北京邮电大学
本科毕业设计（论文）中期进展情况检查表
Mid Term Check Form

学院 School	International Schoo	专业 Programme	Telecommunications w	班级 Class	2012215104
学生姓名 Name	QIAN Cheng	学号 BUPT student no	2012212860	学号QM student no.	120721267
设计（论文）编号 Project No.	RN_2860				
设计（论文）题目 Project Title	User behavior analysis based on DPI(Deep Packet Inspection)				
题目分类 Scope	Research	Networks	Simulation		

主要内容：（毕业设计（论文）进展情况，字数一般不少于1000字）

Main body: The progress of the research on the project. Total number of words is no less than 1000.

目标任务 Targets set at project initiation	Deployment of NDPI platform; A programme to implement reorganization of packets and data stream; A database of user behavior information; A primary analysis front-end program; A part of final paper.
是否完成目标 Targets met? Yes/No	Yes
目前已完成任务 Finished Work	Until the midterm check session, I will have finished all the expected midterm task. 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. I have read <pcapreader-source code analysis>, <registration and maintenance of protocols> and <FTP deep packet inspection>. 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. Then I have stored part of user information and behavior-related factors (IP, URL, timestamp and accessing page content) into a database. I am managing to make a experimental demo of conducting simple statistical analysis and data mining on user behavior details.

<table>
<tr><td>尚需完成的任务
Work to do</td><td colspan="2">The expected date of finishing all the rest part of the whole project will be approximately on May 10th. After fully conversant with the data format, I plan to make a research on several data mining algorithms and select one that is most suitable for our data format and user behavior analysis system. Continuing work will be programming a primary analysis front-end platform to generate user portrait, for querying every user's access behavior details and regional integrated situation. The majority of the rest time will be spent on white/black box testing to ensure the robustness and accuracy of the system. Evaluation of the effectiveness and usability of the system will be carried out. I will compare the results of the analytic system with the actual user behaviors and habits to verify the reliability. Then I will get the conclusion of whether the deep packet inspection will give the expected result and the extent to be employed in user behavior analysis. After that I will real-time collect the packets of a period of time and display the result, which is for the final demo and viva. Finally, I will organize the results and write the entire final paper. Because of the characteristic of my project is research, I might provide some possible applications but I am not going to implement that. For those ideas, I will also write into the additional works in my final paper.</td></tr>
<tr><td></td><td>能否按期完成设计（论文）
Can finish the project on time or not:
Yes/No</td><td>Yes</td></tr>
<tr><td rowspan="2">存在问题和解决办法
Problems and Solutions</td><td>存在
问题
Problems</td><td>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. Another question is that what we can get from the header of HTML code is just the name and category (film or TV series) of the video, instead of the type (romantic or realistic), which make it difficult to detect the user's preference on video.</td></tr>
<tr><td>拟采取
的办法
Solutions</td><td>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. For the second problem, I pictured to connect the user behavior database with an external open source database filled with video details information to make a more comprehensive analytics.</td></tr>
<tr><td>最终论文结构
Structure of the final report</td><td colspan="2">Specification;
Abstract (Both in English and Chinese);A short overview of the whole part.
Keywords;
Table of contents;

CHAPTER 1:
Introduction;

CHAPTER 2:
Motivation and background;
Briefly describe the project;
Highlight the creative point and field;
Schedule of the project and show that I have met the aims stated in the specification.

CHAPTER 3:
Design and implementation;
Introdution of Deep Packet Inspecation technology and my design of User Behavior Analysis system;
The working process of DPI platform of version nDPI1.6 on Ubuntu system and how it real-time grasped the Internet packets and analyzed the protocol type in Data Link layer, Network layer, Transport layer and Application layer;
Principle of extract user behavior-related information from network packets;
A survey on six dominating video website' s HTML code and summarize its format into Regular Expression;
Several possible data mining algorithms suitable for user behavior analysis.
Proposed analysis model and scheme;
Implementation of the analysis system;

CHAPTER 4:
Results and discussion;
Evaluation of the effectiveness and usability of the system;</td></tr>
</table>

	Evaluation of the effectiveness and usability of the system; Comparation of the results of the analytic system with the real user behaviors and habits to verify the reliability; The problem I met and how I resolved; CHAPTER 5: Conclusion,a clear summary of the design and result; Further work, the discussion of extending applications and several solutions to improve the efficiency and automation of the system; References; Acknowledgements; Appendices; Some annotation and explanation important code fragments and functions; Risk assessment, the robustness and accuracy of the system, privacy infringement discussion; Environmental impact assessment. (Approximately 50 pages)
日期 Date	06/03/2016

Fill in the sub-tasks and select the cells to show the extent of each task

	Nov		Dec		Jan		Feb		Mar		Apr		May	
Task 1: Understand the principle of deep packet inspection														
Handle NDPI working structure;														
Deeply inspect all layers of the packet with NDPI source code in C;														
Task 2: Analysis of the open source code--NDPI														
Program to implement reorganization of packets and data stream;														
Grasp real-time Internet packets by NDPI and write condition to filter out the useless packets;														
Focus on research and analysis on HTTP message in application layer and gain the useful information for user behavior														
Task 3: Analysis of user behavior and habits based-on the captured packets														
Store user information and behavior-related factors (IP, URL, timestamp and accessing page content) into MySQL database;														
Conduct simple statistical analysis and data mining on user behavior details;														
Program front-end platform to generate user portrait, for querying every user's access behavior details and regional integrated situation;														
Organized results and write paper.														
Task 4:														