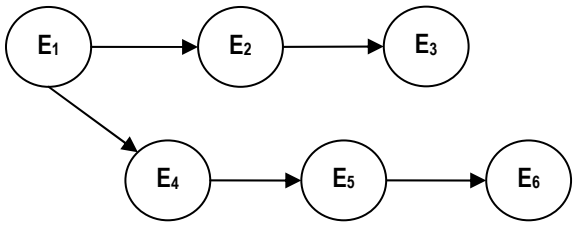


Figure 6. Graph of scheduled events

So the mistake localization may develop according to different scenarios (may conform to different rules). These rules are contained in knowledge base. More precisely cognitive agents for each type of the mistake with

appropriate rules are used for mistake localization. Debugger needs additional information: simulation model structure which is represented by $P = \{U, V, W\}$. This information is necessary to define the path message transfer. Besides, the information mentioned above is helpful to use the graph of scheduled events in order to determine the sequence of events participating in message processing (it is presented on fig.6). Debugger may use such additional information as graph of calculations. This graph provides the determination of sequence of local variables processing.

Expert system for mistake localization includes a set of procedures for debugging, knowledge base with rules 'if….then …' using all additional information mentioned above. Besides, it includes inference engine, explanation module, editor for rules and meta rules. A set of procedures for localization of mistakes may be extended because an investigator may create new information procedures using appropriate linguistic tools.

It is necessary to tell about ontology which may be used for mistake detection. Debugger will form request to ontology if the mistake of the specific type is detected. Appropriate cognitive agent with specific rules needed for an algorithm of processing mistake of specific type will be started. If there are some paths of an algorithm then some cognitive agents will be started.

Conclusion

So the authors presented the linguistic and program tools of CAD Triad.Net system needed not only for design of model but for its monitoring, data collections, analyses, validation and verification. The authors consider the linguistic and intellectual program tools for simulation model verification and validation more precisely. The appropriate linguistic tools – information procedures – are under discussion.

The architecture of the debugger is presented. The debugger is used not only to detect mistakes in the simulation model but to localize them. The debugger detects mistakes using information procedures and determines the appropriate rules of localization thanks to ontology. Proper cognitive agents which act in accordance with these specific rules start to localize mistakes. So the authors suggest solving the problem of validation and verification using multiagent approach and ontology.

So presented approach permits to automate simulation model verification and validation. The process of mistake detection becomes more effective and more flexible, but this investigations are under consideration of authors nowadays.

Acknowledgements

This work was fulfilled due to financial support of the grant of RFBR №12-07-00302-a and the grant of the Ministry of Education and Science № 8.5782.2011.

Bibliography

[Bagrodia, 1999] Bagrodia R.,Takai M. Position paper on validation of network simulation models, Proceedings of *DARPA/NIST Network Simulation, Validation Workshop*, May 1999.

[Balci, 1998] Balci O. Verification, Validation, And Accreditation. Proceedings of the 1998 Winter Simulation Conference. D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, eds.. Grand Hyatt Washington, Washington DC, pp. 41-48.

[Balci, 2002] Balci O, Nance R.E., Arthur J.D. Expanding Our Horizons In Verification, Validation And Accreditation Research And Practice Proceedings of the 2002 Winter Simulation Conference E. Yücesan, C.-H. Chen, J.L. Snowdon, and J.M. Charnes, eds., WSC 2002, San Diego, California, pp.653-663.

[Law, 2001] Law, A. M. and M. G. McComas. 2001. How to build valid and credible simulation models. In Proc. 2001 Winter Simulation Conf., ed. B.A. Peters, J. S. Smith, D. J Medeiros, and M. W. Rohrer, Piscataway, New Jersey: IEEE. pp. 22-29.

[Mikov, 1995] Mikov A.I. Formal Method for Design of Dynamic Objects and Its Implementation in CAD Systems // Gero J.S. and F. Sudweeks F.(eds), Advances in Formal Design Methods for CAD, Preprints of the IFIP WG 5.2 Workshop on Formal Design Methods for Computer-Aided Design, Mexico, Mexico, 1995. pp. 105 -127.

[Salmon, 2011] Salmon S, El Aarag H. Simulation Based Experiments Using Ednas: The Event-Driven Network Architecture Simulator. In Proceedings of the 2011 Winter Simulation Conference S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds. The 2011 Winter Simulation Conference 11-14 December 2011. Grand Arizona Resort Phoenix, AZ, pp. 3266-3277.

[Sargent, 2005], Sargent, R. G. Verification and validation of simulation models. Proceedings of Winter Simulation Conf., ed. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, Piscataway, New Jersey,2005: IEEE. pp.130 -143.

[Sargent, 2007], Sargent, R. G. Verification And Validation of Simulation Models. Proceedings of Winter Simulation Conf., S.G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds., J.W. Marriott Hotel, Washington, 2007, pp.124-137

[Замятина, 2012] Замятина Е.Б., Миков А.И., Михеев Р.А. Лингвистические и интеллектуальные инструментальные средства симулятора компьютерных сетей TRIADNS. International Journal “Information theories & Applications (IJ ITA). Vol 19, Number 4, 2012, pp.355-368.

[Миков, 2009] Миков А.И., Замятина Е.Б. Проблемы реализации системы распределенного моделирования с удаленным доступом. «Методы и средства обработки информации», Труды третьей Всероссийской научной конференции, 6-8 октября 2009 г. Москва, МАКС ПРЕСС, 2009. стр. 38-44.

[Соколов, 2005] Соколов Б.В., Юсупов Р.М. Концептуальные основы квалиметрии моделей и полимодельных комплексов // Имитационное моделирование. Теория и практика: Сборник докладов второй всероссийской научно-практической конференции ИММОД-2005. Том 1. СПб.: ЦНИИТС. 2005. – с. 65-70.

Authors' Information

Elena Zamyatina *– Perm State National Researching University, Associate Professor; Bukirev St., 15, Perm, Russia; e-mail: e_zamyatina@mail.ru.*

Fields of Scientific Research: Simulation, Distributed and Parallel Simulation, Artificial Intelligence, Computer Networks Simulation.

Alexander Mikov *– ACM Member, professor, head of the computing technologies chair, P.O. Box: Kuban State University, 149, Stavropolskaya str., Krasnodar, 350040, Russia; e-mail: alexander_mikov@mail.ru.*

Major Fields of Scientific Research: Distributed information systems, Simulation systems and languages

Roman Mikheev *– Perm State National Researching University, Junior Scientific Fellow; Bukirev St., 15, Perm, Russia; e-mail: miheev@prognoz.ru.*

Fields of Scientific Research: Simulation, Artificial Intelligence, Computer Networks Simulation, Ontologies.

