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.

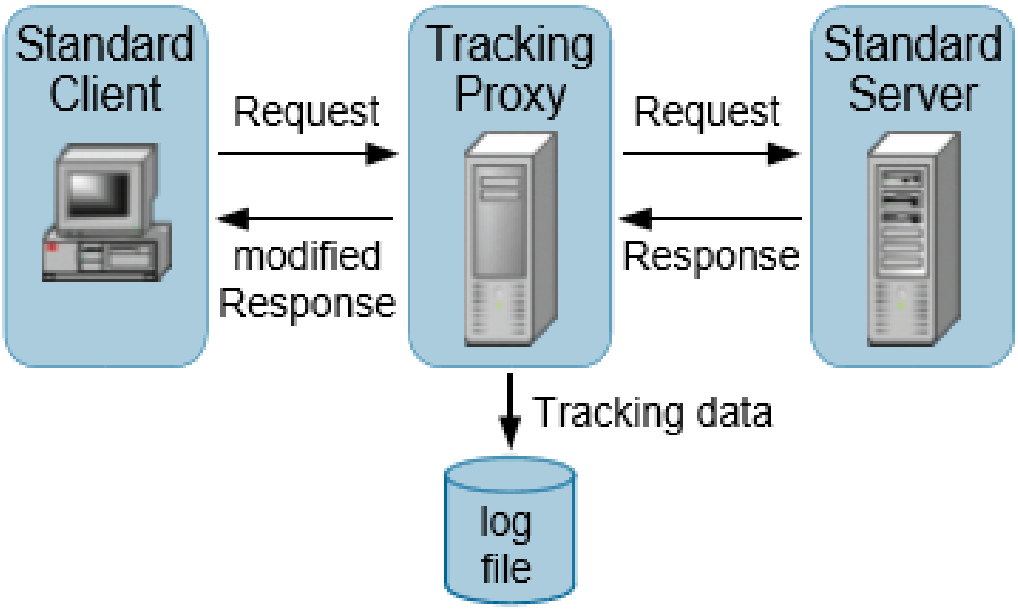

Figure 1. HTTP proxy for inserting JavaScript code into HTML pages for tracking users activity

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.

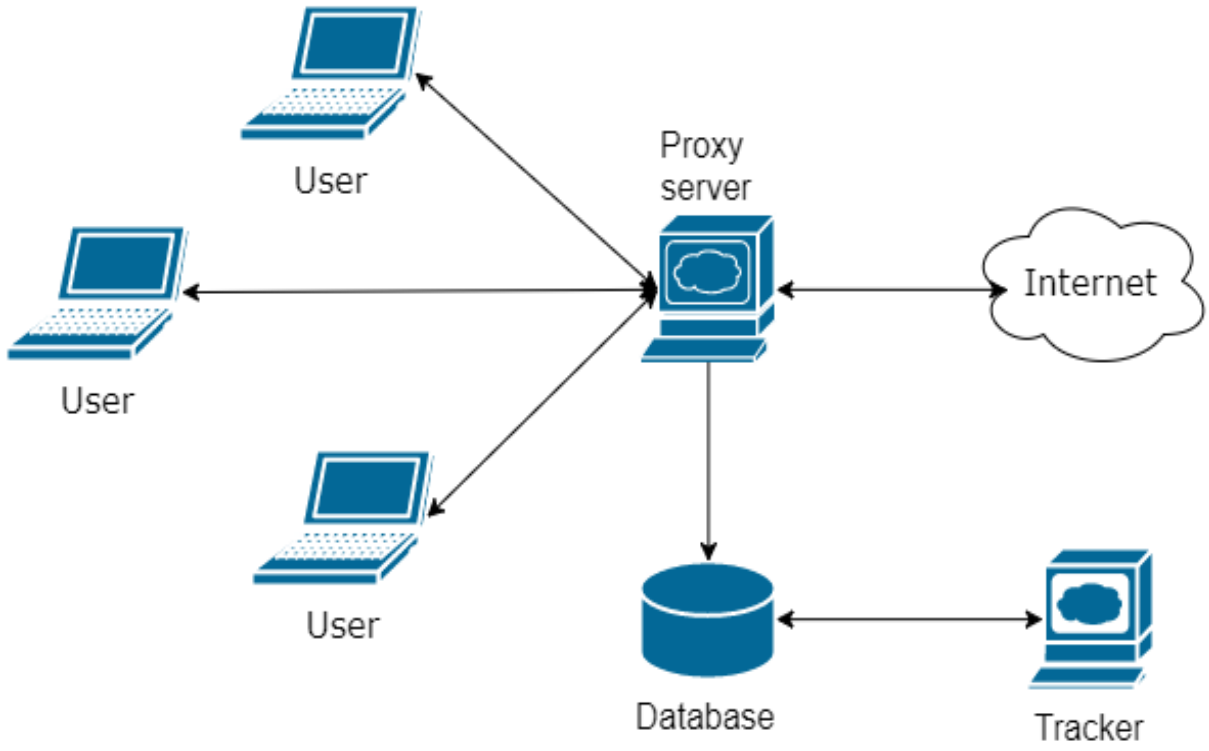

Figure. 3. General scheme of work

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.

