Computer Networks 119 (2017) 102–111

Contents lists available at ScienceDirect

Computer Networks

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

Novel feature selection and classification of Internet video traffic based on a hierarchical scheme

Yu-ning Dong[a,*], Jia-jie Zhao[a], Jiong Jin[b]

[a] College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[b] School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, VIC 3122, Australia

ARTICLE INFO

Article history:
Received 4 October 2016
Revised 15 March 2017
Accepted 24 March 2017
Available online 29 March 2017

Keywords:
Statistical features
QoS
Video traffic classification
k-Nearest Neighbor classification

ABSTRACT

Accurate traffic classification is critical for efficient network management and resources utilization. Different video traffics have different QoS (Quality of Service) requirements. To provide Internet video services with better QoS support, a fine grained classification scheme for network video traffic is proposed in this paper. Through extensive statistical analysis of typical video traffic flows with a consistency-based method, several new flow statistical features are extracted. They are found to be more effective in discriminating different video traffics, especially from the QoS perspective, than commonly used features available in the literature. A hierarchical *k*-Nearest Neighbor (*k*NN) classification algorithm is then developed based on the combinations of these statistical features. Experiments are performed to evaluate the effectiveness of the proposed method on a large scale real network video traffic data. The experimental results show that the proposed method outperforms existing methods applying commonly used flow statistical features.

1. Introduction

In the recent years, Internet video services are widely used for the rapid development of network and multimedia technologies. According to Cisco's forecast report [1], the proportion of Internet video traffic in 2020 will reach 82% of consumer Internet traffic where consumer includes fixed IP traffic generated by households, university populations, and Internet cafes. Meanwhile, a variety of new applications and diverse protocols make the network environment extremely complex, which undoubtedly causes a series of problems such as efficient network management and QoS (Quality of Service) guarantees for multimedia services. In order to cope with these challenges from the perspective of Internet Service Providers (ISP) and network regulators, accurate classification of network traffic is believed to be the key [2]. For different traffic types, operators can allocate network resources according to their QoS requirements.

This work is motivated by the fact that different video applications may require different QoS support and network resources, while video flows belonging to the same category usually have similar QoS requirements. The correct identification of video flows can thus help ISPs to understand what level of QoS they need and make appropriate resource allocation for them in order to improve the end user's quality of experience. The paper presents an Internet video traffic classification scheme based on hierarchical *k*-Nearest Neighbor (*k*NN) method. This scheme first identifies the most effective QoS related discriminative features and feature combinations by applying statistical analysis and data mining techniques to the acquired video data; then classifies the video flows with a hierarchical classification algorithm using the above features or features combination. To the best of our knowledge, this is the first work that attempts to tackle the video service classification issue with finer granularity from a QoS perspective.

Part of our work has been reported in [3]. However, this paper extends the work of [3] substantially in several aspects. Our main contributions include: 1) a consistency-based feature analysis and selection method is presented to systematically find some new and effective features for video traffic classification; 2) a new hierarchical *k*NN classification scheme is proposed which uses the feature combinations. We also give out more detailed descriptions of flow features analysis in this work. Experimental results show that this scheme can achieve better classification accuracy than existing methods using commonly statistical features.

The rest of the paper is structured as follows. Related works are reviewed in Section 2. Section 3 presents the consistency-based method of feature analysis and selection. Section 4 gives the description of hierarchical *k*NN classification algorithm. Section 5 reports the experiment results. Finally, the paper concludes in Section 6.

* Corresponding author.
E-mail addresses: dongyn@njupt.edu.cn (Y.-n. Dong), fzzjj2008@126.com (J.-j. Zhao), jiongjin@swin.edu.au (J. Jin).

http://dx.doi.org/10.1016/j.comnet.2017.03.019
1389-1286/

2. Related work

In the last decade, there has been extensive research on exploring traffic classification method [4]. Here we first review relevant works on ML (machine learning) methods for traffic classification, then focus on related researches on Internet video traffic.

2.1. Internet traffic classification

The relevant research on network traffic classification mainly focuses on ML methods based on flow statistical characteristics [5], including supervised and unsupervised methods.

2.1.1. Supervised methods

The supervised classification methods are provided with a collection of traffic datasets and their pre-identified classes. Zhang et al. [6] proposed a Naïve Bayes based classification scheme using feature discretization and flow correlation. A large scale real-world network dataset with different protocols was carried out in their work. Neural networks have a built-in ability to modify their synaptic connections and weights to adapt to the surrounding environment. Their attributes of non- linearity and adaptability are especially desirable for P2P traffic identification [7]. Jaiswal et al. [8] developed a reduced statistical feature dataset and compared six ML algorithms for traffic classification. Various internet applications (VoIP, Multimedia Streaming, bulk data transfer, Interactive traffic, Email services, WWW traffic and Database traffic) are used in their work. Du et al. [9] put forward a method of P2P traffic identification based on support vector machines (SVM). The method could effectively detect the P2P traffic network flows with three statistical characteristics.

