Efficient Payload Signature Structure for Performance Improvement of Traffic Identification

Woo-Suk Jung[1], Jun-Sang Park[1] and Myung-Sup Kim[1]
[1]Dept. of Computer and Information Science
Korea University
Korea
{hary5832, junsang_park, tmskim}@korea.ac.kr

Jae-Hyun Ham[1,2]
[2]The 2nd R&D Institute-1
Agency for Defense Development
Korea
jhham@korea.ac.kr

***Abstract*— The traffic identification is a preliminary and essential step for stable network service provision and efficient network resource management. While a number of identification methods have been introduced in literature, the payload signature-based identification method shows the highest performance in terms of accuracy, completeness, and practicality. However, the payload signature-based method's processing speed is much slower than other identification method such as header-based and statistical methods. In this paper, we first classifies signatures by matching type based on range, order, and direction of packet in a flow when each signature matches to payload. By using this classification, we suggest a novel method to improve processing speed of payload signature-based identification by reducing searching space.**

***Keywords*—traffic analysis; signature matching type; payload signature; processing speed**

I. Introduction

As network acceleration and development of various services and applications, dependency on network represented by internet was growing to companies and individuals. Application-level traffic monitoring and analysis for the efficient operation and management of the network, the need is growing in various fields such as management and usage identify and develop plans to expand the network. In this reason, method which can classify a various application level traffic in accurate and fast time is required. While a number of classification methods have been introduced in the literature, the payload signature-based classification method shows the highest performance in terms of accuracy, completeness, and practicality [1,2]. However, the processing speed of the current payload signature-based classification system is insufficient for real-time handling of the huge amounts of traffic data in high-speed networks. Given the increasing number of applications as well as the increasing use of applications that generate large amounts of traffic, the inadequate processing speeds of payload-based analysis is a problem that must be solved. Ongoing studies on payload signature-based classification systems aim to accelerate the classification process. However, most studies focus on improving the performance of the pattern-matching algorithm [3-6]. Performance improvement of pattern-matching algorithm is restricted. Generally algorithm check entire payload in the process of matching the signature to the payload. If the number of payload signature increases application traffic analysis rate is increased, but number of searching the unnecessary range is increased and it cause reduce of performance.

To improve the performance of payload signature-based classification, IDS such as snort restricted searching range by using offset rule. But use of offset rule in snort is not essential but an option, so the quality of the signature is different by the experience or ability of person who made the signature. Also, when we extract the offset manually an accuracy and time problem arises.

In this paper, we automate the extraction process, and propose a method for optimizing the searching range of pattern-matching algorithm by categorizing the signature matching type. It restrict the searching range more efficient than snort.

The remainder of this paper is organized as follows. Section 2 describes related research. Section 3 describes the problem of traditional payload signature-based classification method. In section 4, our solution based on categorizing the signature matching type. In section 5, the proposed method is applied to our classification system and its validity is proven. Finally, Section 6 describes conclusions and future research directions.

II. Related Work

Many applications try to bypass the firewall for a seamless service by frequently changing their traffic patterns. For this reason, the signatures that identify these applications from traffic data appear in complex forms. In addition, due to the increase in network-based applications and L7 protocols, the number of signatures necessary for identifying applications has been increasing. As the number of signatures and their

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 by Brain Korea 21 Plus (BK21+).

complexity increase, the processing speed of the payload signature-based classification becomes an important element in determining the performance of the traffic classification system. Many ongoing studies on payload signature-based classification systems aim to accelerate the classification process. However, most studies focus on improving the performance of the pattern-matching algorithm. But the performance of pattern-matching algorithm is wholly dependent on configuration of input data, resulting in limited performance improvement [4]. Many methodologies were proposed to improve the performance for NFA and DFA algorithm which were automata-based pattern-matching algorithm. But automata-based methodology has a performance decline problem depends on frequency of use the wildcard such as '.*' [5,6]. In studies for defining the factors that affect the processing speed of the analysis system and presenting the best classification structure to improve the processing speed proposed the removal of redundant signature to minimize the searching range of the input data, way to structure a hierarchical signature and way to limit the number of the packet to check for the optimization of the searching range [7,8]. Many ongoing studies to improve performance on payload signature-based classification system aims to improve pattern-matching techniques by software and hardware, and by defining the characteristics of traffic and grouping them. However, this method seems limited improve performance relative to the increase of network bandwidth, or it is difficult to apply to the current network environment.

