PRACTICAL TRAINING DATASET GENERATION AND RETRAINING MECHANISM FOR ON-LINE PEER-TO-PEER TRAFFIC CLASSIFICATION

ROOZBEH ZAREI

UNIVERSITI TEKNOLOGI MALAYSIA

PRACTICAL TRAINING DATASET GENERATION AND RETRAINING MECHANISM FOR ON-LINE PEER-TO-PEER TRAFFIC CLASSIFICATION

ROOZBEH ZAREI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Electrical - Electronics & Telecommunications)

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

JANUARY 2012

Dedicated with deep gratitude feeling to my beloved parents for their endless love, support and encouragement.

ACKNOWLEDGEMENT

First and foremost, I give thanks and praise to Almighty God for giving me the life, wisdom and good health to successfully complete this study.

My utmost deep and sincere gratitude goes to my supervisor Dr. Muhammad Nadzir, for his enthusiasm, inspiration, encouragement, knowledge, support and constructive comments throughout this work. Thank you very much Dr. Muhammad Nadzir, for your help, your priceless advices and being my friend.

I would like to express my deep appreciations to my parents who have always supported me and taught me how to strive to achieve my goals and dream. Also, I like to thank my siblings for their love and encouragement.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the project.

Roozbeh Zarei , Johor Bahru, Malaysia

ABSTRACT

Peer-to-Peer (P2P) detection by Machine Learning (ML) classification is affected by the quality and recency of training dataset. Hence, to classify P2P traffic on-line requires the removal of these limitations. In this research work, a novel practical training dataset generation and automatic retraining mechanism for on-line P2P traffic classification are proposed. These two proposals are integrated in a system that removes the limitations of ML classification and makes them suitable for on-line P2P traffic classification. For the first part, a novel two-stage training dataset generation is proposed by combining a 3-class heuristic and a 3-class statistical classification to accurately generate training dataset. In the heuristic stage, traffic is classified as P2P, nonP2P or unknown. In statistical stage, a dual-Decision Tree (DT) is built based on dataset generated in heuristic stage to classify unknown traffic into three classes in order to reduce the amount of classified unknown traffics. The final training dataset is generated based on all flows which are classified in these two stages. In the second part of the system, an automatic retraining mechanism is proposed to satisfy the needs of retraining ML classifier by detecting the changes of traffic behavior and updating the on-line ML classifier with recent accurate training dataset. This mechanism evaluates the accuracy of the on-line ML classifier based on flows labeled by the two-stage training dataset generation. The on-line ML classifier is retrained if its accuracy falls below a predefined threshold. The proposed system has been evaluated on traces captured from the Universiti Teknologi Malaysia (UTM) campus network between October and November 2011. The overall results shows that the two-stage training dataset generation can generate accurate training dataset by classifying more than 95% of total flows with high accuracy (98.59%) and low false positive (0.91%). The on-line ML classifier which is built based on (J48) algorithm and training dataset generated by the two-stage training dataset generation classifies traffic with high accuracy (99%) by using the 25 feature extracted from first 5 packets of each flow. The results also show that using automatic retraining mechanism allow the on-line ML classifier able to maintain its accuracy above a set threshold over time.

