Computer Networks 55 (2011) 3915–3931

Contents lists available at SciVerse ScienceDirect

Computer Networks

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

Session-based classification of internet applications in 3G wireless networks

Seongjin Lee [a], Jongwoo Song [b], Soohan Ahn [c], Youjip Won [a,*]

[a] Dept. of Electronics and Computer Engineering, Hanyang University, Republic of Korea
[b] Dept. of Statistics, Ewha Womans University, Republic of Korea
[c] Dept. of Statistics, University of Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 23 April 2010
Received in revised form 20 July 2011
Accepted 5 August 2011
Available online 25 August 2011

Keywords:
Traffic classification
CART
SVM
Clustering
HSDPA

ABSTRACT

Accurately classifying and identifying wireless network traffic associated with various applications, such as Web, VoIP, and VoD, is a challenge for both service providers and network operators. Traditional classification schemes exploiting port or payload analysis are becoming ineffective in actual networks, as many new applications are emerging. This paper presents the classification of HSDPA network traffic applications using Classification and Regression Tree (CART) and Support Vector Machine (SVM) with the session information as a basic measure. The session is bidirectional traffic stream between two hosts that is used as a basic measure and a unit of information. We acquired and processed HSDPA traffic from a real 3G network without sanitizing the data. CART and SVM are used to classify six application groups (download, game, upload, VoD, VoiP, and web) with a set of twelve easily retrievable features. These features are composed of simple statistical pieces of information, such as the standard deviation of the packet sizes, the number of packets, and the duration of a session. Compared to results of a flow-based application classification, session-based classification produces 11.07% (CART) and 21.99% (SVM) increases in the true positive rate. This feature set is further reduced to two principal components using Principal Component Regression. This paper also takes the initiative to compare CART to *K*-Means, the wired network traffic clustering scheme, and shows that CART is more accurate for classification than is *K*-Means.

1. Introduction

1.1. Motivation

Decisions on deploying wireless network resources, such as access networks and base stations, are difficult to make because the networks are costly and require extensive planning in terms of the capacity and quality of services (QoS). The QoS and network performance of wireless network is sensitive to many environmental factors, such as mobility, signal power, and interference. Wireless network must work within these constraints to provide various types of content over HSDPA or WiMAX networks, also known as WiMAX. As wireless services are becoming widely available, the classification of services, such as web and VoIP, is becoming increasingly important because service providers need user behavior and statistical information to enhance the QoS at a lower cost. If applications used in wireless environments can be identified, user behavior can be analyzed to create a better and more efficient charge model.

One classification method reads each packet's payload and analyzes which application it belongs to [1,2]; another

* Corresponding author. Address: Dept. of Electronics and Computer Engineering, Hanyang University, Haengdang 1(il)-dong Seongdong-gu, 133-791 Seoul, Republic of Korea. Tel.: +82 2 2220 4579; fax: +82 2 2220 4579.
E-mail addresses: insight@hanyang.ac.kr (S. Lee), josong@ewha.ac.kr (J. Song), sahn@uos.ac.kr (S. Ahn), yjwon@hanyang.ac.kr (Y. Won).

1389-1286/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2011.08.010

method compares the port numbers with the Internet Assigned Numbers Authority (IANA). However, these methods are not the most efficient because opening a packet's payload not only involves privacy issues but is also time-consuming, and applications like Skype encrypt the payload [3], making this step unfeasible. Comparing the port numbers with IANA port numbers is not efficient because there are numerous new applications, such as P2P programs, which are not assigned to a specific port number but which have dynamic port numbers in order to hide application's behavior.

Dahmouni et al. [4] use a Markovian signature-based approach to detect anomalies in the traffic and automatically identify the associated application. The first step in the Markov chain is to identify the state space. Dahmouni et al. use control packets generated by an application as the states of the Markov chain. They reorder the packets in the emission order of a flow. The final step is to estimate the transition probabilities between each of the states. Other than the fact that the Markovian signature-based approach necessitates an assumption that the order of the packets in a flow does not change between emission and reception, it seems to be a promising solution for the classification problem. However, the reordering of the packets means that it is necessary to inspect all of the packets in a flow and in the whole traffic to figure out its emission time. This inspection becomes a problem, especially in the wireless environment where connections are frequently disturbed and packets lost.

