ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 210-216

Article Title: **Network Traffic Identification Based On Machine Learning and Deep Packet Inspection**

Network Traffic Identification Based On Machine Learning and Deep Packet Inspection

M. B. Anushlin Leena[1*], R. Jegana[2], P. Abitha Rose[3]

[1]Assistant Professor/ IT (anushlinit@gmail.com)
[2]Assistant Professor /IT (jeganaalvin1@gmail.com)
[3]Assistant Professor /IT (abi.mtech22@gmail.com)
[1, 2, 3] Department of Information Technology, Bethlahem Institute of Engineering, Karungal.

ABSTRACT

Accurate network traffic identification is an important basic for network traffic monitoring and data analysis and is the key to improve the quality of user service. In this project, through the analysis of two network traffic identification methods based on machine learning and deep packet inspection, a network traffic identification method based on machine learning and deep packet inspection is proposed. The deep packet inspection based on the feature library RuleLib, conducts in-depth analysis of data traffic through pattern matching and identifies specific application traffic. Machine learning method is used to assist in identifying network traffic with encryption and unknown features, which makes up for the disadvantage of deep packet inspection that cannot identify new application and encrypted traffic. Experiments show that this method can improve the identification rate of network traffic.

Keywords: Deep Packet Inspection(DPI), Transmission Control Protocol(TCP),User Datagram Protocol(UDP), File Transfer Protocol(FTP), Internet Protocol(IP), Hyper Text Transfer Protocol(HTTP), The Internet Assigned Number Authority(IANA), Peer to Peer(P2P).

1 Introduction

The rapid development of network technology, network users are demanding higher and higher speed and quality of network services. Therefore, it has become one of the challenges in the field of network operation and maintenance management to manage and control various network business traffic through effective technical means, distinguish different services, provide different quality assurance, and meet users business means network traffic identification provides an effective technical means to distinguish traffic of different application .By classifying, identifying and differentiating the applications of network traffic, the traffic of different applications can be subdivided to provide users with personalized network services and improve the network service quality and user satisfaction.

Network traffic identification refers to the identification of bidirectional TCP or UDP flowsgenerated by network communication according to the types of network applications in the internet based on TCP/IP protocol [1]. At present, there are three commonly used methods:

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 210-216

Article Title: Network Traffic Identification Based On Machine Learning and Deep Packet Inspection

Identification based on port matching; Identification based on deep packet inspection (DPI); Identification based on

Firstly, the port-based traffic identification method does not need complicated calculation and analysis, and its implementation principle is simple. It can meet the requirements of fast identification of high-speed network. However, due to the development of new network applications, especially the emergence of P2P applications, most of them use random ports and camouflage ports to protect their network communications, which reduces the accuracy of port-based traffic identification method, which has been gradually eliminated by history.

The identification method based on feature field is able to detect the effective field in the load without relying on the port settings of the applications. It can well identify network flows and specific network application, and the detection accuracy is high. This method can detect network traffic quickly by only detecting the first few specific packets of network traffic. However, because this method depends on the feature field of the application protocol, it can only recognize known applications and cannot recognize new applications. In addition, this method cannot identify the network traffic of load encryption.

The machine learning identification method based on the flow statistics features uses the identification technology in data mining to realize traffic identification through the machine learning method, which overcomes the difficulties that cannot be solved by the first two methods, is free from the influence of port changes and protocol feature changes and can identify new applications. But, this kind of method based on machine learning in both Bayesian identification based on SVM (support vector machine) identification method, cannot identify specific application, need, depending on the type of multiple packet flow to detect traffic detection relatively lags behind, and easily affected by flow length, with less than a certain long flow the misdiagnosis rate is high. In addition, the accuracy of this identification method is easily affected by dynamic network changes and traffic attribute set, and the disadvantage of this kind of method is that it is computationally intensive and not suitable for real-time traffic identification of high-speed network.

Based on the analysis and comparison of the above traffic identification method, a network traffic identification method based on machine learning and DPI technology is proposed according to the principle of feature field based identification method and flow statistics based machine learning method.

2 Related Work

The existing systems the DPI technology can identify specific application trafficand improves the accuracy of identification. But, the DPI technology that cannot identify new applications and encrypted traffic.

