Spring 3-23-2018

WEB PROXY LOG TO DEVELOP A REQUIREMENT BASED RESOURCE ALLOCATION FOR WEB TRAFFIC

Md Baitul Al Sadi
Georgia Southern University, ms12508@georgiasouthern.edu

WEB PROXY LOG TO DEVELOP A REQUIREMENT BASED RESOURCE ALLOCATION FOR WEB TRAFFIC

Md Baitul Al Sadi
Georgia Southern University
Ms12508@georgiasouthern.edu

ABSTRACT

Network packets are busy exchanging information like data from human activities, machine (M2M communication), cellular network, Internet of Things (IoT), business and in many other forms. Network infrastructure plays the role of backbone in the whole communication system. Therefore, it is an integral point of interest not only to analyze network traffic but also to expose security incident or business critical event. The goal of this research is to discover unknown network parameters and event from the network logs, specifically from web proxy log, to explore the relationship between different parameters. This work highlights unique findings and relations among log parameters. For example, how server action varies with SC-byte or CS-byte or RS and relations between the time taken and RS (Content-Type). Furthermore, visualizing the busiest period in terms of bandwidth consumption and available number of live IP addresses, and so on.

Keywords

Network log Analysis, web proxy, resource allocation, bandwidth, cs/sc-content, cs/sc-type

INTRODUCTION

In this era of Information Technology, while globalization takes place, it becomes impossible to imagine a single moment without the internet. The analysis of internet event log is taking more attention to the network analysts as the concept like IaaS (Infrastructure-as-a-Service) is evolving. The IaaS service providers require keeping track of their overall performance regarding indicators including session events, outages, access grants, access failure information, device failure, device malfunctioning, and so on to maintain the Service Level Agreement (SLA) as well as to determine the Key Performance Indicator (KPI). Apart from the view of the service provider, analysis of network logs are also essential for the personnel who deal with network security or network forensic. Although nowadays sensitive information is being transferred by enforcing cryptography, it is still not enough to protect the data from malpractice. Protecting networks requires performing network log analysis prior to and after malicious activities. Pre-analysis aids in preventing attacks or malicious activities while post-analysis helps predict how the incident happened and to advise precautions for prevention.

Web proxy server data can provide the network activity through analysis. As a result, a network administrator can obtain a clear view of the network activity, including user's activity (e.g., what kind of contents or websites are most accessed by the users), resource utilization (in terms of usages of IP addresses and bandwidth), etc. Apart from this, network administrators become aware of security issues such as the ratio of HTTPS traffic versus HTTP traffic, or user-wise access list. Combined, this information can have a significant influence in establishing an accurate resource allocation policy. Simultaneously, an analysis of proxy logs may also help identifying potential security breaches and vulnerabilities of the network.

Network logs can be categorized as a standard web proxy log as outlined by the World Wide Web Consortium (W3C). For this project, logs available on www.honeynet.org were used to investigate concerns related to bandwidth use. Honeynet is a non-profit project devoted to pursue security research on latest attacks and build open source tools to advance network security (honeynet.org, 2017), which makes an ideal source to study bandwidth usage. Honeynet authority collected these logs from a network laboratory. These logs are a combination of production traffic and simulation traffic that is gathered from a web proxy appliance called BlueCoat. The web proxy logs contain columns called date, time, time-taken, c-ip, sc-status, s-action, sc-bytes, cs-bytes, cs-method, cs-uri-scheme, cs-host, cs-uri-path, cs-uri-query, cs-username, s-hierarchy, s-supplier-name, rs(Content-Type), cs(User-Agent), sc-filter-result, sc-filter-category, x-virus-id, s-ip, s-sitename, x-virus-details, x-icap-error-code, and x-icap-error-details. The following Table 1 shows the column name and the detail information of the column of web proxy log. This research work intends to explore the relationships among different internet resources, parameters, services, and protocols. In the result section, several analyses have been shown including how bandwidth differs for various server action for both cases

(server to client and client to server), the amount of various traffic, the distribution of different content type, and resource utilization.

