



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	姚舜宇		院系	计算学部		
班级	1903602		学号	1190202107		
任课教师	李全龙		指导教师	李全龙		
实验地点	格物 207		实验时间	2021.10.30		
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：

熟悉并掌握 socket 编程的过程与技术，深入理解 HTTP 协议，理解代理服务器的工作原理，编程实现代理服务器。熟悉 cache 缓存技术，用于加速访问。

实验内容：

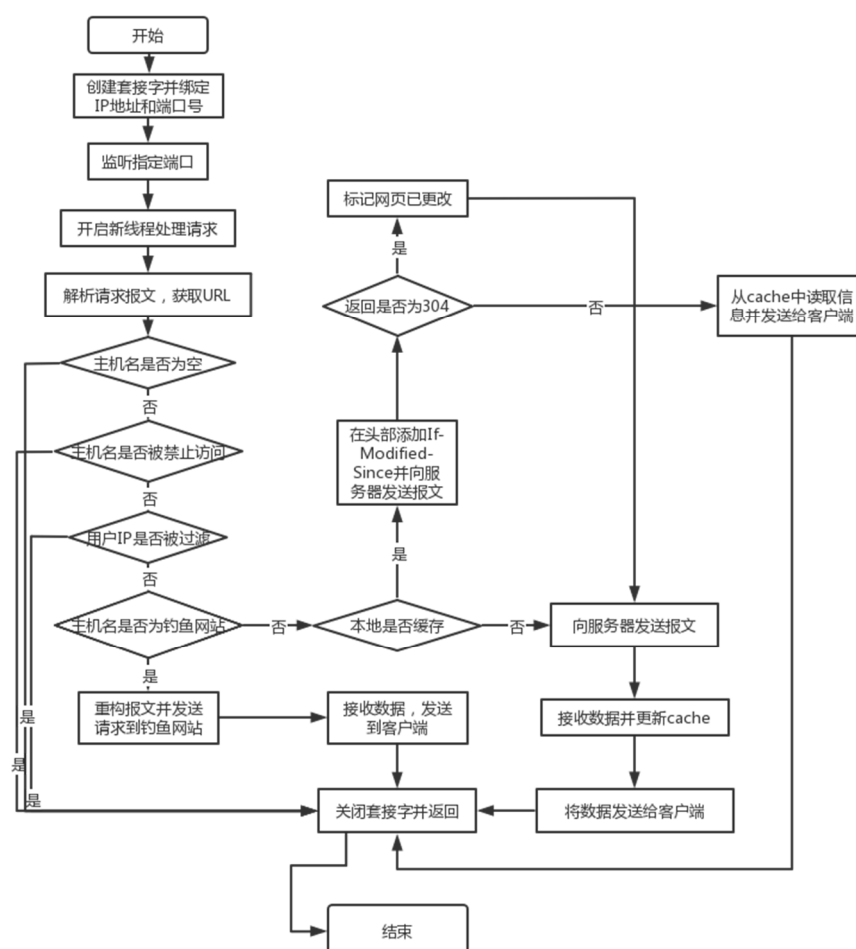
1. 设计并实现一个基本的HTTP代理服务器。在指定端口接收来自客户的HTTP请求并且根据其中的URL地址访问该地址指向的HTTP服务器，接收HTTP服务器的响应报文并转发给对应的客户进行浏览。
2. 在1的基础上增加cache功能，能够缓存原服务器响应的对象，并能够通过修改请求报文向原服务器确认缓存对象是否是最新版本。
3. 扩展HTTP代理服务器，使其能够支持网站过滤、用户过滤、钓鱼网站的功能。