In this paper, we automate the extraction process for extract the correct offset. Also, we proposed categorizing the signature matching type to optimized searching range by using offset value such as order of packet, transmission direction for improve the performance.

III. PROBLEM OF TRADITIONAL PAYLOAD SIGNATURE-BASED CLASSIFICATION

In this chapter we describes the problems with the existing signature-matching method and propose why categorizing the signature matching type is necessary.

In normal application identification system use the partial matching method which end the matching when application signature is matched to the flow. Figure 1 represent a searching range of normal traffic identification. When the signature is matched, system search from beginning of the payload to the matching point. If the signature is not matched, system search the whole of the payload and this affects the analysis time.

To solve the performance degradation problem cause by unnecessary searching, IDS such as snort offers the offset rule which can restrict offset, transmission direction and depth. In snort offset rule is not essential but an option, and this make a difference in searching range of signature. Also, offset accuracy and process time problem can be caused by manually operated extraction process.

For categorizing the signature matching type, we extract the offset value such as order, direction and offset information of packet when signature matched and we automate this process for quick and accurate extraction. Categorizing the signature into the type by using extracted information such as order of packet, direction and matching range, and assigned the searching range that was optimized for each type of signature then we can reduce unnecessary searching range, and it can lead to improve of traffic identification performance

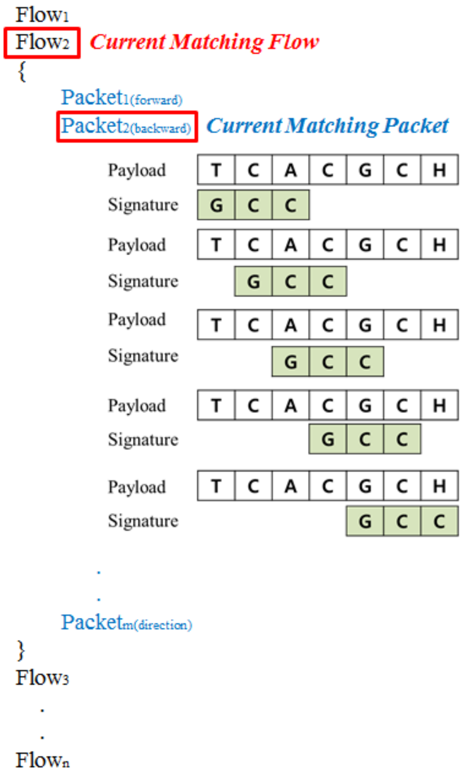


Figure 1. Searching range of established traffic analysis

IV. CATEGORIZATION OF SIGNATURE MATCHING TYPE

In this chapter, we describe how to classify signatures by matching type based on matching range, order, and direction of packet in a flow when each signature matched to payload.

Figure 2 is a diagram of the signature categorizing system. First, we get the traffic trace and signature list as an input data and extract the order, direction and matching range of packet from automatic offset generator. Second, classify the signature by its matching type. Third, we analyze the traffic by considering classified signature matching type.

A. Offset value

This section describes the matching offset value of the signature that is used as the parameter value of the signature algorithm for divide into six types by matching type.

Table 1 is a summary of the offset value. Packet offset is used when the order of matched packet is fixed. Direction is used when the transmission direction of matched packed is fixed. First offset is used when offset of matched packet is fixed. First range is used when offset of matched packet is not fixed. Last offset is maximum matching termination value of matched packet. Depth is obtained by subtracting the minimum matching starting offset from maximum matching terminating offset.