Column Name	Detail
time-taken	Time in milliseconds; duration from the time when the server first receives a request to the time of completion.
c-ip	The IP address of the client who sends the request
sc-status	Http status code like error code
s-action	Server-Action
sc-bytes	bytes: requires to send to connect to the remote server from the client
cs-bytes	bytes: requires to sent to connect from client to remote server
cs-method	Client to Server Method
cs-uri-scheme	Client to Server URI (Unified resource Identifier) scheme
cs-uri-path	Client to Server URI (Unified Resource Identifier) path
cs-uri-query	Client to Server URI (Unified resource Identifier) query
cs-username	Client to Server username
s-hierarchy	Server hierarchy
s-supplier-name	Server supplier name
rs(Content-Type)	Remote Server Content Type
cs(User-Agent)	Client Server (User agent)
sc-filter-result	Server Client-filter-result
sc-filter-category	Server Client filter-category
x-virus-id	Virus or threat description
s-ip	server IP address
s-sitename	Server Site Name
x-virus-details	Detain info about virus or threat
x-icap-error-code	Internet Content Adaptation Protocol (ICAP) error code

Table 1. Column details (https://technet.microsoft.com, 2017)

SIGNIFICANCE OF WEB PROXY LOG ANALYSIS

Typical uses of web proxy logs, which contain legislative information about network parameters, are needed to determine the unusual activities of a particular system or user (Rahaman, 2016; Mahanti, Williamson and Eager, 2000). Most of the web proxy logs contain general information like access time, bandwidth, IP addresses of end users and remote servers, URL, content type and so on. However, the combination of these logs may reveal fascinating unknown facts. For example, the most popular contents of internet traffic are application data, multimedia data like image or video, text data including HTML, CSS, XML, javascript, plain text, metadata, etc.

Web proxy analysis has a benefit to system administrators to help organizations determine the use of a network. The purpose of the log analysis can include the evaluation of network activities of users, making access policies for individual user groups, intranet, or internet resources, identifying and protecting security issues, determining the relationship between network parameters, etc. Web proxy log enables an administrator to view which user groups frequently access resources in an extensive form which can include access time, duration, resource utilization (in terms of bandwidth), content, type of content, size of content, session information as some examples. Web proxy logs not only gives administrator access to view the resources but also allows them to observe user platforms including operating systems, browser information, and IP addresses. All this information together can be a good reference to make policy for an organization. There is no better alternative than understanding the characteristics of internet traffic to improve performance and scalability of internet traffic (Mahanti et al., 2000).

In the context of network security, the log of a web proxy is a potential resource to find out and trace the fingerprint of an attacker who is hidden from the network administrator. An attacker may use different methods to become anonymous such as

the use of a TOR (The Onion Route) network, encrypted proxy, or VPN (Virtual Private Network) to name a few. However, even in these circumstances, there is a possibility that they can leave their footprint (X.Liu, Q. Liu, Wang, and Jia, 2016).

Large-scale log analysis enables the network administrator to perform suspicious activity detection by identifying patterns that might appear in the proxy log repetitively (Yen, Oprea, Onarlioglu, Leetham, Robertson, Juels, and Kirda, 2013; Kiatkumjounwong, Ngamsuriyaroj, Plangprasopchok and Hoonlor, 2014; Rahaman, 2016; Murata, Yamanishi, 2017). The pattern detection is useful to determine security events and predicting user behavior. Weblog analysis can be an efficient way to identify predictive user behavior when it deals with a session (Neelima and Rodda, 2016; Rao, Arora, 2017). This paper intends to analyze web proxy logs to demonstrate a practice by which an administrator can develop requirements for a resource allocation policy. To allocate internet resources, an administrator needs to identify resource consumption for both directions upstream and downstream in different time. Apart from this, it is essential to identify the traffic pattern, content type, uses of various protocols and methods. In this research, several illustrations have been shown on the critical parameters which have significant influence in resource allocation.

