Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 3
August 2009

Study on the TOPN Abnormal Detection Based on the NetFlow Data Set

Hongzhuo Zhang

School of Computer Science and Technology, University of Electronic Science and Technology of China
Chengdu 610054, China

Abstract

In recent years, with the increase of the scale and the complexity of the network, various abnormity flows begin to occur in the network. To know the running state of the network, the technology of NetFlow emerges as the times require. The NetFlow data are transmitted directly by the router which supports the function of NetFlow. Comparing with traditional data acquirement technology, the NetFlow technology needs not deployment in advance and acquires data conveniently, and it is gradually turning into the important data sources for the network management, maintenance, supervision and control. At present, there are still few abnormity detection methods based on the NetFlow data set. In this article, we introduced the principle and functions of the NetFlow, and put forward the TOPN flow abnormity detection method based on the NetFlow technology. This method can effectively detect the flow state in the network, reflect the network state and offer the information about abnormal network flows.

Keywords: NetFlow, TOPN, Abnormity detection

1. Introduction

In recent years, as Internet is using in the life, learning and work by more and more people, and the running of the whole society has been carved by the sign of Internet, and the Internet has been developed from single industry tool to the social and popular toll entering into various industries in China. The scale and complexity of Internet increase day by day, and multifarious network applications are correspondingly emerging. However, with the normal application flows of Internet, various abnormity flows successively occur, which have influenced the normal running of Internet and threatened the security and normal use of the host computer. In order to know the running and use state of network and find out the possible abnormal flows in the network, we need an effective flow detection method, so we put forward the TOPN abnormity detection method based on the technology of NetFlow. Most traditional detection methods need attacking the samples to detect the flow by describing the special mode of each attack. By analyzing the data packet in virtue of the net capture data software to compare the attach samples and complete the detection, this method will possesses higher precision rate. But its abilities of capturing data and analyzing speed especially in the high speed network are not ideal, and the invasion detection system based on characteristics is implemented only depending on the alarming rules established by human beforehand, so it lacks in the self-learning function when it faces changing network attacks continually. The method we put forward in this article is based on the NetFlow technology, and it can realize the data acquirement and analysis in high speed network, and the TOPN detection model we established has certain self-learning ability and ideal abnormality detection ability.

2. Brief introduction of the NetFlow technology

The technology of NetFlow is a set of network flow monitoring technology developed by Darren Kerr and Barry Bruins of Cisco Company in 1996, and it have obtained the supports from many mainstream manufacturers such as Juniper and Extreme. At present, it has been embedded in most Cisco routers, and is gradually becoming into the industrial standard. It is not only an exchange technology, but also a flow analysis technology, and it is one of mainstream charging technologies in the industry. It can solve following questions about the flow such as who interviews whom, where it interviews, which protocol is used and how much the concrete flow is. Because of the technology and the market occupancy ratio of the Cisco network product, the NetFlow technology has been one of mainstream flow analysis technologies at present. Its work principle includes storing the data into the cache by the format of flow record, and educing the data through the protocol of UDP until the data fulfill the conditions. The NetFlow data in this article are transmitted to the data capture host computer through the protocol of UDP by the router of Cisco, and because the data scale is huge, we write the data by the form of file to store in order to ensure the effective acquirement of the data. To be convenient for data analysis, we established a set of Oracle data sever which could translate the written data file into the

Oracle database and analyze the data by operating the Oracle database.

There are many data formats for NetFlow, we adopt the data with NetFlow Edition 5 to study, and the data fields used in this article mainly include unix_secs (the second amount from 0000 UTC 1970 to now), srcaddr (source IP address), dstaddr (destination IP address), dOctets (the total amount of byte on the third layer in the data packet of information flow), srcport (TCP/UDP source port number), dstport (TCP/UDP destination port number), and prot (IP protocol, such as 6=TCP and 17=UDP).

3. Abnormity detection based on TOPN analysis

The detection principle of TOPN includes the statistical analysis of TOPN and the model establishment based on the acquired NetFlow data. The TOPN analysis is to implement statistical ordering aiming at certain selected index, and select the top parameters which can fulfill the conditions. In the article, we mainly explain six TOPN statistical analysis items including host-computer initiating connection amount ratio TOPN analysis, host-computer emitting data amount ratio TOPN analysis, protocol use ratio TOPN analysis, destination port use ratio TOPN analysis, source IP and destination IP pair ratio TOPN analysis, and destination IP ratio TOPN analysis.

(1) The host-computer initiating connection amount ratio TOPN analysis is mainly used to find out the IP addresses, rankings and connection amount ratios of the top N host-computers which initiate the most connections in certain period.

(2) The host-computer emitting data amount ratio TOPN analysis is mainly used to find out the IP addresses, rankings and connection amount ratios of the top N host-computers which emit the most data in certain period.

