ITHEA

International Journal INFORMATION MODELS & ANALYSES

2017 Volume 6 Number 2

International Journal
INFORMATION MODELS & ANALYSES
Volume 6 / 2017, Number 2

EDITORIAL BOARD
Editor in chief: **Krassimir Markov** (Bulgaria)

Alberto Arteta	(Spain)	**Levon Aslanyan**	(Armenia)
Albert Voronin	(Ukraine)	**Luis Fernando de Mingo**	(Spain)
Aleksey Voloshin	(Ukraine)	**Liudmila Cheremisinova**	(Belarus)
Alexander Palagin	(Ukraine)	**Lyudmila Lyadova**	(Russia)
Alexey Petrovskiy	(Russia)	**Martin P. Mintchev**	(Canada)
Alfredo Milani	(Italy)	**Nataliia Kussul**	(Ukraine)
Anatoliy Krissilov	(Ukraine)	**Natalia Ivanova**	(Russia)
Avram Eskenazi	(Bulgaria)	**Natalia Pankratova**	(Ukraine)
Boris Tsankov	(Bulgaria)	**Nelly Maneva**	(Bulgaria)
Boris Sokolov	(Russia)	**Olena Chebanyuk**	(Ukraine)
Diana Bogdanova	(Russia)	**Olga Nevzorova**	(Russia)
Dmytro Progonov	(Ukraine)	**Orly Yadid-Pecht**	(Israel)
Ekaterina Solovyova	(Ukraine)	**Pedro Marijuan**	(Spain)
Evgeniy Bodyansky	(Ukraine)	**Rafael Yusupov**	(Russia)
Galyna Gayvoronska	(Ukraine)	**Sergey Kryvyy**	(Ukraine)
Galina Setlac	(Poland)	**Stoyan Poryazov**	(Bulgaria)
George Totkov	(Bulgaria)	**Tatyana Gavrilova**	(Russia)
Gurgen Khachatryan	(Armenia)	**Tea Munjishvili**	(Georgia)
Hasmik Sahakyan	(Armenia)	**Valeria Gribova**	(Russia)
Ilia Mitov	(Bulgaria)	**Vasil Sgurev**	(Bulgaria)
Juan Castellanos	(Spain)	**Vitalii Velychko**	(Ukraine)
Koen Vanhoof	(Belgium)	**Vladimir Ryazanov**	(Russia)
Krassimira B. Ivanova	(Bulgaria)	**Yordan Tabov**	(Bulgaria)
Leonid Hulianytskyi	(Ukraine)	**Yuriy Zaichenko**	(Ukraine)

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.

International Journal "INFORMATION MODELS AND ANALYSES" Volume 6, Number 2, 2017

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,

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

Printed in Bulgaria

Copyright © 2017 All rights reserved for the publisher and all authors.
® 2012-2017 "Information Models and Analyses" is a trademark of ITHEA®
® ITHEA is a registered trade mark of FOI-Commerce Co.

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

MAJOR TERMS FOR THE EFFECTIVE FUNCTIONING OF THE EXPORT SUPPORT STRATEGIES

Giorgi Gaganidze

Abstract*: The field of the proposed article has been the author's major point of interest for the last 16 years. The paper reviews specific topics of the formulation and management of the Export Support Strategies. The author reviews the role of the export alliances, export management companies and trading companies in the functioning of the efficient export support strategies.*

Keywords: *export support strategy, export alliances, export management company (EMC), export trading companies.*

ITHEA Keywords*: A.1 Introductory and Survey*

Introduction

Globalization of the world economy has made competition very tough. Companies couldn't feel safe and secure even on their domestic markets. Thus, from the second half of the 20th century different export support activities were launched, which could be divided into three main blocks: financial, informational and specific. Financial support activities included special tax exemptions, low interest rate credits, export insurance schemes etc. These activities have been qualified as discrimination practice and the World Trade Organization member states take obligation not to use them. Information services included general information as well as specific market research activities provided by respective Governments or with their support. Under the informational activities could be considered special training programmes provided for the export oriented companies' staff, while charges were claimed to the Government Agencies. Specific activities include support for companies in participation in different Trade Fairs and Shows and organization of Promo actions.

Nowadays different mechanisms and their combinations are used; special attention is paid to staying in line with WTO Rules and Procedures. At the same time one question always asked is: how efficient are these support activities? How can we prove by outcomes that Government funds have been utilized in the best possible way?

When considering these questions, majority of authors put emphasis on an identical problem of all the export support strategies, this is the fact, that the same approach is used with all exporting companies.

Using the same approach had its one positive aspect; by using a standard approach we are providing standard activities, preventing different treatment for different exporters and thus preventing lobbying of interests. At the same time identical approach didn't allow considering specificity of different industries. In practice even two companies interested in the same export market needed different treatment. In Georgian reality besides the above mentioned problems, additional specific factors should be considered. First of all, because of the low number of exported products in some industries Georgia has only one exporter. Thus, supporting export of such products actually would mean support of a single company. All these problems of the export oriented companies transformed into a major decision: which model the companies are going to use: standardization of international support programs or their adaptation. Adapted programs are better focused on specific market needs, but their costs are high. Standardized programs are cheap but do not take into account market differences. Attention to this topic is quite high both from theoretical and practical points of view. Some researches showed positive relations between adaption and performance [1]; some researches showed no relations between the two, [2]; and some researches showed negative relations [3]. To review this topic in detailed manner it would be better to range exporters in order to see a bigger picture and thus make adequate decisions.

Ranging exporters by some major characteristics

Exporters could be ranged by different factors. "Drawing from contingency and the organizational learning perspective, the authors develop and test a model of the effects of different forms of international experience - duration, scope and intensity – on the performance outcomes of promotion adaptation" [4].

The above mentioned factors should be defined as follows: duration factor means, when company entered the export operations, volume defines the number of the export markets, intensity defines dependence on exports. These factors create practical experience of the exporting company and heavily influence identification of the best support forms for their export activities. For efficient functioning of the export support strategies crucial factor is - who provides the support activities. Low efficiency of the export support strategies could be explained by the fact that usually providers are governmental structures. More recently specialized structures have been more frequently used. World practice revealed efficiency by using specialized trading companies. It is very important to use the most appropriate trading intermediary. Just review them in general.

To begin with, we should make a distinction between trading companies and export management companies, as they are providing different functions. Trading companies buy and sell goods, while export management companies based on their market intelligence and practice assist exporters in defining the best buyers. Thus, trading companies perform all the functions of the seller – packaging, marking, forming selling party of the product etc. Trading companies are very depended on financial

resources, which are vital for reducing costs and taking adequate profit. Export management companies didn't undertake these functions and mainly assist exporters in meeting importers' needs. For sure, they give recommendations for exporters, but the major field of their activities is creation of a good selling environment based on the market conditions. We also should review creation of export alliances. Broadly speaking, there are different types of export alliances. The simplest one is, when the same product exports create an alliance in order to better utilize opportunities of the new export market.

Quite often these alliances are called horizontal alliances, as the members of the alliance are on the same stage of the value chain. In this alliance exporters are pursuing the same objectives: information on the export country, co-financing of the promo actions and joint market research. This form of the export alliance plays less and less significant role. Vertical variant of the export alliance is the alliance where members are on different stages of the value chain, such as suppliers, producers, trading companies, international marketing intermediaries and financial-banking structures. Participation of the banks becomes crucial, as it supports solution of a tough financial problem, associated with high expenses of financial resources for exporting operations. High economic and political risks are putting banks in the position, when they are asking for higher interest rates for financial resources needed for export operations. The same is true about participation of the insurance companies in the export alliances. In this case, members of the export alliances perform their general functions and thus assist each other in reaching the major goal - increase export sales or enter new export markets. When using export alliances, you face different theoretical and practical problems, for example: how profits would be distributed, who would sell on the export market etc. As the practice proves there is not a single model, which could solve all problems. You should consider industry and export markets specifics and take into consideration existing selling practices. Generally speaking, export management companies would better support exporters in obtaining relevant information, in other cases it would be better to use trading companies or export alliances. As for Georgian realities, problems with the legislation should be solved in the first place. This is absolutely necessary in the case of the export alliances, where we don't have legislation of any type. With re-trading companies we could use the status of the special trading companies, but the sphere of the use should be widened. Majority of trading companies are profitable while working in both directions such as exports and imports. Also we should consider the fact, that in some products (fruits, vegetables) seasonal fluctuations are high. Also we should bear in mind that in Georgian exports participation of the imported products is very high. Thus the effective use of the Special Trading Company status needs quite a lot specific decisions and changes in legislation, first of all in the Tax Code.

Also we should base any type of the export support strategy on the practical needs of the exporters, thus they should define the nature and scope of the support strategies. So the recommendations from the exporters should serve as the basis for creation of the export support strategies.

Exporters Survey Results

In order to find out views and ideas of exporters in 2016 the exporters' interviews were organized. The research was undertaken under Tbilisi State University Scientific Grant Project "The Ways to Improve Trade Balance of Georgia". Electronic survey was sent to exporters, whose export was more than 1 mln. $ per year. For the sake of clarity re-exporters have been excluded. The exclusion of re-exporters was done in order to have clear picture of the needs of the exporting community. In the organization of the research, Georgian Chamber of Commerce and Industry was actively involved.

The First survey was sent to 433 exporters on May 30, 2016. A second time surveys were sent on June 14, 2016, to 424 companies. 9 companies have been excluded due to the irrelevant contact details. The response was obtained from 44 companies, which represent 10% of the interviewed companies. According to the Georgian business practice this response rate could be considered as normal and thus the obtained information is valid.

The analysis of the survey results leads to some interesting judgments. First of all, the total majority of the exporters indicates trade missions on the target markets as the major priority for the export support strategies, also they admit financial support, as for the geographic direction far east was identified as the major priority, at the same time mentioning CIS and EU as the major destinations for Georgian exports. Trade missions organized with the Government support could be considered as combination of the information and financial support. As the best international practice, Government supports exporters in participating in the Trade fairs or organizes trade mission on the target market. It should be mentioned that in the frame of the Tax reform in Georgia costs of the participation in trade missions could be considered as special type of investments and excluded from the profit tax. This would be a serious step for stimulating exporters. Participation of the export management companies in this process would be crucial. Market information and knowledge of the export management companies should serve as the basis for planning these activities. In addition, we should review financial assistance topics. Setting up and development of the export trading companies could be nominated as number one priority. Trading companies will buy and then sell products on their own. This model could quite efficiently work with non-brand products, as for the brand products this model wouldn't be effective. Also export alliances could solve financial problems, mainly when banks and insurance companies are participating in the alliances. It should be underlined that these models are increasing sales on the existing markets or utilizing potential of the new export markets. Unfortunately none of these models are contributing to the creation of new export products. The low number of exported products could be identified as the major obstacle for development of Georgian exports. Detailed analysis of this problems revealed that the creation of the new export products could be achieved only when all directions of the export supporting strategies work together. These directions could be identified as investment, detailed

analysis of the market needs and creation of the unique market offer. Taking into consideration Georgian economic realities, majority of these problems could be solved through specific export alliances, where international companies would also participate. In this case it would be preferable to introduce export alliances international EMC-s, active participation of the banks should be welcomed, as they would serve as major investors for the creation of the new export products. Preparing export offer should be considered a high portion of the services, where Georgia has objective competitive advantages.

We should also review export support strategies for the services. It would be preferable to create export alliances; this should be done everywhere where putting product and service together would be possible. In the services export, where costs of the materials do not play important role stress should be put on the certification activities, this would increase competitiveness of Georgian service exporting companies. For a short time period some service exports should be free from profit tax.

Bibliography

1. Shoham (1999), "Bounded Rationality, Planning, Standartization of International Strategy and Export Performance: A Structural Model Examination," Journal of International Marketing, 792), 24-50
2. Lages, Sandy D. Jap, and David A. Griffith (2008), "The Role of past Performance in Export Ventures:A Short-TermReactive Approach," Journal of International Business Studies, 39, (2), 304-325
3. Cavusgil, S. Tamer and Shaoming Zou 91994), "Marketing Strategy-Performance Relationship: An Investigation of the Empirical Link in Export market Ventures,' 42(1), 225-245
4. Magnus Hultman, Constantine S. Katsikeas and Matthew J. Robson, "Export promotion Strategy and Performance: The Role of International Experience." Journal of International Marketing, Vol.19, No.4 (2011),pp 17-39.

Authors' Information

Giorgi Gaganidze *-Professor, Ivane Javakhishvili Tbilisi State University, Georgia*

E-mail*: giorgi.gaganidze@tsu.ge*

Major Fields of Scientific Research*: The field of the proposed article has been the author's major point of interest for the last 16 years. The paper reviews specific topics of the formulation and management of the Export Support Strategies. The author reviews the role of the export alliances, export management companies and trading companies in the functioning of the efficient export support strategies.*

ANALYSIS OF WEB USER ACTIVITY DATA

Oleksandr Kuzomin, Tetiana Tolmachova, Oleh Astappiev

Abstract*: The goal of this paper is to investigate and to analyze tracked information from visited web pages from the users. It is important to have an instrument for collecting data about users' activity.*

We present a project - an implementation for tracking user activity on web pages. First of all, we created a proxy server, that helps us to provide a JavaScript tracking code on each web page. Using that JavaScript code we collect all types of user activity, time of active and stay time on the page and some additional information about the session of the user. Than was developed tracker to look through all collected information and to analyze it.

Keywords*: proxy server, tracker, user activity, mouse tracking, data analyzes.*

ITHEA Keywords*: J.1 Administrative Data Processing, J.3 Social And Behavioral Sciences.*

Introduction

The web grows very quickly every day and with it grow a number of users of different web sites, that are interested in different fields and interests of theirs lives.

Web analytics is a fast growing field and because of this every day more and more owners of company sites want to know what the customer of their product is interested in and what can be improved regarding interests of their customers.

Research shows that companies using analytics for decision making are 6% more profitable than those that don't [McAfee 2012]. Companies understand the value of using web analytics tools. The programs, which provide web analytics, give an opportunity to retrieve information about user location, their actions, their interactions with the site and products, and also helps to understand how to attract users more and again. In addition, there are plenty number of tools, which can predict user activity in your website in near future.

Web analytic it is a part of user activity monitoring (UAM), which means monitoring and recording user actions. UAM software can make video recordings of sessions, log and analyze the data, capturing file or screenshots. All of that researchers also use in the field of information security to detect and stop external threats.

In this work, we analyze most popular recent researches and tools in the field of tracking user activity. After that for a deeper understanding describes and illustrates foundations of the building of proxy server, tracker, their interaction with each other, its usefulness, how everything work and how we use them in our developed program. In addition, we analyze results of retrieved data from user's actions.

Foundations of tracking users activity

First of all, it is very interesting to look what users are searching for, what they have choose regarding the search (results from those parts can be used in the field of Personalization and User Modeling) or even when the system try to predict what user will choose on the next step.

Web Mining plays a big role. It can be translated as "data mining on the Web". Web Intelligence or Web-Intellect is ready to "open a new chapter" in the rapid development of e-business. The ability to determine the interests and preferences of each visitor, observing his behavior, is a serious and critical advantage of competition in the e-commerce market.

Web Mining systems can answer on many questions, for example, which of the visitors is a potential customer of the Web-store, which group of customers the Web-store brings the most revenue, what are the interests of a certain visitor or group of visitors.

Web Mining technology encompasses methods that are able to discover new, previously unknown knowledge based on the site's data and which can later be used in practice. In other words, Web Mining technology uses Data Mining technology to analyze the unstructured, heterogeneous, distributed, and a large amount of information contained on Web sites.

Analysis of the use of web resources is also can be very helpful. This direction is based on the extraction of data from the logs of web servers. The purpose of the analysis is to identify the preferences of visitors when using certain Internet resources.

It is extremely important to carry out a thorough preprocessing of the data: delete the extra log entries that are not interesting for analysis. Web Usage Mining includes the following components:

– Preliminary processing;
– Operational identification;
– Tools for detecting patterns;
– Tools for analyzing templates.

Each user of the network has his/her own individual tastes, views, depending on which he visits those or other resources. Having identified which pages and in what sequence the user opened, one can draw a conclusion about his preferences. Analysis of the general trend among all visitors shows how efficiently

the electronic portal works, which pages are visited most, what less. Based on this analysis, you can optimize the site: find previously not noticed problems in the functioning, design, and so on. This direction of Web Mining is also sometimes called click stream analysis, an ordered set of page visits that a user viewed when he came to a website.

The data required for analysis is found in server logs and cookies. When the web page is loaded, the browser also requests all objects inserted into it, for example graphic files. In this regard, there is a problem with the fact that the server adds to the log records of each such request. Hence the need for preprocessing data. After the individual page views are highlighted by the user, they are combined into a session.

Once the data has been cleaned and prepared for analysis, it is necessary to ask the following questions like which page is the common entry point for users? Do visitors visit the site through a specially designed page, or do they immediately reach other pages? In what order were the pages viewed? Does this order correspond to what the developers expect from users? What other web portals are sending users to the site being researched? Which sites receive the largest and smallest number of users? How many pages does the user usually view? How long have visitors been on the site? How is the page the most frequent point of departure of users from the site? Why do visitors leave the site with this country? Is it specifically foreseen for this, or are there any reasons that frighten the user off the site?

We can answer on all of those questions. Let us now follow to the next question: why do you need to preprocess web data:

– Data set must be filtered from records generated automatically together with the page load;

– Delete records that do not reflect user activity. Web bots automatically scan many different pages on the network. Their behavior is very different from the human, and they are of no interest from the point of view of the analysis of the use of web resources;

– Definition of each individual user. Most of the portals on the Internet are accessible to anonymous users. You can apply information about registered users, available cookies to determine each user;

– Identify user session. This means that for each visit, the pages that were requested and their order of viewing are determined. Also try to evaluate when the user left the website;

– Finding the full path. Many people use the "Back" button to return to the previously viewed page. If this happens, the browser displays a page that was previously stored in the cache. This leads to "holes" in the log of the web server. Knowledge of the topology of the site can be used to restore such omissions.

The initial data obtained from the log now need to be preprocessed. We can extract:

1. Page views;
2. Identification of each user;
3. User session;
4. The order of pages viewed;
5. Duration.

Researches and developments in the field of tracking users' activity

A group of researchers made one of the most popular research in our selected field. Authors propose a great solution for user activity tracking. In comparison to other approaches that were proposed by other developers, developed system does not require any manual preparation of web pages for tracking them. Its architecture allows adding some more functionality – for example, it would be possible to add code, which forces the user to move the mouse where they are looking at the same moment [Atterer 2006].

It is necessary to take into account the fact that recording of all types of activity that are connected with some background works on the site or recording cookies must be asked firstly for a permission from user side.

In order to collect all data about users' activity, authors use the approach to create a proxy, which make a connection between client and server (Figure 1). It saves log data with details about any requests sent to servers and the replies that a server sends back.

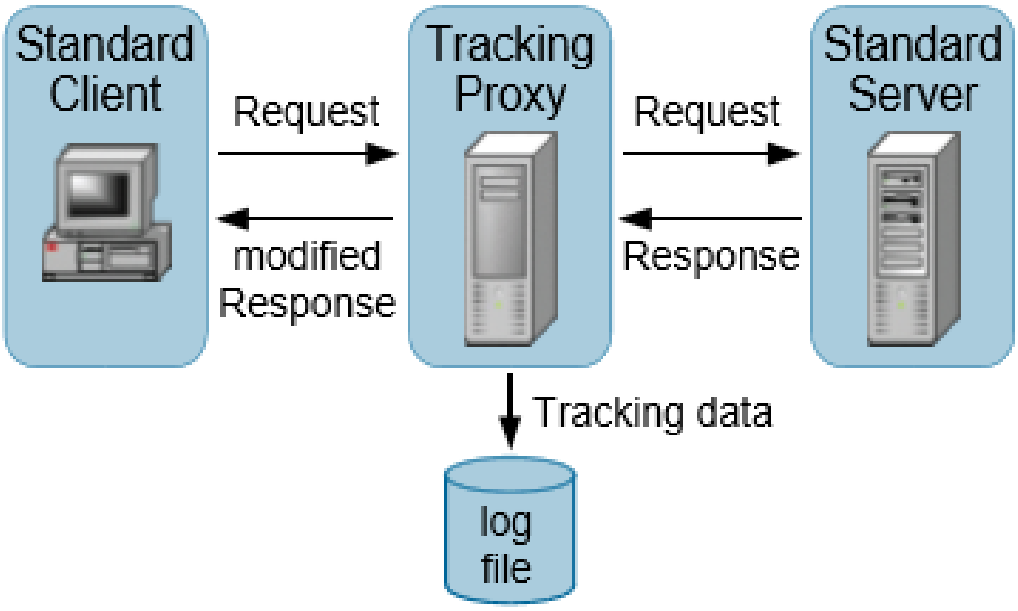

Figure 1. HTTP proxy for inserting JavaScript code into HTML pages for tracking users activity

Regarding of analyzing user activity there is another good example, which is platform Zooniverse from University of Southampton - it is a web-based citizen science platform with a userbase of over 1 million participants, which are Citizen Science. That means that such kind of participant take part and contribute to scientific discovery without the need for specific knowledge or expertise [Tinati 2014]. All of those citizen science participants are volunteers that help the platform in their free time without any reward. That platform includes now 77 smaller projects, where participants have to do some tasks. For instance, in the project "Elephant expedition" the goal is to find elephant or other animals (if they are exist) on the photo that was made by a photo-trap. In total, there are 11 categories, including animals, human and vegetation (no animal). The weakness of this project is that the biggest part of the photos were taken by photo-traps due to the fact that the wind sways the grass and trees and participants of the project have to go through and check all these photos.

The developers of this platform carried out a number of studies to collect users' data for five months. In total 61833 users have been on the platform. On Figure 2 shows regions with corresponding top three regions contributed countries, where users done classifications.

Region	Classifications
Europe (UK, Germany, France)	3688453 (48.2%)
North America (USA, Canada, Mexico)	3071134 (40.2%)
Oceania (Australia, New Zealand, Tanzania)	347818 (4.6%)
Asia (Singapore, India, Japan)	277536 (3.6%)
Far East	*37278 (0.5%)*
Middle East	*15318 (0.2%)*
South America (Brazil, Argentina, Chile)	154807 (2.0%)
Africa (South Africa, Egypt, Kenya)	50045 (0.7%)

Figure. 2. Classifications made aggregated by geographical region

Review of the developed program

The main idea of the development system is to develop a program for recording detailed data for the analysis of user actions without many limitations of existing tools.

General requirements:

1. A detailed record of user actions on web pages;
2. Independence from the operating system and browser version;
3. "Transparent action" - the page view of the user should not be changed;
4. Record additional user information such as operating system version, browser version, screen size, default language and IP address.

The solution is to develop a proxy server – WAPS Proxy and tracker – LWTracker. The general scheme is presented in Figure 3.

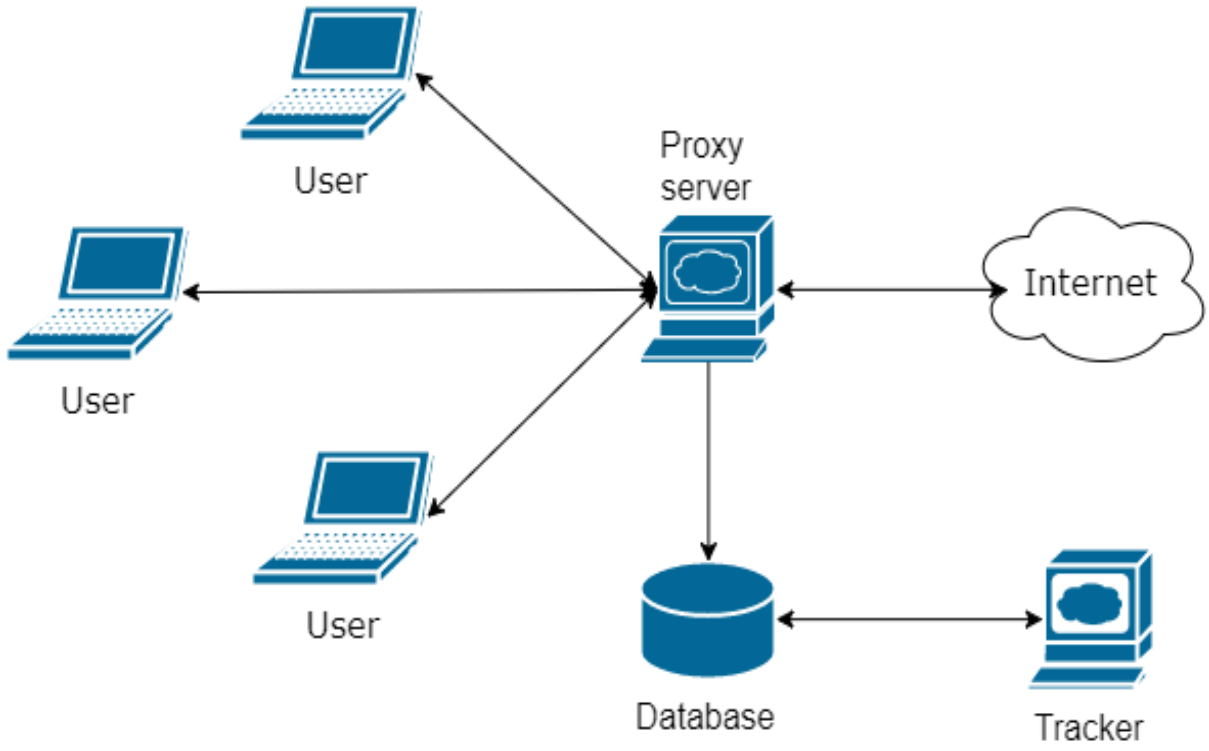

Figure. 3. General scheme of work

A proxy server is a server in computer networks that allows clients to perform indirect requests for network services. First, the client connects to the proxy server and asks for some resource located on another server. Then the proxy server connects to the specified server and receives a resource from it. In some cases, a proxy server may modify a client request or server response for a specific purpose. It is in the case of collecting user activity data that we use a proxy developed between the client and the server.

