UDK 004.852

Otrokh S. I. *State University of Telecommunications, Kyiv*
Kuzminykh V. O., Shvets Ye. Yu. *Igor Sikorsky Kyiv Polytechnic Institute, Kyiv*

ANALYSIS OF INFORMATION IN THE TASKS OF CONSOLIDATING FROM OPEN SOURCES

The article discusses the issues and possibilities of using existing parsing algorithms for obtaining and consolidating data from open sources for monitoring the ecological situation in Ukraine. A study was conducted on the time of operation and the use of memory, depending on the number of iterations. Thanks to this research, it is possible to choose the best algorithm for tasks, when data consolidation is necessary.

Keywords: *consolidating data, open sources, monitoring, parsing algorithms, number of iterations.*

Отрох С. І. *Державний університет телекомунікацій, Київ*
Кузьміних В. О., Швець Є. Ю. *НТУУ «КПІ ім. Ігоря Сікорського», Київ*

АНАЛІЗ ІНФОРМАЦІЇ У ЗАДАЧАХ КОНСОЛІДАЦІЇ ДАНИХ З ВІДКРИТИХ ДЖЕРЕЛ

Розглядаються питання та можливості використання існуючих алгоритмів синтаксичного аналізу для отримання та консолідації даних з відкритих джерел для моніторингу екологічної ситуації в Україні. В основі консолідації лежить процес збору та організації зберігання даних у вигляді, оптимальному з точки зору їх подальшої обробки. Проведено аналіз основних алгоритмів парсинга, що існують на сьогодні, було проведено дослідження і зроблені висновки, в яких випадках той чи інший алгоритм краще використовувати. Для проведення аналізу було використано декілька методів з реалізацією на PHP 7 версії. Інструментом тестування була HTML форма з вибором веб-документа за допомогою AJAX запитів при обмеженні часу виконання одного запиту. Інструменти тестування всіх засобів парсинга розроблені за допомогою сценаріїв bash і PHP в операційній системі Linux і за допомогою пакетних файлів batch, сценаріїв PowerShell і PHP в Windows.

Проведено дослідження щодо часу роботи та використання пам'яті в залежності від кількості ітерацій. У статті приведені обрані для тестування веб-документи і їх характеристики (формат, розмір, URL). Для кожного засобу парсинга було проведено тестування з кожним з веб-документів при п'яти різних варіантах їх обробки. На підставі проведеного аналізу зроблено висновок, що для вирішення задач консолідації інформації з відкритих джерел найбільш відповідним та ефективним може бути використання засоби парсинга зі зберіганням ієрархії елементів. Такі засоби надають найбільш зручний спосіб доступу до елементів веб-документа за допомогою запитів (CSS і XPath) і підходять для виконання багаторазових запитів у задачах консолідації інформації з відкритих джерел для моніторингу екологічної ситуації в Україні.

Ключові слова: *консолідація даних, моніторинг, алгоритми парсингу, кількість ітерацій.*

Отрох С. И. *Государственный университет телекоммуникаций, Киев*
Кузьминых В. А., Швец Е. Ю. *НТУУ «КПИ им. Игоря Сикорского», Киев*

АНАЛИЗ ИНФОРМАЦИИ В ЗАДАЧАХ КОНСОЛИДАЦИИ ДАННЫХ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ

Рассматриваются вопросы и возможности использования существующих алгоритмов синтаксического анализа для получения и консолидации данных из открытых источников для мониторинга экологической ситуации в Украине. Проведен анализ времени работы и использования памяти в зависимости от количества итераций. Благодаря этому можно выбрать лучший алгоритм для тех задач, которые могут возникнуть при консолидации данных.