THE INVERSE MASLOV METHOD AND ANT TACTICS FOR EXHAUSTIVE SEARCH DECREASING

Tatiana Kosovskaya, Nina Petukhova

***Abstract**: Algorithms for soling artificial intelligence problems allowing to formalize by means of predicate language are presented in the paper. These algorithms use the modified Maslov's inverse method. One of these algorithms is directly based on the inverse method, and the second one uses the inverse method and Ant algorithm's tactics. The model example using the described algorithms is shown.*

***Keywords**: artificial intelligence, pattern recognition, predicate calculus, inverse method of S.Yu.Maslov, complexity theory, Ant algorithms.*

***ACM Classification Keywords**: I.2.4 ARTIFICIAL INTELLIGENCE Knowledge Representation Formalisms and Methods – Predicate logic, I.5.1 PATTERN RECOGNITION Models – Deterministic, F.2.2 Nonnumerical Algorithms and Problems – Complexity of proof procedures.*

Introduction

NP-hardness of artificial intelligence problem, in particular, that allowing formalization by means of predicate calculus language [Kosovskaya, 2011] imposes the requirements to the algorithms for their solving. This requirements focused on optimize the sorting that occurs when searching for a suitable substitution, ensuring fulfillment of the target formula. Maslov's inverse method is developed to make more effective the procedure of deducibility in predicate calculus. An algorithm, which is a restriction of the inverse method to the proof of a special case formulas appearing in the decision of many artificial intelligence problems is described in [Kosovskaya, Petukhova, 2012]. A more detailed version of such an algorithm is considered in this paper. A model example of its application to the recognition of a contour image is described.

The use of multi agents involved in solving of the same problem simultaneously allows to find an analogy in the actions of a scout-ants of a real anthill. The use of this analogy in the literature has been called the Ant algorithm [Dorigo, Birattari, Stutzle, 2006]. This tactic allows to "parallelize" process of gathering information and to throw off obviously dead-end search solutions. It is suggested to apply the Ant tactic to the deduction search using the inverse method.

Initial definitions

The solving of many Artificial Intelligence problems permitting the use of predicate language may be reduced to the proof of a logical sequence in the form

$$S(\omega) \Rightarrow \exists \bar{x}_{\neq} A(\bar{x}),$$

where $\omega = \{a_1,...,a_k\}$ is a list of constants, $S(\omega)$ is a set of constant atomic formulas or their negations, $A(\bar{x})$ is an elementary conjunction of atomic formulas of the form $P_{k_i}(\bar{x})$ [Kosovskaya, 2011].

Such a logical sequence is equivalent to the truth of the formula

$$(\&\, S(\omega)) \to \exists \bar{x}_{\neq} A(\bar{x})$$

for every value of constant ω. Using the fact that ω is an arbitrary constant set this formula may be reduced to the formula

$$\forall a_1,...,a_k \exists x_1,...,x_n \left(\overset{\delta}{\underset{i=1}{\&}} \left(\vee \neg S(a_1,...,a_k) \vee P_{k_i}(x_1,...,x_{n_i}) \right) \right),$$

which is a particular case of a formula intended for the use of inverse Maslov method [Kosovskaya, Petukhova, 2012; Orevkov, 2003].

The formula (1) is deducible if and only if there exists such a substitution of the terms $t_1,...,t_{n_i}$ instead of variables $x_1,...,x_{n_i}$ that every elementary disjunction has a contrary pair. In the other words it is required to solve a system of equations in the form

$$i = \overline{1,\delta} \left\{ \begin{array}{l} \overline{1,s} \left[\begin{array}{l} \{ \\ \vdots \\ \{ \end{array} \right. \\ \vdots \\ \overline{1,s} \left[\begin{array}{l} \{ \\ \vdots \\ \overline{1,s} \left\{ \begin{array}{l} t_1 = a_1 \\ \vdots \\ t_{n_i} = a_{n_i} \end{array} \right. \\ \vdots \\ \{ \end{array} \right. \\ \vdots \\ \overline{1,s} \left[\begin{array}{l} \{ \\ \vdots \\ \{ \end{array} \right. \end{array} \right.,$$

where *s* is the number of atomic formulas in $\neg S(a_1,...,a_k)$.

A solution of such a system may be found in an exponential number of steps. The inverse method is oriented on its essential decreasing. In the addition to the ideas of the inverse method it is suggested to order all the formulas $\vee \neg S(a_1,...,a_k) \vee P_{k_i}(x_1,...,x_{n_i})$ in (1) in the following way. As every formula $\vee \neg S(a_1,...,a_k) \vee P_{k_i}(x_1,...,x_{n_i})$ contains only one disjunctive number $P_{k_i}(x_1,...,x_{n_i})$ with variables then

form groups with the same names $P_{k_i}(x_1,...,x_{n_i})$ and then order these groups according to their sizes. The same ordering must be done with the set $S(\omega)$.

First of all we search such a substitution which permits to assign variables in atomic formulas with the rarest predicate name. If such a substitution does not exist then the formula (1) is dot deducible.

If the substitution is found then it is made in all F-sets.

Repeat the procedure until all variables are changed by constants.

IMA – algorithm based on the inverse method

Definition. A list Γ of not repeated formulas in the form $\vee \neg S(a_1,...,a_k) \vee P_{k_i}(x_1,...,x_{n_i})$ is called an F-set for a formula (1) [Kosovskaya, Petukhova, 2012; Orevkov, 2003].

Definition. An F-set is called an empty one if all its formulas do not have variables and are tautological ones. [Kosovskaya, Petukhova, 2012; Orevkov, 2003].

Definition. An F-set is called a deadlock one if it contains at least one false formula without variables or a formula which is not a tautology nor a contradiction. [Kosovskaya, Petukhova, 2012]

IMA-algorithm

1. Construct an F-set corresponding to the formula (1). I.e. write down δ elementary disjunctions of the form $\vee \neg S(a_1,...,a_k) \vee P_{k_i}(x_1,...,x_{n_i})$.

2. Assign all variables in the following way:

2.1. Cancel all marks about deleting of a predicate formula from $S(\omega)$.

2.2. Consider a predicate formula $\vee \neg S(a_1,...,a_k) \vee P_{k_i}(t_1,...,t_{n_i})$ from the F-set containing such an atomic formula $P_{k_i}(t_1,...,t_m)$ that the list $t_1,...,t_m$ has at least one variable.

2.3. Check if $S(\omega)$ contains $\neg P_{k_i}(v_1,...,v_m)$ for some constants $v_1,...,v_m$. If it is so then mark $\neg P_{k_i}(v_1,...,v_m)$ as deleted and go to 2.4. Otherwise go to 3.

2.4. Solve the system of equations identifying the list of variables and constants $t_1,...,t_m$ with the list of constants $v_1,...,v_m$. If the system has a solution[1], then go to 2.5. Otherwise go to 2.3.

2.5. Substitute the values received in point 2.4 instead the variables of $t_1,...,t_m$ into all formulas in the F-set.

2.6. Delete repetitions of formulas in the received F-set.

2.7. If the received F-set is empty them the algorithm halts.

2.8. If the received F-set is a deadlock one then go to 3.

[1] Here the system may don't have a solution if some value for a variable is yet assigned to the other variable.

2.9. If all formulas containing variables is marked as deleted then go to 4. Otherwise go to 3.

3. Cancellation of assignments.

3.1. Cancel the last action of point 2.5 (if it is possible) and go to 2.3.

3.2. If cancellation of the last action of point 2.5 is not possible then mark $P_{k_i}(t_1,...,t_m)$ as deleted and go to 2.

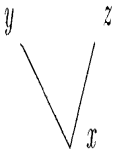
4. If all formulas in the f-set are marked as deleted then the formula is not deductible. The algorithm halts.

An upper bound of a similar algorithm is presented in [Kosovskaya, Petukhova, 2012]. It is $O(s^5)$ where s is the number of atomic formulas in $S(\omega)$. The upper bound of the presented algorithm is a similar one.