Most often, a proxy server used for:

– Caching data;

– Data compression;

– Protection of the local network from third-party access;

– Limited access from the local network to the external one;

– Anonymous access to various resources.

Before the proxy passes HTML data back from server to client – the HTML page changes. This modification causes the JavaScript to be loaded into the context of the page. While the page is being displayed, the JavaScript code collects the data and sends them to the proxy for writing to the database. It should be noted that the display of the page does not change.

Consider the principle of the proxy server. When a user goes to a link, for example www.wikipedia.org, we are redirecting to a subdomain on our server, that is, the link will look like this: www.wikipedia.org.waps.io.

To be able to reconstruct user activity on a web page, it was decided to track the following user actions:

– Navigation, that is, the path from one website to another;

– Timelines, such as the active time on the page (when certain actions were taken), passive (total) time and every action in milliseconds for subsequent playback;

– User actions on the web page related to the mouse. This includes the absolute (window ratio) position of the mouse. That is why we record the size of the user screen;

– Other user actions, such as movement of the cursor on a web page, vertical navigation (scrolling), window size change, text input in text fields, clicking on buttons, pressing any key and clicks on the page.

A brief description of the proxy developed is also available on the official website www.waps.io. The LearnWeb platform is currently actively using a proxy server and tracker.

Evaluation and analysis

In our dataset, we have 120581 number of tracks. Regarding stored data from users, first of all we want to analyze how much tracks were done by users in different periods of time. Based on this, we calculated how much number of tracks were performed by users: from 12 PM to 8 AM (non-woring

hours) – 8044 number of tracks, from 6 PM to 12 PM – 35241 number of tracks and in working hours from 9 AM to 5 PM – 77296 number of tracks.

The next step is to find the number of unique users, that done search queries in the Web in each hour.

Regarding the results, the fewest number of users were at 4 AM – 21 users, the largest number of users were tracked at 11 AM – 570 users. Also, the number of users more than 500 were tracked at 10 AM – 518 users, 12 AM – 561 users, 3 PM – 544 users, 4 PM – 527 users, 5 PM – 543 users and at 6 PM – 533 users.

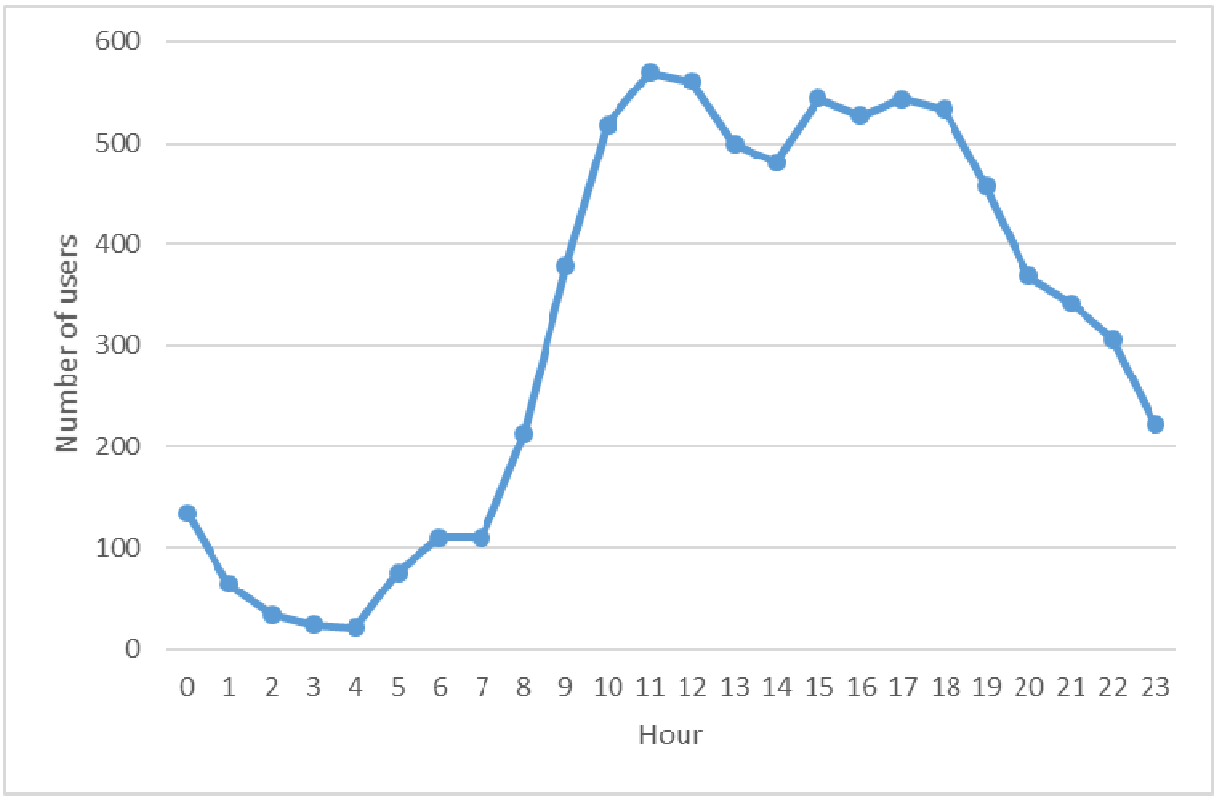

Figure. 4. Ratio between number of users and time3

With the cooperation with teachers from one of Italian University, we have collected user's actions on web pages from one of the courses, that uses LearnWeb platform, and scores, which receives users at the end of the course.

In total in course participated 40 students: five of them failed the course, 19 students did not participate in the exam, and 16 passed the course with the grade. To make an analysis of received data were chosen only those students, which passed and failed the course – 21 students.

To find and to analyze data from that selected users it was decided to group users by their scores. We have divided by the following groups:

- Users, which receive grade 0 – 5 users;
- Grades 20-23 – 3 users;
- Grade 24 – 4 users;
- Grades 25-26 – 4 users;
- Grades 27-29 – 5 users.

On order to build a relation between which score receive group of users, number of tracks, that were performed during the time of searching and hour we make a request to our database. We have done several requests regarding users, which were devided to several groups above.

On Fig. 5 is shown us the result of the request, where we want to find the correlation between number of tracks, which were performed by groups of users at specific period of time.

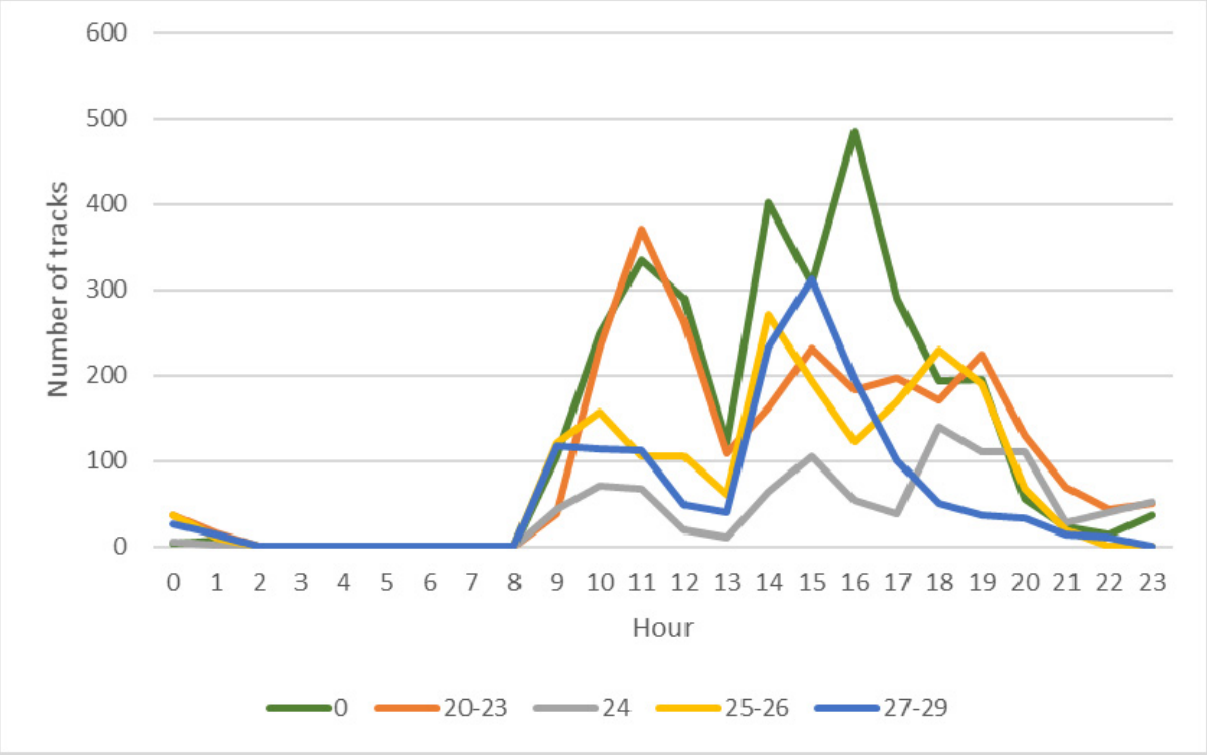

Figure. 5. Ratio between number of tracks and hour

We find out that nobody was doing exercises at night and early morning (from 2 AM to 8 AM). In addition, one interesting fact is that at 1 PM, we have a big recession data in all groups – many users do a break.

From retrieved data and from our charts, that we built to analyze data, we find, that users with greatest scores (from 27 to 29) spent less time for searching information, which they need to complete all tasks set by the teacher and pass the exam. On the same time users, which have not pass the exam or received minimum score spent more time and they search needed information in more number of sources.

Conclusion

In this paper conducted a user study to analyze movements, user activity at all and investigate behavior trends. Also, it was discussed detailed tracking of user interaction on web pages. Going beyond the usual application of tracking technologies for user tests, we have also looked at a large number of tracks and what users did: from general information of the track to analyzing how much time they spend and why.

We have introduced a transparent solution for tracking and analyzing user activity. In comparison to other approaches, our solution does not require manual preparation of web pages for tracking – our JavaScript code for tracking user activity will be added on each webpage, where proxy server will be used. But furthermore, every single user can switch on the Chrome extension and try out the developed proxy server and tracker by himself.

As for future work with the help of machine learning we can find which movements were performed on the web page: by teacher or by student. Furthermore, we can insert additional JavaScript code on each web page, which will track eye movements (does user looking for something else while we track, that user have not done any movements on the web page).

Acknowledgement

The research was done with the help of L3S Research Center, Hannover, Germany.

Bibliography

[McAfee, 2012] A. McAffee, E. Brynjolfsson, T. H. Davenport. Big data: the management revolution. Harvard business review, 2012. pp. 60-68. http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf

[Atterer, 2006] R. Atterer, M. Wnuk, A. Schmidt. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In Proceedings of the 15th international conference on World Wide Web, 2006. pp. 203-212. https://dl.acm.org/citation.cfm?id=1135811

[Tinati, 2014] R. Tinati, M. Luczak-Roesch, E. Simperl, N. Shadbolt. Motivations of citizen scientists: A quantitative investigation of forum participate. In Proceedings of the 2014 ACM conference on Web science, 2014. pp. 295-296. https://dl.acm.org/citation.cfm?id=2615651

Authors' Information

Prof. Dr.-hab. Oleksandr Kuzomin *– Informatics chair of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine; e-mail: kuzy@daad-alumni.de; tel.: +38(057)7021515*

Major Fields of Scientific Research: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.

Tetiana Tolmachova *– Master student in Informatics of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine; Master student in ITIS of Leibniz University; Hannover, Germany; e-mail: tetiana.tolmachova@gmail.com*

Major Fields of Scientific Research: Big Data, Data Mining, Data Analyses.

Oleh Astappiev *– Master student in Informatics of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine; Master student in ITIS of Leibniz University; Hannover, Germany; e-mail: astappev@gmail.com*

Major Fields of Scientific Research: Big Data, Data Mining, Web Ranking.

APPLYING THE HITS ALGORITHM ON WEB ARCHIVES

Oleksandr Kuzomin, Oleh Astappiev, Tetiana Tolmachova

Abstract: *The aim of the work is to study of ranking algorithms on data from Web Archives. A developed software use web archives and extracted data to create a graph of relations between web pages. A HITS algorithm applied on the graph, allows to find meaningful hubs among the pages. The hubs allow calculating authority of each page and ranking them by using the value.*

Link pairs from German Web Archive Data are used as input data for graph (incoming and outgoing links), and German Wikipedia articles as search topics to evaluate results.

The results evaluated using survey with results of HITS algorithm in comparison with results of Bing search engine and competing algorithm PageRank.

The work use Java programming language for implemented algorithms and support software, the destination graph stored in-memory by using Redis. The data extracted from Web Archives by using Hadoop framework and stored in Hive database.

Keywords: *web archives, web search, ranking, hyperlink induced topic search, page rank.*

ITHEA Keywords: *E.2 Data Storage Representations, H.3.1 Content Analysis and Indexing.*

Introduction

Web Archives contains web pages from the past years and save new pages for future researchers, historians, and the public. They are very important for learning how interned is developing. They allow to use knowledge from the past and apply it today to extract new knowledge and very popular for hundreds of research tasks as playground.

Search is the most demanded tool in today's web. There is a trillions of pages stored somewhere in the web and search engines are like road signs before the navigator was developed. By typing simple search term, they navigate to destination pages.

A web search engine is a software system that is designed to search for information on the World Wide Web. The common approach for web search engines is to analyze content of the pages in the web and create index for them. Then, using that index and different retrieval methods, they select all pages that

match a required search term and to return relevant results they apply ranking algorithms to order results.

All the same is true for web archives, but in addition, they also contain some specific attributes, which include crawling information and a history of the page. That can be used to improve search results quality.

The algorithms are improving with years and modern search engines use very advanced technology with Artificial Intelligence (AI) and machine learning. But unfortunately, those high-end algorithms only available for lead companies that specified on search engines.

Therefore, the goal of the work is to apply available ranking algorithms on Web Archives and investigate potential ability to improve search results by including new attributes to the algorithm.

Motivation and problem statement

Improving search results is interesting and demanded task, which can be applied on Web Archives for research.

In the work, one of the popular algorithms for ranking search results is applied on German Web Archives and learned how archived data can change search results.

Despite the fact that for the web search simple algorithms are no longer enough, nevertheless they are increasingly used for research, personal and low-middle level commercial purposes. They should have good enough results and lower maintenance costs.

As a problem statement we have got web archives that contains web pages (articles, news, etc.), which stored as natural language text (unstructured text). With the purpose of building an efficient search results, we are confronted with following problems:

1. Different formats of URLs and aliases.
3. Using diacritic characters and language specific characters in URLs.
2. A huge amount of data, which should be processed in real-time.
4. Slow processing time for Web Archives.

Topic research

Recent development of the Internet and computing technologies makes the amount of information increasing rapidly. That is why it is necessary to retrieve the best of the web pages that are more relevant in terms of information for the query entered by the user in search engine. In recent years,

semantic search for relevant documents on web has been an important topic of research. Many semantic web search engines have been developed that helps in searching meaningful documents presented on semantic web. To relate entities, texts and documents having same meaning, semantic similarity approach is used based on matching of the keywords, which are extracted from the documents.

For example, groups of authors presented a new web ranking system by using Semantic Similarity and HITS algorithm along with AI technique [Bansal, 2014]. In this paper, author proposed Intelligent Search Method (ISM) - a ranking system with improved HITS and Semantic Similarity techniques. It is used to rates the web pages and also known as Hubs and Authorities. A good hub represented a page that pointed to many other pages and a good authority represented a page that was linked by many different hubs. Therefore, its authority value, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages.

Author developed new method to index the web pages using an intelligent search strategy in which meaning of the search query is interpreted and then indexed the web pages based on the interpretation. Comparison of HITS Algorithm, Semantic Similarity Algorithm and ISM method is shown in Table 1.

Table 1. Comparison of Techniques

Parameter / Technique	HITS Algorithm	Semantic Similarity Algorithm	Proposed System
Time Efficiency	72%	87%	91%
Accuracy	79%	91%	95%
User specific Page Generation	No	No	Yes
Relevance Ratio	90%	92%	96%
High Relevance Ratio	30%	41%	51%

New ISM method can be integrated with any of the Page Ranking Algorithms to produce better and relevant search results.

Preparing dataset

On the input of the task, we have Hive table with two columns: source_url and desctination_url. Each row contains one edge of future graph (see Figure 1 for example of DB row).

distinct_a_links.source_url	distinct_a_links.destination_url
de,0-kongress,web2)/?tag=trends	de,0-kongress,web2)/?tag=marktforschung-2-0
de,0-kongress,web2)/?tag=verbraucherkommunikation	de,0-kongress,web2)/?tag=mitarbeiterweblogs
de,0-kongress,web2)/author/buettner	de,0-kongress,web2)/tag/gehalt
de,0-kongress,web2)/kongress-2011	de,0-kongress,web2)/kongress-2011/programm-als-pdf
de,0-kongress,web2)/kongress-2011	de,webstandards-magazin)/
de,0-kongress,web2)/tag/appstore-business-mashups	de,0-kongress,web2)/tag/kongress-2011
de,0-kongress,web2)/tag/gehalt	de,0-kongress,web2)/presse/kongress-2011/sponsor
de,0-kongress,web2)/tag/relevanz	ly,bit)/ehqfg3
de,0-kongress,web2)/tag/ruckblick	de,0-kongress,web2)/?page_id=1130
de,0-kongress,web2)/tag/web-tv	com,twitter,search)/search?q=%23webcenter
de,0-kongress,web2)/tag/web-tv	de,0-kongress,web2)/sponsoring-und-ausstellung
de,0-kongress,web2)/tag/wertewandel	com,twitter)/4punkt0media
de,0-n)/	de,0-n)/2012/11

10 rows selected (1.134 seconds)

Figure 1. Sample results from distinct_a_links table

As the size of the graph is very big (more than 130 billion edges, about 10 terabyte size), first we need to optimize it. The first step was to extract all unique URLs from the table and replace them with IDs, then we will have much smaller graph where edges will be long to long instead of string to string.

We extracted about 6 billion unique links (see Figure 2 for sample data), to rewrite out initial table from string to string to use short IDs of the pages we need very fast access to database which contains those IDs, therefore we decided to use in-memory databases.

links_distinct_by_source.url
de,007box,gaestebuch)/index.php?gbname=gb19882
de,1000dokumente)/index.html?c=glossar_de&l=de&viewmode=1
de,1000ferienwohnungen)/deutschland/friedrichshafen-ferienwohnung-ferienhaus.html
de,1000ferienwohnungen)/lecce-ferienwohnungen-ferienhaeuser.html
de,123tequila)/tequila-arette-anejo-p-397.html
de,12gebrauchtwagen)/hyundai/ix55
de,12travel)/ie/packages/self-drive/southern_costal1.html
de,1stplan,hilfe)/istdaten-erfassen/bilanz/passiva/summe-verbindlichkeiten.html
de,1und1)/
de,1und1,hilfe-center)/article/787345

10 rows selected (0.257 seconds)

Figure 2. Sample results from distinct_links_all table

However, 6 billion URLs still was a lot to fit in-memory (about 4 terabyte), we decided to store SHA-1 hash instead of URLs as key and ID long value as value for our Redis database.

During that work, we noticed that some of URLs has different formats, for example, some of them use http and some use https, some use www prefix and some not. We decided to remove such difference and added pre-processing of URLs which include:

1. Removing all unsafe ASCII characters, if they appears in domain names we replace them into Punycode domains and if such characters appears in path then replace them by using "%" followed by two hexadecimal digits;
2. Removing protocol prefixes, like: http, https;
3. Removing www prefixes;
4. Removing port numbers, like: 80, 443;
5. Decoding URLs from SURT format.

After replacing URLs to hashes, it has 21 153 collisions on our dataset. The hashes was extracted for investigation.

Table 2. Hash collisions over links dataset

Hash value	Original URL	Normalized URL
qEb3qCpvWQthRNLdkKTOiHInVmg=	de,merkspruch)/	merkspruch.de/
qEb3qCpvWQthRNLdkKTOiHInVmg=	de,merkspruch,)/	merkspruch.de/
ZLPnZaYBNeerff/5PH5ip3XXq40=	de,wuhletal,kirche)/	kirche.wuhletal.de/
ZLPnZaYBNeerff/5PH5ip3XXq40=	de,wuhletal,http://kirche)/	kirche.wuhletal.de/

As we can see in table above, that is false negative results. In the middle column you can see URL as it stored in database, right column as it was normalized and left column is hash of the URL. As we can see, after normalization, we have same strings and as result same hashes. Also original URLs from database are not valid SURT format.

Now when assigned short ID (long value) to each URL we need to update our whole dataset of pairs. That should significantly reduce size of it. Also, to create a graph we will need to go over all the pairs again. To decrease number of operations, we created module that read pairs, normalize URL and retrieve its ID from Redis database from previous task.

To create graph, we need to know incoming and outgoing links from each link. Unfortunately, hash table, which we used for previous task can't store multiple values per single key. Therefore we used two new databases, one which store linkID and list of incoming linkIDs and other one with linkID and list of outgoing linkIDs.

The data stored in Hive database is unsorted, but to decrease number of operation on Redis server we need to order the database by source_url or destination_url depending of which table we want to fill. If the pairs are ordered for example by source_url, then during going through them we can collect all neighbors for same source_url (just compare if previous value is equal current) and merge destination_url values, put them to Redis in single operation.

As we already can access to incoming and outgoing links for any page, we can calculate hubs and authorities, which requires for HITS algorithm.

HITS ranking results

To evaluate HITS algorithm we use search results from Bing search engine. We have pre-saved results for all German Wikipedia articles. We selected most popular 3000 pages and used their title as search term. For each search term, we have about 100 search results from Bing.

For our HITS algorithm we use 100 pages from Bing as root set. Then we use base set of pages and all pages which linked or links to them as base set. For that base set we calculate authority and hubs values for thee steps. Then order pages from root set by authority and compare results with original Bing results.

We also implemented PageRank algorithm for better evaluation of HITS results. It will add additional set of results to compare. The implementation of PR algorithm is very simple, we use 10 as default score for each page and split the score between all outgoing pages.

There are some limitations associated with Archived Data, for example the actives mainly contains German Internet pages (in .de domain zone), when Bing provides results regardless of the domain zone. But, we still have such pages in results because we have German pages that have outgoing links to different domain zones and we can calculate authority value for them.

The example results of applying HITS algorithm on search term "Kassel" you can see in the Table 3, the comparison with PageRank score displays in Table 4.

