Chapter 11
Classification of Peer-to-Peer Traffic Using a Two-Stage Window-Based Classifier with Fast Decision Tree and IP Layer Attributes

Bijan Raahemi
University of Ottawa, Canada

Ali Mumtaz
University of Ottawa, Canada

ABSTRACT
This paper presents a new approach using data mining techniques, and in particular a two-stage architecture, for classification of Peer-to-Peer (P2P) traffic in IP networks where in the first stage the traffic is filtered using standard port numbers and layer 4 port matching to label well-known P2P and NonP2P traffic. The labeled traffic produced in the first stage is used to train a Fast Decision Tree (FDT) classifier with high accuracy. The Unknown traffic is then applied to the FDT model which classifies the traffic into P2P and NonP2P with high accuracy. The two-stage architecture not only classifies well-known P2P applications, but also classifies applications that use random or non-standard port numbers and cannot be classified otherwise. The authors captured the internet traffic at a gateway router, performed pre-processing on the data, selected the most significant attributes, and prepared a training data set to which the new algorithm was applied. Finally, the authors built several models using a combination of various attribute sets for different ratios of P2P to NonP2P traffic in the training data.

DOI: 10.4018/978-1-61350-474-1.ch011
1. INTRODUCTION

Peer-to-Peer (P2P) is a type of Internet application that allows a group of users to communicate with each other, directly access and download files from the peers’ machine, and share computing resources (i.e., building a distributed computing environment). P2P traffic and its characteristics have changed the original assumptions under which the data networks were designed. P2P traffic is more symmetric (contrary to the assumption on which Asymmetric Digital Subscriber Line (ADSL) was designed); P2P traffic is less “bursty” which makes it difficult to take advantage of statistical multiplexing (under which the original data networks were designed). Also, P2P traffic lasts longer than typical web or email traffic, and packet lengths are mostly large, which keeps the queues in intermediate switches and routers more utilized, and consume more bandwidth and processing resources in the network devices. Various issues of traffic models associated with traffic in wireless networks are discussed in Doci et al. (2008) and Rohm et al. (2009).

Classification of Internet traffic is a fundamental requirement in areas such as network provisioning, network security, traffic engineering, and network management. Many efforts are made to classify the internet traffic for various applications including classification of P2P traffic by Internet Service Providers (ISPs) and corporate networks. P2P applications bypass central server control implemented by service providers and poses threats in terms of network congestion, and creating an environment for malicious attacks on networks. P2P applications may use randomly selected non-standard ports to communicate and consumes network resources (Shield, 2007). The volume and patterns of P2P traffic put pressure on service providers’ networks in terms of congestion and business models. For example, maintaining Quality of Services (QoS) planned in the access network requires the provisioning of additional bandwidth sooner than expected.

One key challenge in this area is to adapt to the dynamic nature of Internet traffic. With the growth in Internet traffic, in terms of number and type of applications, traditional classification techniques such as port matching, protocol decoding or packet payload analysis are no longer effective. For instance, P2P applications may use randomly selected non-standard ports to communicate which makes it difficult to distinguish them from other types of traffic by inspecting only port numbers. As such, several data mining techniques are proposed to classify the internet traffic based on its statistical characteristics such as packet length, packet inter-arrival time, session duration, and source and destination IP addresses. These include both offline (for static data) and online (for stream data) data mining techniques. Stream data mining represents an important class of data-intensive applications where data flows dynamically in large volumes, often demanding fast and real-time responses. Many of the established data mining algorithms perform well on static data. However, unlike data processing techniques for stored datasets, methods for analyzing stream data require fast, memory efficient and computationally inexpensive algorithms producing results concurrent with the flow of the stream with acceptable accuracy.

Further efforts have been made to develop techniques utilizing window-based algorithms. In this paper, we present a widow-based approach to capture and classify internet traffic using a two-stage classifier with fast decision tree. We captured Internet traffic at various time intervals, preprocessed the data, and selected the most significant attributes for classification which include IP packet length, source IP address and destination IP address. We ran several experiments using different attribute sets and various ratios of P2P and NonP2P traffic and measured the performance of the classifier. The results demonstrated that we can classify the traffic with accuracy higher than 90%.

The rest of the paper is organized as follows. Section 2 gives an overview of research related to classification problems. Section 3 presents