METHODOLOGY

The data for this project was collected from the Bluecoat web proxy located at www.honeynet.org. The data are a collection of web traffic logs from different users. The Honeynet organization allows researchers and other professionals to conduct their independent research using their web proxy log data. The obtained data was converted to a .csv format for cleaning and processing prior to analysis. The cleaning process involved removing unwanted and unrequired data and then the data was in a format ready for analysis. The form of raw data is illustrated in figure 1.

```
#Software: SGOS 3.2.4.8
#Version: 1.0
#Date: 2005-03-31 01:19:51
#Fields: date time time-taken c-ip sc-status s-action sc-bytes cs-bytes cs-method cs-uri-scheme cs-host cs-uri-path cs-uri-query cs-username s-hie
2005-04-05 18:27:21 75 10.0.1.2 200 TCP_NC_MISS 186 231 GET http www.ositis.com /tests/testconnectivity.asp ?Test=16246 - DIRECT www.ositis.com te
2005-04-05 18:36:57 157 10.0.1.2 200 TCP_CLIENT_REFRESH 372 247 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusEngineLocation ? - [
2005-04-05 18:36:57 226 10.0.1.2 200 TCP_CLIENT_REFRESH 388 248 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusPatternLocation ? -
2005-04-05 18:52:57 106 10.0.1.2 200 TCP_NC_MISS 234 213 GET http www.ositis.com /cgi-bin/DNSList.asp - - DIRECT www.ositis.com text/html "Mozilla,
2005-04-05 18:54:01 133 10.0.1.2 200 TCP_NC_MISS 521 508 GET http download.bluecoat.com /release/ProxyAV/2.2.1/2000E/update-22.asp ?6331CC061172FD
2005-04-05 19:01:29 280 10.0.1.2 200 TCP_NC_MISS 186 231 GET http www.ositis.com /tests/testconnectivity.asp ?Test=22634 - DIRECT www.ositis.com te
2005-04-05 19:07:21 164 10.0.1.2 200 TCP_CLIENT_REFRESH 388 248 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusPatternLocation ? -
2005-04-05 19:07:21 252 10.0.1.2 200 TCP_CLIENT_REFRESH 372 247 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusEngineLocation ? - [
2005-04-05 19:09:29 84 10.0.1.2 200 TCP_CLIENT_REFRESH 388 248 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusPatternLocation ? - [
2005-04-05 19:09:29 86 10.0.1.2 200 TCP_CLIENT_REFRESH 372 247 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusEngineLocation ? - D
2005-04-05 19:35:37 196 10.0.1.2 200 TCP_NC_MISS 186 231 GET http www.ositis.com /tests/testconnectivity.asp ?Test=21010 - DIRECT www.ositis.com te
2005-04-05 19:39:53 339 10.0.1.2 200 TCP_CLIENT_REFRESH 372 247 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusEngineLocation ? - [
2005-04-05 19:39:53 403 10.0.1.2 200 TCP_CLIENT_REFRESH 388 248 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusPatternLocation ? -
2005-04-05 19:42:01 107 10.0.1.2 200 TCP_CLIENT_REFRESH 372 247 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusEngineLocation ? - [
2005-04-05 19:42:01 107 10.0.1.2 200 TCP_CLIENT_REFRESH 388 248 GET http www.ositis.com /UpdateLocation/ProxyAV/McafeeAntivirusPatternLocation ? -
```

Figure 1. Raw data of web proxy log

Initially, data was separated into four different files in a .log format compressed in a *.zip file. The large size of the decompressed original file was 2.46GB requiring a python script to open and read the data and transforming the output to a .csv format. An additional python script was constructed to transform the data in a readable form (Python 2.7.0 Release, 2010). Several cleaning steps involving the removal of unwanted and unrequired data. This included repetivive data, blank or null fields, and those unessary for analysis. The final data set was 2.02GB in size and management for analysis in software tools available in Microsoft Excel.

Data Analysis

There are a number of tools available for visual analytics including Tableau, Microsoft Excel, PowerPivot, SAS Visual Analytics, Python (matplotlib). In this research, Tableau is used to conduct visual analytics, because it enables the visualization and presentation of data existing in forms (Tableau Desktop, 2017). Tableau offers flexibility in the variety of data types accepted and its user-friendly interface that allows for expedient analysis in a variety of ways. Tableau enables the users to visually interact with the data providing insight based on the user's liking. The visual illustration of the data helps users to analyze data quickly and make appropriate decisions on a specific event. For example, in this research, Tableau is used to visually depict the amount of bandwidth required to process for different action by the server or needed to transform different type of content. In the following result section, the outcome of data visualization analysis is discussed.

Results

In the field of network communication, one of the most valuable resources is Bandwidth. Optimizing bandwidth is considered as challenging parts of maintaining such communication. To optimize the bandwidth for both directions upstream (client to server) and downstream (server to client) it is essential to identify which action and content type are responsible for consuming most of the bandwidth. From the visual presentation in figure-2 it is observed that an S-Action called TCP_MISS consumes the highest amount of bandwidth which is essentially used for the traffic those are not in the cache of the web proxy server. It

implies that the network administrator should configure the web proxy server in a way by which it can cache more web traffic or allocate more bandwidth for such traffic to avoid the congestion during busy hours. Also in figure 2, it is observed that over 90% of the bandwidth is occupied by the server action called TCP_MISS (traffic that is not in web proxy cache) and TCP_HIT (the traffic available in web proxy cache and it is requested by a client) during server to client data transformation. On the other hand, when it comes to client to server communication, 80% of bandwidth consumed by the server action called TCP_NC_MISS (non-cacheable traffic) and TCP_TUNNELED (traffic that uses HTTPS).

