百度精确搜索的南京大学网页链接正确性和多浏览器兼容性测试报告

王子成 DZ1833026

目录

一 背景和要解决的问题

从 google 搜索引擎出现至今，搜索引擎的技术不断发展，搜索的范围不断被拓宽、搜索准确性不断被提升，因此人们已经逐渐放弃门户网站，转而选择搜索引擎已经作为当下互联网的主要入口，而搜索结果页面呈现信息，尤其是网页链接的准确性尤为关键。南京大学作为中国“双一流”重点大学，学校内每个机构、每个部门通常都有自己的主页作为面向公众的窗口刊登信息。

这些网页链接多数情况下都通过搜索引擎获得，为确保搜索引擎中获取的链接准确无误并能够被正常访问，我们需要对搜索引擎中获取的全部链接进行测试，根据测试收集的结果分析链接及对应网页信息。同时，使用当下被广泛使用的 Chrome 和 Firefox 两种浏览器同时对这些链接进行测试，检测南京大学网页兼容性。

二 研究问题

通过百度搜索引擎的精确搜索功能（在搜索框中输入 site:nju.edu.cn 即可精确搜索南京大学的相关链接）获取南京大学的全部链接，使用 Chrome 和 Firefox 两种浏览器对这些链接进行访问测试，通过检测链接指向网页的 title 信息，链接地址和加载时间，分析链接信息是否准确，链接地址是否属于南京大学顶级域名，链接指向网站是否能够顺利加载，以及链接指向网页是否使用 https 加密增强安全性。收集以上全部信息，对比两个浏览器的相应结果，检测南京大学链接指向网页的兼容性。

由于链接数目较大，因此使用 python 语言和 selenium 包编写测试脚本，自动化访问搜索页面收集的全部链接，并分析收集到的相关数据。

三 实验设计

3.1 总体设计

根据 wikipedia 统计 2018 年桌面端浏览器使用比例，由于 IE 浏览器已经基本过时，我们选取 Chrome 和 Firefox 两款时下较为流行的浏览器进行测试工作。

Usage share of desktop browsers

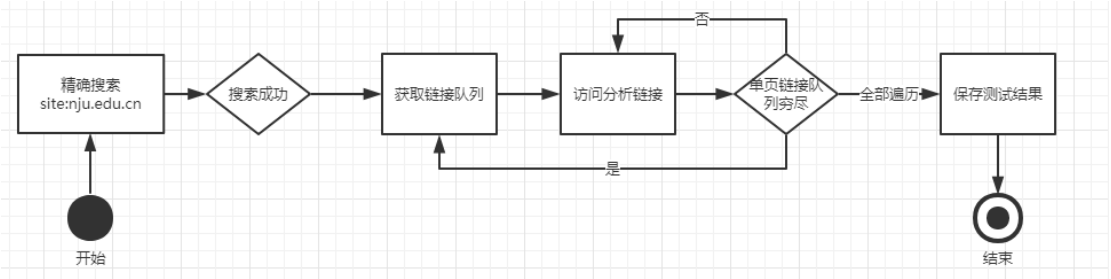
Browser	NetMarketShare[22] November 2018	W3Counter[23] November 2018	StatCounter[24] November 2018
Chrome	65.49%	63.5%	72.45%
IE	9.66%	4.7%	5.36%
Firefox	8.99%	6.6%	9.08%
Edge	4.23%	2.6%	4%
Safari	3.76%	13.9%	5.05%
Opera	1.56%	3%	2.18%
Others	6.31%	5.7%	1.89%

利用 Python 语言和 selenium 自动化测试包对链接进行自动化测试。编写测试脚本，自动化获取并访问在百度精确搜索页面获取的全部链接。首先下载并安装最新更新的稳定 release 版本的 Chrome 和 Firefox 浏览器，并下载 selenium 官方网站提供的浏览器自动执行驱动。使用 selenium 分别创建控制两种浏览器的 webdriver。

之后使用 webdriver 打开 https://baidu.com，搜索 “site:nju.edu.cn” 获取有关南京大学的全部热门链接（经测试，百度搜索引擎的爬虫共获取超过一百万个链接，精确搜索仅提供热门链接，两浏览器均能获取 76 页共 757 个南京大学相关热门链接。

获取单独页面中的全部链接，分别使用 webdriver 模拟自动化操作点击每个链接，若单页链接被全部遍历则自动化点击下一页重复此过程，点击链接并访问指向网页后，首先检测该网页是否能够被正常加载，其次分析该链接 title 是否符合搜索引擎结果页面提供的信息，之后分析链接地址信息，是否使用 https 加密，顶级域名是否属于 nju.edu.cn。

