The Method of Clustering Network Traffic Classifications for Extracting Payload Signature by Function

Kyu-Seok Shim, Young-Hoon Goo, Min-Seob Lee, Huru Hasanova and Myung-Sup Kim
Dept. of Computer and Information Science, Korea University
Sejong, Korea
{kusuk007, gyh0808, chenlima2, hhuru, tmskim}@korea.ac.kr

Abstract— Today, many applications and services have different traffic patterns. In addition, applications that use network functions have many functions. Extraction of these applications is essential for network management. Although the visual is traditionally derived by manually extracting signature, studies of automatic visual extraction are active to reduce the time it takes to extract a visual. However, the automatic signature generation system does not extract signatures by functional. Therefore, this thesis proposes a method to extract functional signatures by clustering traffic by function. The system proposed in this thesis aims to classify traffic by function and put it into automatic signature generation system, and to partition the classified traffic into response and request to extract sophisticated signatures.

Keywords—automatic, clustering, signature, network management

I. Introduction

Today, the most applications utilize network resources. The utilization rate of network resources is increased and the amount of network management target traffic exponentially increases. Applications are increasing the type of service according to user needs. Network administrators can classify traffic for each application. However, it is not possible to classify traffic for each service in the application. Therefore, efficient network management becomes difficult[1,2].

Network monitoring is to find out the traffic amount of the specific application and establish a management policy that matches it. In network monitoring, traffic classification is essential for providing high quality of service to a user and for receiving high quality of service from a network provider with minimum network resources. A signature is an essential feature that used to classify traffic by application in traffic classification process. Obviously, there are wide range of signatures which can be classified by the characteristics of the traffic.

Signature is the key to classifying traffic by application. There are various types of signatures for traffic classification. Port-based signatures can be classified using the port numbers used by the application. Header-based signatures can be classified using the IP address, port number, and protocol used by the application. Statistical-based signatures can be classified using the statistical information that arises from the application. Finally, payload signatures can be classified using sub-string that occurs only in the application, among the payloads that occur in the application.

The payload signature is a unique and continuous substring in payload of the same application traffic. The payload signatures can classify applications most accurately among many signature types. However, in order to generate the payload signature, a lot of time and money is needed. Most of the existing methods generate the payload signature in similar manner. First, a manager gathers traffic of an application to extract the signature. Second, the user finds the substring that commonly occurs while comparing the contents of the payload. After extracting the common substring, the string can uniquely be used to performs the verification of an application. Therefore, depending on the extraction operator, there can be a difference between the quality of the signature, which leads to disadvantages in signature objectivity.

Therefore, the system is being studied that payload signature automatic generates[3-8]. These studied can reduce the amount of time spent on the generation of signatures that were previously done manually. However, there are also problems with the automatic signature generation system. Application traffic, which is inputs from system, must be manually collected. This process may include other application traffic. In addition, it is almost impossible to extract signatures by function. Even if traffic is collected using only that function, it is very easy and frequent to mix traffic with other features.

To solve these problems, this paper proposed the automatic payload signature generation model that includes clustering step for traffic classification by function. Not only does this model extract functional signatures through functional traffic classifications, it also provides more detailed signatures than conventional methods. Since the automatic payload signature system to be used in this paper is an Apriori algorithm based system, detailed signatures can be extracted as traffic is divided in detail. For example, because only the entire keyword can be extracted from integrated traffic, while more detailed keywords can be extracted from the traffic divided by function.

In section 2, we review the previous work in the traffic classification and the automatic signature generation. In section 3, we propose the automatic signature update system

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea Government(MSIT) (No.2018-0-00539-001,Development of Blockchain Transaction Monitoring and Analysis Technology) and by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2018R1D1A1B07045742)

with clustering step. Finally advert the conclusion and future work in section 4.

II. Related Work

The payload signature has the high accuracy and coverage. However, the extracting signature is very difficult and time consuming. Therefore, studies of automatic payload signature generation are in the limelight in the field of network management. The existing methods are LASER (LCS-based Application Signature ExtRaction), Autosig, and SigBox.

LASER automatically generates an application signature, in the form of a sequence of substrings, in the payload of packet by using a modified version of the LCS (longest common subsequence) algorithm. The inputs of this algorithm are two distinct byte streams of packet payloads that belong to two flows. In order to improve the system's performance in terms of execution time and accuracy, this method only considers the first N packets of a flow and groups these packets by their size, since large packets are not likely to carry the same kind of information as the small ones. Finally, the method compares two inputs to get the longest common subsequence between them, and then compare it with another subsequence iteratively to refine it.

Autosig also generates an application signature automatically, which extracts multiple common substring sequences from input flows as application signature. First, it divides the payload of a set of flows into short substrings called shingles. After extracting all of the relevant, common shingles, Autosig merges them if they are neighbors or overlap. Next, a substring tree is constructed to create all possible combinations of substrings. These combinations are considered as signatures.

