International Journal of Engineering

Journal Homepage: www.ije.ir

Behavioral Analysis of Traffic Flow for an Effective Network Traffic Identification

M. Gandomi, H. Hassanpour*

Department of Computer Engineering and IT at Shahrood University of Technology , Iran

P A P E R I N F O

Paper history:
Received 20 July 2017
Received in revised form 27 August 2017
Accepted 08 September 2017

Keywords:
Network Traffic Identification
Behavioral Analysis
Data Mining
Machine Learning
Flow Statistical Featur

A B S T R A C T

Fast and accurate network traffic identification is becoming essential for network management, high quality of service control and early detection of network traffic abnormalities. Techniques based on statistical features of packet flows have recently become popular for network classification due to the limitations of traditional port and payload based methods. In this paper, we propose a method to identify network traffics. In this method, for cleaning and preparing data, we perform effective preprocessing approach. Then effective features are extracted using the behavioral analysis of application. Using the effective preprocessing and feature extraction techniques, this method can effectively and accurately identify network traffics. For this purpose, two network traffic databases namely UNIBS and the collected database on router are analyzed. In order to evaluate the results, the accuracy of network traffic identification using proposed method is analyzed using machine learning techniques. Experimental results show that the proposed method obtains an accuracy of 97% in network traffic identification.

***doi**: 10.5829/ije.2017.30.11b.15*

1. INTRODUCTION

There are huge amount of network traffics from various applications exchanged over the internet. In dealing with this volume of network traffic, network management is very important. Traffic classification is a basic technique which is used by internet service providers to manage network resources and guarantee internet security. In addition, growing bandwidth usage, at one hand, and limited physical capacity of communication lines, at the other hand, lead providers to improve utilization quality of network resources. In fact, classification or identification of network is a critical task in network processing for traffic management, anomaly detection and also to improve network quality-of-service (QoS) [11].

Port and payload based methods are two classical techniques which are effective under traditional network conditions [1-6]. However, many internet applications use dynamic port numbers for communications, which leads to difficulty of identifying traffic by port numbers [7]. Also many applications encrypt the transmitting data to avoid being detected [8]. Therefore, payload-based techniques become inefficient for these traffics. In recent years, traffic flow statistical feature-based identification methods (TFSIM) have attracted the interest of many researchers. The most important part of a TFSIM is the selection of efficient statistical features [2].

In this paper, we propose an effective and applicable network traffic identification method. Thus, preprocessing is performed to remove noise, the incomplete flows and the discovery of sessions before networh traffic identification. Also a robust feature extraction method using behavioral analysis of applications is applied to extract statistical features from packets. For this purpose, two network traffic databases namely UNIBS and our collected database on routers are analyzed. These databases consist of several applications such as Skype, Telegram, Firefox, Chrome, Safari, Internet Explorer, Download Manager, Thunderbird, Bittorrent and Mail.

Following the behavioral analysis of the applications, appropriate flow statistical features are selected. In order to evaluate the results and select the best machine learning techniques, the accuracy of the proposed network traffic identification method is

*Corresponding Author's Email: h.hassanpour@shahroodut.ac.ir (H. Hassanpour)

Please cite this article as: M. Gandomi, H. Hassanpour, Behavioral Analysis of Traffic Flow for An Effective Network Traffic Identification, International Journal of Engineering (IJE), TRANSACTIONS B: Applications Vol. 30, No. 11, (November 2017) 1740-1745

analyzed using four algorithms, i.e. Support Vector Machine (SVM), Bayesian Network, Bosting and Random Forest.

The remainder of the paper is organized as follows, Section 2 reviews the literature about the network traffic identification methods, Section 3 discusses about the proposed method, Section 4 deals with standard UNIBS database and evaluates the results of network traffics identification and Section 5 concludes the paper.

2. RELATED WORKS

The literature in network traffic identification using statistical features can be divided into two groups, those offering one or more features to differentiate between different application traffics, and those employing machine learning techniques to identify the applications. A classification method based on Markov model with convergence criteria of Kullback-Leibler (a similarity measure for probability distributions) has been proposed by Kim et al. [9] to overcome traffic overlapping. This paper argues that the suggested method is effective for those application traffics that have much overlapping, whilst prior methods are not efficient in their classification. The accuracy of the method on 100 samples was reported as 90%. In this method a supervised learning has been used and a Markov model is constructed using the states defined via the first four packets based on the flow direction and packet size.

