Implementation feature selection method for detection packet data on Fasilkom Unsri computer network

To cite this article: Ahmad Fali Okilias et al 2019 J. Phys.: Conf. Ser. 1196 012063

View the article online for updates and enhancements.
Implementation feature selection method for detection packet data on Fasilkom Unsri computer network

Ahmad Fali Okilas1*, Vieronica Varbi Sununianti2, Tasmi3
1 Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
2 Department of Sociology, Faculty of Social and Political Science, Universitas Sriwijaya, Indonesia
3 Computer Engineering Computer Science Department, Universitas Indo Global Mandiri, Indonesia
* fali@ilkom.unsri.ac.id

Abstract. Network traffic generates data traffic accessed by users. These data sometimes have nothing to do with the work area of an office or institution. A lot of data traffic as an example at a university is caused by diverse and numerous user populations. They access various data. For this reason, it is necessary to monitor data traffic in order to control and make decisions for network administrators. To get data packages that are often accessed and the most dominant, it is necessary to build a monitoring system on the network. In building this system, stages are made using the feature selection method. The results of this method to get the detection of data packet engines and applications. Furthermore, these results can be displayed visually in the form of graphs. This result is a dashboard for network monitoring systems. This dashboard can be used to help a network administrator to consider network monitoring decisions.

1. Introduction

The use of the internet in a computer network in an institution or office has become a necessity for today's internet users. At any time during working hours, employees or students must always access the internet using their office computer network or institution. A large number of internet data packages will cause problems in internet usage. This is a concern for network administrators to monitor the use of data packages that are most frequently accessed by users. The most dominant type of data package and the percent amount will be considered by the network administrator in making decisions for network monitoring. Retrieval of data from data packages that are accessed by the user can be displayed visually in order to facilitate monitoring. This visual graphics display can be used as a dashboard of a built-in monitoring system to make a decision on the network.

Research conducted by several previous researchers stated that users have often used HyperText Transfer Protocol Secure (HTTPS) applications, such as those made [1]. They stated problems arising for the classification of the estimated package. From the paper [2] obtained the classification of traffic types as a way to optimize bandwidth so that internet connections are reliable and stable.

At this time a computer network in an institution or company uses heterogeneous networks. One of the meanings of heterogeneity is where the network uses different settings and operating systems. The research conducted by [3] uses the Nagios application to monitor the network with the characteristics of scalability, application (daemon) and data security. In addition, other studies monitor user quality bandwidth and service of quality (QoS) [4].
The dataset from capture results using the TCPDUM tool produces several attributes that must be selected as the basis for recognizing a data packet pattern. The feature selection is one method used in determining the correlation between attributes in one class. There have been many studies conducted using feature selection in identifying one type of data package in network and data science cases. One of them is the technique used is to use entropy and information gain in passing feature selection. Information gain is part of the selection method based on ranking attributes and based on the highest values and the relevance of the class. Research conducted by [5] and [6] uses FE algorithms to detect traffic and attacks.

Identification of data packets is needed to properly identify package patterns. For this purpose, a research on data packet detection with a feature selection method. Where is the case study in this paper takes traffic from Computer Science Faculty Universitas Sriwijaya This paper is also intended to produce a real-time system for monitoring and grouping https traffic on the network of Computer Science Faculty Universitas Sriwijaya. This research was conducted at the Computer Science Faculty Universitas Sriwijaya because of the large amount of data traffic that occurred but unknown packages and what applications were most accessed by the user. The speed of data access there is sometimes felt slow. Besides that, we also want to see the tendency of user behavior about what they often access.

2. Feature Selection

In paper [7] to find entropy values. In this study, the stages were carried out, namely determining the attributes of network traffic by capturing data. Next, find the relevant attributes of the data. The method used is the feature ranking method or information Gain (IG). This method is to find a ranking that will be validated. In the validation method used is the classification of Naive Bayes. The results will get the accuracy of each attribute from the data obtained.

