International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

A Survey on Analysis of service usage of application using Multistage classifier for network traffic classification.

Madhura R Manthale, Varalakshmi B D

PG student, Dept. of CSE, Acharya Institute of technology, Karnataka, India
Associate Professor, Dept. of CSE, Acharya Institute of technology, Karnataka, India

---

Abstract - *Identifying and then classifying network traffic plays a vital role to provide quality of service (QoS) for network management. Hence a multistage classifier is being used for network traffic classification with the help of which a service usage is analyzed. The analysis of application usage aims at identifying customers behaviors, this has become a challenging task for service providers. Traditional approaches were mainly concentrated only on packet inspection for internet traffic classification, which has imposed some performance challenges. At this end, a CUMMA for classifying service usage of mobile messaging apps being developed which jointly modelled with behavior patterns, network traffic and temporal dependencies for messaging Apps service usage classification in mobiles.*

Key Words*: *Network management, Packet inspection, Multistage classifier, CUMMA

1. INTRODUCTION

With the growth in the internet, service providers attempt to provide privacy, reliability, multiple service qualities, security and thereby developing a best-effort application architecture [1]. Then according to application a proper classification of traffic [5] over network should be done such as to prioritize, prevent or protect some traffic.

Initially techniques for network classification were based on the packet port numbers called Port-based approach (PBA) [2] where traffic classification depends on port number usage in transport layer which is found in headers of UDP (User Datagram) or TCP (Transmission Control) protocols [3]. To define a well-known application these are registered with IANA. Classification with port numbers are simpler and faster, but researchers proved with poor performance. To overcome the issues of classification using PBA a Deep packet inspection(DPI) technique[4] was developed where many of network devices identify the traffic type which the packet represents using session and application layer information. Here packets with same protocol and having source and destination address same belongs to same flow. This may lead for privacy data leak in some way. Though these techniques were employed for intrusion detection and for P2P application identification but still Encrypted payloads were not subjected for such techniques. To this end a Statistical signature based approach (SBA) [5] is developed for traffic classification using statistical features [8] like Interarrival time and Packet length. The main aim is to classify traffic or to provide traffic behavior.

Multistage classifier [9] as shown in the fig1 below which depicts that it incorporates all the three i.e., Port based, Deep packet inspection and Statistical signature based approach techniques to classify traffic over the network. Here the algorithm is developed that uses two databases named: (a) port database containing port numbers along with its corresponding application class and (b) signature database consisting of packets with its signature associated also along with its corresponding application. Initially traffic is classified into sessions, this algorithm checks whether new packet arrived belongs to existing session or any new session to be created. Then it checks with port database for whether port number of that particular packet exist in it. If present then packet is classified with its corresponding application of that port. If both methods go wrong in traffic classification, then statistical approach is being used. Later packets in a session are classified to achieve service usage. Using these classification techniques of traffic service providers can identify which endpoints at that particular time are sending packets and also can identify application or in particular a messaging apps that a particular person of interest is using at any time instance.

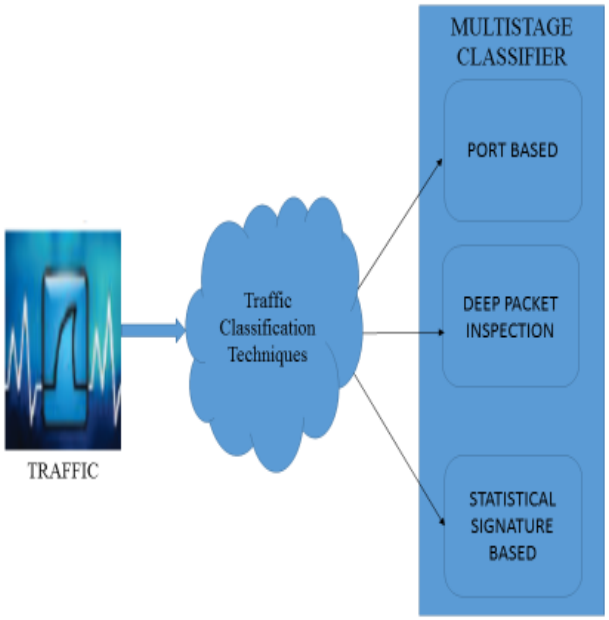

Fig 1. Traffic classification Techniques

At this end, a CUMMA [10] being developed which jointly modelled with behavior patterns, network traffic and temporal dependencies for messaging Apps service usage classification in mobiles. The classification starts from traffic flows into sessions with number of dialogs hierarchically. Once classified the traffic records are stored in a database and hence accessed accordingly. Two perspectives [10] to extract traffic data features are: (i) packet length and (ii) time delay. Then, to classify segmented dialogs, a predictor of service usage being learnt to segment dialogs into single type usage.