Lightweight Source Authentication and Path Validation Introduces (1) DRKeys as efficient and dynamically recreatable key setup protocols, and (2) OPT as an extremely lightweight,

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 210-216

Article Title: Network Traffic Identification Based On Machine Learning and Deep Packet Inspection

scalable, and secure protocol that provides source authentication and path validation. OPT achieves performance improvements.

Flexible Deterministic Packet Marking: An IP Traceback System to Find the Real Source of Attacks, FDPM is suitable for not only tracing sources of DDoS attacks but also DDoS detection. The main characteristic of DDoS is to use multiple attacking sources to attack a single victim (the aggregation characteristic).

Traceback of DDoS Attacks Using Entropy Variations, Proposed an effective and efficient IP traceback scheme against DDoS attacks based on entropy variations. It is a fundamentally different traceback mechanism from the currently adopted packet marking strategies.

3 Proposed System

The proposed system in DPI technology is according to the principle of feature field based identification method and flow statistics based machine learning method. Machine learning method is used to assist in identifying network traffic with encrypted and unknown features.

3.1 Algorithm Design

The network traffic identification method adopts the combination of machine learning and DPI technology to realize network traffic identification.

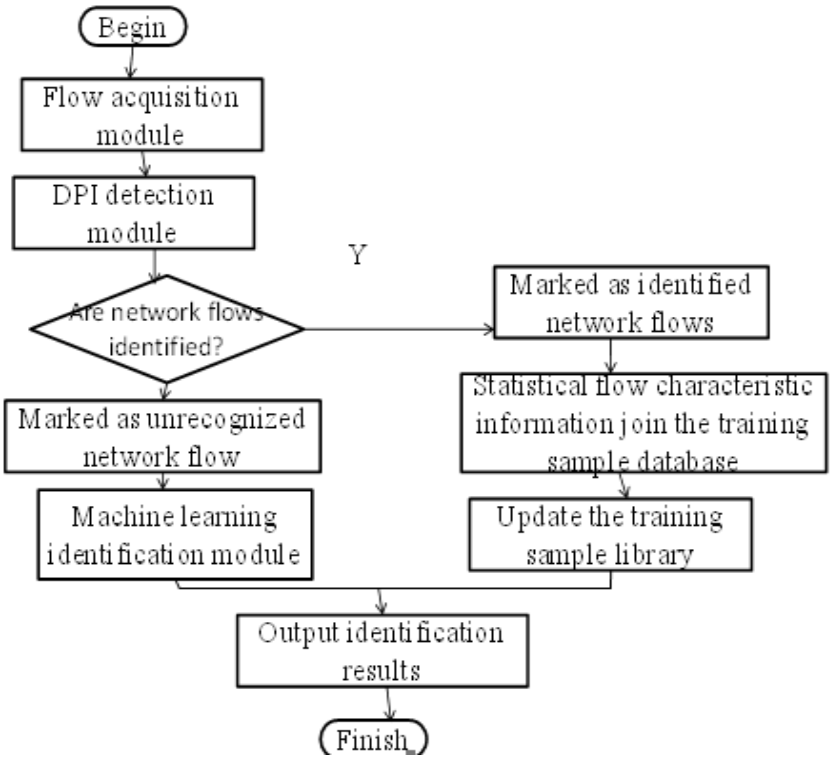

Figure1: *Network traffic identification process based on DPI and machine learning*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 210-216

Article Title: Network Traffic Identification Based On Machine Learning and Deep Packet Inspection

In the DPI technology stage, pattern matching detection is carried out for network data flow according to the protocol feature library loaded. If the corresponding protocol field can be matched, the traffic will be identified. Otherwise, no identification will be made.Afterthe network flow is identified by DPI, the flow statistical feature acquisition module sets a fixed collection time and begins to collect the feature information of the message.When the feature acquisition module completes the acquisition, the statistical feature information of unrecognized network traffic is sent to the trained traffic classifier for identification.For the network flow identification by DPI, its flow statistical characteristic information is added to the training sample library, which is used as the training sample set to relearn the classifier.

3.2 Machine Learning and Dpi Technical Identification Methods

3.2.1 Network Formation [Routing Information Protocol (RIP)

The Routing Information Protocol (RIP) is one of a family of IP Routing protocols, and is an Interior Gateway Protocol (IGP) designed to distribute routing information within an Autonomous System (AS).RIP is a simple vector routing protocol with many existing implementations in the field. In a vector routing protocol, the routers exchange network reachability information with their nearest neighbours. In other words, the routers communicate to each other the sets of destinations ("address prefixes") that they can reach, and the next hop address to which data should be sent in order to reach those destinations. This contrasts with link-state IGPs; vectoring protocols exchange routes with one another, whereas link state routers exchange topology information, and calculate their own routes locally.

