



哈尔滨工业大学  
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	王科龙		院系	计算机学院		
班级	1803104		学号	1180801203		
任课教师	聂兰顺		指导教师	聂兰顺		
实验地点	致知 207		实验时间	2020/10/31		
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						

**实验目的：**

本次实验的主要目的。

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：**

概述本次实验的主要内容，包含的实验项等。

- (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。（选作内容，加分项目，可以当堂完成或课下完成）
- (3) 扩展HTTP 代理服务器，支持如下功能：（选作内容，加分项目，可以当堂完成或课下完成）
  - 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：**

以文字描述、实验结果截图等形式阐述实验过程，必要时可附相应的代码截图或以附件形式提交。

基本的代理服务器功能实现：

1. 服务器监听客户端的请求，获取和客户端连接的套接字
2. 新建线程处理
3. 读取客户端的http请求
4. 解析http请求，获取host和port和method和url
5. 构造发送给远程服务器的请求
6. 新建和远程服务器的套接字
7. 向远程服务器发送请求
8. 在远程服务器和客户端之间转发数据

扩展功能

1. cache

第一次访问时，缓存文件不存在，在读取到客户端的请求之后，利用url创造缓存文件，在转发服务器和客户端之间的请求时，同时向该文件内写响应信息。

第二次访问时，缓存文件存在，在读取到客户端的请求之后，利用url构造相同的文件名查找文件，在文件中读取到Last-Modified信息，将后面的时间信息作为自己要获取的If-Modified-Since信息，将其加入发送到服务器的请求报文中。

在将请求报文发送到服务器后，服务器机器上的文件的Last-Modified没有迟于If-Modified-Since时间，服务器会返回给代理服务器304，不携带数据。然后代理服务器获取的304之后，使用自己缓存的文件，将之前保存的响应信息返回给客户端。

2. 禁止主机访问/禁止访问网站/钓鱼

禁止某些主机的访问需要对请求信息的host进行分析，如果这个主机在被禁止的列表内，

向客户端返回"HTTP/1.1 403 Forbidden\r\n\r\n", 浏览器就会解析这条报文并提示403错误。

禁止访问某些网站, 需要对访问的url进行禁止, 解析到被禁止的url之后, 同样向客户端返回"HTTP/1.1 403 Forbidden\r\n\r\n"即可。

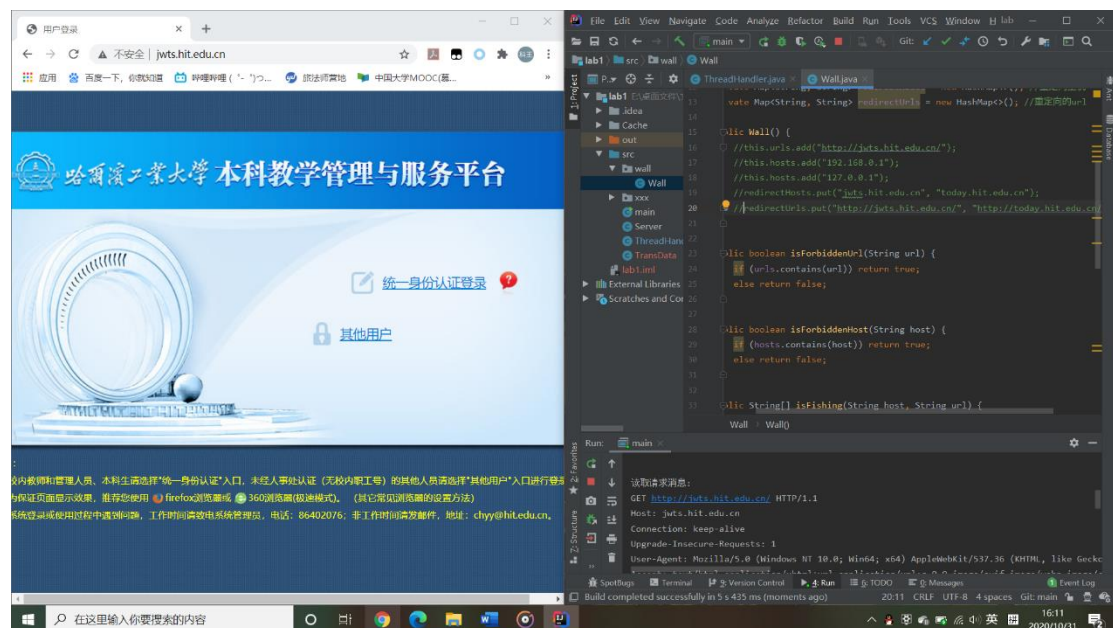
钓鱼, 钓鱼的列表有2个, 一个是对应的主机, 一个是对应的url, 如果主机和url都在钓鱼列表内, 那么就将请求报文里的主机和url都进行替换, 让服务器返回钓鱼网站的响应信息。