A model example

Show that in spite of a high upper bound of the IMA algorithm number of steps real number of steps is rather smaller.

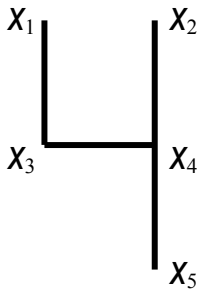
Let a set of contour images is done. It is described by means of the following predicates.



$V(x,y,z) \leftrightarrow$"$\angle yxz < \pi$"

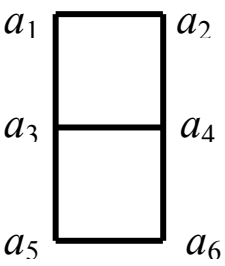
$L(x,y,z) \leftrightarrow$"*x belongs to the* int*reval with the ends y and z*"

We have the class of images of the number four



which has the description $A(x_1,\ldots,x_5) = V(x_3,x_1,x_4)\ \&\ V(x_4,x_2,x_3)\ \&\ V(x_4,x_3,x_5)\ \&\ L(x_4,x_2,x_5)$.

It is required to extract an image of the number four from the image



This image has a description $S(a_1,\ldots,a_6) = V(a_1,a_2,a_3)\ \&\ V(a_2,a_1,a_4)\ \&\ V(a_3,a_1,a_4)\ \&\ V(a_3,a_4,a_5)\ \&\ L(a_3,a_1,a_5)\ \&\ V(a_4,a_2,a_3)\ \&\ V(a_4,a_3,a_6)\ \&\ L(a_4,a_2,a_6)\ \&\ V(a_5,a_3,a_6)\ \&\ V(a_6,a_4,a_5)$.

An estimate of complexity for this example is

$$O\left(10^4\right) = O(10000).$$

To prove the belonging of a part of the image to the class of "number four" it is needed to prove deducibility of the formula

$$S(a_1,\ldots,a_6) \Rightarrow \exists(x_1,\ldots,x_5)_{\neq} A(x_1,\ldots,x_5).$$

After reducing to the form (1) we have

$$\forall a_1,\ldots,a_k \exists x_1,\ldots,x_n \left(\begin{array}{l} \left(\begin{array}{c} V(x_3,x_1,x_4) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \\ \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \\ \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \end{array} \right) \& \\ \left(\begin{array}{c} V(x_4,x_2,x_3) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \\ \neg V(a_3,a_4,a_5) \vee \neg L(a_3,a_1,a_5) \vee \neg V(a_4,a_2,a_3) \vee \neg V(a_4,a_3,a_6) \vee \\ \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \end{array} \right) \& \\ \left(\begin{array}{c} V(x_4,x_3,x_5) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \\ \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \\ \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \end{array} \right) \& \\ \left(\begin{array}{c} L(x_4,x_2,x_5) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \\ \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \\ \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \end{array} \right) \end{array} \right)$$

The next F-set we have the result of point 1. It was received in 4 steps.

$$\left(\begin{array}{l} \left(\begin{array}{c} V(x_3,x_1,x_4) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg L(a_3,a_1,a_5) \end{array} \right) \\ \left(\begin{array}{c} V(x_4,x_2,x_3) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg L(a_3,a_1,a_5) \end{array} \right) \\ \left(\begin{array}{c} V(x_4,x_3,x_5) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg L(a_3,a_1,a_5) \end{array} \right) \\ \left(\begin{array}{c} L(x_4,x_2,x_5) \vee \neg V(a_1,a_2,a_3) \vee \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg L(a_3,a_1,a_5) \end{array} \right) \end{array} \right)$$

Point 2.1 is not applicable. According to the point 2.2 take the formula $V(x_3,x_1,x_4)$, according to the point 2.3 take the formula $\neg V(a_1,a_2,a_3)$ (+1 step) and mark it as a deleted one (+1 step). According to the point 2.4 solve the system of equations

$$\begin{array}{l} x_3 = a_1 \\ x_1 = a_2 \\ x_4 = a_3 \end{array} \text{ (+3 steps).}$$

According to the point 2.5 F-set has the form

$$\left(\begin{array}{c}\left(V(a_1,a_2,a_3)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\neg V(a_2,a_1,a_4)\vee\neg V(a_3,a_1,a_4)\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\ \left(V(a_3,x_2,a_1)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\neg V(a_2,a_1,a_4)\vee\neg V(a_3,a_1,a_4)\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\ \left(V(a_3,a_1,x_5)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\neg V(a_2,a_1,a_4)\vee\neg V(a_3,a_1,a_4)\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\ \left(L(a_3,x_2,x_5)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\neg V(a_2,a_1,a_4)\vee\neg V(a_3,a_1,a_4)\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\end{array}\right)$$

(+12 steps)

According to the point 2.9 (+4 steps) return to the points 2.2 and 2.3. The formula $V(a_4,x_2,a_2)$ does not have a pair for unification (+22 steps), that is why go to 3 (+9 steps). F-set differs from the initial one only by marking the formula $\neg V(a_2,a_1,a_4)$ as a deleted one. According to the point 2.2 take the formula $V(x_3,x_1,x_4)$, according to the point 2.3 take the formula $\neg V(a_3,a_1,a_4)$ (+2 steps) and mark it as a deleted one (+1 step).

According to the point 2.4 solve the system of equations

$$\begin{array}{l}x_3=a_3\\ x_1=a_1\\ x_4=a_4\end{array}$$ (+3 steps).

According to the point 2.5 F-set has the form

$$\left(\begin{array}{c}\left(V(a_3,a_1,a_4)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\boxed{\boxed{\neg V(a_2,a_1,a_4)}}\vee\boxed{\boxed{\neg V(a_3,a_1,a_4)}}\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\ \left(V(a_4,x_2,a_3)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\boxed{\boxed{\neg V(a_2,a_1,a_4)}}\vee\boxed{\boxed{\neg V(a_3,a_1,a_4)}}\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\ \left(V(a_4,a_3,x_5)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\boxed{\boxed{\neg V(a_2,a_1,a_4)}}\vee\boxed{\boxed{\neg V(a_3,a_1,a_4)}}\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\ \left(L(a_4,x_2,x_5)\vee\boxed{\boxed{\neg V(a_1,a_2,a_3)}}\vee\boxed{\boxed{\neg V(a_2,a_1,a_4)}}\vee\boxed{\boxed{\neg V(a_3,a_1,a_4)}}\vee\neg V(a_3,a_4,a_5)\vee\neg V(a_4,a_2,a_3)\vee\right.\\ \left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\end{array}\right)$$

(+12 steps)

According to the point 2.9 (+4 steps) return to the points 2.2. Take the formula $V(a_4,x_2,a_3)$ and according to the point 2.3 take the negation $V(a_4,x_2,a_3)$ (+9 steps) and mark it as a deleted one (+1 step). According to the point 2.4 solve the system of equations

$$x_2=a_2$$ (+1 step)

According to the point 2.5 F-set has the form

$$\left(\begin{array}{l}
\left(V(a_3,a_1,a_4)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\
\left(V(a_4,a_2,a_3)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\
\left(V(a_4,a_3,x_5)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\
\left(L(a_4,a_2,x_5)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\neg V(a_4,a_3,a_6)\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)
\end{array}\right)$$

(+4 steps)

