Application of KMP Algorithm in Customized Flow Analysis

Qingzhu Meng, Zhenming Lei, Dazhong He
Beijing Laboratory of Advanced Information Networks,Beijing Key Laboratory of Network System Architecture and Convergence
Beijing University of Posts and Telecommunications
Beijing,China
e-mail:victor_mqz@qq.com

Huan Wang
Haohandata Technology Co., Ltd
Beijing, China
e-mail: wanghuan@haohandata.com.cn

***Abstract*—Internet technology has developed rapidly today, the Internet has become an indispensable part of the way everyone online more and more diverse, there are PC, mobile phones, flat and even watches, etc .; and the speed of the network with the increase in bandwidth and rapid growth The The traffic data completely records the behavior of each person online, if you can effectively use the flow of data analysis, it will be a huge treasure. The usage of virtualization technology can improve the utilization rate of data center equipment and bring convenience to cloud computing applications [1]. Based on the analysis of traffic, this paper introduces the basic idea of traffic analysis system, what is custom traffic analysis, and then explores how to improve the efficiency of traffic analysis. KMP algorithm is proposed for initialization based on the K-means partition algorithm[2], experiments show that KMP algorithm significantly reduces the disparity between the initial solution and the actual solution[3], so you can consider using it to improve efficiency ,such as the application of KMP algorithm in traffic analysis.**

Keywords-traffic analysis; flow; KMP algorithm

I. INTRODUCTION

The 21st century is a high-speed information network era, the rapid development of the Internet on the national economy and life had a profound impact. In recent years, with the popularity of the Internet, a variety of mobile and web pages of explosive growth, network traffic into exponential growth. The Internet into the people's lives, changing the people's living habits, the emergence of electronic banking, social networks such as WeChat, the rise of microblogging, the popularity of e-commerce caused nearly five years, the global mobile Internet traffic An increase of nearly 20 times, the three major operators to increase the speed of network traffic.

Recently, it's more getting that volume of data size has grown to terabytes and petabytes, and the types of data generated by applications become richer than before[4].This high-speed development of the community, not only showed a flow of exponential growth, data processing capabilities are also developing rapidly. Now the community has undoubtedly been fully into the era of large data, data processing speed has risen several grades, from hadoop technology few people try to now spark data processing technology is only mature less than five years. In the big data, the traffic data has become a treasure house for each company trying to dig. And with the large data technology matures, cloud technology has also been popular in the Internet, so the flow of data processing, making it easy to calculate the data is more important.

Traffic data is mainly from the network interface, port and protocol packets intercepted by the preservation of the message. According to the different protocol and the network environment, the message file has different message format. Therefore, the message file can not be processed directly. If the processing is forced, it will take a lot of time. In order to meet the needs of rapid and convenient processing, the need to extract the necessary information in the message file to form a convenient format to deal with the file, can greatly improve efficiency.

The information in the message includes the IP address of the user in each network, and the IP address that the user wants to access, as well as the response of the server and the response of the server. In short, the traffic data packets contain most of the data in the Internet, as long as good at using, you can get the user's preferences, needs, Internet habits and other information. In order to be able to quickly read the message in the message, you need a systematic tool to achieve this function. High-performance network protocol analysis and output key technology is precisely for this demand generated by the technology, to the message file as an input file, according to the needs of users, you can output different keywords, according to their own needs to change the key Word output format, and can also accurately locate the traffic information of different users, select the number of specific population flow. In short, high-performance network protocol analysis and output key technology is now a necessary tool for the community.

II. THE BASIC IDEA OF CUSTOMIZED FLOW ANALYSIS

Network protocol analysis refers to the binary format of data packets transmitted on the network are resolved to restore all the network protocol information and the transfer of technology [5]. The customization flow analysis of this paper is a technology that can analyze the real-time traffic message data in the local area network into a customized output file. The interface to process pcap is function pcap_loop which is used to cycle to capture network packets,

978-1-5090-6352-9/17/$31.00 ©2017 IEEE

and transfer received data into data processing function packet_handler[6]. The main idea is as shown in Fig. 1.

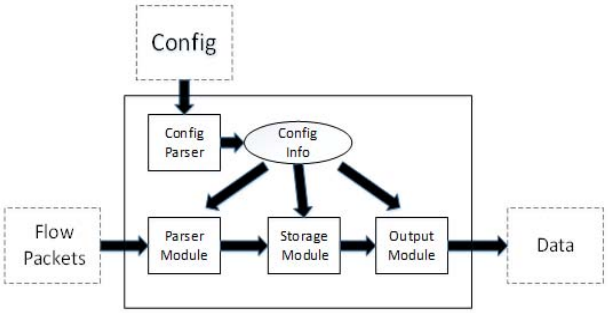

Figure 1. Customized flow analysis architecture.