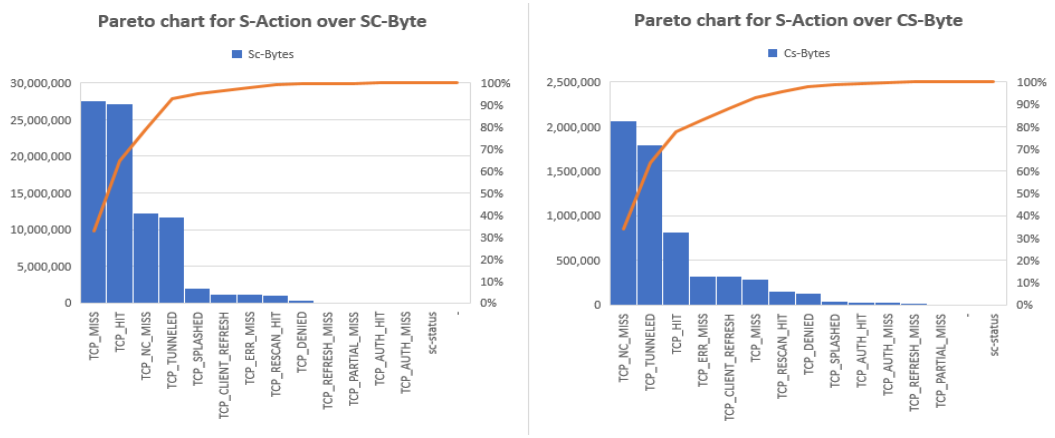

Figure 2. Pareto chart for S-Action over SC-Byte and CS-Byte

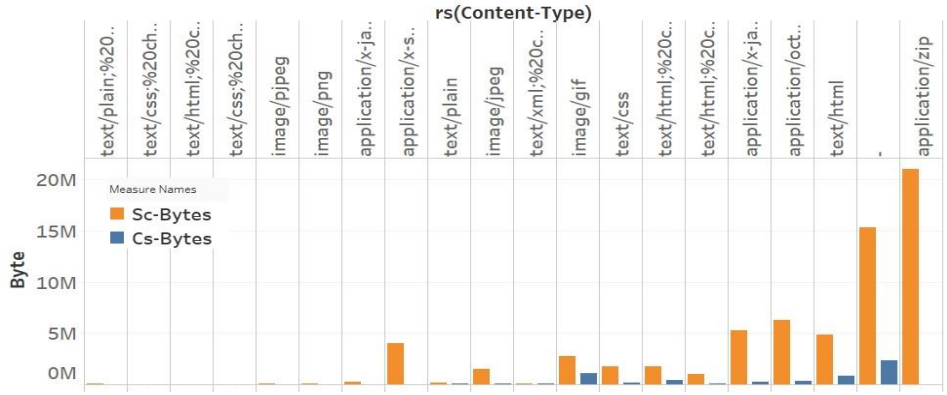

Figure 3. Bandwidth requirement from server to client and client to server in terms of rs-content-type (Remote Server content type)

Web traffic contains a variety of content including text, image, video, and application. Figure 3 shows that in terms of RS content type, the category named applications consumes maximum amount bandwidth, which gives an idea to the network administrator that there should be more allocated bandwidth for the category application. However, the administrator should focus more on downstream traffic as the ratio of downstream traffic over upstream traffic is very high (figure 2, figure 3). Figure 4 represents the CS-URI-Scheme. The packed bubbles depict that most of the traffic uses HTTP and HTTPS which implies that a significant amount of traffic is web traffic. If the size of the bubble is considered, then it is not a hard to determine the ratio of secure web traffic versus non-secure web traffic. The CS-Byte consumption for secure (HTTP) web traffic is 996,986 bytes where this number goes to 2,913,020 bytes for non-secure (HTTPS) web traffic. The second most of the traffic uses TCP protocol, followed by CS-Method, POST, and Head respectively.

