中国互联网审查研究

杨淑颖, 董一夫, 郑界涵

2010–2011 学年度

贡献

研究

杨淑颖　配合分析词汇封锁原因、进行总结

董一夫　分析词汇封锁原因、提供额外关键词列表

郑界涵　提出测试方法并进行实现、提取词库、协助分析数据

论文

本论文各部分由该部分负责的同学撰写、郑界涵排版完成。

目录

章节 1

选题原因

在日常生活中，我们发现在中国访问某些网站十分不方便，这不仅给我们的学习带来影响，更阻断了与国外朋友的一切联系。因此，为了解我国互联网现状，宣传普及互联网自由的精神，同时按照中国特色提出合适的建议，我们决定对此问题进行研究，并解决问题。

章节 2

计划

2.1 步骤

对于整个研究过程，我们分成了几个步骤:

- 研究有关法律条文
- 进行敏感词调查
- 测试访问全球 50 个最热门网站
- 开发突破封锁及用户管理系统
- 进行分析和总结

章节 3

公开法律条文研究

为进一步调查中国互联网审查，首先我们应该分析他的合理性与合法性。以下是我们找到的有关法律以及条文的内容。对于如下内容，我们的组员表示不发表任何评论。

3.1 《中华人民共和国互联网管理条例》

（2002 年 11 月 15 日生效）

第十四条 互联网上网服务营业场所经营单位和上网消费者不得利用互联网上网服务营业场所制作、下载、复制、查阅、发布、传播或者以其他方式使用含有下列内容的信息：

（一） 反对宪法确定的基本原则的；
（二） 危害国家统一、主权和领土完整的；
（三） 泄露国家秘密，危害国家安全或者损害国家荣誉和利益的；
（四） 煽动民族仇恨、民族歧视，破坏民族团结，或者侵害民族风俗、习惯的；
（五） 破坏国家宗教政策，宣扬邪教、迷信的；
（六） 散布谣言，扰乱社会秩序，破坏社会稳定的；
（七） 宣传淫秽、赌博、暴力或者教唆犯罪的；
（八） 侮辱或者诽谤他人，侵害他人合法权益的；
（九） 危害社会公德或者民族优秀文化传统的；
（十） 含有法律、行政法规禁止的其他内容的。

第三十七条 本条例自 2002 年 11 月 15 日起施行。2001 年 4 月 3 日信息产业部、公安部、文化部、国家工商行政管理局发布的《互联网上网服务营业场所管理办法》同时废止。

3.2 《互联网上网服务营业场所管理办法》

（2001 年 4 月 3 日生效，2002 年 11 月 15 日废止）

第十二条 互联网上网服务营业场所经营者和上网用户不得利用互联网上网服务营业场所制作、复制、查阅、发布、传播含有下列内容的信息:

（一） 反对宪法所确定的基本原则的;
（二） 危害国家安全，泄露国家秘密，颠覆国家政权，破坏国家统一的;
（三） 损害国家荣誉和利益的;
（四） 煽动民族仇恨、民族歧视，破坏民族团结的;
（五） 破坏国家宗教政策，宣扬邪教和愚昧迷信的;
（六） 散布谣言，扰乱社会秩序，破坏社会稳定的;
（七） 散布淫秽、色情、赌博、暴力、凶杀、恐怖或者教唆犯罪的;
（八） 侮辱或者诽谤他人，侵害他人合法权益的;
（九） 法律、行政法规禁止的其他内容。

3.3 《中华人民共和国宪法》

第一条 中华人民共和国是工人阶级领导的、以工农联盟为基础的人民民主专政的社会主义国家。

社会主义制度是中华人民共和国的根本制度。禁止任何组织或者个人破坏社会主义制度。

第三十五条 中华人民共和国公民有言论、出版、集会、结社、游行、示威的自由。

第四十条 中华人民共和国公民的通信自由和通信秘密受法律的保护。除因国家安全或者追查刑事犯罪的需要，由公安机关或者检察机关依照法律规定的程序对通信进行检查外，任何组织或者个人不得以任何理由侵犯公民的通信自由和通信秘密。

