Finding the Highly Efficient Application Signature through Payload Signature Quality Evaluation

Sung-Ho Lee, Young-Hoon Goo, Baraka D. Sija, and Myung-Sup Kim
Dept. of Computer and Information Science
Korea University
Korea
{gaek5, gyh0808, sijabarakajia25, tmskim}@korea.ac.kr

Abstract— **Internet traffic identification is an essential preliminary step for stable service provision and efficient network management. The payload signature-based-classification is considered as a reliable method for Internet traffic identification. But its performance is highly dependent on the number and the structure of signatures. If the numbers and structural complexity of signatures are not proper, the performance of payload signature-based-classification easily deteriorates. Therefore, in order to improve the performance of the identification system, it is necessary to regulate the numbers of the signature. In this paper, we propose a novel signature quality evaluation method to decide which signature is highly efficient for Internet traffic identification. We newly define the signature quality evaluation criteria and find the highly efficient signature through the method. Quality evaluation is performed in three different perspectives and the weight of each signature is computed through those perspectives values. And we construct the signature map(S-MAP) to find the highly efficient signature. The proposed method achieved an approximately fourfold increased efficiency in application traffic identification.**

Keyword— Application Identification, Application Signature, Signature Quality Evaluation, S-MAP Introduction

I. INTRODUCTION

Due to the high speed technology of current Internet, a number of private and company applications providing different services are being drastically developed. In such environment, application level traffic monitoring and analysis for efficient operations and management of a particular network is in a great demand. Although several methods have already been proposed to solve the problem, real time methods for handling large amounts of traffic in high-speed link that can accurately classify various types of applications level traffic are still required.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A3A01018057), and Institute for Information & communications Technology Promotion (IITP) grant funded by Korea government (MSIP) (No. B0101-15-0300, The Development of Cyber Blackbox and Integrated Security Analysis Technology for Proactive and Reactive Cyber Incident Response)

In the application level traffic identification, payload signature-based analysis method showed a relatively higher identification accuracy compared to other analytical methods. But, low processing speed still remains a problem [1,2]. With the increasing usage of the Internet applications, payload processing speed issues are challenges to be resolved when considering the network trends.

In this paper, we propose a novel method for quality evaluation of signatures which are generated by the automatic signature generation system. The quality evaluation method is to quantify the content of each signature and search a relatively highly efficient signature in terms of application traffic identification.

The quality evaluation is performed based on some criteria of each signatures. These criteria are performance, redundancy and characteristic. Redundancy is about numbers of traffic flows that each signature can identify. Characteristic is for considering the readable or meaningful character portion in signature. Performance is about identification speed that how fast the signature can identify the traffic, so offset and searching depth of signature is considered. By those three evaluations criteria, we calculate a quality weight value and construct the signature map, the S-MAP. Finally, through S-MAP browsing, highly efficient signatures were determined.

An experiment for the verification of the proposed method was conducted through the Torrent application traffic. Signatures were extracted by automatic signature generation system and the proposed signature quality evaluation method is applied to those extracted signatures. As result, total signatures are reduced about 83% when quality evaluation is applied. At the same time, the number of signatures decrease but traffic identification completeness is nearly maintained. When compared to non-evaluated signatures that quality evaluation is not applied, signatures that found by quality evaluation have four-fold traffic identification efficiency on average. So the proposed signature quality evaluation can be determined as an effective method to exclude unnecessary signatures and search for the highly efficient signatures.

The remainder of this paper is organized as follows. Section 2 describes related researches. And the problems handled in this paper are defined in Section 3. The details of the proposed method applied in the study is described, in

Section 4. In Section 5, our proposal is applied to the identification system and its validity is proven. Finally, Section 6 describes conclusions and future research directions.

II. Related Work

In this section, we describe the requirement of payload signature-based identification system in the perspective of application traffic identification.

When Internet service providers try to provide seamless service to the users, the signatures gradually become complex and appear in a variety of forms because the configuration of the application-level protocol becomes a complex structure to bypass the network security devices. In addition, an increase in the number of Internet applications causes the increase of the application signatures. As the signature increases in both, complexity and number, the processing speed of the payload signature-based application traffic identification system has become an important factor in determining the overall performance of the system.

Therefore, several studies related to traffic identification have been in progress, as stated above, by improving inefficiency of signatures expect a fundamental improvement in application traffic identification system.

Various pattern matching algorithms for improving the speed of the application identification systems have been proposed. However, performance of the pattern matching algorithm is dependent on the configuration of the input data, indicating a limited performance improvement [5]. There was also a limit that depends on the configuration or structure of the signature because there was no consideration of the offset or the pattern having the matching step for each signature.