According to the point 2.9 (+4 steps) return to the points 2.2. Take the formula $V(a_4,a_3,x_5)$ and according to the point 2.3 take the formula $\neg V(a_4,a_3,a_6)$ (+10 steps) and mark it as a deleted one (+1 step). According to the point 2.4 solve the system of equations

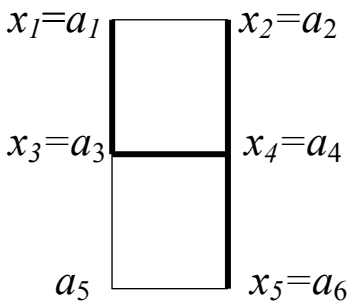
$$x_5 = a_6 \text{ (+1 step).}$$

According to the point 2.5 F-set has the form

$$\left(\begin{array}{l}
\left(V(a_3,a_1,a_4)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\overline{\neg V(a_4,a_3,a_6)}\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\
\left(V(a_4,a_2,a_3)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\overline{\neg V(a_4,a_3,a_6)}\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\
\left(V(a_4,a_3,a_6)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\overline{\neg V(a_4,a_3,a_6)}\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)\\
\left(L(a_4,a_2,a_6)\vee\overline{\neg V(a_1,a_2,a_3)}\vee\overline{\neg V(a_2,a_1,a_4)}\vee\overline{\neg V(a_3,a_1,a_4)}\vee\neg V(a_3,a_4,a_5)\vee\overline{\neg V(a_4,a_2,a_3)}\vee\right.\\
\left.\overline{\neg V(a_4,a_3,a_6)}\vee\neg V(a_5,a_3,a_6)\vee\neg V(a_6,a_4,a_5)\vee\neg L(a_4,a_2,a_6)\vee\neg L(a_3,a_1,a_5)\right)
\end{array}\right)$$

(+ 2 steps)

The empty F-set is received. According to point 2.7 the algorithm halts (+4 steps). An object “four” is extracted. The extracting needed 166 steps.



This example shows us the necessity of ordering formulas within F-set. If we order formulas before the algorithm run (as it was mentioned at the begin of the paper) then the initial formula would have the form

$$\forall a_1,\ldots,a_k \exists x_1,\ldots,x_n \left(\begin{array}{l} \left(\begin{array}{c} L(x_4,x_2,x_5) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg V(a_1,a_2,a_3) \vee \\ \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \end{array} \right) \& \\ \left(\begin{array}{c} V(x_3,x_1,x_4) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg V(a_1,a_2,a_3) \vee \\ \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \end{array} \right) \& \\ \left(\begin{array}{c} V(x_4,x_2,x_3) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg V(a_1,a_2,a_3) \vee \\ \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \end{array} \right) \& \\ \left(\begin{array}{c} V(x_4,x_3,x_5) \vee \neg L(a_3,a_1,a_5) \vee \neg L(a_4,a_2,a_6) \vee \neg V(a_1,a_2,a_3) \vee \\ \neg V(a_2,a_1,a_4) \vee \neg V(a_3,a_1,a_4) \vee \neg V(a_3,a_4,a_5) \vee \neg V(a_4,a_2,a_3) \vee \\ \neg V(a_4,a_3,a_6) \vee \neg V(a_5,a_3,a_6) \vee \neg V(a_6,a_4,a_5) \end{array} \right) \end{array} \right).$$

In such a case the problem would be solved after one execution of all algorithm steps without cancellation ones. We have not ordered the F-set for more clearness.

Ant tactics

An ant algorithm [Dorigo, Birattari, Stutzle, 2006] is sufficiently new approach to the solving of Artificial Intelligence problems. It reflexes actions of insects. Behavior of ants is the basis of a series of methods the most successful of which is optimization of ant colony.

Optimization of ant colony simulates actions of some kinds of ants. They put pheromone in the ground in order to mark successful ways for other colony members moving. More ants use the same way more the pheromone concentration on this way. More the pheromone concentration is on the way more preferable this way is in comparison with the other ones. In such a way an “ant logic” allows to choose a shorter way between two points.

Ant algorithms are iteration ones. Every iteration takes into account actions of artificial ants. Every ant constructs its decision during making some actions and does not repeat the same action. For every step an ant chooses the next action in dependence of pheromone quantity used for marking the actions before their fulfilling.

Ant algorithms successfully solve some NP-hard problems, for example, the problem of “traveling salesman”.

Idea of an algorithm using the ant tactics

Artificial ant is a program agent using to solve some problem and being a member of a big colony of artificial ants. Any ant has a set of simple rules which allow him to choose an action. It has a list of taboos, i.e. a list of actions which it has already done or cannot do at all.

A real ant put some pheromone while moving along the way. Artificial ant increases a mark of already fulfilled action.

The population of artificial ants distributes the actions between themselves in equal parts. Such a distribution in equal parts is necessary because every action may be the first.

In the problem under consideration it is needed to prove or to refute deducibility of a predicate formula. The most difficult stage of such a proving is to find a list of distinct values for variables $\overline{x}$ the existence of which is claimed by the formula. So the ant tactic will be used by means of decreasing the action mark in the following way.

Initially the mark of formulas $P(t_1,...,t_m)$ and $\neg P(a_1,...,a_m)$ unification equals 1. If the unification is possible and does not lead to a deadlock F-set then the mark their unification increases by 1. If a deadlock F-set is received then the mark becomes 0. Otherwise it decreases by 1.

Every artificial ant begins its actions with its own disjunct using rules of IMA algorithm. While assigning variables ants connect each other and compare results. Comparison of different ants results consists in the checking of non-contradictoriness of these results. The results of ant actions are contradictory if they assign the same variables with different values or the same value is assigned to different variables. If the result of two ants actions does not contradict each other then the assignment is fulfilled in the disjunctions of the both ants.

If an ant cannot assign any variable then this ant does not make any action further.

If deducibility of the formula is proved or all marks are 0, then the algorithm halts.

Bibliography

[Dorigo, Birattari, Stutzle, 2006] M. Dorigo, M. Birattari, T. Stutzle. Ant Colony Optimization. Artificial Ants as a Computational Intelligence Technique. In: IRIDIA – Technical Report Series, 2006, № 23, pp 1-2.

[Kosovskaya, 2011] T.Kosovskaya. Discrete Artificial Intelligence Problems and Number of Steps of their Solution. In: International Journal on Information Theories and Applications, Vol. 18, Number 1, 2011. P. 93 – 99.

[Kosovskaya, Petukhova, 2012] T.Kosovskaya, N.Petukhova. The Inverse Method for Solving Artificial Intelligence Problems in the Frameworks of Logic-Objective Approach and Bounds of its Number of Steps. In: International Journal «Information Models and Analyses», Vol. 1. 2012. P. 84-93

[Orevkov, 2003] Orevkov V.P. The inverse method of logical derivation In: Adamenko A. Kuchukov A. Programming Logic and Visual Prolog. St. Petersburg, "`BHV-Petersburg"'. (2003), pp.. 952-965. (in Russian)

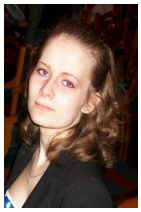
Acknowledgement

The paper is published with financial support of the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Authors' Information

Tatiana Kosovskaya *– Dr., Senior researcher, St.Petersburg Institute of Informatics and Automation of Russian Academy of Science, 14 line, 39, St.Petersburg, 199178, Russia; Professor of St.Petersburg State Marine Technical University, Lotsmanskaya ul., 3, St.Petersburg, 190008, Russia; Professor of St.Petersburg State University, University av., 28, Stary Petergof, St.Petersburg, 198504, Russia, e-mail: kosov@NK1022.spb.edu*

Major Fields of Scientific Research: Logical approach to artificial intelligence problems, theory of complexity of algorithms.



