High performance payload signature-based Internet traffic classification system

Sung-Ho Lee, Jun-Sang Park, Sung-Ho Yoon, and Myung-Sup Kim
Dept. of Computer and Information Science
Korea University
Korea
{gaek5, junsang_park, sungho_yoon, tmskim}@korea.ac.kr

Abstract— **Internet traffic classification is an essential step for stable service provision and efficient network management. The payload signature-based-classifier is considered as a reliable method for Internet traffic classification, but is prohibitively and computationally expensive for real-time handling of large amounts of traffic on high-speed network. To solve this problem, most studies focused on the pattern matching algorithm or hardware-based approaches such as FPGA and network processor. However, in order to improve the performance of the classification system, It is also necessary to consider the classification criteria and signature model in accordance with the characteristics of various application protocols. In this paper, we newly define the classification criteria and signature model, and propose an optimized classification architecture in perspective of input data minimization and complexity of pattern matching algorithm to improve the processing speed of classification system. Each of them can be applied individually, or in any combination. The proposed method achieved an approximately 5-fold increase in processing speed over existing baseline classification system.**

Keyword—Traffic Classification; Payload Signature; Processing speed

I. INTRODUCTION

As individual and corporate users are becoming increasingly dependent on the Internet, network speeds are increasing and a variety of services and applications are being developed. Thus, there is a growing need for monitoring and analyzing Internet traffic from an application perspective, in order to achieve efficient network operation and management in various areas such as pay-for billing, CRM (Customer Relationship Management), SLA (Service Level Agreement), etc.

The payload signature-based classification method has been known to exhibit the highest levels of performance in terms of accuracy, completeness, and practicality [1-4]. However, the processing speed of the method is not sufficiently fast for the real-time handling of the large volume of traffic data generated from high-speed networks [5-8]. Taking into consideration of the increasing number of Internet-based applications and the expanding use of applications that generate high volumes of traffic, the inadequate processing speeds of payload-based analysis systems are an important issue that needs to be addressed.

The signature model without considering the characteristic of the application-level protocol is an important cause of performance bottleneck of classification system, so It is necessary to use the field-based signature model(e.g. URI, Host, User-agent) in order to classify HTTP traffic into detailed services. This approach will improve the accuracy and processing speed of the classification system.

In this paper, we suggest a multi-dimensional criteria for traffic-classification that considers the various utilization purposes of the traffic-classification. Furthermore, we define 3 types of signature model based on each protocol characteristic. The proposed signature models are applied to the different classification routines in accordance with the input data and pattern matching algorithm to improve the processing speed. Finally, we propose an optimal classification system architecture in terms of classification criteria, signature model, input data minimization and complexity of pattern matching algorithm to maximize the processing speed of classification system. The proposed method achieved an approximately 5-fold increase in processing speed over the baseline classification system.

II. PROPOSED CLASSIFICATION SYSTEM

In this section, we describe the processes implemented in the proposed classification method to improve the processing speed of the payload signature-based classification system.

Figure 1 shows the diagram of classification system. Preprocessing module separate the HTTP traffic from the input trace and parse the HTTP traffic by each field to be applied to the field-based simple string. We have to predefine the classification criteria and signature model. The signature model determines input data minimization method and pattern matching algorithms.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2015R1D1A3A01018057), Next-Generation Information Computing Development Program through the National Research foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2010-0020728).

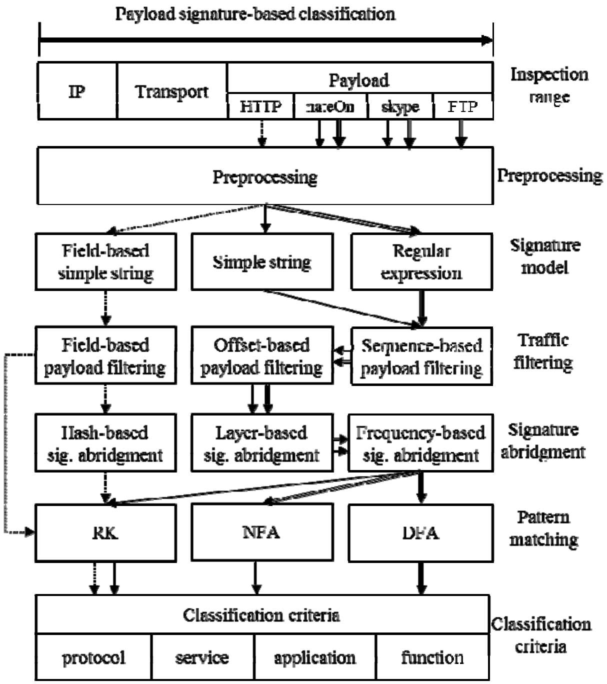

Figure 1. Diagram of classification system

A. Signature model

Payload-signature model should define the extraction position from the payload and the representation method that reflects the feature of the application-level protocol that is the purpose of the classification. In this paper, we use field-based simple string, simple string and regular expression.