*k*NN classifier is a simple yet often effective supervised method, which includes many advantages (as given in 4.1). A large amount of literature has been reported on researches with *k*NN. Zhang et al. [10] proposed three improved Nearest Neighbor algorithms by incorporating correlated information into the classification process. Their experiments are carried out on two real-world traffic data sets. Patwary et al. [11] presented and implemented PANDA, a parallel and distributed kd-tree based KNN algorithms. The results show that their methods are more suitable for state-of-the-art Big Data analytics problems. Silas et al. [12] designed and implemented a real time flow-based network traffic classification system (NTCS). Their identified application traces include: Www, Https, Ftp, Xvttp and Isakmp Protocol. The modules of the system are built as concurrent processes, which are more effective on Internet traffic monitoring. They use some machine learning methods, such as *k*NN, C4.5 Decision Tree and AdaBoost, to validate the reliability of the approach.

2.1.2. Unsupervised methods

The Unsupervised classification method is essentially a synonym for clustering. Clustering techniques have been applied in the context of Internet traffic analysis for a long time. Zhang et al. [13] adopted a semi-supervised ML technique to effectively discriminate zero-day application. Liu et al. [7] applied FCM clustering to classify P2P application. Their work aimed at reducing the computational complexity of FCM (Fuzzy C-Means) while keeping the clustering accurate. Zhang et al. [14] proposed an encrypted traffic classification scheme based on improved *K*-Means that helps reduce the impact of random initial clustering centers. Their dataset sampled from 5 classes: Skype, QQ, SSH, SSL, MSN.

2.2. Internet video traffic classification

Previous works mostly classified video traffic into one or two classes with coarse granularity. For example, Mu et al. [15] proposed a parallelized network traffic classification scheme using hidden Markov model, where they divided video services into conversational and streaming videos; Gonçalves et al. [16] merely distinguished peer-to-peer (P2P) video traffics. Nguyen et al. [17] make an identification of first-person-shooter online game and VoIP traffic by Naive Bayes and C4.5 Decision Tree ML algorithms. They achieved IP traffic classification by using statistics derived from sub-flows–a small number of packets taken at any point in a flow's lifetime. Claypool et al. [18] proposed a scenario to analyze the network performance of the OnLive thin client game system. Their scenario verifies the differences among traditional network game, thin client game, pre-record video and on live video in terms of bit rate, packet size and inter-packet time. They found that different video streams have different features. Datta et al. [19] presented a case study of Google Hangouts (a semi peer-to-peer application). They used application semantics to identify a set of features and three conventional classification to assess the performance. Wang et al. [20] find two features (downstream/upstream bandwidth) appropriate to classify Internet video traffics, and they propose a modified K-SVD classification algorithm to get QoS classes. Zink et al. [21] proved that trace statistics is relatively stable over short time period while long term trends can be observed. They also showed that P2P paradigm can reduce network video traffic significantly and allow for faster access to video clips.

Nevertheless, our work is substantially different from all the previous ones. Because most of them are either for a particular type of traffic, or just to emphasize on the improvement of algorithms without considering the key problem, that is, how to mine meaningful features from the original flow to enhance the performance [2]. This paper will thus be dedicated to find useful statistical characteristics that can better distinguish different types of video traffic flows.

Table 1
Dataset of video applications.

Application	Traffic class	Volume
ASD	Asymmetric standard definition videos	0.56 GB
AHD	Asymmetric high definition videos	1.23 GB
HTTP-download	HTTP-download videos	2.95 GB
QQ	Interactive video communication class	1.19 GB
Xunlei	P2P video data sharing	4.48 GB
Sopcast	Network live TV	2.62 GB

3. Dataset and feature selection

We have adopted *k*NN algorithm to classify six types (as given in 3.1) of typical video traffics. The key and novel idea we leverage is to select some novel features in a systematic way and build a hierarchical scheme to classify flows from our data acquisition. In the following we further describe our data acquisition and features we selected.

3.1. Data acquisition

We captured real-world video flow data using WireShark[1] in NUPT campus network environment in different time (morning, afternoon and evening) from October 2013 to July 2014. In our study, a flow sample refers to sequences of packets captured in 30 min during the flow's life time of a video application. We grabbed 60 video flows for each application, namely 360 flows in total. Total size of captured data is 13.03GB (see Table 1).

The captured flow trace data contains five columns: the packet arrival time, source and destination IP addresses, protocol and

[1] http://wiki.wireshark.org/;.