A typical high speed traffic archiving and querying system has three critical functions: packet capture, packet indexing, and storage management [7]. This article is mainly concerned with the content of customized ideas, the so-called customization, that is, the output content can be customized, that is, users can customize their own needs in the output file field number, order, and the contents of the field operations. The user needs to write the configuration file according to the custom syntax before the specific output can be made according to the configuration information.

First read the configuration module using the custom syntax parsing engine to extract the configuration file information, and then save the extracted configuration information. And then call the PF_RING library function to read the network card to receive the message file, PF_RING can be efficient to receive the kernel card message flow into user mode. When a message from the kernel state to user mode will go through three processes: the analysis module according to the configuration information content for specific analysis, such as IP only need a field, only to resolve the IP address; storage module will parse the field The output module will take out the linked list of the queue to be processed, traverse the linked list, and organize the fields according to the field processing method in the configuration information, after which the output module will sort the list of the queue to be processed, and then sort the list according to the field processing method in the configuration information. The collated fields are spliced into a single bill for output. As compute power and storage capacities continue to rise, and costs continue to decline, Big Data and analytics are playing an increasingly important role [8]. The output will be sent to HDFS which is work for bigdata.

Do not worry about the speed of the input module, because we use PF_RING to reveive packet. PF_RING ZC's processing rate reached line rate 10Gbps for all packet lengths[9].In the whole process, the field processing is undoubtedly the most frequent implementation, such as a message may be output dozens of fields, each analysis of a message may be executed dozens of field processing functions. And in the case of large traffic in the LAN, there may be tens of thousands of messages per second, so the efficiency of the field processing function to be as high as possible to ensure that the message can be dealt with in a timely manner. In the field processing, the string lookup is a relatively time-consuming work, such as the URI field is usually longer (to make the length of n), in the matching field is not very short case (the length of m), if the use of BF Algorithm to find the processing time complexity o (mn), so can not meet the high performance requirements, this paper needs to explore more suitable for high-speed network in the string search algorithm.

III. THE STRING MATCHES IN THE FIELD PROCESSING

String matching refers to the length of a text in the text to find the length of m sub-string Partion the emergence of the location, the application of custom traffic analysis process in the field processing part. As the field processing frequency is very high, so the need for higher performance in order to ensure the entire flow analysis process efficient.

To find the substring, the easiest way to think is the BF algorithm, also known as brute force matching algorithm, the efficiency of this algorithm is very low, the process is: the text from the first position in turn followed with the characters of Partion, until the case To a different character, then the text start position to advance a character, and then continue to compare Partion; if not encountered different, to record the start of the match position. Although this algorithm is easy to understand, but the efficiency is very low, the average time complexity to o (mn), because in the traversal process Text pointer often back to the previous match the beginning of the position, which led to the algorithm is very inefficient The The following describes an algorithm that can effectively reduce this repetitive operation.

A. KMP Algorithm

However, it is necessary to employ bit-oriented frame header searching when the stream is byte-misaligned, which induces a more complex high-level language programming as well as a lower efficiency (the time complexity of the bit-oriented sliding window frame synchronization method is o(m× n) , where m, n are, in bit, the lengths of the stream and the frame header, respectively)[10]. The biggest advantage of the KMP algorithm relative to the BF algorithm is to eliminate the backtracking of the Text pointer. The matching result can be used to move the position of the Lastion string to the next match. Therefore, the time complexity of KMP algorithm is o (m + n), the efficiency is much higher than the BF algorithm.

The basic process of the KMP algorithm is to set the pointer t and p at the beginning of the Text and Partion, and compare the current pointer to the corresponding character from the starting position. If they are the same, move the t and p respectively; if they do not match, p Pointer jumps to points that are as far away from the starting position as possible. It is because of this jump can make the t pointer does not have to backtrack, it will not miss the effective match. So the first step in the algorithm is to get the p pointer in different positions when the match failed to jump to the location, such a position in the Partion string itself to find the law. The core idea is to find the current position of the Partion string for the end of the length of the sub-string k can be with the Partion start k-bit string match, such as abcdabcbcd substring, when matching to 5 to 7 bits of abc, with the start position abc match, then that if Text and

Partion match to the first 7 of Partion failed, then the text of the current pointer for the end of the two must be able to Ab and Partion the beginning of the two match, so the next comparison only need to use Partion The third is in the current position of Text to compare, so there is no need to let Text back to the two before starting to compare.

In order not to miss all the matching situation, so the p pointer jump distance to be as small as possible, for a Partion, the location of their matching failure is the only jump, and only their own. By the following algorithm can get the next array, next [i] that Partion match to the i-bit failure after the p pointer to jump to the location.