Nina Petukhova *– PHD student, St.Petersburg State Marine Technical University, Lotsmanskaya ul., 3, St.Petersburg, 190008, Russia, e-mail: ninka_mat@mail.ru*

Major Fields of Scientific Research: Logical approach to Artificial Intelligence problems.

APPLICATION OF SOME CYBERNETIC MODELS IN BUILDING INDIVIDUAL EDUCATIONAL TRAJECTORY

Borislav Lazarov

Abstract. *The individual educational trajectory (IET) is a medium-term didactic complex that provides optimal opportunities for developing the creative potential of the learner by taking into account his/her personality. It includes: 1) developing an individual informational environment; 2) tuning the didactical resources; 3) personalization of the interim goals; 4) planning personal learning and research activities; 5) considering the self-organization. In the paper we submit a model of IET called DMT. In contrast to the cognition trajectory developed by Kolyagin and Ganchev the DMT has no linear structure. The IET in our model consists of sections due to the interim goals and each section of it could be perceived as a beam whose spectrum consists of the listed above 5 components. So the architecture of DMT could be associated with the one of the skyscraper Taipei 101 (the abbreviation DMT stands for Didactic Model Taipei). The educational process in any section in the IET could be conducted in different manner that allows different didactic approaches to be applied with respect to the individualities of the learner. The particular didactic approaches we had applied in a case study held in 2011/2012 academic year are strongly influenced by some cybernetic ones in the Consequence Driven Systems (CDS) theory.*

Keywords: *didactical models, individual educational trajectory, elliptic arbelos, Socratic style teaching.*

ACM Classification Keywords: *D.4.8 Performance (Modeling and prediction), I.6.5 Model Development (Modeling methodologies), J.1 ADMINISTRATIVE DATA PROCESSING (Education)*

0. Introduction

The cybernetic models in education are in use as long as the cybernetic idea appears. Among the earliest ones is the model of Hodge [Hodge, 1970] which architecture is shown at figure 1. The model represents in a very simple and clear way the general stream of the educational process from the organizational perspective. The mystery of learning remains out of sight and the process of learning is affected indirectly.

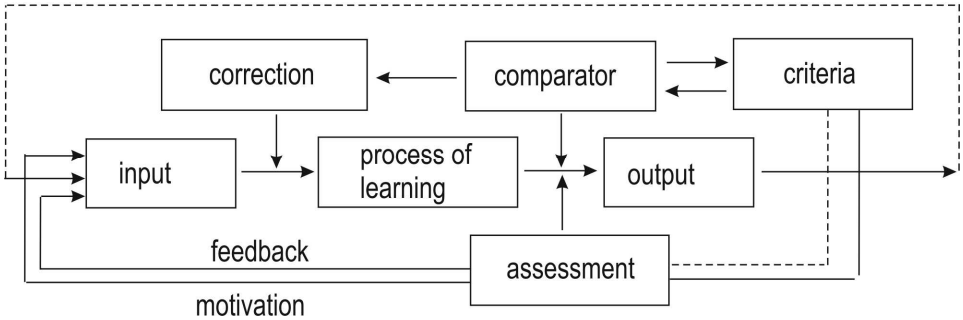


Figure 1. The architecture of the model proposed by Hodge

Quite more sophisticated cybernetic model is the one submitted by Garrison and Magoon [Garrison and Magoon, 1972]. It consists of five blocks each of them representing the mechanism of one particular cognitive process. The interaction of the blocks gives the big picture of the learning process. However, the model is too complicated to

be put in use in a regular pedagogical practice. Below we are going to present a model which complexity is somewhere in-between the ones of the two just pointed marginal models. Key role in it play a system of didactic schema that are analogs of cybernetic ones. Our interest in cybernetic models was provoked after attending a lecture of N. Ackovska about taxonomy of the learning agent [Ackovska, 2010].

1. A didactical heritage from modern perspective

The individual approach in education is considered mainly as contrapuntal to the didactical technology. However, some general technological rules could be put in the fundament of any individual teaching-learning process and the first systematic attempt in this direction (as far as we know) is described more than two millennia ago in the Plato's *Dialogues*. In the dialogue called *Meno* [Plato, 4th century BC] we can observe a didactic approach that becomes classics in teaching – the Socratic Method (Fig. 2).

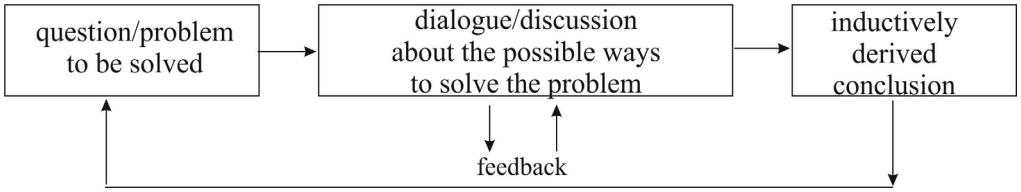


Figure 2. The architecture of the Socratic structure as given by Andreev [Andreev, 1966]

It is emblematic that Socrates illustrated his method with a mathematical example. In modern times this method is developed by various prominent mathematicians and educators like Freudenthal (explicitly [Freudenthal, 1982]) or Polya (implicitly [Polya, 1961]). The genuine Socratic Method of stating questions to rethink the initial problem is too limited for the modern didactics. However, we adopt the style of inductive questioning that leads the student to build his/her knowledge through small steps, to increase understanding through inquiry. Further we refer to such type of teaching as the *Socratic style*. Nowadays close to this style in Europe is so called *inquiry based science education* (for mathematics *problem-based approach*) [Rocard et al, 2006]. The inquiry based education was declared to be potentially effective as general classroom practice. Despite the enormous resources that European educational structures spend to implement the ideas of inquiry based education we are not known convincing evidences for positive breakthrough in mass mathematics education. On the contrary, our belief is that the Socratic style is more effective when it is applied to advanced students combined with the individual approach. The first experimental work that we have done witness such thesis [Lazarov, 2012].

2. The cognition trajectory in teaching mathematics

The concept of *cognition trajectory* is elaborated by Ganchev [Ganchev, 1996] to describe the process of education as a manageable object. The cognition trajectory is think to be an ideal educational process illustrated with a curve that connect the initial cognition status of a learner T_0 with the educational goal G for a (mead term) period. The learner is supposed to 'move along' the cognition trajectory learning the material in full scale (the desired case); if not then 'the learner declines the cognition trajectory' (the case that is much closer to the reality). Two examples of such discrepancy between the desired educational process and the real situation illustrated with cognition trajectory are shown on figure 3.

1) The teacher organizes education supposing the student's knowledge is in a neighborhood of T_1 but the student's knowledge is still in a neighborhood of T_0 (the left picture).

2) The teacher organizes education supposing the student's knowledge has been reached the status T_1 but the student's knowledge went into another direction (in a neighborhood of T') from the beginning T_0 (the right picture).