ABSTRAK

Pengesanan jaringan rakan-ke-rakan (P2P) oleh klasifikasi jenis Pembelajaran Mesin (ML) dipengaruhi oleh kualiti dan pembaharuan dalam set data latihan. Oleh itu, untuk mengelaskan trafik P2P secara *on-line* kelemahan-kelemahan ini harus dihapuskan. Dalam kerja penyelidikan ini, kaedah generasi set data latihan yang baru dan praktikal serta mekanisme latih-semula automatik bagi klasifikasi trafik P2P *on-line* dicadangkan. Kedua-dua cadangan disatukan dalam sistem untuk mengatasi kelemahan klasifikasi ML dan menjadikannya sesuai untuk klasifikasi trafik *on-line* P2P. Untuk bahagian pertama, kaedah dua-peringkat generasi set data latihan yang baru dicadangkan dengan menggabungkan heuristik kelas-3 dan klasifikasi kelas-3 statistik untuk menjana set data latihan dengan tepat. Pada peringkat heuristik, trafik dikelaskan kepada P2P, bukan-P2P atau tidak-diketahui. Pada peringkat statistik, Pepohon dwi-Keputusan (DT) dibina berdasarkan set data yang dijana dalam peringkat heuristik untuk mengelaskan lalu lintas yang tidak diketahui kepada tiga kelas untuk mengurangkan jumlah lalu lintas yang dikelaskan sebagai tidak-diketahui. Set data latihan akhir yang dihasilkan adalah berdasarkan kepada semua aliran yang dikelaskan dalam kedua-dua peringkat ini. Dalam bahagian kedua sistem, mekanisme latih-semula automatik adalah dicadangkan untuk memenuhi keperluan latih-semula pengelas ML dengan mengesan perubahan tingkah laku lalu lintas dan mengemaskini pengelas *on-line* ML dengan set data latihan yang tepat dan terbaru ini. Mekanisme ini akan menilai ketepatan pengelas *on-line* ML yang berdasarkan aliran yang dilabel oleh generasi set data latihan dua peringkat. Pengelas *on-line* ML dilatih semula jika ketepatan jatuh di bawah tahap yang dipratentukan. Sistem yang dicadangkan telah dinilai berdasarkan data yang diperoleh dari rangkaian kampus Universiti Teknologi Malaysia (UTM) dari bulan Oktober hingga November 2011. Keputusan keseluruhan menunjukkan bahawa generasi set data latihan dua peringkat boleh menjana set data latihan yang tepat dengan mengklasifikasikan lebih daripada 95% jumlah aliran dengan ketepatan tinggi (98.59%) dan kadar *false positive* yang rendah (0.91%). Pengelas "on-line" ML yang dibina berdasarkan algoritma (J48) dan set data latihan yang dihasilkan oleh kaedah generasi set data latihan dua peringkat berjaya mengklasifikasikan lalu lintas dengan ketepatan yang tinggi (99%) dengan menggunakan 25 ciri-ciri yang diekstrak dari 5 paket pertama dalam setiap aliran. Keputusan juga menunjukkan bahawa penggunaan mekanisme latih-semula automatik membenarkan pengelas *on-line* ML mampu mengekalkan ketepatan melebihi tahap yang dipratentukan dalam jangka masa tertentu.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DNS	–	Domain Name System
HTTP	–	Hypertext Transfer Protocol
IANA	–	Internet Assigned Numbers Authority
IMAP	–	Internet Message Access Protocol
IRC	–	Internet Relay Chat
ML	–	Machine Learning
NetBIOS	–	Network Basic Input/Output System
NTP	–	Network Time Protocol
OC48	–	Optical Carrier-Level 48
OC192	–	Optical Carrier-Level 192
POP	–	Post Office Protocol
P2P	–	Peer-to-Peer
QoS	–	Quality of Service
SMTP	–	Simple Mail Transfer Protocol
SSH	–	Secure Shell
DT	–	Decision Tree
VPN	–	Virtual Private Network

LIST OF SYMBOLS

fn	–	False Negative
fp	–	False Positive
n_n	–	Total number of nonP2P packets
$n_{n \to n}$	–	the number of nonP2P packets classified as nonP2P
$n_{n \to p}$	–	the number of nonP2P packets classified as P2P
n_p	–	Total number of P2P packets
$n_{p \to n}$	–	the number of P2P packets classified as nonP2P
$n_{p \to p}$	–	the number of P2P packets classified as P2P
tn	–	True Negative
tp	–	True Positive

CHAPTER 1

INTRODUCTION

In recent decades, peer-to-peer (P2P) applications became widespread among network users and consume a large proportion of total network bandwidth [1]. This poses new challenges for network administers to manage network resources [2]. From quality-of-service (QoS) point of view, accurate traffic classification can serve as a tool for network resources' identification and QoS utilization for different applications [3]. The request-of-bandwidth management techniques which optimize network performance and provide QoS guarantees have grown rapidly [4]. Accordingly, accurate on-line network traffic classification is important in traffic engineering and intrusion detection.

1.1 P2P Traffic classification

Traditionally, network traffic can be simply classified by using the port-based technique [5–7]. This solution classifies the traffic according to the well-known ports defined by Internet Assigned Numbers Authority (IANA) [8]. Unfortunately this technique has become unreliable [6, 9] since new applications like P2P applications use arbitrary ports numbers. To overcome this problem, payload-based classification techniques [10, 11] that inspect the payload of packet have been proposed. Although payload-based classification can be extremely accurate [6], but it consumes a lot of process resources and fails to classify encrypted traffic. The heuristic classifications [4, 12, 13] are based on the host sending receiving behaviors have proposed to solve the limitation of the payload-based classification. In order to classify applications correctly, this method needs to study the behavior of particular hosts within a certain duration. Therefore, it cannot be applied in real-time traffic classification.