Table 3. Authority and hub values of HITS for search term "Kassel"

Link	Authority	Hub
Kassel Marketing | Tourismus-Informationen für Kassel kassel-marketing.de/	58929	148127497
Wetter Kassel - aktuelle Wettervorhersage wetteronline.de/wetter/kassel	46581	114631041
KSV Hessen Kassel e.V. - Die offizielle Homepage dasbesteausnordhessen.de/	45712	125635663
Kassel: Information für Kassel bei meinestadt.de home.meinestadt.de/kassel-documenta-stadt	45666	125741104
Stadtportal - Startseite www.kassel.de kassel.de/	45666	125741104

Table 4. HITS Authority and PageRank score

Link	HITS Authority	PageRank Score
Kassel Marketing | Tourismus-Informationen für Kassel kassel-marketing.de/	58929	0.09894597
Wetter Kassel - aktuelle Wettervorhersage wetteronline.de/wetter/kassel	46581	0.027334956
KSV Hessen Kassel e.V. - Die offizielle Homepage dasbesteausnordhessen.de/	45712	0.023083081
Kassel: Information für Kassel bei meinestadt.de home.meinestadt.de/kassel-documenta-stadt	45666	0.025142923
Stadtportal - Startseite www.kassel.de kassel.de/	45666	0.025142923

As we can see in Table 4, the website of Kassel's football team has lower PageRank score. That can be explained that in total the website has smaller number of incoming links, but the links is from better sources.

Some other results that illustrate the difference between HITS and PR are displayed in Table 5. Also the results compared to Bing results in table 6.

Table 5. Results of HITS and PR for search term "Volkswagen AG"

Link	HITS Authority	PageRank Score
Volkswagen AG - Home - SSI SCHÄFER https://www.ssi-schaefer.com/de-de	184900	9.485999 (1)
VOLKSWAGEN AKTIEN News | 766403 Nachrichten… http://www.finanznachrichten.de/nachrichten-aktien/vo...	7462	0.10863955 (5)
Volkswagen Aktie | Aktienkurs | Chart | 766400 wallstreet-online.de/aktien/volkswagen-aktie	6456	0.025147859 (11)
Volkswagen Konzern Startseite volkswagenag.com/	2002	0.82785743 (2)
Volkswagen Personal volkswagen-karriere.de/de.html	1898	0.30051792 (3)

Table 6. Results of HITS and PR for search term "Volkswagen AG"

HITS results	Bing results
Volkswagen AG - Home - SSI SCHÄFER https://www.ssi-schaefer.com/de-de	Volkswagen Konzern Startseite volkswagenag.com/
VOLKSWAGEN AKTIEN News | 766403 Na… http://www.finanznachrichten.de/nachrichten-a...	Wie gut klingt das denn. volkswagen.de/de.html
Volkswagen Aktie | Aktienkurs | Chart | 766400 wallstreet-online.de/aktien/volkswagen-aktie	Volkswagen AG – Wikipedia de.wikipedia.org/wiki/Volkswagen_AG
Volkswagen Konzern Startseite volkswagenag.com/	Volkswagen Group Homepage volkswagenag.com/content/vwcorp/con…
Volkswagen Personal volkswagen-karriere.de/de.html	Volkswagen International de.volkswagen.com/de.html

Evaluating results

To evaluate results we created survey page, which contains ten results from Bing, and ten reordered results by using authority value of HITS algorithm. Also ten results of HITS with ten results of PageRank algorithm.

The survey asks users to compare results which is more relative, as they think and make decision by clicking on one of the submit buttons on the bottom.

Survey is available online for everyone, we asked some students to participate in and 31 people accept proposal. In average one-person answers for 50 topics, and 1541 in total.

The results approximate expectations and Bing search engine provides better results than re-ranked results by using HITS algorithm. The results of that survey illustrated on Figure 3.

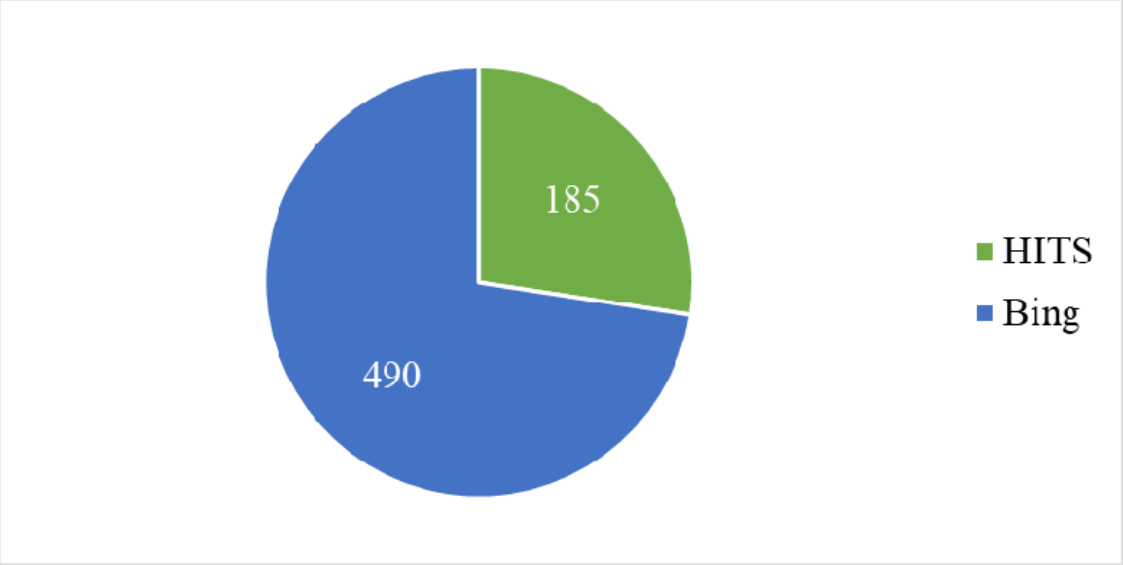

Figure 3. The results of survey HITS vs Bing results

Comparing results of HITS algorithm and PageRank algorithm (see Figure 4) give a little more points in favor of HITS algorithm.

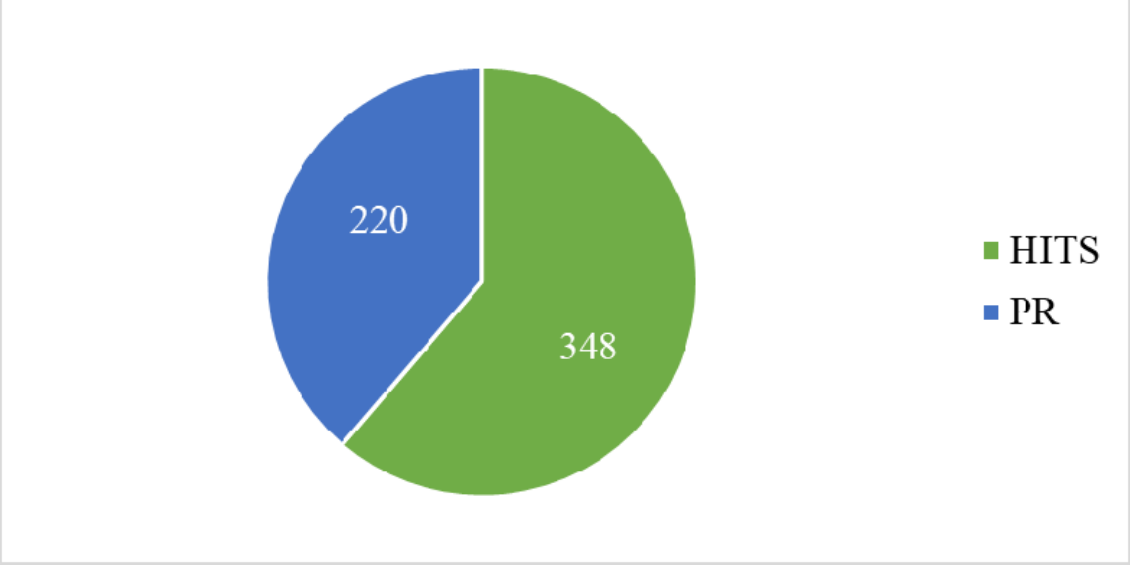

Figure 4. The results of survey HITS vs PR results

Comparing results of PageRank vs Bing (see Figure 5), gives most of the points to Bing results. It should be noted that in comparison with Bing results, HITS results take a bit more points than PageRank results.

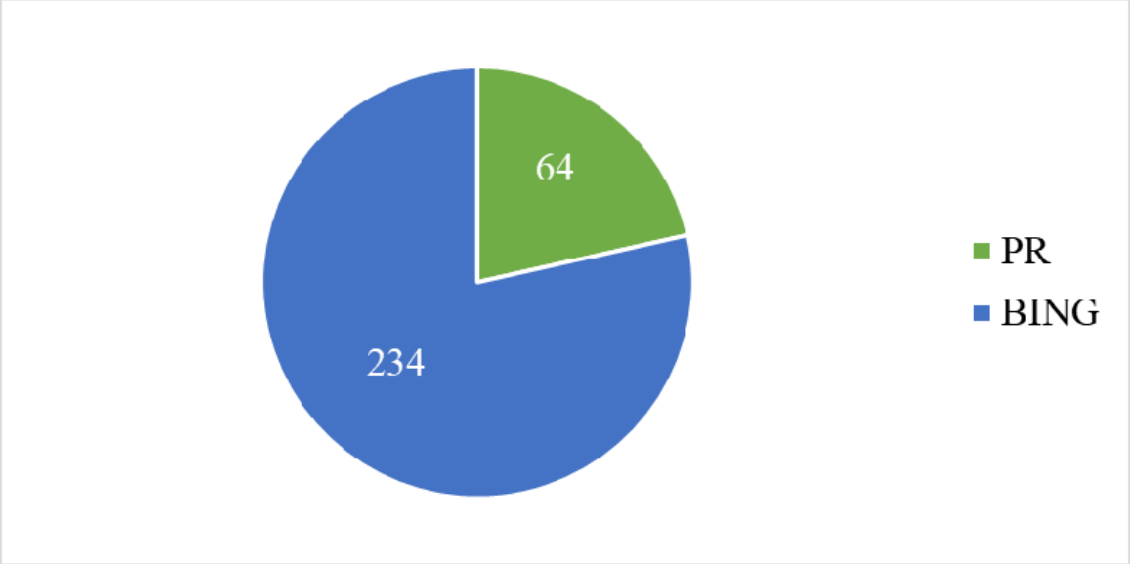

Figure 5. The results of survey PR vs Bing results

As we can see on table 6 the results of HITS and PageRank results worst results than Bing, six of ten top results contains pages with information about company shares. That pages appears on top due to specific content of the Web Archives that was used. For some other research purposes that archives contains a lot of pages with information about trades and shares.

In addition to that, some other results also contains very specific pages only for used Web Archives. Using a simple survey among unprepared users was not the best way to evaluate the quality of the results.

Conclusion

The aim of the work is to begin research in the direction and show some first results.

During the work was implemented two algorithms for ranking web results, the initial HITS algorithm was compared with PageRank algorithm.

Our results confirmed that modern search engines use very sophisticated technologies that include not only ranking algorithms, they also use AI and machine learning techniques to improve our daily Internet search experience.

Nevertheless, HITS algorithm that was developed slightly later than PageRank and using more depth scanning gives relatively better results. And that also gives us motivation to continue our work, we have plans to improve the results.

The amount of data that we have on input is very huge, and several first tries ware failed. We were need to experiment with different techniques and technologies to work with given data. Moreover, even now, when we have prepared relations graph, each iteration in the program must be justified, otherwise, everything works very slowly.

The search in Web Archives is not for everyday use and we do not expected that results will completely satisfy us. The key idea the work is research of additional attributes of web archives, unfortunately, we do not have enough time to present them in the work.

In the work, we finished only first part of our goal. Right now, we implemented simple HITS and PageRank algorithms, they allow us to make some small researches over retrieved data.

The next step will be to include first crawl date and last crawl date into HITS algorithm. Potentially we can find some hubs that existed before, but no longer exists today. By using those properties, we also can know age of the pages and we do not know how it change results.

Acknowledgement

The research was done with the help of L3S Research Center, Hannover, Germany.

Bibliography

[Bansal, 2014] N. Bansal, S. Paramjeet. Improved Web Page Ranking Algorithm Using Semantic Similarity and HITS Algorithm. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2014. pp. 346-348. http://www.ijettcs.org/Volume3Issue4/IJETTCS-2014-08-26-146.pdf

[Sharnagat, 2014] R. Sharnagat. Named Entity Recognition: A Literature Survey, 2014. 27 p. https://pdfs.semanticscholar.org/83fd/67f0c9e8e909dc7b90025e64bde0385a9a3a.pdf

[Ridings, 2002] C. Ridings, M. Shishigin. Pagerank Uncovered, Technical report, 2002. 56 p. http://www.voelspriet2.nl/PageRank.pdf

[Miller, 2001] J. C. Miller, G. Rae, F. Schaefer, L.A. Ward, T. LoFaro, & A. Farahat. Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, 2001. pp. 444-445. https://dl.acm.org/citation.cfm?id=384086

Authors' Information

Prof. Dr.-hab. Oleksandr Kuzomin *– Informatics chair of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine; e-mail: kuzy@daad-alumni.de*
tel.: +38(057)7021515
Major Fields of Scientific Research: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.

Oleh Astappiev – Master student in Informatics of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine; Master student in ITIS of Leibniz University; Hannover, Germany; e-mail: astappev@gmail.com
Major Fields of Scientific Research: Big Data, Data Mining, Web Ranking.

Tetiana Tolmachova *– Master student in Informatics of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine; Master student in ITIS of Leibniz University; Hannover, Germany; e-mail: tetiana.tolmachova@gmail.com*
Major Fields of Scientific Research: Big Data, Data Mining, Data Analyses.

SOME QUALITY CHARACTERISTICS AND METRICES IN OVERALL TELECOMMUNICATION NETWORKS

Emiliya Saranova, Stoyan Poryazov

Abstract: More precise definitions of some ITU-T traffic concepts are used. On their base, three new more precise QoS traffic indicators are proposed and used for QoS characterization of a services delivered by a pool of resources (service phase concept). The quality of composition of two connected (sequentially or in parallel) service phases is presented as aggregation function of the qualities of the phases. Four different metrics are proposed. The results allow more precise prediction of service composition quality as function of sub-services quality. The approach is applicable for the overall telecommunication network QoS estimation.

ITHEA Keywords*: Service composition, Causal aggregation, Quality of service composition, Quality metrics*

Introduction

The composition of informational services (especially in the internet) is an important topic from many years [Stegaru et al. 2012]. In the survey [Kondratyeva et al 2013] the quality of a composition as a function of the composed services' qualities is considered. Most of the metrics, for quality of composition estimation, use weighting coefficients proposed from the Network Administration. The objective metrics, based on direct analysis of flow, time and traffic parameters are rare. At the present, the state of the art is not very forwarded, in this direction [Otsetova, A., Saranova, 2017].

In this paper, the approach developed from authors in [Poryazov et al 2018a] and [Poryazov et al 2018b] is advanced for estimation of quality of service composition as an aggregation function of qualities of the service components (sub-services). More precise definitions of some ITU-T traffic concepts are used. On their base, three new more precise QoS traffic indicators are proposed and used for QoS characterization of a services delivered by a pool of resources (service phase concept). The quality of composition of two connected (sequentially or in parallel) service phases is presented as aggregation function of the qualities of the phases. Four different metrics are proposed. The results

allow more precise prediction of service composition quality as function of sub-services quality. The approach is applicable for the overall telecommunication network QoS estimation.

Basic virtual devices

At the bottom of the structural model presentation, we consider “basic virtual devices” that do not contain any other virtual devices. Basic virtual devices, used in this paper, have graphic representations as shown in Figure 1.

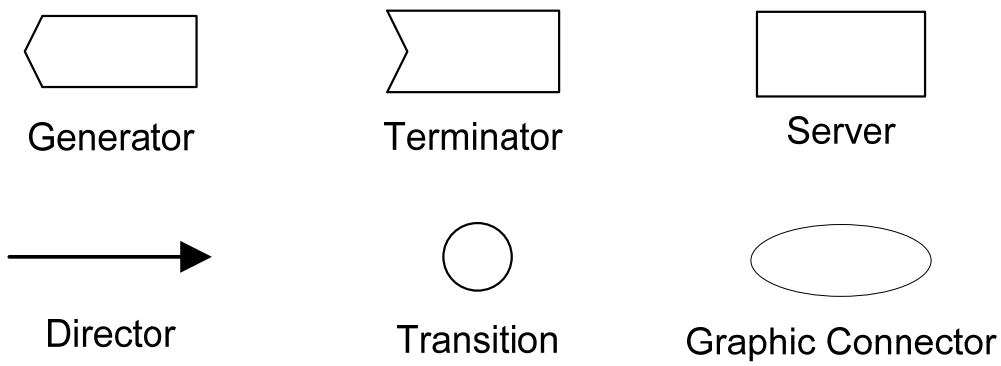

Figure 1. A graphical representation of the main basic virtual mono-functional devices used.

- *Generator* – this device generates calls (requests, transactions);
- *Terminator* – this block eliminates every request entered (so it leaves the model without any traces);
- *Server* – this device models the delay (the time duration of a service, the holding time) of requests in the corresponding device without their generation or elimination. It models also traffic- and time characteristics of the requests processing (c.f. Figure 2);
- *Director* – this device unconditionally points to the next device, which the request shall enter, but without a transfer or delay of it;
- *Transition* – this selects one of its possible outputs for each request entered, thus determining the next device where it shall go to;
- *Graphic Connector* – this is used to simplify the graphical representation of the conceptual model structure. It has no modeling functions.

The flows and main parameters of a basic virtual device (e.g. server) have the graphic representation as shown in Figure 2.

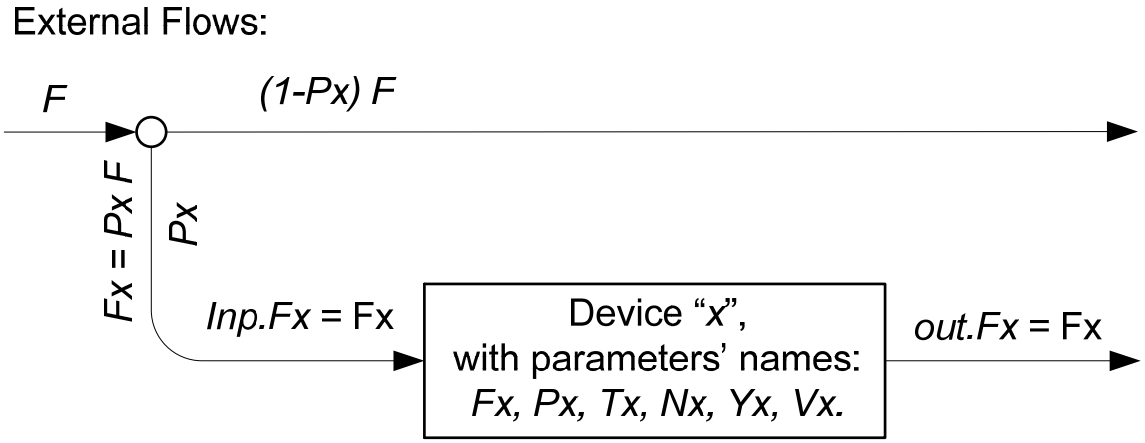

Figure 2. A graphical representation of a basic virtual device ***x*** and its main parameters. In stationary state $inp.Fx = out.Fx = Fx$.

Parameters of the basic virtual device x are the following (c.f. [ITU-T E.600, 1993] for terms definition):

Fx = Frequency (intensity or incoming rate) of the flow of requests (requests per time unit) to device x;

Px = Probability of direction of the requests towards device x;

Tx = Time duration of servicing of a request in by device x;

Nx = Number of lines (service places, positions, capacity) of device x;

Yx = Traffic intensity [Erlang];

Vx = Traffic volume [Erlang - time unit].

Stationary state

By assumption, every device, in the models, in this paper, is in a stationary state. So one may apply the Little's theorem [Little 1961] and for each device:

$$Yx = Fx\,Tx \tag{1}$$

In stationary state, input $(inp.Fx)$ and output $(out.Fx)$ flow intensities of a server coincide, because requests are not generated or terminated in a server. Obviously:

$$inp.Fx = out.Fx = Fx \tag{2}$$

Parameters' Qualification

In Fig. 2, one may see notations $inp.Fx$ and $out.Fx$. Traffic qualification is necessary and it is used in [ITU-T E.600, 1993], but without any attempt for including the qualifiers in the parameters' names. Since [Poryazov, Saranova 2006] we use up to two qualifiers as a part of the parameter's name. In this paper we use 'rep.' for 'repeated', 'ofr.' for 'offered', "srv"for 'served', "crr" for 'carried'etc. We expand the meaning of the traffic qualifiers to the other parameters determining the traffic, e.g. in our notations, $ofr.Ys = ofr.Fs \ srv.Ts$ means: 'the offered traffic intensity to the switching system is a product of the offered requests' frequency (rate) and the service time duration in the switching system'.

Service phase concept. Causal normalization

Based on the ITU-T definition of a service, provided in [ITU -T.E.800] (Term 2.14), i.e. "A set of functions offered to a user by an organization constitutes a service", we have proposed [Poryazov et al 2018b] the following definition of a service phase.

Definition 1: The Service Phase is a service presentation containing:

- One of the functions, realizing the service, which is considered indivisible;
- All modeled reasons for ending/finishing this function, i.e. the causal structure of the function;
- Hypothetic characteristics, related to the causal structure of the function (a well-known example of a hypothetic characteristic is the offered traffic concept).

Note that a service phase corresponds to service delivered by one pool of resources. We may present the service phase in device s by means of k+1 basic virtual causal devices, each representing a different reason for ending this service phase (Figure 3).

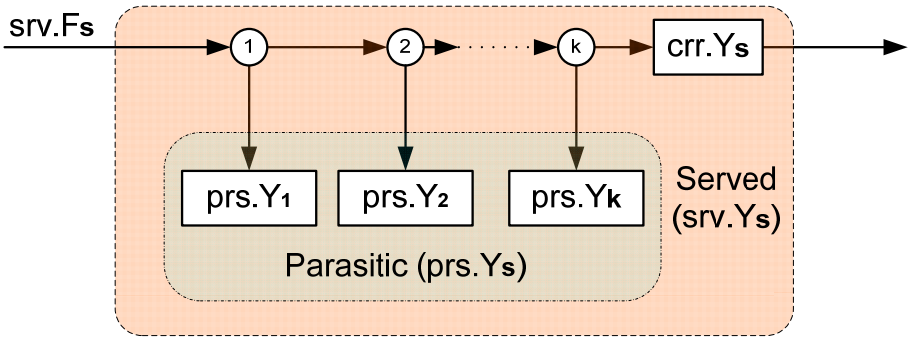

Figure 3. Traffic characterization of a service phase, represented as device ***s***, by means of ***k+1*** basic virtual causal devices.

In Figure 3, only one causal device represents successful completion of the service in device *s* – with carried traffic (*crr.Ys)*, whereas the remaining causal devices represent *k* different reasons for unsuccessful ending of the service – respectively with traffics $prs.Y_1, prs.Y_2, ..., prs.Y_k$.

Generalizing, for more precise traffic characterization in a pool of resources, we propose the following definitions:

Definition 2: The Served Traffic in a pool of resources is the traffic, occupying (using) resources in the pool.

In Fig. 3, the served traffic in device *s* ($srv.Ys$) is the following sum:

$$srv.Ys = prs.Y1 + prs.Y2 + ... + prs.Yk + crr.Ys \quad . \qquad (3)$$

Definition 3: The Carried Traffic in a pool of resources is the traffic, which was successfully served in the pool (and carried to the next service phase).

In Figure 3, the carried traffic in device ***s*** is $crr.Ys$.

Definition 4: The Parasitic Traffic in a pool of resources is the traffic, which was unsuccessfully served in the pool.

In Figure 3, each of traffics $prs.Y_1, prs.Y_2, ..., prs.Y_k$ is a parasitic one. Parasitic traffic occupies real resources of the pool, but not for an effective service execution.

In Definitions 2 and 3, the served- and carried traffic are different terms, despite the ITU-T definition of the carried traffic as "The traffic served by a pool of resources" ([ITU-T E.600, 1993], Term 5.5). We believe that this distinction leads to a better and more detailed traffic- and QoS characterization.

Causal aggregation in a service phase

The causal aggregation is understood as an aggregation of all cases in the model, corresponding to different reasons for service ending (referred to as unsuccessful cases).

