The Research on the Analysis of the Redundant Data

Tan Shunhua and Chen Miao
School of Information Engineering，Southwest University of Science and Technology
Mianyang, SiChuan ,China
{Tanshunhua,chenmiao}@swust.edu.cn

ABSTRACT

The data of port-80 is an important part of network data, which is also one of the data services that we study accurately at this stage. Through mining data packets of port-80 deeply, mining characteristics of data flow horizontally, and vertically mining the data business and behavioral characteristic of high-level application protocol packets, the behavior of network users and network redundancy and other aspects of data transmission can be more comprehensive analyzed. As a result of completely holding the characteristics of network data at this stage, the study presents a good foundation for the development of network transfer technique.

Keywords: *Data mining; IP packets; URL; Redundant Data*

1. INTRODUCTION

For the network measurement which is the foundation of the applications, TCP/IP Protocol suites play an overwhelming role in inter-connected networks, a study based on IP packets is one of main research directions of network traffic. As the aggregation of IP packets, IP flows cannot express the characteristics of IP packets, but reflect the users' behaviour in higher layers. And so there are more and more studies focused on the influences in network payloads and performances which are given by protocols and users by IP flows analysis [1].

In the IP flow, data of port-80 is a significant part. Although it is not the biggest one in the proportion of data, it is a very important studying object in the area of researching users' behavior and identifying network data services, especially the study of mining data of port-80 is the hot topic in network data analysis at present.

2. ANALYSIS OF EXISTING WORK

Nowadays, according to the processing data source and method, the analysis of network data is divided into two categories: web log server-based analysis of user behaviour and network packet-based analysis of flow characteristics and identifying data business.

A. Analysis of network log

Web consists of semi-structured documents, includes text, hypertext mark-up, hyperlinks, and other important information, in recent years, numerous scholars use the features of the page itself to conduct classification research, for example, JIANG Baolin [2] holds the belief that the HTML (e.g. TITLE, Hn, B etc.) have different importance, so the information can be weighted when feature selection is in process. Chen Shengrong [3] used the sub-vectors of relative important classification of neighbour pages to amend the source pages classification results by comparing the similar neighbour Web pages. Ghani [4] tried to use part of a page and content of web pages which are linked to it to indicate that page, however, the result of classification is not as effective as using the method which only uses part of the page. Oh et al. [5] also combined part of page content and linked web pages, but they only chose a small part of the content that closer to the original page by text-based similarity, not introduced all the pages which they have linked.

TABLE 1: WEB LOG MAIN CONTENT

Field	Description
Date	Date and time – zone of request
Clien IP	Remote host IP or DNS entry
Username	Remote login name of the user
Byte	Bytes transferred sent or received
Server	Server name IP address and port
Request	URL query
Status	HTTP status code returned to the client
Service name	Requested service name
Time cost	Time taken for transaction to complete
Protocol and version	Used transfer protocol and its version
User agent	Service provider
Cookie	Cookie ID
Reference	Previous page

Although these methods have achieved some good results, all of their analysis objects are the web logs; the main contents of web log are shown in Table 1, which mainly reflects the information of application layer, and is closer to users' information which is also the main data source for the analysis of user behaviour. Admittedly the heat log information of network can be reflected, but it cannot reflect data quantity of the specific type, for example, through the log analysis, the subject of hot news can be obtained, but it cannot dig out the amount of data. As a result, we cannot analyze how many of such data transferring repeatedly in network, consuming network bandwidth and reducing network performance.

B. Analysis of IP data flow

Most of the ongoing research is focused on the flow classification algorithms that are based on flow statistical characteristics. According to different flow characteristics, the popular flow classification algorithm is divided into two categories: The first method only uses the statistical characteristics of all the packets collections that are contained by unspecific single stream. Researchers [6-7] are more representative; in addition, [8] concludes the features of these characteristics: the characteristic parameters of various corresponding business are linear indivisible, Thus it can only use the higher complexity pure Bayesian and Bayesian neural network classification algorithm to assort. The second one utilizes the statistical characteristics of one collection that are associated with unspecific streams to conduct classification, which makes fully use of social characteristics of different businesses, as [9] mentioned, the method analyzes the host behaviour of a stream from society, function and application three different levels, and classifies the network flow by the stream host behaviour.

These characteristics can effectively identify Peer-to-Peer (P2P) traffic in network flow classification, and improve the recognition rate of the traditional algorithms. These also enable the use of multinomial logistic regression algorithm to classify the network flow, and reduce the complexity of the traditional algorithms.