(3) The protocol use ratio TOPN analysis is mainly used to find out the top N used protocols, rankings and the ratios relative to the total protocols in certain period.

(4) The destination port use ratio TOPN analysis is mainly used to find out the top N used destination port numbers, rankings and the ratios in the all ports in certain period.

(5) The source IP and destination IP pair ratio TOPN analysis is mainly used to find out the top N used source IP and destination IP pairs, rankings and the ratios in all source IP and destination IP pairs in certain period.

(6) The destination IP ratio TOPN analysis is mainly used to find out the top N used destination IP addresses, rankings and the ratios in all destination IP addresses.

3.1 TOPN ranking detection

TOPN ranking detection is a sort of usual TOPN detection method, and it is mainly used to confirm a general ranking sequence, and once the ranking by the statistics has certain transition with the consulted ranking, the result is treated as abnormity. Because we adopt the technology of NetFlow, we aim at the wide area network (WAN). There are numerous IP addresses in a huge WAN, and in the short term, the occurrences of top IP addresses of TOPN in different time periods will be dispersed, it is not ideal for ranking detection, so we can find out the destination ports and protocol types with few changes by the experiment to establish the model which would perform the abnormity detection.

3.2 TOPN ratio detection

There are numerous IP addresses in a huge WAN, and in the short term, the occurrences of top IP addresses of TOPN in different time periods will be dispersed, but the use ratios of top N IP addresses accord with the normal distribution, so we can establish the model according to the change range of ratio to implement the abnormity detection.

The mass data are adopted to establish the model first, and the mass data in one week can cover the abnormity. After that, the model should be renovated periodically, and the time of one weak probably is used to renovate the model, and the data to renovate the model are the normal data selected by the model.

The main method establishing the model is to use the average value to add the n times of standard deviation.

(1) The host-computer initiating connection amount ratio TOPN model mainly includes the top N TOPN and the corresponding connection amount ratios.

(2) The host-computer emitting data amount ratio TOPN model mainly includes the top N TOPN and the corresponding data amount ratios.

(3) The protocol use ratio TOPN model mainly includes the top N TOPN and the corresponding protocol use ratios.

(4) The destination port use ratio TOPN model mainly includes the top N TOPN and the corresponding destination port use ratios.

(5) The source IP and destination IP pair ratio TOPN model mainly includes the top N TOPN and the corresponding source IP and destination IP pair ratios.

(6) The destination IP ratio TOPN model mainly includes the top N TOPN and the corresponding destination IP ratios.

When the ratio of relative TOPN by the statistic is higher than the corresponding ratios of the model, the data in this period are regarded as abnormity, and the system will emit alarm and perform the abnormity detection aiming at the abnormity and data. By the experiment, we found that the alarm rate was a little high when the model used the single standard deviation, and the detection effect was good when the model used the double standard deviations, and the detection effect was bad when the over-double standard deviations were used.

4. Experiment and analysis

By setting up the router of the school network information center, we use it to transmit the NetFlow data to the data capture host-computer by UDP protocol, and store the data into the Oracle data server, and analyze the server to perform relative data analysis and detection. The experiment environment is seen in Figure 1.

4.1 TOPN ranking detection

In the experiment, we can find out the ranking distribution rules of protocol and port in the network. The main distribution sequence tendency of the protocol generally is UDP, TCP, ICMP and Reserved. Sometimes UDP and TCP will exchange their places, which are related to the flow of P2P, and the ICMP and Reserved will also exchange their places sometimes. The main distribution sequence tendency of the port generally is from 80 (HTTP), 8000 (Tencent QQ sever opens this port), 015000, 28000, 0 (Reserved), 443 (Https), 8080 (WWW agent opens this port), 10000 (network data management protocol), 26000 to 21 (FTP). According to the sequence, we establish the normal ranking model, and when the statistical port ranking and protocol ranking change and the abnormity data occur, the system will emit alarm and perform the abnormity detection.

4.2 TOPN ratio detection

Table 1 ~Table 4 are the ratio detection models in certain period.

Table 5 and Table 6 respectively listed the TOPN detections of the destination port and protocol in five minutes, and when the system detected the abnormity, the system would compare the abnormity data with the sample base of attack type to judge the attack type.

The ratio values of the second model and the eighth model exceed the corresponding values of the normal model, so the system emitted alarm and compared the data of abnormity with the sample base, and the detection results respectively were “Teardrop 68: refusing service attack” and “TCP SYN 61: refusing service attack”.

The ratio value of the first model exceeded the corresponding value of the normal model, the system emited alarm and compared the data of abnormity with the sample base, and the detection result was “Teardrop 68: refusing service attack”.