In the case of the second array, if the p pointer is at the end of the Partion, the i-j + 1 bit of the text is If the two pointers do not match, the p pointer jumps to the next [j] bit of the Partion. If the pointer does not match, the p pointer jumps to the next [j] bit of the Partion. The algorithm process is shown below.

The above is the use of Partion's next array to find Partion in the text of the matching position of the KMP algorithm the whole process, t pointer position i do not have to back, so the time complexity is o(n + m), when Text can be approximated to o(n).

B. Field Processing in Customized Traffic Analysis

The message is stored in the storage parsing module and the result is stored. In order to save the time of packet parsing, the module will only dump the contents of the field without parsing and storing it as a string. If not processed, the final output of the message is the original content of the cut, can not be understood by the user, such as the contents of the IP address field in the message "61 61 61 61", then the user can understand the IP The address should be "97.97.97.97", but if the field is not processed, the displayed IP address will be "aaaa". Similar to the situation there are IPv6 address, TCP layer and UDP layer port number, data link layer mac address and other information need to be converted.

Algorithm I get next[]

```
Input:  partion
Output: next[]
getNext
  j <- 0
  k <- -1
  m <- partion.length
  next[0] <- -1
  while j < m then
  do
   if k = -1 or partion[j] = partion[k] then
   do
    let j + 1
    let k + 1
    next[j] <- k
   else
    k <- next[k]
   end if
  end of while
Return next
```

Algorithm II KMP core

```
Input:  text、partion and next[]
Output: the position of match
KMP
  i <- 0
  j <- -1
  n <- text.length
  m <- partion.length
  while j < m and I < n then
  do
   if j = -1 or partion[j] = text[i] then
   do
    let j + 1
    let k + 1
   else
    j <- next[j]
   end if
   if j >= m then
   do
    result <- i – m
   else
    result <- 0
   end if
  end of while
Return result
```

Another situation is that the field results may be very long, but only need to focus on one of them, you need to truncate, such as field URL will generally be longer, but only need more than a dozen characters in front of the user can know What kind of website.

In addition, the user may need to deal with the results of the analysis, such as the TCP sequence number seq is often very large, users want to get tens of thousands of units, it is necessary to seq divided by ten thousand, to facilitate the user to determine the value of the order of magnitude.

From the above analysis we can see that the field processing function is very useful, not only can significantly improve the usability, but also can improve the efficiency of the analytical process. This article is mainly concerned with the field processing function in the field matching class method, the method used to determine whether there is a specified string in the field, the specified string is often not particularly short, the function has only one parameter, to find the string. Such as the common user-agent field or referer field will be longer, in the field field search performance requirements will be higher. Next, the performance of the search for the test.

C. Application of KMP Algorithm in Field Processing

In order to verify that the KMP algorithm is capable of field processing, performance testing is used to evaluate its efficiency. Performance test input file for the 50,000 messages (ordinary LAN traffic), the configuration file to set the output field for the user-agent, the use of field processing methods for the field search, find sub-string "Firefox" and "Chrome", the use of Intel The vTune test tool tests the total time-consuming of the field matching method.

TABLE I. TIME-CONSUMING OF EACH ALGORITHM

Algorithm	Find "Firefox" cpu time consuming	Find "Chrome" cpu time consuming
BF	0.116 seonds	0.109 seconds
KMP	0.074 seconds	0.056 seconds

In the actual development process, hardware selection is reasonable or not will directly affect the data acquisition performance [11]. In this paper, we make process on the same device which is working for our lab. A single variable method is used to test the BF algorithm and KMP algorithm respectively. The test result is time consuming higher than KMP algorithm. Considering the special case, ten fields in a message need to use the character matching method, then the BF algorithm takes more than 1 second, can not completely deal with the message in the network; and the use of KMP algorithm to ensure that in extreme cases can work smoothly.

IV. CONCLUSION

This paper introduces the main idea and architecture of the custom flow analysis system, and introduces the process of converting the network traffic from the message to the form of formatted data. In this process the main focus on the field processing, what is the field processing, why should this step, and this paper presents a vision, that is, the string search algorithm applied to the feasibility of this step. After testing the network packets, it is found that the benefits obtained using the KMP algorithm can meet the needs of routine traffic analysis. The whole algorithm depends little on hardware and can have high efficiency and accuracy [12]. Accordingly, this paper concludes that the KMP algorithm can be applied in the field processing of customized traffic analysis.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China (61671078, 61701031), Director Funds of Beijing Key Laboratory of Network System Architecture and Convergence (2017BKL-NSAC-ZJ-06), and 111 Project of China (B08004, B17007). This work is conducted on the platform of Center for Data Science of Beijing University of Posts and Telecommunications.

