北 京 邮 电 大 学

BBC6521 Project 毕业设计 2015/16

Early-term Progress Report

初期进度报告

学院 School	International School	专业 Programme	Telecommunication Engineering and Management	班级 Class	2012215104
学生姓名 Student Name	Cheng Qian	学号 BUPT Student No.	2012212860	学号 QM Student No.	120721267
设计（论文）编号 Project No.	RN_2860	电子邮件 Email	tom10ye@bupt.edu.cn		
设计（论文）题目 Project Title	User behavior analysis based on DPI(Deep Packet Inspection)				

已完成工作：

Finished Work：

Regular meeting was arranged with my supervisor on every Tuesday. With the help from Professor Yao, I made a schedule of the whole project and checked our milestones every week, which guaranteed the progress of the project. Until now, I have deployed a DPI platform of version nDPI1.6 on Ubuntu system and successfully real-time grasped the Internet packets and analyzed the protocol type in Data Link layer, Network layer, Transport layer and Application layer. Due to the open nature of Deep Packet Inspection technology, I read a lot of blogs on CSDN forum to understand the principle and working structure of the nDPI model. Then I analyzed its source code in C with GDB test. Tracking the change of parameters and functions, I handled the basic execution flow of the code. For that most of the user behaviors could be exclusively recognized according to the protocol type in Application layer, my research mainly focused on the HTML format of video websites. To accurately identify and extract the user behavior information keyword, I made a survey on six dominating video website's HTML code and summarize its format into Regular Expression. The subsequent test showed this method with high sensitivity on the categories of videos, which are, combined with the timestamps, the legible indicators of user habits. The problem I am facing is the reorganization of packets and data stream after transmission in the network. It required an analysis on the special fields of IP packet header and TCP header, especially the Fragment Offset in IP and SYN/ACK segment in TCP and also the message protocol in HTTP. The solution I was working on is to allocate a big part of memory for initializing an array to sort the packets, but the efficiency is enormously declined. And I am still exploring the optimized method.

是否符合进度？On schedule as per GANTT chart?	[YES]

下一步：

Next steps:

Firstly, I am going to store user information and behavior-related factors (IP, URL, timestamp and accessing page content) into a MySQL database. Then I will conduct simple statistical analysis and data mining on user behavior details. After fully conversant with the data format, I plan to program a primary analysis front-end platform to generate user portrait, for querying every user's access behavior details and regional integrated situation. And I will also organize the results and write a part of the final paper.