Application Traffic Classification in Hadoop Distributed Computing Environment

Kyu-Seok Shim, Su-Kang Lee and Myung-Sup Kim
Dept. of Computer and Information Science
Korea University
Sejong, Korea
{kusuk007, sukanglee, tmskim}@korea.ac.kr

***Abstract*—Today, network traffic has increased because of the appearance of various applications and services. However, methods for network traffic analysis are not developed to catch up the trend of increasing usage of the network. Most methods for network traffic analysis are operated on a single server environment, which results in the limits about memory, processing speed, storage capacity. When considering the increment of network traffic, we need a method of network traffic to handle the Bigdata traffic. Hadoop system can be effectively used for analyzing Bigdata traffic. In this paper, we propose a method of application traffic classification in Hadoop distributed computing system and compare the processing time of the proposed system with a single server system to show the advantages of Hadoop.**

Keywords—Hadoop; Payload; Traffic; Distribute; Signature;

I. INTRODUCTION [1]

Today, network is becoming more complex and diverse because of the emergence of various applications and services. Therefore, network traffic is growing rapidly. But, methods for network traffic analysis are not developed to catch up the trend of increasing usage of network. Most methods for network traffic analysis are operated on a single server environment. But if the amount of traffic data is increased, the existing method has limit in terms of memory, processing speed, and storage capacity.

Application traffic classification is important parts in the fields of network traffic analysis. Among them, method for application traffic classification based on payload signature promises high classification accuracy rate. But there are disadvantages in payload signature based analysis that requires high processing load comparing to other analysis method. Therefore, as the traffic volume is increased the application traffic classification using payload signature suffers from the burden in memory, processing speed, storage space.

In this paper, we provide an application traffic classification based on payload signature in Hadoop distributed computing environment. Application traffic classification based on payload signature is a method to determine the application level identity of traffic data by comparing the payload that can be known from the packet information with pre-structured signatures. In this paper, we propose an application traffic classification method in Hadoop environment.

The structure of this paper is as follows. First, we describe the related research and prove the need for this study in Section II. In section III, we describe the structure of the proposed application traffic classification system in Hadoop distributed computing environment. In section IV, we prove the feasibility of the proposed system through experimental work conducted in campus Internet traffic. Finally, in section V, we describe the conclusion and future work.

II. RELATED WORK

In the Internet environment of today there are various applications, so application traffic classification is important issue. In the past, the applications that accounted for most Internet traffic such as HTTP, telnet, e-mail, FTP and SMTP have port numbers under 1024. So, traffic classification based on the port information from the IANA definition was enough to get results of high reliability and accuracy. However, in the network environment of today, data session port number is dynamically generated such as streaming program. Therefore, analysis based on the port cannot guarantee high reliability like before.

To compensate for this problem in the application classification of Internet traffic, several methods have been proposed. Conventional methods can be divided into three. They are the signature-based classification method, the traffic correlation-based classification method, and the machine-learning based classification method.

The traffic correlation based classification method uses relational information among traffic flows and uses this information to classify traffic. It uses features such as address system (IP address, port number, protocol), occurrence time and occurrence form of the traffic.

The machine-learning based classification method uses classification and clustering techniques of machine-learning to classify the traffic. It uses items that can be features of Internet traffic (port number, flow duration, inter-packet arrival time, packet size, etc.).

[1] This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (2010-0020728), Brain Korea 21 Plus (BK21+), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2007483).

The signature based classification method analyze the traffic that is generated in a particular application for extracting features called signature that can distinguish them from other applications, and it classifies the traffic by using the signatures. This method show the result with high accuracy, but processing speed and system load exist disadvantage in the process of analysis[2][3].

This paper suggests a new way of traffic classification based on signature method in Hadoop distributed system to improve the system load and processing speed. There are existing studies for reducing the processing speed. However, most studies focus on the improvements of algorithms. Therefore, this paper suggests application-level classification based on signature based method in Hadoop distributed system and apply the campus network traffic, to prove its validity.