In solving the limited performance problem of the conventional matching algorithms, new method that minimize the search space of the signature to reflect the occurrence pattern of the application level traffic identification system has been proposed [6]. Reflecting the characteristics of the traffic generation pattern, or increasing the efficiency of the application identification is similar to the method proposed in this paper as well.

However, previous studies considered only the HC (Hit Count) of each signature and did not conduct evaluation by the numerical importance or value. As a result, the search space minimization was not an optimal method in searching and sorting signatures. Therefore, it could not fundamentally solve the inefficiencies and the method was also limited.

Without relying on the signature pattern matching algorithm or signature based identification method, using the correlation between the local traffic has also been proposed [3,4,7]. However, such as CDN (Content Delivery Network) traffic that have correlation, but because each of the threshold for identifying traffic shows different application and it only considers correlation between traffic without using the signature identification method causing the accuracy to be low. Therefore, there is a limitation in the configuration of the overall traffic analysis system.

Existing researches suggest methods that mainly define the character of traffic to improve the pattern matching techniques and grouping to minimize the signature search range in order to improve the processing of the payload signature-based traffic identification system. However, these methods neither determine inefficient signatures nor quantify the value of each signature. Therefore, they could not achieve a lot of performance improvements in application traffic identification method or full traffic analysis system.

III. Problem Definition

Current payload signature generated by automatic traffic signature generation system is string that lists up the character coming out from the common traffic file which aims to identify. Among the automatically generated signature, some signatures may have enough quality as application signature, but other signatures might be nothing but common character string, simply have no meanings. Previously there was no clear criteria for evaluating the quality and the meaning of the signature.

As the result, unwanted signatures were also included in identification process. Nevertheless, all signatures satisfied high identification rate but the methods applied did not manage to determine which signatures are important and which ones are not.

In this paper, we propose a method to define this phenomenon as inefficiency of the signature. Quality evaluation method is to quantify the importance of each signature and find the efficient and significant signatures specifically.

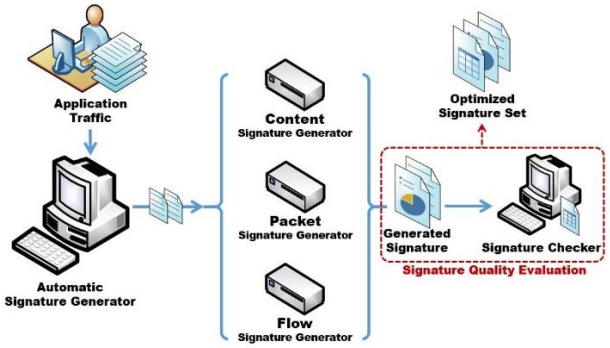

Figure 1. Proposed evaluation system

Figure 1 shows the current automatic payload signature generation system. When specific application traffic files are entered, strings that are common in each of the traffic file are extracted by the automatic signature generator. Consequently, flow, packet and content signatures for the particular application are generated. The core of the proposed method in this paper is the boxed part, the signature quality evaluation in Figure 1. The signature quality evaluation method is done based on three criteria: redundancy,

characteristic and performance. When signature checker receive the generated signature from the signature generator, it applies the proposed signature quality evaluation method to signatures. This is done through browsing the optimized signatures set which have higher efficiency and identification capacity than other signatures.

IV. PROPOSED QUALITY EVALUATION METHOD

The proposed quality evaluation method first quantifies the importance of each signature and second finds the efficient signatures based on the weight value. The quality evaluation process is performed through three-steps as shown in Figure 2. First, we apply three evaluation criteria to each signature and calculate the specific quality value of each criteria. Second, we calculate the signature quality weight by referring to criteria value. Third, we construct the Signature Map (S-Map) by using signature quality weight to search high quality signatures.

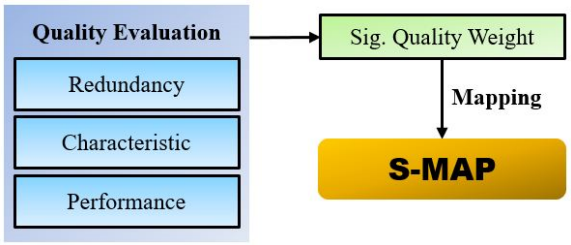

Figure 2. Signature quality process

The first criterion is the availability (Redundancy in Figure 2) of the signature. If a certain traffic unit such as flow or packet is matched by a number of signatures, the flow or packet are matched several times redundantly. Therefore, the redundancy value is relatively low. In the opposite manner if a signature matches a number of flows or packets, the signature will be determined as highly efficient with high utilization. The quality evaluation calculates the results of this utilization and availability of all signatures under Equation (1), defining the final result as 'Redundancy Value (RV)'.

