哈爾濱工業大學
Harbin Institute of Technology

计算机网络
课程实验报告

实验名称	HTTP 代理服务器的设计与实现				
姓名	张瑞		院系	计算机科学与技术	
班级	1903104		学号	1190201421	
任课教师	刘亚维		指导教师	刘亚维	
实验地点	格物 207		实验时间	2021 年 10 月 31 日	
实验课表现	出勤、表现得分(10)		实验报告得分(40)		实验总分	
	操作结果得分(50)				
教师评语					

计算机科学与技术学院 SINCE 1956...
School of Computer Science and Technology

实验目的：

熟悉并掌握 Socket 网络编程的过程与技术；
深入理解 HTTP 协议，掌握 HTTP 代理服务器的基本工作原理；
掌握 HTTP 代理服务器设计与编程实现的基本技能。

实验内容：

(1)设计并实现一个基本 HTTP 代理服务器。

要求在指定端口（例如 8080）接收来自客户的 HTTP 请求并且根据其中的 URL 地址访问该地址所指向的 HTTP 服务器（原服务器），接收 HTTP 服务器的响应报文，并将响应报文转发给对应的客户进行浏览。

(2)设计并实现一个支持 Cache 功能的 HTTP 代理服务器。

要求能缓存原服务器响应的对象，并能够通过修改请求报文（添加 if-modified-since 头行），向原服务器确认缓存对象是否是最新版本。

(3)扩展 HTTP 代理服务器，支持如下功能：

a)网站过滤：允许/不允许访问某些网站；
b)用户过滤：支持/不支持某些用户访问外部网站；
c)网站引导：将用户对某个网站的访问引导至一个模拟网站（钓鱼）。

实验过程：

(1)Socket编程的客户端和服务器端主要步骤

1)客户端：

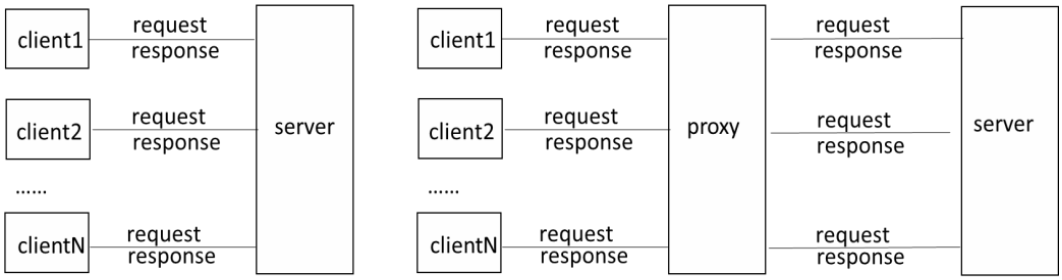
a)创建一个套接字并与服务器连接
b)发送请求报文
c)接收响应报文
d)关闭套接字

2)服务器端：

a)创建一个套接字
b)将套接字与端口绑定
c)监听端口
d) 取出监听的套接字等待队列中的第一个连接请求，为每个请求创建一个新的套接字，并创建一个新的线程来处理其请求
e)接收请求报文
f)发送响应报文
g)关闭该套接字，退出线程，继续监听端口，处理新的请求

(2)HTTP代理服务器的基本原理

代理服务器，俗称“翻墙软件”，允许一个网络终端（一般为客户端）通过这个服务与另一个网络终端（一般为服务器）进行非直接的连接。下图为普通Web应用通信方式与采用代理服务器的通信方式的对比。