Here a causal generalization is proposed, as an aggregation of all successful ($crr.Ys$) and all unsuccessful cases ($prs.Ys$).

$$prs.Ys = \sum_{i=1}^{k} prs.Yi ; \tag{4}$$

By Definition 2, the served traffic is a sum of the parasitic and carried traffic (c.f. Figure 1):

$$srv.Ys = prs.Ys + crr.Ys ; \tag{5}$$

$$srv.Fs = prs.Fs + crr.Fs . \tag{6}$$

By using the Little's formula we have: $prs.Ys = prs.Fs\ prs.Ts$ and $crr.Ys = crr.Fs\ crr.Ts$. Hence:

$$srv.Ys = srv.Fs\ srv.Ts = prs.Fs\ prs.Ts + crr.Fs\ crr.Ts \tag{7}$$

The causal generalization of Figure 3 is presented in Figure 4:

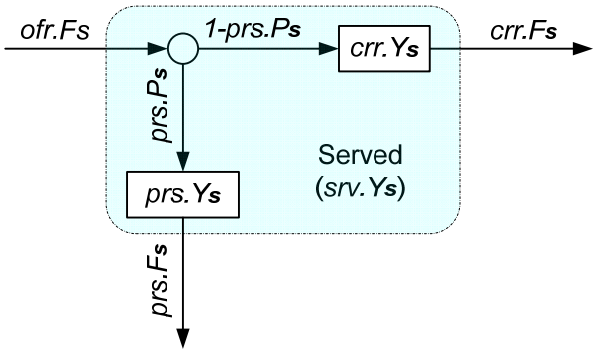

Figure 4. Causal generalized presentation of a service phase.

The causal generalized presentation of a service phase (Figure 4) consists of two virtual devices – one presents successful (carried) service and the other – unsuccessful (parasitic) service. $ofr.Fs$ is offered (incoming) intensity of requests; $crr.Fs$ and $prs.Fs$ are carried and parasitic respectively; $prs.Ps$ is the probability for parasitic service.

If we assume that $ofr.Fs$, $prs.Ps$, $prs.Tx$ and $crr.Fs$ are known, then directly from Figure 4 and Little's theorem, we receive:

$$prs.Fs = ofr.Fs\ \ prs.Ps \tag{8}$$

$$prs.Ys = prs.Fs\ \ prs.Ts = ofr.Fs\ prs.Ps\ \ prs.Ts \tag{9}$$

$$crr.Fs = ofr.Fs\ (1 - prs.Ps) \tag{10}$$

$$crr.Ys = crr.Fs\ crr.Ts = ofr.Fs\ (1 - prs.Ps)\ \ crr.Ts \tag{11}$$

$$srv.Fs = prs.Fs + crr.Fs = ofr.Fs \tag{12}$$

The service phase is considered as a device comprising carried and parasitic causal devices (c.f. Figure 4), so the served traffic intensity $srv.Yx$ is a sum of their traffics:

$$\begin{aligned} &srv.Ys = srv.Fs\ \ srv.Ts = prs.Ys + crr.Ys = \\ &ofr.Fs\,(prs.Ps\ \ prs.Ts + (1 - prs.Ps)\ crr.Ts) \end{aligned} \tag{13}$$

Therefore the mean service time of the phase ($srv.Ts$) is:

$$srv.Ts = prs.Ps\ \ prs.Ts + (1 - prs.Ps)\ crr.Ts \tag{14}$$

Quality factors and indicators of a service phase

There are many aspects of Quality of Service (QoS) [ITU -T.E.800], many factors influencing the QoS and corresponding indicators. We divide two types of factors' effects: degradative and terminative causing different effects. Degradative factors may cause degradation of the QoS, but the service is classified as successful. Terminative factors (i.e. interruption of the connection) cause termination of the service, which is classified as unsuccessful. A factor may be classified as degradative or terminative depending of its intensity, (i.e. noise, distortions, packet losses in speech communication, etc.).

In this paper we consider terminative effects only. Consequently, the quality indicators may be expressed in the concepts and terms of causal presentation of the services, considered in the previous two sections.

Following ITU-T approaches of traffic qualification and indicators definition ([ITU-T E.600, 1993] [ITU-T E.425]), and causal presentation of the services explained (c.f. Figure 4), we propose the following quality indicators:

Flow Quality (Q_f):

$$Q_f = \frac{carried\ requests'\ flow\ intensity}{served\ requests'\ flow\ intensity} = \frac{crr.F}{srv.F} \tag{15}$$

Traffic Quality (Q_y):

$$Q_y = \frac{carried\ traffic\ intensity}{served\ traffic\ intensity} = \frac{crr.Y}{srv.Y} \tag{16}$$

Time Quality (Q_t):

$$Q_t = \frac{total\ successful\ service\ time\ of\ requests}{total\ service\ time\ of\ requests} \tag{17}$$

The expression of the defined indicators using known parameters follows.

From (15), (10) and (12) follows

$$Q_f = \frac{crr.F}{srv.F} = \frac{ofr.F\ (1 - prs.P)}{ofr.F} = 1 - prs.P \tag{18}$$

From (16), (11) and (13) follows

$$Q_y = \frac{crr.Y}{srv.Y} = \frac{(1 - prs.P)\ crr.T}{prs.P\ \ prs.T + (1 - prs.P)\ crr.T} \tag{19}$$

Taking into account (14), from (19) follows:

$$Q_y = \frac{(1 - prs.P)\ crr.T}{srv.T} \tag{20}$$

Proposition 1: Numerically $Q_y = Q_t$ in a service phase.

Proof: From (17) and (20) follows:

$$Q_t = \frac{total\ successful\ service\ time\ of\ requests}{total\ service\ time\ of\ requests} = \frac{ofr.F\ (1 - prs.P)\ crr.T}{ofr.F\ srv.T} = Q_y \qquad (21)$$

Quality aggregation of two consecutive service phases

Let Device x comprises two consecutively connected service phases named 1 and 2 (Figure 5)

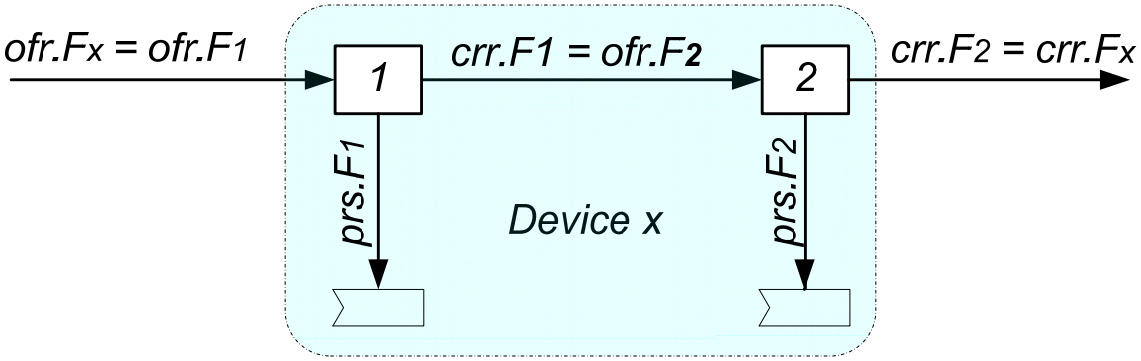

Figure 5. Two consecutively connected service phases named 1 and 2, and comprised in Device x .

We'll express the quality of the comprising Device x as function of its internal phases 1 and 2 (Figure 5). Obviously:

$$ofr.Fx = ofr.F1\ ;\ crr.F1 = ofr.F2\ ;\ crr.F2 = crr.Fx \qquad (22)$$

By definition (15), the flow qualities (Q_{fx} , Q_{f1} , Q_{f2}) of Device x and internal phases 1 and 2 are:

$$Q_{fx} = \frac{crr.F_x}{srv.F_x} \qquad (23)$$

$$Q_{f1} = \frac{crr.F_1}{srv.F_1} \qquad (24)$$

$$Q_{f2} = \frac{crr.F_2}{srv.F_2} \qquad (25)$$

Proposition 2: The flow quality Q_{fx} of the consecutively connected service phases is a multiplication of the flow qualities (Q_{f1} , Q_{f2}) of the phases:

$$Q_{fx} = Q_{f1} Q_{f2} \qquad (26)$$

Proof: From (12), (22) and (24):

$$crr.F_1 = ofr.F_1\ Q_{f1} \tag{27}$$

From (12), (22) and (25):

$$crr.F_2 = ofr.F_2\ Q_{f2} \tag{28}$$

From (22), (23), (27) and (28) follows

$$crr.F_x = crr.F_2 = ofr.F_2\ Q_{f2} = crr.F_1\ Q_{f2} = ofr.F_1\ Q_{f1} Q_{f2} = ofr.F_x Q_{f1} Q_{f2} \tag{29}$$

From (23) and (29) follows (26).

Proposition 3: The traffic quality (Q_{yx}) of the consecutively connected service phases is a weighted mean of the traffic qualities (Q_{y1}, Q_{y2}) of the phases:

$$Q_{yx} = Q_{y1}\ \alpha_1 + Q_{y2}\ \alpha_2 \tag{30}$$

where:

$$\alpha_1 = \frac{srv.Y_1}{srv.Y_x} \tag{31}$$

$$\alpha_2 = \frac{srv.Y_2}{srv.Y_x} \tag{32}$$

$$\alpha_1 + \alpha_2 = 1 \tag{33}$$

Proof: By definition (16), the traffic qualities (Q_{yx}, Q_{y1}, Q_{y2}) of Device x and internal phases 1 and 2 are:

$$Q_{yx} = \frac{crr.Y_x}{srv.Y_x} \tag{34}$$

$$Q_{y1} = \frac{crr.Y_1}{srv.Y_1} \tag{35}$$

$$Q_{y2} = \frac{crr.Y_2}{srv.Y_2} \tag{36}$$

In the scheme in Figure 5, every served (carried or parasitic) request in devices 1 and 2 is served in comprising device x , therefore:

$$crr.Y_x = crr.Y_1 + crr.Y_2 \tag{37}$$

$$srv.Y_x = srv.Y_1 + srv.Y_2 \tag{38}$$

From (34), (37) and (38) follows:

$$Q_{yx} = \frac{crr.Y_1 + crr.Y_2}{srv.Y_x} = \frac{crr.Y_1}{srv.Y_x} + \frac{crr.Y_2}{srv.Y_x} \tag{39}$$

From (39), (35) and (36) follows:

$$Q_{yx} = \frac{Q_{y1}\ srv.Y_1}{srv.Y_x} + \frac{Q_{y2}\ srv.Y_2}{srv.Y_x} = Q_{y1}\ \alpha_1 + Q_{y2}\ \alpha_2 \tag{40}$$

From (38) and (40) follows Proposition 3.

Proposition 4: Traffic quality metrics of the consecutively connected service phases depends of flow quality metrics.

Proof: Expressing (31) and (32) by the parameters of devices 1 and 2, and using (11), (13), (15), (24) and (25) we obtain:

$$srv.Y_1 = ofr.Fx\,[prs.P_1\ prs.T_1 + Q_{f1}\ crr.T_1] \tag{41}$$

$$srv.Y_2 = ofr.Fx\ Q_{f1}[prs.P_2\ prs.T_2 + Q_{f2}\ crr.T_2] \tag{42}$$

$$srv.Yx = ofr.Fx\,\{prs.P_1\ prs.T_1 + Q_{f1}[crr.T_1 + prs.P_2\ prs.T_2 + Q_{f2}\ crr.T_2]\} \tag{43}$$

Quality aggregation of two parallel service phases

Let two service phases are connected in parallel (Figure 6) and considered as a comprising Device x . An offered request may be served alternatively: with probability P_1 in Service Phase 1 or with probability P_2 - in Service Phase 2. Obviously:

$$P_1 + P_2 = 1 \tag{44}$$

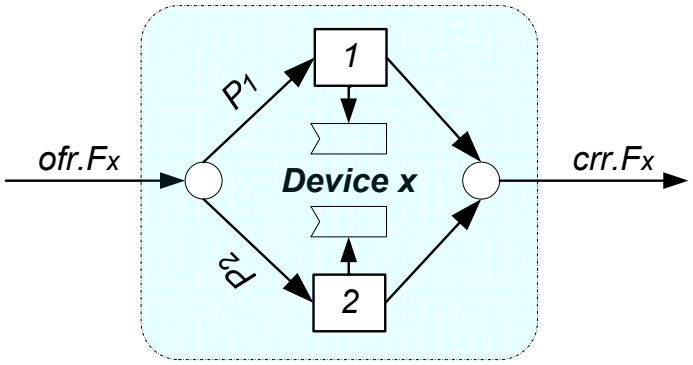

Figure 6. Two parallel connected service phases named 1 and 2, and comprised in Device x.

Proposition 5: The flow quality (Q_{fx}) of Device x depends from the flow qualities of two parallel connected service internal phases 1 and 2, (Q_{f1}, Q_{f2}) by the expression:

$$Q_{fx} = Q_{f1}\ P_1 + Q_{f2}\ P_2 \tag{45}$$

Proof: By definition (15):

$$Q_{fx} = \frac{crr.F_x}{ofr.F_x} \tag{46}$$

From Figure 6 follows:

$$crr.F_x = crr.F_1 + crr.F_2 \tag{47}$$

From (46):

$$crr.F_1 = Q_{f1}\ ofr.F_1 \tag{48}$$

$$ofr.F_1 = ofr.F_x\ P_1 \tag{49}$$

Therefore:

$$crr.F_1 = Q_{f1}\ crr.F_x\ P_1 \tag{50}$$

Analogously:

$$crr.F_2 = Q_{f2}\ crr.F_x\ P_2 \tag{51}$$

From (47), (50) and (51) follows:

$$crr.F_x = ofr.F_x\ (Q_{f1}\ P_1 + Q_{f2}\ P_2) \tag{52}$$

From (52) follows (46) – the Proposition 5.

Proposition 6: The traffic quality (Q_{yx}) of Device x depends from the flow qualities of two parallel connected internal service phases 1 and 2, (Q_{y1}, Q_{y2}) by the expression:

$$Q_{yx} = Q_{y1}\ \beta_1 + Q_{y2}\ \beta_2 \tag{53}$$

where:

$$\beta_1 = \frac{srv.Y_1}{srv.Y_x} = \frac{P_1\ srv.T_1}{srv.T_x} \tag{54}$$

$$\beta_2 = \frac{srv.Y_2}{srv.Y_x} = \frac{P_2\ srv.T_2}{srv.T_x} \tag{55}$$

$$\beta_1 + \beta_2 = 1 \tag{56}$$

Proof: By definition (16):

$$Q_{yx} = \frac{crr.Y_x}{srv.Y_x} \tag{57}$$

From (37) and (57) follows:

$$Q_{yx} = \frac{crr.Y_x}{srv.Y_x} = \frac{crr.Y_1 + crr.Y_2}{srv.Y_x} = \frac{crr.Y_1}{srv.Y_x} + \frac{crr.Y_2}{srv.Y_x} \tag{58}$$

From (58), (35) and (36):

$$Q_{yx} = \frac{crr.Y_1}{srv.Y_x}\ \frac{srv.Y_1}{srv.Y_1} + \frac{crr.Y_2}{srv.Y_x}\ \frac{srv.Y_2}{srv.Y_2} = Q_{y1}\ \frac{srv.Y_1}{srv.Y_x} + Q_{y2}\ \frac{srv.Y_2}{srv.Y_x} = Q_{y1}\ \beta_1 + Q_{y2}\ \beta_2 \tag{59}$$

Where, following (49) and analogous expression for $ofr.F_2$ follows that:

$$\beta_1 = \frac{srv.Y_1}{srv.Y_x} = \frac{P_1\ ofr.F_x\ srv.T_1}{ofr.F_x\ srv.T_x} = \frac{P_1\ srv.T_1}{srv.T_x} \tag{60}$$

$$\beta_2 = \frac{srv.Y_2}{srv.Y_x} = \frac{P_2\ ofr.F_x\ srv.T_2}{ofr.F_x\ srv.T_x} = \frac{P_2\ srv.T_2}{srv.T_x} \tag{61}$$

Conclusion

The four metrics defined (26), (30), (45) and (53) are useful for QoS estimation of different parts of the telecommunication network as well as of the overall network (considered as consists of aggregated served and carried traffics).

For further work, estimation of composed QoS degradative factors is very important, because it is not investigated enough.

Acknowledgement

This work is coordinated under EU COST Action IC 1304 entitled "*Autonomous Control for a Reliable Internet of Services*" (ACROSS) and is financed by Bulgarian NSF Project DCOST 01/20.

Bibliography

[ITU -T.E.800] ITU-T ITU-T Recommendation E.800 (09/08), Definitions of terms related to quality of service.

[ITU-T E.425] ITU-T Rec. E.425 (03/2002). Network management – Internal automatic observations.

[ITU-T E.600, 1993] ITU-T Recommendation E.600 (03/93), Terms and definitions of traffic engineering.

[Kondratyeva et al 2013] Olga Kondratyeva, Ana Cavalli, Natalia Kushik, Nina Yevtushenko. Evaluating Quality of Web Services: a Short Survey. 2013 IEEE 20th International Conference on Web Services. pp.587-594, DOI: 10.1109/ICWS.2013.83, https://www.researchgate.net/publication/261317755

[Little 1961] Little J. D. C., 1961. A Proof of the Queueing Formula L=λW. Operations Research, 9, 1961, 383-387.

[Poryazov et al 2018a] Poryazov, S., Saranova E., Ganchev, I.. Conceptual and Analytical Models for Predicting the Quality of Service of Overall Telecommunication Systems. In: Ivan Ganchev, Rob van der Mai, J.L. (Hans) van den Berg (Editors). Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer, LNCS, 2018, pp. 27.(in printing)

[Poryazov et al 2018b] Poryazov, S., Saranova E., Ganchev, I.. Scalable QoS Indicators towards Overall Telecom System QoE Management. In: Ivan Ganchev, Rob van der Mai, J.L. (Hans) van den Berg (Editors). Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer, LNCS, 2018, pp. 26. .(in printing)

[Poryazov, Saranova 2006] S. A. Poryazov, E. T. Saranova. Some General Terminal and Network Teletraffic Equations in Virtual Circuit Switching Systems. Chapter 24 in: A.Nejat Ince, Ercan Topuz (Editors) Modeling and Simulation Tools for Emerging Telecommunications Networks. Springer Sciences+Business Media, LLC, USA 2006, pp. 471-505. ISBN-10: 0-387-32921-8 (HB).

[Stegaru et al. 2012] Georgiana Stegaru, Cristian Danila, Ioan Sacala, Mihnea Moisescu, Aurelian Stanescu. Quality Driven Web Service Composition Modeling Fra
mework. Luis M. Camarinha-Matos; Lai Xu; Hamideh Afsarmanesh. 13th Working Confeence on Virtual Enterpries (PROVE), Oct 2012, Bournemouth, United Kingdom. Springer, IFIP Advances in Information and Communication Technology, AICT-380, pp.87-95, 2012, Collaborative Networks in the Internet of Services. <10.1007/978-3-642-32775-9_9>. <hal-01520449>

[Otsetova, A., Saranova, 2017] Otsetova, A., Saranova E.. Quality of a composite service as a function of the qualities of the comprised sub-services. International Journal "Information Technologies & Knowledge", Volume 11, Number 2, 2017, ISSN 1313-0455 (printed) ISSN 1313-048X (online), 103-112.

Authors' Information

Emiliya Saranova *– Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Block 8, 1113 Sofia, Bulgaria]*

University of Telecommunications and Post, Sofia, 1 Acad. St. Mladenov Str, Sofia 1700, Bulgaria, E-mail: e.saranova@utp.bg

Major Fields of Scientific Research: Information modeling, General theoretical information research, Multi-dimensional information system

Stoyan Poryazov – *Bulgarian Academy of Science, Institute of Mathematics and Informatics, Sofia, Bulgaria, stoyan@cc.bas.bg*

Major Fields of Scientific Research: Informational modeling, Quality of services, Quality of experience, Overall telecommunication network

REQUIREMENT ANALYSIS OF USER INTERFACE COMPONENTS FRAMEWORK FOR MOBILE DEVICES

Yurii Milovidov

Abstract. *The article suggests approaches to simplify the design of a user-friendly and modern graphical user interface for mobile devices. Comparative analysis of mobile application development environments is represented. The selection of the development environment for creating the user interface is justified. Conclusions are made regarding the choice of the Unity environment, which supports the compilation of applications for various platforms, as well as the developer-friendly mechanism for creating composite controls.*

Key words: *graphical user interface, framework, mobile development, cross-platform development, user control.*

ITHEA keywords: *D.2 Software engineering, D.2.1 Requirement/Specification, D.2.13 Reusable Software.*

Introduction

Designing convenient and effective graphical user interface (GUI) is one of the most crucial aspects of mobile application development. It is a precondition for creating various GUI components for simplifying of user interface planning. But limitations of development environment may cause extra difficulties while implementing the GUI. There are three most common ways to surpass such difficulties: creating self-developed solutions, paying for ready-to-use non-free libraries or looking for some options on the marketplace.

In any case, stakeholders should set up the process of reusing GUI libraries.

Preparing new GUI libraries is performed by combining default user interface (UI) components into certain composite objects. Such objects, possessing needed functionality, are close to what developer needs in the majority of cases. Such approach is time-consuming - developer has to join components and set up all the interactions and visual parts. So, it is a proper solution to create a library of assets that are reusable and cover common cases and demands.

This grounds the actuality of the problem of implementing a library for designing composite user elements in game development engines. This article is devoted to solving this task for Unity3D environment.

Other widespread effective requirement analysis techniques are based on Model-Driven Development Approach using visual modeling languages as UML [Chebanyuk, 2014a], [Chebanyuk, 2014b], [Chebanyuk and Shestakov, 2017].

Related work

While development frameworks for mobile devices provide rich support for sophisticated input mechanisms like gestures, etc., they lack support for graphical editors. In the paper [Buchmann and Pezoldt, 2014] authors present a lightweight framework for graphical editors which empower the user to easily build touch-enable graphical editors for android devices.

The development and maintenance of mobile applications for multiple platforms is expensive. One approach to reducing this cost is model-driven engineering. In the paper [Jia and Jones, 2012] authors present a novel model-driven approach to cross-platform mobile application development using a Domain Specific Language (DSL), called AXIOM (Agile eXecutable and Incremental Object-oriented Modeling). This approach could significantly reduce the development cost and increase the product quality of mobile applications.

Unity is a feature-rich, fully-integrated development engine that provides out-of-the-box functionality for the creation of interactive 3D content. The book [Thorn, 2015] shares extensive and useful insights to create animations using a professional grade workflow.

The book [Smith and Queiroz, 2013] helps build successful games with the Unity game development platform using the powerful C# language, Unity's intuitive workflow tools, and a state-of-the-art rendering engine to build and deploy mobile, desktop, and console games.

Task

Design a framework of graphical user interface components that is answer to the next requirements:

- convenient for reuse;
- containing elements that are not present in standard UI framework.

Grounding of choice game engine for user interface designing

Currently there are several environments for Android applications development, and they vary in many ways. The majority of such environments require purchasing a specific license for commercial usage. However, there are some non-proprietary solutions with open source code. The most widespread are:

- android studio;
- unity.

The major advantages of these environments are:

- presents of debugging tools;
- out-of-box software for hardware emulation;
- integration with modern integrated development tools.

Peculiarity of Unity3D is a possibility of creating cross-platform mobile applications, while Android Studio is designed only for implementing Android apps.

Asset store libraries overview

Represent an analysis of free libraries that are represented on Asset Store market in “Scripting/GUI” category.

Consider the next uGUI Windows Extension, Skill UI, NGUI Infinite Pickers and GS Custom Multipurpose Dynamic Listview.

uGUI Windows Extension is a library by Motion Entertainment that is designed to use modal and dialog windows in a Unity application. This library includes customizable templates and ready-to-use features like window transitions. The appearance of components in uGUI Windows Extension is highly customizable.

Skill UI by Hamed’s Games is a set of components for designing user interfaces includes the following:

- grid for displaying text, images and other interface elements in a tabular view in Unity3D;
- uniform grid for displaying the very same information as grid when all components, and grid cells have the same size;
- dock panel - interface area used for placing child elements in horizontal or vertical rows;
- wrap panel - interface area that places child elements horizontally from left to right;
- stack panel - interface area that places child elements in a line that may be aligned horizontally and vertically.
-

NGUI Infinite Pickers (developer Cregzo) is a big collection of GUI widgets:

- date picker - interactive scrollable widget for selecting date and time values;
- image picker - interactive widget for selecting an image (i.e. game avatar);
- item picker - interactive widget for selecting one of text field options (i.e. character class selecting).

Requirement specification for GUI components library

A requirement specification for graphic elements' user library is presented in the table 1.