## 实验结果:

采用演示截图、文字说明等方式, 给出本次实验的实验结果。

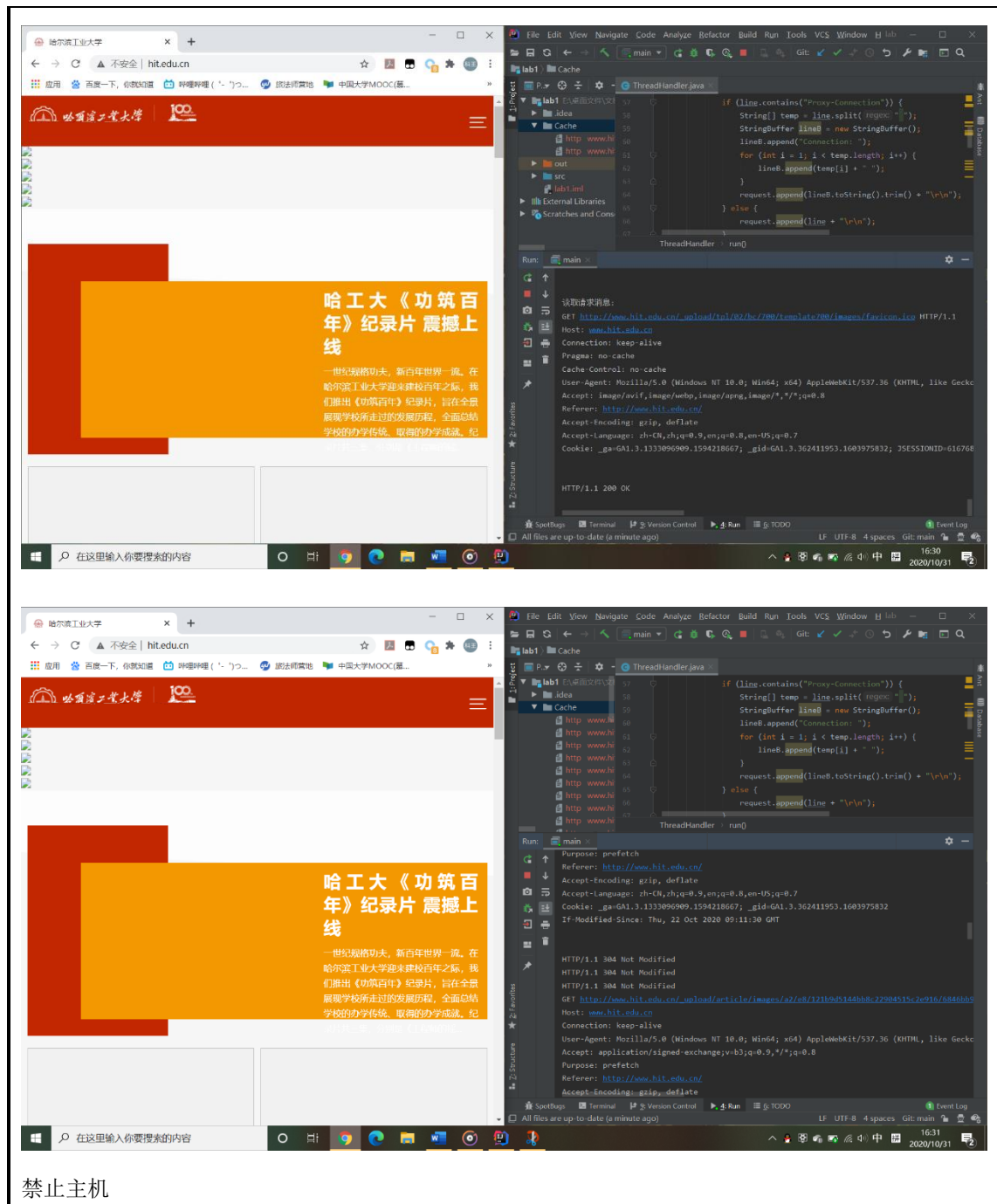
### 基本功能

将代理服务器设置为本机ip地址127.0.0.1, 访问jwts.hit.edu.cn

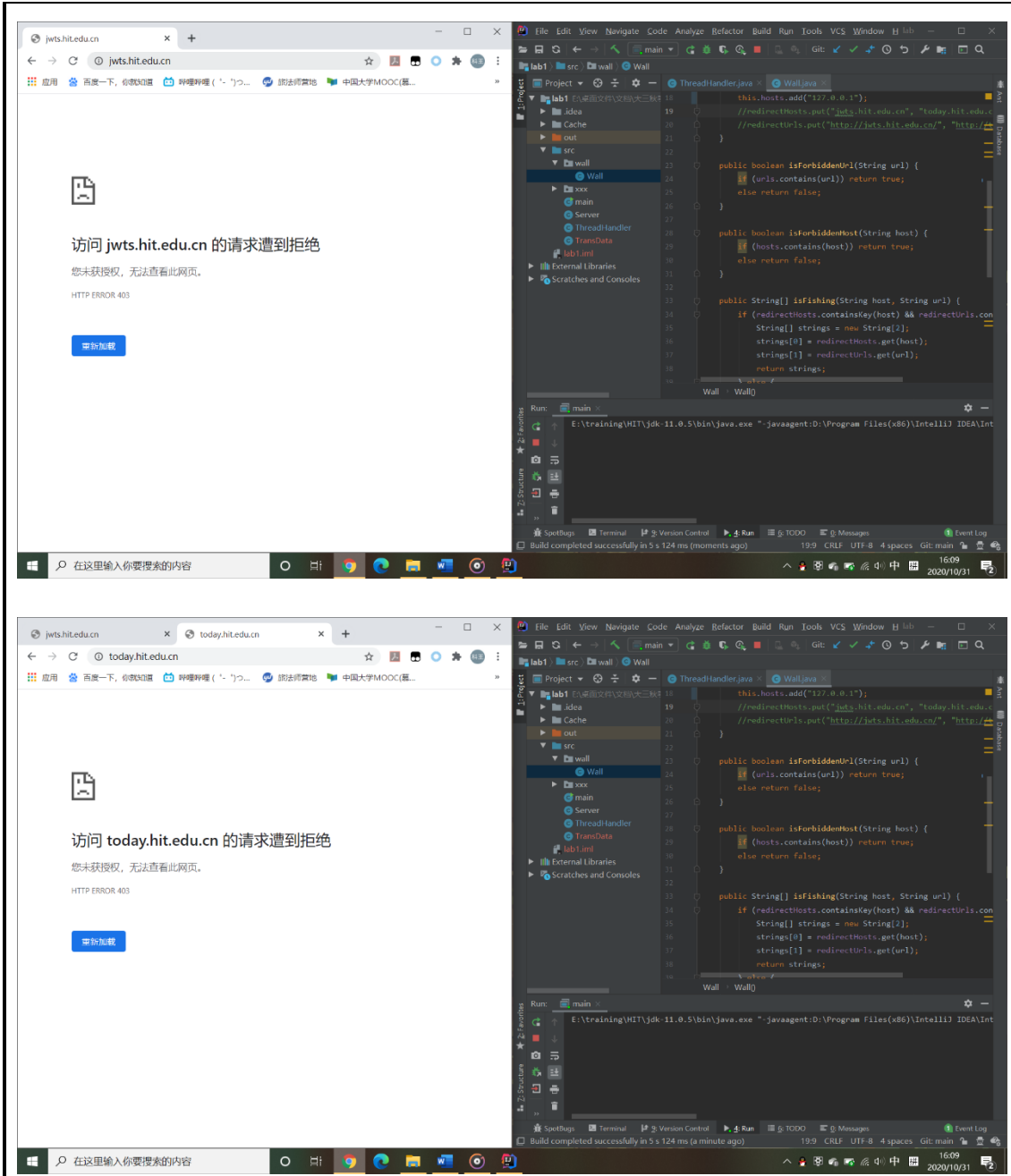


### 使用缓存功能

演示2次访问today.hit.edu.cn的效果



禁止主机



The image displays two screenshots of a web browser and an IDE (IntelliJ IDEA) showing a Java-based firewall implementation.

