A network traffic classification method based on quintuple feature and regular expression matching

Shujuan Qiao[1, a], Yi Zhou[2, b] , Leiyue Zhou[3, c] and Liancheng Zheng[3, d]

[1] China Power Information Technology of Beijing, Beijing 100085, China;

[2] Hubei Electric Power Company, Hubei 430077, China.

[3] School of control and Computer engineering, North China Electric Power University, Beijing 102206, China.

[a] qiaoshujuan@sgitg.sgcc.com.cn, [b]12940237@qq.com, [c]zlyddygyx1102@163.com, [d]865048915@qq.com

Keywords: traffic classification, quintuple, regular expression.

Abstract. A traffic classification method, based on quintuple feature and regular expression matching, is presented for the requirement of instantaneity and accuracy in the process of network traffic classification. First of all, the method proposed uses quintuple information recognition technology to classify network traffic rapidly. Then, the data packets are deeply, accurately classified and identified by the characteristics of regular expression matching protocol. Proved by examples, the method adopted in this paper can effectively improve the instantaneity and accuracy of network traffic classification.

1. Introduction

In recent years, with the rapid development of Internet, more and more new network applications are emerging. The network scale expands unceasingly and the network environment is becoming more and more complex, which makes the analysis, monitoring and management of network traffic become more complex and difficult. The correct identification and analysis of network traffic characteristics are the prerequisite of our in-depth understanding of the network status, user behavior and the status of the Internet. It is of great importance whether it is for network administrators or users and service providers.

Network traffic classification technology is one of the basic means to analyze the characteristics of network traffic and enhance the network controllability which is widely used in traffic monitoring, network security detection, user behavior analysis, and other network activities. Literature [1] analyzes the requirement of deep packet inspection, describes the workflow of the deep packet inspection engine, and introduces the method of regular expression pattern matching, which has a high value to strengthen the detection capability of the security gateway. Literature [7] summarizes the principles and content of network traffic monitoring, as well as network traffic monitoring technology, but there is no reasonable and effective monitoring techniques proposed. Literature [2] uses machine learning algorithms to classify convection, which consumes more resources, needs prior knowledge of data samples, and has certain difficulty to apply to online environment .

This paper aims at the requirements of instantaneity and accuracy in the process of network traffic classification. A method based on quintuple feature and regular expression matching method is proposed, considering the accuracy of the traditional traffic classification method, the application of the scene and the applicability of high-speed network traffic classification when used independently. Using quintuple information and traffic flow table for primary classification of flow, and further using regular expression matching method to identify application layer protocols for unknown traffic, which can classify the network traffic quickly and accurately.

2. Traffic Classification Method

In this paper, the method uses quintuple (source IP, destination IP, source port, destination port and protocol number) as the characteristic to carry on the primary classification to the traffic, in order to identify such traffic according to common features of similar traffic. Extracting quintuple information at first. Then, calculating the hash index value by using quintuple information with certain hash algorithm. Finally viewing the flow table according to the index value. According to the results of the view, if the flow is known which means the flow is identified by the protocol type, then mark the flow, and there's no need to detect the subsequent data packet. Using regular expression matching method to accurately identify the application layer protocol for those unrecognized traffic based on quintuple feature . The regular expression engine has multiple versions, this paper uses Henry Spencer's V8-routines(a subset of Perl compatible regular expressions). V8-routines regular expression is not case sensitive, and can't recognize the value of NULL (0). Since the network packets are processed as a common string , so the network packet "\ x00" such zero value is ignored, because "00" represents that there is no, and cannot match null characters in the data packets. Therefore, it should be pretreated to remove the "\ x00" value before the matching. This paper considers the powerful function of regular expression in string processing, which greatly simplifies the complexity of string manipulation.

3. Implementation of traffic classification

The method in this paper classify network traffic by traffic classification method based on the quintuple feature at first. Then, the unrecognized traffic will be submitted to application layer recognition processing program through packets handled by the network layer protocol, namely use the regular expression matching method to identify application layer protocols.

3.1 Extract quintuple information.

Data packets captured from the network equipment are decapsulated according to the TCP/IP four layer structure (link layer, network layer, transport layer and application layer), and then analyze the type of each layer protocol to get quintuple information.