2. RELATED WORK

REVIEW OF TECHNIQUES

Prior literature survey on techniques of network classification was based on the packet port numbers called PBA where traffic classification depends on port number usage in transport layer which is found in UDP or TCP header [3]. To define a well-known application these are registered with IANA. Classification with port numbers are simpler and faster, but researchers proved with poor performance. Moore and papagiannaki [4] founded that only 70% byte accuracy is acquired using IANA list.

To overcome the issues of classification using PBA a DPI technique [4] was developed where many of network devices identify the traffic type which packet represents using session and application layer information. As shown in table1 the paper An Overview of Network Traffic Classification Methods describes packets with same protocol and having source and destination address same belongs to same flow. This may lead for privacy data leak in some way. Though these techniques were employed for intrusion detection and for P2P application identification but still Encrypted payloads were not subjected for such techniques. A Survey on Network Traffic Classification Techniques paper from table1 generalizes techniques of Classification based on the packet port numbers being used in earlier days of Internet. To avoid the problems with port based approach the payload based technique is used which is used in inspection of the payload content. However, privacy of the packet is not secured. Due to such security issues, the statistical features of packet were being examined to identify the application that generated them.

Statistical protocol classification [7] algorithm is based on fingerprint classifies traffic using three features of packet, they are packet size, interarrival time and order of arrival. Threshold of an anamoly score of algorithm is used for classification and by using the training set which are pre-labelled flow from the application are analyzed then a protocol fingerprint is constructed. The PDF (probabilistic density function) vector is used for protocol fingerprint. For all ith packet a pair of {si, ∇ti} of PDFi is built, where si denotes the packet size and ∇ti denotes interarrival time between ith and i-1th packet. For unknown flow Zhan et.al, proposed An Effective Network Traffic Classification Method[8], where the algorithm checks whether there is atleast one PDF whose description is compatible with behavior. The technique has disadvantage that classifier is unaware of packet loss.

MULTISTAGE CLASSIFIER

Multistage classifier [9] uses all these three techniques to classify traffic over the network. Here the algorithm is developed that uses two databases named: (a) port database containing port numbers along with its corresponding application class and (b) signature database consisting of packets with its signature associated also along with its corresponding application. Initially traffic is classified into sessions, this algorithm checks whether new packet arrived belongs to existing session or any new session to be created. Then it checks with port database for whether port number of that particular packet exist in it. If present then packet is classified with its corresponding application of that port. If both methods go wrong in traffic classification, then statistical approach is being used. Later packets in a session are classified to achieve service usage. Using these classification techniques of traffic service providers can identify which endpoints at that particular time are sending packets and also can identify application or in particular a messaging apps that a particular person of interest is using at any time instance.

Sl.No	TITLE	Publisher and Author	Remarks
1.	Machine learning based encrypted traffic classification.	Alshammari, Heywood IEEE symposium-2009	The main objective of this work is to assess the machine learning robustness based traffic classification. This work provides information about port based network classification method.
2.	An Overview of Network Traffic Classification Methods	Zeba Atique Shaikh and Prof. Dr. D.G. Harkut, Ram Meghe College of Engineering	Traffic classification over the network is used for identification of different protocols and applications that is present in the network. A comparison of multiple network traffic classifiers, that depend on Deep Packet Inspection is shown
3.	A Survey on Network Traffic Classification Techniques.	Aafa J S and Soja Salim Computer Science and Engineering Sree Chitra Thirunal College of Engineering, Trivandrum 2014	This paper generalizes techniques of Classification based on the packet port numbers being used in earlier days of Internet. To avoid the problems with port based approach the payload based technique is used which is used in inspection of the payload content. However, privacy of the packet is not secured. Due to such security issues, the statistical features of packet were being examined to identify the application that generated them.
4.	An Effective Network Traffic Classification Method with Unknown Flow Detection	Zhan et.al, IEEE Transactions-2013	Here a new method is proposed to tackle with unknown application problems in situation that are crucial of the small supervised set of training data. This method also possesses the detection of unknown flows that are generated by applications, the correlation information is utilized among network traffic in the real-world for boosting the classification information.

Table 1: Traffic classification.

Once the packets are classified, the router can apply appropriate service policies for those packets by considering the features for predicting usage.

3. USAGE PREDICTION

TRAFFIC FEATURE EXTRACTION AND USAGE TYPE PREDICTION

Given a set of dialogs, the objective is to identify their usage types. To this end, first mine the discriminative features of the network traffic data from two perspectives:

(i) packet length and (ii) time delay [7]. Once the packets are classified, the router can apply appropriate service policies for those packets.