$$\mathrm{RV(s)} = \frac{count(flows\ identified\ by\ s)}{count(flow)} \qquad (1)$$

The second criterion takes proportion of human readable characters in the content of signature into account in terms of uniqueness (Characteristic) of signature. Among the generated signatures, some signatures including padding bits or random value might have neither meaning nor uniqueness in respect of application traffic identification. Therefore, it is necessary for the method to find the uniqueness of the signature and its specific features. The quality evaluation assesses the presence of the application name and length of distinguishable character and numeric string. This criterion is defined and calculated under Equation (2) as 'Characteristic Value' of the signature.

$$\mathrm{CV(s)} = \begin{cases} \frac{len(readable\ characters\ in\ S)}{lengh(S)} \\ 1 \quad if\ app.name\ is\ in\ s \end{cases} \qquad (2)$$

The third criterion is the signature matching rate (Performance). In order to determine the signature matching rate, a matching offset should be considered. That means consideration of an offset location in the packet payload where the signature appears. The reason why consider signature matching offset and searching depth is for reducing the time required for the traffic identification and improving the total traffic identification system performance. If an offset is fixed to the front of payload, signature matching process will be done more quickly and it can result in improving of the traffic identification system performance. Therefore, the faster the speed of the signature matching the better the signature defined. The 'Performance Value' criterion is calculated under Equation (3).

$$\mathrm{PV(s)} = \frac{\mathrm{max}len(payload\ with\ s) - maxoff(S)}{maxlen(payload\ with\ s)} \qquad (3)$$

As a result, the quality evaluation calculates the quality weight of each signature on the basis of utilization (Redundancy), uniqueness (Characteristic) and the signature matching rate (Performance) criteria, using the Equation (4).

$$\mathrm{QV(S)} = PV(s) + CV(s) + RV(s) \qquad (4)$$

We calculate the quality weight value on the basis of defined three criteria. Calculation method of quality weight is shown in Equation (4). The reason for configuration shown in Equation (4) is to consider the all three quality evaluation criteria value equally. Hence, quality weight value has range from minimum 0 to maximum 3.

Table 1. S-MAP Table(Example)

Sig ID	QV Weight	Identified Flow ID
1	2.7	3,5,7,10,11,12
2	2.5	2,5,9,11
3	2.1	3,5
4	0.6	10,12
5	2.3	1,3,5,7,9,10
6	0.8	11
7	0.5	6
8	2.9	1,2,5,8,9,12
9	0.9	5
10	2.6	2,4,5,6,9
11	1.1	4
12	0.3	1

Quality weight value as shown in Figure 2, is used to make S-MAP. The S-MAP is configured based on the quality weight value defined above by the quality evaluation method.

Table 1 and Figure 3 show an example of S-MAP. As shown in Table 1, each signature has its own quality weight and identified flow (packet or content) ID. In Figure 3, two-dimensional map is drawn based on Table 1. Each matched point (red dot) indicates identified flow ID and has quality weight value. For searching the optimal signature set, we draw a line connecting the point which has the highest quality weight in the Y-axis along the sequence (X-axis).

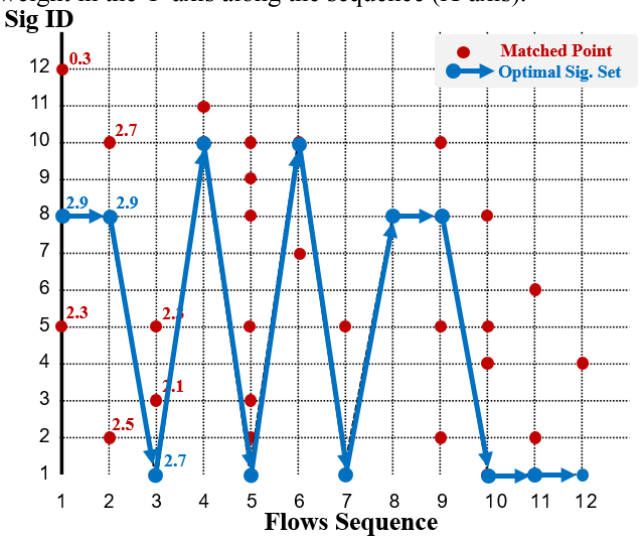

Figure 3. S-MAP(Example)

When referring to Figure 3, the line of each point is connected along the sequence (X-axis) and the signature set used at this time (Y-axis) is {1, 8, 10} signature set. Thus, the signatures of the high efficiency are 1, 8 and 10 signature. In this case the efficiency of the selected signatures is obtained through the signature average efficiency (SAE) under Equation (5).

$$\text{SAE} = \frac{\text{count(Flows identified by S)}}{count(\text{S})} \qquad (5)$$

SAE (Signature Average Efficiency) refers to average number of flows identified by a signature set S. Since, the number of signatures have been reduced to a quarter value after quality evaluation compared with existing signature set (from 12 to 3), SAE is increased four times. Therefore, the signature set searched through the S-MAP has fourfold higher efficiency on average than previous existing signature set.