3.1.1 Analysis of the transport layer protocol

This paper focuses on Ethernet, and only analyze Ethernet protocol due to the various type of link layer protocol. Network layer protocol type is stored in the Ethernet header, including the source MAC address, destination MAC address, type field, data and so on. IP packets parsed out is transmitted to the data packet processing function for Further analysis, this function will analyze the transport layer protocol type according to the protocol fields in IP data packet header structure shown in Fig. 1, and records the source IP address and destination IP address. When the value of the protocol field is 6, the transport layer protocol is TCP, and the transport layer protocol is UDP when the protocol field is 17.

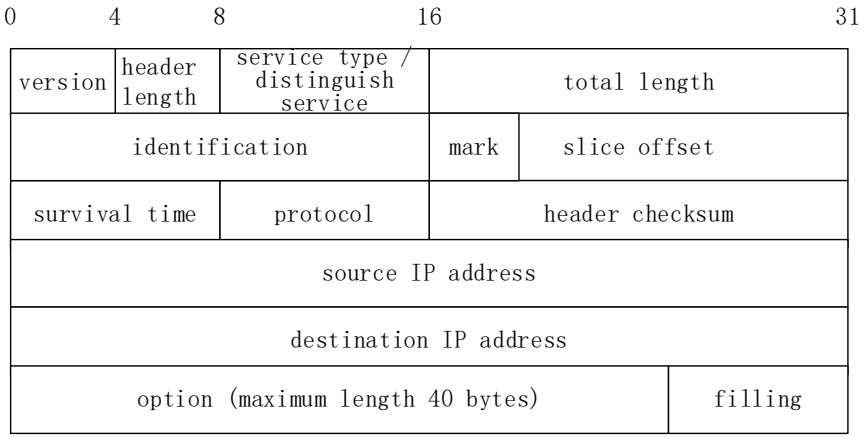

Fig. 1 IP packet header structure

3.1.2 Extract the source port number and destination port number

Extracting the TCP or UDP protocol source port number and destination port number. The structure of the TCP data packet is shown in Fig. 2, parsing the TCP packets according to the structure of TCP packets, and extracting the source port and destination port. For UDP packets, similarly, extracting the required information according to the structure of Fig. 3.

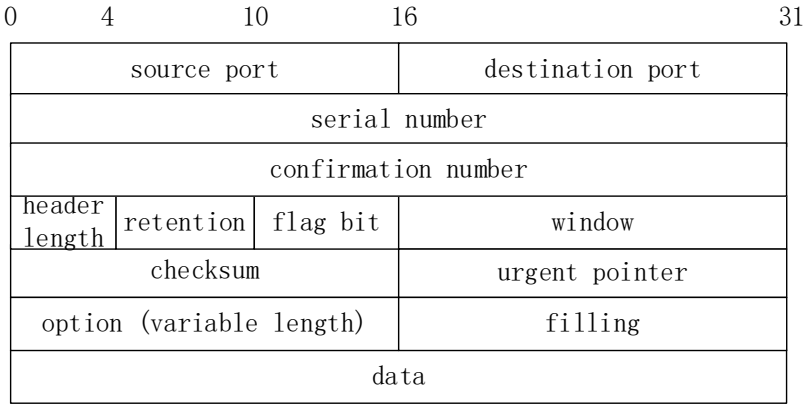

Fig. 2 TCP packet structure

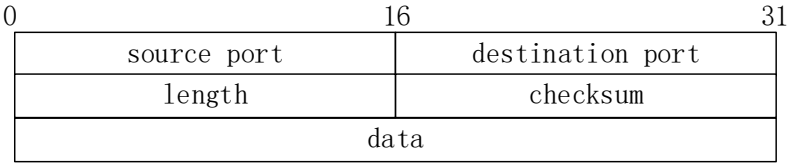

Fig. 3 UDP packet structure

At this point, quintuple information is formed according to the extracted information.

3.2 View flow table

Calculating the hash index value according to quintuple information and certain hash algorithm, and checking the flow table to judge whether this flow is known. Marking the flow if it is known, and there is no need for checking subsequent data packets.

The design of flow table uses a certain hash algorithm, collects quintuple information, and makes the flow with the same hash value link to the same node, as shown in Fig. 4. The situation may appear that different quintuple compute the same hash value due to the hash algorithm itself, which just need special processing when operating the flow table.

