Neurocomputing 156 (2015) 252–267

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Effective packet number for early stage internet traffic identification

Lizhi Peng, Bo Yang, Yuehui Chen*

Shandong Provincial Key Laboratory for Network Based Intelligent Computing, University of Jinan, Jinan 250022, PR China

ARTICLE INFO

Article history:
Received 13 May 2014
Received in revised form
29 October 2014
Accepted 24 December 2014
Communicated by D. Liu
Available online 5 January 2015

Keywords:
Feature extraction
Early stage traffic classification
Machine learning

ABSTRACT

Accurately identifying Internet traffic at the early stage is very important for the applications of traffic identification. Recent years, more and more research works have tried to build effective machine learning models to identify an Internet flow with the few packets at its early stage. However, a basic and important problem still needs to be studied in depth, that is how many packets are most effective in early stage Internet traffic identification. In this paper, we try to resolve this problem. Three Internet traffic data sets are applied. And the sizes of the first 10 packets are extracted for study. We firstly apply mutual information to analyze the information that the first n packets provide to the flow type. Then correlation analysis of each pair of adjacent packets is carried out to find out the feature redundancies. And then we execute a number of crossover identification experiments with different numbers of packets using 11 well-known supervised learning algorithms. Finally, statistical tests are applied for the experimental results to find out which number is the best performed one. Our experimental results show that 5–7 are the best packet numbers for early stage traffic identification.

1. Introduction

Recent years, early stage traffic identification has caught enough interests at the research community. Most traditional machine learning based traffic identification techniques extract features on a whole network flow [15,29,34]. The most widely followed feature extracting method is presented by Moore et al. in 2005 [33]. They extract 248 statistical features based on a whole flow, such as the maximum, minimum and average values of the packet sizes, and RTT. Classifiers using such statistical features can get very high identification performances [36], and such statistical features even have been successfully applied in anomaly detection [16]. However, in real circumstances, it makes no sense to recognize Internet flows when they have ended. Thus, we must identify them accurately in their early stages so that we can apply the subsequent management and security policies. Therefore, some researchers have turned to find effective models which are able to identify Internet flows at their early stages. This makes early stage identification to become a hot topic in traffic identification researches [10]. Qu et al. have studied the problem of accuracy of early stage traffic identification, and found that it is possible to identify traffic accurately at their early stages [40].

However, an important problem still should be further studied: How many packets are most effective for early stage traffic identification? As far as we know, there are only a few studies that concern on this issue. In this study, we set out to study selecting the most effective number of packets in early stage traffic identification using both information analysis methods and empirical methods. Three traffic data sets and eleven widely used classifiers are applied for our study. We use the application layer payload size of each packet as the feature. Mutual information analysis is firstly applied to discover how much identification information the first n packets can provide. Then we try to find the redundances between each pair of adjacent packets using correlation analysis. And then all selected classifiers are applied in a group of identification experiments using different number of packets. At last, the experimental results are analyzed using statistical hypothesis tests to find out the best behaved packet number.

The rest of the paper is organized as follows: We firstly review the related work in Section 2, and then introduce the data sets used in the study in Section 3. Section 4 illustrates the framework and detailed steps of our study. Then the methods applied are depicted in Section 5, include the basic theories of mutual information and Pearson's correlation coefficient, selected classifiers, and statistical tests used in the study. The details of results and analysis are given in Section 6. Some discussions will be presented in Section 7. Finally, we make some conclusions in Section 8.

2. Related work

It is relatively hard to recognize an Internet flow only using the few early stage packets. Thus, a key problem of early stage traffic identification is to extract effective features. Bernaille et al. presented

* Corresponding author.
E-mail addresses: plz@ujn.edu.cn (L. Peng), yangbo@ujn.edu.cn (B. Yang), yhchen@ujn.edu.cn (Y. Chen).

http://dx.doi.org/10.1016/j.neucom.2014.12.053
0925-2312/© 2015 Published by Elsevier B.V.

a famous early stage traffic identification technique in 2006 [2]. They use the size of the first few data packets of each TCP flow as the features, and by applying K-means clustering technique, they got high identification rates for 10 types of application traffics. Este et al. have proved in 2009 [14] that early stage packets of an Internet flow carry enough information for traffic classification. They analyzed round trip time (RTT), packet size, inter-arrival time (IAT) and packet direction of early stage packets and found that the packet size is the most effective feature for early stage classifications. Huang et al. have studied the early stage application characteristics and used them for classification effectively in 2008 [21]. Recently, they extracted early stage traffic features by analyzing the negotiation behaviors of different applications, and applied these features for machine learning based classifiers with high performances [22]. Hullár et al. proposed an automatic machine learning based method consuming limited computational and memory resources for P2P traffic identification at the early stage [24]. Dainotti et al. [9] construct high effective hybrid classifiers for early stage traffic classification. Nguyen et al. use statistical features derived from sub-flows for timely identification of VoIP traffic [35], they extend the concept of early stage to “timely”, since a sub-flow refers to a small number of most recent packets taken at any point in a flow's lifetime. Rizzi et al. proposed a highly efficient neuro-fuzzy system for early stage traffic identification [41].