Hadoop is a platform that supports both distributed storage and distributed computing capabilities[1][5][7]. Distributed computing capabilities supported by MapReduce and distributed storage supported by HDFS (Hadoop Distributed File System). After the papers about distributed file system by Google was published, Hadoop have been developed as a system for MapReduce corresponding to the structure.

MapReduce is a key concept in Hadoop. Operation procedure of MapReduce is as follows. Input files are divided into multiple pieces depending on size and distributed among multiple cooperating storage nodes. When data is input to HDFS, it is replicated to the three. When a slave node becomes faulty, the loss of data prevented by the same replicated files. Multiple maps conduct distributed computing function and reduce combine distributed data from multiple maps.

Several maps run the same distributed computing function. One reduce run the merge function for distributed data. Therefore, the role of MapReduce is very important in a distributed processing in Hadoop.

III. Application Traffic Analysis based on Hadoop

In this section, we describe the structure of the application traffic classification based on payload signature in Hadoop distributed computing environment. The system takes flow information as input and reveals the amount of flows, packets, byte of traffic as output . First, we refer to the process of collecting traffic data and the construct of the input file to Hadoop system. Next, we suggest method of application traffic classification in Hadoop distributed computing environment. The overall process of this system is a Fig 1.

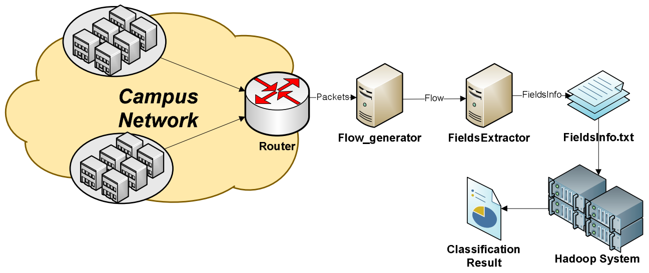


Fig 1. Application Traffic Analysis System based on Hadoop

A. Traffic Acquisition

We collect the Packet units of traffic that occurs on campus network. Collected Packets are converted into Flow format through the Flow generator. The Flow is defined as the same of 5-tuple (source IP, sour port, protocol number, destination IP, destination port) set of Packets. Flow is composed of the 5-tuple, the amount of packets, bytes and payload of each packet.

B. Implementation of Experimental Environment

Text file is input to the Hadoop system. Also, text file is output by three conditions field extractor using the Flow file.

- First, we analyze only HTTP traffic. Because encrypted non-HTTP traffic is limit to analyze the payload.
- Second, we ignore continuing flow. Flow file is generated every 1 minute. For example, when the traffic occurs during the continuing 3 minutes by a particular application, flow is generated 1 minute of the first occurring. After flow is generated for the same application, it will have the same meaning as in the flows that were created before. Therefore, we analyze the first flow only and ignore the continuing flow.

  Third, we use only the first payload packet that is first request packet of a flow. Because it is possible to grasp the information of a specific application in the first request packet of a flow.

C. Configure the Input File

By the three conditions, flows are printed in a text file by FieldsExtractor. Text file is constructed several Flow information. The one Flow's information is constructed with "sourceIP, sourcePort, protocol number, destinationIP, destinationPort, First Request Packet Payload, the number of packets, the number of bytes".

D. Application Classification using MapReduce

Procedure for application classification using Map and Reduce of the Hadoop system are as follows using the text flow file as input.

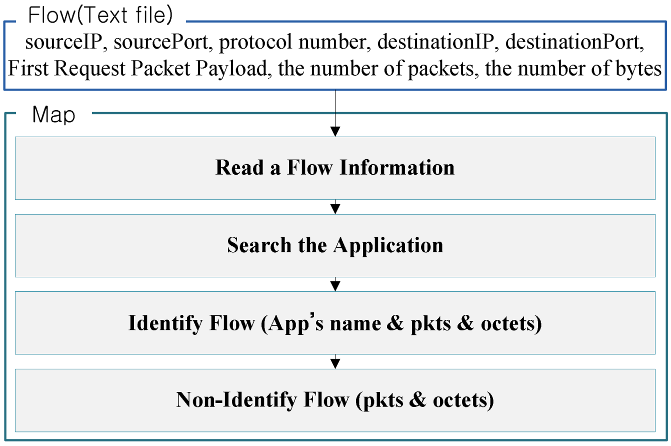

Flow(Text file)
sourceIP, sourcePort, protocol number, destinationIP, destinationPort, First Request Packet Payload, the number of packets, the number of bytes
Map
Read a Flow Information
Search the Application
Identify Flow (App's name & pkts & octets)
Non-Identify Flow (pkts & octets)

Fig2. Process of the Map function.