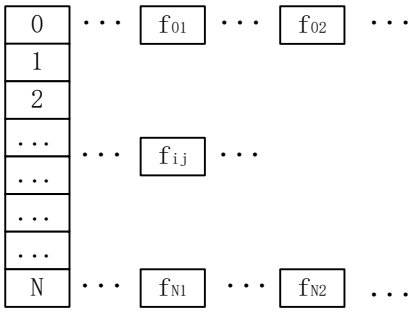

Fig. 4 Overall structure of flow table

3.3 Regular expression matching

The V8-routines regular expression matching restrictions should be considered in the realization of the data packets of regular expression matching. Firstly, the regular expression compilation should be carried on. Using regcomp() function to compile it into the regex_t structure before the comparison of a string and regular expression in order to improve efficiency. Secondly, the kernel of the regular expression function library is not case sensitive, so it's necessary to sort the data packet application layer load at first, remove the string terminator '\ 0' and do case conversion. Finally call regexec () function to complete pattern matching.

The overall flow chart based on quintuple feature flow classification method and the regular expression matching method is shown in Fig. 5 below.

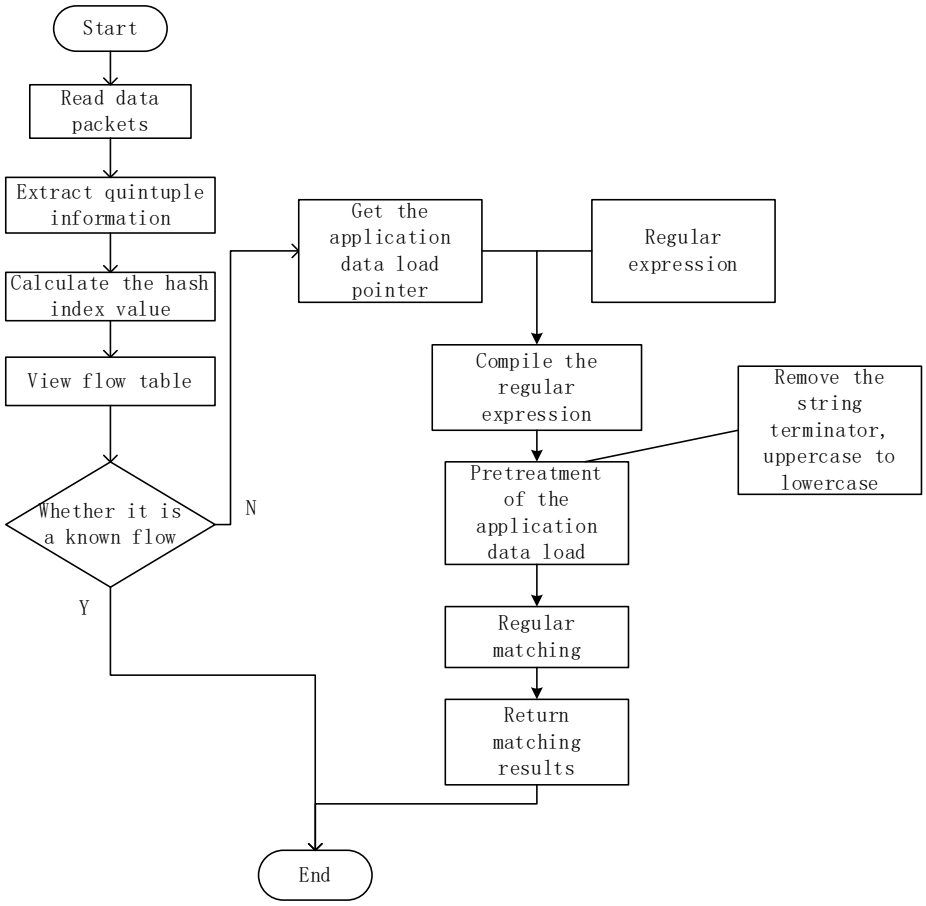

Fig. 5 Overall process

4. Case analysis

4.1 Test environment

Experimental environment is shown in Fig. 6. Traffic sending server sends traffic by using Tcpreplay tools read traffic file, and makes the flow to the gigabit switches. 24 port switch can mirror the multiple gigabit traffic onto a network port. The method proposed in this paper monitors the mirror port to analyze traffic, and outputs flow identification result. The flow used in the experiment is captured in the export of a university laboratory network, which contains a representative Internet traffic and covers a wide range of protocols in the hundreds of networks. The size of the traffic is 20G, containing 2561500 data flows. In the test, the public DARPA traffic is also used as a measure of recognition rate and accuracy.