The initial stages are pre-processing data to process data cleaning to produce accuracy. Then the process is continued by dividing the dataset into training data and testing data. This stage is a learning process that predicts dominant attributes. To get the data package pattern, this learning process is carried out.

The results of entropy and ranking of each attribute are obtained by using the information Gain (IG) method. Ranking results are obtained that the Source_port attribute is the first rank. This means that this attribute has a large contribution to the attribute service that is used as a class. While the three attributes are Header_length, Urgent_Pointer and Protocol are the three lowest ranks. This result means these attributes have no contribution.

The selection of attributes in this study uses the filtering method, where one method used is the Mutual Information [8],[9]. The advantages of this method are based on a simple principle, the more information obtained the better.

3. Result and Analysis

Data packet detection for feature selection produces two output stages, namely the detection of data packet engines and applications; visualization of detection of data packet engines and applications with PHP. These results are described as follows.

3.1. Detection of data packet engines and applications

Data packages are taken in real-time by building a network on the Fasilkom Unsri campus using public IP. The resulting data is obtained by using the TCPDUM application to obtain raw data. This is to prove that feature selection is correct. Then tested with a system that was built using python programs, MYSQL and PHP as media to display the results of monitoring into graphic form.

This research results in retrieving data packets and applications from engine detection. Learning systems need to be done to update the knowledge of the system so that it can detect the types of packages. Rules developed are the result of feature selection to update knowledge in identifying packages and applications. The experimental results shown in Figure 1 are graphs from the detection of data packet engines and applications.
In this study, the engine experiment was conducted four times to ensure that the rules built could recognize data packages and applications on the internet. From Figure 1, the results of the first experiment are shown by the blue line detected by the data packet engine and application. Where the dominant data packets that are accessed by users are Google, https, HTTP, YouTube, and others. In the second experiment, indicated by the red line, the dominant data packet accessed by the user is almost the same as the first experiment, but the percentage is higher. But for HTTP is almost the same as https, and Facebook starts to increase compared to the first trial, even higher than YouTube. The third experiment was shown by the green line and the fourth experiment which was shown by blue, almost the same pattern, but the percent was higher than the previous one. Significant results appear high percent data packet values accessed by the user, DHCP. Where the first and second experiments are not so high the DHCP value.

The amount of data from the engine that was built is shown in Figure 1. In the experiment that was carried out found for the type of data package on average the most https type package that is equal to 60.87 percent, Other 22.58 percent, HTTP 8.53 percent, DHCP 0.01 percent, DNS 0.00 percent, and SSH 0.00 percent. While the biggest average application is Google at 4.19 percent followed by YouTube at 2.30 percent and the last is Facebook at 1.53 percent.

This is a technological intervention for humans. The choice of consumer social actions also shows consuming the virtual world in various ways. Finally, cultural differences occur where a sense of security is a challenge. Facebook is known to still be one of the active social media used by consumers in the millennial and Z-generation categories. Facebook is still the user's goal to add and maintain relations with the real world and virtual interactions. Actually, here, technology is able to facilitate a number of people to create social capital regardless of form.

From the results of the experiment shows the engine has not been able to recognize packet types for DNS and SSH. This is evidenced by the results obtained from four experiments yielding 0.00 percent. Many applications still produce packages that have lost or cannot be identified as in the third experiment; there was 37.57 percent of the amount of data that has not been detected by the system.

3.2. Monitoring of detection of data packet engines and applications with PHP
This paper shows the monitoring of detection of data packet engines and applications with PHP. Figure 2 is the result of the detection of the data packet and application detection using PHP. The system built is a system that can calculate and analyze presentations from protocols and internet applications.

This system consists of two modules to send files to the MySQL database and one to do data calculations and analysis. The first module is used to read .csv files that will be stored in the database in table form. The goal for the calculation process is easier to do. The first module will transfer data to the results table in the form of id, Application Name and number of packages. The second module is used
for calculation and displays the visualization of the calculation of the number of packages and applications in the form of Percent.

Figure 2. Monitoring of engine detection Data packages and applications with PHP and Mysql.