Muehlstein et al. [10] have shown that an attacker can determine the operating system type, browser type and application type of HTTP data and encrypted data of HTTPS using network traffic. The authors claimed, their work is the first contribution to identifying operation system, browser and application in an encrypted network. They evaluated a database containing 20 thousand sessions Windows operating system traffic, Ubuntu, Iphone operating system (IOS), Chrome Explorer, Firefox and Safari, applications such as You tube, Face book, Twitter, etc. Each session including a quintuple (source and destination port number, source and destination address, related protocol) is mapped to a triad (operating system, explorer and application) and uses the Support Vector Machine (SVM) learning algorithm and the Radial Basis Function (RBF) as the Kernel function to identify operating system type, browser type and application type.

Loo et al. [11] have proposed an algorithm based on incremental K-means clustering that uses both labeled and unlabeled samples for learning. In this method, both Manhattan and Euclidean distance metrics were evaluated to measure similarity degree of samples. The test was conducted on 671 thousand flows, and accuracy of clustering was reported to be 94%. Execution speed using Manhattan distance metric was three times better than Euclidean.

Qin et al. [12] have suggested a model named Bi-flow, which can capture mutual behavioral characteristics among various terminals. Packets (out and in) with the same source and destination are regarded as Bi-flow. Packet size distribution existing in Bi-flow was used as a feature. It was shown that packet size distributions vary in different applications.

The method proposed by Aliakbarian et al. [13] seeks the best number of features for traffic identification using supervised machine learning algorithm. For classification purposes, algorithms such as decision tree, Artificial Neural Network, Bayes learning and Bagging and Boosting were used. The authors concluded that in the case of M classes of applications, features space in best is reduced to M-1.

Zhou et al. [14] have suggested a method to classify network using Feed-Forward neural networks, which uses an extraction algorithm based on most similarities to choose appropriate features.

3. THE PROPOSED STATISTICAL TRAFFIC IDENTIFICATION FRAMEWORK

Consider three users, named A, B and C, are connected to the Internet (Figure 1) and the destination reaches their information through the routers. Traffic analyst is located on the router connection, and began recording the data connection. The proposed network traffic identification method is modeled in Figure 2. This model, same as supervised classification methods has both learning and testing phases. In the learning phase, labeled application packets are recorded and preprocessing operation is performed on the packets. Then statistical features is extracted and the machine learning algorithm is trained using labeled samples. In the testing phase, unknown packets are sniffed and after preprocessing and feature extraction, the trained system predicts their labels.

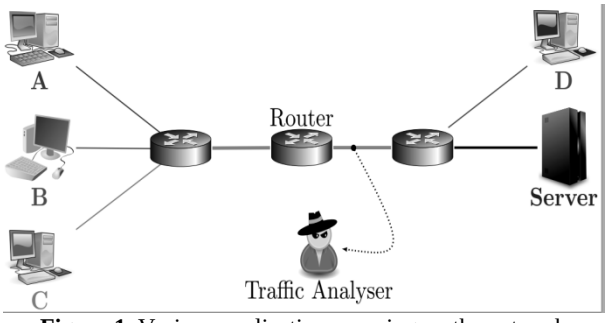

Figure 1. Various applications running on the network

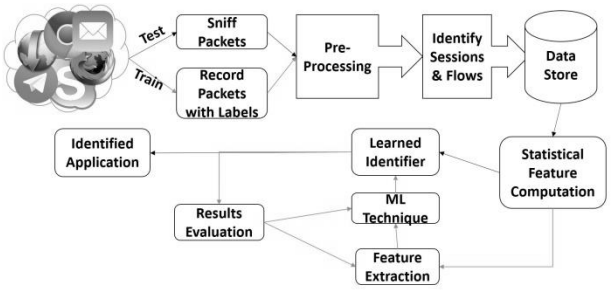

Figure 2. Porposed framework model

In this paper, to identify the applications using four different classification techniques, i.e. SVM, Bayesian Network, Bosting and Random Forest the results are evaluated and the best techniques are selected.