3.4 《世界人权宣言》

（1948 年 12 月 10 日联合国大会通过，中国已签署）

第十九条 人人有权享有主张和发表意见的自由；此项权利包括持有主张而不受干涉的自由，和通过任何媒介和不论国界寻求、接受和传递消息和思想的自由。

章节 4

测试封锁情况

4.1 关键词搜索测试

为了调查我国网络审查机制的审查重点，以及它对人们，特别是我们这个年龄的中学生生活的影响，我们决定进行关键词分析。一开始，有很多种备选方案浮现出来：比如制作监视器，供同学安装；设计问卷邀请同学填写等等。

后来，我们发现，其实输入法的词库就是一个非常好的工具——它会记录用户输入的所有词，以及该词被输入过的次数。这应该是最方便，误差最小的一种方式。通过提取用户词库，逐个测试其中的词，即可找到在生活常用词中被屏蔽的那些关键词，同时还可计算出对于生活常用词，我国审查系统封锁的百分比。

调查关键词，要先从提取输入法关键词开始。

4.1.1 提取输入法关键词

这里以谷歌拼音输入法为例。在输入法的设置页面，可以将谷歌拼音输入法的词库进行导出。导出后的文件使用 GBK 编码，转换编码为 UTF-8，可以看到文件 `jiehan-utf8.dic` 有 3 栏，使用一个制表符分隔每栏。截取几行如下，用于研究：

```
啊 666 a
阿 5 a
啊啊 24 a a
啊啊啊 87 a a a
啊啊啊啊 25 a a a a
啊啊啊啊啊 30 a a a a a
啊啊啊啊啊啊 2 a a a a a a
啊啊啊啊啊啊啊 6 a a a a a a a
```

啊啊啊啊啊啊啊啊啊啊啊 3 a a a a a a a a a a a
啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊啊 1 a a a a a a a a a a a a a a a a a a a

可以看到数据的格式：第一栏是词语；第二栏是词语被用户输入的次数；第三栏是词语的汉语拼音。

通过简单的浏览，可以看到，这样的词库，显然是不太有用的。因此，我们截取了频数，也就是第二栏的内容，大于等于 30 的所有词来做测试。这样一定程度上避免了无意义词汇被包含在我们的测试范围之内。在 bash 中使用 Awk 代码：

```
jiehan@jiehan-laptop:~$ cat jiehan-utf8.dic | awk '{ if (
    $2 > 30 && length( $1 ) >= 2 ) print $1"\t"$2 }' > freq-
    more-than-30
```

来截取出这样符合条件（频数大于等于 30，且长度不小于 2）的词语。再次截取过滤之后的词汇如下：

啊啊啊 87
是啊 68
可爱 35
答案 35
胡岸 46
晚安 263
好吧 81
是吧 43
我把 45
掰掰 35

4.1.2 Google 搜索的自动进行

下一个步骤是进行测试。测试显然需要在中国的网络连接下进行。我们编写了测试程序，来对类似如上文件中的词语或站点 URL 逐个进行测试，然后再将结果写入到专门的文件中，再使用编写的分析程序进行数据统计。

下面提及的代码都可以在随附数据 CD 中的 **关键词测试/** 下的 4 个目录中找到，名称分别为 `test.py` 和 `analyze.py`。虽然文件名相同，但其内容略有不同：两个输入法词库测试的 `test.py` 和 `analyze.py` 内容相同；敏感事件和敏感人物测试的 `test.py` 和 `analyze.py` 相同。

封锁方法人工测试以及测试方法的挑选

根据多次人工测试，以及查阅各种资料[1]，我们发现，在使用境外搜索引擎搜索的过程中，屏蔽的方法只有重置连接（发送 RST 包）一种。这将非常便于测试——这也是使用境外搜索引擎的原因之一，因为便于判断。而国内搜索引擎（如百度），通常对于敏感词的处理方法是删除搜索结果，有时甚至不对用户进行提示[2]。这会给自动化判断带来一定的麻烦。

我们首先测试**词语样本** `1`，它共有 851 个词语信息。为了分次测试方便，我们为测试数据进行了分组，100 个一组，分别以编号（1–9）存入目录 **关键词测试/输入法样本** `1/` 中。

在第 7 页 subsection 4.1.1 提取用户词库的过程中，我们的待测试文件（输入文件）有 2 栏。两栏均取自用户词库。程序在开始运行时会将每行的内容读取并赋值给名为 `words` 的 list 中。在处理每一行的过程中，程序将把制表符左侧和右侧的内容分开，分别存入 `word` 和 `freq` 变量当中，这样我们就得到了关键词及其频数。

