DOI: 10.5604/01.3001.0010.8543

CONTROL MODEL OF DATA STREAM TRANSMITTED OVER A NETWORK BASED ON PROXYING TECHNOLOGY

Olesia Barkovska, Vitaliy Serdechnyi
Kharkiv National University of Radioelectronics, Electronic Computer Department

Abstract. *Network traffic control model is described in the work, including data mining modules transmitted in the network as well as further qualified data analysis module based on artificial intelligence methods and means, namely recursive and convolutional neural networks. The topicality of work is proved by the intensive scientific research conducted in the field of information security, big data intelligent data processing tools, and stipulated by growing necessity to limit access of children to aggressive information, which can impact on the child's psychoemotional state. Particular attention is paid in the article to proxy-server development for HTTP query receipt, search for a match in white and black lists and decision making as to data legality.*

Keywords: communication system traffic control, data transfer, access protocols, learning systems

MODEL KONTROLI DANYCH PRZEKAZANYCH PRZEZ SIEĆ W OPARCIU O TECHNOLOGIĘ PROXY

Streszczenie. *W artykule rozważono model kontroli ruchem sieciowym, który zawiera moduł do zbierania danych przesyłanych przez sieć, a także moduł do dalszej analizy przygotowanych danych w oparciu o metody i środki sztucznej inteligencji, a mianowicie rekurencyjne i splotowe sieci neuronowe. Znaczenie tej pracy jest potwierdzone przez dużą liczbę badań naukowych prowadzonych w dziedzinie bezpieczeństwa informacji, środków intelektualnych do przetwarzania dużej ilości danych, a także ze względu na rosnącą potrzebę ograniczenia dostępu dzieci do agresywnych informacji z sieci, które mogą wpływać na ich stan psychoemocjonalny. Artykuł koncentruje się głównie na opracowaniu serwera proxy do odbierania żądań HTTP, wyszukiwania dopasowań na „czarnych" i „białych" listach i podjęcia decyzji o zezwoleniu na obejrzenie.*

Słowa kluczowe: kontrola ruchu w systemie komunikacyjnym, transfer danych, protokoły dostępu, systemy uczenia się

Introduction

Control and analysis of data stream transmitted in Internet channels gains actuality every year. The World Wide Web is used for data mining, information exchange, rendering services, in particular entertainment and inconsistent with censorship. Access to the above-mentioned information is available for people of any age, gender and emotional condition. A specific instance is the information education function of electronic resources, access to which is often granted to children. When searching for valuable information, a child may encounter information, which can have a negative influence on the immature personality, e.g. bullying, aggressive images, explicit or graphic descriptions etc.

Let us call the kind of information, which causes agitation of the child's psyche, anxiety or aggression state, unwanted information (UI). Thus, limiting child's access to such kinds of information together with the creation of additional information filtration means in accordance with the set rules (filters) is considered highly relevant.

Methods of traffic classification into acceptable for viewing and inacceptable were studied by [6].

One of the methods to solve the set task is the use of the parent control function enabled by various extensions for Google Chrome, Mozilla Firefox, Opera browsers. A white list of sites authorized for lookup and a black list of certain sites unacceptable for child sessions are concurrently formed. Such popular browser extensions as Adult Blocker and TinyFilter may conduct morphological analysis of the page text content and filter pages on the basis of the downloaded content. A considerable shortcoming of the given method is their affinity to a definite browser and support of the most popular browsers only. For example, it is possible to avoid this information filtration method via the use of Internet Explorer installed on all OS Windows computers.

Limits to child's access to the content are possible via search engine tuning. Along with this, it is impossible to define the type of unwanted content – search engines usually delete only porn resources from the search results effectively.

Content filtration function can be enabled on the router. Almost all the leading manufacturers (ASUS, ZyXEL and TP-Link) provide for this capability [7]. The parent computer is identified by the router via MAC address. The functionality of filtration by the router is limited to a temporary filter (scheduled internet connection), the white and black lists of sites, incoming traffic analysis by means of verification with various unreliable resource databases.

Current control and data streams demarcation methods analysis enabled to make a conclusion that the existing methods have a number of substantial drawbacks such as the insufficient level of control and data streams demarcation methods integration, mostly passive nature of the audit processes and limited capabilities of intellectual text and image message analysis in the short run [7].