Ключевые слова: *консолидация данных, открытые источники, мониторинг, алгоритмы парсинга, количество итераций.*

1. Introduction

The amount of information produced by a person is constantly increasing, which significantly increases the problem of storing large volumes of information and extracting from them the necessary important information. Every day, the problem of its processing becomes more relevant. With a large number of electronic materials, the user will soon be unable to find the right information without effective methods of processing information [1]. Consolidating information [2] can help solve these problems. In a broad sense, consolidation can be understood as a process of searching, selecting, analyzing, structuring, transforming, storing, cataloging and providing information to consumers on specified topics. The task of consolidating information is one of the most important tasks of processing large volumes of data [3].

Collection of information from the open source is one of the standard methods for gathering information in various spheres of life of modern society. As a rule, specialists collect and analyze information from the media, public reports, official data, press conferences, professional and academic reports, conferences, reports, articles in periodical electronic sources. Electronic data carriers largely determine the approaches and methods of targeted information search and increase the effectiveness of individual information seeking procedures. Consolidation of information is usually considered as a set of methods and procedures aimed at extracting data from different sources, providing the necessary level of their informativity and quality, transforming into a single format in which they can be downloaded to a data warehouse or analytical system.

In some cases, data consolidation is the initial stage of the implementation of an analytical task or a project program information and analytical system [2]. The basis of the consolidation is the process of collecting and organizing data storage in the form optimal in terms of their processing on a particular analytical platform or solving a specific analytical task. Also, consolidation tasks are the assessment of the quality of data and their enrichment, in order to reduce the amount of information that I have to process in an information retrieval or information-analytical system.

The main results that should ensure consolidation of data for further thorough processing is:

– compact storage of large volumes of data;
– support for the integrity of the data structure;
– high-speed access to large volumes of data;
– control of the relevance of data.

The main feature of collecting information on the basis of open sources is the dynamic volatility of the information content of these sources, the lack of reliable a priori information about content and relevance, and, as a rule, a small accuracy of a priori estimates of the state and the relevance of these sources to the topics and query parameters.

For the effective processing of data from open source information, it is necessary to consolidate data using specialized software that will provide:

– justified choice of the most relevant information sources inquiry;
– the possibility of accumulation of information about the correspondence of sources during the execution of the request;
– analysis as the most promising in terms of relevance, and less relevant sources;
– taking into account the possibility of changing the state of the sources upon repeated request;
– taking into account the information evaluation of the perspective from the point of view of the relevance of the sources for their arrangement.

In the process of consolidating information from one of the open sources, the stream of information resources, such as the media, which includes various news sites (RSS feeds) and semantic sites (or electronic versions of the media), plays a special role.

2. Analysis of parsing methods

Parsing is generally understood as the process of analyzing and decomposition certain data into components using special programs or scripts [4]. Sometimes parsing is confused with grabbing. These are pretty close concepts, but they still have different meanings. The grubber only allows you to download information from the network (HTML-pages, RSS-tapes, XML-documents) into your database, and the parser can identify useful information from this bunch and process it, depending on the tasks [5, 6].

Automated collection of large volumes of information at the highest possible speed are performed by special programs that relate to so-called bots [7].

The information extracted by the bots is accumulated in order to select according to certain criteria or analysis, the backup in case of loss of access to the primary sources, for periodic updating and further processing.

Currently, there are a number of different methods of parsing, which are based on various parsing tools. The choice of this or that method for a specific task requires a serious justification. The qualitative and quantitative results of parsing depend to a large extent on the correctness of this choice. This is especially important for such large tasks as the task of consolidating information from open sources.

Several methods were used to conduct the analysis with the implementation of PHP 7 version. The testing tool was the HTML form with the choice of a web document using AJAX queries. The time limit for executing a single AJAX request was 300 sec. Testing tools for all parsing tools are designed using bash and PHP scripts in the Linux operating system, using batch files, PowerShell scripts, and PHP in Windows.