实验过程：

一．代理服务器原理

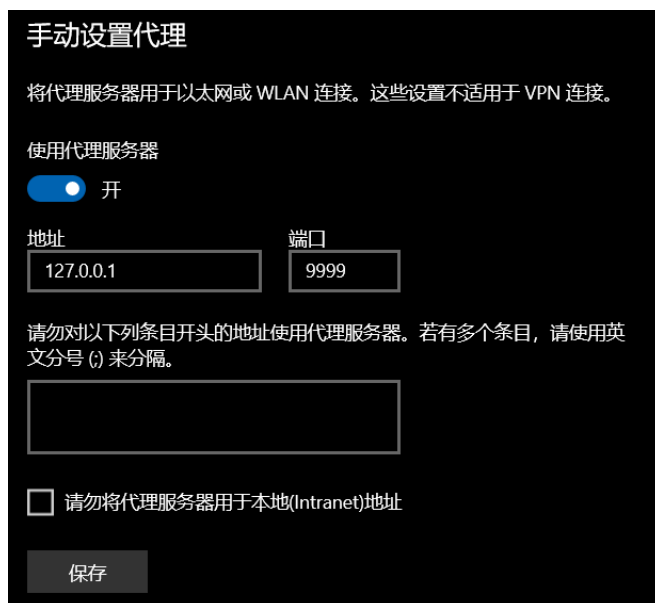
代理服务器是网络的一个中间实体，同时扮演服务器和客户端的角色。对于客户端，可以讲请求发送到代理服务器，代理服务器再发送请求到目标服务器，最后将结果返回。代理服务器起到了中间桥梁的作用，可以根据用户的要求增加功能。如cache缓存功能，可以大大加速访问的速度。也可以进行网站、用户过滤，以及钓鱼网站等等。

本实验中代理服务器的流程图：



二．代理服务器的配置

在设置界面里选择手动设置代理，适用代理服务器，并设置好地址和端口号，保存。



The screenshot shows the 'Manual Setup Proxy' (手动设置代理) window. It includes a title bar, a subtitle '将代理服务器用于以太网或 WLAN 连接。这些设置不适用于 VPN 连接。', a toggle switch for 'Use Proxy Server' (使用代理服务器) which is turned 'On' (开), and input fields for 'Address' (地址) set to '127.0.0.1' and 'Port' (端口) set to '9999'. Below these is a text box for additional addresses, a checkbox for 'Do not use proxy server for local (Intranet) addresses' (请勿将代理服务器用于本地(Intranet)地址), and a 'Save' (保存) button.

三. 本实验中代理服务器的实现过程

首先定义代理服务器的各个参数

```
# 代理服务器相关参数
PARAMETERS = {
    'HOST': '127.0.0.1',
    'PORT': 9999,
    'MAX_LISTEN': 50,
    'MAX_LENGTH': 4096,
    'CACHE_SIZE': 1000
}
```

初始化socket并且在循环中监听指定端口，当接收到客户端的请求则创建一个新线程进行处理。其中tcp_link是代理服务器的核心函数。

```
s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.bind((PARAMETERS['HOST'], PARAMETERS['PORT']))
s.listen(PARAMETERS['MAX_LISTEN'])
while True:
    sock, address = s.accept()
    print('address: ' + str(address))
    threading.Thread(target=tcp_link, args=(sock, address)).start()
```

接收来自客户端的http请求报文，对报文进行解码，获取请求行进行格式化，通过字符串的分析获得url地址，并且通过传入参数获得主机IP。

```
message = so.recv(PARAMETERS['MAX_LENGTH'])
if len(message) == 0:
    return
message = message.decode('utf-8', 'ignore') # 对报文进行解码，忽略错误
request_line = message.split("\r\n")[0].split() # 获得请求行，去掉前后空格
url = urlparse(request_line[1]) # 获得URL
hostIP = address[0] # 获得主机IP
```