In order to eliminate the given drawback, development of a multifunctional model to control active data stream transmitted in the network as well as implementation of a proxy-server based on HTTP-requests and HTTP-responses intellectual analysis aimed at securing the child against threats from the Internet are proposed in the work.

The aim of the work is provision of inability of child's access to the information education electronic resources containing information that may cause child's psyche agitation, anxiety and aggression state, i.e. to unwanted information.

Main feature of the final software product is a primary scanning and subsequent deep analysis opened web-page content's based on existing methods of artificial intelligence. The advantage of this analysis is that the proposed approach, which is described by the control model of data stream transmitted over a network based on proxying technology [2], is able to restrict the child from viewing the site, analyzing not only the name, keywords or URL, but also the tonality of the content of the requested web page.

1. Description of the control model of data stream transmitted over a network based on proxying technology

Proceeding from the existent solutions, their advantages and disadvantages, a control model of data stream transmitted over a network based on the proxying technology was elaborated. In the given model, the proxy-server is an intermediary between the Internet and the clients, namely all Web browsers installed on a local computer. The model is presented in Fig. 1.

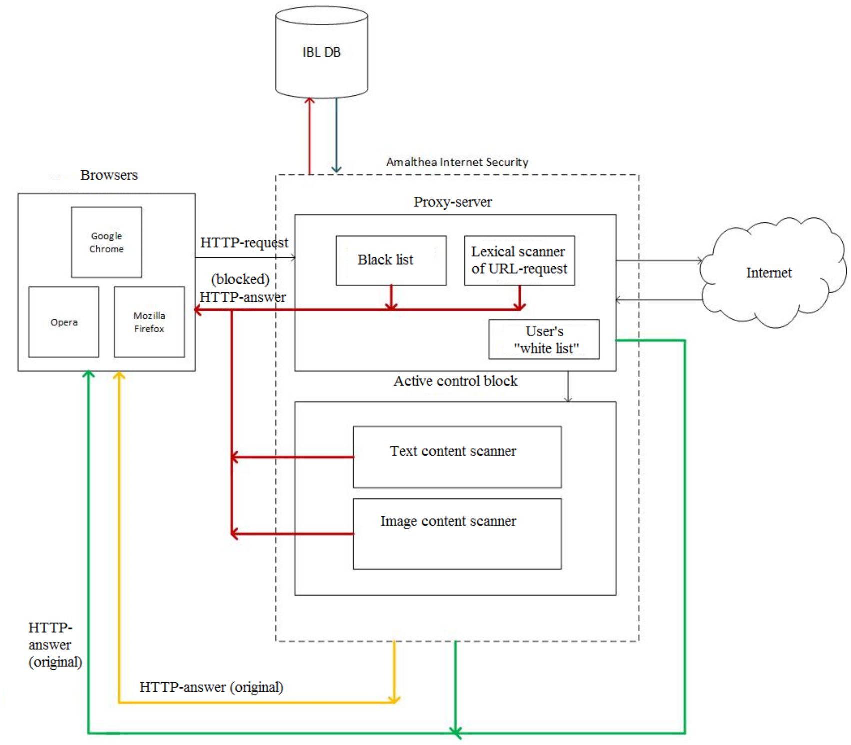


Fig. 1. Generic Control Model of Data Stream Transmitted over a Network Based on Proxying Technology

Below the description of modules in the proposed system is offered:

- the proxy server provides for the first stage of query analysis: all queries sent by any browser are received by the proxy server. Inside the server, the HTTP header is parsed and URI is distinguished, which is further compared with the white and black lists and studied by the lexical scanner. If the validation stage is successful, query analysis is accomplished and it goes to the addressee network. The response from the addressee server is also received by the proxy-server, which is compared with the white list. If a match is found, the HTTP response is sent directly to the browser avoiding the in-process control stage. Otherwise, the in-process web-page control stage begins, which presupposes intellectual content analysis;
- the black list is the list, which consists of two components: user exceptions and the exceptions generated by the system in the functioning process. If a match of the distinguished URI with the black list is found in the validation process, an HTTP-response is generated by the server to the browser with the message that the requested resource is banned. The URI further goes to the lexical query scanner;
- the lexical scanner checks on the basis of the requested URI if the page may contain forbidden materials. The procedure is based on the primitive regular expressions by means of comparison with the given letter combinations, which clearly indicate that the resource is unwanted. If this check is successful, the HTTP query is sent to the addressee;
- page content scanner conducts in-process content control via incremental intellectual analysis of the received HTTP response. Upon this, the content is divided into text content and graphic content. Text content analysis is brought to text sentiment analysis (if the text expresses a positive or a negative point of view about something) or text diversification according to topics. If the text analysis is successful, all the available images from the site received from the remote server are also analyzed for presence of the forbidden graphic information. In case of the positive analysis of the two above-mentioned content types, the response in sent to the browser and the URI of the analyzed resource is automatically added to the white list. If at least one of these stages results in the negative response, a message is formed for the browser to block the resource and the given URI is added to the black list. It should be noted that in the system being developed, a variant of partial editing of the opened Internet resource is also proposed. For example, if a web page is not a prohibited site and does not contain unwanted materials, and at the end of the page there is a comment block with obscene language, then this block will be "cut" from this page and it will be displayed in the browser without it. This approach makes it possible to safely view pages without completely blocking access to them. The criterion determining the decision to completely block or partially edit a web page can be the number of detected unwanted materials and their scatter across the page. If all of them are in one place, which is typical for comments with obscene language, then you can apply partial editing. If unwanted materials are scattered more evenly throughout the page, which is typical of a typical web resource with pornographic materials, then such a page must be completely blocked. In case the page to be opened is subject to partial editing, it is not entered either in the "white" or "black" list, so that the page is checked by the system at each visit. Automatic exceptions generation enables to minimize delays in the user's work with regularly revisited resources and to avoid extra checks if the resource is marked unwanted;
- DBL DB (Distributed Black Lists Database) is a database to which banned sites are added. Program samples are regularly compared with this database. Thus, the more program samples function for various users, the more increased the unwanted site base, and the installed program can access the complete base from the moment of use.

2. Tasks performed by the proxy-server

The proxy-server functionality is limited by the following tasks:

- data provision for in-depth analysis;
- banning the unwanted web resources based on the analysis results.

The given tasks are performed in the function blocks of the proxy server presented in Fig. 2.

In Fig. 2, all blocks are conditionally classified into 5 groups: Configuration, Black and White Lists, SSL, Proxy kernel and Analysis. Each group contains two or more conventional function blocks.

In the Configuration group, there is the operating system and separate programs are fine-tuned for log-on via a proxy-server. Definition of the separate Firefox Proxy Configuration block is explained by the fact that Mozilla Firefox browser does not conform to the internal OS Windows proxy settings and there is the need to elaborate separate methods of its configuration.

The blocks of the Black and White Lists group put into effect work with the white and black exceptions lists. All the sites added to the black list are blocked at the request stage and the sites in the white list will be fully accessible without any time-consuming analysis.

The SSL group realizes the features for processing self-signed SSL certificates. The blocks of this group generate and manipulate such certificates as well as perform caching in order to accelerate establishment of a secured connection.

Proxy kernel is the core of the entire proxy server. Here all the necessary networking functions are realized: transport connection, establishment of a secure connection, HTTP queries / responses processing and a series of other functions, which directly provide for the work with Internet traffic.

All the features connected with data preparation for the further stage are performed by the function blocks of the Analysis group. If there is the need in analyzing a Web resource, a series of preliminary stages must be executed; formation of a "loop" on the side of the proxy-server (browser work simulation), necessary data acquisition, their decryption, transformation and formatting. Also, The Analysis group has the Preliminary Analysis function block, which verifies queries at the stage of their receipt using the set of ready regular expressions.