5. Conclusions

Through analyzing the principle and functions of NetFlow, we put forward a sort of TOPN abnormity detection method based on NetfFow, which could effectively detect the abnormity flows in the network by means of the traditional ranking transition detection method and the corresponding ranking use ratio value detection method, and be effectively applied in the wide area network with high speed. Nowadays, this method has been successfully applied in certain practical project of University of Electronic Science and Technology of China.

References

Lai, Jibao, Wang, Huiqiang & Jin, Shuang. (2007). Study of Network Security Situation Awareness System Based on NetFlow. *Application Research of Computer,* No.8.

Li, Xiangguo & Fei, Lingling. (2008). The Research of NetFlow-based Flow Analyzing Technology. *Control & Automation,* No.15.

Wang, Hua. (2004). Key Software Technique and Algorithm of Analyzing NetFlow Traffic Data. *Computer Engineering,* No.z1.

Yan, Jiahao, Ma, Rui & Wu, Yibo. (2006). Application and Design for Internet Traffic Monitoring Technology. *Designing Techniques of Posts and Telecommunications,* No.4.

Table 1. TOPN model of host-computer initiating connection amount

TOPN ranking N	Ratio
1	1.7237%
2	1.4924%
3	1.3819%
4	1.3179%
5	1.1980%
6	1.1417%
7	1.0807%
8	1.0516%
9	1.0327%
10	1.0071%

Table 2. TOPN model of destination IP ratio

TOPN ranking N	Ratio
1	1.4124%
2	1.2801%
3	1.1746%
4	1.1195%
5	1.0332%
6	0.9919%
7	0.9486%
8	0.9216%
9	0.9012%
10	0.8783%

Table 3. TOPN model of protocol use ratio

TOPN ranking N	Ratio
1	57.9436%
2	44.7775%
3	1.2337%
4	0.8144%
5	0.0197%
6	0.0017%

Table 4. TOPN model of destination port use ratio

TOPN ranking N	Ratio
1	12.9209%
2	3.0820%
3	2.7714%
4	2.3662%
5	1.8860%
6	1.3052%
7	1.0643%
8	1.0027%
9	0.9635%
10	0.9272%

Table 5. TOPN statistics of destination port

TOPN ranking N	No. of destination port	Ratio	Start time
1	80	5.29374271142224	2008-12-10 10:56:03
2	28000	3.35425422559238	2008-12-10 10:56:03
3	15000	1.92983437718383	2008-12-10 10:56:03
4	0	1.42244514436732	2008-12-10 10:56:03
5	53125	1.41290074150137	2008-12-10 10:56:03
6	22222	1.16968302708965	2008-12-10 10:56:03
7	8000	1.05471136957793	2008-12-10 10:56:03
8	33333	1.0075378841485	2008-12-10 10:56:03
9	30433	0.848354797269204	2008-12-10 10:56:03
10	4576	0.755763118891445	2008-12-10 10:56:03

Table 6. TOPN statistics of protocol

TOPN ranking N	No. of protocol	Ratio	Start time
1	17(UDP)	59.6278341116958	2008-12-10 10:56:03
2	6(TCP)	39.1158152949604	2008-12-10 10:56:03
3	1(ICMP)	1.24362472285577	2008-12-10 10:56:03
4	47	0.0127258704879384	2008-12-10 10:56:03

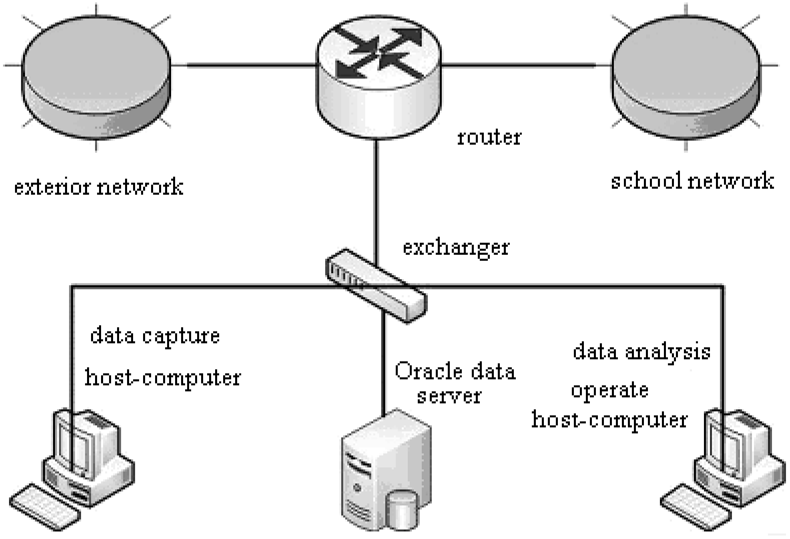

Figure 1. Experiment Environment