之后就会进入 Google 搜索测试的过程。这一过程是通过 Python 3.2 的 `urllib` 实现的。在编写程序的时候，通过观察用户在浏览器的搜索，找出 Google 搜索的关键词参数为 `q`。同时，经过测试发现，Google 会判断 `User-Agent` 字符串，依此只接受一般浏览器的请求。因此，我们截取了 Google Chrome 浏览器的 `User-Agent` 字串，也在发送请求时一并发给 Google，以伪造一般用户的行为。

设计测试程序

经过测试和浏览 `urllib` 的文档，我们发现在连接被重置后，`urllib` 会抛出一个 `urllib.error.URLError` 错误。我们使用 `try...except` 来处理这样的错误，判断连接是否被重置。

联系 Google 的代码如下：

Listing 4.1: 关键词测试/输入法样本 1/test.py: test_by_google()

```
def test_by_google(keyword):
        # FIXME 假 User-Agent
        params = urllib.parse.urlencode({'q':
            keyword})
        headers = {'User-Agent': 'Mozilla/5.0 (X11
            ; Linux x86_64) AppleWebKit/534.24 (
            KHTML, like Gecko) Ubuntu/11.04'}
```

[1]维基百科《防火长城》中的 “关键字过滤阻断” 一节。

[2]来自 Wikileaks《China: censorship keywords, policies and blacklists for leading search engine Baidu, 2006-2009》。

```
try:
        req = urllib.request.Request("http
            ://www.google.com.hk/search?%s"
             % params, None, headers)
        result = urllib.request.urlopen(
            req, None, 5)
        content = result.read().decode('
            utf-8')

        # 检查返回搜索结果中是否包含我们搜索的关
            键词
        if re.search(keyword, content) ==
            None:
                print("[!] 收到回
                    应,但 content 中未找
                    到 \"%s\" % keyword")
                return False

        print("[~] 无误,content 中找
            到 \"%s\"" % keyword)
        return True
except:
        # 发生了异常,代表我们被重置了
        print("[!] \"%s\" 被重
            置" % keyword)
        return False
```

经过测试和阅读有关文献[3]，我们发现，如果被我国网络审查系统重置了连接，那么在约 90 秒内，随后所有的连接都将被重置。因此，我们在程序中设置了一个等待机制。如果一个词被重置，那么程序就会等待一定的时间之后再继续下一个测试。保险起见，我们不仅将延时设置为了 300 秒，同时还使用了一个 “安全词” 的检测。在 300 秒等待结束后，会搜索一个 “安全词”，以测试封锁期是否已经结束。在本例中，这个 “安全词” 为 “`bhsf`”。程序将会不停等待，直到搜索 “安全词” 时可以正确地接收到来自 Google 的搜索结果。这部分的代码如下：

Listing 4.2: 关键词测试/输入法样本 1/test.py: jiehan_wait() & wait_until_unblock()

```
def jiehan_wait(interval):
```

[3]维基百科《防火长城》中的 “关键字过滤阻断” 一节。

```
            pbar = ProgressBar(widgets=[Percentage(),
                Bar()], maxval=300).start()
            for i in range(300):
                    time.sleep(interval / 300)
                    pbar.update(i+1)
            pbar.finish()

    def wait_until_unblock(wait=True):
            sys.stdout.write('[X] 清理: 解封测试:\n')

            if wait:
                    jiehan_wait(TEST_INTERVAL)

            current_wait = 1

            sys.stdout.write('重
                试[ #%d]\n' % current_wait)
            sys.stdout.flush()

            while test_by_google(TEST_KEYWORD) ==
                False:
                    current_wait = current_wait + 1

                    jiehan_wait(TEST_INTERVAL)
                    sys.stdout.write('[RETRY #%d]\n' %
                        current_wait)
                    sys.stdout.flush()

            print("[~] 当前解封!\n")
```