SigBox uses the Apriori algorithm[9,10] to solve the above disadvantage. The above methods are necessary preprocessing and postprocessing in order to compare two strings. The preprocessing is setting the order of traffic and grouping the traffic. The postprocessing integrates the generated substring into one rule. However, SigBox extracts substring likely to become signatures by increasing the length-1 all the substrings candidates. Therefore, this method does not take much time to the extraction process and does not required preprocessing and postprocessing. In this paper, we extract signature using Sigbox.

Earlier, we mentioned the automatic signature generation system. However, all system cannot extract payload signature by function. Therefore, clustering methods are very important in this system. The quality of the extracted payload signature depends on the results of clustering. The most common method of clustering is K-means clustering.

The k-means clustering algorithm is one of the division of clustering. Division divides a given data into groups. The algorithm divides the data into k groups of one or more data objects. In the process, the cost function, such as distance based cross-group, is minimized. Therefore, k-means algorithm sets the sum of squares of data within each group as a function of cost and minimizes the value of each function in the clustering[11].

However, the performance of k-means clustering depends on the initial value. In 2007, David Arthur and Sergei Vassilvitskii proposed k-means ++ to reduce the damage caused by these properties[12]. The k-means ++ algorithm is selecting the initial value of the k-means algorithm. Select one random data from the data set and center it. For each of the following data sets, the distance D(x) is calculated from the data and the closest center of interest selected. Select any data using a distribution of biased probabilities relative to $D(x)^2$, then set it to the n-center. Repeat this process, performing k-means clustering with selected k-centers as initial value.

III. Proposed Method

In this paper, the proposed method is the automatic signature generation system with clustering steps. Existing methods extract signatures common to all traffic traces without traffic clustering. Therefore, incorrect signature is extracted, and the extracted signature is classified as signature for the application.

However, this process must be performed in order to extract signature-specific signatures used in the application, and this process reduces the chance of inaccurate signatures. The process of extracting features of traffic is required for clustering. Therefore, the traffic collection team includes the traffic feature extraction process.

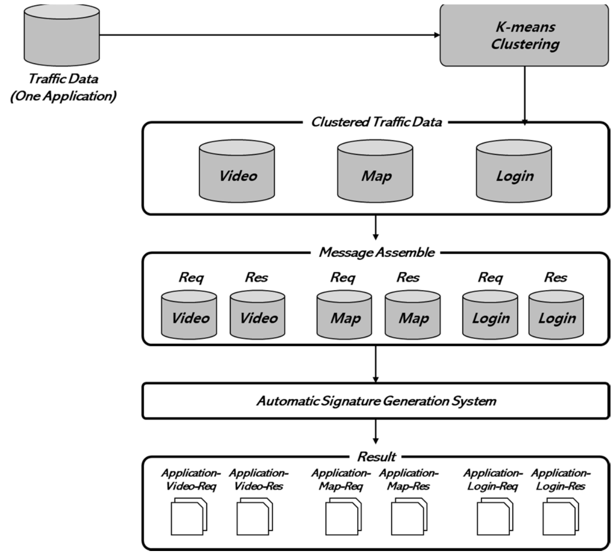

Figure 1. Automatic payload signature generation model that includes clustering step

The traffic features used in this study include the size of the flow, the number of packets contained in the flow, the duration of the flow, and the delay time. Cluster similar flows based on the following features. For example, in the case of video services, only large flows of flow can be collected and signatures for video services can be extracted.

However, the signature that corresponds to the result to be extracted from the automatic generation of signature shall state which service of any application. The application is addressed through the answer sheet traffic, but the service cannot be specified in the process. In this study, these applications and services descriptions are defined as labels. Labels are fields that specify that when traffic is classified by extracted signatures, the traffic is " B service of application A ". To define a service's label, this study uses a Domain Name Server (DNS) response packet.

Because the information in the response packet has a multi-dimensional address scheme, it is possible to extract service keywords from that information. Automatic Signature Generation System accepts clustered traffic and automatically extracts signatures using the Apriori algorithm. In the corresponding part, a payload signature is extracted from all three phases, first of which the content signature means a series of strings commonly occurring in each Traffic Trace relative to the traffic payload. Thus multiple content signatures can be extracted from a single packet or flow. Second, packet signal means the set of content signatures that occur in the same packet.

Create sophisticated signatures of the packet units to increase the accuracy of the signature, and to generate a flow signature, a third, not a break in the packet unit. Flow signal means the set of packet signatures that occur in the same flow. Using the Apriori algorithm, which is a type of data mining, all of the above processes will be used to develop a mechanism that will skip the pre-processing and postprocessing process and allow all strings to be compared at once.

