Identification of Network Applications based on Machine Learning Techniques

Pere Barlet-Ros, Valentín Carela-Español, Eva Codina, Josep Solé-Pareta
Advanced Broadband Communications Center (CCABA)
Universitat Politècnica de Catalunya (UPC)
Jordi Girona 1-3, 08034 Barcelona, Catalunya, Spain
{pbarlet, vcarela, ecodina, pareta}@ac.upc.edu

Keywords: Application identification, Traffic classification, Passive monitoring

Passive network monitoring systems are widely used by network operators to observe in real-time the traffic from operational networks. A representative example of such a system is the SMARTxAC platform [1], which is used by the Supercomputing Center of Catalonia (CESCA) for continuously monitoring the Catalan Research and Education Network (Scientific Ring).

One of the main usages of these systems is the identification of network applications. Traditionally, this identification was carried out by using the port numbers present in the packet headers, given that most applications were associated with a predefined set of port numbers (e.g., well-known ports).

However, nowadays it is widely accepted that port numbers no longer provide reliable enough information to accurately identify network applications. The main causes include the existence of a large number of web-based services, applications that use dynamic ports (e.g., FTP in passive mode) and the usage of tunnelling. Moreover, users tend to change the port numbers associated with their applications (e.g., P2P) to bypass firewalls, evade detection or avoid security attacks.

So far, several research works have proposed different solutions to the problem [2-6], most of them with relatively limited success. As a result, most network monitoring systems still use the port numbers to classify the network traffic. One of the alternatives most commonly adopted by network operators is the inspection of the packet contents by using pattern recognition algorithms (e.g., L7-filter [7]). Nevertheless, this technique presents three main limitations. First, it is not suitable for high-speed links, because existing pattern matching algorithms are computationally very expensive. Second, the new generation of P2P applications already support encryption and protocol obfuscation techniques that hinder the recognition process. Finally, the collection and inspection of packet contents may present privacy issues.

In this work, we present a novel identification method based on supervised machine learning techniques that addresses the problems described above. Machine learning is a broad branch of the Artificial Intelligence field that allows computers to extract knowledge from data provided in the form of examples (i.e., *training set*). In particular, the training set consists of pairs <*object*, *class*>, where the *object* is usually represented as a vector of features, whereas the *class* is the value to be predicted. Therefore, the task of a supervised learner is to find out those features that better predict the class of an object based on the examples within the training set.

In our case, the training set is composed of actual traffic flows collected in our network, while the feature vector contains particular features of each flow that we consider as relevant to predict the application they belong to (e.g., average packet size, flow duration or average interarrival time). The basic requirement of our method is that the monitoring system must be able to extract these features in real-time and they should not depend on the packet contents in order to avoid performance and privacy issues. Once the features are extracted, the traffic flows are classified according to the application they belong to. This phase is carried out by using pattern matching or by manually inspecting the packet contents. This detailed analysis is possible during the training phase given that it is executed offline. The training phase finishes with the generation of a decision tree using the C4.5 algorithm [8]. In this work, this phase was performed using the Weka [9] open source software, developed by the University of Waikato. A complete list of the features used in this work is available in [10].

Finally, the monitoring system uses this decision tree to identify in real-time the application of each collected flow, without inspecting the packet contents or relying only on the port numbers. Furthermore, once the system is trained, the identification method is lightweight enough to be used in high-speed networks without packet loss.

Figure 1 presents the preliminary results obtained with an actual implementation of this method in the SMARTxAC platform. The figure shows the identification accuracy broken down by application class (the average accuracy is of 97.14%) when monitoring two full-duplex Gigabit Ethernet links that connect the Scientific Ring to RedIRIS (the Spanish NREN). Figure 2 plots the application breakdown statistics reported by the original version of SMARTxAC (which uses the port numbers) compared to those obtained with the new method. While in the original version almost 50% of the traffic could not be classified (A_UKNWN), when using the proposed method most of this traffic was classified as P2P.

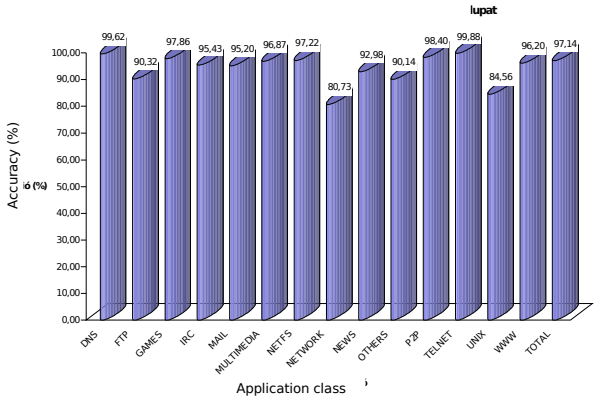

Figure 1: Identification accuracy broken down by application class using the proposed method based on machine learning techniques