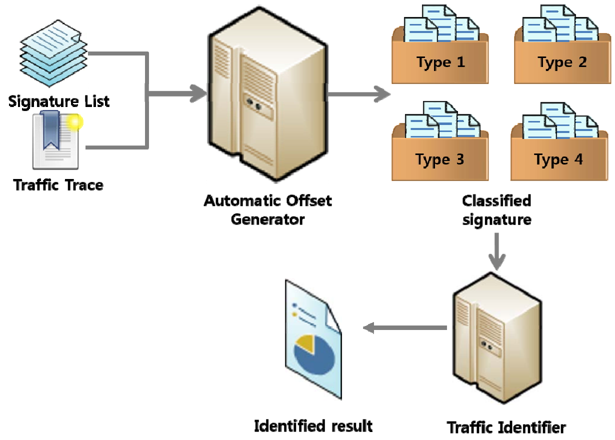


Figure 2. Diagram of signature categorizing system

Figure 3 is an algorithm which decide the offset value and searching range by receive the signature matching offset information as an input data. Decision of searching range is roughly divided into three cases, when the packet offset is fixed, when the direction is fixed and both variables are not fixed. If the set of packet offset has only one value then we extract it as packet order. Else if set of direction has only one value then we extract it as direction. And if the matching offset of signature is fixed we extract first offset, else if we extract the first range.

Table 1. Definition of Offset value which used in signature categorizing

Offset value	explanation
Packet Offset	Order value of packet when signature matching sequence of packet is fixed
Direction	Value that packet's transmission direction is to-server or to-client when signature matched
First Offset	Position value when starting position of signature matching is fixed
First Range	Minimum position value when starting position of signature matching is not-fixed
Last Offset	Last position value when signature matched
Depth	Range information which signature is matched. (Last Offset minus First Offset or Last Offset minus First Range)

Figure 4 is an example of signature model which was made, including signature matching type information and offset value. We can interpret that ".*BitTorrent protocol.*" signature is determined signature matching type 1, the transmission direction of the packet destined for a server and has a range from offset value 2 to length of 19.

```
mS: Flow matched signature
PO(mS) : Packet offset set of mS
Dir(mS) : Direction set of mS
FO(mS) : first offset set of mS
LO(mS) : Last offset set of mS
Input : Signature info container
Output : Signature matching range

for each mSi do
    //first offset decision
    if| FO(mSi) | == 1
        FO(mSi) is fixed offset
    else
        find FOmin(mSi) value among FO(mSi)
    //Last offset decision
        find LOmax(mSi) value among LO(mSi)
    //Packet offset decision
    if| PO(mSi) | == 1
        mSi is matched only in PO(mSi) then,
        //range decision
        FOmin(mSi) ≤mSi.Range ≤LOmax(mSi)
    //Packet Direction decision
    else if| Dir(mSi) | == 1
        mSi is matched only in Dir(mSi) then,
        //range decision
        FOmin(mSi) ≤mSi.Range ≤LOmax(mSi)
    else
        //range decision
        FOmin(mSi) ≤mSi.Range ≤LOmax(mSi)
done
```

Figure 3. Searching range decision algorithm for automatic Offset extraction

```
payload=".*BitTorrent protocol.*"
Offset_Type="1" Direction="forward" First_Offset="2" Depth="19"
```

Figure 4. Signature model which contains Offset value

A. *Type classification by Offset value*

Table 2 is a type decision table by using six values that we explained. We have six types for classification. In type 1, 2 the packet offset is fixed. In type 3, 4 the direction of packet is fixed. In type 5, 6 both packet offset and direction of packet are not fixed.

Table 2. Signature type decision table

Pattern Type	Packet Offset	Direction	First Offset	First Range	Depth
Type 1	O	X	O	X	O
Type 2	O	X	X	O	O
Type 3	X	O	O	X	O
Type 4	X	O	X	O	O
Type 5	X	X	O	X	O
Type 6	X	X	X	O	O

Figure 5 is a searching range when packet offset is fixed. We simply search the corresponded packet because the order of packet is fixed.

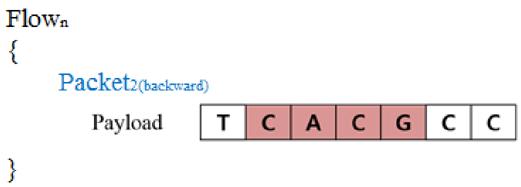


Figure 5. Searching range of signature matching type 1, 2.