In [2], the authors say five packets are enough to distinguish the application behaviors. The authors in [14] use the first six packets of a flow for their study, but they also did not say why six packets are used. In [22], the authors extract early traffic features from 20 packets, and the number is also an empirical value. Dainotti et al. apply a hybrid feature extraction method for their work [9]. They use packet size (PS) and inter-packet time (IPT) of the first 10 packets for some classifiers, while for other classifiers, they use average and standard deviation values of PS and IPT of the early packets. For all of these studies, the number of packets in the early stage of flows is selected empirically. It is obvious that the number should be neither too large nor too small. Using too many packets will increase the computational complexity of the feature extraction procedures and decrease the efficiency of the identification models. While using too few packets will reduce the identification accuracies since they cannot contain enough characteristics of the flows.

L. Bernaille et al. have concerned about the problem of effective packet number for early traffic identification in 2006 [3]. They have studied the influences of the size and direction of the first 10 packets of Internet TCP connections, and have used K-means, GMM and HMM models for early stage traffic classification. They firstly think that packets exchanged after the application negotiation phase are not standardized and their sizes no longer help to recognize the application, and drawn the conclusion that $P=4$ (using the first four packets) is the most effective for the selected clustering methods. They also have executed a number of experiments using eight traffic traces collected on eight different networks, and the results verify that all of the three selected algorithms are able to get the highest application identification performances using the first four packets for most cases. However, they have not studied the influences of the first few packets for supervised learning techniques. Sena and Belzarena also have analyzed the size of the first few packets on both directions of Internet flows as a relevant statistical fingerprint [42]. Different from the unsupervised techniques used in the researches of Bernaille, support vector machine was applied as the baseline model in Sena's work. A centroid clustering algorithm also was applied in this paper. The author executes their experiments on the traffic traces collected on a ISP's network, and compared the classification accuracies using the first five, six and seven packets. Their results of the both selected algorithm show that five or six packets in each flow direction are enough for early stage traffic identification. It is obvious that the packet number range studied in this work is narrow, and the coverage of learning techniques is also narrow. Lim et al. have carried out an empirical study on the topic in 2010 [30]. They use a number of data sets including five Internet traffic traces collected at two backbone links of different countries, and the features they studied including not only the size of the first 10 packets, but also some connection level features and statistical features. They have used Naive Bayes, k-nearest neighbors, C4.5 decision tree and support vector machine for the experimental validations. They have found in their experiments that the size of the first packet contributes the most in identifying UDP application flows, and for TCP flows, the sizes of the second-sixth packets are the key features to identify the causing applications. Unfortunately, their work also mainly stays on empirical studying.

3. Data sets

3.1. Auckland II traffic traces

Auckland II is a collection of long GPS-synchronized traces taken using a pair of DAG 2 cards at the University of Auckland which is available at [47]. There are 85 trace files which were captured from November 1999 to July 2000. Most traces were targeted at 24 h runs, but hardware failures have resulted in most traces being significantly shorter. We selected two trace files captured at February 14, 2000 (20000214-185536-0.pcap and 20000214-185536-1.pcap) for our study. The traces include only the header bytes, with a maximum amount of 64 bytes for each frame, while the application payload is fully removed. And all IP addresses anonymized using Crypto-Pan AES encryption. The header traces were captured with a GPS synchronized mechanism using a DAG3.2E card connected to a 100 Mbps Ethernet hub interconnecting the University's firewall to their border router.

Since the application payloads were not recorded in Auckland II, DPI tools are invalid to obtain ground truths. The only way to pick out the original application type is using port numbers. In this study, we only accounted TCP case since TCP is the predominant transport layer protocol. Each flow is thus assigned to the class identified by the server port. We selected 8 main types from Auckland II traces and filtered mouse flows with no more than 10 non-zero packets. Table 1 lists all selected types and their instance and total bytes distributions.

3.2. UNIBS traffic traces

UNIBS is another opening traffic traces developed by Gringoli and his research team, available at [46]. They developed a useful system namely GT [27] to application ground truths of captured Internet traffics. The traces were collected on the edge router of the campus network of the University of Brescia on three consecutive working days (September 30, October 1 and October 2, 2009). They are composed of traffic generated by a set of 20 workstations running the GT client daemon. Traffics were collected

Table 1
Characteristics of Auckland II traces.

Type	# instances	Total bytes
ftp	251	136,241
ftp-data	463	5,260,804
http	23,721	139,421,961
imap	193	86,455
pop3	498	98,699
smtp	2602	1,230,528
ssh	237	149,502
telnet	37	21,171