Admittedly, using this kind of analysis has achieved a better result in the field of flow characteristic analysis, network security, and classification of network data business etc.; however, it is not conductive to mine the data deeply for it neglecting to analyze the URL etc. application layer protocols of the data. In order to acquiring more valuable information, we must combine the two kinds of analysis technique together.

3. ANALYSIS OF DATA PACKETS

A. The description of packet captured system for intermediate node network data [11]

The network topology of network data collection test platform is depicted in Figure 1. The platform for testing data collection is set up in the centre network node of a university, which can get all the information in any school network nodes by the function of “network-centric CISCO6509” mirroring in the device as shown in Figure 1. As recently network services are focused on the application of TCP Protocol on port 80, the early stage of study on redundancy in network is based on the characters of TCP Protocol on port 80, and the reliable and efficiency of capture system is being analyzed.

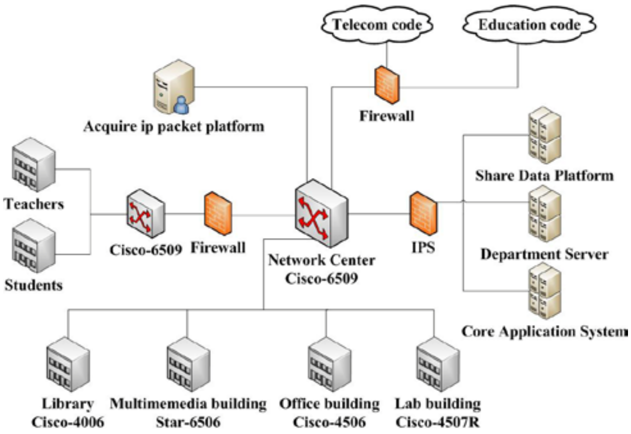

Figure 1: Network topology of network platform

Capture technology relies on WinPcap’s packet technology. It obtains the most complete and authentic data from the bottom of network, since its library works in the bottom of network analysis system module. Network packet captured system, a network communication procedure, implements network communication by programming on network card.

The main function of WinPcap is that it sends and receives original packets independently from the host agreements. In other words, WinPcap does not block, filter or control other application packages sending/receiving; it merely monitors the transmission data in the sharing network. If it needs to capture all the packets and frames which flow through the NIC, NIC must be set up promiscuous mode. That means it can monitor network communications real-time and get a copy from packet which is engaged without affecting any normal sending and receiving communications. It can monitor and capture network data without affecting the load of current network and the configuration of network node system by developing data collection module under the WinPcap. The data frames, captured by WinPcap, are the Ethernet frames which package through the transport layer, network layer and data link layer; hence the data frames can be further analyzed. IP packets captured process is shown in Figure 2.

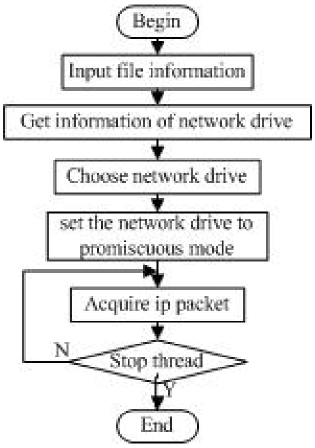

Figure 2: Acquire packet flow chart

B. Mining URL information[12]

To achieve the statistics of redundant traffic, the number of network resource and specific location (URL) must be analyzed first; and then calculate the amount of each resource; finally, conduct cluster analysis. We extracted packets information by processing captured data packets one by one. Relevant information of every packet was written in the database, till the end of files. In the MySQL database, all data was written into the table according to the source and destination IP, destination port, and URL which are grouped and cluster analyzed, all resources can be separately analyzed the number of visits and ranking number of visit, the size of same data resources, and the size of redundant data, ratio analysis and much more information to reflect user behaviours in networking. The detail of processing is shown in Figure 3:

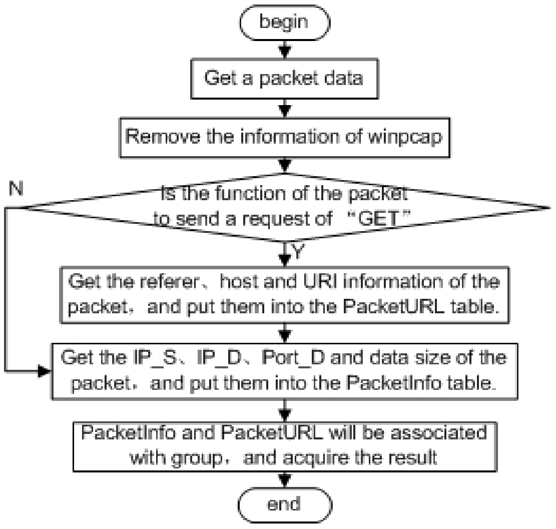
begin
Get a packet data
Remove the information of winpcap
Is the function of the packet to send a request of "GET"
N
Y
Get the referer、host and URI information of the packet，and put them into the PacketURL table.
Get the IP_S、IP_D、Port_D and data size of the packet，and put them into the PacketInfo table.
PacketInfo and PacketURL will be associated with group，and acquire the result
end