Table 1. Requirement specification for graphic elements user library

Requirement code	Requirement description
Functional requirements	
F1	Tabular display of text and numeric data;
F2	Tabular data scrolling
F3	Table is updated as soon as the data are changed. After this, the table is available for new review.
F4	Vertical and horizontal content alignment
F5	Text areas display depending on state of other widgets
F6	Text areas displaying and collapsing data
Non-functional requirements	
NF1	To be well documented using behavioral UML diagrams
NF2	To be designed for easy reuse
NF3	To be extensible
NF4	Adapting UI components for different resolutions of mobile screens
NF5	Portability

Consider process of software requirement elicitation taking an example functionality of UI component supporting representation of table supporting scrolling. Activity diagram for UI table with scrolling is shown in the figure 1. Sequence diagram is shown in the figure 1.

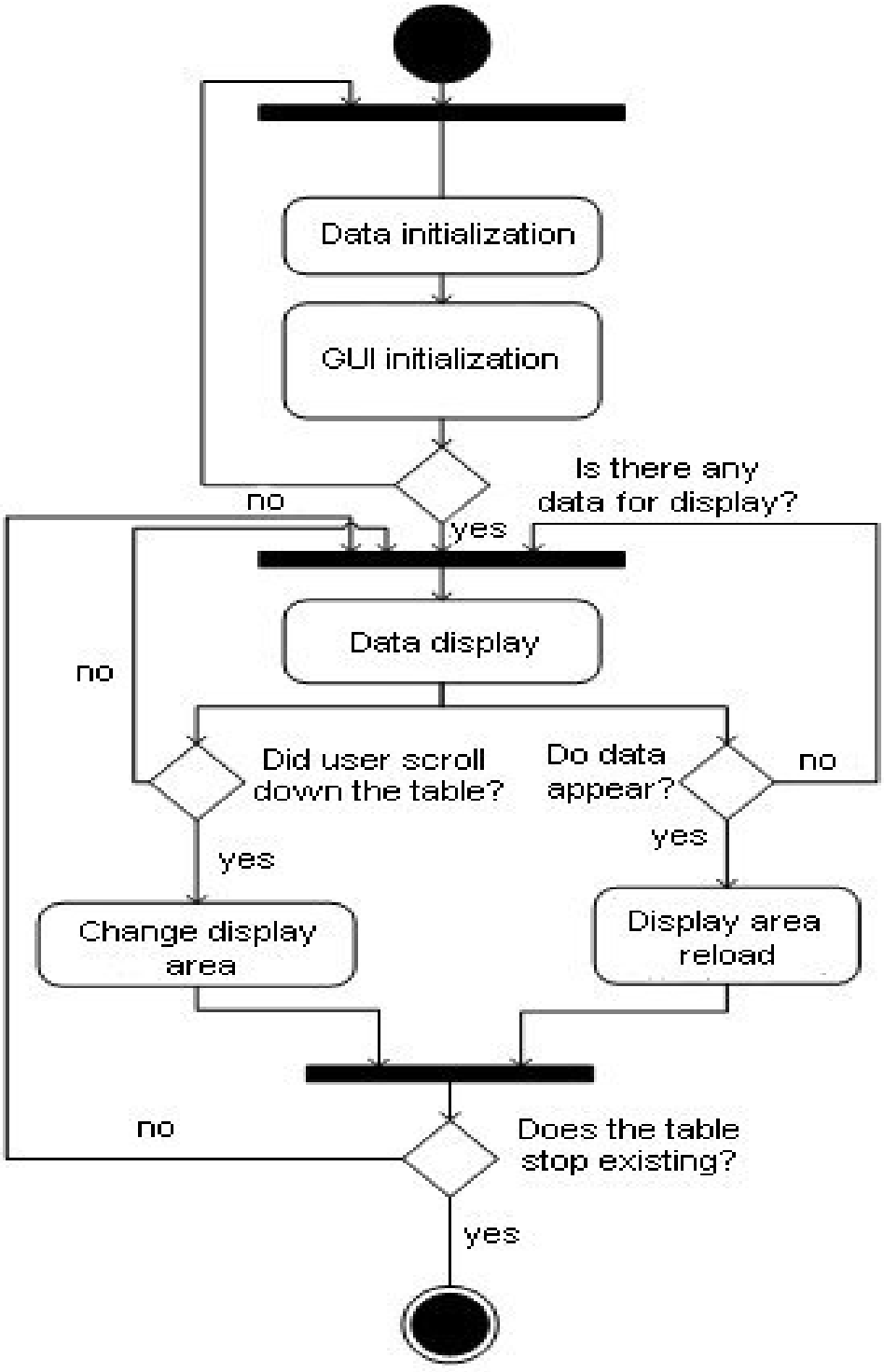

Figure 1. Activity diagram for UI table with scrolling

During table draw the table data are initialized and displayed. Table is bind to database, file or other data structure defined by developer.

During the data display, user is able to scroll through the component, is the amount of records is bigger than the display area. For the detailed analysis of algorithm for user interface working sequence diagram is designed. User passes data inside the table, which should check the data and display it in a tabular representation with indents and cell alignments.

All the alignment and indentation parameters are set up by the component developer. Upon the update, data is rewritten and resent to the table. After this, the table is filled with up-to-date information. As soon as user leaves the table it is destroyed. However, all the data are saved and renewed table have to be redrawn further.

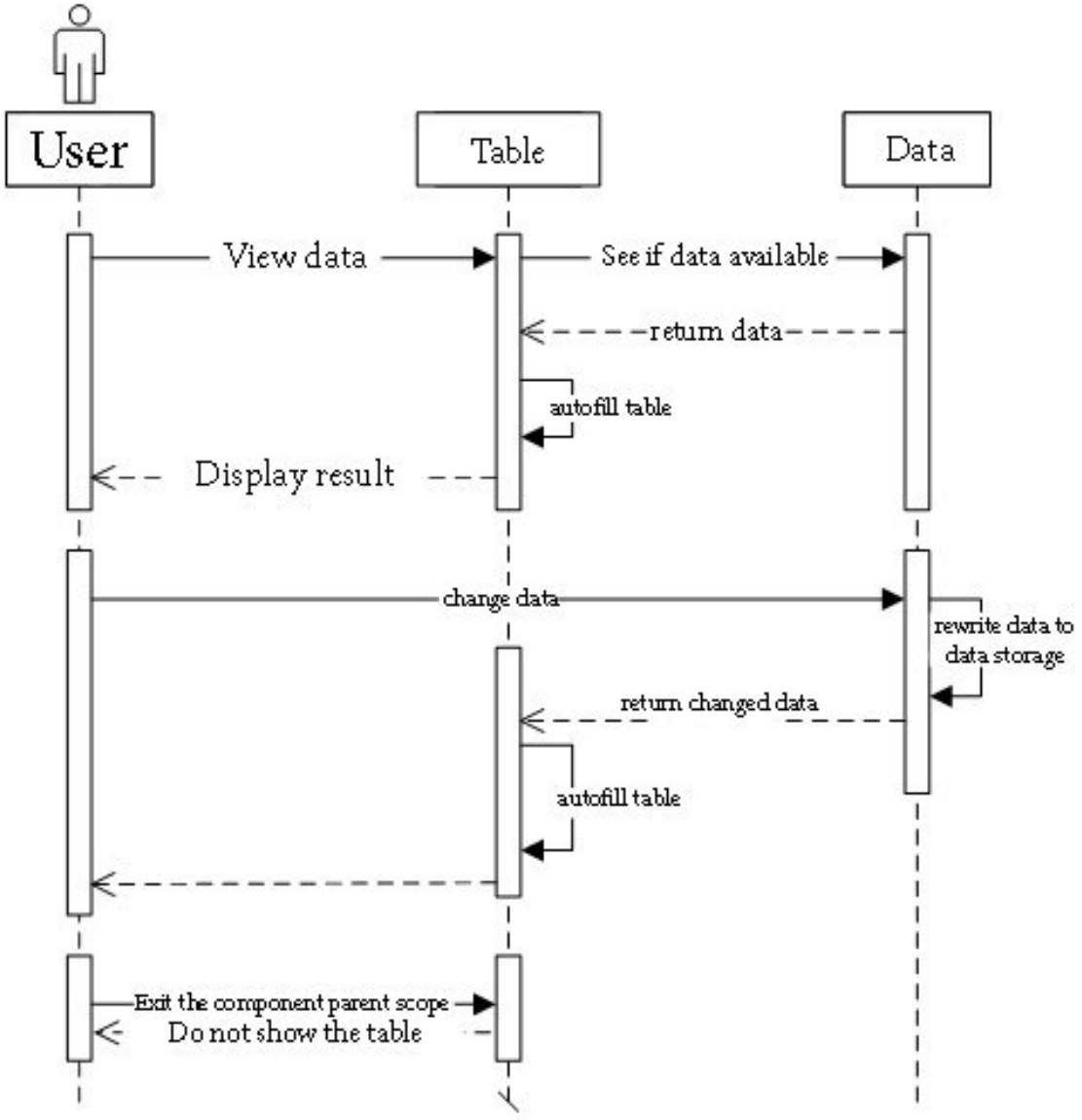

Figure 2. Sequence diagram for UI table with scrolling

Elicitation of non functional requirements

Convenient using

All the library elements are composed from the basic Unity GUI elements. It has a simple-to-understand API for developers acquainted with the Unity GUI system. In terms of ergonomics, library components can match the default GUI system components.

Portability

Library components may be used in any Unity project, which implies it can be used on various platforms and multiple operating systems.

Unity v5.3 is available on PC, Mac, Linux, Android, IOS, Tizen, PS3, PS4, Xbox 360, Web GL, Unity Web Player, Apple TV, Samsung TV, which gives an opportunity to present the product to a large audience.

Conclusions

In scope of a given paper, the following results have been accomplished:

1. a specific development environment for library creation has been chosen;
2. functional requirements for library components have been defined;
3. a fundamental library structure has been established;
4. the dependencies and interconnections between libraries have been set up;
5. prototypes of user components have been designed according to the functional requirements.

Bibliography

[Buchmann and Pezoldt, 2014] Thomas Buchmann, Patrick Pezoldt. A Lightweight Framework for Graphical Editors on Android Devices, Proceedings of the 9th International Conference on Software Engineering and Applications, August 29-31, 2014, in Vienna, Austria, University of Bayreuth, Germany ISBN: 978-989-758-036-9 – Access mode: http://scitepress.org/PublicationsDetail.aspx?ID=5zuIhFNKzes=&t=1

[Chebanyuk, 2014a] Olena Chebanyuk Method of behavioral software models synchronization International journal Informational models and analysis. 2014, Vol. 3, № 2, pp. 147-163

[Chebanyuk, 2014b] Olena Chebanyuk. Method of domain models designing. International Journal Informational Models and Analysis, 2014, Vol. 3, № 3. pp. 233-245

[Chebanyuk and Shestakov, 2017] Olena Chebanyuk, Kyrylo Shestakov. An Approach for Design of Architectural Solutions Based on Software Model-To-Model Transformation “International Journal Informational Theories and Applications”, Vol 24, № 1 – 2017, p.60-84

[Jia and Jones, 2012] Xiaoping Jia, Christopher Jones. A Model-driven Approach to Cross-platform Application Development, Proceedings of the 7th International Conference on Software Paradigm Trends - Volume 1: ICSOFT, 24-33, 2012, Rome, Italy. ISBN: 978-989-8565-19-8 – Access mode: http://scitepress.org/PublicationsDetail.aspx?ID=sbAVxcydWJ0=&t=1

[Smith and Queiroz, 2013] Matt Smith, Chico Queiroz. Unity 4.x Cookbook. – UK: Packt Publishing, 2013. – 386p.

[Hocking, 2015] Joe Hocking. Unity in Action: Multiplatform game development in C# with Unity 5, ISBN: 9781617292323, 2015. – 352 p. – Access mode: https://www.manning.com/books/unity-in-action

[Thorn, 2015] Alan Thorn. Unity Animation Essentials, ISBN113: 9781782174813, 2015. – 200 p. – Access mode: https://www.packtpub.com/game-development/unity-animation-essentials

[Lukosek, 2016] Greg Lukosek. Learning C# by Developing Games with Unity 5.x, ISBN13: 9781785287596, 2016, 230 p. – Access mode: https://www.packtpub.com/game-development/learning-c-developing-games-unity-5x-second-edition

Authors' Information

Yurii Milovidov *- National University of Life and Environmental Sciences of Ukraine, Kyiv, scientific Department of Programming Technologies. Senior Lecturer.*

e-mail: milovidov@email.ua

Major Fields of Scientific Research: Programming technologies, web-design, Internet – technologies.

COMPUTER-BASED BUSINESS GAMES' RESULT ANALYSIS

O. Vikenteva, A. Deriabin, N. Krasilich, L. Shestakova

***Abstract**: Given research considers the Business Intelligence analysis of computer based business games. A tool environment, called Competence-based Business Game Studio (CBGS), is applied for business games' design and development. An approach is proposed that allows designing and conducting business games based on enterprises business processes. Consequently, CBGS may be considered as a universal product with respect to domain. Competence-based Business Game Studio consists of several subsystems. The Analysis Subsystem makes possible to exclude human factor from the process of player skills and knowledge assessment, the latter are scored employing an automated approach based on formal parameters. This paper defines the source data for Analysis Subsystem as well. Data warehouse containing multidimensional data marts was designed for the evaluation of player's competency. Two info-cubes were developed: the first info-cube is proposed to assess players' actions, the second one - to identify bottlenecks within business processes using efficiency assessment of Decision Making Points. In order to collect information about players Complex Analysis methods are proposed for implementation: such as aggregation, navigation and filtering. To evaluate business game quality three types of Decision Making Points should be distinguished. Decision Making Points completed by players are allocated to the aforementioned types using cluster analysis (PAM-algorithm) and supervised classification.*

***Keywords**: business intelligence methods, data warehouse, competencies, active learning methods, business-game.*

***ACM Classification Keywords:** K.3 Computers and Education: K.3.2 Computer and Information Science Education – Information systems education. I. Computing Methodologies: I.2 Artificial Intelligence: I.2.1 Applications and Expert Systems – Games.*

Introduction

The implementation of game mechanics implies an increase of a player's involvement into the learning process by simulation of real-life conditions. Moreover, player's actions are evaluated in accordance with the set of competencies and criteria. There are a lot of researches in the business games area. For instance, one of the most popular and complex business games products are SimulTrain, Innov8, BrandPro. However, most of such systems focus on a certain domain.

The proposed approach to the creation of competence-based educational environment consists of the development of design, technical, organizational and methodological tools for implementing one of the active methods of forming competencies that is named competence-based business games [Vikentyeva, 2013]. The approach is multi-model and is based on the development of domain-specific models applied in design and execution stages [Vikentyeva, 2015].

Competence-based educational system (CBGS – Competence-based Business Games Studio) should consist of several subsystems. The CBGS structure is presented in Figure 1. [Vikentyeva, 2013].

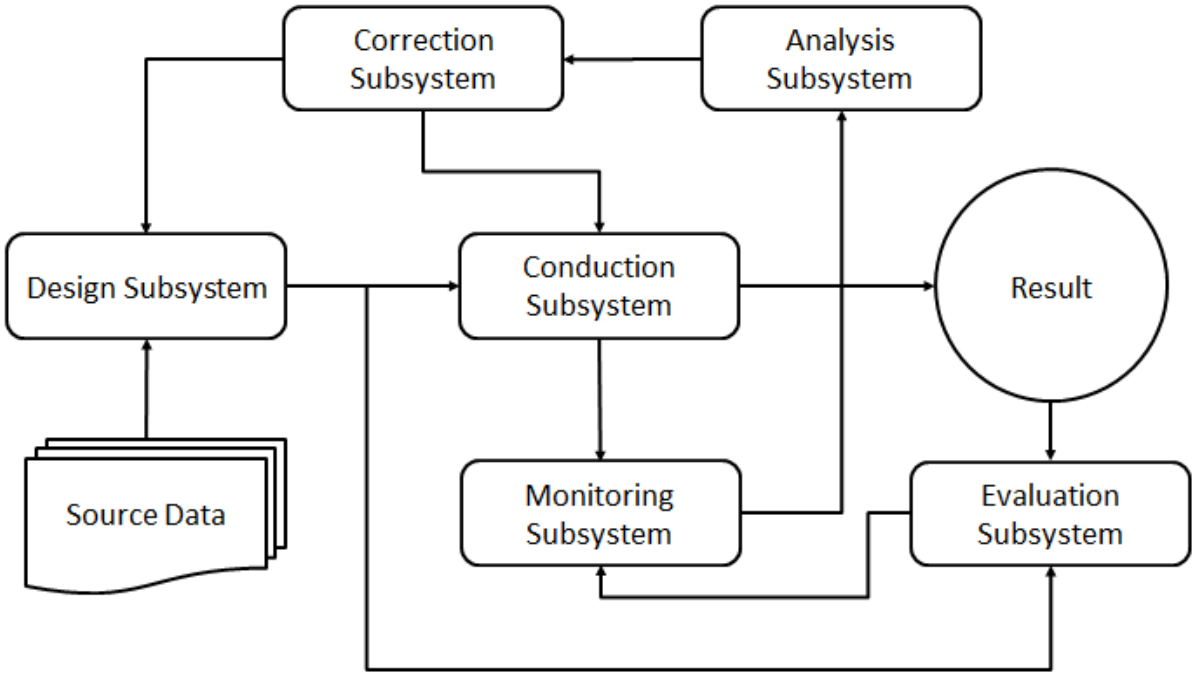

Figure 1. Structure of CBGS

Nowadays prototypes of following subsystems are developed:

- Design Subsystem. Business Processes Models are building within the Design Subsystem. These models are transformed from weakly formalized format based on real business processes models into formalized form with the use of graphical models editor.
- Conduction Subsystem. Source data for the subsystem are game plan and information about resources used during the game. The mechanism testing users is named Decision Making Point (DMP). DMP determines the course of game when a user has chosen resources.
- Evaluation Subsystem. It allows evaluating player's actions based on tests.

– Monitoring Subsystem. The subsystem implements two modules for working with databases: one is design to work with the database of operational data obtained during the game, the second works with a database of the results of players' testing.

This research issues related to business games results analysis using Business Intelligence methods are considered.

The process of human resources knowledge evaluation is subjective since it implies the influence of human factor. The CBGS's Analysis Subsystem allows excluding human factor due to automated approach to assess the trainee competency based on formal parameters. Nowadays there is a lot of research in the field of Educational Data Mining (EDM). EDM aims to apply Data Mining methods to extract information related to the learning process [Hung, 2012], [Jeong, 2013], [Sahedani, 2013].

The Analysis Subsystem should assess player's competences (knowledge, skills, experiences) based on his choice of resources within DMPs. DMPs allow the player to choose the sequence of operations of a business process. Data of passed games have to be compared with the reference model developed within the Design Subsystem.

It is important to take into account that the reason of a trainee inability to complete the game with 100% success might be the Game bottlenecks. Some algorithms may be not trivial even for experts of a corresponding business process as model of unified educational business process (UEBP) including DMP is automatically generated. Business game scenario is being built based on UEBP.

Analysis Subsystem should perform two major analysis procedures [Vikentyeva, 2016]:

– Player's actions analysis that allows providing player's characterization based on all business games, which the player participated.
– Game analysis to its correction in the case of bottlenecks identification. Such analysis has to be conducted for all DMPs.

Data Sources for Analysis Subsystem

The reference model of business process is created within the Design Subsystem. Business Process Design database stores the correct sequence of operations for each business process as well as a set a set of resources for every operation. Data for tables «Business Process», «Operation», «Resources» have to be loaded from this database [Vikentyeva et al., 2015].

Competence is a set of knowledge, skills, experience and personal characteristics, that are needed for successful performance of tasks [Kozodaev, 2015].

The concept of competence is considered in the learning process. It is important to understand that personal characteristics and experience of players are not considered within the project, because it is extremely difficult to evaluate experience level in a short time. Therefore, competence will be defined as a set of knowledge, skills necessary for successful passage of a business game.

Process of competences planning implies creation of matrix defining the dependence between operations of business processes and competences [Vikentyeva et al., 2013]. Within different business processes the same operations can be characterized by different competences.

It is possible to identify the relationship between operations and set of competences. In order to determine to what extend is competence formed and what knowledge and skills a player has, the resources that the player chooses to perform operations should also be included in the multidimensional array of competencies, since they are the ones that determine whether the player possesses the necessary set of knowledge and skills to perform the operation (the player knows which resources to choose and can apply them).

This structure can easily be formed in a multidimensional data warehouse, developed within the framework of the analysis subsystem [Vikentyeva, 2016]. The schema of database storing the results of passing games is also considered in the research [Vikentyeva, 2016].

Based on the data that can be extracted from the Design Subsystem and the Conduction Subsystem, it can be determined that the evaluation of the players' actions should be carried out according to three criteria:

- Correspondence of the sequence of operations performed by the learner during the game to the reference model.
- Competence of the player (within a single game).
- Satisfactory time of passing the game.

Data Warehouse Info-objects

Data warehouse info-objects are divided into two types [Kolb, 2012]:

- A characteristic is a sequence of values of one of analyzed parameters. Characteristics may include master data, texts and hierarchies;
- A key figure is a data quantitatively characterizing the set of characteristics.

Table 1 presents characteristics that are created within the designing data warehouse.

Table 1. Characteristics Developed in the Data Warehouse

Characteristic Name	Type	Amount of Symbols
Time of a Game	Time	–
Game Number for the Player	Integer	–
Player	String	5
Business Process	String	255
Operation	String	255
Resource Type	String	255
Resource	String	255
Competence	String	255
Competence Type (Knowledge/Skill)	String	6
Knowledge/Skill Name	String	255
Operation Number in the Reference Model	Integer	–
Actual Operation Number	Integer	–

Table 2 presents key figures that are created within the designing data warehouse.

Table 2. Key Figures Developed in the Data Warehouse

Key Figure Name	Type	Unit of Measurement
The Deviation in Operations Sequence	Integer	–
Operation Performance Indicator	Integer (0 or 1)	–
Resource Selection Indicator	Integer (0 or 1)	–
Formed Percentage of Knowledge/Skill	Number	Percentage
Maximum Percentage of Knowledge/Skill	Number	Percentage

Data Warehouse Info-Providers

Within the developed data warehouse multidimensional data marts (info-cubes) are used.

In accordance with the functional requirements for the Analysis Subsystem it is necessary to design two info-cubes:

- Evaluation of Players' Actions.
- Search of Business Game Bottlenecks.

Table 3 represents the set of info-objects that are included into the info-cube designed for evaluation of players' actions [Vikentyeva, 2016].

Table 3. Structure of Info-cube Designed for Evaluation of Players' Actions

Dimension	Characteristics
Time	Time of a Game
Game	Player
	Game Number for the Player
	Business Process
	Operation
	Resource
Competence	Competence
	Competence Type (Knowledge/Skill)
	Knowledge/Skill Name
Key Figures	
	The Deviation in Operations Sequence
	Operation Performance Indicator
	Formed Percentage of Knowledge/Skill

With the use of this set of data, the following reports can be obtained:

- The percentage of each competence formation for the player. The report will show aggregated data on competences.
- List of knowledge and skills that a player possesses or does not possess.
- Correspondence of actual operation sequence of a game to the reference model.

Table 4 represents the set of info-objects that are included into the info-cube designed for searching business game bottlenecks [Vikentyeva, 2016].

Table 4. Structure of Info-cube Designed for Searching Business Game Bottlenecks

Dimension	Characteristic
Game	Player
	Business Process
	Game Number for the Player
Decision Making Point	Operation
	Resource Type
	Resource
Key Figures	
	Resource Selection Indicator
	Maximum Percentage of Knowledge/Skill

By applying clustering to the data bottlenecks in decision making point (DMP) can be detected.

The Process of Loading Data into Data Warehouse Info-providers

Into the info-cubes data is loaded from the following databases:

- Database for business processes' modeling.
- Database for competence planning.
- Database of actual results of game.

The algorithms for loading data into the info-cube designed for evaluation of players' actions are presented in Table 5.