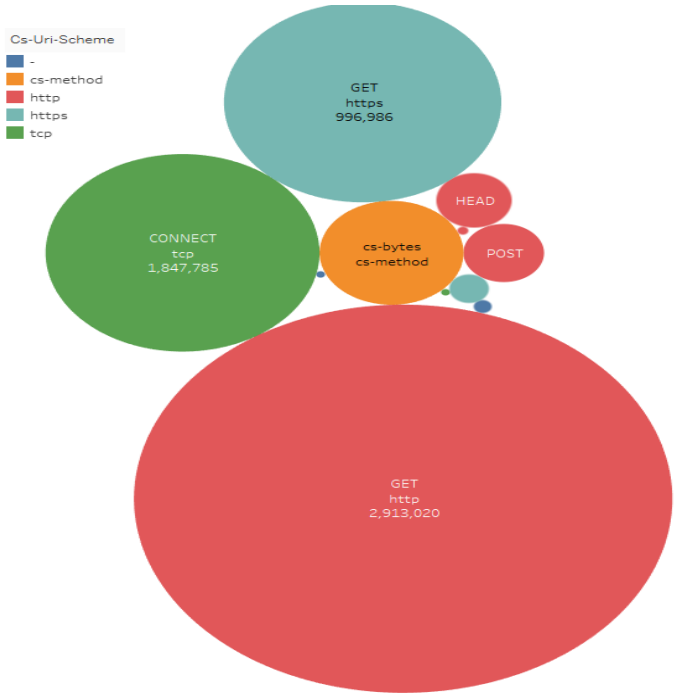

Figure 4. CS-URI-Scheme

The number of live IP address represents the number of active users in a network who are using the network resources concurrently. Derived from the web logs, figure 5 (a,b) depicts that highest number of live IP addresses and maximum consumption of bandwidth are being observed in between 10:00 AM and 3:00 PM. From both figures, it is observed that the most resource consumed period, the busiest hour, is between 11:00 AM and 2:00 PM, and 2:00 PM is observed as the busiest hour. It can be concluded that employees become most active right after the lunch period. These figures give an idea of resource utilization during a busy hour which is essential information for a network administrator to allocate network resources for different portion time in a day.

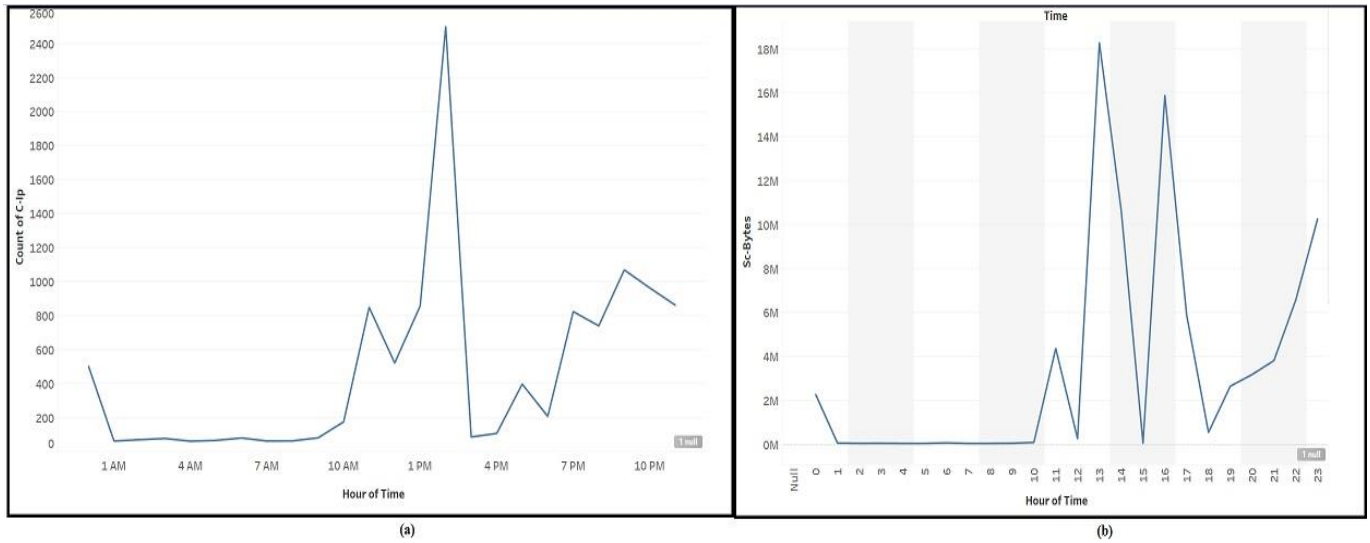

Figure 5. (a) Hourly variation of live IP addresses (b) Hourly bandwidth consumption