Figure 3: The detail of processing

1. Move back the non-data content of the data packet automatically added by Wincap,24 bytes in total.
2. Move back the non-data content of the data packet automatically added by Wincap,16 bytes in total. This part of the packet has recorded the capture time and length of the packet. Transform the time data to a decimal base integer and then to the time format character. As China is in the time zone 8, add 8 hours to the time data. Use the function memcpy (&m_ph,m_Data,16) to get the packet length.
3. Copy the data segment which loaded by TCP Protocol to the temporary strings container. Firstly, find whether the content of
4. First three bytes in the data container is "GET"; if it is "GET", then look for the information of two consequent "\0x20" characters which behind "GET", URL resource location can be extracted. Secondly, find it whether contains "Referrer:"string sign information, then Referrer information can be extracted. And then find it whether contains "Host:" string sign information, then Host information can be extracted. Finally, under the premise of containing "GET" information check it whether has at least one "Referrer" or "Host" information, If it has, then write the extracted information, related source and destination IP, and port information to table Packet URL in My SQL database.
5. If data packet has port 80 characters, but does not have any information about "GET", "Referrer" or "Host" at the same time, then write the extracted information, related source and destination IP, and port information to table Packet Info in My SQL database.
6. In table Packet Info, the amount of resource can be analyzed by grouping IP_S, IP_D, and PORT_D.
7. The URL data size of every resource can be come out to table Packet Result by associating table Packet Info and Packet URL in the order of IP_S, IP_D, PORTD.
8. In table Packet Result, the number of repeat visits of resource and the amount of data size of redundant network traffic through grouping URL

4. EXPERIMENT

The network data collection test platform is able to get all the information in any school network nodes by the function of "network-centric CISCO6509" mirroring in the device, which is shown in Figure 1. We obtained experimental data from a building real-time network data, which started on 2010-12-18 8:00 and ended on 2010-12-19 8:00, totally for 24 hours capturing of data. Since we only captured data packets from port 80, the size of the captured data is 133.8GB. According the processing method above, results are as follow:

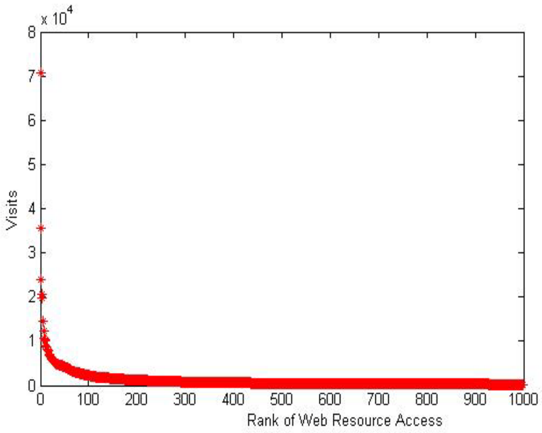

Figure 4: The detail of processing

In Figure 4, it is a twenty-four-hour observation result of users accessing Internet through http Protocol in one building in SWUST (Southwest University of Science and Technology) campus. Unlike the observation result in [13] which mainly analyzed the visits of web sits from web log data, in this study, we mined the information more deeply, and the analysis of visits is based on data packets. During the experiment over 110 thousand visits of different web sites were observed, visit heats are very difference. The most popular site received more than 70 thousand visits, while the coldest one only had one visit, which drops over five orders of magnitude. When the web sites sorted by visit heat, it is found that top 10 web sites attracted about 230 thousand visits (occupied 8.3%), top 50 web sites attracted 470 thousand visits (occupied 16.7%),top 100 web sites attracted 620 thousand visits(occupied 22.1%), top 1000 web sites account for 1.17 million visits (occupied 41.8%). It can be seen that the majority of users concentrated their interests in a few web sites [13].

The research object is web pages, and then we conducted further mining of page data, the resource (the resource here refers to the transmission data that apply independently, which is as large as a independent web page, or as small as an icon) of these web pages was studied as the unit. In the experiment, 3.6 million visits of resources and totally 1.93 million resources are observed by the study of the minimum logical unit in the network data. Through researching the data of top 10 visits resources, it is found that some web sites' advertisements or messages occupied most part of the data. For example, the top 1 visits (swf.flash2.minigame.xunlei.com/flash_game/recommend_for_xunlei7.xml) and the third one (http://msg.baidu.com/ms?ct=18&cm=3&tn=bmSelfUsrStat&mpn=13227114&un=sucjhwaxp) is one web site's advertisement and message respectively.