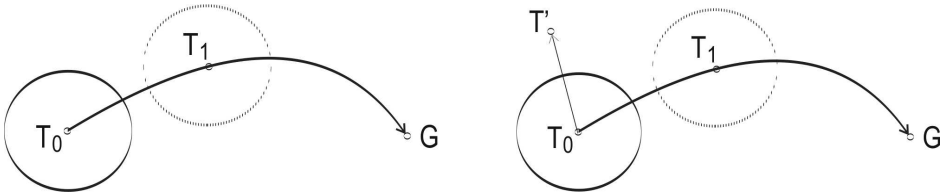


Figure 3. Two cases of discrepancy between desired learning process and reality as given by Ganchev [ibid]

Teacher's intervention is relevant and effective only in the case when student's knowledge is coherent with the cognition trajectory. Ganchev calls such education *properly directed* and lists a system of six main steps which should be performed during the math lessons to ensure proper direction of teaching-learning process. The general idea of the system refers to the Vygotsky's findings about *zones of actual and proximal development* [Vygotsky, 1978] and follows the rule: any educational activity should not leave the student's *zone of proximal development*.

3. The individual educational trajectory in teaching mathematics

We agree in general with Ganchev's conclusions about the building math lessons as fundamental classroom experience [ibid]. However, the development of the global educational environment allows organizing math education much closer to the individual specifics of the learner. A convenient interpretation of the cognition trajectory for describing the individual approach is the *individual educational trajectory*. By ***individual educational trajectory*** (IET) we will understand

organizational frame and plan for realization of a medium term educational process that is coherent with the individual specifics of the learner and provides opportunities for the optimal development of his/hers creative potential [Lazarov, 2012].

The concept of IET refers to the educational microcosmos of the student which is immersed into the global educational environment. The design and implementation of the IET is a complex process that includes the following components.

1. Formation of an individual informational environment.
2. Individualization of the didactical resources, including selection of the individual (re)searching instruments.
3. Individualization of setting the educational goal, including flexible approach to achieve it.
4. Individualization of the learning temps, investigation activities, layout style.
5. Taking into account the individual reflexive abilities and self-organization aptitude in searching a synergetic effect. [ibid]

The process of building IET is a step-by-step (iterative) procedure and any of the listed components is supposed to be actualized according to the student's interim achievements. The developing of the entire process goes in two directions, let call them vertical and horizontal. The description of the horizontal movement will be given in the next section of the article.

The vertical movement in the k-th iteration is formed of the following steps.

…(k – 1) →

- A near educational goal (of learning, investigative or research type) is mapped out with respect to the actual knowledge, skills and competences (KSC) of the learner (actual development).
- A (very limited) informational resource is determined which is focused on the goal.
- All needed activities to extend the actual KSC to a level required for reaching the stated goal are performed.
- Student proceeds to the goal in a *specified manner*.
- The achievements are analyzed.

→ (k+1) …

The educational goal in the (k+1)-th iteration should require learner's KSC among the elaborated ones in the k-th iteration. So the starting KSC of the (k+1)-th iteration are among a subset of the achieved KSC in the k-th one. The transition

$$(k-1) \rightarrow (k) \rightarrow (k+1)$$

supposes actualization of the components (1)-(5). This vertical movement along IET could be illustrated as shown in the figure 4 – the architecture of our model in building IET reminds of the architecture of the skyscraper Taipei 101. Further we will call this model DMT, which is an abbreviation of *didactical model Taipei*.

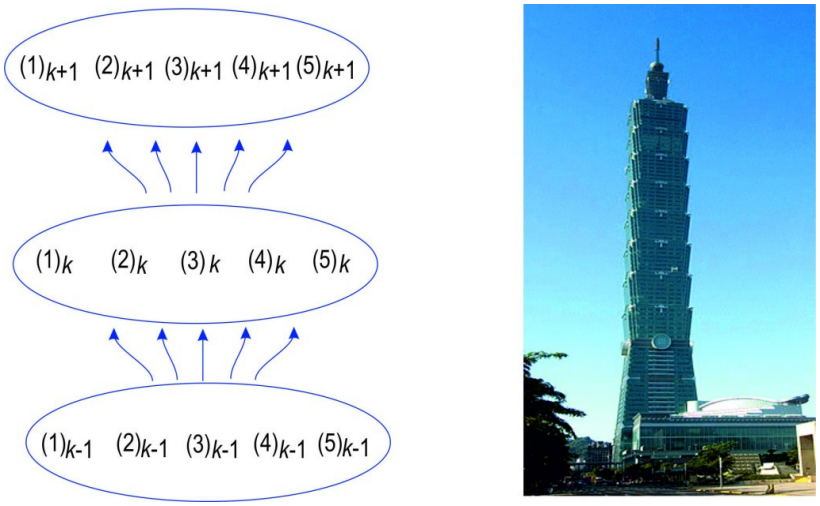


Figure 4. Iterative structure of building IET and the skyscraper Taipei 101 (the photo is taken from http://en.wikipedia.org/wiki/File:Taipei101.portrait.altonthompson.jpg).

As one can see DMT provides a frame to turn education in a manageable process of flexible type. The flexibility of the architecture in vertical perspective is guaranteed by the revision of the components (1) – (5) on any iteration. Additional flexibility is added by the horizontal structure of the DMT.

4. Basic concepts related to the horizontal movement

The process of building IET in any iteration could be described with some auxiliary models. It refers to the horizontal movement in DMT structure. Further we are going to clarify the 'specified manner' pointed in the fourth bullet from the description of the k-th iteration in the above section. This manner can vary significantly due to the development of the student's KSC. The auxiliary didactical models we apply are under the influence of the cybernetic ones from the Consequence Driven Systems (CDS) theory [Ackovska, 2010]. The basic concepts we are going to use in any of our auxiliary models refer to some analogues in CDS.

First let us say that any particular student acts in a ***local behavioral environment*** (LBE) which is a complex socio-economical and cultural structure. For our purposes we consider LBE including:

- people related to the student's behavior (teachers, parents, classmates etc.);
- institutions that organize education and creative work (school, clubs etc.);
- events that provides opportunities to manifest the achievements (tournaments, conferences etc.);
- system of values that form the cultural context of the student (motivation factors, anticipation about the future professional realization etc.).

DMT restricts the local behavioral environment to the listed components as far as they affect most directly the student's educational behavior but we are clear about the simplification of the reality. For instance our model neglects the emotional status of the student that sometimes is crucial in taking decision. The change of LBE depends on the student development. E.g. if (s)he succeeds with a project (s)he can attend some conferences to present it where (s)he can: meet new people, join new clubs, see new opportunities for future professional realization etc.

The ***learning and creativity interface*** (LCI) is the next composite concept that appears in the auxiliary models. We consider LCI as the triad (EC,SS,ER) where EC stands for the ***educational context***, SS is the ***Socratic style interaction*** between the teacher and the student, ER are the ***educational resources***. The EC is the refraction of the LBE through the educational goal, a set of signals that are sent from LBE ingredients and affect the student's educational activeness related with the stated goal. SS is that component that describes the personal site of the teaching-learning process, communications between the student and the teacher related to the educational goal. ER include the didactical and technical tools, sources of information etc. that are implemented during the study and research activities. All three ingredients of the triad LCI ensures the real time development during the movement along an IET. They interact and the momentary magnitude of any part depends on the momentary status of the educational process.