V. EVALUATION

In this section, we apply the proposed method to traffic data collected from a real campus network, and we then prove the validity of the method.

Table 2. Traffic Trace

Trace ID	Size(MB)	Flow	Pkt
1	113	2500	114,705
2	110	405	104,785
3	46	1452	48,867
4	34	1503	52,069
5	55	501	52,302
6	54	646	52,048

Traffic used for the quality evaluation are shown in Table 2. All 6 traces are *Torrent* application traffic collected while downloading multimedia files.

We entered the traffic traces of Table 2 to the system shown in Figure 1 and applied the quality evaluation to the signatures made. As a result, we got the results as shown in Figure 4, 5. Signatures used for the evaluation are Content Signature in Figure 1. The reason why we used the lowest level Content Signature is the fact that, it can be easily applied to higher packet and flow level signatures when an efficient signature set is selected.

Figure 4 is an S-MAP implementation through the method defined in the previous section. A total of 139 content signature and approximately 14,000 packets sequence were generated by the traffic traces in Table 2. They were reduced to 24 signature by applying the proposed quality evaluation.

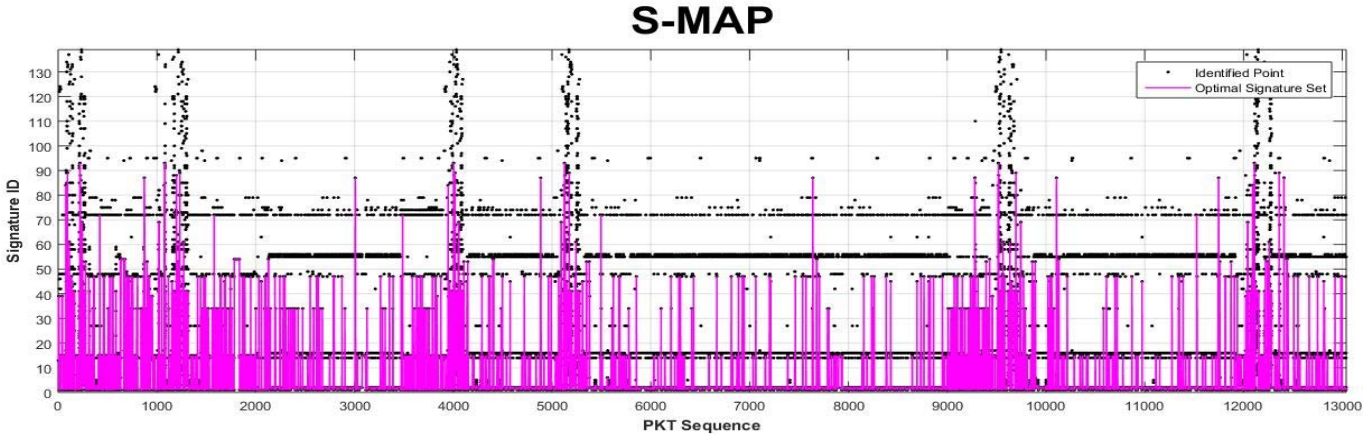

Figure 4. S-MAP

As a result of the previously defined aspects SAE, about four times more efficient signature set were searched compared to the existing signature set that does not apply the quality evaluation. In Figure 4, Each point refers to the sequence of packets identified by the signature in the S-MAP. As shown in Figure 4, the actual level of connected points of the Y-axis was observed to be 24 levels when confirmed by S-MAP.