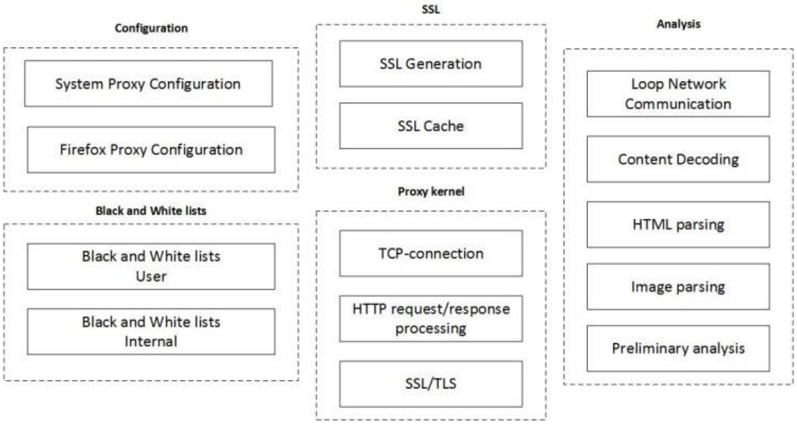
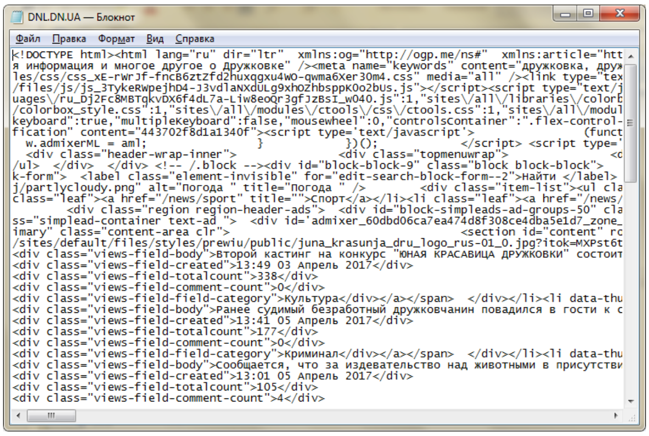

Configuration
System Proxy Configuration
Firefox Proxy Configuration
SSL
SSL Generation
SSL Cache
Analysis
Loop Network Communication
Content Decoding
HTML parsing
Image parsing
Preliminary analysis
Black and White lists
Black and White lists User
Black and White lists Internal
Proxy kernel
TCP-connection
HTTP request/response processing
SSL/TLS

Fig. 2. Function Blocks of the Proposed Proxy Server



```
<!DOCTYPE html><html lang="ru" dir="ltr"  xmlns:og="http://ogp.me/ns#"  xmlns:article="htt
я информация и многое другое о Дружковке" /><meta name="keywords" content="дружковка, друж
les/css/css_xE-rWrJf-fncB6ztZfd2huxqgxu4WO-qwma6Xer30m4.css" media="all" /><link type="tex
/files/js/js_3TykeRwpejhD4-J3vdlaNXdULg9xhOZhbsppKOo2bUs.js"></script><script type="text/j
uages\/ru_Dj2Fc8MBTqkvDX6f4dL7a-Liw8eoQr3gfJzBsI_w040.js":1,"sites\/all\/libraries\/colorb
/colorbox_style.css":1,"sites\/all\/modules\/ctools\/css\/ctools.css":1,"sites\/all\/modul
keyboard":true,"multipleKeyboard":false,"mousewheel":0,"controlsContainer":".flex-control-
fication" content="443702f8d1a1340f"><script type='text/javascript'>              (funct
  w.admixerML = aml;              }             })();            </script> <script type='
  <div class="header-wrap-inner">                  <div class="topmenuwrap">            <d
/ul>  </div>  </div> <!-- /.block --><div id="block-block-9" class="block block-block">
k-form">  <label class="element-invisible" for="edit-search-block-form--2">Найти </label>
j/partlycloudy.png" alt="Погода " title="Погода " />        <div class="item-list"><ul cla
class="leaf"><a href="/news/sport" title="">Спорт</a></li><li class="leaf"><a href="/news/
        <div class="region region-header-ads">  <div id="block-simpleads-ad-groups-50" cla
ss="simplead-container text-ad ">  <div id='admixer_60dbd06ca7ea474d8f308ce4dba5e1d7_zone_
imary" class="content-area clr">                                   <section id="content" rc
/sites/default/files/styles/prewiu/public/juna_krasunja_dru_logo_rus-01_0.jpg?itok=MXPst6t
<div class="views-field-body">Второй кастинг на конкурс "ЮНАЯ КРАСАВИЦА ДРУЖКОВКИ" состоит
<div class="views-field-created">13:49 03 Апрель 2017</div>
<div class="views-field-totalcount">338</div>
<div class="views-field-comment-count">0</div>
<div class="views-field-field-category">Культура</div></a></span>  </div></li><li data-thu
<div class="views-field-body">Ранее судимый безработный дружковчанин повадился в гости к с
<div class="views-field-created">13:41 05 Апрель 2017</div>
<div class="views-field-totalcount">177</div>
<div class="views-field-comment-count">0</div>
<div class="views-field-field-category">Криминал</div></a></span>  </div></li><li data-thu
<div class="views-field-body">Сообщается, что за издевательство над животными в присутстви
<div class="views-field-created">13:01 05 Апрель 2017</div>
<div class="views-field-totalcount">105</div>
<div class="views-field-comment-count">4</div>
```