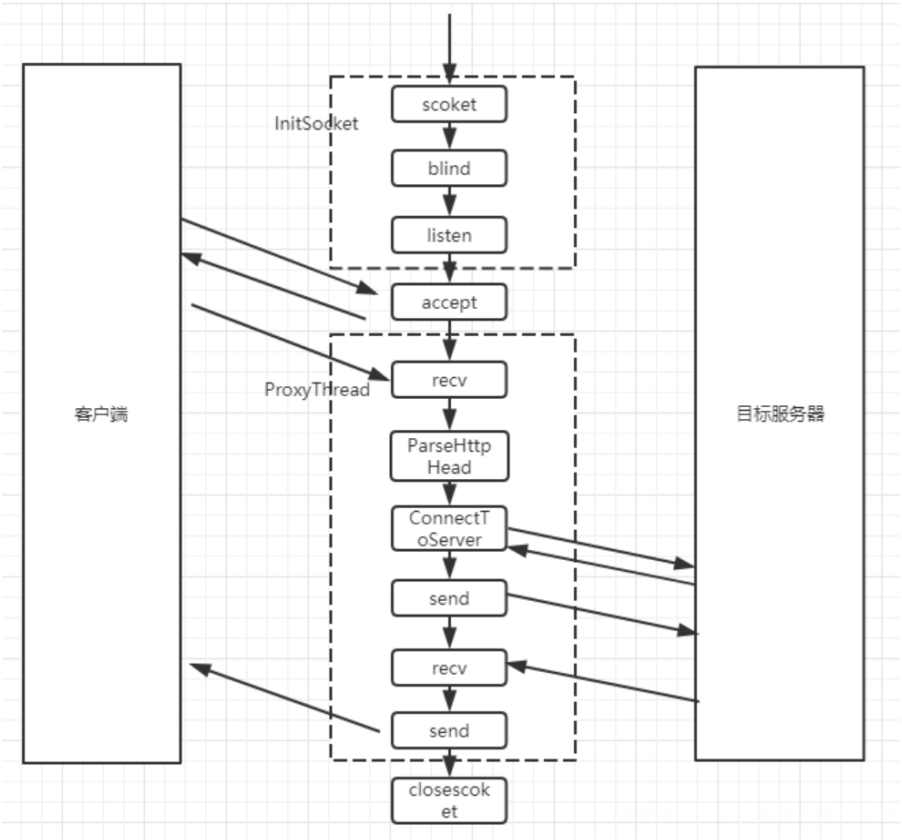


一般B/S　　使用代理的B/S

代理服务器在指定端口（例如8080）监听浏览器的访问请求（需要在客户端浏览器进行相应的设置），接收到浏览器对远程网站的浏览请求时，代理服务器开始在代理服务器的缓存中检索URL对应的对象（网页、图像等对象），找到对象文件后，提取该对象文件的最新被修改时间；代理服务器程序在客户的请求报文首部插入<If-Modified-Since: 对象文件的最新被修改时间>，并向原Web服务器转发修改后的请求报文。如果代理服务器没有该对象的缓存，则会直接向原服务器转发请求报文，并将原服务器返回的响应直接转发给客户端，同时将对象缓存到代理服务器中。代理服务器程序会根据缓存的时间、大小和提取记录等对缓存进行清理。

(3)HTTP代理服务器的程序流程

运用Socket编程实现代理服务器时，具体流程如下图所示：



(4)实现HTTP代理服务器的关键技术及解决方案

1)设计并实现一个基本HTTP代理服务器：

这一部分主要借鉴了实验指导书上参考代码的设计思路，修改个别细节后即可实现，下面还是简单介绍下这一部分的设计与实现过程。

a)初始化套接字，利用bind()函数将该套接字与服务器地址绑定，同时，监听端口按实验指导书上的要求设置为10240，然后利用listen()函数在该端口进行监听。

b)用accept()函数取出监听的套接字等待队列中的第一个连接请求，为每个请求创建一个新的套接字，并创建一个新的线程来处理其请求。

c)用recv()函数接收来自客户端的HTTP请求；用ParseHttpHead()函数解析HTTP头部；用ConnectToServer()函数创建套接字，并连接目标服务器；用send()函数将客户端的请求转发给服务器；用recv()函数接收服务器的响应报文；最后再用send()函数将该响应发送给客户端。

d)完成上述处理后，代理服务器等待一段时间后关闭套接字，并退出为该请求创建的线程。

2) 设计并实现一个支持Cache功能的HTTP代理服务器:

