THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

An Overview of Traffic Classification Using Payload-Based Technique

Max Bhatia (corresponding-author),
Department of Computer Science Engineering,
Lovely Professional University, Punjab (India)
max.16870@lpu.co.in

Vikrant Sharma
Department of Computer Science Engineering,
Lovely Professional University, Punjab (India)
vikrant.23738@lpu.co.in

Abstract: For a long period of time, classifying internet traffic is of utmost important task. With the rapid growth of internetas well as with the emergence of various applications & services for communication, the traffic classification approaches have also been evolving.Many techniques have been proposed over the years for classifying the traffic, one of which is payload-based (also called Deep Packet Inspection) technique which consists of analysing the contents of captured packets of traffic in order to find the application associated with it and produces most accurate classification results. There have been many research works which make use of Deep Packet Inspection (DPI) technique for classifying traffic or finding the ground-truth of the traffic and many open-source tools are available for performing such task. This paper presents an overview of DPI technique for classifying network traffic and various modules available for it.

Keywords: internet traffic classification, Deep Packet Inspection

1. Introduction

Classification of internet traffic is a critical task from the view point of network management and security. Accurate traffic classification enables network administrators or Internet Service Providers (ISPs) to manage and deploy the network resources efficiently. With its help, various task related to network management such as implementing network security policies,maintaining Quality-of-Service (QoS) for applications, implementing pricing, etc. can be accomplished in a better way. Various traffic classification approaches exists such as port-based, payload-based and statistical-based; which have been used (either independently or in combination) by the researchers over the years to address various issues such as timely traffic identification, data encryption,data privacy, etc. and proposed their novel techniques for traffic classification. This paper presents an overview of payload-basedtraffic classification technique and various open-source modules available for its deployment.

The remaining paper is organised as follows: Section 2 presents traffic classification techniques. Section 3 presents various open-source traffic classification available for deployment of DPI technique. Section 4 presents the analysis of these open-source modules. At last, conclusion has been presented in Section 5.

2. Traffic classification techniques

This section presents various traffic classification techniques and some of the relevant work published.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

A) Port-based classification

This is most common and the oldest method of traffic classification which classifies the traffic by analysing the port numbers contained in the UDP/TCP packet header. The well-known ports ranging from 0-1023 are assigned by IANA[1] to various protocols such as HTTP, SMTP, DNS, FTP, etc. using which the traffic flows can be easily classified. It is the fastest approach since the port numbers can be easily accessed from the packet headers and are not usually affected by encryption. This technique can be easily implemented and hence is most often used to build certain rules in firewalls and access control list[2]. However, port-based technique cannot be used to classify all kinds of traffic flows since there exists some protocols such as Peer-to-Peer (P2P) and FTP (operating in passive mode) which use random or ephemeral port numbers for communication. In addition to this, many P2P applications masquerade the traffic to avoid their detection. Finally, this technique is also unable to work on Network Address Port Translation. The authors in [3] and [4] found that only 30%-70% of the traffic could be detected by using port-based approach.

B) Statistical-based classification

Statistical-based classification technique makes use of parameters such as inter-arrival time of packet, packet length, duration of flow, etc. [5] in order to classify the traffic. Unlike payload-based (DPI) technique, it does not depend upon payload of the packets for classification, since access of packet payload may not be possible in various scenarios due to legal restrictions to preserve user privacy, encrypted packets, etc.There are many research works which make use of machine learning algorithms along with statistical-based technique. Machine learning algorithms help to learn and build a model automatically from a dataset which is called learning phase. Such data model selects the suitable threshold values and attributes by making a decision table or decision tree in order to classify the traffic. Machine learning algorithms use one of the two learning strategies: supervised approach or un-supervised approach. In supervised learning approach, pre-labelled training dataset is provided to train the model which is then used to classify the unknown traffic. In unsupervised learning approach, no prior training dataset is provided and the records are grouped according to their feature similarity [7].