3.2.2 Deep Packet Inspection

The DPI module mainly based on the feature library RuleLib, conducts in-depth analysis of data traffic through pattern matching and identifies specific application traffic. The feature library stores the features of various protocol in the form of XML files. The detection flow table stores the detected data flow according to the quintuple information of the flow. When the quintuple information of the subsequent flow is the same as the existing flow information in the detection flow table, it can be directly determined as the same application flow.

DPI technology generally consists of two parts, one is scanning algorithm, the other is feature library. The scanning algorithm is to match the content and feature library of IP packet load word by word. The string matching algorithm commonly used in DPI technology. DPI detection is similar to feature matching in anti-virus software. The antivirus software matches the scanned current file to its own virus library word by word. If the same characteristic code is found, the type and name of the virus will be determined.

This approach can accurately identify network flows based on the feature library, and can be accurate to the specific application to which the network flows belong, with high detection accuracy. But DPI is unable to identify application traffic that has not yet been recorded in the

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 210-216

Article Title: Network Traffic Identification Based On Machine Learning and Deep Packet Inspection

feature library, lags behind the release of new applications, and does not recognize encrypted network data streams.

3.2.3 Machine Learning

The core of the network traffic identification method based on machine learning is that computer program can constantly improve their performance with the accumulation of learning experience, so as to complete tasks that cannot be completed by conventional methods. In network traffic identification, this kind of prior knowledge can be different characteristics of network traffic and supervisory information of people. Selecting appropriate machine learning algorithm can make good use of prior knowledge to complete traffic identification.

The flow of network traffic identification method based on machine learning is shown in Fig.2. Firstly, the training data set is used to train the identification model, and then the recognizer is established according to the training model. After the recognizer is established, the identification of traffic can be realized.

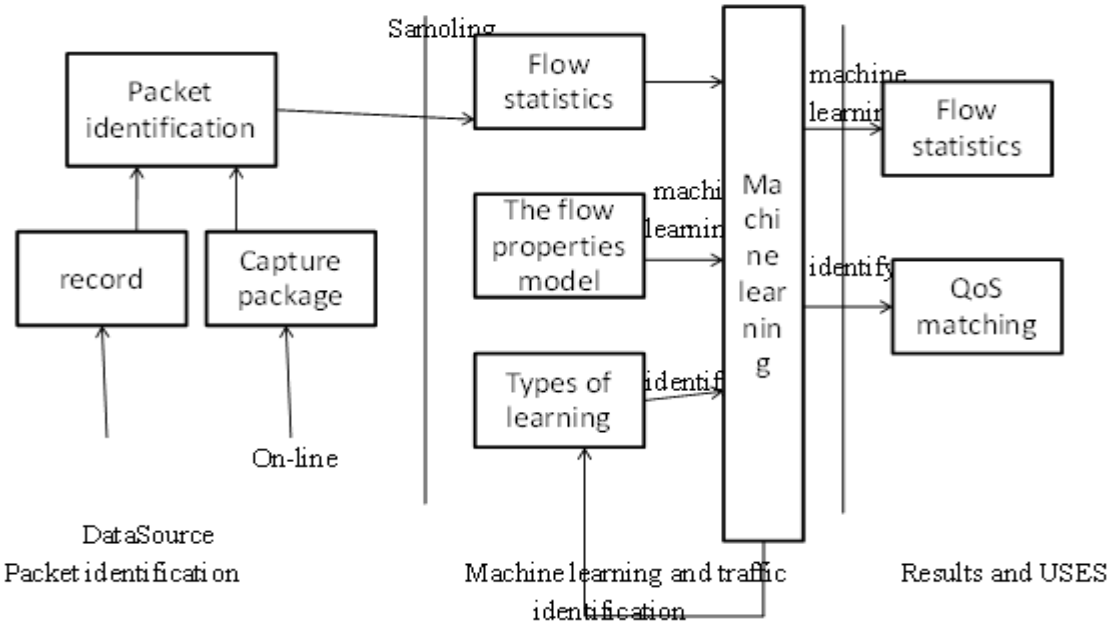

Figure 2: *Flow identification flow chart based on machine learning*

The naive bayes identification method is to determine the category of the sample by calculating the posterior probability. Its basic ideas is based on bayes formula and conditional independence assumption in probability theory, and the combined probability of attribute and category is used to estimate the category of the new sample. In the identification module of machine learning, Naive Bayes identification method is used to classify network traffic.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 210-216