- *Field-based simple string:* A signature that consists of a sequence of characters in an application-level protocol field(e.g. HTTP Host, User-agent).
- *Simple string:* A signature that consists of a sequence of characters in a payload. And it can appear at any position in a payload.
- *Regular expression:* A signature that consists of a set of sequence of characters and includes wildcard characters in a signature.

Table 1 describes the signature model and examples. The applications using HTTP protocol such as YouTube are suitable to use field-based simple string model. The simple string is most appropriate for applications which signature extracted a sequence of characters that appear in a specific location in a payload. Regular expression signature model is useful for repetition of characters within a specific range, pattern with length constraints and wildcards such as skype.

Table 1. Signature model

Model	Example	Criteria
Field-based simple string	User-agent : Mozilla; Host : www; Domain : YouTube; Uri : login	app:IE svc:YouTube prot:HTTP funct:login
Simple string	string : NCPT;	app:NateOn svc:nate prot:NateOn funct:login
regular expression	RE : ^(\x01.? .? .? .? .? .? .? .?\x01| \x02.? .? .? .? .? .? .? .?\x02| … \xff.? .? .? .? .? .? .? .?\x0ff)	app:skype svc:skype prot:skype funct:voice-call

B. Input data minimization

We propose the input data minimization method to optimize the search space of traffic trace and signatures. Figure 2 shows the input traffic and signature minimization methods that separated whether HTTP traffic or not.

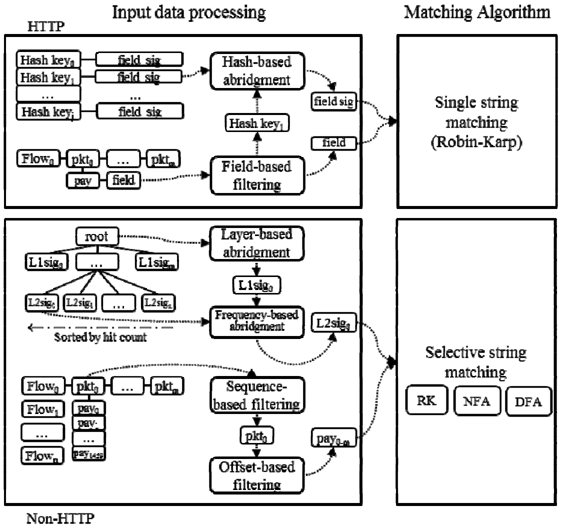

Figure 2. input data filter and pattern matching algorithm

Our classification system uses the hash-based, layer-based and frequency-based abridgment method in order to minimize the signature search space and field-based, sequence-based and offset-based filtering to minimize the traffic search space.

Hash-based signature abridgment

We define a payload-signature model that can classify HTTP traffic into multi-dimensional classes. HTTP traffic classification is achieved by minimizing the load obtained from the pattern matching by performing the pattern matching after comparing the hash key of a string during the pattern-matching process. For this approach, we need to preprocessing module that parse the traffic trace by HTTP protocol field to match the signature. The method for parsing is similar to the signature-extraction method. And field-based abridgment converts the strings that are parsed by field to key values through hash function. Hash-based abridgment module searches the same key value in the signature hash table and forwards field signature to pattern matching module. However, the hash-based signature abridgment method will not apply to all cases. The string that is extracted from matching the traffic and the string that is extracted when the signature is created should be extracted under the same policy without an intervention of administrator. The processing of URI signature extraction requires the intervention of an administrator. Therefore, the signature field values and matching traffic can be different. Thus, we do not compose the URI fields using hash-key value. The URI signature is composed of a string form that is not a key value. The URI field cannot be key matched with the URI field that exists in the HTTP traffic because it does not use the entire field as a signature; rather, it defines the necessary parts according to the administrator's decision.

Layer-based signature abridgment

We propose two-level hierarchical signature structure to reduce the signature search space and to determine an application protocol name and an application name for each flow. This consists of application protocol-level signatures at the first level, and application-level signatures at the second level. An application-level protocol could be commonly used by a number of applications for various purposes. In our signature hierarchy, the RTSP traffic is detected at the first level and then the application name is determined at the second level. The signature hierarchy is defined by an inclusion relationship. If all of the traffic identified using signature S_X can be classified using signature S_Y, then S_Y includes S_X. S_Y is termed an application protocol-level signature and S_X is termed an application-level signature.