In type 3, 4 the direction of matched packet is fixed, so we simply search the packet which the direction is same as matched packet. In type 5, 6 both packet offset and direction are not fixed, so we have to search every packet of each flows. The type which packet offset is fixed, direction is fixed and else can divided into two cases whether it has first offset or first range. When signature have first offset, we can check success or failure of matching by comparing the first byte which match has started. If the signature has first range, we search from first range to length of the depth. The faster the type number narrow the search range.

V. PERFORMANCE EVALUATION

In this chapter, we applied the signature classification by its matching type which was described at chapter 4 to traffic identification, and evaluate the performance by comparing the traditional identification method and the proposed method. In the experiment, we used the offset value of signature which was gained through the target traffic trace analysis for categorizing the signature by its matching type. Then applies the categorized signature to the traffic identification and evaluate whether proposed method is much more efficient than traditional method.

We determine the performance based on the analysis rate and analysis time. Analysis rate measured that how much did the signature analysis the traffic. Analysis time measured the user-time that CPU actually takes in the process of matching the signature in identification system.

A. *Experiment environment*

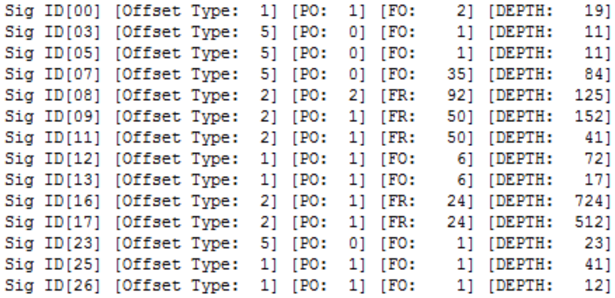
In this section, we described the experiment environment for performance evaluate experiment of traffic identification by using the categorizing the signature matching type. We collected the traffic by using WireShark and Net Monitor, and exclude the other application traffic except torrent application traffic. Collected traffic was 15 total traces and it was used to test set for performance evaluation. Table 3 is an information of test set. Table 4 shows 32 signatures that we used in our experiment. Every 32 signatures were used for analysis the torrent application traffic.

Table 3. Test set information which used in experiment

Traffic Trace ID	Content Size(MB)	Duration (min.)	Flow	Pkt	Byte
006_UT_FP02	119	24	10,874	2.1E+06	1.8E+09
006_UT_FP03	42.1	6	3,988	1.0E+06	8.4E+08
006_UT_FP04	355	7	2,996	1.5E+06	1.3E+09
006_UT_FP05	46.1	9	2,961	1.1E+06	9.7E+07
006_BT_FP02	119	18	4,946	2.0E+06	1.8E+09
006_BT_FP03	42.1	8	3,329	9.4E+05	8.4E+08
006_BT_FP04	355	10	3,603	1.5E+06	1.3E+09
006_BT_FP05	46.1	5	2,619	1.3E+05	1.2E+09
008_UT_FP06	20	1	278	4.8E+04	3.5E+07
008_UT_FP07	961	21	2,505	1.1E+06	1.2E+09
008_BT_FP08	899	5	1,760	5.3E+05	4.5E+08
008_BT_FP09	1,490	3	2,262	6.2E+05	6.8E+08
008_UT_FP10	1,520	1	1,363	1.0E+06	9.6E+08
008_UT_FP11	25	1	330	3.3E+04	2.8E+07
008_UT_FP13	954	5	3,204	1.0E+06	8.5E+08
Total	6693.4	124	47,018	1.46E+07	1.34E+10