为了实现这个功能，我引入了新的结构体数组cache来存储缓存内容，如下所示：

```
//缓存数据
struct Cache
{
    char url[1024];  //url地址
    char time[40];  //上次更新时间
    char buffer[MAXSIZE];  //缓存的内容
    Cache()
    {
        ZeroMemory(this, sizeof(Cache));
    }
} cache[CACHE_NUM];
```

a)当客户端提出访问请求时，代理服务器查找cache中是否有该请求的响应：如果没有，则后续流程同基本HTTP代理服务器；如果有，则修改客户端的请求，向其中增加“If-Modified-Since”字段，并将时间设为在cache中查找到的time，再将修改后的请求转发给服务器。

b)代理服务器提取出从服务器返回的响应报文中的状态码：若状态码为304，说明网站尚未更新，代理服务器直接将本地缓存发送给客户端；若状态码为200，说明网站已更新，代理服务器将该响应发回给客户端，并更新本地缓存。

3)扩展HTTP代理服务器，支持如下功能：

a)网站过滤：先创建一个invalid_websites.txt，其中存放禁止访问的网站。对于客户端发来的HTTP请求，若检测到其中URL为invalid_websites.txt中的任意一个，代码直接跳转到结束部分，拒绝连接，关闭套接字，退出线程。

b)用户过滤：先创建一个invalid_users.txt，其中存放禁止访问外网的用户。用accept()函数取出监听的套接字等待队列中的第一个连接请求时，检查发出该请求的客户端的IP地址，若检测到其与invalid_users.txt中的任意一个，则拒绝该连接请求，继续从等待队列中的取出下一个连接请求。

c)网站引导：解析完HTTP头部之后，若检测到其中URL为钓鱼的源网站，代理服务器修改HTTP头部中的URL和Host字段为钓鱼目的网站URL和Host，并且修改客户端请求报文中的URL和Host后转发给服务器，后续流程同增加网站过滤功能的HTTP代理服务器。