在程序开始时，以及每次测试被重置时，`jiehan_wait()` 会被调用。

我们采用一个标记机制来记录程序的测试情况。标记将放在输出文件 `report-N` 中的第三列。标记的含义如下：

- R ——被重置（**R**eset）：该词被使用重置的方法封锁；
- P ——正常（**P**ass）：没有发现针对该词的异常。

运行程序

程序将花费一定时间。运行截图请见第 12 页上的 Figure 4.1。

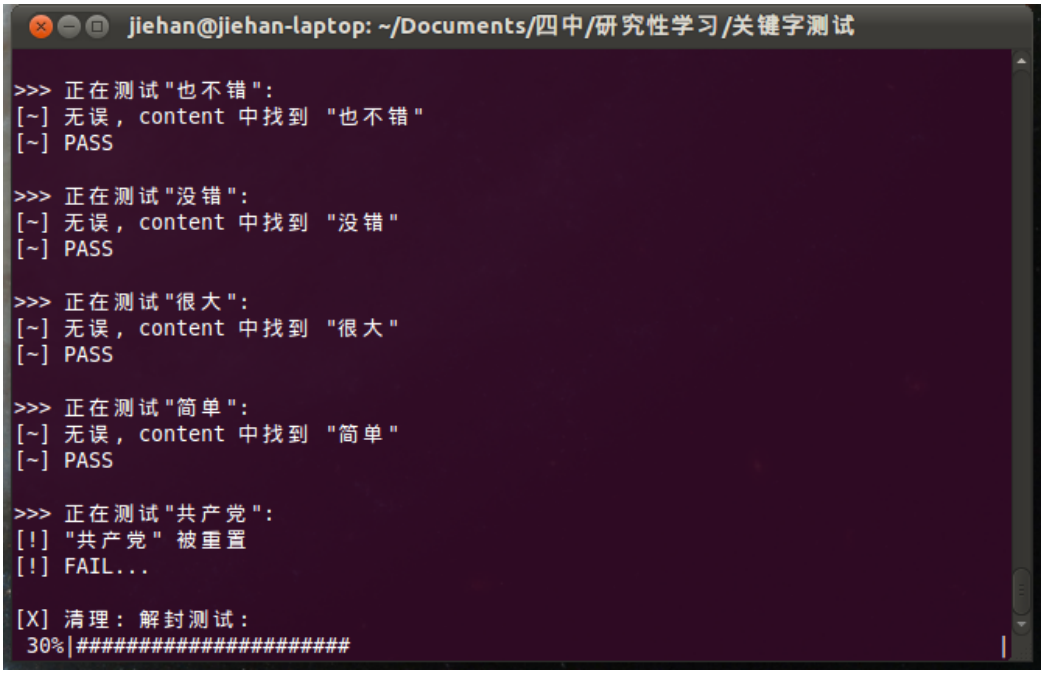

Figure 4.1: 测试过程中的截图

完成后，输出文件的样式大致如下：

简单 63 P
共产党 39 R
看不到 41 P
找不到 36 P
看到 228 P

设计统计程序

得到了结果，但是数据量非常的大，很难手工统计。因此，我们还设计了一个专用统计程序，统计各词的被屏蔽状况。输出结果有三：测试文件中被封锁的所有词、被封锁词的频数和总频数、被封锁词的个数和词总数。

这里使用了非常简单的 4 个计数器：`freqSum`、`freqBlockedSum`、`wordsSum`，和 `wordsBlockedSum`。由于过于简单，在此不再赘述。

该部分程序的源代码可以在随附数据 CD 的 **关键词测试/** 的各个子目录找到。

程序输出结果形如：

被封关键词：共产党，麦当劳，代理，王雨萌，周年，翻墙，
封锁频数（被封/所有）： 377/83576
封锁词数（被封/所有）： 6/851

引入第二个输入法词库测试样本

为了保证关键词测试部分的准确性，我们还进行了第二个输入法词库的测试，以尽量减小个人爱好因素造成的偏差。这一个输入法词库取自某一个组员，和上一位组员的爱好相差较大。

不同点 由于这组数据的提供者使用谷歌拼音输入法的时间不长，若仍沿用上一次测试所用的提取方法 “词频高于 30”，则待测词语就会非常少。因此，我们对于这位同学的词库，提取词频高于 5 的所有二字或以上词语。提取完毕后共计 352 个。对于这 352 个词语，我们的测试方法和第一次测试完全一样。

4.1.3 追加的两组关键词测试

引入敏感事件和敏感人物词汇表

在分析数据的过程中，我们对仅输入法提供的结果不完全满意。因此我们引入了这份由我们自行准备的 “敏感事件词汇表” 和 “敏感人物词汇表”，再次进行测试。更加详细的内容请见第 18 页上的 section 5.2 追加调查的分析，在此只描述测试方法上的变化。