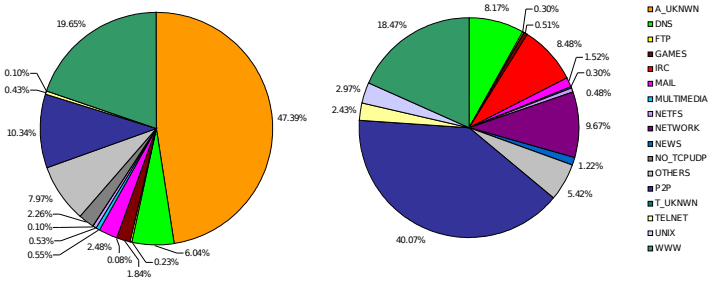

Figure 2: Application breakdown using port numbers (left) and using the proposed method based on machine learning techniques (right)

In the full paper version of this extended abstract we would describe our method in detail and include a comparison with existing alternative techniques. Currently, we are working on the definitive deployment of this method in the Scientific Ring network and on its validation with more datasets. The results obtained in this evaluation and the conclusions drawn from the continuous usage of this technique by our network operators would be also included in the full paper.

Acknowledgements

This work is supported in part by the Supercomputing Center of Catalonia (CESCA), under the SMARTxAC agreement, and by the Spanish Ministry of Science and Technology, under contracts TSI2005-07520-C03-02 (CEPOS) and TEC2005-08051-C03-01 (CATARO).

References

[1] BARLET-ROS, P; SOLÉ-PARETA, J; BARRANTES, J; CODINA, E; DOMINGO-PASCUAL, J. "SMARTxAC: A passive monitoring and analysis system for high-speed networks". TERENA Networking Conference 2006, Catania, Italy.

[2] KARAGIANNIS, T; BROIDO, A; FALOUTSOS, M; CLAFFY, K. C. "Transport layer identification of P2P traffic". Internet Measurement Conference. 2004. Taormina, Italy.

[3] MOORE, A. W; ZUEV, D. "Internet traffic classification using bayesian analysis techniques". ACM Sigmetrics. 2005. Banff, Canada.

[4] MOORE, A. W; PAPAGIANNAKI, K. "Toward the accurate identification of network applications". Passive and Active Measurement Conference. 2005. Boston, United States.

[5] KARAGIANNIS, T; PAPAGIANNAKI, K; FALOUTSOS, M. "BLINC: Multilevel traffic classification in the dark". ACM Sigcomm. 2005. Philadelphia, United States.

[6] ANTONIADES, D; POLYCHRONAKIS, M; ANTONATOS, S; MARKATOS, E; UBIK, S; ØSLEBØ, A; "Appmon: An application for accurate per application network traffic characterization". IST Broadband Europe 2006, Geneva, Switzerland.

[7] L7-FILTER. Application layer packet classifier for Linux. http://l7-filter.sourceforge.net, 21-10-2007.

[8] QUINLAN, J. R. "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers. 1993.

[9] WEKA 3. Data mining with open source machine learning software in Java. http://www.cs.waikato.ac.nz/~ml/weka/, 21-10-2007

[10] CODINA, E. "Identificación de aplicaciones de red basada en técnicas heurísticas". Master Thesis. Universitat Politècnica de Catalunya. 2006.

Vitae

Pere Barlet-Ros received a M.Sc. degree in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 2003. He is an Assistant Professor and Ph.D student at the Computer Architecture Department of UPC. He was also a visiting Ph.D. student at the National Laboratory for Applied Network Research (2004), Intel Research Cambridge (2004) and Berkeley (2007). His research interests are in the fields of network measurements, traffic analysis and evaluation of network performance.

Valentín Carela-Español received a M.Sc. degree in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 2007. He is currently a research assistant at the Computer Architecture Department of UPC. He was also a visiting student for six months at Ghent University in Gent, Belgium (2007). His research interests are in the fields of network measurements and traffic analysis.

Eva Codina received a M.Sc. degree in Computer Science from the Universitat Politècnica de Catalunya (UPC) in 2006. From 2003 to 2006, she was a Projects Scholarship Holder in the Computer Architecture Department of UPC.

Josep Solé-Pareta was awarded his M.Sc. degree in Telecommunication Engineering in 1984, and his Ph.D. in Computer Science in 1991, both from the Universitat Politècnica de Catalunya (UPC). In 1984 he joined the Computer Architecture Department of UPC, where he is currently a Full Professor in Computer Science and Communications. He did a 'Postdoc stage' (summers of 1993 and 1994) at the Broadband and Wireless Networking Lab. of the Georgia Institute of Technology. His current research interests are in broadband Internet, high-speed and optical networks, with emphasis on traffic engineering, traffic characterisation, traffic management, QoS provisioning and MAC protocols for legacy and optical metro networks.