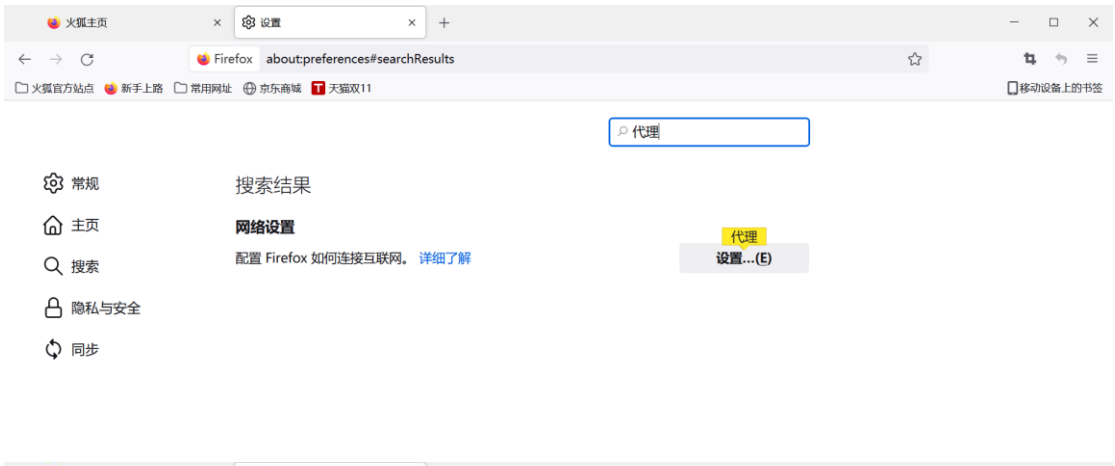
(5)HTTP代理服务器源代码（带有详细注释）

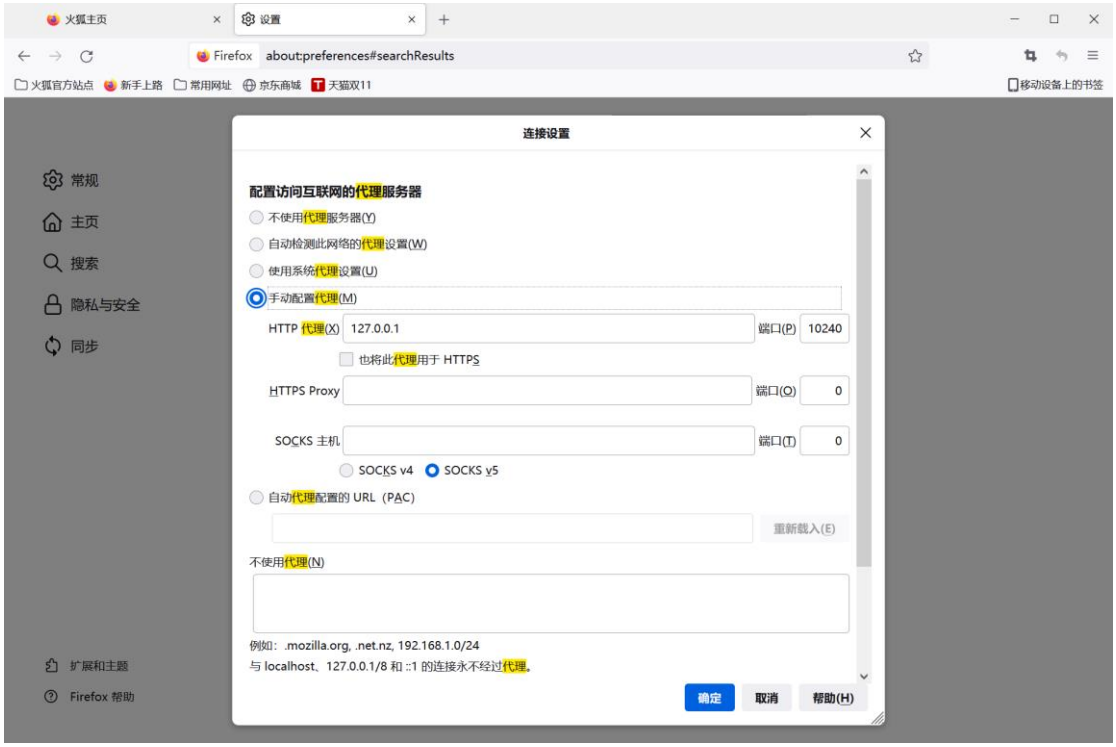
详见随报告一同上传的源代码文件夹

实验结果：

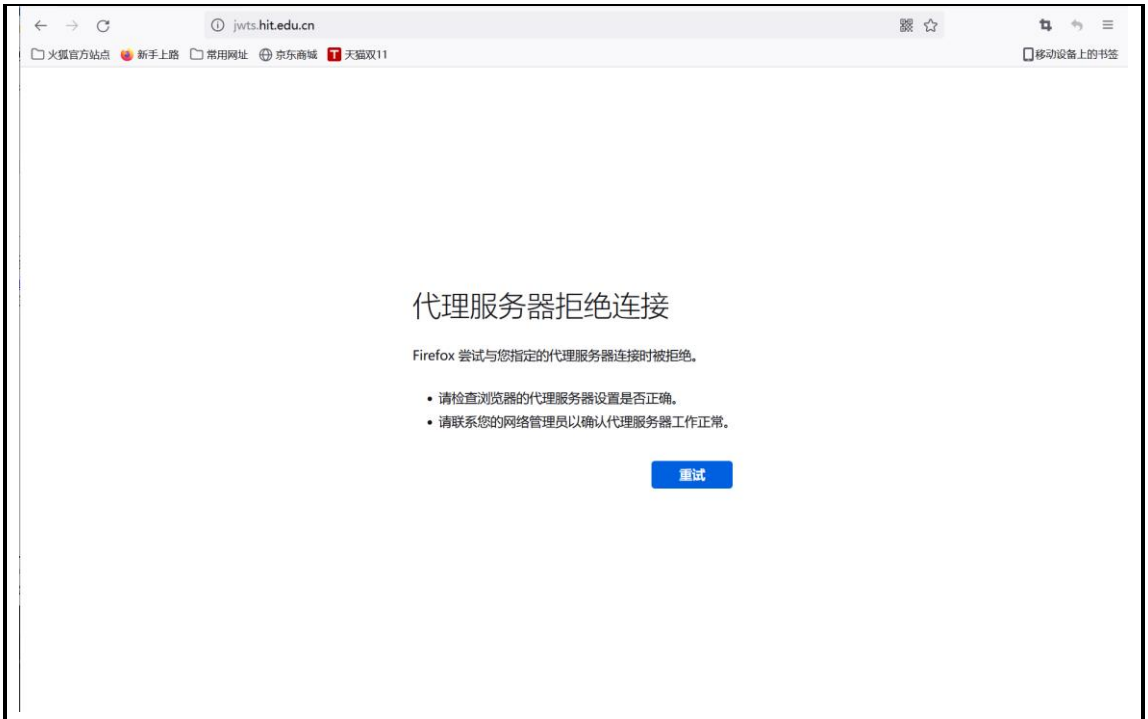
(1)实验前浏览器代理设置：

打开火狐浏览器，在设置中搜索“代理”，按实验指导书要求配置地址为本机IP地址127.0.0.1，端口为10240。





设置完成后，若我们访问哈工大教务处的网站http://jwts.hit.edu.cn，会发现浏览器显示代理服务器拒绝连接（我们尚未运行程序，打开代理服务器），这说明代理服务器设置成功。