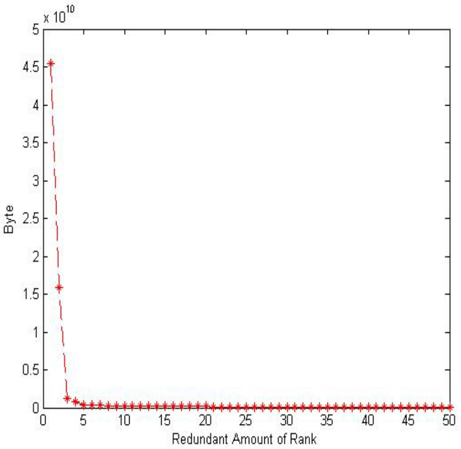

Figure 5: Redundant Network Traffic

The experiment proves that there exists a great deal of redundancy in the network traffic. The redundant network traffic is defended that the same content data transferred repeatedly between two network nodes in the Internet, which not only wastes network bandwidth, reduces the quality of network service, but also increases the network servers' load. Especially audio and video information has become mainstream media in the network; the visited number of video files are thousands of times of text files, also users visits popular sits thousands of times more than common sits. Then, Conduction surfaced by sharing hot video information. It is hoped that sharing videos national wide will be a consideration in the next generation. However, it will never meet the requirement of the whole country; even if further increase the bandwidth of ten times, hundred times in the classic server/client (C/S) structure. During the building of next generation Internet, "satellite broadcasting and distribution storage" and "broadcast storage structure" will be added as a sub-structure, which is based on considered keeping TCP/IP as the main basic structure in the Internet. Thus, the next generation Internet (CNGI) will have a unique dual structure around the world [14].

5. CONCLUSION

Data of port-80 is an important part of network data, which is also one of the data services that need to be analyzed and studied accurately. Through mining data packets of port-80 deeply, mining characteristics of data flow horizontally, and vertically mining the data business and behavioural characteristic of high-level application protocol packets, the behaviour of network users and network redundancy and other aspects of data transmission can be more comprehensively analyzed. As a result of completely holding the characteristics of network data at this stage, the study provides a foundation for the development of network transfer techniques.

REFERENCES

[1] Zhou mingzhong, study of large-scale network ipflows behavior characteristics and measurement algorithms, 2010-07-060

[2] Jiang baolin; shen zhan; zhang chuan; ge jiaxiang; hu yunfa;. web log mining considering web content and web structure computer engineering,2004,30(16):30-32.

[3] Chen sheng-rong; dong shou-bin. research on chinese web pages classification based on preferential links journal of zhengzhou university (natural science edition),2007,39(2):78-82.

[4] Ghani r,slattery s,yang y m.hypertext categorization using hyperlink patterns and meta data. San Francisco:eighteenth international conference on machine learning,2001:178-185.

[5] Oh h,myaeng s,lee m.a practical hypertext categorization method using links and incrementally available class information[c].athens:23rd annual international acm sigir conference on research and development in information retrieval,2000:264-271.

[6] Moore a w, zuev d. internet traffic classification using bayesian analysis techniques; acm sigmetrics. New York: acm press, 2005: 50-60.

[7] Auld t, moore a w, gull s f. bayesian neural networks for internet traffic classification[j]. ieee trans on neural network, 2007, 18(1):223-239.

[8] Okabe t, kitamura t, shizuno t. statistical traffic identification method based on flow-level behavior for fair voip service//[c]ieee workshop on voip management and security. vancouver: ieee press, 2006: 35-40.

[9] Karagiannis t, papagiannaki k, faloutsos m. blinc:multilevel traffic classification in the dark; ACM sigcomm 2005. philadelphia: acm press, 2005:229-240.

[10] Lin ping; yu xun-yi; liu fang; lei zhen-ming; a network traffic classification algorithm based on flow statistical characteristics, journal of beijing university of posts and telecommunications, 2008-02-003

[11] Miao chen, shun-hua tan, guo-hai yang , yi-zhi wang ;research on network business identification technology based on ip packets, international conference on apperceiving computing and intelligence analysis 2010

[12] Shunhua tan, miao chen,guohai yang and yizhi wang; research on redundant network traffic, 2011 international conference on innovation and information management

[13] Tan shunhua, user behavior mining on large scale web log data , 3rd international conference on computer design and applications

[14] Li youping, research of complementary architecture network , journal of swustwest university of science and technology, 2006(3):1-5