3. 1. Data Preparing In recording packets, waste packets are one of the most important problems. The order of the waste packets is operating systems packets, firewalls packets and network components packets. Each of the mentioned items, to inform and synchronize with each other, send control packets over the network periodically. Also, redundant packets are another example of waste packets. These packets arise in case of incomplete sending or not successfully reaching at destinations. Thus, sometimes a packet is sent repeatedly. As a result, we need to apply some filters on packet capturing.

In this paper, four preprocessing steps are considered as follows: 1) removing all operating systems packets, firewalls packets and network components packets, 2) removing redundant packets (repeated packets), 3) removing unnecessary parts of packets and 4) removing incomplete flows and sessions (this part of preprocessing is performed after sessions identification).

3. 2. Pre-processing Network traffic has different levels. To communication between a source and destination, a session is established. A session may be established in a long time, hence it is divided into flows with short intervals. Flow duration is considered to be 30 minutes [12]. The initial packets of each session have valuable information for identifation methods. For this reason, before network traffic identifocation operation, flows and sessions should be identified. As reported [15], after the flow sepration, first packet size distributions are compared, then second packet size distributions and so forth. Packets that have same Ip address and port number of the sender and recipient are considered as a session.

3. 3. Feature Selection Traffics of network applications are divided into P2P and web application traffic (Client-Server). P2P has a distributed application architecture that partitions tasks or workloads between peers. Peers are equally privileged, equipotent participants in the application. The Client-Server is a distributed application architecture that partitions tasks or workloads between providers of a resource or service, called servers, and service requesters, called clients. Often clients and servers communicate over a computer network on separate hardware, but both client and server may reside in the same system.

In the method of identification based on statistical features, packets content (above transport layer) is not considered. So, application's behavior is examined with regard to network layer characteristics. Table 1 presents a comparison between behavioral characteristics of P2P and Client-Server application traffics in the collected database that will be introduced in Section 4.

As seen, with regard to the behavioral analysis of applications, five statistical features were selected, i.e. packet order, packet size, time interval between send and receive of packets, time duration of flows, and packet rate in network layer. Figures 3 and 4 show the packet size and packet rate distribution.

4. EXPERIMENTAL RESULTS

In previous section, five statistical features were selected with respect to behavioral analysis of the applications. In this section, at first, two databases of UNIBS and our collected database are examined. Having evaluated the features using machine learning algorithms, the features are examined in identification of six applications.

4. 1. UNIBS Database These data were collected from the router inside Bersica University.

TABLE 1. Behavioral features of Client-Server and P2P applications

Characteristics	Traffic (P2P)	Traffic (Client-Server)
Bite transmission rate in flows	High	Low
Self-similarity property parameter	<0.5	0.7-0.8
Data distribution	Heavy trail, on/off	Long tail
Flow duration	<100 milsec	>100 milsec
Data transmission rate bit/sec	Low	High
Time interval between packet send and receive	<0.3 milsec	>0.3 milsec
Packet size	>800 byte	<800 byte
Entropy	High	Low
Packet Transmission rate	$<3*10^5$	$> 2*10^5$