In automatic mode, HTML, XHTML, XML formats of web documents of different sizes and various iterations are processed. The working time of each parsing tool was 300 seconds. The web documents selected for testing and their characteristics (format, size, URL) are presented below:

– the default Wordpress blog page, HTML, 75 KB, https://en.blog.wordpress.com/;
– google's search server (SERP) for parsing php, HTML, 333KB, https://www.google.com/search? hl = en & q = parsing + php;
– the top-level list of top-level domains for the online encyclopedia Wikipedia, HTML, 934 KB, https://en.wikipedia.org/wiki/ List_of_Internet_top-level_domains;
– an article page containing over 900 comments on the Habrahabr resource, HTML, 2.45 MB, https://habrahabr.ru/post/70330/;
– an official list of all MIME types on the IANA resource, XHTML, 584 KB, http://www.iana.org/assignments/media-types/media-types.xhtml;
– catalogs of the Ozon.ru store of various sizes, XML, 164KB, 1.63MB, 23.1MB, 159MB, http://www.ozon.ru/context/detail/id/2643202/.

For each parsing tool, we tested each web document with five different processing options. In Fig. 1 as an example, parsing tools are shown and signed in order of decreasing processing time.

In Fig. 2 shows the difference in processing time of an XML document in the volume of 23.1 MB between the similar type of parsing tools – Stream Expat and Expat with conversion to the structure, the SimpleXML standard and SimpleXMLReader, phpQuery and QueryPath, using the PhantomJS console web browser and the standard DOM API.