这次测试，方法大体上相同，只是我们在各个测试和分析程序中去除了词频统计的部分。具体的源代码和测试数据及结果可以在随附数据 CD 的 **关键词测试/敏感事件/** 和 **关键词测试/敏感人物/** 目录中找到。

4.2 50 个最热门网站的可访问性测试

通过互联网使用经验，众所周知的是，除了关键词的封锁，我国还拥有特定网站的访问机制。为了更好地从另一个方面了解封锁情况，我们决定对前 50 个网站进行可访问性测试。

4.2.1 从 Alexa 拷贝热门网站域名

Alexa Internet, Inc. 是一家加利福尼亚州的公司，隶属于 Amazon.com[4]。它提供较为准确的站点网络流量排名。

在 `http://www.alexa.com/topsites` 可以浏览前 500 个热门网站的列表。我们手工将前 50 个域名进行了复制，粘贴并存入到了 `the_list` 文件中，这个文件可以在随附数据 CD 的 **热门网站访问测试/** 目录中被找到。获取的时间是 2011 年 4 月 29 日，可能和现在的排名情况有出入。

文件的前 5 行形如：

[4]根据 Wikipedia “Alexa Internet” 页面。

```
google.com
facebook.com
youtube.com
yahoo.com
blogspot.com
```

4.2.2 访问测试的自动进行

测试方法的设计

测试站点的可访问性和测试关键词能否搜索，有一些区别。比如，测试关键词能否搜索在算法上的实质是构造 Google 搜索结果页面的 URL，然后检查能否访问，并检查其返回的结果是否是我们期待的；然而，站点的可访问性受多个因素影响，比如站点是否被 DNS 污染（被返回了错误的 IP 地址），以及站点的 IP 地址本身是否已经被屏蔽，导致无法连接到主机。但是，在本次测试中，我们并不关心一个站点到底是被哪种，或者哪几种方法屏蔽。我们关心的是它是否被屏蔽。同时，对于远程站点服务器和我国的屏蔽设备来说，自动访问测试和真实的用户访问是完全一样的，它们并不知道远端的是自动测试程序，还是用户的浏览器。因此，在我们的研究中，自动进行访问测试的原理，大体上与第 8 页 subsection 4.1.2 中介绍的，进行 Google 搜索的方法一致。

设计测试程序

这部分的程序代码是从进行关键词测试的代码修改而来的。

访问函数将在从文件读取出来的内容前面加上 `http://`、在后方加上 `/`，同时删去了测试网站时所不需要的测试步骤。考虑到网站本身的不确定性，同时还将超时时间由 5 秒改为了 10 秒。另外，还去除了第 10 页 4.1.2 节中描述的等待机制，因为在这里我们访问的总是不同的主机。

修改之后的访问函数的代码是：

Listing 4.3: 热门网站访问测试/test.py: try_access()

```
def try_access(domain_name):
    testee = 'http://%s/' % domain_name
    headers = {'User-Agent': 'Mozilla/5.0 (X11;
        Linux x86_64) AppleWebKit/534.24 (KHTML,
        like Gecko) Ubuntu/11.04'}
    try:
        req = urllib.request.Request(testee, None,
            headers)
```

```
            result = urllib.request.urlopen(req, None,
                10)

            return True
        except:
            # 发生了异常，该站点无法访问
            print("[!] 无法访问")
            return False
```

在这个情况下，我们还修改了标记机制所使用的标记。在本例中，标记将放在输出文件 `report-1` 中的第二列。标记的含义如下：

- F ——被重置（**F**ail）：出于各种原因，该站点无法访问；
- P ——正常（**P**ass）：没有发现针对该站点首页的异常。

这个测试仅针对站点首页，并不能说明站点的每个页面都没有被屏蔽。同时，如果站点在境外，它将依旧受到关键词的检查。

运行测试

程序将花费一定时间。完成后，输出文件的样式大致如下：

```
google.com P
facebook.com F
youtube.com F
yahoo.com P
blogspot.com F
```

修改统计程序

统计程序 **热门网站访问测试/**`analyze.py` 也进行了一些小修改，这个文件可以在随附数据 CD 中找到。

运行统计程序

程序输出结果形如：

```
被封站点: facebook.com, youtube.com, blogspot.com,
twitter.com, googleusercontent.com, fc2.com, imdb.com,
xvideos.com, megaupload.com,
封锁站点数（被封/所有）: 9/50
```