Table 5. The Algorithms for Loading Data into the Info-cube Designed for Evaluation of Players' Actions

Dimension	Characteristics	Algorithm of Data Loading	Source Database
Time	Time of a Game	Formula: End Time-Start Time of a Game	Database of actual results of game
Game	Player	Direct assignment	Database of actual results of game
	Game Number for the Player	Count distinct Business Process ID with the actual Business Process ID, Player ID and Start Time less or equal the Game Start Time	Database of actual results of game
	Business Process	Direct assignment	Database for business processes' modeling
	Operation	Direct assignment	Database for business processes' modeling
	Resource	Direct assignment	Database for business processes' modeling
Competence	Competence	Direct assignment	Database for competence planning
	Competence Type (Knowledge/Skill)	Defined by table type	Database for competence planning
	Knowledge/Skill Name	Direct assignment	Database for competence planning

Key Figures			
Dimension	Characteristics	Algorithm of Data Loading	Source Database
	The Deviation in Operations Sequence	Formula: Operation Number within the Reference Model for the Game – Actual Operation Number	Database for business processes' modeling Database of actual results of game
	Operation Performance Indicator	If the operation is present in the database of the actual results of games for a particular game and for a specific player, then 1, otherwise 0	Database of actual results of game
	Formed Percentage of Knowledge/Skill	If the resource characterizing knowledge/skill is selected within the specified business process and operation, then the percentage of knowledge/skill within the competence is assigned, otherwise 0	Database for competence planning Database of actual results of game

Table 6. The Algorithms for Loading Data into the Info-cube Designed for Searching Business Game Bottlenecks

Dimension	Characteristics	Algorithm of Data Loading	Source Database
Game	Player	Direct assignment	Database of actual results of game
	Game Number for the Player	Count distinct Business Process ID with the actual Business Process ID, Player ID and Start Time less or equal the Game Start Time	Database of actual results of game
	Business Process	Direct assignment	Database for business processes' modeling
Decision Making Point	Operation	Direct assignment	Database for business processes' modeling
	Resource	Direct assignment	Database for business processes' modeling
	Resource Type	Direct assignment	Database for business processes' modeling
Key Figure			
	Resource Selection Indicator	Direct assignment (if resource was selected, then 1, otherwise 0)	Database of actual results of game
	Maximum Percentage of Knowledge/Skill	Direct assignment	Database for competence planning

Data Analysis Algorithms for the Info-cube Designed for Evaluation of Players' Actions

Complex Analysis method is applied for Evaluation of Players' Actions. The player's competence within a single business process may be defined by several ways:

- The total competence of player based on actual results of game. Aggregation on Business Process and calculation of average percentage of competence are necessary for this analysis. Other characteristics are not considered. Sample of data includes Game Number, Player, Business Process, Competence, Formed Percentage of Knowledge/Skill.
- The percentage of competence obtained by a player within a single game. Such a sample will determine the degree of competence obtained by the player within the operation. Sample of data includes Game Number, Player, Business Process, Operation, Competence, Formed Percentage of Knowledge/Skill.
- Possession of certain knowledge and skills. For this type of analysis, the data should be fully detailed. Sample of data includes Game Number, Player, Business Process, Operation, Resource, Competence, Competence Type (Knowledge or Skill), Knowledge/Skill Name, Formed Percentage of Knowledge/Skill (the key figure is restricted by condition «>0»).
- Unformed knowledge and skills of the player. For this type of analysis, all the data within a single game must be aggregated by Operations and Knowledge/Skills. Sample of data includes Game Number, Player, Business Process, Operation, Competence, Competence Type (Knowledge or Skill), Knowledge/Skill Name, Formed Percentage of Knowledge/Skill (the key figure is restricted by condition «==0»).
- In addition to the degree of the player's competence, the data set of the Info-cube also allows to determine the deviation of actual operations' sequence from the reference model. Sample of data includes Game Number, Player, Business Process, Operation, The Deviation in Operations Sequence (the key figure is restricted by condition «<>0»).
- In addition, it is possible to identify which operations from the reference model were not performed. Sample of data includes Game Number, Player, Business Process, Operation, Operation Performance Indicator (the key figure is restricted by condition «==0»).
- A player progress. This type of analysis is performed by comparison of all results of passing a particular game if the player participates in the game not for the first time.

Data Analysis Algorithms for the Info-cube Designed for Searching Business Game Bottlenecks

Models of real business processes performed at enterprises can't be used in the design of business games, therefore the concept of a model of a unified educational business process (UEBP) is introduced [Vikentyeva, 2015]. UEBP reflects the essential invariant characteristics of business processes of enterprises. UEBP can be quite complex and include not only consistent actions, but also various business conditions, repetitive operations. UEBP must contain operations that simulate the learning situation in the Business Game. The learning situation is understood as the situation in which decisions are made in the process of selecting resources for performing operations and/or the next operation of business process, etc. The learning situation allows to form or verify the player's competencies.

The Business Game is an interactive test for each player, and, as it is known, the tests should include questions, the correctness of the answers to which has a normal distribution. Therefore, there are two types of Decision Making Points taking a role of bottlenecks in Business Game or UEBP. Types of such points are the following:

- Simple DMPs are DMPs in which almost nobody makes mistakes even passing a game for the first time.
- DMPs of increased complexity are DMPs in which even the most competent players make the same mistakes.

The data analysis for the search for "bottlenecks" in Business Game should be implemented using one of the Data Mining methods - clustering. At this stage of the design, we are looking for clusters of three types of DMPs:

- Simple DMPs.
- DMPs of normal complexity.
- DMPs of increased complexity.

Set of characteristics of the same type is used for each combination of Business Process and Decision Making Point. Training sample is formed for all Business Processes.

The following characteristics must be used for DMPs' clustering:

- Amount of mistakes made when selecting mandatory resources.
- Amount of mistakes made when selecting optional resources.
- Formed percentage of the player's competence within each game.

During factor analysis it was revealed that the characteristics must be normalized in order to reduce the amount of data. Normalization of data is performed by calculating the following values for each DMP:

- Average rate of mistakes made when selecting mandatory resources.
- Average rate of mistakes made when selecting optional resources.
- Average rate of the player's competence within each game.

These average values represent three dimensions in the characteristic set for clustering.

After that it was necessary to identify the most appropriate clustering algorithm for finding bottlenecks in Business Game.

It is important to understand that the search for problem Decision Making Points needs to be done in two stages, that is, clustering is performed two times. Simple Decision Making Points need to be identified in a sample that includes the results of absolutely all games, including games of players with low level of competencies. Decision Making Points of increased complexity should be identified only among those games for which users have received high assessment, that is, the average player's competence within a business process is at least 75%. The second sample allows clearing the data from the unsuccessful traineeship due to a lack of knowledge of business processes.

The paper [Barsegian, 2004] provides a description of the clustering algorithms that is later is used for algorithms' comparison.

Comparison of clustering algorithms will be performed according to the following criteria:

- The total number of clusters is known (three clusters: simple DMPs, DMPs of normal complexity and DMPs of increased complexity).
- The volume of data sets may vary.
- The form of the clusters is arbitrary.
- Ease of work with multidimensional objects.
- The distance between clusters is small.

These criteria were singled out on the basis of the initial data (data in info-cube for searching bottlenecks in Business Game), requirements for the result of data analysis and analysis of clustering methods.

The comparison is made by the method "from the inverse", that is, it is determined which algorithms do not satisfy the criteria in question. A comparison of clustering algorithms is presented in Table 7.

Table 7. Clustering Algorithms Comparison

Algorithm Name	Known Total Number of Clusters	Variable Volume of Data Sets	Arbitrary Form of Clusters	Ease of Work With Multidimensional Objects	Small Distance Between Clusters
AGNES (Agglomerative Nesting)	No	Yes	Yes	Yes	Yes
CURE	Yes	No	Yes	Yes	Yes
DIANA (Divisive Analysis)	No	Yes	Yes	Yes	Yes
BIRCH	No	Yes	No	Yes	Yes
MST	No	Yes	Yes	No	Yes
K-means	Yes	Yes	Yes	Yes	No
Maximin	No	Yes	Yes	Yes	Yes
PAM	Yes	No	Yes	Yes	Yes
CLOPE	No	No	Yes	Yes	Yes
Self-organizing Map	Yes	No	Yes	Yes	Yes
HCM	Yes	No	Yes	Yes	Yes
Fuzzy C-means	Yes	No	Yes	Yes	Yes

The results of the comparison show that no algorithm fully meets all the criteria. However, it is important to take into account that under large volumes of data, databases with a multimillion-number transactions and a large set of characteristics are understood. As factor analysis allowed reducing characteristic set to three dimensions and OLAP technology allows getting aggregated data, it is assumed that actually the volume of data is not large in the general sense. Thus, evaluating the characteristics of the algorithms, it was decided to use the PAM algorithm for DMPs' clustering, since the training sample will not include a huge number of objects, the number of clusters is set and equal to three, the algorithm is less sensitive to emissions, the occurrence of which cannot be predicted in advance. Clusters will be classified in remoteness from the reference model.

In order to exclude of possible clustering mistakes related to fixed numbers of clusters two technical DMPs should be added into training sample. For instance all DMPs might be of normalized complexity, but PAM algorithm will distribute them into three sets anyway as this condition is set initially. Technical DMPs with following parameters {0; 0; 100} and {100; 100; 0} representing simple DMP and DMP of increased complexity properly allows getting rid of this problem. Here the first parameter is average rate of mistakes made when selecting mandatory resources, the second - average rate of mistakes made when selecting optional resources and the third - average rate of the player's competence gained during the DMP performance. Such technical DMPs should not be displayed to the player as an output, but allow avoiding errors associated with a fixed set of clusters.

In addition to clustering, supervised classification should also be applied to evaluate the quality of DMPs' design. Since DMPs distributed in clusters «Simple DMP» and «DMP of increased complexity» might be simple or complex but have not worthless id their parameters are not equal to {0; 0; 100} or {100; 100; 0}. The decision about such points redesign has to be made by the developer of UEBP, however DMPs having parameters equal to {0; 0; 100} or {100; 100; 0} should be highlighted singularly since they require redesign doubtlessly.

Conclusion

Since each resource is related with knowledge or skill analysis of player's competency is possible.

To conduct analysis of player's competency corresponding info-cube was designed. Applying Complex Analysis methods such as aggregation, navigation and filtering following reports regarding player's competency can be obtained:

– Player's competency within a business process.

- The percentage of different player's competences.
- Bottlenecks in player's knowledge and skills.
- The reasons of the lack of player's competency.
- The list of the most qualified participants.
- Weaknesses of players.
- Knowledge and skills not acquired by players previously.
- The average time taken to complete a single iteration of the business process.
- The progress of players in time.

To conduct analysis of Business Game another info-cube was designed. The analysis of the Business Game includes an assessment of the degree of successful DMPs performance in order to determine if DMPs were designed correctly. To assess the quality of the business game, it is proposed to distinguish three types of Decision Making Points:

- Simple DMPs.
- DMPs of normal complexity.
- DMPs of increased complexity.

Clustering is used to distribute all DMPs to these types. To determine the most appropriate clustering algorithm, a characteristic set was determined:

- Amount of mistakes made when selecting mandatory resources.
- Amount of mistakes made when selecting optional resources.
- Formed percentage of the player's competence within each game.

In order to increase operating speed of clustering algorithm, it was necessary to reduce the number of analyzed transactions. Therefore, it was decided to normalize the analyzed indicators and characteristic set was reformulated as follows:

- Average rate of mistakes made when selecting mandatory resources.
- Average rate of mistakes made when selecting optional resources.
- Average rate of the player's competence within each game.

Since the number of clusters is knows and the volume of training sample is not large due to normalization of analytic set it was decided to use the PAM in order to assess DMPs. In addition to clustering supervised classification should be applied.

Based on implementation of clustering and supervised classification algorithms the UEBP developer is able to identify DMPs that are recommended to revision as well as DMPs that must be revised as they distort the results of player's competence assessment.

Bibliography

[Barsegian, 2004] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004, С. 67-75.

[Hung, 2012] Hung J.L., Rice K., Saba A. An Educational Data Mining Model for Online Teaching and Learning. Journal of Educational Technology Development and Exchange, 2012, pp. 77-94.

[Jeong, 2013] Jeong H. Educational Data Mining. How Students' Self-motivation and Learning Strategies Affect Actual Achievement. Department of Computer Science, Indiana University-Purdue University Fort Wayne, 2013.

[Kolb, 2012] Kolb E. BW310: BW - Enterprise Data Warehousing – Germany: SAP AG, 2012.

[Kozodaev, 2015] Козодаев М.А. Оценка проектного персонала: не забыть бы, для чего это делается (часть 1). Управление проектами и программами, 2015.

[Sahedani, 2013] Sahedani K.S., Supriya Reddy B. A Review: Mining Educational Data to Forecast Failure of Engineering Students. International Journal of Advanced Research in Computer Science and Software Engineering, 2013, pp. 628-635.

[Vikentyeva, 2013] Викентьева О.Л., Дерябин А.И., Шестакова Л.В. Концепция студии компетентностных деловых игр. Современные проблемы науки и образования, № 2, 2013, http://www.science-education.ru/108-8746.

[Vikentyeva et al., 2013] Викентьева О.Л., Дерябин А.И., Шестакова Л.В. Функциональные требования к студии компетентностных деловых игр. Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. № 8, 2013, С. 31-40.

[Vikentyeva, 2015] Викентьева О.Л., Дерябин А.И., Шестакова Л.В., Лебедев В.В. Многомодельный подход к формализации предметной области. Информатизация и связь. №3, 2015, С.51-56.

[Vikentyeva et al., 2015] Викентьева О.Л., Дерябин А.И., Шестакова Л.В., Красилич Н.В. Проектирование редактора ресурсов информационной системы проведения деловых игр. Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. №16, 2015, С. 68-87.

[Vikentyeva, 2016] Vikentyeva O., Deryabin A., Krasilich N., Shestakova L. Employment of Business Intelligence Methods for Competences Evaluation in Business Games. International Journal "Information Technologies & Knowledge", V. 10, №3, 2016, pp. 286-299.

Authors' Information

Olga Vikentyeva *– National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: oleovic@rambler.ru.*

Major Fields of Scientific Research: Information Systems Design, Software Engineering, Data Mining

Alexandr Deryabin – *National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: paid2@yandex.ru.*

Major Fields of Scientific Research: Multi-dimensional information systems, Data Mining

Nadezhda Krasilich *– National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: mefaze@yandex.ru.*

Major Fields of Scientific Research: Business Informatics, Business Intelligence

Lidiia Shestakova *– National Research University Higher School of Economics, City of Perm, Perm, Russia, e-mail: L.V.Shestakova@gmail.com.*

Major Fields of Scientific Research: Computational Mathematics, Business Informatics.

МЕТОДЫ АНАЛИЗА ЗАШИФРОВАННОГО ТРАФИКА ДЛЯ ОБНАРУЖЕНИЯ СКРЫТЫХ УГРОЗ

Тамара Радивилова

***Аннотация**: Злоумышленники и вредоносное программное обеспечение используют зашифрованный протокол SSL/TLS для осуществления несанкционированной активности, что создает проблемы при обнаружении вторжений. Существует два подхода к обнаружению вторжений в зашифрованном трафике: без его дешифрования и после его дешифрования. В ходе работы проведен анализ основных методов дешифрования трафика протокола SSL/TLS. В работе представлены методы и технологии обнаружения вредоносной активности в зашифрованном трафике, которые используются ведущими компаниями. Также предложен метод перехвата и дешифровки трафика, передаваемого по протоколу SSL/TLS, который можно применять при удаленном прослушивании сети, что позволяет дешифровывать передаваемые данные в режиме приближенному к реальному времени.*

***Ключевые слова**: протокол SSL/TLS, угрозы, уязвимость, системы обнаружения сетевых вторжений, методы дешифрования.*

***ITHEA Keywords**: E.3 Data Encryption, C.2 Computer-communication networks - C.2.2 Network Protocols, I.5 Pattern recognition - I.5.4 Applications, K.6 Management of computing and information systems – K.6.5 Security and Protection.*

Введение

В последнее время наметился тренд увеличения доли шифрованного трафика. По оценкам компании Cisco на данный момент 60% трафика в Интернет зашифровано, а согласно прогнозам Gartner к 2019-му уже 80% трафика будет таковым [Cisco, 2014, Orans, 2016]. Шифрование необходимо для обеспечения приватности граждан, сохранения тайн в секрете, выполнения требований законодательства. Но злоумышленники также используют шифрование для обхода механизмов детектирования их несанкционированной активности, скрывая взаимодействие с командными серверами вредоносных программ и для других задач. [Cisco, 2014, Orans, 2016, Лукацкий, 2018]

С одной стороны средства защиты не могут видеть, что происходит в зашифрованном трафике (по данным Ponemon Institute 64% компаний не могут детектировать вредоносный код в зашифрованном трафике). Для проникновения в зашифрованные соединения организации часто используют атаку Man-in-the-Middle (MITM), которую они осуществляют в легальных целях, но это не является легальным (нарушение тайны переписки, требования законодательства по обеспечению конфиденцильности информации). Обычно на периметре корпоративной или ведомственной сети устанавливается шлюз или кластер из шлюзов, которые осуществляют “перехват” и дешифрование данных внутри сети компании, чтобы защититься от атак, использующих протокол SSL/TLS (SSL - Secure Sockets Layer, TLS - Transport Layer Security) для передачи вредоносного содержимого, а также для анализа передаваемых данных системами обнаружения вторжений (IDS). [Ponemon, 2016, D'Hoinne, 2013]. После дешифрования трафик проверяется на наличие вредоносных активностей, зашифровывается снова и отправляется на IP-адрес назначения. Обнаружение вредоносных активностей в SSL/TLS трафике является трудоемким и сложным, поскольку шифрование мешает эффективности классических методов обнаружения и является сложной проблемой для IDS.

Целью этой работы является обзор методов анализа трафика и разработка метода дешифрования трафика SSL/TLS для обнаружения скрытых угроз.

Основные механизмы проверки трафика SSL/TLS и разработки ведущих компаний

Разработчики ведущих компаний и ученые ведут активную работу по разработке методов обнаружения вредоносной активности в зашифрованном трафике.

Cisco Encrypted Traffic Analytics извлекает и анализирует четыре основных элемента данных: последовательность длин и времени пакета, распределение байтов, специфичные для TLS функции и исходный пакет данных. Уникальная архитектура специализированных интегральных схем Cisco (ASIC) обеспечивает возможность извлечения этих элементов данных без замедления работы сети передачи данных. [Cisco, 2014]

Cisco Stealthwatch Enterprise использует NetFlow, прокси-серверы, телеметрию конечных точек, механизмы политики и доступа, сегментацию трафика и многое другое, чтобы установить базовое «нормальное» поведение для хостов и пользователей на предприятии. Stealthwatch может коррелировать трафик с глобальными угрозами, чтобы автоматически идентифицировать зараженные хосты, командные и контрольные коммуникации и подозрительный трафик.

Stealthwatch поддерживает глобальную карту рисков - очень широкий поведенческий профиль о серверах в Интернете, идентифицирующий серверы, связанные с атаками, может быть использован как часть атаки в будущем. Это не является черным списком, а представляет собой целостную картину с точки зрения безопасности. Stealthwatch анализирует новые зашифрованные элементы данных трафика в расширенном режиме NetFlow, применяя методы машинного обучения и статистическое моделирование, чтобы выявлять вредоносные шаблоны в зашифрованном трафике, для выявления угроз и улучшения реакции на инциденты.

The SANS Institute предлагает использовать четыре подхода к дешифрованию соединений SSL/TLS: 1) выполнение проверки на самом сервере; 2) прокси-сервер терминалов; 3) дешифрование самим IDS; 4) автономный инструмент для дешифрования соединения [Butler, 2013, Bakhdlaghi, 2017].

1. Выполнение проверки на самом сервере. Самый простой способ проверить зашифрованный трафик - использовать IDS на основе хоста (HIDS) на самом сервере, где дешифруется трафик, принадлежащий этому серверу. HIDS может отслеживать действия сервера и искать необычное поведение, изменения в базах данных, системных файлах или любых критически важных данных. Установка HIDS может добавить дополнительную нагрузку, которая может негативно повлиять на производительность, особенно для нагруженного сервера.

2. SSL/TLS терминальный прокси (обратный прокси). Обратный прокси - это сервер, который выступает в качестве посредника между серверами бекэнда и клиентами. Он принимает запросы клиентов и извлекает ресурсы, эффективно скрывающие бекэнд сервера от клиентов. Обратный прокси сервер может быть настроен для выполнения шифрования SSL/TLS, выступающего в качестве SSL/TLS терминального проски, который снимает нагрузку с дешифровки соединений SSL/TLS, передавая незашифрованный трафик на ассоциированные сервера. Однако использование прокси-сервера SSL / TLS позволяет использовать IDS внутри локальной сети серверов.

3. IDS выполняющий дешифрование. IDS предоставляется возможность выполнения процесса дешифрования при закрытом ключе. Это может быть предварительный процессор или плагин, который поддерживает дешифрование и нормализацию трафика перед тем, как перейти к механизму обнаружения. В настоящее время нет препроцессора для Snort для выполнения процесса дешифрования, хотя теоретически возможно разработать такой предварительный процессор или подключаемый модуль (Snort FAQ, n.d.). Однако функция дешифрования доступна в некоторых устройствах IDS, таких как Juniper IDP.

4. Автономный инструмент выполняющий дешифрование. Инструмент Viewssld использовался для дешифрования соединения SSL/TLS, использующего обмен ключами RSA. Viewssld - это бесплатный инструмент с открытым исходным кодом, который может дешифровывать трафик SSL / TLS для IDS. Он работает, прослушивая интерфейс на определенном IP-адресе, дешифруя зашифрованный трафик с помощью закрытого ключа сервера и выдает дешифрованный трафик на порт прослушивания IDS. Он не поддерживает обмен ключами Диффи-Хелмана, а поддерживает только обмен ключами RSA.

Компания Symantec использует решение Encrypted Traffic Management для устранения зашифрованного скрытого трафика [Symantec, 2017]. Ключевым компонентом этого набора решений SSL Visibility Appliance является высокопроизводительное средство проверки, дешифрования и управления SSL, масштабирование до 9 Гбит/с SSL-дешифрования и способное одновременно передавать дешифрованную информацию нескольким инструментам безопасности. Возможности проверки и дешифрования SSL, предоставляемые SSL Visibility Appliance, позволяют существующим средствам безопасности и сети (IDS/IPS - Intrusion Prevention Systems, DLP, анализаторам вредоносных программ, Next Gen Firewalls - NGFW, криминалистике, платформам аналитики безопасности), получать доступ к открытым текстам в потоках SSL, тем самым позволяя устройству безопасности эффективно выполнять свою работу, даже с SSL-зашифрованным трафиком.

Кампания Gigamon предлагает Security Delivery Platform GigaSECURE, в которой операциям безопасности разрешается использовать уникальный архитектурный подход «зоны дешифровки» для решения проблемы дешифрования SSL/TLS [Gigamon, 2017]. В «зоне дешифровки» трафик SSL/TLS дешифруется один раз и подается на несколько защищенных инструментов для дальнейшего анализа и проверки, тем самым устраняя ненужные и повторяющиеся циклы дешифрования и повторного шифрования в инфраструктуре. Благодаря расширенному дешифрованному решению дешифрования SSL/TLS, Gigamon обеспечивает полную видимость сети для выявления вредоносных угроз и предоставления дешифрованного трафика, представляющего интерес для соответствующих инструментов безопасности для немедленного анализа.

Необходимо отметить, что системы дешифровки трафика, т.е. устройства, реализующие функции SSL–прокси SSL разгрузки, в дальнейшем будут все более востребованными, учитывая рост использования протокола шифрования SSL/TLS. Они позволяют не только дешифровать трафик для снижения нагрузки на конечные серверы, но и отправить его на дополнительный анализ с привлечением сторонних средств защиты информации.

Также, существует много ситуаций, когда администраторы ИТ должны использовать проверку пакетов, например Wireshark. Обычно самым простым способом дешифрования данных является использование закрытого ключа для соответствующего открытого ключа. Wireshark предоставляет еще одно средство для дешифрования данных, а также с использованием пре-мастер ключа.

В работах [McGrew, 2016; McGrew1, 2016; Strasák, 2017] рассмотрены методы обнаружения вредоносного трафика HTTPS без его дешифрования. Такие методы очень важны, так как в этом случае отпадает необходимость в каком-либо перехватчике трафика HTTPS, соблюдалась бы конфиденциальность и безопасность сообщений, и обнаружение вторжений происходило бы быстрее. Кроме того, эти методы могут использоваться совместно с некоторым перехватчиком трафика HTTPS в качестве первого уровня обнаружения вторжений в сетевом трафике, и если какой-либо трафик будет подозрительным, тогда для дешифрования будет использоваться перехватчик трафика HTTPS.

В работе [McGrew, 2016] авторы предлагают обнаружение вредоносной активности в HTTPS трафике без его дешифровки, однако их метод основан на сборе данных из незашифрованных сообщений TLS-рукопожатия. В отличие от них, в работе [Strasák, 2017] используются данные без дешифрования. Авторы работы [McGrew1, 2016] используют без дешифрования потоки TLS, потоки DNS, HTTP заголовки и незашифрованную информацию заголовка TLS для обнаружения вредоносного трафика HTTPS. Однако данные методы применимы только после детального статистического анализа трафика в сети и его дальнейшего анализа методами машинного обучения, так как трафик в каждой сети имеет свои характерные особенности.