Conclusion

This study presents an analysis of web proxy log traffic using Tableau as a data visualization tool for analysis. Using data from www.honeynet.org, we assessed web proxy log traffic data to clean, prepare, and process for our analysis. Web proxy logs keep track of the history of user activities allowing an administrator to develop requirements based on a resource allocation. Using the collected data, we observed that more than 90 percent of bandwidth is occupied by TCP_MISS and TCP_HIT action during server to client data transformation and the amount of both kinds of traffic are similar. Hence, most of the web traffic is a collection of cached traffic and non-cached traffic, which implies the web proxy server is required to configure in an efficient way so that it can store most of the data requested by the users which will play a role to optimize the overall bandwidth consumption. Also, we saw 80 percent bandwidth is consumed by TCP_NC_MISS and TCP_TUNNELED in client to server communication suggesting that upstream traffic demands more privacy then downstream traffic as most of the traffic is either non-cacheable or secure. From packed bubbles, it can be concluded that most of traffic request are web-based traffic that appear to be mostly insecure. Furthermore, we observed that most resources are consumed in between the time 10:00 AM and 2:00 PM where the number of live IP addresses and bandwidth peak most at 2:00 PM showing that at the busy hour when maximum live IP addresses are available, the resource consumption is also higher.

Some of the benefits garnered from this research include exposing the vulnerability and security breaches of the whole network, making use of available web proxy logs. These logs are a useful source from where a network administrator can investigate people's activities and a network's resource utilization. Web proxy log analysis makes it possible to create an efficient IT network policy serving both resource management and security needs effectively.

In addition, this study serves as a good baseline for future work in preparing an efficient IT resource management policy. Researchers can conclude from this work that the analytics on web proxy has the potential to reveal unspoken interesting facts on web traffic. Network administrators can benefit by gathering knowledge of the network system like the resource consumption rate, most favored internet content and applications, and how resource consumption may vary on different factors.

REFERENCES

1. Mahanti, A., Williamson, C., & Eager, D. (2000) Traffic analysis of a web proxy caching hierarchy. *IEEE Network*, 14, 3, 16-23.
2. Honeynet.org (2017, November 25) Retrieved from www.honeynet.org.
3. Neelima, G., & Rodda, S. (2016) Predicting user behavior through sessions using the web log mining. In *2016 International Conference on Advances in Human Machine Interaction (HMI)*, 1-5.
4. technet.microsoft.com (2017) Select W3C Fields to Log (IIS 7), Retrieved from: https://technet.microsoft.com/en-us/library/cc754702(v=ws.10).aspx
5. Rahaman, M. A., Hebert, C., & Frank, J. (2016) An Attack Pattern Framework for Monitoring Enterprise Information Systems, *2016 IEEE 25th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, 173-17.
6. Murata, M., & Yamanishi, K. (2017) Detecting Drive-by Download Attacks from Proxy Log Information using Convolutional Neural Network.
7. Kiatkumjounwong, N., Ngamsuriyaroj, S., Plangprasopchok, A., & Hoonlor, A. (2014) Analysis and classification of web proxy logs based on patterns of traffic rates, *TENCON 2014-2014 IEEE Region 10 Conference*, 1-5
8. Python 2.7.0 Release. (2010) Retrieved from: https://www.python.org/download/releases/2.7/
9. Rao, R. S., & Arora, J. (2017) A Survey on Methods used in Web Usage Mining, *Res. J. Eng. Technol*, 4, 5.
10. Tableau Desktop. (2017) Tableau Desktop, Retrieved from: https://www.tableau.com/products/desktop/download
11. Yen, T. F., Oprea, A., Onarlioglu, K., Leetham, T., Robertson, W., Juels, A., & Kirda, E. (2013) Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks, In *Proceedings of the 29th Annual Computer Security Applications Conference*, 199-208.
12. Liu, X., Liu, Q., Wang, X., & Jia, Z. (2016) Fingerprinting web browser for tracing anonymous web attackers, *IEEE International Conference of Data Science in Cyberspace (DSC)*, 222-229.