Procedure to the application traffic classification by using the text file and MapReduce is as follows. First, one flow of information is input to the map. The one flow has information of eight. And we look for the payload. We compare the payload of the flow and signatures of applications. If the payload and signature matches, then it returns the name of the application, and outputs the name of the application in the Map. For example, flow of the Naver application contains a Key: "s-portal-naver" in its first request packet, the map output a Value: "1". If map process complete its analyzing the all flow information, reduce is performed. The execution procedure of reduce is as follows.

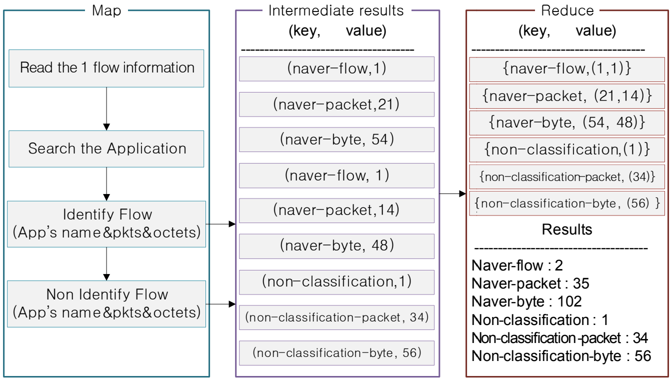
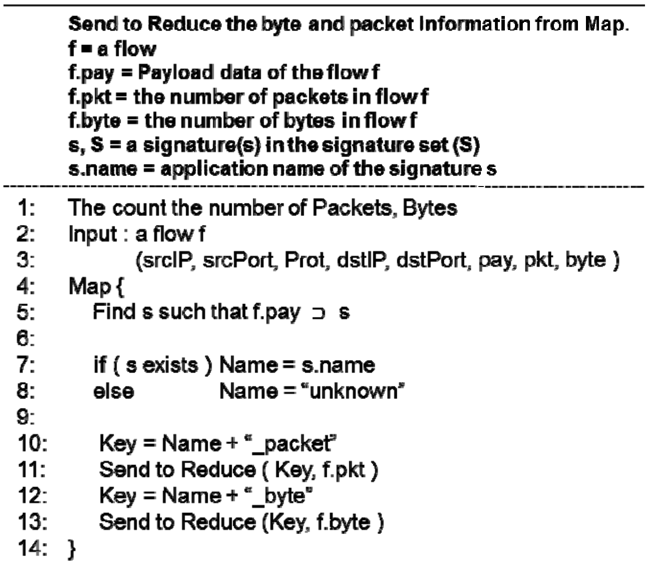


Fig3. Process of the MapReduce

Reduce add to Value of the same Key sent from the Map. Therefore, it is possible to know the number of flow of specific application. But, there is a difference to process that prints out between the number of packets, bytes and the number of flows. The output procedure of the number of packets, bytes for specific application is as in Fig4.



```
Send to Reduce the byte and packet Information from Map.
f = a flow
f.pay = Payload data of the flow f
f.pkt = the number of packets in flow f
f.byte = the number of bytes in flow f
s, S = a signature(s) in the signature set (S)
s.name = application name of the signature s
-----------------------------------------------------------------
The count the number of Packets, Bytes
Input : a flow f
        (srcIP, srcPort, Prot, dstIP, dstPort, pay, pkt, byte )
Map {
   Find s such that f.pay ⊃ s

   if ( s exists ) Name = s.name
   else            Name = "unknown"

    Key = Name + "_packet"
    Send to Reduce ( Key, f.pkt )
    Key = Name + "_byte"
    Send to Reduce (Key, f.byte )
}
```

Fig4. Algorithm of Application Detection Distriduted System in Map

Packet and byte of an application output algorithm are as follows. If the signature of the application is consistent with the payload of flow, after outputting the name of the application, it sends to reduce a Key: "Application' packet" Value: "the amount of flow's packet". For example, if the flow is occurred by Naver portal site, than it send to reduce Key: "s-portal-naver-packets", Value: "the amount of flow's packet". Therefore, when send to Reduce, Values that Key is "s-portal-naver-packets" are summed. Finally, it is possible to determine the sum of the Packet corresponding to Naver. The amount of byte is the same as when outputting the amount of transfer of Packet.