Several traffic classification based on Machine Learning (ML) have proposed to classify the Internet traffic based on the statistical characteristics [14–16]. They are able to identify encrypted traffic and applications that use dynamic ports. However, there are several limitations of ML classification. First, they require training data set to build classifier model. Second, their accuracy become low if the traffic behavior changes or new applications are released. Recently, the combinations of these methods have proposed [17–19] to cover the weaknesses of one methods with the advantage of the other.

1.2 Research Motivation

Based on various researches, more than 50% of internet traffic are P2P traffic [19, 20]. The performance of networks is affected by the drawbacks of P2P traffic and it is worse when P2P applications geting more popular. P2P applications consume network resources and hence, the measurement of actual P2P traffic is very important. The use of dynamic port numbers, masquerading techniques, port hopping and encryption by P2P application [21] cause the detection of P2P traffic to be very complex and challenging. As a result, accurate online identification play a vital role in many areas such as traffic engineering, QoS, and intrusion detection.

Although a lot of researchers have tried to classify P2P traffic [19,22,23], there is a need for improved method. Network traffic classification based on traditional techniques such as port-based or payload-based cannot be used for traffic classification, whereas port-based techniques fail to classify applications which use port hopping. Payload-based classification not only is computationally expensive but also fails to classify encrypted traffic. Heuristic classification has high accuracy but has slow speed it is slow (cannot be use for on-line traffic). On the other hand, statistical classification has the best accuracy but it needs prior knowledge and frequent training. A combination of heuristic and statistical can provide a hybrid classification framework where dataset can be generated by the heuristic classifier and on-line classifying be done by a statistical classifier.

1.3 Problem Statement

Recently, several hybrid traffic classification have been proposed [17–19, 24] to classify P2P traffic. That shows promising results such as high accuracy. This is mostly due to the use of statistical classification in hybrid traffic classification. On the other hand, there are some limitations rely on applying statistical classification in on-line traffic that were not addressed. The limitations of statistical classification are that they require training data set to generate classifier model and also their accuracy becomes low over time if the traffic behavior changes over time.

The performance of ML classifiers is very sensitive to the quality of the training dataset. They can classify traffic with high accuracy if they trained with accurate training dataset [25]. Generally, training dataset which used for statistical classification are generated by using payload-based classification [26]. Hence it is inaccurate due to limitations of payload-based classification. The work in [24] tried to address this issue by using heuristic classification to generate training dataset, but it generates a lot of unknown traffic and achieves high false positive (fp).

Another limitation is relied on the fact that the behavior of the traffic is changed over time due to two reasons. First, the network itself is dynamic and its parameters may change over time. Second, new trends of applications may appear and become popular at any given time in network [27]. Therefore the classifier generated with training dataset in fixed time can not classify accurately traffic over time and may become outdated. The work in [26] tried to solve this problem by generating new ML classifier when the current one becomes outdated, but the accuracy achieved for ML classifier was not high (around 88-97$\%$). This is probably due to using inaccurate training dataset to build the classifier since training dataset was generated by payload-based classification in this technique.

1.4 Research Contribution

This research work presents a system consisting of two parts to remove the limitations of ML classification and make them able for on-line P2P traffic classification. The first part (two-stage training dataset generation) is proposed based on combination of heuristic and statistical classification to remove the requiring of training dataset by generating accurate training dataset. While the second part

(automatic retraining mechanism) satisfies the needs of retraining by detecting the changes of traffic behavior and updating ML classifier.

A novel two-stage training dataset generation is proposed. In the proposed two-stage method, an improved heuristic classification is coupled with a dual-Decision Tree(DT) classifier to generate accurate training dataset. In first stage, traffic is classified as P2P, non-P2P, or unknown by using heuristic classification similar to [24]. Afterwards, the dual DT in 3-class statistical classification are built based on training dataset generated in first stage to classify unknown traffics. The role of the dual-DT is to reduce the unknown traffics. Our proposed two-stage training dataset generation is applied on many traces captured in UTM campus between October and November 2011 in order to access the performance and robustness of the system. Dual-DT classifier enhanced the training dataset by classify more than 90% of unknown traffic with more than 98% accuracy. Our proposed two-stage training dataset generation shows the ability to genetarte accurate training dataset by correctly classifing around 95% of total flows with high accuracy(1.41% class noise).