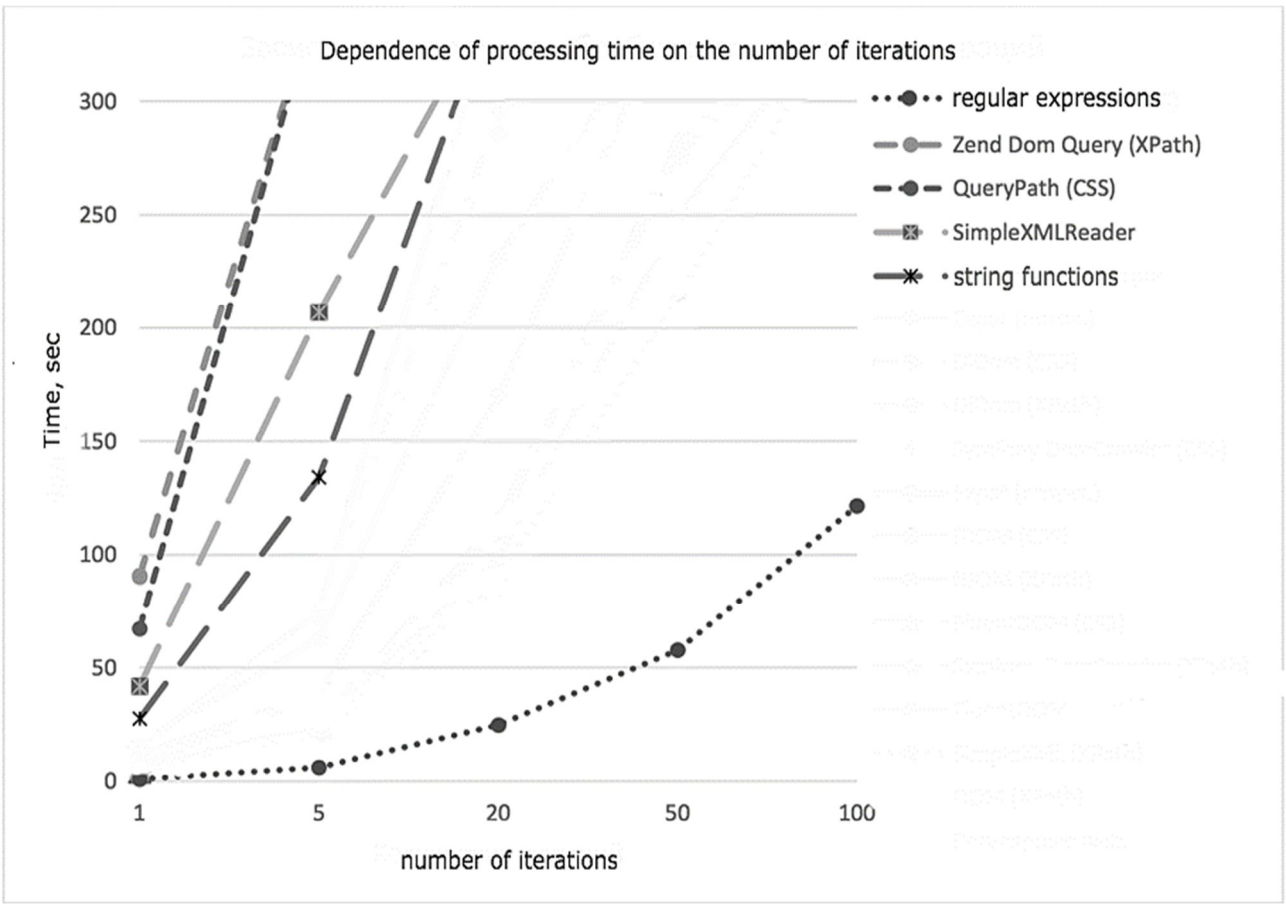

Fig. 1. Dependence of the processing time of the XML document of the "Ozon.ru" directory in the size of 159 MB surrounded by Windows with PHP 7 version from the number of iterations