The authors in [6] proposed traffic classification method by using the parameters such as average, minimum, maximum &standard deviation with respect to the packets'inter-arrival time in a flow, transport layer protocol used, total number of bytes & packets in a flow. The authors follow the idea of Vector Space Model where each packet and flow forms a vector from these values. They were able to achieve true positive rate of 90% and false positive rate of 7% and also point out the concern that the detection rate of this technique can varysince variance of inter-arrival time can temporarily fluctuate the usage rate as it depends on the kind of network.The authors in [8] used a simple method to enhance the Quality of Service for the Voice-over-IP (VoIP) traffic where they marked the packet as VoIP if the packet-length ranges between 60-150 bytes. They argue that Quality of Service of the encrypted VoIP traffic can be enhanced with this technique since there is only a small amount of non-VoIP traffic whose packet-length appears within this range. Hence, this technique can be used to enhance Quality of Service of VoIP traffic but cannot be used for traffic classification purpose.

The authors in [9] mentioned an approach using machine learning algorithms for identifying the Skype characteristics from version 2 to 4. They used an approach used by Nguyen et al. [10] to divide the flows into sub-flows for improving the classification results and deploying

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

in real-time environments. The parameters utilized by authors are: packet length, packets' inter-arrival time and ratio of packets of flow in forward & reverse direction. Their proposed technique was able to achieve 86% recall and 97% accuracy in identifying Skype traffic with the help of J48 algorithm.

C) Payload-based classification

i) Pattern matching:

Deep Packet Inspection (DPI) approach is deployed in Network Intrusion Detection System and classifying network traffic. In addition to that, it is also used for analysing the ground truth of the traffic. By using DPI, contents of captured packets, which are generally in string format, can be compared with the set of rules. But, using such a format poses a strong limitation such as not able to handle complex services and limited expressiveness. Hence, regular expression (RegEx) is preferably used to simplify string comparison due to their scalability, flexibility and powerful approach in inspecting the contents of captured traffic.

In [11], authors review the pattern matching work for DPI where they mention that although string matching approach is intensively used for filtering internet content and detecting intrusions, it still needs to address various implementation issues such as supporting large volume of signatures and ability handle and process multi-gigabit connections efficiently. Further, the authors divided this approach into various parts: heuristic-based approach, filtering-based approachand automation-based approach. They proposed various enhancements that can be applied to these categories. The authors also discussed various ways to deploy DPI technique and how that can efficiently assist in implementing string-matching approaches which are scalable.

In [12], a light-weight DPI scheme is proposed by the authors which removes the limitation of packet loss that acts as performance bottleneck while operating with DPI-based systems. The authors collected signatures of more than 60 applications using L7-filter which is a well know tool [13] for identifying the traffic. The proposed technique was evaluated by inspecting fraction of the contents of given packet-payload of a flow and also by inspecting full-payload of a limited number of packets of a flow. They also performed in-depth analysis regarding the requirements for deploying DPI-based approach. They performed the experiments by capturing the traffic from various sources such as LAN, country-wide academic network and ISPs and the results show that by inspecting fraction of the contents of given packet-payload or by reducing the number of packets being inspected for a given flow, the processing time of DPI will be reduced without compromising the classification accuracy.

ii) Protocol decoding

This approach statefully reconstruct application and session information from the contents of captured packets. It utilize characteristics of protocol headers and sequence of packets to identify the protocol. Due to the inspection of protocol header, this method is usually called DPI. In addition to that, it also analyses the behaviour of protocol which is modelled as protocol state machine. This approach usually has low false negative rates and provides high accuracy [14], but it has a limitation where it is difficult to build good protocol decoder as it requires in-depth knowledge about the protocol as well as the network environment. The main issue with this approach is that there are large amount of network protocols in existence and the decoders have to be kept updated [14] with them. In addition to that, it is unable to work on encrypted traffic. Hence, this approach is mostly used for dedicated protocols such as HTTP, email, etc. as used in Cisco'sCNBAR which is network-based product [15].

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Besides,other vendors also develop and sell theirDPI-based products such as Juniper, Ipoque, Qosmos, Amesys, Utimaco, Sandvine, etc.

In [16], the authors proposed an approach to automatically find communication state machine and characteristic payload of a given protocol. They observed the protocol from raw network traces and gathered all the necessary information about it. In the first step, they analysed the payload to find the typical and static parts of the protocol. In the second step, protocol state machine was constructed by using State Machine Builder which make use of the packet direction and some useful fields of the packet to build the machine's initial state. The number of states are then reduced to a small number which contains all transitions; after which redundant states and unnecessary edges are eliminated from the state machine. The generated state machine of state machine builder produces a graph as an output.