An automatic retraining mechanism is proposed to maintain the accuracy of on-line ML classifier above a certain threshold over time by detecting the changes of behavior traffic and retraining on-line ML classifier with recent accurate training dataset. Our proposed system consists of three phases. Packets are labeled using the two-stage training dataset generation in the first phase. Then an on-line ML classifier is built. In second phase, the accuracy of on-line ML classifier is evaluated based on the flows which are labeled by the two-stage training dataset generation. Finally the on-line ML classifier is retrained if the accuracy of current one falls below a predefined threshold. The system is evaluated with different accuracy thresholds in order to access the performance of the system. A synthesis dataset which consists of 25,000 flows is used to evaluate the accuracy of on-line ML classifier. The results show that the on-line ML classifier can classify P2P traffic based on the first 5 packets in each flow and maintain it's accuracy above a set threshold over time by utilizing of the proposed retraining mechanism.

1.5 Organization of the Report

The reminder of this report is organized as follows. In Chapter 2 the related works on P2P traffic classification and also the limitations of ML classification are discussed. Chapter 3 introduces a framework for Practical training dataset generation and automatic retraining mechanism for On-line P2P traffic classification. This chapter includes research methodology, experimental set-up, and benchmark criteria. Chapter 4 presents a two-stage training dataset generation, discussing the accuracy of proposed system, and assessing the performance of the heuristics and DTs. In Chapter 5, An automatic retraining mechanism is proposed and its effect on accuracy of the on-line ML classifier is discussed. Chapter 6 concludes the research work and point out potential future directions for this work.

CHAPTER 2

BACKGROUND AND RELATED WORKS

This chapter presents related literatures in the P2P traffic classification. The limitations of P2P traffic classification techniques are discussed. We review different P2P traffic classification techniques such as port-based, payload-based, heuristic, and statistical to investigate the advantages and disadvantages of each approach. Afterward, we discuss the limitations of P2P ML classifications which are related to applying ML classification in on-line traffic classification and also some techniques which used to remove these limitations are explained.

This chapter is organized as follow. Section 2.1 reviews different methods of P2P classification. Section 2.2 investigates the limitations of ML classifications and presents some methods which have been proposed to remove these limitations. In Section 2.3, the motivation for extended work is discussed.

2.1 P2P Classification Techniques

There are numbers of researches in P2P traffic classification [4, 11, 28–31]. Different approaches have been proposed by researchers to classify P2P traffic including port-based method [32], payload-based method [11], heuristic method [4,29] and statistical method [30, 33]. Several researches showed that some of these techniques can not be applied for classifying P2P traffic, since P2P traffic uses various obfuscation techniques such as port hopping and encrypted payloads. Hybrid techniques have also been proposed [4, 19]. These methods indicate promising result in P2P classification.

REFERENCES

1. Chen, Z., Yang, B., Chen, Y., Abraham, A., Grosan, C. and Peng, L. Online hybrid traffic classifier for Peer-to-Peer systems based on network processors. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., vol. 9. 685–694. 2009.

2. Mellia, M., Pescapè, A. and Salgarelli, L. Guest Editorial: Traffic classification and its applications to modern networks. *Comput. Netw.*, 2009. 53: 759–760.

3. Soysal, M. and Schmidt, E. G. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. *Performance Evaluation*, 2010. 67(6): 451 – 467.

4. Karagiannis, T., Broido, A., Faloutsos, M. and claffy, K. Transport layer identification of P2P traffic. In: *IMC '04: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. New York, NY, USA. 121–134. 2004.

5. Moore, D., Keys, K., Koga, R., Lagache, E. and Claffy, K. C. The CoralReef Software Suite as a Tool for System and Network Administrators. *Proceedings of the 15th USENIX conference on System administration*. Berkeley, CA, USA: USENIX Association. 2001. 133–144.