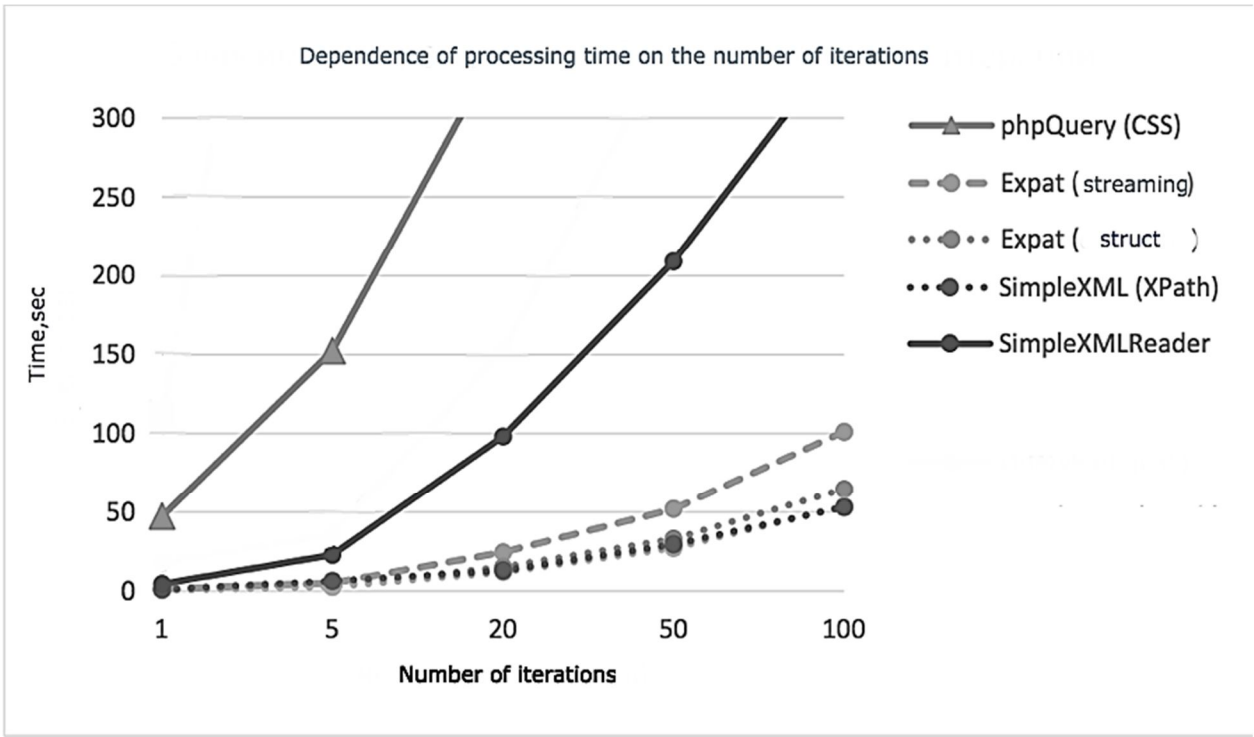

Fig. 2. Dependence of the processing time of the XML document of the directory "Ozon.ru" size 23.1 MB surrounded by Linux with PHP 7 version of the number of iterations

An analysis of the effectiveness of work of the means of parsing in three environments allows us to talk about their adaptability. The graphs show examples of measurements only for some of the environments.

Use of volumes of RAM by means of parsing is shown in Fig. 3. If memory consumption increases with an increase in the number of iterations, it indicates an ineffective use of memory or problems with the collection of unnecessary information.