I would like to thank the laboratory teachers for my growth brought great help and support. Mr.Lei highly respected, over seventy years of age is still conscientious, every day on time to the laboratory to lead students to academic research, to guide the research direction of the laboratory for each of us to establish a good example. At the same time, I would also like to thank my professional teacher Huan Wang, Mr.Wang always in the face of my problems to give me professional guidance, and according to my personal situation to my developmental advice, so that I can quickly get started and steadily improve my ability. What about teacher Dazhong He, with a lot of practical work, so that I can understand the significance of the experiment and role. It was with the help of these teachers that I could learn something that made me feel like working at home in the lab.

REFERENCES.

[1] S.J.Ding. Network Resource Fault Detection Based on Traffic Analysis[A]. Advanced Science and Industry Research Center.Proceedings of 2015 International Conference on Automation,Mechanical and Electrical Engineering(AMEE 2015)[C].Advanced Science and Industry Research Center:,2015:8.

[2] XUAN Hengnong,ZHANG Runchi,SHI Shengsheng. An Efficient Cuckoo Search Algorithm for System-Level Fault Diagnosis[J]. Chinese Journal of Electronics,2016,25(06):999-1004. [2017-09-23].

[3] Yoon-joo CHAE. Towards Conceptual Big Data Software Platform[A]. Science and Engineering Research Center.Proceedings of 2015 International Conference on Control, Automation and Artificial Intelligence(CAAI 2015)[C].Science and Engineering Research Center:,2015:5.

[4] CAI Yu-chen,WANG Zhen-hua,ZHANG Guo-feng,CHEN Zhao-hui. PC-based frame synchronization method for byte-misaligned stream[J]. The Journal of China Universities of Posts and Telecommunications,2014,21(03):23-28+34. [2017-09-23].

[5] Qing-Xiu Wu. The Network Protocol Analysis Technique in Snort[A]. Information Engineering Research Institute, USA.Proceedings of 2012 International Conference on Solid State Devices and Materials Science(SSDMS 2012 V25)[C].Information Engineering Research Institute, USA:,2012:5..

[6] Xiaofan Lu Graduate School Changchun University of Technology Changchun,China Weijia Sun School of Computer Science and Engineering Changchun University of Technology Changchun,China Huiping Li Graduate School Changchun University of Technology Changchun,China. Design and Research Based on WinPcap Network Protocol Analysis System[A]. IEEE Industrial Electronics Society Beijing (Shenzhen) Chapter、Changchun University of Technology, China、Intelligent Information Technology Application Research Association (IITA Association), Hong Kong.Proceedings of 2010 International Conference on Computer,Mechatronics, Control and Electronic Engineering (CMCE 2010) Volume 1[C].IEEE Industrial Electronics Society Beijing (Shenzhen) Chapter、Changchun University of Technology, China、Intelligent Information Technology Application Research Association (IITA Association), Hong Kong:,2010:3.

[7] WANG Zhe School of SoftWare of BeiJing Jiaotong University School of SoftWare of BJTU BeiJing,ChinaCHEN Jun,YUAN Gang Zhao ZhouYang School of SoftWare of BeiJing Jiaotong University School of SoftWare of BJTU BeiJing,China. A Quick algorithm for planar fragmented objects stitching Base on KMP algorithm[A]. IEEE Beijing Section, China、Chongqing Computer Federation, China、Chongqing University, China、Beijing University of Technology, China、Southwest University, China、Chongqing Three Gorges University, China.Proceedings of 2011 13th IEEE Joint International Computer Science and Information Technology Conference(JICSIT 2011) VOL.03[C].IEEE Beijing Section, China、Chongqing Computer Federation, China、Chongqing University, China、Beijing University of Technology, China、Southwest University, China、Chongqing Three Gorges University, China:,2011:4.

[8] Zhen Chen,Linyun Ruan,Junwei Cao,Yifan Yu,Xin Jiang. TIFAflow: Enhancing Traffic Archiving System with Flow Granularity for Forensic Analysis in Network Security[J]. Tsinghua Science and Technology,2013,18(04):406-417. [2017-09-23].

[9] K, Anders, Ericsson. Protocol Analysis[M]. Colorado:Departments of Psychology University of Colorado and Carnegie-Mellon University, 1981.

[10] David, R, Musser. STL Tutorial and Reference Guide: C++ Programming with the Standard Template Library[M]. America:Addison-Wesley Professional, 2009.

[11] Rick, Hofstede, Pavel, Čeleda, Brian, Trammell. Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX[D]. America:IEEE Communications Surveys & Tutorials, 2014.

[12] Haipeng Wang. Comparison of High-Performance Packet Processing Frameworks on NUMA[A]. The Institute of Electrical and Electronics Engineers、IEEE Beijing Section.Proceedings of 2016 IEEE 7th International Conference on Software Engineering and Service Science（ICSESS 2016）[C].The Institute of Electrical and Electronics Engineers、IEEE Beijing Section:,2016:5.