6. Moore, A. W. and Papagiannaki, K. Toward the Accurate Identification of Network Applications. *PAM*. 2005. 41–54.

7. Paxson, V. Bro: A System for Detecting Network Intruders in Real-Time. *Computer Networks*. 1999. 2435–2463.

8. IANA. http://www.iana.org/assignments/port-numbers, 2012.

9. Thomas Karagiannis, N. B. k. c. M. F., Andre Broido. Is P2P dying or just hiding? *GLOBECOM, Texas, USA*, 2004. Vol.3: 1532–1538.

10. Haffner, P., Sen, S., Spatscheck, O. and Wang, D. ACAS: automated construction of application signatures. *Proceedings of the 2005 ACM SIGCOMM workshop on Mining network data*. New York, NY, USA: ACM. 2005, MineNet '05. ISBN 1-59593-026-4. 197–202.

11. Sen, S., Spatscheck, O. and Wang, D. Accurate, scalable in-network identification of p2p traffic using application signatures. *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA. 2004. 512–521.

12. Karagiannis, T., Papagiannaki, K., Taft, N. and Faloutsos, M. Profiling the end host. *Proceedings of the 8th international conference on Passive and active network measurement*. Berlin, Heidelberg: Springer-Verlag. 2007, PAM'07. 186–196.

13. Xu, K., Zhang, Z. and Bhattacharyya, S. Profiling internet backbone traffic: behavior models and applications. 2005: 169–180.

14. Bernaille, L., Teixeira, R. and Salamatian, K. Early application identification. *Proceedings of the 2006 ACM CoNEXT conference*. ACM. 2006, CoNEXT '06. 6:1–6:12.

15. Moore, A. W. and Zuev, D. Internet traffic classification using bayesian analysis techniques. *Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*. ACM. 2005, SIGMETRICS '05. 50–60.

16. Williams, N., Zander, S. and Armitage, G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *SIGCOMM Comput. Commun. Rev.*, 2006. 36(5): 5–16.

17. Xu, K., Zhang, M., Ye, M., Chiu, D. M. and Wu, J. Identify P2P traffic by inspecting data transfer behavior. *Computer Communications*, 2010. In Press, Corrected Proof.

18. Lu, W., Tavallaee, M. and Ghorbani, A. Hybrid Traffic Classification Approach Based on Decision Tree. *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*. 2009. 1 –6.

19. Keralapura, R., Nucci, A. and Chuah, C.-N. A novel self-learning architecture for p2p traffic classification in high speed networks. *Computer Networks*, 2010. 54(7): 1055 – 1068.

20. Joun F.Buford, E. K. L., Heather Yu. *P2P:Networking and Applications*. Morgan Kaufmann. 2009.

21. Madhukar, A. and Williamson, C. A Longitudinal Study of P2P Traffic Classification. In: *MASCOTS '06: Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation*. Washington, DC, USA. 179–188. 2006.

22. Chunzhi, W., Wei, J., Hong, C., Luo, W. and Fang, H. Research on a method of P2P traffic identification based on multi-dimension characteristics. *Computer Science and Education (ICCSE), 2010 5th International Conference on*. 2010. 1010 –1013.

23. John, W. and Tafvelin, S. Heuristics to Classify Internet Backbone Traffic based on Connection Patterns. In: *Information Networking, 2008. ICOIN 2008. International Conference on Busan, Korea*. 1–5. 2008.

24. Hassan, M. and Marsono, M. A three-class heuristics technique: Generating training corpus for Peer-to-Peer traffic classification. *Internet Multimedia Services Architecture and Application(IMSAA), 2010 IEEE 4th International Conference on*. 2010. 1 –5. doi:10.1109/IMSAA.2010.5729416.

25. Thuy T.T. Nguyen, G. A. A Survey of Techniques for Internet Traffic Classification using Machine Learning. *IEEE Communications Surveys and Tutorials*, 2008. 4th edition.

26. Mula-Valls, O. *A practical retraining mechanism for network traffic classification in operational environments*. Master thesis. Universitat Politècnica de Catalunya.

27. Tian, X., Sun, Q., Huang, X. and Ma, Y. A Dynamic Online Traffic Classification Methodology Based on Data Stream Mining. *Computer Science and Information Engineering, 2009 WRI World Congress on*. 2009, vol. 1. 298 –302. doi:10.1109/CSIE.2009.904.