**Top Screenshot:**

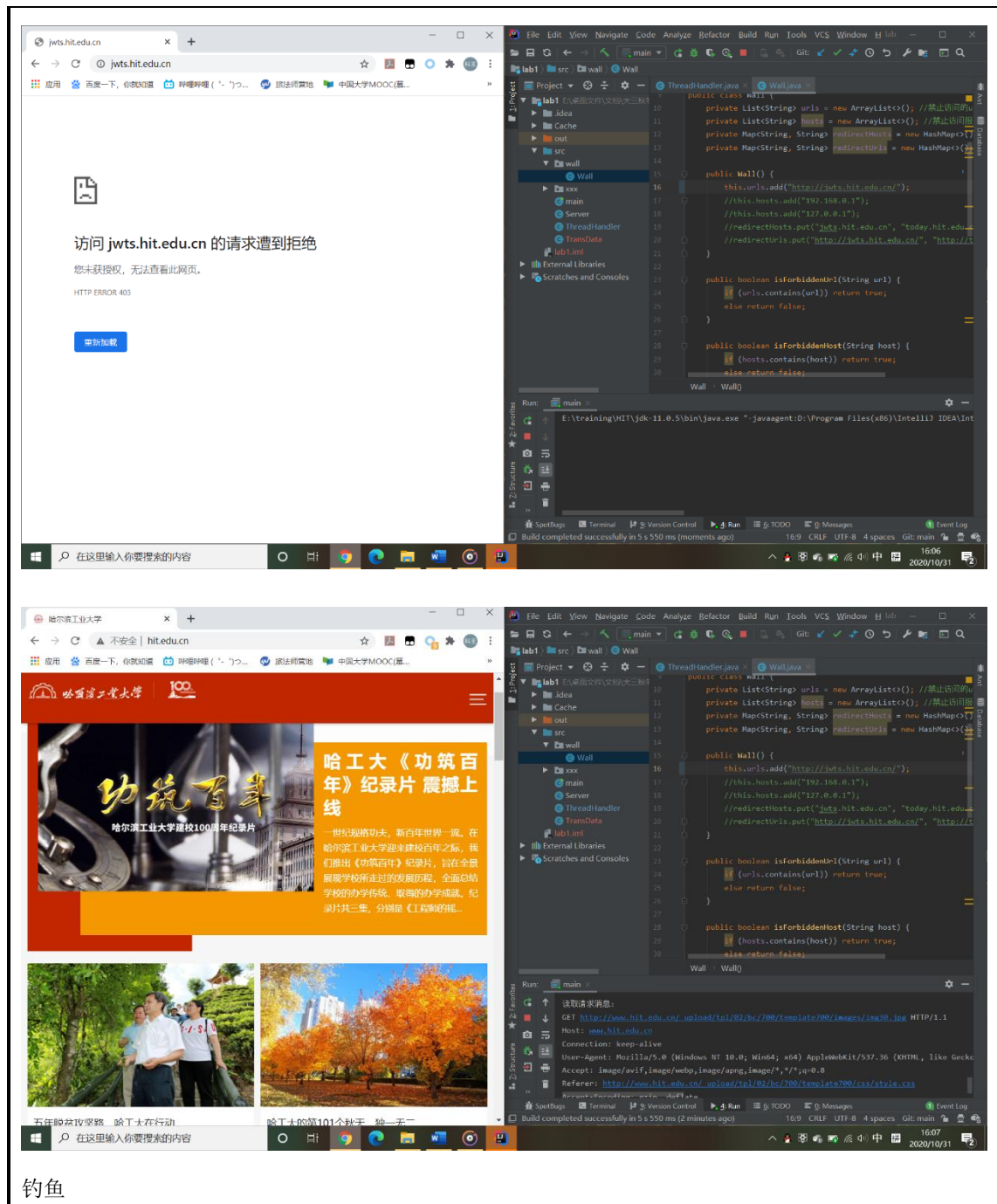
- Browser:** The address bar shows `jwts.hit.edu.cn`. The page content indicates a request to `jwts.hit.edu.cn` was denied, with the message "访问 jwts.hit.edu.cn 的请求遭到拒绝" (Request to jwts.hit.edu.cn is denied) and "HTTP ERROR 403".
- IDE:** The code editor shows the `Wall.java` file. The code defines a `ThreadTandler` class (note the typo) with methods for handling requests and checking if a URL or host is forbidden. The `isForbiddenUrl` method checks if the URL contains `"jwts.hit.edu.cn"`. The `isForbiddenHost` method checks if the host contains `"jwts.hit.edu.cn"`. The `isFishing` method checks if the host and URL are in a list of fishing sites.

**Bottom Screenshot:**

- Browser:** The address bar shows `today.hit.edu.cn`. The page content indicates a request to `today.hit.edu.cn` was denied, with the message "访问 today.hit.edu.cn 的请求遭到拒绝" (Request to today.hit.edu.cn is denied) and "HTTP ERROR 403".
- IDE:** The code editor shows the same `Wall.java` file. The `isForbiddenUrl` method now checks if the URL contains `"today.hit.edu.cn"`. The `isForbiddenHost` method checks if the host contains `"today.hit.edu.cn"`. The `isFishing` method checks if the host and URL are in a list of fishing sites.