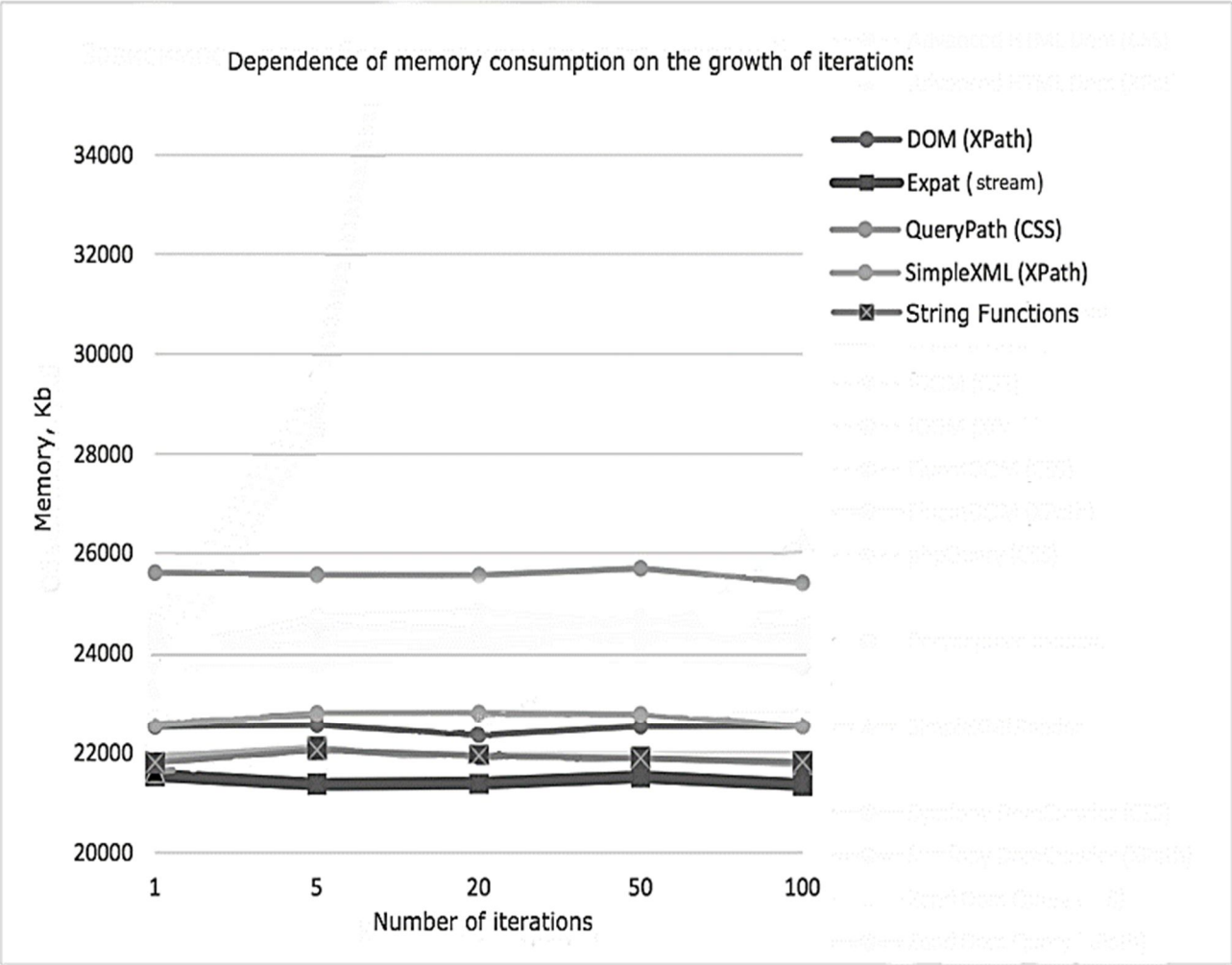

Fig. 3. Dependence of memory consumption when processing an XML document in an environment of Linux with PHP 5 version from the number of iterations

3. Conclusions

Main conclusions from the analysis:

– Use of volumes of RAM by means of parsing is shown in Fig. 3. If memory consumption increases with an increase in the number of iterations, it indicates an ineffective use of memory or problems with the collection of unnecessary information.Parsing methods that store the hierarchy of elements in which each web document being analyzed are stored in memory as a hierarchy of its elements (nodes); when re-accessing different parts of the document, there is less time to process the web document, since it is not necessary to reload and build a tree.

– Streaming parsing methods are applicable in tasks where there is no need to reapply. In this case, there is a sequential overflow of the nodes to the desired and alternate boot into the memory of only the current node. These methods can analyze very large documents.

– A parsing method with regular expressions can be used for tasks where it is necessary to extract a specific part of a document with a pre-known structure. Such a search guarantees an effective result under the condition of well-defined expressions. Such a parsing method is more appropriate to use where the funds that use standard extensions of the LibXML library.

Thus, for solving consolidation problems of open source information [8], the use of parsing with the storage of elements hierarchy can be most appropriate and effective. These methods provide the most convenient way to access Web document elements using queries (CSS and XPath) and are suited to any task of reusing a sample of information from a web document, which is very important when performing multiple requests for consolidation of open source information.

References (MLA)

1. Koval O., Kuzminykh V., Khaustov D. "Using stochastic automation for data consolidation." *Research Bulletin of NTUU "KPI"*. Engineering 2 (2017): 29-36.

2. Shakhovska N. B. "Methods for processing consolidated data using data spaces." *Problemy prohramuvannia* 4 (2011): 72-84.

3. Chernyak L. "Big data – a new theory and practice." *Moskva: Open Systems* 10 (2011): 36-41.

4. Schrenk M. "Webbots, spiders, and screen scrapers 2nd Edition: a guide to developing internet agents with PHP/CURL." *No starch press inc.* (2012): 362.

5. "Benchmark HTML parsers/" [*Electronic resource] The site of "Habrahabr"*. (2012): URL: https://habrahabr.ru/post/163979/.

6. "Morgan C. XML for PHP developers: Part 2. Advanced XML parsing techniques." *[Electronic resource] Community developerWorks* (2010):
URL: http://www.ibm.com/developerworks/ru/library/x-xmlphp2/index.html.

7. Rasti Harold E. "XML Parsing Analysis in PHP." *[Electronic resource] Community developerWorks* (2007. 11 October (2007):
URL: http://www.ibm.com/developerworks/ru/library/xpullparsingphp/index.html

8. Kuzminykh V. O., Koval A. V., Osypenko M. V. "Methods of machine training on the basis of stochastic automatic devices in the tasks of consolidation of data from unsealed sources." *[Electronic resource] CEUR workshop proceedings - Vol-2067 urn:nbn:de:0074-2067-8* (2017): 63-68. URL: [http://ceur-ws.org/Vol-2067/paper9.pdf]

Список використаної літератури (ДСТУ)

1. Kuzminykh V. Using stochastic automation for data consolidation / V. Kuzminykh, O. Koval, D. Khaustov // Research Bulletin of NTUU "KPI". Engineering. – 2017. – №2. – С. 29-36.