接下来是完成附加功能的部分。

包括判断主机名是否允许被访问、用户IP是否被过滤、是否是钓鱼网站并且如何处理。如果主机名禁止访问或用户IP被过滤，则直接输出提示信息，然后关闭套接字并返回。如果是钓鱼网站，首先输出提示信息，然后根据给定的参数重构请求报文，发送到被引导到的网站服务器，从中接收数据，转发给客户端，之后关闭套接字并返回。

```
if url.hostname in fishing: # 主机名为钓鱼网站
    print('fishing from ' + str(url.hostname) + ' to ' + str(fishing[url.hostname]))
    new_hostname = fishing[url.hostname] # 新的目标主机名
    message = message.replace(request_line[1], 'http://' + new_hostname + '/')
    message = message.replace(url.hostname, new_hostname) # 将报文重构
    fish_socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
    fish_socket.connect((new_hostname, 80))
    fish_socket.sendall(message.encode())
    while True:
        buff = fish_socket.recv(PARAMETERS['MAX_LENGTH'])
        if not buff:
            fish_socket.close()
            break
        so.sendall(buff)
    so.close()
    fish_socket.close()
return
```

接下来完成cache的功能模块。

首先确定缓存路径和文件名，并初始化标记为未修改。

进行判断，如果本地已经存在该文件，说明以前访问过该网站并且进行了缓存，此时需要查看最后一次缓存之后网站内容是否有发生变化并且更新最后一次缓存的时间。通过向服务器发送报文，解析返回数据的方式进行判断。如果返回数据的响应码为304，则表示发生未变化，直接从本地文件中读取信息发送到客户端。如果响应码不为304，则表示发生了变化，直接将标记修改为已修改。

接下来进行判断，如果本地缓存中不存在该文件或者标记为已修改，则表示需要更新缓存。向服务器发送数据，获取服务器返回的数据，并且写入到缓存中，然后将数据转发给客户端，最后关闭套接字。

当本地缓存已经存在相应文件，判断网页是否被修改时的逻辑实现：

```
path = cache_dir + url.hostname # 缓存路径和文件名
modified = False # 第一次标记为未修改
if os.path.exists(path): # 当已经存在该文件，需要判断网页是否被修改
    modified_time = os.stat(path).st_mtime # 缓存文件最后修改的时间
    headers = str('If-Modified-Since: ' + time.strftime('%a, %d %b %Y %H:%M:%S GMT', time.gmtime(modified_time)))
    # 把modified-time按报文要求格式化
    message = message[:-2] + headers + '\r\n\r\n' # 把If-Modified-Since字段加入到请求报文中
    # 向服务器发送报文
    server_socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
```

```
server_socket.connect((url.hostname, 80))
server_socket.sendall(message.encode())
data = server_socket.recv(PARAMETERS['MAX_LENGTH']).decode('utf-8',
'ignore')
print(data)
server_socket.close()
if data[9:12] == '304': # 响应码为304，表示网页未变化，从cache中读取网页
    print('the web is not modified, read from cache.')
    with open(path, "rb") as f:
        so.sendall(f.read())
else: # 网页变化，标记为已修改
    modified = True
```

当没有该网页的缓存或者网页被修改时的逻辑实现：

```
if not os.path.exists(path) or modified: # 如果没有该网页的缓存或者网页已被修改
    # 向服务器发送数据，才能接收到服务器发回来的数据
    server_socket = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
    server_socket.connect((url.hostname, 80))
    server_socket.sendall(message.encode())
    print('Update cache.')
    f = open(path, 'wb') # 重写缓存
    while True:
        buff = server_socket.recv(PARAMETERS['MAX_LENGTH'])
        if not buff:
            print(buff)
            f.close()
            server_socket.close()
            break
        f.write(buff) # 将接收到的数据写入缓存
        so.sendall(buff) # 将接收到的数据转发给客户端
    so.close()
```

实验结果：

一. 正常状况

首先访问今日哈工大today.hit.edu.cn。

访问结果如下：

[←](#)
[→](#)
[×](#)
不安全 | today.hit.edu.cn

[Gmail](#)
[YouTube](#)

[跳转到主要内容](#)
[菜单](#)
[首页](#)

- [首页](#)
- [活动播报](#)
- [院部导航](#)
- [公告公示](#)
- [新闻快讯](#)
- [访问排行](#)
- [空间服务](#)
- [站内搜索](#)

recommendation

网站推荐

[更多](#)