Fig. 3. Dearchived GZIP Data

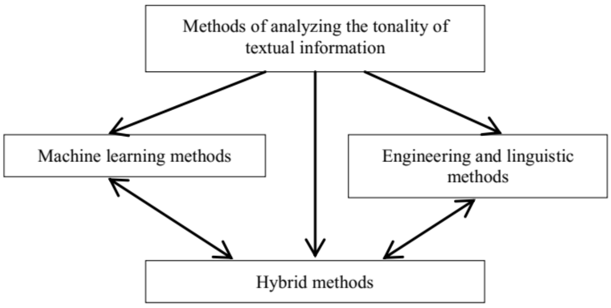

Methods of analyzing the tonality of textual information
Machine learning methods
Engineering and linguistic methods
Hybrid methods

Fig. 4. The methods of analyzing the tonality of textual information

The HTML page in the HTTP response body may be transmitted by the server as clear or archived. If clear data are transmitted, their body transformation into text representation is performed after its definition. If archived data are transmitted, their decompression is needed. One of the most popular compression methods is GZIP. This method is based on the DEFLATE algorithm. Text content of the site under scrutiny after its decompression is presented in Fig. 3.

After dearchiving GZIP data and preparing input text for next group next step will be to make deep analyze of image and text content using intellectual methods. Approaches used to analyze the tonality of textual information in natural language, are divided into two main groups: engineering and linguistic methods and methods based on machine learning (Fig. 4). Machine learning methods, the list of which includes (but is not limited to) the method of conditional random fields (Condition Random Fields, CRF), Support Vector Machine (SVM) method, artificial neural networks (Artificial Neural Networks, ANN) and others. This group of methods uses mathematical models that automatically determine the optimal set of parameters for the solution of a particular problem, in this case, the definition of a key [3, 4].

Considering together the monitoring of social media and the task of automated intellectual analysis of text publications, it can be argued that today it is one of the most relevant and actively developing applied areas of sociology and computer linguistics. This fact can be confirmed by the fact that this direction has become the leading topic of international conferences in the field of computer linguistics and automatic word processing.

A number of recent works on the analysis of the tonality of English texts demonstrate the superiority of deep learning models over shallow algorithms, which include linear and logistic regressions, CRF, Bayesian classifier, and the widely used SVM method. Richard Socher has the best results in this field, in which the algorithm of machine learning for the first time, without the use of linguistic dictionaries or rules, was able to correctly recognize semantic and syntactic negations, correctly select the negative key of the phrase containing a list of positive words and denials between them. Richard Saucer, who has the best results in this field, showed that the machine learning algorithm is able to correctly recognize semantic and syntactic negations, correctly identify the negative key of the phrase containing a list of positive words and negations between them, without the use of linguistic dictionaries or rules. An additional argument in favor of methods of in-depth training (as well as methods of machine learning in general) is the ability to analyze data without an in-depth study of linguistics and/or attracting linguistic experts.

3. SSL Certificate Generation

In the elaborated proxy server, both nonsecure and secure (HTTPS) connections are realized [1].

The problem of secure connection establishment within the framework of this design lay in the field of proper SSL certificates operation. SSL functioning analysis resulted in the need to immediately generate a certificate for the requested resource and create a sort of a non-existent certification center. This is stipulated by the fact that the browsers will “know” that such certification center exists and the certificates signed by this center can be trusted.

The very mechanism of signature generation functions very simply. Equation 1 describes signature generation for *M* message.

$$SIGNATURE = M^{d} (\bmod N) \quad (1)$$

where *d* and *N* are the author’s private key to the message.

M message and *SIGNATURE* can further be sent to the recipient. Signature verification on the second side is described by equation 2.

$$M = SIGNATURE^{e} (\bmod N) \quad (2)$$

where *e* and *N* are the author’s public key to the message.

Thus, everybody can verify the signature of the message but it can be signed only by the person with the private key. The browser authenticates the incoming certificate by means of using the public key of the individual who signed it. Therefore, by creating and storing a certificate, it is necessary to save the private key and, then, in the proxy server work, it is possible to sign SSL certificates immediately using this key. The browser, authenticating such certificates, will verify them with the public key from the certification center, the certificate of which resides in Windows storage.