(2)基本HTTP代理服务器功能验证：

打开CodeBlocks运行程序，再访问哈工大教务处网站，我们可以看到此时网站能成功加载出来，且控制台输出相关信息，说明代理服务器的基本功能正常实现。

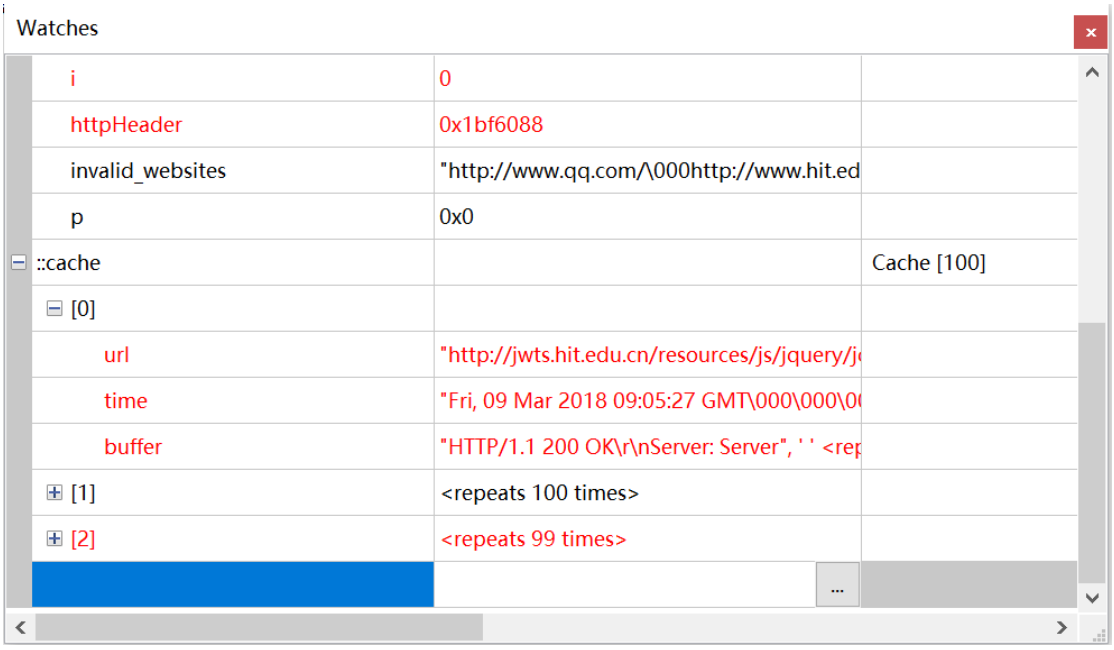

```
代理服务器正在启动
初始化...
代理服务器正在运行，监听端口10240
GET http://jwts.hit.edu.cn/ HTTP/1.1
客户端请求访问的URL是：http://jwts.hit.edu.cn/
代理连接主机jwts.hit.edu.cn成功
关闭套接字
--------------------------------------------------

GET http://jwts.hit.edu.cn/resources/js/jquery/jquery-1.7.2.min.js HTTP/1.1
客户端请求访问的URL是：http://jwts.hit.edu.cn/resources/js/jquery/jquery-1.7.2.min.js
代理连接主机jwts.hit.edu.cn成功
成功添加缓存
关闭套接字
--------------------------------------------------
```

(3)HTTP代理服务器缓存功能验证：

可以看到刚才访问教务处网站时，控制台已经输出“成功添加缓存”的信息，这是因为我们加入了缓存功能，当我们紧接着再次访问教务处网站时，控制台告知我们缓存命中，并且缓存信息未过期，代理服务器直接将缓存内容返回给我们。

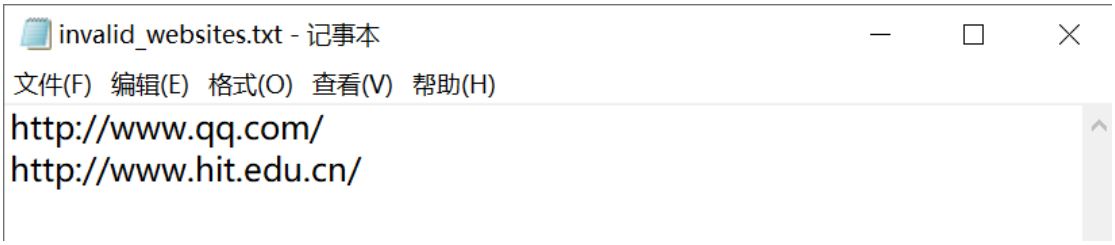
```
----------------------------------------------


GET http://jwts.hit.edu.cn/ HTTP/1.1
客户端请求访问的URL是：http://jwts.hit.edu.cn/
代理连接主机jwts.hit.edu.cn成功
关闭套接字
----------------------------------------------


GET http://jwts.hit.edu.cn/resources/js/jquery/jquery-1.7.2.min.js HTTP/1.1
客户端请求访问的URL是：http://jwts.hit.edu.cn/resources/js/jquery/jquery-1.7.2.min.js
代理连接主机jwts.hit.edu.cn成功
缓存命中，正在验证网站是否修改过...
网站未修改过，将由缓存提供网站信息
成功返回缓存
关闭套接字
----------------------------------------------
```

同时，通过设置断点查看变量值的方法，我们能看到cache数组中确实存入了我们所需要的网站信息，说明代理服务器的缓存功能正常实现。



(4)HTTP代理服务器网站过滤功能验证：

首先我们将禁止访问的网站写入invalid_websites.txt。



然后在浏览器中输入http://www.qq.com/，会发现无法访问页面。

同时控制台输出信息提醒网站被屏蔽。

```
GET http://www.qq.com/ HTTP/1.1
客户端请求访问的URL是：http://www.qq.com/
该网站已被屏蔽
关闭套接字
----------------------------------------
```