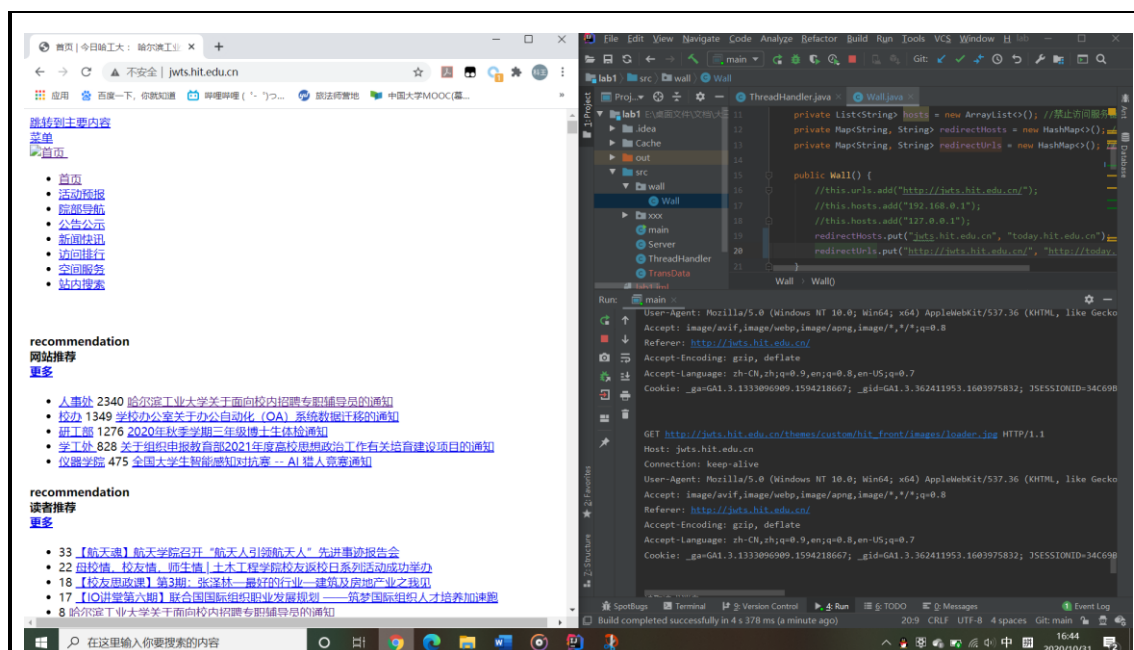
**禁止访问一些网站**

禁止访问 `jwts.hit.edu.cn`, 但是可以访问 [www.hit.edu.cn](http://www.hit.edu.cn)



钓鱼





网站钓鱼时，我们需要对伪造网站了如指掌，我们才有可能呈现出真正的伪造网站，但是在这个实验中，我使用的伪造网址都是学校的网站，我只对一个html文件进行了重定向和修改，因此某些图片是有可能出现404响应消息的，因为这些都是在原主机上寻找钓鱼网站的资源，自然会找不到文件。

## 问题讨论：

对实验过程中的思考问题进行讨论或回答。

1.有时会出现Bad Request响应。

这些都是由于报文格式不正确引起的，实验过程中需要对http报文进行解析，http请求报文和响应报文的格式非常固定，多一个空格少一个空格都会影响网页的呈现效果，严重了服务器就无法解析。

2.在服务器和客户端之间中转数据时，在Java中推荐使用字节流，而不要使用字符流。因为有些数据是图片等等非字符数据，由于字符流一次读取布置一个字节，因此很有可能在最后出错。

3.在本次实验中，每次读取请求报文时，都会控制在读到空行时跳出，对于GET来说，是没有请求实体的，但是POST方法是有可能有请求实体的，里面可能存放登录的账号，密码等等信息，由于没有读取，这也就是实验用的程序无法进行网页登录的原因。

## 心得体会：

结合实验过程和结果给出实验的体会和收获。

网络编程中，为了在服务器和客户端之间联通，要打通的障碍非常多，对于除了http来说，除了基本的请求-响应之外，还有为了保证安全的协议https，而https是消息中间劫持人(代理服务)的克星，代理服务没有证书私钥，完全无法代替客户端和服务器通信，也没法读取数据，只能做盲目的转发。

代码不在报告中提供，详见附件。