章节 5

分析

5.1 生活词汇分析

5.1.1 说明

这部分内容的数据来自第 9 页 section 4.1.2 中描述的第一、二次对于输入法词库的测试。

频数 用户打过的所有词的次数总计。假设，“翻墙” 一词打过 10 次，那么它的频数就是 10。频数的比例可以更好的反映实际用户被封锁的可能性。

封锁词数 所有词中（不考虑频率）被封锁的个数。

5.1.2 结果

如下两个样本可以在随附数据 CD 中的 **关键词测试/** 目录下找到。

样本 1

被封关键词：共产党、麦当劳、代理、王雨萌、周年、翻墙。
封锁频数 (被封/所有)：377/83576
封锁词数 (被封/所有)：6/851

样本 2

被封关键词：刘本林、周末、吴康妮、周五、复习、学习。
封锁频数 (被封/所有)：65/6767
封锁词数 (被封/所有)：6/352

5.1.3 具体分析

总体情况

此次调查被封锁的总频率约为 0.49%；被封锁次数约占词库的 1.00%.

逐词分析

共产党 维基百科中 “共产党”、“中国共产党” 的链接（内容中立）；中国共产党官方网站；歌颂中国共产党及共产党干部的中国大陆新闻网站。

麦当劳 麦当劳官方网站；中外网站对麦当劳的介绍；关于麦当劳的新闻。其王府井餐厅是近期敏感的 “茉莉花革命” 在北京的活动地点。

代理 链接大部分为代理服务器供应商（帮助用户突破网络信息封锁）。

王雨萌 链接均关于名字是 “王雨萌” 的人。未发现被封锁原因。

周年 链接为中华人民共和国建国、中国共产党成立、五四运动、武昌起义、汶川地震以及各种学校、博物馆等机构的周年纪念。只有一个 “文化大革命”40 周年是敏感内容。

翻墙 链接大部分为代理服务器供应商（帮助用户突破网络信息封锁）以及对突破网络封锁的介绍。

刘本林 链接均关于名字是 “刘本林” 的人。其中有一个犯罪分子。“刘本昕”（相近的名字）是一个饱受争议的政治敏感人物。

周末 链接有《南方周末》的网页。

吴康妮 链接均关于名字是 “吴康妮” 的人，无敏感人物。“吴康轩”（相近的名字）为有争议的政治人物。

周五 未发现被封锁原因。

复习 链接均为教育网站的中考、高考复习网页。未发现被封锁原因。

学习　链接均为各种考试备考网页。未发现被封锁原因。

通过对以上被封锁词的分析，我们对中国互联网审查封锁关键词类型做出以下猜测：

1. 政治敏感词。其中包括：政治敏感事件、日期、机构、人物。
2. 与敏感词相近的词。
3. 按一定比例封锁输入的词。

5.1.4　总结

测试时，输入法输入的内容很多，而很多生活上的词语是肯定不会拿去搜索引擎搜索的，所以导致了被封锁的词比我们预期要低。这次测试是对生活中常用的词语的抽样调查。

5.2　追加调查的分析

5.2.1　说明

第一轮调查，我们运用的词库中有很多口语词，敏感词较少。追加调查中，我们测试了草拟的政治敏感词与政治敏感人物表，并进行测试。测试搜索引擎：Google。

这部分内容的数据来自第 13 页 section 4.1.3 中描述的第三、四次对于输入法词库的测试。

5.2.2　结果

敏感事件词汇表样本

被封关键词：六四、天安门事件、诺贝尔和平奖、共产党、中国共产党、零八宪章、天安门母亲、法轮功、九评共产党、退党、三退、退党保平安、大纪元、茉莉花、六四戒严、叛徒内奸工贼刘少奇、China's Human Rights Record。

封锁词数 (被封/所有)：17/42

敏感人物词汇表样本

被封关键词：方励之、李锐、胡继伟、王若水、吴国光、胡平、刘晓波、达赖喇嘛、赵紫阳、鲍彤、高智晟、胡佳、姜力钧、李海、李智、刘荻、谭作人、王丹、王军涛、王小宁、王炳章、王有才、魏京生。