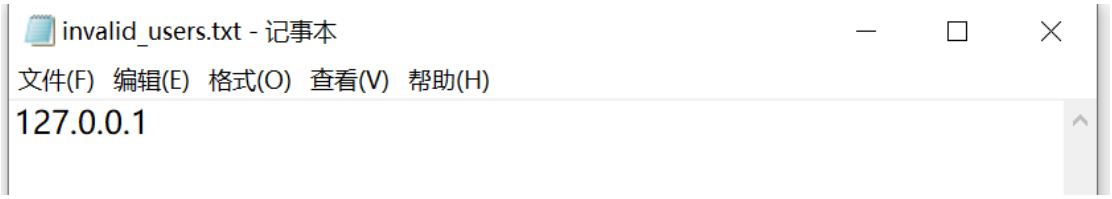
再输入http://www.hit.edu.cn/，发现是一样的情况。

```
GET http://www.hit.edu.cn/ HTTP/1.1
客户端请求访问的URL是：http://www.hit.edu.cn/
该网站已被屏蔽
关闭套接字
----------------------------------------
```

但与此同时，其余网站仍可正常访问，说明代理服务器的网站过滤功能正常实现。

(5)HTTP代理服务器用户过滤功能验证：

首先我们将禁止访问外网的用户（本机127.0.0.1）写入invalid_users.txt。



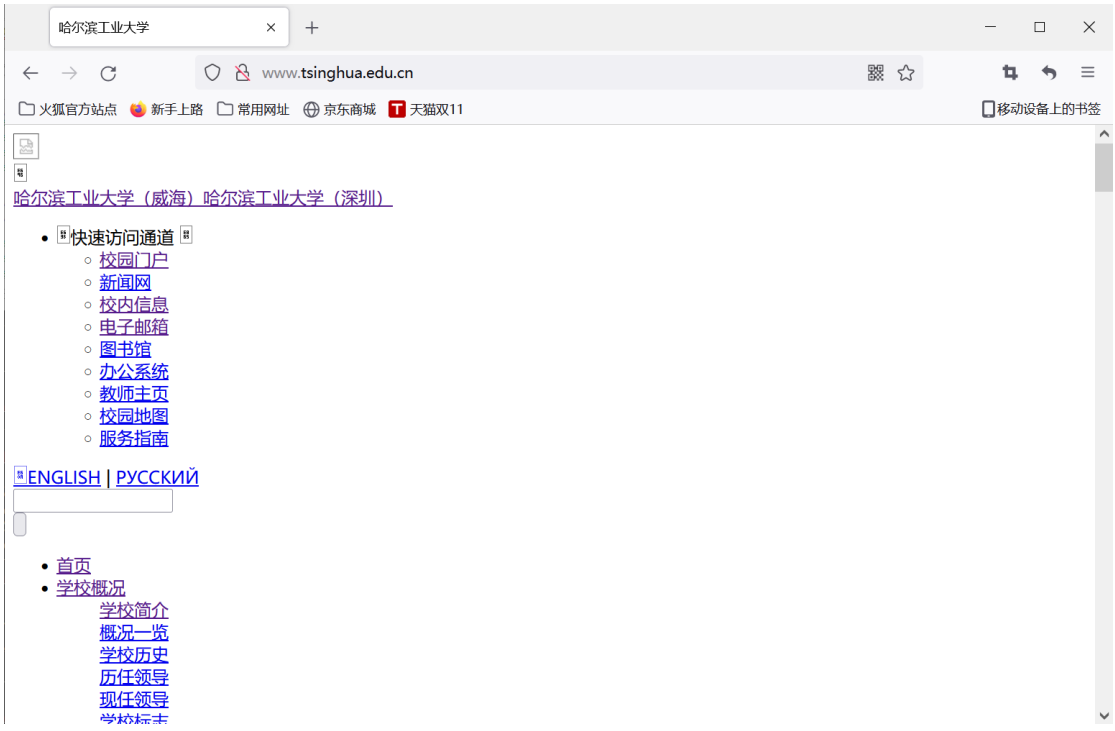
此时访问教务处网站，发现已被禁止，控制台输出相关信息。

```
代理服务器正在启动
初始化...
代理服务器正在运行，监听端口10240
用户访问受限
```