2. Шаховська Н. Б. Методи опрацювання консолідованих даних за допомогою просторів даних/ Н. Б. Шаховська // Проблеми програмування. – 2011. – № 4. – С. 72-84.

3. Черняк Л. Большие данные – новая теория и практика / Л. Черняк // Москва: Открытые системы. – 2011. – № 10. – С. 36-41.

4. Schrenk M. Webbots, spiders, and screen scrapers: a guide to developing internet agents with PHP/CURL. / M. Schrenk // No Starch Press Inc. – 2012. 362 p.

5. Бенчмарк HTML парсеров [Електронний ресурс] // Сайт «Хабрахабр». – 2012. URL: https://habrahabr.ru/post/163979/.

6. Морган К. XML для PHP-разработчиков: Часть 2. Расширенные методы парсинга XML: [Електронний ресурс] / К. Морган // Сообщество developerWorks. – 2010. URL: http://www.ibm.com/developerworks/ru/library/x-xmlphp2/index.html.

7. Расти Хэролд Э. [Rusty Harold E.] Синтаксический анализ XML в PHP: [Електронний ресурс] // Сообщество developerWorks. – 2007.
URL: http://www.ibm.com/developerworks/ru/library/xpullparsingphp/index.html.

8. Kuzminykh V. O., Koval A. V., Osypenko M. V. Methods of machine training on the basis of stochastic automatic devices in the tasks of consolidation of data from unsealed sources [Електронний ресурс] / V. O. Kuzminykh, A. V. Koval, M. V. Osypenko // CEUR Workshop Proceedings - Vol-2067 urn:nbn:de:0074-2067-8. – 2017. – P. 63-68. URL: [http://ceur-ws.org/Vol-2067/paper9.pdf]

Автори статті

Отрох Сергій Іванович – кандидат технічних наук, професор, завідувач кафедри мобільних та відеоінформаційних технологій, Державний університет телекомунікацій, Київ, Україна. Тел.: +380 (66) 981 24 39. E-mail: 2411197@ukr.net.

Кузьміних Валерій Олександрович – кандидат технічних наук, доцент кафедри автоматизації проектування енергетичних процесів і систем, НТУУ «КПІ ім. Ігоря Сікорського», Київ, Україна. Тел.: +380 (95) 012 66 63. E-mail: vak0202@ukr.net

Швець Єгор Юрійович – студент магістратури кафедри автоматизації проектування енергетичних процесів та систем, НТУУ «КПІ ім. Ігоря Сікорського», Київ, Україна. Тел.: +380 (99) 973 32 23.
E-mail: egor.shvetz2012@gmail.com

Authors of the article

Otrokh Serhii Ivanovych – candidate of sciences (technic), professor, head of mobile and video information technologies department, State University of Telecommunications, Kyiv, Ukraine. Tel .: +380 (66) 9812439. E-mail: 2411197@ukr.net.

Kuzminykh Valerii Oleksandrovych – candidate of sciences (technic), associate professor of design automation of energy processes and systems department, Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine. Tel.: +380 (95) 012 66 63. E-mail: vak0202@ukr.net

Shvets Yehor Yuriiovych – master's degree student of design automation of energy processes and systems department, Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine. Tel.: +380 (99) 973 32 23. E-mail: egor.shvetz2012@gmail.com

Дата надходження
в редакцію: 12.01.2018 р.

Рецензент:
доктор технічних наук, професор
О. В. Барабаш
Державний університет телекомунікацій, Київ