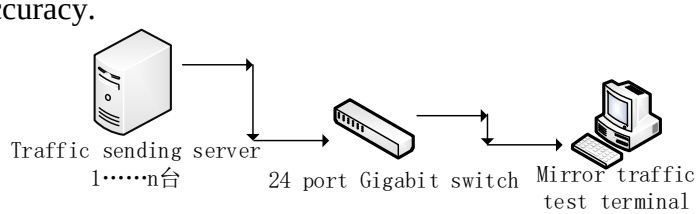

Fig. 6 Test environment

4.2 Results and analysis

Under the same experimental conditions, the recognition effect on the application layer protocol of this method is analyzed by comparing the recognition results of this method with the flow detection tools OpenDPI system. Using the public traffic DARPA in 1999 to continue testing the recognition accuracy of this method, and explain the advantages of recognition results. DARPA traffic totaled 347,987 packets, which contain 79179 TCP connections , 6565 HTTP protocols, 20 NBNS protocol connections , 4 SNMP protocol connections , 292 SMTP protocol connections, 37 POP protocol connections ,202 NTP protocols , 138 FTP protocol connections ,21 Finger protocol connections ,19769 DNS protocol connections and 90 TELNET protocols . This method and

OpenDPI system are both used to process the DARPA traffic file in order to carry out the contrast experiment, and get Table 1 and Table 2 through recording protocol identification results .

Table 1 Comparison of the number of identified packages using the proposed method and the OpenDPI system

Protocol Type	Using the proposed method	OpenDPI
HTTP	653428	622551
NBNS	10031	10031
FTP	687932	602023
SMTP	110032	96437
POP	275	275
SNMP	3183	3183
MPEG	48544	-
DNS	15378	15378
Flash	90181	-
NETBIOS	130	130
QQ	82459	82313
Windowsmedia	776	-
MMS	1026	-
ICMP	126544	-
MSN	9975	8060
Telnet	12302	12302
GRE	396270	396270
GNUTELLA	3634	-
SSH	106593	101698

As can be seen from table 1, using the proposed method to identify the number of packets per protocol in DARPA traffic is not less than the OpenDPI system, and can identify the protocol that OpenDPI cannot identify.

Table 2 Identification results using the proposed method and OpenDPI system

Protocol Type	Inherent connection number	Using the proposed method	OpenDPI system
HTTP	6565	6565	6500
NBNS	20	20	18
SNMP	4	4	4
SMTP	292	287	254
POP	37	34	30
NTP	202	196	188
FTP	138	130	130
Finger	21	21	20
DNS	19769	19547	18431
TELNET	90	90	80

As can be seen from Table 2, the use of the proposed method has achieved good results for identifying each protocol in DARPA traffic, and the average accuracy rate reached 95% or more.

5. Summary

The proposed method in this paper combines the traffic classification method based on quintuple feature with the regular expression matching method. It can make preliminary and rapid classification, and do accurately identify for unknown traffic flow which preliminary classification unrecognized. Through the experiment test, it has realized the fast and accurate classification of network traffic to a certain extent. With the development of encryption technology, encrypted traffic proportion in the

entire network traffic is getting higher and higher, it should be further studied that identifying the encryption protocol and fining classification of encrypted traffic in the future.

Acknowledgment

This paper is funded by the project of The State Grid Corporation of China in 2014 “Research and development on the information safety threat analysis technology during the network access process”

References

[1]. LaiJianhua, Research on security gateway deep packet inspection engine. vol. 1(2010), p.98-100.

[2]. WangYixuan:Network traffic classification method research and traffic characteristics analysis(Master degree, Beijing University of Posts and Telecommunications,China 2013).

[3]. Wangqian, Design and implementation of P2P traffic controller based on deep packet inspection ,May.2013.

[4]. Yanke,Chenjian,Zhengyue, et al. Research on real time similarity matching algorithm in high speed Ethernet, vol. 3(2014), p.739-742.

[5]. Xiawei:Research and Implementation of Protocol Identification and Traffic Classification on Application Layer(Master degree, North China Electric Power University,China, 2014).

[6]. DengKaiyuan,Jianglei,Performance analysis of different regular expressions matching engine ,vol.7(2011) p.105-107.

[7]. JuGengui,LiYanyan, Summary of network traffic monitoring technology and its application .vol. 7(2011), p.86-88.

[8]. Yanhao:Analysis of user behavior based on the network traffic monitoring(Doctoral degree,Beijing University of Posts and Telecommunications,China, 2011).