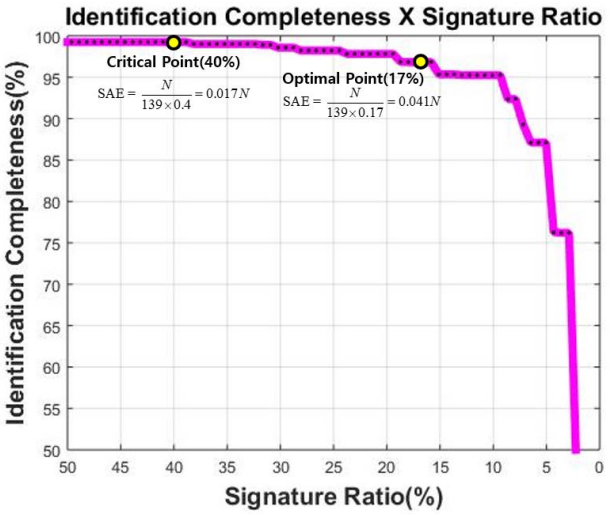

Figure 4. Signature ratio according to identification rate

In Figure 5, x-axis shows the signature ratio by contrast to total generated signatures and y-axis indicate the identification completeness according to the signature ratio. When considering the generated 139 signatures to 100%, exclude low quality weight signature sequentially from total signature set until identification completeness is lower. The point which identification completeness begins lower is defined as critical point. At critical point, we obtained a reduced signature set. Referring to Figure 5, at the time when the signature ratio is 40%, it can be seen that the identification rate gradually declined. By maintaining the identification rate of the 139 signatures, we managed to reduce them to 55 signatures which is 40% of total signatures. However, when using the proposed method for quality evaluation, 24 signatures are found at the same time making 23% lower signature ratio than the critical point. We define that point as optimal point. In optimal point, however identification completeness is lower less than 3% compared with critical point.

When the signature ratio is lower than 17%, Signatures with a higher quality weight value which is optimal signatures are excluded from the optimal signature set. And when the time that the signature ratio is about 10%, it can be seen that identification completeness is drastically decreasing. It means that optimal point is well designated point and signature quality evaluation is done carefully.

When calculated based on the previously defined SAE, we assume that the number of identified packets in the critical point is *N*. In Figure 5, critical point has a value of SAE 0.017*N*. On the other hand, the SAE value of optimal point has an 0.04*N*. Thus, it was clearly possible to search for the optimal signature set having high traffic identification efficiency of about three times more than the critical point signature set.

Furthermore, when comparing with total signature set having 100% signature ratio, optimal signature set has six-fold higher traffic identification efficiency.

VI. Conclusion and Future Work

In this paper, we proposed finding highly efficient application traffic payload signature set through quality evaluation method.

As a result, a set of optimal signatures is found through the quality evaluation. In terms of identification completeness optimal signature set has a high traffic identification efficiency almost equal to original signature set. In addition, when considering traffic identification capacity of the signature, optimal signature set found by quality evaluation has four-fold higher traffic identification capacity than non-evaluated signature set in average.

The proposed payload signature quality evaluation method for application traffic identification shows to be very positive in terms of the traffic identification. However, our main future work is to correct some weaknesses and reinforce the heuristic process of the evaluation method. Moreover, we plan to apply the proposed quality evaluation method to a real time automatic signature generating system in future.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2015R1D1A3A01018057).

References

[1] J. S. Park, J. W. Park, S. H. Yoon, Y. S. Oh, M. S. Kim, "Development of signature Generation system and Verification Network for Application Level Traffic Classification", in Proc. KIPS conf. Apr. 23-24, 2009, pp. 1288-1291, PuSan, Korea.

[2] S. H. Yoon, H. G. Roh, M. S. Kim, "Internet Application Traffic Classification using Traffic Measurement Agent ", in Proc. KICS Jul. 2-4, 2008, pp.618. Jeju Island, Korea.

[3] Fnag Yu, Zhifeng Chen, Yanlei Dino, T. V. Lakshman, Randy H. Katz, "Fast and memory Efficient Regular Expression Matching for Deep Packet Inspection" ANCS 2006, December , 2006, San jose, California USA.

[4] Christopher L. Hayes , Yan Luo, "DPICO: a high speed deep packet inspection engine using compact finite automata", ACM/IEEE Symposium on Architecture for networking and communications systems, December 03-04, 2007, Orlando, Florida, USA

[5] C. L. Hayes and Y. Luo, "DPICO: A high speed deep packet inspection engine using compact finite automata," in Proc. ACM/IEEE Symp. Architecture Netw. Commun Syst. (ANCS '07), pp. 195-203, Orlando, USA, Dec. 2007.

[6] J. S. Park and M. S. Kim, "Performance improvement of application-level traffic classification system using application traffic pattern," in Proc. KICS Summer Conf., pp. 3-7, Jeju, Korea, Jun. 2011.

[7] J.-S. Park, S.-H. Yoon, and M.-S. Kim, “Performance improvement of the payload signature based traffic classification system using application traffic locality,” J. KICS, vol. 38B, no. 7, pp. 519-525, Jul. 2013.