Article Title: Network Traffic Identification Based On Machine Learning and Deep Packet Inspection

Preparation Stage

Feature selection is to get the most effective feature combination on the premise of ensuring the original value of the application data. Feature attributes selection adopt FCBF algorithm to select 8 feature attributes of network flow as feature set.

Training Stage of Classifier

The naive bayes method belong to the supervised identification method, which preprocesses the training samples by offline training method. In this experiment. Java is used to analyze the sample data set and calculate the prior probability and conditional probability of the corresponding network traffic application type.

Identification Stage

In the identification phase, the classifier is generated according to the decision model of the training phase to realize the online identification of network traffic.

3.3 Statistical Report

The machine learning method based on the statistical characteristics of flow is used to assist the identification of network flows with encryption and unknown features, which makes up for the shortcomings of DPI technology in identifying new application and encrypted traffic and improves the identification rate of network traffic.

3.4 Algorithm Implementation

Implementation is the final and important phase. It is the phase where theoretical design is turned into working system, which works for the user in the most effective manner. It involves careful planning, investigation of the present system and the constraints involved, user training, system testing and successful running of developed proposed system.
The implementation process begins with preparing a plan for the implementation of the system. According to this plan the activities are to be carried out, discussions made regarding the equipment and resources and the additional equipment has to be acquired to implement the new system. The user tests the developed system and changes are made according to their needs. The testing phase involves the testing of a system using various kinds of data. This method also offers the greatest security since the old system can take over if the errors are found or inability to handle certain type of transactions while using the new system.

4 Conclusion

By analyzing two kinds of network traffic identification methods based on feature field and flow statistic, a network traffic identification method based on machine learning and DPI technology is proposed. This method uses DPI technology to identify most network traffic, reduces the future workload that needs to be identified by machine learning method, and DPI

technology can identify specific application traffic, and improves the accuracy of identification. The machine learning method based on the statistical characteristics of flow is used to assist the identification of network flows with encryption and unknown features, which makes up for the shortcomings of DPI technology in identifying new application and encrypted traffic, and improves the identification rate of network traffic.

References

1. He Deng, Year: 2009, "Research on Network traffic identification Based on Machine Learning Method", Zhuzhou: Human University of Technology.

2. Hongboshi, Year: 2005, "Research on Bias identification", Beijing: China Science and Technology press, pp. 3 – 4.

3. Jianguo Wang; Wenxing Zhang, Year: 2015, "Support vector machine modeling and intelligent optimization", Beijing: Tsinghua University press, pp. 25 – 30.

4. Xiuxia Huang; Li Sun, Year: 2016, "Optimization of C4.5 algorithm", Computer engineering and design, pp. 1 – 3.

5. Ting Hu; Yong Wang Xiaoling Tao, Year: 2010, "A comparative study of network traffic identification methods", Journal of Guilin University of Electronic Science and Technology, pp. 216 – 219.

6. S. Sen; O. Spatscheck; D. Wang, Year: 2004, "Accurate, scalable internet work identification of P2P traffic using application signature", Proceedings of the 13th International Conference on World Wide Web.

7. Lihua Zhou, Year: 2013, "Implementation of network intrusion detection system based on SBOM algorithm", Intelligent computers and Applications, pp. 90 – 92.

8. Yang Ming-Hour; Ming-Chien Yang, Year: 2012, "RIHT: A novel hybrid IP traceback scheme", IEEE Trans. On Info. Forensics and Security, Vol: 7, no: 2, pp. 789-797.

9. Tiffany Hyun-Jin kim; Cristina Basescu; Limin Jia; Soo Bum Lee; Yih-Chun Hu; Adrian Perrig, Year: 2014, "Lightweight source authentication and path validation", in Proc. SIGCOMM, pp. 271 – 282.

10. K. Park, H. Lee, Year: 2001, in: IEEE INFOCOM 2001. Twentieth Annual Join Conference of the IEEE Computer and Communications Societies. Proceedings, Vol: 338.