3. Open-source DPI classification modules

There are various DPI-based open-source traffic classification modules. OpenDPI[17] module is implemented as the combination of protocol decoding and behavioural analysis approaches, is considered one of the best DPI-based tool by the community as it achieves high classification accuracy.

A superset of OpenDPI module is nDPI[18] which is maintained by ntop. It was released under GPL license and added the new protocols in OpenDPI, thus extending the original library. It is cross-platform DPI tool which is also suitable to monitor real-time traffic. It is used by nProbe and ntop for detecting the applications without any regard to the ports used by them and hence identifies known protocols which use non-standard ports also.

A module based on Linux kernel named HiPPIEsupports both the functionalities of a firewall and DPI [19]. Itcan identify traffic generated by 31 protocols and supports alsoIPTables and netfilter kernel modules.

Libprotoident[20][21] is a software library which performs the identification of application layer protocols from the captured flows. It uses TCP & UDP port numbers and only first 4 payload-bytes transferred in both directions. It is possible that the same initial 4-bytes could be present in other protocols hence some constraints can be added to such rules in order to discriminate between various such protocols. Libprotoident is less memory-intensive and faster when compared with existing DPI modules and also gives comparable classification results. The developers can use the simple API to write the code which can use the libprotoidentrules for identifying more than 200 application protocols.

L7-filter [13] is Linux netfilter classifier which identifies the captured packets by using application layer data. It is implemented as userspace& kernel modules and identifies network protocols by using RegEx and can be used together with Linux QoS system in order to perform traffic shaping which is independent of the port and application specific management of bandwidth.

IPP2P [22]tool identifes P2P traffic. It extends the netfilter/iptablesand is easily integrated in existing Linux firewalls. This tool helps the network administrators for dynamically filtering the traffic in intelligent way. It searches for P2P signalling patterns in all TCP and UDP packet-payloads and works together with connection marking and connection tracking.

4. Analysis of payload-based classification modules

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

The in-depth analysis and evaluation of traffic classification modules that are discussed in previous section has been performed by the authors in [23]. In that, they created their own testbed environment and utilized a common set of traffic traces in order to evaluate the classification modules based on: classification accuracy, early detection, memory usage and CPU utilization. Hence, we refer the user to [23] to study the details of evaluation.

The most common problem with traffic classification using DPI technique (i.e. using protocol decoding and regular expression) is encrypted traffic. Various types of encryption used in protocol or application are:

1) traffic encrypted after performing handshake
2) packet payload encrypted
3) complete payload and parts of the packet header encrypted
4) complete flow encrypted

The example of first kind of encryption are email protocols. The protocols which comes under second kind of encryption are BitTorrent or SSL/TLS. The representative of third kind of encryption is XMPP (i.e. Extensible Messaging and Presence Protocol) which is XML-based protocol (where PDU is an XML document) and is also known as Jabber. The last kind of encryption is used in all types of encrypted tunnels such as IPSec, TLS, etc. Therefore, protocol decoding or pattern matching approach can be used for traffic classification if any of the first 3 kinds of encryption is involved. However, the fourth kind of encryption (where whole flow is encrypted) makes these approaches ineffective.

The authors in [23] found that by the 8^{th} packet, most of the network flows can be detected with the use of protocol decoding or pattern matching approach; after which the detection is mostly not possible. Hence, these approaches work only if complete handshake of the protocol is captured since afterwards, the packets that are exchanged generally do not contain protocol header, as in the case of: FTP, HTTP, BitTorrent, eDonkey, SOCKS, etc. This suggests that these approaches cannot detect most of the flows which have already started.

5. Conclusion

Internet traffic classification has been of utmost importance since the evolution of internet itself. Throughout the years, various traffic classification techniques have been proposed to address various techniques issues. The main purpose of the paper is to give an overview of various payload traffic classification approaches and open-source DPI modules available for performing such task. The open-source classification modules mentioned in this paper has already been evaluated on the basis of classification accuracy, early detection, memory usage and CPU utilization by the authors in [23] using a common set of captured traffic traces, which gives an idea about the significance and limitations of such modules. Because of the various limitations associated with this technique for classifying the traffic, so it is beneficial to use such technique for the purpose of ground truth analysis of the traffic.Therefore, for performing classification of the traffic in real-time, statistical-based technique is most suitable.