- 基础学部 3103 [2021级大一年度项目立项委员会](#)
- 计算学部 2478 [【国家网络安全宣传周】计算学部网络安全空间安全学院关于举办“共话网络安全”学术论坛的通知](#)
- 人事处 2248 [黑龙江省外商企业咨询服务有限责任公司派遣到哈尔滨工业大学空间环境与物质科学研究院 \(国家大科学工程\) 工程师公开招聘公告](#)
- 化工学院 1603 [第九届“聚合杯”化学实验技能竞赛暨第一届哈尔滨工业大学化学知识及实验创新竞赛通知](#)
- 学工处 1590 [2020-2021年度本科生国家奖学金初审名单公示](#)

recommendation

读者推荐

[更多](#)

- 105 [【文化素质教育讲座】南北极科学探索——21 世纪人类面临的机遇和挑战](#)
- 47 [马克思主义学院与外国语学院成功举办教师羽毛球联谊赛](#)
- 44 [哈工大马克思主义学院“青椒论坛”揭牌暨首期讲座成功举办](#)
- 43 [王婉斌教授为哈工大一校三区马克思主义学院师生作辅导报告](#)
- 10 [马克思主义学院举办2021级研究生新生素质拓展团队训练](#)

notices

[←](#)
[→](#)
[↺](#)
不安全 | cs.hit.edu.cn

[Gmail](#)
[YouTube](#)



- [首页](#)
- [学部概况](#)
 - [学部介绍](#)
 - [组织机构](#)
 - [历任领导](#)
 - [现任领导](#)
 - [行政人员信息](#)
 - [行政人员信息](#)
 - [学部历史](#)
 - [主任寄语](#)
- [师资队伍](#)
 - [师资队伍](#)
 - [博士生导师](#)
 - [教学名师](#)
 - [兼职教授](#)
 - [硕士生导师](#)
- [教育教学](#)
 - [本科生培养](#)
 - [硕士生培养](#)
 - [博士生培养](#)
 - [教学成果奖](#)
 - [教学成果奖与国家级精品](#)
- [科学研究](#)
 - [科学研究概况](#)
 - [基地建设](#)
 - [学科设置](#)
 - [研究方向](#)
 - [科研成果](#)

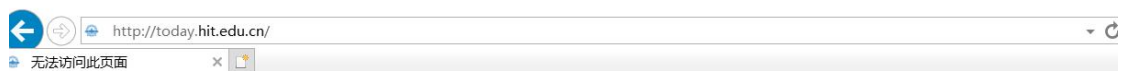
再次访问计算学部官网，发现响应码为304，代理服务器从缓存中读取信息。

```
the web is not modified, read from cache.
HTTP/1.1 304 Not Modified
Date: Mon, 25 Oct 2021 09:16:54 GMT
Server: Server
X-Frame-Options: SAMEORIGIN
Frame-Options: SAMEORIGIN
Last-Modified: Thu, 07 Sep 2017 01:15:34 GMT
ETag: "1793-5588f327a5980"
Accept-Ranges: bytes
Connection: close
```

二. 网站过滤

设置禁止访问的网址：今日哈工大。

```
# 禁止访问的url
No_Access_url = [
    'today.hit.edu.cn'
]
```



无法访问此页面

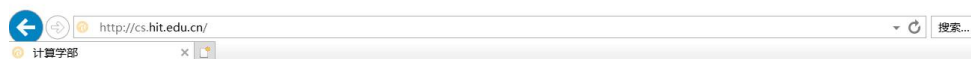
- 确保 Web 地址http://today.hit.edu.cn 正确
- [在必应上搜索此站点](#)
- [刷新页面](#)

[详细信息](#)

[修复连接问题](#)

```
address: ('127.0.0.1', 50632)
today.hit.edu.cn 127.0.0.1
ParseResult(scheme='http', netloc='today.hit.edu.cn', path='/', params='', query='', fragment='') is not accessed.
```