The last composite concept that appears in our general scheme is the triad KSC of ***knowledge, skills and competences***. Here we skip the details and refer to [Winterton et al., 2006].

The architecture of the horizontal movement is shown in the figure 5. In general this architecture represents the most common case. However, some particular applications deserve to be considered separately.

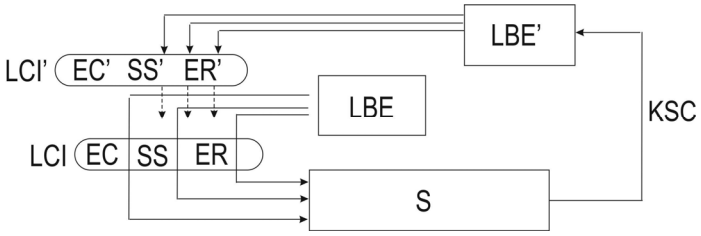


Figure 5. Architecture of the horizontal movement in IET.

5. Some special cases of the general auxiliary model

The general auxiliary model of the horizontal movement allows modifications in any particular stage in building IET. The taxonomy we present below looks like a didactical replica of the CDS one given by Ackovska [ibid]. However, an important difference between the cybernetics and didactics is the nature of the learning agent: our model considers the student as a creative person who is responsible for his/her education. Another difference is the matter of the interaction between the key players: the Socratic style in our model significantly differs from the instructional style in CDS.

Context-independent model. This is an auxiliary model for education with restricted dependence on EC. It is convenient for the starting levels of the IET when the educational goal is in a new area for the student. Student is supposed to generate initial knowledge and to elaborate basic skills in a field that allows stating the educational goal. In such a situation the impact of the LBE onto educational goal is indirect and very limited, thus it can be neglected. The main LCI components are ER and SS. At the output the KSC are also without considerable third component – competences. The Socratic style could be realized mainly as giving advices and encouragement. The educational resources should be user-friendly: the reading materials should be easy; the software should be relevant to the educational status of the student.

EC considering Models. These are auxiliary models that are suitable in various stages of building IET. We are going to introduce them in order of simplicity.

At the final stage of a particular IET the student-teacher interaction could be diminished according to the formed student's competence of synthetic type. The student has got high level skills and is able to create new knowledge. The educational context stimulates him/her to achieve some results relevant to a LBI that includes prestigious events as contests or conferences. Student's reflexive abilities and self-organization aptitude allows him/her to state educational goals by himself/herself. The Socratic style takes the form of equipollent communication. So we can speak for **LCI with restricted teacher's assistance.**

The interim stages of IET are described with auxiliary models in which the components of the LCI interact in full scale. A common feature of all such models is that SS provides a kind of reinforcement. More often than not the reinforcement is **context-dependent**. The teacher has closer look at the student's efforts and organizes an inquiry that allows the student to achieve the stated KSC. The ER should be relevant to the EC. Let us give some examples.

- Tuning the didactical tools: if a problem is too hard at the moment the teacher can decompose it; if a concept is not clear in general case, some particular cases to be proposed for consideration; if a construction is not properly designed, a deductive analyze to be set for discussion etc.
- Fitting the informational sources: if a book is too hard for understanding, it could be substituted with some easier articles; if the student refers to a web-site which is not reliable, a web-search for other sources to be organized etc.

- Conducting the progress in computer skills: organizing a step-by-step procedure for studying the computer program with properly graded examples.

The social microcosmos is of great importance. EC is directly affected from the other students' treatment of the achievements of our student's. Success or failure in some activities could change the student's self-confidence and could influence the educational goals. This leads to some changes in the didactical resources which the teacher applies. Thus we can speak about **consequence dependent model**.

The accommodation of the teacher's interventions along the IET leads to another models. Teacher can give advices or reinforcement immediately but only if it is necessary. In this case we speak about **tutorial teaching model**. The teacher can launch an intervention after student makes several trials to achieve a particular result. This is **delayed intervention model**.

6. How the theory works

In this section we are going to give an idea about how the above models were applied in building a particular IET. We present briefly some parts of a case study carried out in 2011/2012 scholastic year with a 12th grade student. This student (further we call her PL) was involved in a middle term activity during 2010/2011 scholastic year when she performed rather successfully in generalizing an idea from the math tournament Chernorizetz Hrabar. PL was interested also in Wasan geometry and this gave us reason to state the following educational goal: *to develop further student's synthetic competence via studying some sangaku constructions for dynamic stability*. The plan was some dynamic-geometry applets to be designed that illustrate geometrical properties pointed in sagaku problems which are invariant when the parameters of the construction are changed; the results to be proved by methods from the Euclidean and/or analytical geometry; the calculations to be performed using computer algebra system if necessary; eventually a report to be prepared and presented at some conferences.

The first iteration in building this particular IET. Let us explain how the ground floor of the DMT was furnished.

1) The informational environment was restricted to a set of issues of the Bulgarian magazine Education of Mathematics and Informatics where Jordan Tabov edited the column *Problems of the issue* dedicated to sangaku problems.
2) Socratic style was applied every time some questions arise; a considerable amount of short term activities were organized to fill up some gaps in geometry mainly connected with loci.
3) The interim educational goal was: the interface of GeoGebra to be studied, some simple constructions to be performed and examined about dynamic stability; some appropriate pictures of sangaku tablets to be captured from the WWW (no upgrade of some competences was planned). The research process was focused on extracting some common constructions from the sangaku problems among the *problems of the issue*.
4) The temps of learning and elaborating skills were intensive due to the deadlines for submitting reports for an annual student conference. As a matter of fact the in being synthetic competence of the student allows such intensity: she has already passed a similar training process previous year.
5) We also refer to the student's self-control due to the reflexive abilities and self-organization aptitude shown in the previous period.

PL activities during this period were not considerably affected by the education in school or relations with classmates; the student's motivation was not connected with some tests or examines; the eventual participation

in conferences was far enough to have direct impact at this stage. Thus the influence of the LBE was very limited and we can speak that this part of the IET was context independent.

The second iteration. The just described preparatory work drifted a lot of collateral information. Despite the recommended literature was limited, the number of sources used by the student raised, the math methods that appears in the sources were also too many to be studied deeper. So the first steps in the IET on this (second) stage of the DMT were to clean up the collateral information and methods and to focus on the potentially suitable ones. Let us give an example.

A large number of the sangaku problems deal with circles and this is why naturally appear constructions that remain the arbelos. Such constructions are well studied by Fukagava, Okumura, Watanabe and other authors (a comprehensive list of books and articles is given in [Watanabe, 2011]). But in some sangaku tablets appear also constructions with circles and ellipse. The figure compound by an ellipse and two internally touching it circles (called by us ***elliptic arbelos***) is not studied separately (as far as we know). Moreover, some interesting properties of such figures were observed. Since the usual method to simplify an arbelos-likely construction is inversion, PL was recommended to study the elliptic arbelos applying inversion. But the image of the inverted ellipse occurs analytical curve of 4^{th} grade – the method did not work. So we decide to skip elaborating skills in inversion and to focus on analytical methods. Using analytical methods instead of the classical geometry proofs is usually regarded as prejudice of deduction. But the deductive side was not neglected: the proofs of the basic properties of the elliptic arbelos were performed by PL in traditional Euclidean style.