Crotti et al. [5] exploit simple properties of IP packets, such as size, interarrival, and arrival order of the packet, along with a structure called protocol fingerprints. To generate an application layer protocol's fingerprint, the Probability Density Function (PDF) must be estimated. Estimation of PDF requires a set of flows generated by the same known protocol. However, if the PDF is not estimated, packets received by the system will be categorized as "unknown".

HSDPA or WiMAX network applications share limited resources; a connection can be closed at any time that a user desires, and these applications are noise sensitive. Along with the aforementioned characteristics, there are many other properties that lead to different statistical characteristics compared to those of wired networks. Due to the difference in the characteristics of the network, we cannot assume that classifying the Internet application in the HSDPA or WiMAX network is the same as classifying them in wired networks. From the service provider's point of view, knowing the statistical characteristics of usages of various applications is important because this information can be exploited as base knowledge for establishing fee systems. The classification of the applications also helps to identify wireless user behaviors.

In order to accurately detect 3G application groups, we provide a classification method with features based on the basic statistics of a flow. A flow is defined as packets with the same values of five tuples, such as source IP/Port, destination IP/Port, and protocol. This flow provides valuable information on application-specific examination, charging models, and on-line stream analysis. A few papers have suggested the importance of integrating direction in representing packet size, but in this work we separate the ingress and egress traffic stream and combine them into what we term a *Session*. Our approach not only maintains the direction of the packet but also derives basic statistics on traffic in both directions. Establishing a simple set of features is important to identification, as is having the right set of features. In this work we propose twelve basic statistics as features, and we show that this set can be reduced even further to provide similar performance in terms of identification.

Users of 3G Networks use wireless services for many reasons, but the three major reasons may be web search, entertainment, and VoIP. In this paper, we categorized a set of dominant service groups in a 3G Network as Web, Download, Upload, Game, VoD, and VoIP. A well-known telecommunications company and leading 3G network service provider in Korea, SK Telecom, provided traffic traces for these applications. Traffic traces are acquired from a running system and are not sanitized in any form. Although the proposed scheme is examined and analyzed with HSDPA network traffic, it is designed such that it can be adaptively migrated into the WiMAX network. Wireless traffic is certainly very different from wired traffic. Ridoux et al. [6] traced the wide-area CDMA wireless network and a backbone wireline to compare the difference between the two networks using wavelet. They found that wireless-in and wireless-out traffic flows are independent, suggesting each flow to be as important as any other flows. Many have experimented with different machine learning algorithms and clustering algorithms, but there are no publicly available real traffic traces and approved frameworks to test the performance of various classification or clustering schemes. Without such a framework, it is a challenging task to compare various schemes and present a rational result. In this work, however, we evaluate different classification techniques and a clustering technique in an effort to provide knowledge on the performance of various schemes applied to wired networks. Classification and Regression Tree (CART) [7] and Support Vector Machine (SVM) [8] are used as Classification algorithms. We reduce the feature numbers by exploiting Principal Components Analysis (PCA), and we derive a linear regression model with a few PCs. We also evaluated the clustering algorithm introduced in [9], which is a *K*-Means clustering scheme, to show the applicability of a wired network's clustering algorithm to wireless networks.

1.2. Related work

Much of the traffic classification literature is based on the wired network environment, where there are extensively manipulated data mining and machine learning schemes, such as *K*-Means, DBSCAN, and AutoClass [10]. *K*-Means is an unsupervised learning algorithm that classifies a given dataset into *K* clusters by defining *K* centroids. These centroids are defined such that square of the distances between the datasets and the center (the so-called objective function) is minimumized. However, a downside of *K*-Means is that it has spherical shaped clusters, whereas DBSCAN, a Data Mining Algorithm for density-based spatial clustering, may have arbitrary shapes [10]. When the