References

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

[1]IANA, Port Numbers. [Online]. Available:http://www.iana.org/assignments/port-numbers

[2]Qi, Yaxuan, LianghongXu, Baohua Yang, YiboXue, and Jun Li. "Packet classification algorithms: From theory to practice." In *IEEE INFOCOM 2009*, pp. 648-656. IEEE, 2009.

[3] Moore, Andrew W., and KonstantinaPapagiannaki. "Toward the accurate identification of network applications." In *International Workshop on Passive and Active Network Measurement*, pp. 41-54. Springer, Berlin, Heidelberg, 2005.

[4]Madhukar, Alok, and Carey Williamson. "A longitudinal study of P2P traffic classification." In *14th IEEE International Symposium on Modeling, Analysis, and Simulation*, pp. 179-188. IEEE, 2006.

[5] Nguyen, Thuy TT, and Grenville J. Armitage. "A survey of techniques for internet traffic classification using machine learning." *IEEE Communications Surveys and Tutorials* 10, no. 1-4 (2008): 56-76.

[6]Piskac, Pavel, and Jiri Novotny. "Using of time characteristics in data flow for traffic classification." In *IFIP International Conference on Autonomous Infrastructure, Management and Security*, pp. 173-176. Springer, Berlin, Heidelberg, 2011.

[7]Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[8]Yildirim, Taner, and P. J. Radcliffe. "VoIP traffic classification in IPSec tunnels." In *2010 International Conference on Electronics and Information Engineering*, vol. 1, pp. V1-151. IEEE, 2010.

[9]Jesudasan, RozannaNadeera, Philip Branch, and Jason But. "Generic attributes for skype identification using machine learning." *Centre for Advanced Internet Architectures, Swinburne University of Technology, Melbourne, Australia, Tech. Rep. A* 100820 (2010): 20.

[10]Nguyen, Thuy TT, and Grenville J. Armitage. "Training on multiple sub-flows to optimise the use of Machine Learning classifiers in real-world IP networks." In *LCN*, pp. 369-376. 2006.

[11]Lin, Po-Ching, Ying-Dar Lin, Yuan-Cheng Lai, and Tsern-Huei Lee. "Using string matching for deep packet inspection." *Computer* 41, no. 4 (2008): 23-28.

[12]Fernandes, Stênio, Rafael Antonello, ThiagoLacerda, Alysson Santos, DjamelSadok, and TordWestholm. "Slimming down deep packet inspection systems." In *IEEE INFOCOM Workshops 2009*, pp. 1-6. IEEE, 2009.

[13]Application Layer Packet Classifier for Linux. [Online].Available: http://l7-filter.sourceforge.net/

[14]Risso, Fulvio, Mario Baldi, Olivier Morandi, Andrea Baldini, and Pere Monclus. "Lightweight, payload-based traffic classification: An experimental evaluation." In *2008 IEEE International Conference on Communications*, pp. 5869-5875. IEEE, 2008.

[15] Network Based Application Recognition (NBAR). Cisco. [Online]. Available: http://www.cisco.com/en/US/products/ps6616/products_ios_protocol_group_home.html

[16]Trifilo, Antonio, Stefan Burschka, and Ernst Biersack. "Traffic to protocol reverse engineering." In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1-8. IEEE, 2009.

[17] OpenDPI.org. [Online]. Available: http://www.opendpi.org/

[18]nDPI. [Online]. Available: http://www.ntop.org/products/ndpi/

[19]HiPPIE - Linux Software. [Online]. Available: http://www.linux112.com/hippie-p313520.html

[20] WAND Network Research Group. [Online]. Available: http://research.wand.net.nz/software/libprotoident.php

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

[21]Alcock, Shane, and Richard Nelson. "Libprotoident: traffic classification using lightweight packet inspection." *WAND Network Research Group, Tech. Rep.* (2012).

[22]Official IPP2P homepage. [Online]. Available:http://www.ipp2p.org/

[23] Finsterbusch, Michael, Chris Richter, Eduardo Rocha, Jean-Alexander Muller, and Klaus Hanssgen. "A survey of payload-based traffic classification approaches." *IEEE Communications Surveys & Tutorials* 16, no. 2 (2013): 1135-1156.