封锁词数 (被封/所有)：23/49

5.2.3 具体分析

封锁词数占输入词数的一部分。还有一部分敏感词没有被封锁。

5.2.4 人工追加测试

吾尔开希 未被封锁，但是搜索结果的所有链接几乎都被封锁。

苏晓康 未被封锁，但是前两页搜索结果只有 http://wjiausa.blog.hexun.com/38275291_d.html 和 http://article.netor.com/m/jours/adindex.asp?boardid=27420&joursid=42906 的链接可以打开。根据其内容，我们怀疑这是漏封的网页。

5.2.5 总结

有一些词可以并不在搜索中被封锁，而在网页中直接被封锁。这说明，有一些词在即使在搜索引擎中不被封锁，但是会在其他网页中被封锁。一个原因是，也许封网部门漏封了一个关键词，所以搜索引擎得以显示，但是因为网页中此敏感词与其他被封锁敏感词同时出现，所以网页会被封锁。

章节 6

网络封锁突破计划

6.1 VPN 和 OpenVPN 的选定

VPN 是一种较有前景的安全加密通信方法。而 OpenVPN 则是较为开放的 VPN 的实现，比较便于自定义和开发。

6.2 实现目标

我们希望我们的系统具有：

- 安全连接功能：用户可以在国内自由地访问互联网资源。
- 基于用户名和密码认证的连接：用户不需要客户端证书，而使用用户名和密码连接服务器。
- 用户使用及用量监控功能：记录用户流量和连接详情。
- 用户账户管理功能：管理员（出资者）可以随时浏览用户的使用情况，添加或删除账户。
- 高级路由功能：为国内网站设置专门路由，令其不通过 VPN 来访问网络，节约服务器资源，提高浏览速度。

目前，我们实现了前 3 项。后 2 项由于时间等原因还没有进行实现。

6.3 用户名和密码认证的实现

这部分内容已有人实现过，并且发布在了互联网上。我们参照 DozView 的《基于用户名/密码认证和流量控制的 OpenVPN 系统的实现》完成了基于用户名和密码认证，以及流量记录。

我们的数据表结构和该文章不同，我们自行设计了用户和日志数据表，如下：

Listing 6.1: ‘log‘ 数据表

```
CREATE TABLE ‘log‘ (
  ‘username‘ varchar(32) COLLATE utf8_unicode_ci
    NOT NULL COMMENT 'The owner of this log entry
    ',
  ‘started_at‘ timestamp NOT NULL DEFAULT
    CURRENT_TIMESTAMP COMMENT 'Connected at...',
  ‘ended_at‘ timestamp NULL DEFAULT NULL COMMENT '
    Ended at...',
  ‘server_mode‘ text COLLATE utf8_unicode_ci NOT
    NULL COMMENT 'The mode of the OpenVPN server'
    ,
  ‘remote_ip‘ text COLLATE utf8_unicode_ci NOT
    NULL COMMENT 'The IP address connected to the
    server',
  ‘bytes_rcvd‘ bigint(20) DEFAULT NULL COMMENT '
    How much data were sent from the client',
  ‘bytes_sent‘ bigint(20) DEFAULT NULL COMMENT '
    How much data did the client receive'
) ENGINE=MyISAM DEFAULT CHARSET=utf8 COLLATE=
  utf8_unicode_ci COMMENT='Connection log';
```

Listing 6.2: ‘log‘ 数据表

```
CREATE TABLE ‘users‘ (
  ‘username‘ varchar(32) COLLATE utf8_unicode_ci
    NOT NULL COMMENT 'Username of the user',
  ‘password‘ varchar(64) COLLATE utf8_unicode_ci
    NOT NULL COMMENT 'Password of the user',
  ‘role‘ tinyint(4) NOT NULL DEFAULT '1' COMMENT '
    User role: 1 for normal; 9 for admin',
  ‘status‘ tinyint(4) NOT NULL DEFAULT '1' COMMENT
    'User status: 1 for active; 2 for manually-
    set inactive; 3 for exceeded quota',
  ‘quota_30d‘ bigint(20) NOT NULL DEFAULT '0'
    COMMENT 'Quota for this user in bytes. 0 for
    unlimited',
```

```
  `usage_30d` bigint(20) NOT NULL DEFAULT '0'
    COMMENT 'How much traffic did this user make'
    ,
  `note` text COLLATE utf8_unicode_ci COMMENT 'The
    maintainer''s note',
  PRIMARY KEY (`username`)
) ENGINE=MyISAM DEFAULT CHARSET=utf8 COLLATE=
  utf8_unicode_ci;
```