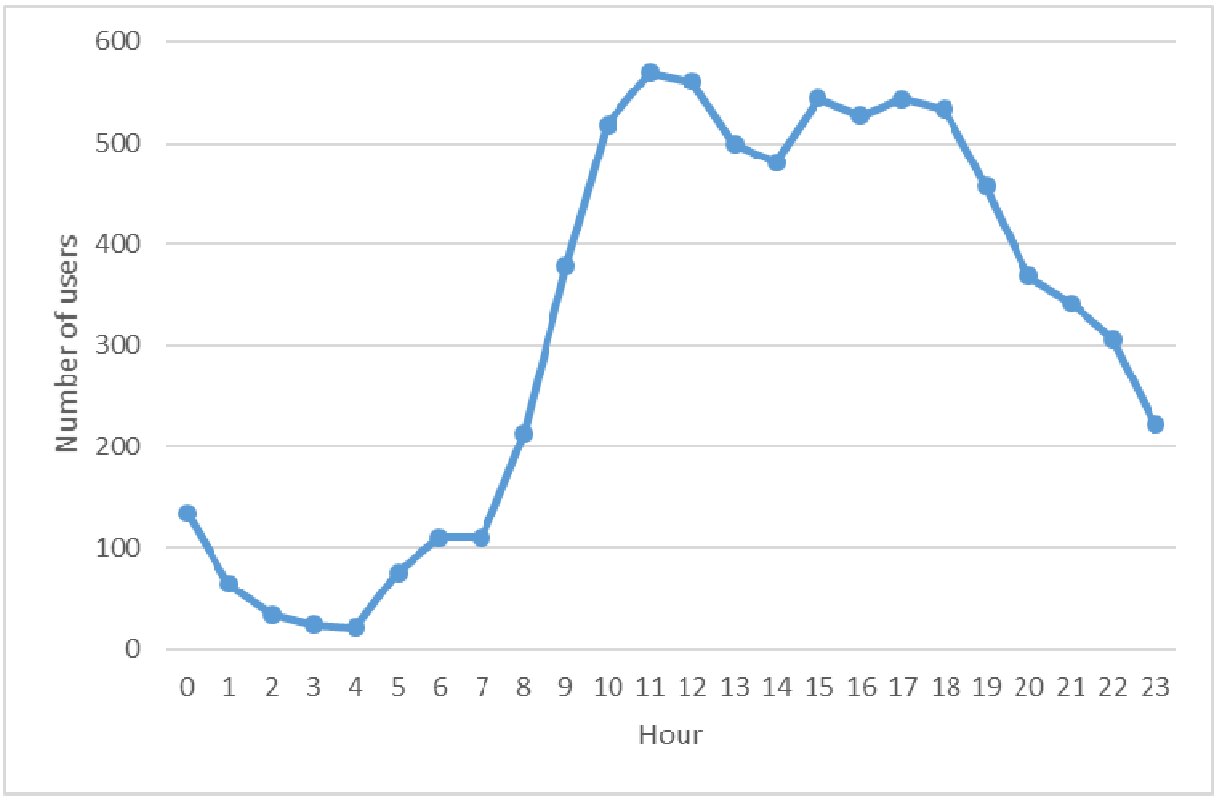

Figure. 4. Ratio between number of users and time3

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.

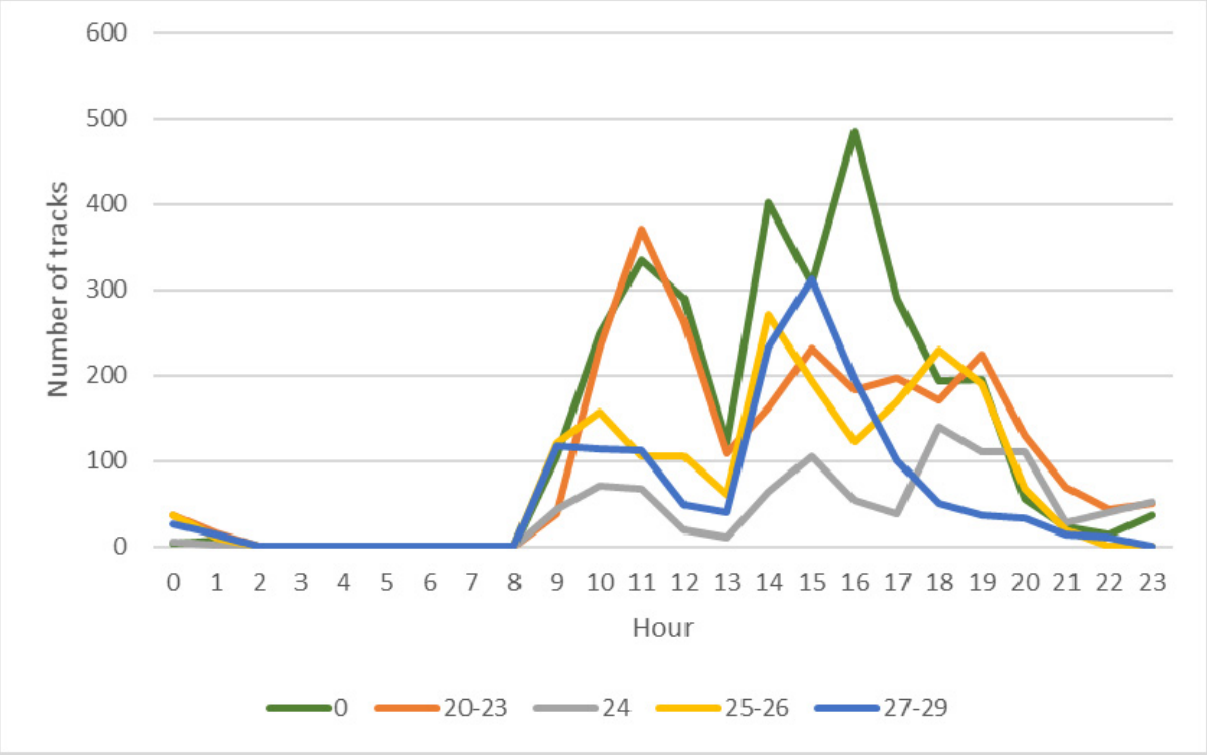

Figure. 5. Ratio between number of tracks and hour

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.