B. *Experiment result*

Figure 6 shows the offset value which was extracted from 006_UT_FP02 trace analysis by signatures in Table 4. 14 of 32 signature were matched in 006_UT_FP02 trace. Matching type 1 which can reduce the searching range most contains 5 signatures, matching type 2 contains 5 signatures and matching type 5 contains 4 signatures.



```
Sig ID[00] [Offset Type:  1] [PO:  1] [FO:    2] [DEPTH:   19]
Sig ID[03] [Offset Type:  5] [PO:  0] [FO:    1] [DEPTH:   11]
Sig ID[05] [Offset Type:  5] [PO:  0] [FO:    1] [DEPTH:   11]
Sig ID[07] [Offset Type:  5] [PO:  0] [FO:   35] [DEPTH:   84]
Sig ID[08] [Offset Type:  2] [PO:  2] [FR:   92] [DEPTH:  125]
Sig ID[09] [Offset Type:  2] [PO:  1] [FR:   50] [DEPTH:  152]
Sig ID[11] [Offset Type:  2] [PO:  1] [FR:   50] [DEPTH:   41]
Sig ID[12] [Offset Type:  1] [PO:  1] [FO:    6] [DEPTH:   72]
Sig ID[13] [Offset Type:  1] [PO:  1] [FO:    6] [DEPTH:   17]
Sig ID[16] [Offset Type:  2] [PO:  1] [FR:   24] [DEPTH:  724]
Sig ID[17] [Offset Type:  2] [PO:  1] [FR:   24] [DEPTH:  512]
Sig ID[23] [Offset Type:  5] [PO:  0] [FO:    1] [DEPTH:   23]
Sig ID[25] [Offset Type:  1] [PO:  1] [FO:    1] [DEPTH:   41]
Sig ID[26] [Offset Type:  1] [PO:  1] [FO:    1] [DEPTH:   12]
```

Figure 6. Offset value information extracted from 006_UT_FP02 trace

Table 5 and Figure 7 shows a result of the comparison between traditional traffic identification and proposed traffic identification which applies categorizing the signature matching type. Overall, the analysis rate of the proposed method compared to the traditional method is the same, and analysis time reduced to 20% on average and 41% on maximum.

As a result, same analysis rate between the traffic identification which used categorizing the signature matching type and which did not means that reduction of searching range was exactly apply to unnecessary searching range.

Table 4. Torrent signature which used in experiment

Sig ID	Payload signature	Protocol	Port
0	.*BitTorrent protocol.*	TCP	N/A
1	.*BitTorrent protocol.*	UDP	N/A
2	.*d1:ad2:id20.*	TCP	N/A
3	^d1:ad2:id20.*	UDP	N/A
4	.*d1:rd2:id20.*	TCP	N/A
5	^d1:rd2:id20.*	UDP	N/A
6	.*find_node.*UT.*	UDP	N/A
7	.*info_hash.*get_peer.*	UDP	N/A
8	.*5:peers.*	TCP	N/A
9	.*User-Agent:.*uTorrent.*	TCP	N/A
10	.*Host: com-utorrent.*	TCP	N/A
11	.*User-Agent: BTWebClient.*	TCP	N/A
12	.* //announce.info_hash=.*peer_id=.*	TCP	N/A
13	.*//scrape.info_hash=.*	TCP	N/A
14	.*d5:added.*	TCP	N/A
15	.*d5:added.*	UDP	N/A
16	.*Host:.*utorrent\.com*."	TCP	N/A
17	.*Host:.*bittorrent\.com*."	TCP	N/A
18	.*bittorrent.com.	UDP	N/A
19	.*tracker.*	UDP	N/A
20	*BT.*announce	UDP	N/A
21	.*UT.*announce.*	UDP	N/A
22	.*d2:ip6:.*1:rd2:id20.*	UDP	N/A
23	^d2:ip6:.*1:rd2:id20.*	UDP	N/A
24	^......................BT.....*	UDP	N/A
25	^......................UT.....*	UDP	N/A
26	.*\x00\x00\x04\x17\x27\x10\x19\x80\x00\x00\x00.*	UDP	N/A
27	^BT-SEARCH.*	UDP	N/A
28	.*BitTorrent.*	UDP	1900
29	.*uTorrent.*	UDP	1900
30	.*Host:.*deluge-torrent\.org.*	TCP	N/A
31	.*User-Agent: Deluge.*	TCP	N/A

Restricting the order of packet is very efficient in optimizing the searching range, but it is difficult to apply to snort because snort is packet based analysis system. In point of performance improvement, flow based analysis system that can efficiently use these various offset value is demanded. As a result, we shows that categorizing the signature type was efficient which was defined and purposed in this paper.