It is worth noting that at the stage of immediate SSL certificate generation by the browser, internal SSL cache is primarily checked. If the certificate for the requested host is not found, a new certificate is immediately generated. Thus, the connection system presented in Fig. 4 will be formed.

4. Conclusions

The control model of data stream transmitted over a network based on proxying technology was proposed in the work. The speed and correct operation of the proxy server currently depend on the definite browser. Proxy server test results analysis showed that 100% of the sites in the black list are successfully blocked when accessed with Opera, Internet Explorer, Google Chrome, Yandex.Browser browsers. However, the result of working in Yandex is 30% lower. This is associated with multiple queries of this browser to its servers, which results in the fact that it takes a lot of time for the site to open, or it does not open at all at the end of timeout for waiting for the response from the server set by the browser. The best result in the speed of opening Web pages when working via the elaborated proxy server was obtained with Google Chrome. Working via the proxy server in this browser enables not only to surf the Internet without any significant delay, but also to work with media traffic such as watching videos online on YouTube and listening to music.

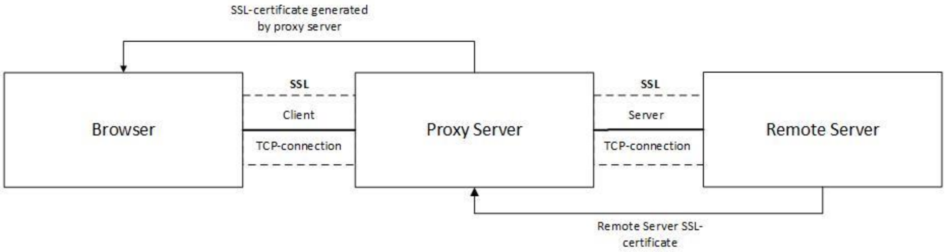


Fig. 4. Scheme of HTTPS Connection via a Proxy Server

References

[1] Barabanov A.B., Markov A.C., Cirlov B.L.: Sertifikatsiya sistem obnaruzheniya vtorzheniy. Otkrytyye sistemy. SUBD 3/2012, 31–33.

[2] Huan T.: The Application of SSL Protocol in Computer Network Communication. Intelligence Computation and Evolutionary Computation. Advances in Intelligent Systems and Computing 180/2013, 779–783.

[3] Kuznetsova E.S., Loukachevitch N.V., Chetviorkin I.I.: Testing rules for a sentiment analysis system. Computational Linguistics and Intellectual Technologies 12(2)/2013, 71–80.

[4] Pak A., Paroubek P.: Language independent approach to sentiment analysis (LIMSI participation in ROMIP ’11). Computational Linguistics and Intellectual Technologies 11(2)/2012, 37–50.

[5] Scherer S., Schwenker F., Palm G.: Emotion Recognition from Speech Using Multi-Classifier Systems and RBF-Ensembles. Studies in Computational Intelligence 83/2008, 49–70.

[6] Sentsova A.Yu., Mashkyna Y.V.: Avtomatizatsiya ekspertnogo audita informatsionnoy bezopasnosti na osnove ispol'zovaniya iskusstvennoy neyronnoy seti. Bezopasnost' informatsionnykh tekhnologiy 2/2014, 65–70.

[7] Wan S., Ren Z.: Research on Full-Proxy Based TCP Congestion Control Strategy. Advances in Intelligent and Soft Computing 133/2012, 371–378.

Ph.D. Eng. Olesia Barkovska
e-mail: olesia.barkovska@nure.ua

Senior lecturer of Electronic Computer Department of Kharkiv National University of Radioelectronics, Ph.D. in Systems and Means of Artificial Intelligence from 2011. Her research interests include data clustering, malware analysis, network security, network and server performance

B.Sc. Vitaliy Serdechnyi
e-mail: serdechny.vs@ukr.net

B.Sc. in Computer Engineering from 2017. Is currently studying for a M.Sc. degree in System Programming in Kharkiv National University of Radioelectronics. His research interests include artificial intelligence, neural networks, Internet security, .NET platform, C#

otrzymano/received: 10.10.2017 *przyjęto do druku/accepted: 15.02.2018*