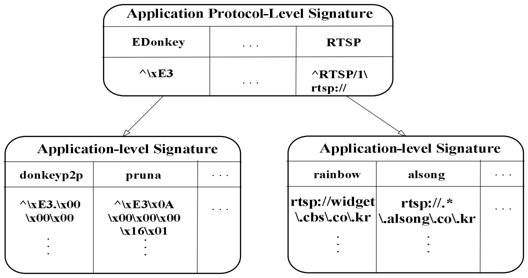

Figure 3. Two-level hierarchical signature structure.

In the two-level hierarchical signature structure, the classification system first identifies the input flow via the application protocol-level signature. If a flow was classified by an application protocol-level signature, then the classification system can identify the flow via the application-level signatures included in the application protocol-level signature. This hierarchical analysis can reduce the signature search space of the classification system, and reduce the processing time.

Frequency-based signature abridgment

The popularity of various applications could be highly uneven, due to the existence of well-known services, e.g. popular websites, e-mail, etc. These phenomena motivate us to determine the signature matching order in the classification system.

Figure 4 is a CDF graph that represents the signature hit rate. Signature hits occurred for only 90 of the 542 non-HTTP signatures for one minute of traffic trace at a certain time, and 80% of the traffic flows of trace were matched by only 50 or less signatures.

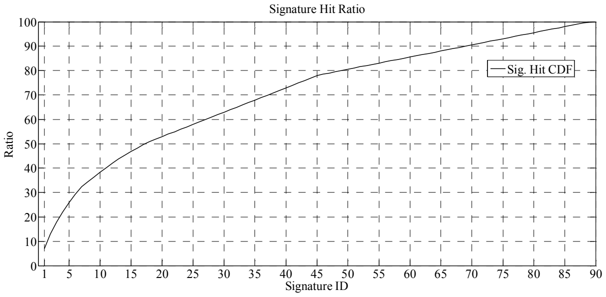

Figure 4. The classified rate of traffic flows in CDF.

Most traffic can be classified using a few signatures during that specific period. We can minimize the search space by first examining frequently occurring signatures and dynamically changing the signature memory ordering according to the signature hit ratio.

Sequence & offset-based payload filtering

Flow-based analysis, rather than packet-based, is popularly used in traffic classification. In addition, we need to minimize the number of packets in a flow and limit the byte size in the packet that will be searched.

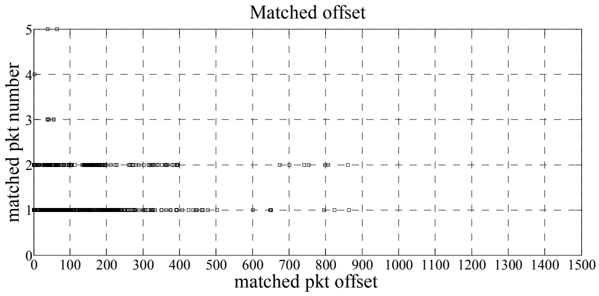

Figure 5. Distribution of the matched offset of the signature

Figure 5 shows the distribution of the matched offset of non-http signatures, in terms of packet and byte position in flows. Most signatures were found in the first 5 packets of the flow, and within the first 1,000 bytes in the packets. We can utilize this experimental result to reduce the search space of the input data for the pattern matching.

According to the analysis results, the classification accuracy and completeness increases as the number of packets inspected increases, but they are almost identical after the fifth packet within the first 1,000bytes in the packets. That is, most connections transmit a low number of control packets that are common amongst all of the same type of connection, before sending the content packets. Therefore, the classification result can be sufficiently accurate and the classification time can be reduced, by limiting the number of packets and bytes.

III. 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

Type	Duration	Flow	Packet	Byte
HTTP	1 Day	41,365K	2,712M	2,263G
Non-HTTP		10,672K	1,144M	1,088G

Table 2 provides the details of the traffic trace making up the full payload that was used to analyze the performance of the proposed method in the experiment.

Table 3. Summary of proposed method

	Baseline method	Proposed method	
		HTTP	Non-HTTP
# of sig.	1,588	1,046	542
sig. model	RE	field based simple string	simple string, RE
traffic filtering	1st~10th packet in flow, max bytes in a packet	field-based	1st~5th packet in a flow, 1000 bytes in a packet
sig. abridgment	linear	hash-based	layer & frequency-based
matching algorithm	NFA-partial	RK	selective matching

Table 3 presents the classification methods we applied to optimize the speed of our classification system, based on the experimental results presented in Section 3.