若将invalid_users.txt中的用户改为192.0.0.1，由于该地址不是本机，所以此时的浏览器又可正常访问网站了，说明代理服务器的用户过滤功能正常实现。

(6)HTTP代理服务器网站引导功能验证：

首先将钓鱼的源网站设为清华大学官网http://www.tsinghua.edu.cn/，然后将钓鱼的目的网站设为哈工大官网http://www.hit.edu.cn/，当我们在浏览器中请求访问清华官网时，发现网站跳转到了哈工大官网，同时控制台输出相关信息，说明代理服务器的网站引导功能正常实现。



```
GET http://www.tsinghua.edu.cn/ HTTP/1.1
客户端请求访问的URL是：http://www.tsinghua.edu.cn/
修改后的HTTP报文如下：
GET http://www.hit.edu.cn/ HTTP/1.1
Host: www.hit.edu.cn
User-Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:93.0) Gecko/20100101 Firefox/93.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,*/*;q=0.8
Accept-Language: zh-CN,zh;q=0.8,zh-TW;q=0.7,zh-HK;q=0.5,en-US;q=0.3,en;q=0.2
Accept-Encoding: gzip, deflate
Connection: keep-alive
Upgrade-Insecure-Requests: 1


成功从源网站：http://www.tsinghua.edu.cn/转到目的网站：http://www.hit.edu.cn/
代理连接主机www.hit.edu.cn成功
关闭套接字
------------------------------------------
```

问题讨论：

(1)pragma comment(lib,"Ws2_32.lib")命令是静态链接Ws2_32.lib库，但Codeblocks使用的MingGW不支持该写法，可以在设置里手动加上libws2_32.a

(2)goto语句后面，不能再新定义变量，需要把变量定义全部放到goto之前。

(3)当设置完代理服务器后，发现使用HTTPS的请求仍可以被正确响应，这是由于我们实现的是HTTP代理服务器，对HTTPS报文无法处理。查阅资料我们发现两者其实属于不同的协议：HTTP是一种用于分布式、协作式和超媒体信息系统的应用层协议，以明文方式发送内容，不提供任何方式的数据加密，安全性低；而HTTPS是一种透过计算机网络进行安全通信的传输协议，它经由HTTP进行通信，但利用SSL/TLS来加密数据包，更加安全。仔细观察就能发现，我们用HTTP访问网站时，网址栏前方的图标也会提醒我们连接不安全。

(4)当我们进行网站引导，若在修改客户端的HTTP请求报文时，只能简单将URL和Host设置为钓鱼目的网址的URL和Host，则可能会出现图片和链接加载不全等问题。这一问题应该能通过代理服务器事先缓存钓鱼目的网址的所有相关信息而解决，当检测到客户端请求访问钓鱼源网站时，直接传回已缓存好的目的网站信息。

心得体会：

通过本次实验，我对socket编程有了初步的了解并掌握了相关编程技能，深刻地理解了HTTP代理服务器的基本原理与程序流程，同时我还学习了到了线程、监听、阻塞/非阻塞模式方法等知识，还在读取与解析HTTP头部的过程中更加深入地理解了其格式和内涵。在自己动手“搭梯子”，实现网站过滤、用户过滤、网站引导和缓存等功能的过程中感受到了快乐与成就感。