Table 5. Torrent signature which used in experiment

Traffic Trace ID	not apply signature matching type categorizing		apply signature matching type categorizing	
	Analysis rate(%)	Analysis time(sec)	Analysis rate(%)	Analysis time(sec)
006_UT_FP02	94.64	0.59	94.64	0.49
006_UT_FP03	98.57	0.23	98.57	0.28
006_UT_FP04	99.52	0.18	99.50	0.14
006_UT_FP05	99.32	0.17	99.32	0.12
006_BT_FP02	89.77	0.31	89.77	0.24
006_BT_FP03	96.67	0.19	96.67	0.15
006_BT_FP04	98.36	0.21	98.36	0.17
006_BT_FP05	99.01	0.14	99.01	0.11
008_UT_FP06	85.61	0.02	85.61	0.02
008_UT_FP07	99.28	0.25	99.28	0.19
008_BT_FP08	85.11	0.12	85.11	0.10
008_BT_FP09	96.91	0.14	96.91	0.10
008_UT_FP10	84.23	0.08	84.23	0.07
008_UT_FP11	94.24	0.03	94.24	0.02
008_UT_FP13	85.96	0.22	85.96	0.17
Total	94	2.88	94	2.27

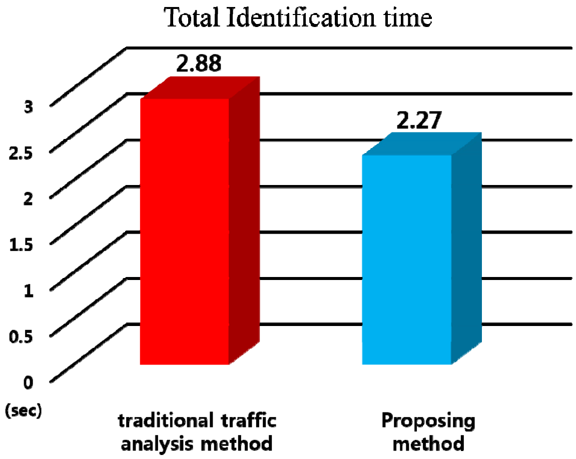


Figure 7. Total analysis time comparing

VI. Conclusion and Future Work

In this paper, we extract more various offset value such as packet order and direction automatically. And we proposed traffic identification by categorizing the signature matching type into 6 types. We shows that proposed traffic identification maintains the analysis rate and reduce the analysis time to 20% on average and 41% on maximum. Through these result, we draw a conclusion that restriction of searching range only applied to unnecessary searching range. Also, we show that use of offset rule in snort is not an option but an essential and we need systematic signature structure for restrict the searching range. In future work, we plans to study about definition of multi sequence signature structure that several single sequence signature were connected by ".*".

References

[1] J. S. Park, J. W. Park, S. H. Yoon, Y. S. Oh, and M. S. Kim, "Development of signature generation system and verification network for application level traffic classification," in Proc. KIPS Conf., pp. 1288-1291, Pusan, Korea, Apr. 2009.

[2] S. H. Yoon, H. G. Roh, and M. S. Kim, "Internet application traffic classification using traffic measurement agent," in Proc. KICS Summer Conf., pp. 1747-1750, Jeju Island, Korea, July 2008.

[3] F. Yu, Z. Chen, Y. Dino, T. V. Lakshman, and R. H. Katz, "Fast and memory efficient regular expression matching for deep packet inspection," in Proc. ACM/IEEE Symp. Architecture Networking Commun. Syst. (ANCS '06), pp. 93-102, San Jose, U.S.A., Dec. 2006.

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

[5] G. Vasiliadis, M. Polychronakis, S. Antonatos, E. P. Markatos, and S. Ioannidis, "Regular expression matching on graphics hardware for intrusion detection," in Proc. 12th Int. Symp. Recent Advances Intrusion Detection (RAID '09), pp. 265–283, Saint-Malo, France, Sep. 2009.

[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to Algorithms, 2nd Ed., MIT Press and McGraw-Hill, 2001.

[7] J.-S. Park, S.-H. Yoon, J.-W. Park, H.-S. Lee, S.-W. Lee, and M.-S. Kim, "Performance Improvement of the Payload Signature based Traffic Classification System." in Proc. KICS journal, vol. 35, no. 09, pp1287-1294. Sep. 2010.

[8] J.-S. Park, S.-H. Yoon, and M.-S. Kim, "Performance Improvement of Signature-based Traffic Classification System by Optimizing the Search Space." in Proc. KSII journal vol.12 no.3, pp. 89-99. Jun. 2011.