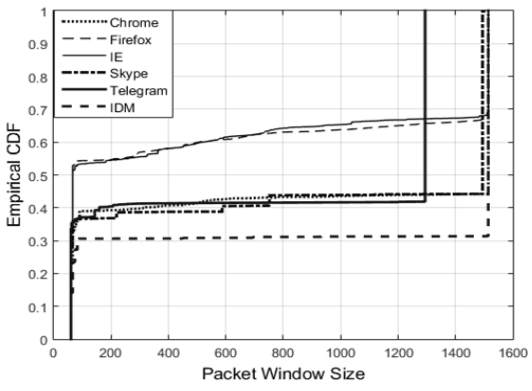

Figure 3. Packet size distribution

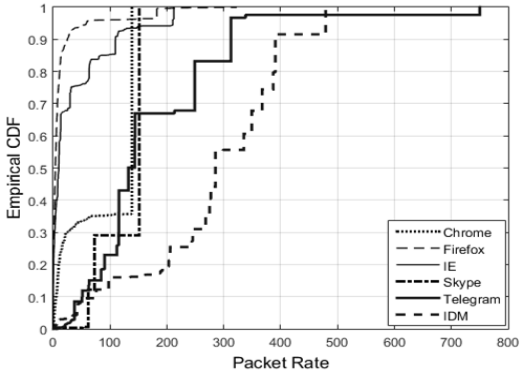

Figure 4. Packet rate distribution

Ethernet of the university includes several 1000Base-TX (one of the most prevalent and fastest forms of old Ethernet so far) and the routing was carried out by two-core Linux processors. Data have been collected from link 100Mb/s in router with the help of Tcpdump software.

The data were collected in a two days period in rush hours. Since there was full access to the router, 250 bites of each frame were saved. Labeling the data was performed using ground truth algorithm and by examining packet content. For this reason, labels for applications were provided with a high reliability. Table 2 shows the number of samples containing six applications from this database.

4. 2. The Collected Database This data was collected from a local area network router.

TABLE 2. UNBS database

Application	Count of samples
Firefox	3516410
Thunderbird	65660
Skype	181550
Safari	3482650
Bittorrent	1244240
Mail	1007790

The data have been collected from link 1gb/s in router using the Wireshark software; traffic of six applications have been captured. The data were collected in a three days period in rush hours. Data labeling was performed using Deep Packet Inspection (DPI) method and by examining packet content. Hence, labels of applications were provided with a high reliability. Table 3 shows the number of samples containing six applications from this database.

4. 3. Identification Results In this section, we use the selected features from the previous section, and identify the related applications using four different classification techniques, i.e. SVM, Bayesian Network, Bosting and Random Forest.

Tables 4 and 5 show the accuracy of the four different machine learning algorithms on two databases in application identification. The number of samples in the applications are different, so each of the applications is evaluated individually. Table 6 compares the results of using the selected features from the existing approaches. As can be seen, results of the selected features using the proposed approach have a higher detection rate compared to other methods.

TABLE 3. Collected database

Application	Count of samples on router
Firefox	636070
Skype	302130
Telegram	681760
IDM	529580
Chrome	302610
IE	305030

TABLE 4. Results from UNIBS database

Application	RandForest	Bosting	SVM	BayesNet
Firefox	89.90	52.96	26.56	96.17
Thunderbird	93.09	61.77	54.41	93.07
Skype	88.97	59.20	59.43	98.78
Safari	85.11	49.83	66.93	96.18
Bittorrent	97.70	68.11	46.63	89.22
Mail	95.97	49.80	39.97	96.98

TABLE 5. Results from collected database

Application	RandForest	Bosting	SVM	BayesNet
Firefox	96.1	34.41	91.77	93.09
Skype	97	36.56	62.96	74.30
Telegram	97.18	39.83	41.39	77.24
IDM	95.32	50.15	14.25	97.47
Chrome	95.88	26.43	29.20	88.47
IE	95.9	47.65	39.02	84.54

TABLE 6. Comparison results between the proposed method and other existing methods

Ref.	Iden. method	Features	Applications	Accuracy (%)	Database
[16]	Bayesian technique	Flow duration, port number, inter-arrival time, packet size, bandwidth used , inter-arrival time FFT	Database, P2P, Bulck, Mail, Services	96.29	Hand classified traces
[6]	Naive Bayes and Chi-Square test	Packet size in transport layer	Skype traffic	97	Two self collected datasets
[15]	Protocol fingerprints	Packet size, inter-arrival time, packets order	HTTP, SMTP, POP3, SSH	90	Proprietary traces
[17]	C4.5 Decision tree	Packet size, byte rate, packet size (MIN, MAX, AVG, STD)	Torrent, DNS, HTTP, SMTP, SSH, SSL	90	UTM campus network
[18]	C4.5 Decision tree	Fast correlation-based filter algorithm	WWW, Mail, Web, P2P, FTP, IM	95.21	Captured at Ohio University
[14]	Feed-froward	Fast correlation-based filter algorithm	HTTP, P2P,IMAP, POP3,SMTP, MYSQL	95	Hand classified traces
[19]	Naïve Bayes, MLP, C4.5	AdaBoost	Torrent, Edonkey, FTP, HTTP, POP3	90%	Chinese educational site
Proposed	Bayes net, SVM, Bosting, Random Forest	&Packet order, packet size, inter-arrival time , flow time, rate	Skype, Firefox, Chrome, IE, Telegram, IDM	97.2	Collected database, UNIBS