Figure 6 shows the performance improvement of each, and the combination of the proposed methods.

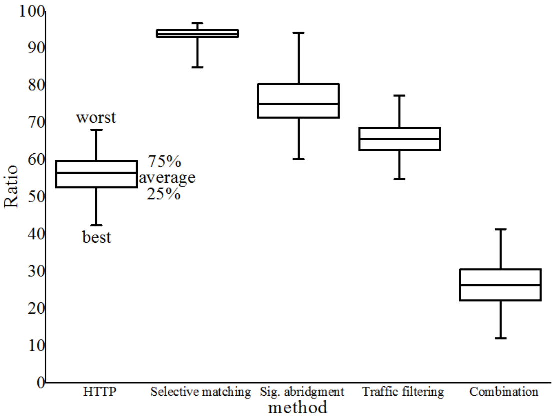

Figure 6. Comparison of classification times.

It shows the rate of how much the processing speed is improved in the worst, average and best case using box plot graphs against the baseline classification system. The proposed HTTP traffic classification method achieves an approximately 2-fold increase in processing speed over the baseline classification system for HTTP traffic. The selective matching, signature abridgment and traffic filtering are applied to non-HTTP traffic. The selective matching is faster than the NFA-partial, which gives the highest average performance for all types of signatures. The traffic filtering can significantly reduce the search space, by limiting the number of packets inspected in a flow, and limiting the byte size in a packet. The traffic filtering shows the best performance improvement for Non-HTTP traffic. The combination of all improves 5 times in processing speed against the baseline classification system.

IV. Conclusion and Future Work

In this paper, we proposed a multi-dimensional criteria for traffic-classification and 3 types of signature model based on protocol characteristics. The each signature model can be applied to the different classification routines in terms of input data minimization and pattern matching algorithm to improve the processing speed. Also we proposed several minimization methods in the search spaces of signatures and input traffic data. Finally, we proposed a selective pattern matching algorithm for improvement of the processing speed. It is possible to design a high-speed Internet traffic classification system. The proposed method achieves an approximately 5-fold increase in processing speed over the baseline system.

This proposed method provides a software-based approach to improve the processing speed of classification systems in a general propose computing environment. We plan to apply our approach to specific hardware system that will allow real-time analysis on a large-scale network.

References

[1] J. S. Park, S. H. Yoon, M. S. Kim, “Software Architecture for a Lightweight Payload Signature-based Traffic Classification System”, Proc. Traffic Monitoring and Analysis Workshop, Vienna, Austria, pp. 136-149, Apr. 2011.

[2] A. Dainotti, A. Pescape, K. Claffy, “Issues and future directions in traffic classification”, IEEE Network: The Magazine of Global Internetworking, Vol. 26, No. 1, pp. 35-40, Jan. 2012.

[3] R. Antonello, S. Fernandes, D. Sadok, J. Kelner, “Characterizing Signature Sets for Testing DPI Systems”, Proc. IEEE GLOBECOM Management of Emerging Networks and Services Workshop, Houston, TX, USA, pp. 678-683, Dec. 2011.

[4] Aceto, G.., Dainotti, A., de Donato, W., Pescape, A., “PortLoad: taking the best of two worlds in traffic classification”, Proc. IEEE INFOCOM Conference on Computer Communications Workshops, San Diego, CA, USA, pp. 1-5, Mar. 2010.

[5] Huang, N. F., Jai, G. Y., Chao, H. C., Tzang, Y. J., Chang, H. Y., “Application traffic classification at the early stage by characterizing application rounds”, Information Sciences, Vol. 232, pp. 130-142, May 2013.

[6] T. Ban, S. Guo, M. Eto, D. Inoue, K. Nakao, “Towards Cost-Effective P2P Traffic Classification in Cloud Environment”, IEICE Transactions on Information and Systems, Vol. E95-D, No. 12, pp. 2888-2897, Dec. 2012.

[7] Khalife, J. M., Hajjar, A., Díaz-Verdejo, J., “Performance of OpenDPI in Identifying Sampled Network Traffic”, Journal of Networks, Vol. 8, No. 1, pp. 71-81, Jan. 2013.

[8] Ji-Hyeok Choi, Myung-Sup Kim, "Processing Speed Improvement of Traffic Classification based on Payload Signature Hierarchy," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2013, Hiroshima, Japan, Sep. 25-27, 2013.