IV. EXPERIENCE AND RESULT

Hadoop is a platform which is supporting distributed data processing as I mentioned before. It is regarded as the essential system to be adapted in the present because the number of network traffic and data has been increasing. Thus, our lab has established Hadoop platform environment. There are six computers, one master node and five slave nodes. Each node consists of dual core, RAM 1G and HDD 200G.

In this chapter, we proved to potential through application classification result in Hadoop distributed computing environment. Application classification experiment did progress using defined signatures of naver, nate, daum, google, facebook and campus portal site. Furthermore, comparing the processing time of the traffic analysis system based on payload signature in distributed data processing environment with on a machine shows the advantage of Hadoop.

TABLE I. RESULT OF APPLICATION TRAFFIC CLASSIFICATION BASED ON PAYLOAD SIGNATURE ON HADOOP DISTRIBUTED ENVIRONMENT

Time	Type	Total	Classification	Non-Classification
1m	Flow	50	20	30
	Packet	942	230	712
	Byte(MB)	0.53	0.11	0.42
0.5h	Flow	3,638	1,003	2,635
	Packet(×1000)	157	37	120
	Byte(MB)	134.49	30.01	104.48
1h	Flow	5,989	1,852	3,837
	Packet(×1000)	334	57	277
	Byte(MB)	285	43.61	241.39
3h	Flow	9,966	3,750	6,216
	Packet(×1000)	657	118	539
	Byte(MB)	555.09	87.42	467.67
6h	Flow	13,985	5,144	8,841
	Packet(×1000)	1,145	212	933
	Byte(MB)	993.61	171.42	822.19
9h	Flow	26,993	7,250	19,743
	Packet(×1000)	1,691	298	1,393
	Byte(GB)	1.44	0.24	1.2
12h	Flow	82,736	28,505	54,231
	Packet(×1000)	6,068	1,289	4,779
	Byte(GB)	5.1	1.02	4.08
18h	Flow	252,683	71,861	180,822
	Packet(×1000)	15,268	3,235	12,033
	Byte(GB)	12.7	2.56	10.14
24h	Flow	331,610	94,852	236,758
	Packet(×1000)	20,094	4,376	15,718
	Byte(GB)	16.74	3.48	13.26

In the Hadoop distributed processing environment, the application classification based on payload signature proceeds by specifying the six applications. It is to analyze flow, packet and byte of total, classified and unclassified traffic in all HTTP traffics. And classified traffic is surveyed amount of flow, packet, and byte by specifying portal site.

Table 1 shows the experimental results of the application traffic classification based on payload signature in the Hadoop distributed processing environment. Time field in Table 1 is the collection time of traffic data. We collect traffic data with various time intervals, from 1 minute to 24 hours, to obtain the changes of processing time over different traffic size. The type field in Table 1 is the type of measurement: flow, packet, and byte. The total field is the amount of total traffic data in flow, packet, and byte that occurred during the traffic collection time. The classification field is the classified amount of traffic data in flow, packet, and byte. The Non-Classification field is the un-classified amount of traffic data in flow, packet, and byte. We applied payload signatures of 6 applications, which results in 30~40% of classification ratio over total traffic. Then, we will analyze the processing speed of a single-server system and the proposed Hadoop-based system.