5. CONCLUSION

In this paper, we proposed an effective network traffic identification method. In this method, initially preprocessing is performed to remove noise, the incomplete flows and the discovery of sessions before identification. Then feature extraction is performed using behaviral analysis of applications. Two network traffic databases namely UNIBS and our collected database on routers were analyzed. Following the behavioral analysis of the applications, flow statistical features were selected. These features are packet order, packet size, time interval between packet send and receive, time duration of flows, and packet forwarding rate. To evaluate the results, the accuracy of application identification method by four algorithms, namely Bosting, SVM, Bayesian Network and Random Forest was analyzed. The evaluation results show a high identification accuracy on the two database containing different versions of applications.

REFERENCES

1. Foremski, P., "On different ways to classify internet traffic: A short review of selected publications", ***Theoretical and Applied Informatics***, Vol. 25, (2013), 147–164.
2. Zhang, J., Xiang, Y., Wang, Y., Zhou, W., Xiang, Y. and Guan, Y., "Network traffic classification using correlation information", ***IEEE Transactions on Parallel and Distributed Systems***, Vol. 24, No. 1, (2013), 104-117.
3. Wang, Y., Xiang, Y. and Zhang, J., "Network traffic clustering using random forest proximities", in Communications (ICC), IEEE International Conference on, IEEE., (2013), 2058-2062.
4. Adami, D., Callegari, C., Giordano, S., Pagano, M. and Pepe, T., "Skype-hunter: A real-time system for the detection and classification of skype traffic", ***International Journal of Communication Systems***, Vol. 25, No. 3, (2012), 386-403.
5. Finamore, A., Mellia, M., Meo, M. and Rossi, D., "Kiss: Stochastic packet inspection classifier for udp traffic", ***IEEE/ACM Transactions on Networking***, Vol. 18, No. 5, (2010), 1505-1515.
6. Bonfiglio, D., Mellia, M., Meo, M., Rossi, D. and Tofanelli, P., "Revealing skype traffic: When randomness plays with you", in ACM SIGCOMM Computer Communication Review, ACM. Vol. 37, (2007), 37-48.
7. Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., *Supervised machine learning: A review of classification techniques*. 2007.
8. AbuHmed, T., Mohaisen, A. and Nyang, D., "A survey on deep packet inspection for intrusion detection systems", ***arXiv preprint arXiv:0803.0037***, (2008).
9. Kim, J., Hwang, J. and Kim, K., "High-performance internet traffic classification using a markov model and kullback-leibler divergence", ***Mobile Information Systems***, Vol. 2016, (2016).
10. Muehlstein, J., Zion, Y., Bahumi, M., Kirshenboim, I., Dubin, R., Dvir, A. and Pele, O., "Analyzing https encrypted traffic to identify user's operating system, browser and application", in Consumer Communications & Networking Conference (CCNC), 2017 14th IEEE Annual, IEEE., (2017), 1-6.
11. Loo, H.R. and Marsono, M.N., "Online network traffic classification with incremental learning", ***Evolving Systems***, Vol. 7, No. 2, (2016), 129-143.
12. Qin, T., Wang, L., Liu, Z. and Guan, X., "Robust application identification methods for p2p and voip traffic classification in backbone networks", ***Knowledge-Based Systems***, Vol. 82, (2015), 152-162.
13. Aliakbarian, M.S., Fanian, A., Saleh, F.S. and Gulliver, T.A., "Optimal supervised feature extraction in internet traffic classification", in Communications, Computers and Signal Processing (PACRIM), 2013 IEEE Pacific Rim Conference on, IEEE., (2013), 102-107.