(i) Packet length:
Normally the data flow pattern is exhibited by packet length sequence, and hence reflects different usage behaviours.
The following features describes about the packet length perspective based on their usage behaviours.

a. Descriptive statistics:

As the packet length basic properties are described from multiple aspects, it is hence descriptive statistics is needed. So, from the Packet length sequence given the first order statistics and second order statistics are extracted (i.e., median, skewness, standard deviation etc., of packet length) as features.

b. Forward and Backward Direction Variances:

The packet size variance is a kind of application behaviours signature. Though with low variation of sequences, this feature helps in capture of the fine grained variances. For the packet length sequence given, select observation positions which are representative from sequence. Later sequence is splitted into subsequences for each position being selected according to backward and forward direction.

c. Packet percentage:

Different application behavior shows different packet length ranges i.e., for example video stream packet length is larger than text message packet length. Packet length feature hence helps in equal distribution which reduces the noise from small fluctuations. For a packet length sequence given first identify the range of IP packet, and split accordingly into min and max subranges. Finally calculate the packet percentage of each and every subranges.

d. Hop counts:

Unlike in text or video stream where Hop count describe the packet pulse, count packet numbers which have greater length than next packet. This number hence is used to characterize the sequence fluctuation.

(ii) Time Delay:
Consider two consecutive packets, extract time interval for each and hence obtain the time delay sequence.
The time delay sequence characterizes patterns into implementation and protocol aspects, where implementation is of application usages and protocols are for data transmission. Similarly, packet length definition are adopted and features are mined from time delay sequence.
Once both features extracted then feed segmented dialogs to a predictor of service usage to segment them into single type usage. Packet inspection for internet traffic classification, such as HTTP header parsing which has imposed some significant performance challenges.
At this end, a CUMMA being developed which jointly modelled with behavior patterns, network traffic and temporal dependencies for messaging Apps service usage classification in mobiles. The classification starts from traffic flows into sessions with number of dialogs hierarchically. As two perspectives to extract traffic data features are: (i) packet length and (ii) time delay. Then, to classify segmented dialogs a predictor of service usage being learnt to segment dialogs into single type usage. Here given with segmented traffic subsequence, main aim is to output the probabilities [10] of each usage type. So that ranking all probabilities can be done such that based on the highest probability in traffic classification subsequence for the service usage can be done. CUMMA also designed for offline application usage analytics without traffic processing, thus providing highest efficiency. CUMMA helps even to identify end user behavior for traffic classification.

4. CONCLUSIONS

The paper surveys about different techniques being used for network traffic classification to provide better quality of service. On comparison with port based and deep packet inspection, the statistical protocol method proved to be better by its performance. Taking into consideration of all these three methods a new classifier technique called multistage classifier being developed which proved to be more efficient by its better performance. There is still a lot space in a research of such methods, while most approaches were implemented on variety of applications.
Initially a data collection platform is built to collect the internet traffic for application usages, once collected traffic is then segmented hierarchically using CUMMA technique with the help of the packet length and Time delay features. Upon classifying into subsequences of the packets, they are then used for application usages. Finally main aim is to output the probabilities of each usage type. Hence by ranking all them based on highest probabilities helps in classification of traffic subsequence for service usage.

REFERENCES

[1] Yanjie Fu, Junming Liu, Xiaolin Li, "Service Usage Analysis in Mobile Messaging Apps: A Multi-Label Multi-View Perspective". IEEE 16th International Conference on Data Mining, 2016

[2] Riyad Alshammari and A Nur Zincir-Heywood, "Machine learning based encrypted traffic classification: identifying ssh and skype". In Computational Intelligence for Security and Defence Applications, 2009. CISDA 2009. IEEE Symposium on, 2009.

[3] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: Multilevel traffic classification in the dark," in Proc. of the Special Interest Group on Data Communication conference (SIGCOMM) 2005, Philadelphia, PA, USA, August 2005.

[4] Zeba Atique Shaikh and Prof. Dr. D.G. Harkut, Ram Meghe College of Engineering, "An overview

Network Traffic classification Methods", International Journal on Recenet and Innovation Trends in computing and communication, February 2015.

[5] Aafa J S and Soja Salim Computer Science and Engineering Sree Chitra Thirunala College of Engineering, Trivandrum "A Survey on Network Traffic Classification Techniques." International Journal of Engineering Research & Technology, 2014

[6] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification," in Proc. ACM/SIGCOMM Internet Measurement, Conference (IMC) 2004, Taormina, Sicily, Italy, October 2004.

[7] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," SIGCOMM Comput. Commun. Rev., vol. 37, no. 1, pp. 5–16, 2007.

[8] Jun Zhan, Chao Chen, Yang Xiang, Wanlei Zhou and Athanasios, "An Effective Network Traffic Classification Method with Unknown Flow Detection", IEEE Transactions on Network and Service Management, Vol. 10, No. 2, June 2013

[9] DU Min, CHEN Xingshu, "Online Internet Traffic Identification Algorithm Based on Multistage Classifier," IEEE Trans. On Inf. Forensics and Security, Vol. 8, No. 1, Jun 2013.

[10] Y. Fu, H. Xiong, X. Lu, J. Yang, and C. Chen, "Service usage classification with encrypted internet traffic in mobile messaging apps," in IEEE Transactions on Mobile Computing, 2016.