Figure 5 shows the performance increase of the proposed Hadoop-based application classification system over a single server-based system. X-axis in Figure 5 is a traffic acquisition time and Y-axis is the processing time by seconds. As the traffic acquisition time is getting longer and longer, the traffic volume and flow data to be processed is increased. And Figure 5 shows that the processing speed is much more improved in the Hadoop-based system. Single server-based system is more efficient than the Hadoop-based system in small amount of traffic data because there is some overhead in MapReduce processes to organize the works among distributed nodes. However, the processing speed in the Hadoop-based system is faster since it analyzes traffic data longer than 3 hours. In addition, the processing time of the single server-based system is rapidly increased when the collection time of traffic data increase over 9 hours.

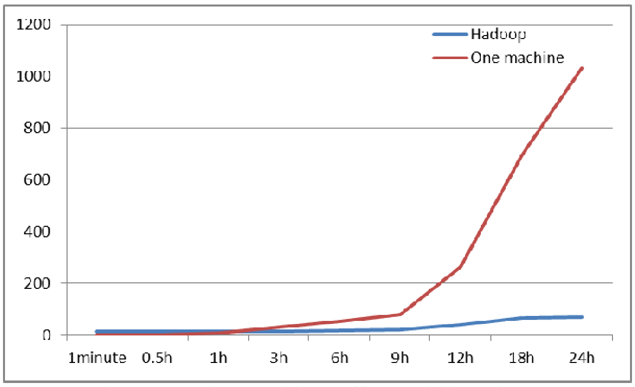


Fig5. Processing speed by Traffic acquisition Time.

However, there is some difference in the experimental environment of Hadoop-based system and a server-based system. The Hadoop-based system analyzes the Text files that have been extracted by FieldsExtractor module, but the server-based system analyzes the flow in the form of binary file. Therefore, it is difficult to expect an accurate comparison in the processing time. When considering the server-based system, the processing time has increased exponentially as the traffic acquisition time increases. However, in the Hadoop-based system, the processing time does not increased significantly as the acquisition time of traffic data increased. Therefore, we can expect that the proposed Hadoop-based system will give about the same performance improvement even though we use of the binary file as an input format to the Hadoop-based system.

V. Conclusion

In this paper, we proposed an application traffic classification system using payload signature in Hadoop distributed computing environment. Also we demonstrated the advantages of the proposed system in processing speed through a comparison between the Hadoop-based system and a single server system. In analyzing a small amount of traffic data, Hadoop-based system is not much effective. But in analyzing a large amount of traffic data, it showed big advantages in the processing speed, memory and storage capacity.

Our current prototype system has a couple of drawbacks in the perspective of the low analysis rate because of a few number of applications applied and the high processing speed because of the input traffic used in the form of text data. In future work, the system will be updated to handle traffic data in the form of binary format in the Hadoop distributed computing environment. Further, we are planning to research the many areas of network management with Hadoop base.

References

[1] Zhao-wen LIN, Yan MA "Research of Hadoop-based data Flow management system, Volume 18, Supplement 2, Dec 2011, pp164-168

[2] JS Park, SH Yoon, JW Park, HS Lee, SW Lee, MS Kim, "Research on the Performance Improvement of Application-Level Traffic Classification System using Payload Signature", KNOM Review, Vol.12 No.2, , Dec. 2009, pp.12-21.

[3] JS Park, JW Park, SH Yoon, HS Lee, MS Kim, "Performance Improvement of Traffic Classification System based on Payload Signature", Proc. of the 20th Joint Conference on Communications and Information (JCCI) 2010, Apr. 28-30, 2010, pp. 148.

[4] Lee Y, Kang W, Lee Y. A hadoop-based packet trace processing tool. In: Proceedings of the third international conference on traffic monitoring and analysis, TMA'11. Berlin, Heidelberg: Springer-Verlag; 2011. p. 51–63.

[5] Apache Hadoop. http://hadoop.apache.org/

[6] Fnag Yu, Zhifeng Chen, Yanlei Dino, T. V. Lakshman, Randy H. Katz, "Fast and memory Efficient Regular Expression Matching for Deep Packet Inspection" ANCS 2006, December , 2006, San jose, California USA.

[7] K. Shvachko, H. Huang, S. Radia, R. Chansler, The hadoop distributed file system, in: 26th IEEE (MSST2010) Symposium on Massive Storage Systems and Technologies, May, 2010.