Ponemon Institute попросил респондентов оценить вероятность возникновения конкретных атак и возможность противостояния этим атакам, которые показаны в таблице 1 [Ponemon, 2016].

Из таблицы 1 видно, что вероятность противодействия атаке достаточно мала, по сравнению с вероятностью ее появления.

Таблица 1. Вероятность возникновения конкретных атак и возможность противостояния этим атакам

Типы атак	Вероятность	
	атаки	Противостояния атаке
1. Злоумышленник делает фишинговые угрозы еще более законными, а даже осведомленые получатели считают, что использование TLS гарантирует им безопасность. Однако, нажав на ссылку, злоумышленник отправляет пользователей к серверу SSL, на который загружено злонамеренное программным обеспечением, которое заражает клиента, поскольку трафик вредоносных программ зашифрованный и не распознается системами обнаружения вторжений.	79%	17%
2. Злоумышленник отправляет зашифрованный поток защищенных, чувствительных и других критических данных, поступающих через брандмауэр через "обычные" порты (443,80 и др.), которые брандмауэр настроен принять, поскольку они являются утвержденными портами.	78%	30%
3. Ряд злоумышленников использует шифрование, чтобы скрыть информацию о сети, включая пароли и конфиденциальные данные, которые они присылают на серверы SSL. Шифрование ослепляет системы мониторинга/инспектирования для этой внутренней сети.	74%	16%
4. Злоумышленник мешает коммуникациям с вредоносным программным обеспечением, когда червь, вирус или ботнет «звонит домой», чтобы отправить украденные данные к главному компьютеру или загрузить инструкции или больше вредоносных кодов.	66%	26%
5. С помощью межсайтового скриптинга злоумышленники похищают файлы cookie, которые могут использоваться для захвата аккаунта или сеанса, изменения настроек, отравления cookie и / или ложной рекламы. Все это можно выполнить, прячась в SSL / TLS трафике.	62%	19%

Предлагаемый метод дешифрования трафика

Описанный в данной статье метод дешифровки TLS-трафика предполагает у злоумышленника наличие доступа к компьютеру или сети, либо же наличие закладки на компьютере жертвы, которая может собрать данные о сессиях. Такие условия нужны для формирования файла сессионных ключей, который будет использован вместе с соответствующим перехваченным трафиком [Волков, 2016]. Перехватить трафик жертвы можно, находясь в любом участке сети на промежутке между сервером и объектом нападения.

Ниже приведено описание реализации предложенного метода к дешифровке TLS-трафика. В реализации использовался анализатор трафика Wireshark, который помогает провести анализ работы сети, диагностировать проблемы, а также имеет много других полезных возможностей.

Для проведения эксперимента была создана локальная сеть, состоящая из трех подсетей, веб-сервера и сервера доступа. Для подключения компьютеров из подсетей к веб-серверу использовался протокол HTTPS с использованием TLS-соединения. В одной из подсетей была добавлена дополнительная точка доступа, и закладка была сделана на одном из компьютеров, отправив ее по электронной почте. Затем, используя Wireshark, данные в сети были прочитаны дополнительной точкой доступа.

Для получения ключей включаем логгирование сессионных ключей, которые используются для зашифровки и дешифровки трафика. Получение таких логов не является трудоёмким. Их можно получить, например, из браузеров – браузеры Firefox и Chrome научились выводить в специально задаваемый файл данные, достаточные для деривации (получения) сессионных ключей, которыми шифруется передаваемый/принимаемый ими трафик, поскольку внутри TLS используется симметричное шифрование. Строго говоря, делают это не сами браузеры, а библиотека NSS в их составе; именно она задает формат записываемых файлов. Для дешифровки TLS необходимо иметь файл с логгированными записями сессионных ключей в NSS-формате и анализатор трафика Wireshark. Wireshark весьма чувствителен к формату NSS-файла, поэтому необходимо тщательно перепроверить сходимость числа байтов в каждом элементе строки и отсутствие лишних пробелов, что может сэкономить время.

Захватывать трафик нужно после того, как начнётся запись ключей в лог-файл, так как в противном случае нам не удастся завладеть сессионными ключами, которые соответствуют захваченным TLS-записям. Также необходимо помнить, что ключи являются временными, т.е. пригодны лишь для одной TLS-сессии. Также необходимо отслеживать обмен трафиком с определённым хостом и фильтрация по нужному протоколу, чтобы изначально отбросить

ненужные пакеты, проходящие через прослушиваемый интерфейс. После того, как удалось сформировать файл с сессионными ключами, нужно его привязать к Wireshark'у.

Теперь в содержимом пакета появилась вкладка «Decrypted SSL Data». Теперь, если перейти в эту вкладку можно увидеть текст запроса. Кроме того теперь можно выбрать любой пакет с протоколом SSL или TLS и в его контекстном меню выбрать функцию «Follow SSL Stream» – в результате получается содержимое пакетов. Как видно, несмотря на то, что общение проходит по HTTPS, мы видим передаваемый трафик и можем экспортировать его для дальнейшего анализа.

Описанный в данной работе метод обладает основным недостатком: он требует существенных затрат времени на составление файла с сессионными ключами. Однако предложенный метод можно формализовать, а в последующем и автоматизировать, что сократит временные затраты на реализацию данной атаки и, возможно, откроет новые возможности для проведения таких атак.

Особенностью описанного метода является то, что не обязательно перехватывать трафик на компьютере, который генерирует TLS-трафик, его можно перехватывать находясь в сети и прослушивая её. А добыть файл с сессионными ключами можно, поставив на компьютер жертвы закладку или просто скопировать его, имея доступ к компьютеру

Выводы

SSL/TLS стал универсальным стандартом для аутентификации и шифрования сообщений между клиентами и серверами. Он широко распространен в организациях и предприятиях и быстро растет из-за быстрого увеличения облачных, мобильных и веб-приложений. Однако SSL создает угрозу безопасности, вводя «слепое пятно», что увеличивает риск проникновения вредоносного ПО в организацию. Проверка SSL/TLS является важной и желательной функцией для аналитиков безопасности, но она имеет свою стоимость.

Для дешифрования трафика желательно выбрать способ дешифрования трафика, основанный на потребностях и структуре сети: на самом сервере, SSL/TLS терминальный прокси или использование автономного инструмента или возможностей, добавленных в IDS. Если HIDS установлен на самом сервере, он может добавить дополнительную нагрузку, которая может негативно повлиять на производительность, особенно для загруженного сервера.

В работе предложен метод дешифровки трафика SSL/TLS, который можно применять даже при удаленном прослушивании сети. Данный метод был автоматизирован и позволяет дешифровывать данные практически в режиме онлайн.

В работе [Anderson, 2017] проведен анализ использования TLS вредоносными и корпоративными приложениями, в ходе которого взяты миллионы зашифрованных потоков TLS и целевое исследование по 18 семействам вредоносных программ, которые состоят из тысяч уникальных образцов вредоносных программ и десяти тысяч вредоносных потоков TLS. Сделан вывод, что использование TLS вредоносными программами отличается от доброкачественного использования в настройках предприятия и что эти различия эффективно используются в правилах и классификаторах машинного обучения. В своей дальнейшей работе мы планируем провести анализ зашифрованного SSL/TLS трафика методами data science на обнаружение несанкционированной деятельности.

Bibliography

[Anderson, 2017] Blake Anderson, Subharthi Paul and David McGrew. Deciphering Malware's use of TLS (without Decryption). Journal of Computer Virology and Hacking Techniques, pp 1–17, 2017. https://doi.org/10.1007/s11416-017-0306-6

[Bakhdlaghi, 2017] Yousef Bakhdlaghi. Snort and SSL/TLS Inspection. SANS Institute. InfoSec Reading Room. P.24. 2017.

[Butler, 2013] J. Michael Butler. Finding Hidden Threats by Decrypting SSL. A SANS Analyst Whitepaper. 2013.

[Cisco, 2014] White paper. Encrypted Traffic Analytics. Cisco public, 2018.

[D'Hoinne, 2013] Jeremy D'Hoinne and Adam Hils. Security Leaders Must Address Threats From Rising SSL Traffic. Gartner, 2013.

[Gigamon, 2017] Whitepaper: Prevent Encrypted Threats and Data Loss with Inline SSL Decryption.

[McGrew, 2016] David McGrew, Blake Anderson, Subharthi Paul. Deciphering Malware's use of TLS (without Decryption). 6 Jul 2016.

[McGrew1, 2016] David McGrew, Blake Anderson. Identifying Encrypted Malware Traffic with Contextual Flow Data. 2016.

[Orans, 2016] Lawrence Orans, Adam Hils, Jeremy D'Hoinne, Eric Ahlm. Gartner Predicts 2017: Network and Gateway Security, 2016.

[Ponemon, 2016] Hidden Threats in Encrypted Traffic: A Study of North America & EMEA. Independently conducted by Ponemon Institute LLC, 2016.

[Strasák, 2017] František Strasák. Detection of HTTPS Malware Traffic. Bachelor project assignment. Czech Technical University in Prague. 2017. P.49.

[Symantec, 2017] A Technology Brief on SSL/TLS Traffic. Symantec Corporation World Headquarters.

[Волков, 2016] В.А. Волков. Об одном из методов атаки на протокол TLS «Young Scientist» • № 5 (32) • may, 2016, с.213-217.

[Лукацкий, 2018] Алексей Лукацкий. Как Cisco анализирует зашифрованный трафик без его расшифрования и дешифрования. January 15, 2018 [онлайн] Gblogs.cisco.com, Доступно: https://gblogs.cisco.com/ru/eta/

Информация об авторах

Тамара Радивилова *– к.т.н., доцент Харьковского национального университета радиоэлектроники; пр. Науки 14, 61166, Харьков, Украина; e-mail:: tamara.radivilova@gmail.com.*

Основные области научных исследований: самоподобные и мультифрактальные временные ряды, телекоммуникационные системы, управление трафиком, информационная безопасность

Annex for papers written in Russian

Methods of analysis encrypted traffic for hidden threats detection

Tamara Radivilova

Abstract*: Attackers and malicious software uses encrypted protocol SSL/TLS to perform unauthorized activity, which creates problems for intrusion detection. There are two approaches to intrusion detection in encrypted traffic: without decrypting it and after decrypting it. The analysis of the main methods of SSL/TLS protocol traffic decryption was carried out. The work presents methods and technologies for detecting malicious activity in encrypted traffic, which are used by leading companies. Also, a method for intercepting and decrypting traffic transmitted over SSL/TLS, which can be used for listening to the network remotely, offers a way to decrypt the transmitted data in almost real-time mode.*

Keywords*: protocol SSL/TLS, threats, vulnerability, Intrusion Detection Systems, decryption methods.*

ОСОБЕННОСТИ ЭВОЛЮЦИИ ПРОЦЕССА УПРАВЛЕНИЯ РИСКАМИ В ИТ ПРОЕКТАХ

Снежана Гамоцкая, Александра Василевская

***Аннотация**: Данная статья посвящена проблеме управления рисками в ИТ-проектах. На базе стандарта ANSI PMBOK были рассмотрены особенности процесса управления проектными рисками. Анализ выполнен по трем последним редакциям стандарта, проанализирована общая схема управления рисками проекта и кратко рассмотрены основные этапы управления проектными рисками. В результате проведенного исследования можно сделать вывод о том, что технология работы с рисками в ИТ-проектах на данный момент в полном объеме отражает потребности по идентификации, анализу, классификации и мониторингу рисков. Соответственно, все изменения в ближайшее время будут сосредоточены на детализации и конкретизации отдельных этапов, а не на изменении количества и сути этапов управления.*

***Ключевые слова**: IT-проект, проектный менеджмент, проектные риски, управление рисками.*

Введение

На сегодняшний день для организаций, занимающихся разработкой программных продуктов, очень важной является проблема повышения уровня качества выполнения своих проектов. И особенно важным становится отслеживание превышения продолжительности, стоимости и обеспечения качества выполнения работ проекта. Неблагоприятные события, которые приводят к возникновению проблем проекта, достаточно трудно отследить и сложно спрогнозировать. Неопределенность порождается различными факторами, которые можно условно объединить в несколько групп:

- полное или частичное отсутствие информации о внутренних или внешних факторах проекта;
- наличие ложной, недостоверной информации о внутренних или внешних факторах проекта;
- недостаточно высокая квалификация специалиста, отвечающего за аналитически-прогнозные расчеты;
- замена объективной информации субъективным восприятием этой информации.

Риск порождается, в основном, неопределенностью внешней среды, поскольку внутренние

факторы проекта легче поддаются мониторингу, анализу и коррекции.

Процесс разработки программного обеспечения является рискованным процессом – жизненный цикл программного проекта (SDLC) постоянно подвергается рискам, от начала проекта до окончательной сдачи программного продукта. Каждая фаза SDLC имеет различные наборы угроз, которые могут помешать успешному завершению процесса разработки. Для управления рисками должным образом, необходимо адекватное понимание процесса разработки программного обеспечения, проблем, рисков и причин их возникновения. Таким образом, первый шаг в управлении рисками заключается в их определении.

Если мы не будем управлять рисками, они начнут управлять нами... Такое утверждение не подлежит сомнению. Не существует вида человеческой деятельности, в которой бы не возникали риски того или иного вида. С другой стороны, любое действие, связанное с риском, должно быть целенаправленным, иначе оно теряет всякий смысл своего существования. Избежать неопределенности, которая и приводит к возникновению рисков, в проектной деятельности нет никакой возможности, поскольку такая неопределенность является одним из элементов объективной действительности.

Как следствие, в литературе приводится много существующих сегодня перечней факторов риска, но большинство из этих списков относительно короткие и слишком обобщенные. Однако, ни один из исследователей не может отрицать тот факт, что факторы риска, возникающие при разработке программного обеспечения, непрерывно меняются с течением времени и появлением новых инструментов и технологий. Кроме того, при формировании каждого из таких списков не исследуются потенциальные факторы риска, которые могут возникнуть на различных этапах SDLC. Таким образом, большая часть выявленных рисков являются общими для всех этапов SDLC, но полного перечня рисков с их привязкой к этапам SDLC в настоящее время не сформировано [Гамоцька, 2016].

Целью написания статьи является анализ стандартов ANSI PMBOK для выявления направлений дальнейшей эволюции процессов управления проектными рисками.

Структура процесса управления проектными рисками

Управление рисками тесно интегрировано в жизненный цикл любого проекта, и выполняется постоянно. Хорошо известно, что риски перейдут в категорию проблем только при отсутствии эффективного управления.

Любая функция управления состоит из пяти относительно самостоятельных видов управленческой деятельности: планирования, организации, координации, активизации, контроля.

Достижение результата каждого предыдущего вида деятельности является необходимым условием выполнения следующего. Эти пять видов выполняются один за другим, пока данная функция не будет полностью реализована. Степень полноты реализации функции управления зависит от комплексности управленческой деятельности [Чорна, 2003].

Возможны различные комбинации планирования, организации, координации, активизации, контроля. Иногда выполняется укрупнение функций за счет объединения нескольких функций управления в один этап. Иногда, наоборот, одна функция управления реализуется в течение нескольких этапов управленческой деятельности.

В прошлом столетии уже существовали технологии по управлению рисками, хотя они были относительно простыми, даже примитивными, по сравнению с существующими в наше время. Как следует из ANSI PMI PMBOK [PMBOK, 1996], управление рисками проекта сначала разделяли на четыре основных этапа (шага): определение перечня рисков (идентификация), количественное определение рисков, планирование и разработка средств реагирования на риски, контроль реагирования на риски, что отражено на рисунке 1.

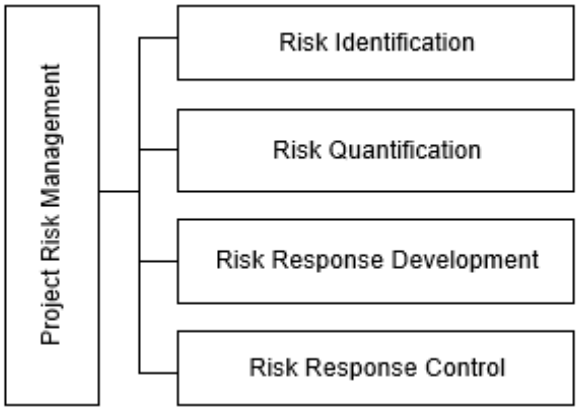

Рисунок 1. Управление рисками ANSI PMI PMBOK 4th Edition [PMBOK, 1996]

Несмотря на то, что рекомендованный алгоритм достаточно простой, он позволял девальвировать последствия рисков, которые возникали в процессе выполнения проекта. Собственно, это было больше устранение и/или смягчение последствий возникновения рисков, чем их предотвращение. Но и такой уровень управления рисков позволял эффективно решать проблемы превышения запланированных ресурсов, времени и стоимости в проекте. Суть его можно описать цепочкой «вижу проблему – определяю ее характеристики – подбираю решение – проверяю результат». Именно наличие пункта проверки устранения влияния риска на проект (risk response control) и дает нам возможность говорить о эффективности принятых и реализованных решений.

Усложнение технологии управления рисками происходило последовательно, в том числе и из-за изменения подхода к разработке программных продуктов, что и демонстрируют нам различные редакции ANSI PMI PMBOK. Уже в 2000 году, при выходе 5й редакции стандарта, алгоритм управления рисками проекта был расширен (рис.2). На рисунке 2 изображены этапы процесса управления рисками в проекте: темно-серым цветом на нем выделено принципиально новые этапы, по сравнению с редакцией 4 (Risk Management Planning, Qualitative Risk Analysis), а светло-серым те, которые уже имелись в прошлой редакции стандарта, но существенно изменились (Risk Monitoring and Control).

Согласно новым рекомендациям перед идентификацией рисков теперь должно проводиться общее планирование управления рисками проекта. Также из количественного анализа рисков выделили проведение качественного анализа, то есть перед оценкой значения риска нужно определить, проанализировать и охарактеризовать все риски, которые могут возникнуть в процессе работы над проектом. Кроме этого, претерпел изменения завершающий этап, который заключается в контроле значений рисков и выполнении всех запланированных действий по реакции на возникшие риски, и как эти действия повлияют на предотвращение или смягчение негативного влияния рисков на результат выполнения проекта. Теперь предполагается не просто проверять результат воздействия, а следить за развитием риска, то есть, проводить его мониторинг.

Риски проекта влияют на все взаимосвязанные ограничения проекта, такие как содержание, качество, расписание, бюджет, ресурсы. Для достижения запланированных результатов и получения при завершении проекта качественного продукта, необходимо эффективно управлять изменениями в проекте и своевременно реагировать на возникшие отклонения. Применение логически сгруппированных процессов управления рисками, объединенных в группы, позволяет оптимизировать этот процесс.

Совсем недавно вышла последняя, шестая, редакция ANSI PMI PMBOK. Вопрос управления рисками остается актуальным, и в данной редакции его также не обошли вниманием. Были внесены изменения в рекомендуемый алгоритм управления рисками, хотя эти изменения нельзя назвать кардинальными (рис.3).

В редакцию стандарта 2017 года не вносились новые этапы, не менялись уже существующие, а были уточнены определенные шаги и несколько по-другому расставлены акценты. В частности, анализ ответа (реакции) на риск разделяется на два последовательных шага - отдельно выделено формирование плана ответов (реакций) на риски, которое теперь отделяется от, собственно, реакции на риски.

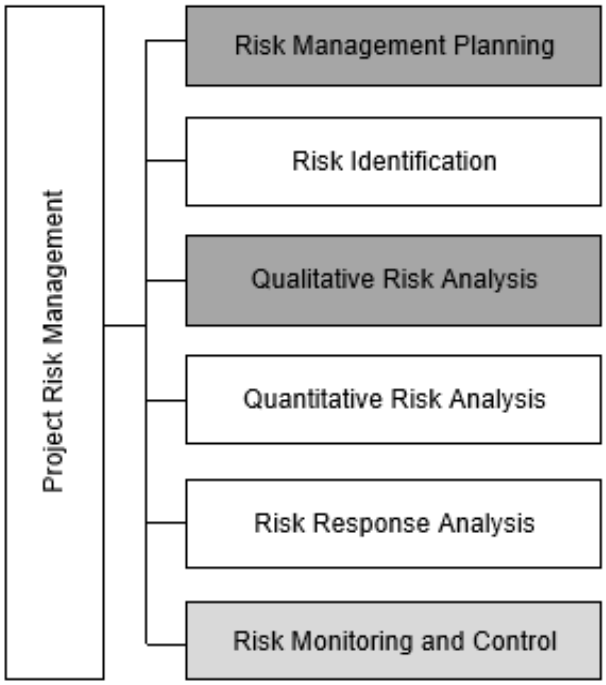

Рисунок 2. Управление рисками ANSI PMI PMBOK 5th Edition [PMBOK, 2012]

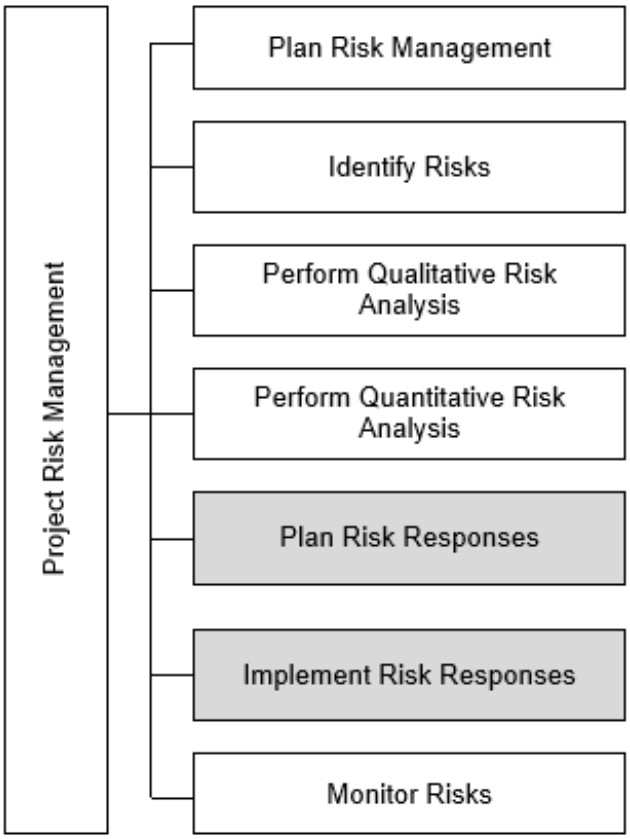

Рисунок 3. Управление рисками ANSI PMI PMBOK 6th Edition [PMBOK, 2017]

Итак, исходя из этого, можно сделать вывод, что в последние годы процедура работы с рисками достигла такого уровня, который характеризует ее как достаточно завершенную. Это значит, что в дальнейшем изменения в процедуре управления будут иметь операционный характер и не окажут большого влияния на сам алгоритм.

Для лучшего понимания принципов управления рисками рассмотрим подробнее этапы этого процесса. Согласно последней, шестой, версии ANSI PMBOK, процесс управления рисками проекта состоит из семи шагов.

Шаг I. Планирование управления рисками

На этом этапе выполняется подготовительная работа по управлению рисками проектной деятельности, определяется базовый перечень рисков проекта, который будет скорректирован (расширен, изменен, уточнен) на следующем этапе.

Шаг II. Выявление (идентификация) рисков

Как один из методов определения рисков целесообразно применить механизм структурной декомпозиции работ (СДР) проекта, который формально относится к сфере управления содержанием и границами проекта [Верес, 2003]. После завершения планирования управления рисками и их идентификации, все работы, определенные с помощью СДР, должны полностью описывать содержание проекта и его ограничения. Это позволит определить все возможные точки возникновения рисков.

В перечень, составляющийся на этом этапе, нужно внести как риски, которые могут проявиться в процессе выполнения отдельных работ проекта, так и такие, которые связаны со сроками выполнения отдельных этапов и проекта в целом. Также оценивается возможное влияние зарегистрированных в списке рисков на финансирование проекта.

Шаг III. Выполнение анализа рисков

На этом этапе проводится качественный анализ, целью которого является определение, классификация и учет рисков. Качественный анализ, как правило, проводится еще на стадии разработки бизнес-плана проекта и включает следующие составляющие:

- матрица вероятностей и влияния;
- оценка качества данных о рисках;
- категоризация рисков;
- оценка срочности рисков;
- ранжирование рисков проекта относительно других проектов;
- список приоритетов рисков, и оценка вероятности их возникновения и воздействия;

- списки рисков для анализа и реагирования;
- списки рисков для наблюдения.