Monitoring of the results of the experiment to facilitate analysis and monitoring system. The results of this visualization become a dashboard of the monitoring system built. From Figure 2 experiment 1 (a), application and protocol monitoring system, the dominant data package is https 63.38 Percent and other 24.30 percent. From Figure 2 experiment 2 (b), application and protocol monitoring system, the dominant data package is https 71.24 percent and HTTP 14.37 percent. From Figure 2 experiment 3 (c), application and protocol monitoring system, the dominant data package is https 46.14 percent and other 37.57 percent. From Figure 2 experiment 4 (d), application and protocol monitoring system, the dominant data package is https 62.70 percent and other 20.35 percent. If the results of this monitoring are analyzed by the network administrator, it can be used for decision making in network monitoring.

The virtual world is the choice of users to try new experiences. With a fluid user identity, it supports freedom of experimentation in this huge virtual world without limits and time. Here, there is its own culture where there are no moral values, religion, nor cultural values that influence the user's social interaction. Reference groups become external forces that influence consumer social behavior in cyberspace, such as lifestyle, the similarity in the background, and social status. The virtual world needs to be seen as the arena of class struggles that fight over each other. So, the need for the presence of the State and educational institutions that become social control in the virtual world, even with limited coverage. It is time to develop information technology aimed at social development.
4. Conclusion
The most common types of data packages from large to small are https, HTTP, DHCP, DNS, and SSH. For the average application from large to small are Google, YouTube, and Facebook. The engine built cannot recognize the type of package for DNS and SSH. Monitoring of the results of the experiment to facilitate analysis and monitoring system. The results of monitoring can be as a tool for network administrators because it is in the form of a dashboard of the monitoring system built. Display of data retrieval results appears in the form of a pie chart. It describes the data package and which applications are most dominantly accessed by the user. This description shows the active users in the network of Computer Science Faculty Universitas Sriwijaya. The use of this dashboard can be used in decision-making systems for network monitoring. With pie charts, network administrators more easily analyze user behavior on the computer network they are monitoring.

References
[1] Husák, M., Cermák, M., Jirsík, T., & Celeda, P. (2016). Open Access HTTPS traffic analysis and client identification using passive SSL / TLS fingerprinting. EURASIP Journal on Information Security. https://doi.org/10.1186/s13635-016-0030-7
[2] Siqueira, I. G., Ruiz, L. B., & Loureiro, A. a. F. (2007). Coverage area management for wireless sensor networks. International Journal of Network Management, (October 2005), 17–31. https://doi.org/10.1002/nem.
[3] KORA, Ahmed D. & SOIDRIDINE, Moussa Moindze. (2012). Nagios Based Enhanced IT Management System. International Journal of Engineering Science and Technology (IJEEST), Vol. 4 No.03 March 2012
[4] Shuhaimi, Mohammad Ali Arsyad bin Mohd. et all. (2011). The New Services in Nagios: Network Bandwidth Utility, Email Notification and sms Alert in Improving the Network Performance. 7th International Conference on Information Assurance and Security (IAS)
[5] Kumar, Koushal & Batth, Jaspreet Singh. (2016). Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms. International Journal of Computer Applications (0975–8887), Volume 150 – No.12, September 2016.
[6] B.Azhagusundari & Thanamani, Antony Selvadoss. (2013). Feature Selection based on Information Gain. International Journal of Innovative Technology and Exploring Engineering (IJITEE). ISSN: 2278-3075, Volume-2, Issue-2, January 2013.
[7] Novaković, J., Strbac, P., & Bulatović, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1), 2334–6043. https://doi.org/10.2298/YJOR1101119N.
[8] H. Yang and J. Moody, “Feature selection based on joint mutual information,” Proc. Int. ICSC Symp. ..., vol. 27, no. 8, pp. 1226–1238, 1999.
[9] S. Foithong, O. Pinngern, and B. Attachoo, “Feature subset selection wrapper based on mutual information and rough sets,” Expert Syst. Appl., vol. 39, no. 1, pp. 574–584, 2012.