使用IE浏览器输入网址，发现无法访问，输出提示信息today.hit.edu.cn is not accessed。与此同时，计算学部官网cs.hit.edu.cn可以正常访问。



- [首页](#)
- [学部概况](#)
 - [学部介绍](#)
 - [组织机构](#)
 - [历任领导](#)
 - [现任领导](#)
 - [行政人员信息](#)
 - [行政人员信息](#)
 - [学部历史](#)
 - [主任寄语](#)
- [师资队伍](#)
 - [师资队伍](#)

三. 用户过滤

设置过滤用户IP:

```
# 是否过滤用户IP
# Blocked_User = False
Blocked_User = True
```

接下来访问今日哈工大和计算学部官网。发现都无法访问。



无法访问此页面

- 确保 Web 地址http://today.hit.edu.cn 正确
- [在必应上搜索此站点](#)
- [刷新页面](#)

[详细信息](#)

[修复连接问题](#)



无法访问此页面

- 确保 Web 地址http://cs.hit.edu.cn 正确
- [在必应上搜索此站点](#)
- [刷新页面](#)

[详细信息](#)

[修复连接问题](#)

提示信息显示: the user 127.0.0.1 is forbidden.

```
address: ('127.0.0.1', 50930)
address: ('127.0.0.1', 50934)
today.hit.edu.cn 127.0.0.1
the user 127.0.0.1 is forbidden.
address: ('127.0.0.1', 50933)
address: ('127.0.0.1', 50941)
cs.hit.edu.cn 127.0.0.1
the user 127.0.0.1 is forbidden.
address: ('127.0.0.1', 50940)
```

四. 网站引导

首先设置钓鱼网站, 将今日哈工大跳转到计算学部官网:


```
# 钓鱼网站
fishing = {
    'today.hit.edu.cn': 'cs.hit.edu.cn'
}
```

输入今日哈工大的网址，发现页面是计算学部官网的内容。



- [首页](#)
- [学部概况](#)
 - [学部介绍](#)
 - [组织机构](#)
 - [历任领导](#)
 - [现任领导](#)
 - [行政人员信息](#)
 - [行政人员信息](#)

```
today.hit.edu.cnaddress: ('127.0.0.1', 63351) 127.0.0.1
```

```
fishing from today.hit.edu.cn to cs.hit.edu.cn
```

```
today.hit.edu.cn 127.0.0.1
```

```
fishing from today.hit.edu.cn to cs.hit.edu.cn
```

输出的提示信息显示从today.hit.edu.cn钓到cs.hit.edu.cn。

五. Cache缓存

首先删除cache目录，保证未访问过任何网站：



访问计算学部官网。

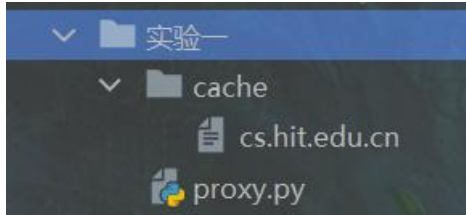


- [首页](#)
- [学部概况](#)
 - [学部介绍](#)

```
cs.hit.edu.cn 127.0.0.1
```

```
Update cache.
```

此时没有缓存，需要向原服务器发送请求并返回数据。提示信息显示更新缓存。更新之后目录如下：



更新后继续访问计算学部官网。

```
the web is not modified, read from cache.  
cs.hit.edu.cn 127.0.0.1
```

提示信息显示，网站内容没有修改，从cache中读取信息并将数据返回给客户端。

问题讨论：

在实现功能中的钓鱼网站时，我把请求的url部分全都换成导向网站的url，但发现这样只能加载出来网页的文字，而各种样式无法加载。后来把url中的主机名全部修改成为导向网站的主机名，而不是修改整个url，发现这样就可以加载出来图片样式等内容。

心得体会：

对socket编程的过程与技术有了一个初步的了解。
更加深入地理解了http协议和代理服务器的基本原理。
掌握了代理服务器的设计与编程的基本实现。