28. Sears, W., Yu, Z. and Guan, Y. An Adaptive Reputation-based Trust Framework for Peer-to-Peer Applications. *Network Computing and Applications, Fourth IEEE International Symposium on*. 2005. 13 –20. doi: 10.1109/NCA.2005.6.

29. Perényi, M., Dang, T., Gefferth, A. and Molnár, S. Identification and analysis of peer-to-peer traffic. *Journal of Communications*, 2006. 1(7): 36–46.

30. Li, W. and Moore, A. W. A Machine Learning Approach for Efficient Traffic Classification. In: *MASCOTS '07 Proceedings of the 2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*. Washington, DC, USA. 310–317. 2007.

31. Crotti, M., Dusi, M., Gringoli, F. and Salgarelli, L. Traffic classification through simple statistical fingerprinting. *SIGCOMM Comput. Commun. Rev.*, 2007. 37(1): 5–16.

32. Sen, S. and Wang, J. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Trans. Netw.*, 2004. 12: 219–232.

33. Raahemi, B., Hayajneh, A. and Rabinovitch, P. Peer-to-Peer IP Traffic Classification Using Decision Tree and IP Layer Attributes. *International Jornal of Buisniss Data Communication and Network*, 2007. 3: 60–74.

34. Kumar, S., Dharmapurikar, S., Yu, F., Crowley, P. and Turner, J. Algorithms to accelerate multiple regular expressions matching for deep packet inspection. *SIGCOMM Comput. Commun. Rev.*, 2006. 36: 339–350.

35. Kim, H., CAIDA, U., Barman, D. and Faloutsos, M. Comparison of Internet Traffic Classification Tools. *ANF Workshop*. 2007, vol. 2.

36. http://www.caida.org/tools/measurement/coralreef/, 2010.

37. Karagiannis, T., Papagiannaki, K. and Faloutsos, M. BLINC: multilevel traffic classification in the dark. *SIGCOMM Comput. Commun. Rev.*, 2005. 35(4): 229–240.

38. WEKA: Data Mining Software in Java. URL `http://www.cs.waikato.ac.nz/ml/weka/`.

39. Zhang, M., John, W., Claffy, K. and Brownlee, N. *State of the art in traffic classification: A research review*. Technical report. 2009.

40. Raahemi, B., Hayajneh, A. and Rabinovitch, P. Classification of Peer-to-Peer Traffic Using Neural Networks. In: Karras, D. A., Li, C., Majkic, Z. and Prasanna, S. R. M., eds. *Artificial Intelligence and Pattern Recognition*. ISRST. 411–417. 2007.

41. Szabo, G., Szabo, I. and Orincsay, D. Accurate traffic classification. *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007*. 2007. 1–8.

42. Wang, H., Fan, W., Yu, P. S. and Han, J. Mining concept-drifting data streams using ensemble classifiers. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, KDD '03. 226–235.

43. Zhang, G., Xie, G., Yang, J., Min, Y., Zhou, Z. and Duan, X. Accurate Online Traffic Classification with Multi-Phases Identification Methodology. 2008: 141 –146.

44. PACE, Protocol and Application Classification Engine developed by Ipoque. URL `http://www.ipoque.com/products/pace-applicationclassification`.

45. L7-filter. Application Layer Packet Classifier. URL `http://l7-filter.sourceforge.net/`.

46. OpenDPI. http://www.opendpi.org/, 2012.

47. Bernaille, L., Teixeira, R., Akodkenou, I., Soule, A. and Salamatian, K. Traffic classification on the fly. *ACM SIGCOMM Computer Communication Review*, 2006. 36(2): 26.

48. Li, J., Zhang, S., Lu, Y. and Yan, J. Real-Time P2P Traffic Identification. *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008, New Orleans, LA, USA*. 2008. 1 –5.

49. Gu, C., Zhang, S. and Sun, Y. Realtime Encrypted Traffic Identification using Machine Learning. *Journal of Software*, 2011. 6(6).

50. Quinlan, J. R. *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1993. ISBN 1-55860-238-0.

51. Witten, I. H. and Frank, E. *DATA MINING: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers. 2005.

52. Tcpdump. http://www.tcpdump.org/, 2012.

53. Quinlan, J. R. http://www.rulequest.com/Personal/, 2012.

54. Wang, Y. and Yu, S.-Z. Machine Learned Real-Time Traffic Classifiers. 2008. 3: 449 –454. doi:10.1109/IITA.2008.536.