6.4 网页管理部分

所有代码均可在随附数据 CD 的 **网页管理系统/** 目录中找到，在此不再赘述。界面请看 Figure 6.1。

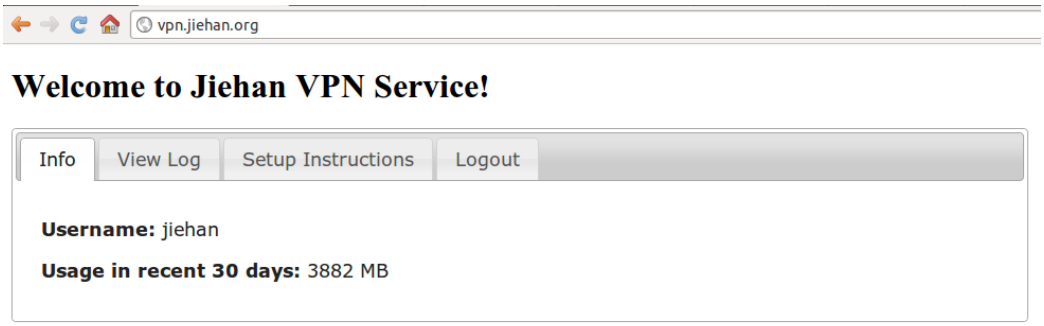

Figure 6.1: vpn.jiehan.org 截图

6.5 帐号发放策略

目前，任何四中校内申请帐号的请求都会被批准。当前共有 7 个帐号。

6.6 成本估计

本次的 VPN 服务运行在 `us2.jiehan.org` 上。它使用 `Thrust::VPS` 的服务，月收费 7.95 美金，提供 1 TB 的流量。由于运行的是 VPN 服务，不仅要接收用户的流量，还要再转发一次，因此实际可以服务 500 GB 的数据。

$$7.95 \text{ dollars} \div 500 \text{ GB} = 0.0159 \text{ dollars/GB}$$

也就是说，用户使用 1 GB，只需要投资者支付 0.0159 美金。

由于目前 VPN 服务用户数量不大，因此无法得出单个用户的预计资金使用情况。

章节 7

整体结论

7.1 封锁内容归纳

从上述的结果，以及对结果的分析中，我们可以看出因政治原因遭封锁的占绝大多数。可见政府对这方面内容公开后果的顾虑还是很大的。这也反映出政府对网民的不信任，他们希望通过网络审查的手段控制社会上主流言论的倾向。

7.2 对现行政策的建议

根据我们对上述问题的研究与分析，并综合中国特色考虑，我们希望中国政府减少对于网络内容的封锁，适当公开相关内容与信息。首先，这有助于减少网民对封锁内容不恰当的好奇心；其次，可以增强网民对政府的监督；最后，还可以较大程度地实现真正的言论自由，为网民提供一个更加自由的讨论平台。

同时，我们认为网民应克制过激言论的发表，防止政府以此为借口封锁相关内容，为创造一个更加和谐的网络环境共同努力。在必要时，网民也可通过 VPN 等方式获取所需信息。

因此我们认为，我国政府对于网络审查的改革是十分有必要的。

章节 8

课题的继续研究

8.1 敏感词封锁的进一步研究

在编写程序进行测试的过程中，我们发现，有些词汇在第一次测试时显示为被封锁，而第二次则可以正常搜索，如随附数据 CD 中 **关键词测试/输入法样本** `1/1` 文件中的 “胡岸” 一词。我们通过人工、自动等方法进行了多次测试，都没有找到对于这类词封锁的规律。

因此，问题研究的下一步，可以针对这些看起来 “偶尔封锁” 的词汇进行进一步的研究。

8.2 VPN 项目的完善

由于时间等诸多限制，“VPN 的廉价解决方案” 并没有完全完成。还有如下几个比较吸引人的功能没有来得及完成。如果时间允许，我们会：

- 实现 “令中国网站的流量不通过 VPN，而直接通过原网关访问互联网。并且中国 IP 地址的列表实现定期自动更新”；
- 将用户用量统计等统计功能完善。