Шаг IV. Выполнение количественного анализа рисков

Это этап количественного анализа. Его целью является измерение определенных на предыдущем этапе рисков.

Количественная оценка риска является, по сути, дополнением к качественной оценке. В результате количественного анализа риска получают числовые значения величины отдельных рисков, а также значение риска проекта в целом. Риск может определяться как в абсолютных, так и в относительных величинах. Измерение степени риска в абсолютных величинах целесообразно применять для характеристик отдельных видов потерь, а в относительных - при сравнении прогнозируемого уровня потерь с реальным уровнем.

Среди распространенных методов количественной оценки степени риска можно отметить такие, как:

- анализ ожидаемого денежного значения;
- метод достоверных эквивалентов;
- метод исторических симуляций;
- статистический метод;
- метод анализа целесообразности затрат;
- метод экспертных оценок;
- метод использования аналогов;
- метод Монте-Карло;
- дерево решений;
- деревья событий;
- деревья отказов;
- анализ ожидаемого значения;
- анализ сценариев;
- распределение вероятностей;
- анализ чувствительности;
- обновление реестра рисков;
- вероятностный анализ проекта;
- вероятность достижения целей по срокам и стоимости;
- список количественно определенных рисков с расставленными приоритетами;
- тенденции результатов количественного анализа рисков и другие.

В общем случае можно сказать, что величина риска напрямую зависит от вероятности его возникновения и угрозы, которую этот риск представляет для процесса протекания и результата проекта.

Шаг V. Планирование и разработка средств реагирования на риски.

Данный этап обеспечивает непрерывность процесса управления рисками. Он структурирует и обобщает полученную на предыдущих шагах информацию и объединяет полученные данные в план реагирования на риски, проявивших себя в данный момент.

Следует заметить, что в каждом проекте, кроме явных, обязательно есть неизвестные (неопределенные) риски, доля которых в общем объеме рисков зависит от того, в какой области реализуется проект. В проектах, связанных с разработкой программного обеспечения, эта доля достаточно высока. Для того, чтобы нивелировать результаты таких рисков, в резерв управления проектом закладываются определенные финансовые и временные ресурсы. Но, в случае, если реагирование на риски возможно только после их проявления в проекте, затраты на компенсацию потерь будут достаточно высокими. Гораздо эффективней опережение событий, получение информации о потенциальном появлении неизвестного риска заранее, когда есть возможность скорректировать процесс разработки и не возвращаться к уже выполненным работам повторно.

Шаг VI. Внедрение разработанного плана реакции на риски.

Данный этап является логическим продолжением предыдущего, на нем воплощаются в практику проведенные ранее теоретические разработки. Реагирование на риски нельзя рассматривать отдельно от анализа и планирования ответов на них, поэтому эти этапы тесно связаны между собой.

Шаг VII. Мониторинг рисков.

На этом этапе выполняется текущий контроль выполнения антирисковых мероприятий, которые были сформулированы и реализованы на двух предыдущих шагах. В случае, если средства реагирования на риски оказываются неэффективными, в них вносятся коррективы, то есть выполняется циклическое возвращение на шаг V.

Также на этом этапе выполняется завершающий мониторинг и контроль, и подводятся итоги по проведенным антирисковым мероприятиям. Выполняется анализ с целью подготовки данных для дальнейшего использования в следующих итерациях текущего проекта. Накопленная статистическая информация может быть использована для последующих проектов, что позволит повысить эффективность процесса их выполнения.

Организация мониторинга рисков дает возможность корректировать текущую деятельность в соответствии с ситуацией на данный момент, а не только влиять на результаты возникновения и развития рисков. Эффективность системы контроля рисков в целом значительно зависит от эффективности системы их мониторинга.

Выводы

Проектный тип управления приобретает все большее значение в области информационных технологий вследствие уникальности выполняемых в каждом проекте работ, которые невозможно унифицировать, быстрого обновления продукции, необходимости эффективной координации ресурсов для достижения цели, а также четко определенных границ во времени для производства каждого продукта отрасли.

В статье мы рассмотрели особенности процесса управления рисками в ИТ-проектах на базе стандарта ANSI PMI PMBOK и провели анализ тенденции развития технологии за последние двадцать лет. По состоянию на текущий момент технология работы с проектными рисками имеет достаточно высокий уровень развития. Это позволяет сказать о том, что содержательное наполнение алгоритма работы с рисками в ИТ-проектах в полном объеме отражает потребности по идентификации, анализу, классификации и мониторингу рисков и все разработки в ближайшее время будут сосредоточены на детализации и конкретизации отдельных этапов, а не на их изменении.

Bibliography

[PMBOK, 2017] A Guide to the Project Management Body of Knowledge (PMBOK ® Guide) 6th Edition. PMI, 2017. 537p.

[PMBOK, 2012] A Guide to the Project Management Body of Knowledge (PMBOK ® Guide) 5th Edition. PMI, 2012. 586p.

[PMBOK, 1996] A Guide to the Project Management Body of Knowledge (PMBOK ® Guide) 4th Edition. PMI, 1996. 216p.

[Верес, 2003] Верес О.М. Управління ризиками в проектній діяльності / О. М. Верес, А. В. Катренко, І. В. Рішняк, В. М. Чаплига // Вісник Національного університету "Львівська політехніка". – 2003. – № 489 : Інформаційні системи та мережі. – С. 38-49

[Hijazi et al, 2014] Haneen Hijazi, Shihadeh Alqrainy, Hasan Muaidi, Thair Khdour. RISK FACTORS IN SOFTWARE DEVELOPMENT PHASES // European Scientific Journal January. – 2014, vol.10, No.3. – p.p.213-232.

[Гамоцька, 2016] Гамоцька С.Л. // Матеріали IV Міжнародної науково-практичної конференції «Обчислювальний інтелект (Результати, проблеми, перспективи)» (ComInt). – Київ. – с.211-212

[Чорна, 2003] Чорна М.В. Проектний аналіз. - Харків: Консум, 2003. — 228 с.

Authors' Information

Снежана Гамоцкая *– Киевский национальный университет имени Тараса Шевченко; аспирант. e-mail: GamotskaSL@i.ua*

Основные направления научных исследований: проектные риски, автоматизация проектирования, современные web-технологии

Александра Василевская *– Киевский национальный университет имени Тараса Шевченко; ассистент. e-mail: vasilevskaya.alexandra@gmail.com*

Основные направления научных исследований: нейронные сети, обработка изображений, искусственный интеллект, автоматизация проектирования, автоматизация систем обработки информации, энергосбережение, энергоэффективность, гидроаэродинамические системы.

Features of the Evolution Risk Management Process in IT-Projects

Snezhana Gamotskaya, Aleksandra Vasylevskaya

Abstract*: This article is devoted to the problem of risk management in IT projects. Based on the ANSI PMBOK standard, the features of the project risk management process were considered. The last three editions of the standard and the overall risk management scheme of the project were analysed. The main stages of project risk management were briefly reviewed. As a result of the research we have made a conclusion that the technology of working at risks in IT-projects at the present moment fully reflects the needs for identification, analysis, classification and monitoring of these risks. Accordingly, all changes in the close future will focus on detailing and specifying individual stages, rather than changing them.*

Keywords*: IT-project, project management, project risks, risk management.*

ITHEA Keywords*: K.6.1 Management of Computing and Information Systems - Project and People Management.*

ON TWO REPRESENTATIONS OF CONCURRENT PROGRAMS

Taras Panchenko, Sunmade Fabunmi

Abstract: *In the paper we show that under generic conditions, for each fixed input data, the observable behavior of a shared memory concurrent program in the Interleaving Parallel Composition Language (ICPL) extended with the Start and Join operations, which allow creating threads and waiting till thread completion during runtime, is equivalent to the observable behavior of a certain execution of a parallel composition of a fixed number of copies of threads from a fixed finite set (power representation), each of which are representable in the pure ICPL language. This result can be useful for formal verification of concurrent programs which allow dynamic creation of threads during runtime, since it reduces latter problem to the problem of verifying correctness of a certain concurrent program with a fixed set of threads and no dynamic thread creation for each fixed input data.*

Keywords: *concurrent programming, Interleaving Parallel Composition Language, formal methods, software verification.*

ITHEA Keywords: *D.1.3 Concurrent Programming, D.2.4 Software/Program Verification.*

ACM Classification Keywords: *F.3.1 Theory of Computation - LOGICS AND MEANINGS OF PROGRAMS - Specifying and Verifying and Reasoning about Programs, D.2.4 Software - SOFTWARE ENGINEERING - Software/Program Verification.*

Introduction

A large number of software systems used today (e.g. operating systems, database management systems, server software, etc.) are implemented as multithreaded programs with shared memory. Many of such systems form a part of the critical infrastructure of various private and public organizations. Their safety, security, and reliability are of paramount importance, which makes it desirable to obtain the strongest possible guarantees of correctness of their implementation. Such guarantees can be provided by formal methods of software development and verification. Formal methods have been in development for over 50 years, starting from the works of Floyd and Hoare on methods of proving (partial) correctness of sequential programs. Still the task of proving correctness of sequential programs remains difficult and far from being widely applied in software development. The problem of verification of concurrent programs is even more difficult. There are many existing approaches to formal verification of concurrent programs [Ashcroft, 1975; Hoare, 1985; Owicki, 1976; Jones, 1981; Jones, 1983; Xu,

1997; Harel, 1997; Pnueli, 1977; Lamport, 1993; Chandy, 1988; Lamport, 1994; Manna, 1992], but there not so many practical results, most notable of which include formal verification of small specialized microkernels and hypervisors such as seL4 and CertiKOS, which, however, have a far lower complexity than the widely used concurrent software.

Thus, the problem of finding scalable ways of verification of concurrent software is still important.

One way to deal with it is to reduce the problem of verification of programs which, potentially, have a very rich set of runtime behaviors which is difficult to analyze, e.g. the programs which can dynamically create and terminate threads, to a series of problems of verification of programs which have a much simpler, easier to analyze set of behaviors, e.g. programs with a fixed set of threads which cannot create or terminate threads during runtime.

In this paper we propose this kind of reduction. More specifically, we focus on programs expressible in the Interleaving Parallel Composition Language [Panchenko, 2008], which is a convenient formal language for expressing and verifying concurrent software. It was used for formal modeling of real world systems such as a distributed presentation software Infosoft e-Detailing [Kartavov, 2015] and proving their correctness. In this paper we prove that under generic conditions, for each fixed input data, the observable behavior of a shared memory concurrent program in IPCL extended with the Start and Join operations (which allow creating threads and waiting till thread completion during runtime), is equivalent to the observable behavior of a certain execution of a parallel composition of a fixed number of copies of threads from a fixed finite set (which we call the power representation), which are representable in the pure IPCL language and have a much simpler and predictable behavior which is easier to check for correctness.

This allows us, under generic conditions, to reduce the problem of verification of correctness of programs in IPCL with dynamic thread creation and join – to the problem of verification of pure IPCL programs, the availability of this kind of reduction allows us to focus on the simpler cases of concurrent software verification, for which specialized efficient approaches are available, and reduce to them the problems of verification of real-world, complex, large-scale concurrent software.

Interleaving Parallel Composition Language with Start and Join Operations

The syntax and semantics of the pure Interleaving Parallel Composition Language (IPCL) were described in [Panchenko, 2008]. Applications of ICPL were given in [Panchenko, 2006; Panchenko, 2004; Kartavov, 2015, Kartavov2, 2015; Polishchuk, 2015]. The syntax and semantics of the Interleaving Parallel Composition Language with Start and Join operations were proposed and described in [Panchenko, 2017].

Here we recall the main definitions.

In general, IPCL is a family of languages with formally defined syntax and semantics which allow expressing concurrent programs with shared memory, execution of which can be interpreted as interleaving execution of a set of sequential threads.

Basic syntax is defined by the following BNF:

$$P ::= \bar{x} := \bar{e} \mid P_1; P_2 \mid \textbf{if } b \textbf{ then } P_1 \textbf{ else } P_2 \mid \textbf{while } b \textbf{ do } P \mid P_1 || P_2$$

where

- P_i denotes programs,
- $\bar{e}$ denotes an expression(s) which evaluates to a value(s) (e.g. a number, string, complex data structure, etc.),
- $\bar{x}$ denotes a variable name(s),
- $:=$ denotes the atomic vector assignment operator,
- ; denotes the sequential execution operator,
- **if – then – else** and **while – do** are the usual sequential branching and loop operators,
- || is the composition of parallel execution of two threads.

Formal and detailed definition of semantics of the language can be found in [Panchenko, 2008].

If P is an IPCL program, we denote as P^n (power operator), where *n* is a natural number, the parallel composition of *n* copies of *P*, i.e. a program which performs interleaving execution of *n* threads each of which executes in accordance with *P*.

Pure IPCL can be enriched with the dynamic thread creation (*start*) and joining of threads (*join*) operations which model common multithreading constructs:

$$P ::= \bar{x} := \bar{e} \mid P_1; P_2 \mid \textbf{if } b \textbf{ then } P_1 \textbf{ else } P_2 \mid \textbf{while } b \textbf{ do } P \mid P_1 || P_2 \mid start(P) \mid join(id)$$

Informally, the $start(P)$ operation takes one argument – a program code *P* and creates a thread which executes a body of *P*. The created thread receives its unique identifier. The $start(P)$ operation can be invoked from any thread and returns immediately after creation of the new thread. Afterwards, execution of both the invoking thread and the newly created thread continues in the arbitrary interleaving fashion.

The $join(id)$ operation takes as an argument a scalar value – an identifier (*id*) of a thread previously created using the $start(P)$ operation and suspends execution of the thread which invokes $join(id)$ until the thread with the given *id* terminates (a thread terminates, if its execution reaches the end of the thread's program code). Afterwards, $join(id)$ resumes execution of the thread which has invoked it.

Formal definition of semantics of the $start(P)$ and $join(id)$ operations can be found in [Panchenko, 2017].

The Main Result

The following theorem uses the notation and terminology defined in [Panchenko, 2008] and [Panchenko, 2017].

Theorem 1. Let *P* be an IPCL program with *start* and *join* operations which takes no input data. Assume that each execution of *P* is terminating. Then the set of lengths of executions of *P* is bounded.

Proof.

Since *P* takes no input data, all executions of *P* start at one program execution state which we denote as q_0 and set of executions of *P* is the set of paths in the labeled transition system, which we denote as *L*, which describes the operational semantics of the IPCL program *P* (with *start* and *join* operations) which starts at *q*. Since at each point of program execution there can be at most finite amount of existing processes and all operations performed by individual processes available in the IPCL language are finitely non-deterministic, the outdegree of any node reachable from q_0 in *L* is finite (i.e. the program can progress from a state *q* to a state from at most finite set of possible successor states that depends on *q*). Suppose that the set of lengths of executions of *P* is unbounded. Then the set of states reachable from q_0 is infinite. Then Konig's lemma implies that *L* has an infinite run starting from q_0 which corresponds to some non-terminating execution of *P*. This contradicts the assumption that each execution of *P* is terminating. Thus, the set of lengths of executions of *P* is bounded.

Theorem is proved.

Corollary. Let *P* be an IPCL program with *start* and *join* operations which takes no input data. Assume that each execution of *P* is terminating.

Then there exist natural number *K* and IPCL programs without *start* and *join* operations $P_1, \dots, P_n$ such that the set of traces of executions of *P* is a subset of the set of traces of executions of $P_1^K \,||\, \dots \,||\, P_n^K$.

Proof (sketch).

By Theorem 1 the set of lengths of executions of *P* is bounded by some natural number *K*. Then the set of numbers of processes created during each execution of *P* is bounded from above by *K*. Let $P_1, \dots, P_n$ be all procedures in the program *P* in which the *start* operation is replaced by a nondeterministic choice of a number from the set {1,...,*K*}. The nondeterministic choice can be modeled by a set of auxiliary processes:

$$A_i: x := n; x := x + 1; n := x;$$

where *n* is a common global variable storing the result and *x* is a local variable for each auxiliary process.

Then it is easy to see that the trace of each execution of *P* corresponds to the trace of some execution of $P_1^K || \dots || P_n^K$ – namely the one in which the nondeterministic choice functions returned the same values as the corresponding *start*() function invocations in *P* (which, as we have shown above, are in range 1,2,...,*K*). Thus, the set of traces of executions of *P* is a subset of the set of traces of executions of $P_1^K || \dots || P_n^K$.

Corollary is proved.

Conclusion

We have shown that under very general conditions, the observable behavior of a shared memory concurrent program in the IPCL language extended with the Start and join operations for a fixed input data is equivalent to the observable behavior of a certain execution of a parallel composition of a fixed number of copies of threads from a fixed finite set (power representation), each of which are representable in the pure IPCL language. This obtained result can be useful for solving the problem of verifying the correctness of concurrent programs which allow dynamic creation of threads. The obtained result reduces this problem to the problem of verification of correctness of a certain concurrent program with a fixed set of threads and no dynamic thread creation for each fixed input data.

We plan to apply this reduction to developed concurrent software systems in the future work.

Acknowledgement

We would like to thank scientific advisors and colleagues for the fruitful discussions held and to appreciate their work. Particularly, Mykola Nikitchenko, Ievgen Ivanov, Dmytro Bui, Volodymyr Redko, and others.

Bibliography

[Ashcroft, 1975] Ashcroft E.A. Proving assertions about parallel programs // Journal of Computer and System Sciences. -- 1975. -- No. 10. -- pp. 110--135

[Chandy, 1988] Chandy K.M., Misra J. Parallel Program Design: A Foundation. -- Reading, MA: Addison-Wesley Publishing Company, 1988. -- 493 p.

[Harel, 1997] Harel D., Pnueli A. On the development of reactive systems // Apt K.R. (ed.) Logics and models of concurrent systems, NATO ASI Series, Vol. F13. -- Springer-Verlag, 1985. -- pp. 477--498

[Hoare, 1969] Hoare, C.A.R. An Axiomatic Basis for Computer Programming. Communications of the ACM. Vol. 12, no. 10, 1969, pp. 576--583

[Hoare, 1985] Hoare C.A.R. Communicating Sequential Processes. -- Prentice Hall International, 1985. -- 238 p.

[Jones, 1981] Jones C.B. Development Methods for Computer Programs Including a Notion of Interference: DPhil. Thesis. -- Oxford University Computing Laboratory, 1981. -- 315 p.

[Jones, 1983] Jones C.B. Specification and Design of (Parallel) Programs // Information Processing Letters: IFIP Information Processing'83 (In IFIP 9th World Congress). -- 1983. -- pp. 321--331

[Kartavov, 2015] Kartavov, M., Panchenko, T. and Polishchuk, N. Properties Proof Method in IPCL Application To Real-World System Correctness Proof. International Journal "Information Models and Analyses". Sofia, Bulgaria, ITHEA. Vol. 4, No. 2, 2015, pp. 142--155

[Kartavov2, 2015] Kartavov, M., Panchenko, T. and Polishchuk, N. Infosoft e-Detailing System Total Correctness Proof in IPCL [in Ukrainian]. Bulletin of Taras Shevchenko National University of Kyiv. Series: Physical and Mathematical Sciences, No. 3, 2015, pp. 80--83

[Lamport, 1993] Lamport L. Verification and Specification of Concurrent Programs // deBakker J., deRoever W., Rozenberg G. (eds.) A Decade of Concurrency, Vol. 803. -- Berlin: Springer-Verlag, 1993. -- pp. 347--374

[Lamport, 1994] Lamport L. The temporal logic of actions // ACM Transactions on Programming Languages and Systems. -- 1994. -- Vol. 16, No. 3. -- pp. 872 -- 923

[Manna, 1992] Manna Z., Pnueli A. The Temporal Logic of Reactive and Concurrent Systems, Specification. -- Berlin: Springer-Verlag, 1992. -- 427 p.

[Nipkow, 2003] Nipkow, T., Paulson, L. C., Wenzel, M. Isabelle/HOL: A Proof Assistant for Higher-Order Logic. Springer, 2003, 226 p.

[Nikitchenko, 1998] M. Nikitchenko, Technical Report IT-TR: 1998-020. A Composition Nominativej Approach to Program Semantics. Technical University of Denmark, 1998

[Owicki, 1976] Owicki S. and Gries D. An Axiomatic Proof Technique for Parallel Programs // Acta Informatica. -- 1976. -- Vol. 6, No. 4. -- pp. 319--340

[Ostapovska, 2016] Ostapovska, Yu., Panchenko, T., Polishchuk, N. and Kartavov, M. Correctness Property Proof for the Banking System for Money Transfer Payments [in Ukrainian]. Problems of Programming. No. 2-3. 2016. pp. 119--132

[Panchenko, 2004] T. Panchenko, The Methodology for Program Properties Proof in Compositional Languages IPCL [in Ukrainian], in Proceedings of the International Conference “Theoretical and Applied Aspects of Program Systems Development” (TAAPSD’2004), 2004, pp. 62–67

[Panchenko, 2006] T. Panchenko, Compositional Methods for Software Systems Specification and Verification [in Ukrainian]. (Panchenko, 2006 Thesis), Taras Shevchenko National University of Kyiv, 2006, 177 p.

[Panchenko, 2007] T. Panchenko, Parallel Addition to Shared Variable Correctness Proof in IPCL [in Ukrainian], Bulletin of Taras Shevchenko National University of Kyiv. Series: Physical and Mathematical Sciences, no. 4, 2007, pp. 187–190

[Panchenko2, 2007] T. Panchenko, Simplified State Model for Properties Proof Method in IPCL Languages and its Usage with Advances [in Ukrainian], in Proceedings of the International Scientific Conference “Theoretical and Applied Aspects of Program Systems Development” (TAAPSD’2007), 2007, pp. 319–322

[Panchenko, 2008] T. Panchenko, The Method for Program Properties Proof in Compositional Nominative Languages IPCL [in Ukrainian], Problems of Programming, no. 1, 2008, pp. 3–16

[Panchenko2, 2008] T. Panchenko, Formalization of Parallelism Forms in IPCL [in Ukrainian], Bulletin of Taras Shevchenko National University of Kyiv. Series: Physical and Mathematical Sciences, no. 3, 2008, pp. 152–157

[Panchenko, 2016] Panchenko, T. Application of the Method for Concurrent Programs Properties Proof to Real-World Industrial Software Systems. Proceedings of the International Conference on ICT in Education, Research, and Industrial Applications (ICTERI'2016), pp. 119--128

[Panchenko, 2017] Panchenko, T., Ivanov Ie., Fabunmi S., Trofimenko Ie., Skidonenko A. Extended Dynamic State and Instances Spawn Model in IPCL. Bulletin of Taras Shevchenko National University of Kyiv, Series Physics & Mathematics, No. 4, 2017

[Polishchuk, 2015] Polishchuk, N., Kartavov, M. and Panchenko, T. Safety Property Proof using Correctness Proof Methodology in IPCL. Proceedings of the 5th International Scientific Conference "Theoretical and Applied Aspects of Cybernetics". Kyiv: Bukrek, 2015, pp. 37--44

[Pnueli, 1977] Pnueli A. The temporal logic of programs // Proc. 18th Annual Symposium on the Foundations of Computer Science (Providence). -- New York: IEEE Computer Society Press, 1977. -- pp. 46--57

[Redko, 1978] V. Redko, Compositions of programs and composition programming [in Russian], Programming, no. 5, 1978, pp. 3–24

[Wiedjik, 2006] Wiedijk F. The Seventeen Provers of the World. Foreword by Dana S. Scott. F. Wiedijk (editor), Lecture Notes in Artificial Intelligence, Vol. 3600, Springer-Verlag Berlin Heidelberg, 2006

[Xu, 1997] Xu Q., de Roever W.-P., He J. The Rely-Guarantee Method for Verifying Shared Variable Concurrent Programs // Formal Aspects of Computing. -- 1997. -- Vol. 9, No. 2. -- pp. 149--174

Authors' Information

Taras Panchenko *– Panchenko, 2006, Associate Professor at the Theory and Technology of Programming Department, Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska Street, Kyiv, Ukraine, 01601*

e-mail: tp@infosoft.ua

Major Fields of Scientific Research: Theory and Technology of Programming, Software Engineering, Software Correctness, Formal Methods

Sunmade Fabunmi *– intern at the Faculty of Computer Science and Cybernetics, Taras Shevchenko National University of Kyiv, 64/13, Volodymyrska Street, Kyiv, Ukraine, 01601*

e-mail: sunmadefabunmi@yahoo.com

Major Fields of Scientific Research: Theory and Technology of Programming, Software Engineering, Software Correctness, Formal Methods

TABLE OF CONTENTS