The goal of K-means clustering algorithm is to have N data objects split into k sets that maximize cohesiveness between objects in each set when given D dimension data objects. At the center point of each set, the goal is to find an aggregation S that minimizes the sum of squares of the objects in the collection.

Cluster the traffic, separate it by function, and input the segmented traffic into the automatic signature generation system. The entered traffic is subdivided into Request messages and Response messages. Request message set extracts a request signature through the automatic signature generation system. Repeat this process for the response message set. The reason why the request and response message are shared is that the visual derived from each message may be different. Separate inputs provide sophisticated signatures because of the support calculations of Apriori, the automatic signature generation system core algorithm.

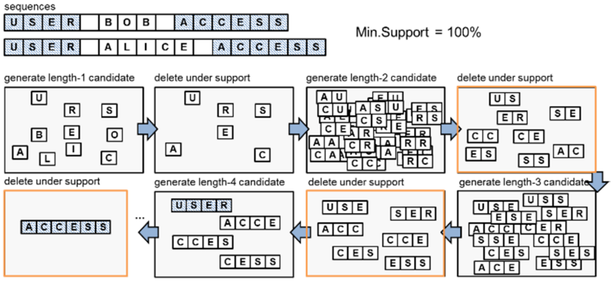

Figure 2. Automatic Signature Generation System based on Apriori algorithm

Automatic signature generation system is used to modify the sequential pattern algorithm (Apriori) for signature extraction. In the process of automatic signature generation, the system generates sequences for the payload traffic extraction. From the extracted payload traffic, length-1 content signatures are an alphabet. From the extracted length-1 content signatures, length-2 content signatures are created with deletion of unwanted content parts. The process continues to length-k until no more content signatures to generate higher lengths in the extraction of common strings.

IV. Conclusion

In this paper, we propose an automatic generation of payload signatures for classification by network traffic service. Traffic clustering uses clustering techniques to cluster each traffic trace using traffic statistical information for each flow. Clustered flows can be divided into flows by service and entered into the automatic signature generation system using flows by functions. In addition, the entered flow is grouped into response and request messages to generate signatures using the Apriori algorithm. These are signatures that allow traffic to be classified by application service.

In future works, the methodology is collected and tested for each application, and the best clustering techniques are studied. It also plans to select a feature type for traffic used in clustering to conduct a study to achieve optimal results.

References

[1] M.-S. Kim, Y. J. Won, and J. W.-K. Hong, "Application-level traffic monitoring and an analysis on IP networks," ETRI journal, vol. 27, pp. 22-42, 2005.

[2] B. Park, Y. Won, J. Chung, M. S. Kim, and J. W. K. Hong, "Fine-grained traffic classification based on functional separation," International Journal of Network Management, vol. 23, pp. 350-381, Sep 2013.

[3] Baraka D. Sija, Kyu-Seok Shim and Myung-Sup Kim, "Automatic Payload Signature Generation for Accurate Identification of Internet Applications and Application Services," KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS Vol. 12, No. 04, Apr. 2018, pp. 1572-1593.

[4] X. Feng, X. Huang, X. Tian, and Y. Ma, "Automatic traffic signature extraction based on Smith-waterman algorithm for traffic classification," in Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on, 2010, pp. 154-158.

[5] Kyu-Seok Shim, Young-Hoon Goo, Sungyun Kim, Mi-Jung Choi and Myung-Sup Kim, "SigManager: Automatic Payload Signature Management System for the Classification of Dynamically Changing Internet Applications," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2017, Seoul, Korea, Sep. 27-29, 2017, pp.350-353.

[6] Y. Wang, Y. Xiang, and S. Z. Yu, "An automatic application signature construction system for unknown traffic," Concurrency and Computation-Practice & Experience, vol. 22, pp. 1927-1944, Sep 2010.

[7] Y. Choi, "An Automated Classifier Generation System for Application-Level Mobile Traffic Identification," 2011.

[8] P. Haffner, S. Sen, O. Spatscheck, and D. Wang, "ACAS: automated construction of application signatures," in Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data, 2005, pp. 197-202.

[9] Jiao Yabing, " Research of an Improved Apriori Algorithm in Data Mining Association Rules", International Journal of Computer and Communication Engineering, Vol.2, No.1, 2013, pp. 25-27.

[10] Abaya, Sheila A. "Association rule mining based on Apriori algorithm in minimizing candidate generation." International Journal of Scientific & Engineering Research, Vol.3.7, 2012, pp. 1-4..

[11] Jain, A.K, “Data clustering: 50 years beyomd K-means.”, Pattern recognition letters, 2010, pp.651-666

[12] Bachem, Olivier, et al. "Approximate K-Means++ in Sublinear Time." Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp.1459-1467.