Now let us highlight the role of the context with the next example. PL had the opportunity to contact directly Prof. Jordan Tabov who is the editor of the column *Problem of the issue*. He gave her genuine Japanese books dedicated to Wasan geometry and encouraged her for further study of the topic. The additional informational resources were important. But more important was the external positive opinion from an international expert – this was another powerful stimulus for PL, who recognized the importance of her activity.

The advices and reinforcement in this stage were given in Socratic style applying the tutorial auxiliary model. The interim goals were covered successfully and the results obtained were enough to be reported at two consecutive math conferences for school students.

The third iteration. The opportunity to present her work at conference changed significantly the EC. Further activities planned were connected with the layout of the content. But during polishing details a lot of questions appeared and answering these questions leaded to a considerable upgrade of PL's synthetic competence. Let us give an example. The proof of one of the main theorems about the elliptic arbelos required an inequality to be verified. After several failures in solving the inequality by hand PL turned to graphical methods and managed to give strong reasons about the solutions of the inequality. (Later an analytical solution was given to her by Prof. Nikolay Nikolov in a private communication). We decided to pay less attention to the technical skills in solving equations, inequalities etc. Instead of this PL turned her efforts to interpret the results obtained about the equations with a computer algebra system. On this stage in building this IET we applied delayed intervention model.

The next iterations in the DMT refer to the application of the consequence dependent model. Performing successfully at several conferences and contests PL became more confident and even when she did not solve some of the stated problems she proposed conclusive graphical or numerical arguments. Her LBI included some students who share the same interests and views. The next moment is indicative: PL needed a kind of animation to demonstrate the change of the radius of the incircle when the elliptic arbelos changes its type from intersecting to tangential and then to non-intersecting. She did not know how to make this animation. She wrote a question in the GeoGebra Forum and received several suggestions. PL was fascinated by the helpfulness of the international math community. When she asked the persons who helped her about their names to write an official

acknowledgement she was surprised with their modesty – no one considered such cooperation as something extraordinary that need acknowledgement.

After several months of studying and working under supervision by the author of this paper PL became enough independent and she got her own view on the theory of the elliptic arbelos. Our IET completed. We think that the educational goal was achieved in general. PL was competent to study a mathematical object using different approaches including advanced analytical, graphical and numerical methods, to prepare a report and presentation, to present the results to competent auditory, to discuss different sides of the findings.

7. Final remarks

The taxonomy of the learning paradigms proposed by Ackovska [ibid] sketches 10 cybernetic models each one approved in some technological processes, i.e. any model works in a real life situation. The educational process of a human is more specific and needs special cares that take into account the personality of the learner. Our experience shows that any attempt to apply strict regulations in teaching-learning process diminishes the effectiveness of the education. On the contrary, a teaching-learning of flexible type gives better outcomes in general but needs quite larger arsenal of didactical instruments than any hard didactical technology. The IET we had build used a variety of auxiliary didactical models that allow us to react adequately in any particular stage. The Ganchev's ideas (pointed in the second section) are more technological and serve the in-class math education directed to covering some educational standards. In contrast to this the IET is directed to a broader field – building a competence of synthetic type (*synthetic competence*) in which the math knowledge (math key competence) is just a component. Via IET we lose in size of the target group but we gain the deepness of the knowledge, skills and competences build.

The complete design of IET is not possible to be made in advance. The collateral information should be cleaned up in any step, i.e. to be neglected when stating the next educational goal. The Socratic style of teacher-student interaction should be coherent with the local behavior environment and the adequate auxiliary didactical analogs of the cybernetic models help to put in practice a desired individual educational trajectory.

Acknowledgement

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Ackovska, 2010] N. Ackovska. *Taxonomy of Learning Agents.* Didactical Modeling Vol. 4 , 2010/2011. http://www.math.bas.bg/omi/DidMod/Articles/Volume04 (active in December 2012).

[Andreev, 1966] Андреев, М. *Процесът на обучението*. Университетско издателство *Св. Климент Охридски*. София, 1996, с. 49.

[Freudenthal, 1982] Фройденталь, Г. *Математика как педагогическая задача*. Часть I, Просвещение. Москва, 1982, с. 75-81.

[Ganchev, 1996] Ганчев, И., Кучинов Организация и методика на урока по математика. Модул, София, 1996, с. 14-19.

[Garrison and Magoon, 1972] Garrison, K. and Magoon, R. *Educational Psychology; an Integration of Psychology and Educational Practices*. Merrill Publishing Company, 1972, p. 290

[Hodge, 1970] П. *Модель, предлагаемая для анализа учебного процесса. Теория и практика обучения*. В Берг, А. (ред.-сост.) *Кибернетика и проблемы обучения*. Прогрес, Москва, 1970. С 266-288.

[Lazarov, 2012] Лазаров, Б. Индивидуална образователна треактория – изследване на частен случай. Математика и информатика. Бр. 3, 2012. С. 238-248.

[Plato, 4th century BC] Платон. *Диалоги*. Мысль, Москва,1986, с. 368-397.

[Polya, 1961] Пойа, Д. *Как решать задачу*.Учпедгиз, Москва, 1961, с.29-39.

[Rocard et al, 2006] Rocard, M. et al. *Science Education Now: A Renewed pedagogy for the Future of Europe.* http://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf (active in Dec 2012) p. 9.

[Vygotsky, 1978] Vygotsky, L. *Mind in Society: Development of Higher Psychological Processes*. Harward University Press, Cambridge, 1978, p. 86

[Watanabe, 2011] M. Watanabe. A Possibility of the Study on Contemporary Geometry through Arbelos and Wasan. PhD thesis, IMI-BAS, Sofia, 2011.

[Winterton et al., 2006] J. Winterton et al. Typology of knowledge, skills and competences: clarification of the concept and prototype. Cedefop Reference series; 64, Luxembourg: Office for Official Publications of the European Communities, 2006, pp 25-29.

Authors' Information

Borislav Lazarov *– Assoc. Professor, Institute of Mathematics and Informatics – Bulgarian Academy of Sciences*

e-mail: lazarov@math.bas.bg

Major Fields of Scientific Research: Didactical models, Gifted education

TABLE OF CONTENT

A Joint Global and Local Tone Mapping Algorithm for Displaying Wide Dynamic Range Images
Alain Horé, Chika A. Ofili, and Orly Yadid-Pecht, Fellow, IEEE 3
Component Modeling: on Connections of Detailed Petri Model and Component Model of Parallel Distributed System
Elena Lukyanova..... 15
Model for Astronomical Dating of the Chronicle of Hydatius
Jordan Tabov 23
Connectivity Control in AD Hoc Systems: a Graph Grammar Approach
Alexander Mikov, Alexander Borisov..... 37
Citation-Paper Rank Distributions and Associated Scientometric Indicators – a Survey
Vladimir Atanassov, Ekaterina Detcheva 46
Sub-Optimal Nonparametric Hypotheses Discriminating with Guaranteed Decision
Fedor Tsitovich, Ivan Tsitovich..... 62
Linguistic and Program Tools for Debugging and Testing of Simulation ModelS of Computer Networks
Elena Zamyatina, Alexander Mikov, Roman Mikheev..... 70
The Inverse Maslov Method and ANT Tactics for Exhaustive Search Decreasing
Tatiana Kosovskaya, Nina Petukhova..... 81
Application of Some Cybernetic Models in Building Individual Educational Trajectory
Borislav Lazarov..... 90
Table of content..... 100