14. Zhou, W., Dong, L., Bic, L., Zhou, M. and Chen, L., "Internet traffic classification using feed-forward neural network", in Computational Problem-Solving (ICCP), 2011 International Conference on, IEEE., (2011), 641-646.
15. Crotti, M., Dusi, M., Gringoli, F. and Salgarelli, L., "Traffic classification through simple statistical fingerprinting", ***ACM SIGCOMM Computer Communication Review***, Vol. 37, No. 1, (2007), 5-16.
16. Moore, A.W. and Zuev, D., "Internet traffic classification using bayesian analysis techniques", in ACM SIGMETRICS Performance Evaluation Review, ACM. Vol. 33, (2005), 50-60.
17. Zhang, J., Chen, C., Xiang, Y., Zhou, W. and Xiang, Y., "Internet traffic classification by aggregating correlated naive bayes predictions", ***IEEE Transactions on Information Forensics and Security***, Vol. 8, No. 1, (2013), 5-15.
18. Hu, L. and Zhang, L., "Real-time internet traffic identification based on decision tree", in World Automation Congress (WAC), 2012, IEEE., (2012), 1-3.
19. Wang, Y., Xiang, Y. and Yu, S., "Internet traffic classification using machine learning: A token-based approach", in Computational Science and Engineering (CSE), 2011 IEEE 14th International Conference on, IEEE., (2011), 285-289.

Behavioral Analysis of Traffic Flow for an Effective Network Traffic Identification

M. Gandomi[a], H. Hassanpour[b]

[a] *Artificial Intelligence, Shahrood University of Technology, Shahrood, Iran*
[b] *Department of Computer Engineering and IT at Shahrood University of Technology , Iran*

P A P E R I N F O

Paper history:
Received 20 July 2017
Received in revised form 27 August 2017
Accepted 08 September 2017

Keywords:
Network Traffic Identification
Behavioral Analysis
Data Mining
Machine Learning
Flow Statistical Featur

چکیده

امروزه با افزایش و توسعه برنامه‌های کاربردی تحت شبکه، شناسایی و طبقه بندی سریع و با دقت ترافیک شبکه جهت بالابردن کیفیت سرویس دهی و شناسایی ناهنجاری ها، نیاز مبرم مدیران شبکه می باشد. تاکنون روش‌های متعددی جهت شناسایی ترافیک شبکه ارائه شده است. در این میان روش‌های مبتنی بر تحلیل آماری بسته‌ها به کمک روش‌های یادگیری ماشین، دارای اهمیت بالایی می‌باشند. در این مقاله روشی جهت شناسایی ترافیک شبکه ارائه شده است. در این روش جهت آماده سازی داده‌ها از یک راه‌حل پیش‌پردازش موثر استفاده شده است. سپس ویژگی‌های بسته ها با استفاده از تحلیل رفتار برنامه‌های کاربردی در ساخت، ارسال و دریافت بسته‌ها جهت تعامل با کاربران استخراج شده‌اند. دلیل این امر را می‌توان عدم تغییر رفتار و مکانیسم اجرایی برنامه‌های کاربردی در نسخه‌های مختلف دانست. به همین منظور ابتدا رفتار شش برنامه کاربردی اسکایپ، فایرفاکس، کروم، اینترنت اکسپلورر، تلگرام و مدیریت دانلود مورد بررسی و ارزیابی قرار گرفته و با توجه به رفتار این برنامه‌های کاربردی ویژگی‌هایی استخراج شده است. برای ارزیابی ویژگی های ارائه شده، داده‌های دو پایگاه داده UNIBS و پایگاه داده جمع آوری شده بر روی مسیریاب، مورد استفاده و تحلیل قرار گرفته است. جهت ارزیابی نتایج، میزان صحت تشخیص درست برنامه کاربردی توسط الگوریتم های یادگیری ماشین مورد تحلیل قرار می‌گیرد. نتایج نشان می‌دهد که روش ارائه شده میزارن دقت شناسایی برنامه های کاربردی را به بیش از ۹۷ ٪ افزایش می-دهد.

***doi**: 10.5829/ije.2017.30.11b.15*