最后汇总全部链接分析测试数据，按照 CSV 格式保存到硬盘中，使用 notepad++等文本编辑软件，最后手动对不符合 CSV 格式规范的数据进行清洗，获得的数据可以使用 Excel 等表格软件查看，并通过 python 脚本或 R 语言脚本进行详细地数据分析总结。



3.2 详细设计

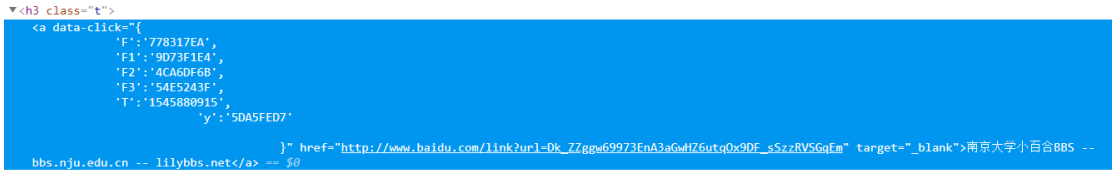
3.2.1 链接数据的获取

获取搜索引擎的搜索框，输入精确搜索内容并确认搜索。

```
elem = driver.find_element_by_name("wd")
elem.send_keys("site:nju.edu.cn")
elem.send_keys(Keys.RETURN)
```

3.2.2 搜索结果页面中的 title 获取

使用浏览器的查看源码功能，发现搜索结果页面中每个链接均按照以下形式呈现。



因此，我们只需要获得 classname 为 t 的数据列表，并根据列表中提取的链接名称获取链接地址。

3.2.3 脚本与函数设计

Test.py 测试脚本：
nextpage(driver)：当搜索结果页面的链接被全部分析之后，调用 webdriver 进行翻页操作。
initdriver(browser) ：根据浏览器的类型创建 selenium webdriver，打开百度搜索页面，等待用户登录，登陆成功后输入 “site:nju.edu.cn” 进行精确搜索，返回 webdriver。
trim() 链接 title 可能存在空格等干扰因素，去除这些干扰因素，返回被清洗过的数据。
operate(driver,browser)：根据 driver 和 browser 自动化测试分析获取的全部链接数据，返回结果列表。
writeTocsv(errorlist, browser)：将结果列表按照 csv 格式写入文件。
Analysis.py 分析脚本：负责读取生成的两个 csv 格式文件，分别提取其中的 titile 和 url 信息作为 index 生成集合，对两个集合进行交和并操作，分析收集到数据的异同。之后分别记录两个 csv 文件中各种 WARNING 和 ERROR 的比例，绘制饼状图。

3.2.4 点击链接测试分析结果数据格式设计

Title	链接指向网页的 title 信息
url	链接地址
WARNING0	链接指向的网页未使用 HTTPS 进行加密
ERROR0	链接指向网页的信息与搜索结果页面链接信息不符
ERROR1	链接指向地址与搜索结果页面链接地址不符
ERROR2	链接指向网页加载发生异常

针对每个链接进行上述问题分析，当链接及链接指向网页不存在问题时则只记录链接 title 和链接地址，一旦出现任何问题则以 csv 格式进行数据记录。

3.2.5 搜索引擎的登录问题

百度搜索引擎在搜索的过程中会需要登录，且登录时要下载“百度 APP”或使用手机验证码，因此需要在打开百度搜索页面后让系统等待一段时间，使用手机进行登录，成功后再进行其他步骤。

由于 selenium webdriver 在相同系统内无法并行执行，因此必须在使用一个浏览器测试全部链接之后再使用另一个浏览器，为了能够提醒测试人员进行登录操作，我们使用 winsound 包调用计算机蜂鸣器，使用 winsound.Beep(3600, 60*1000)函数大音量蜂鸣 1 分钟。确保第二个浏览器的自动化测试能够顺利进行。

3.2.6 标签页切换问题

在 Chrome 和 Firefox 两款浏览器中，每访问一个新的测试链接会产生一个新的标签页，如果放任标签页增长不加以控制的话可能会有影响测试脚本的性能，因此使用 webdriver 提供的 window_handles 获取标签页的句柄，在访问链接进行分析测试时，将 webdriver 的主句柄设置为被访问链接 handle，访问结束后关闭标签页，将 webdriver 主句柄切换回搜索结果页面 handle。

3.2.7 网页加载过程中的等待问题

由于网络状态随时间不断变化，而且不同网页需要加载时间不同，因此使用两种等待方法，

```
driver.implicitly_wait(30)
time.sleep(5)
```

第一种是隐式等待，即在执行访问被测试链接后 webdriver 等待 30s 时间给网页进行加载，如果在 30s 内加载成功，则继续进行之后步骤，如果超出 30s 则直接抛出异常，在本次实验的测试脚本中将被视为网页无法加载，记录 ERROR2 异常。

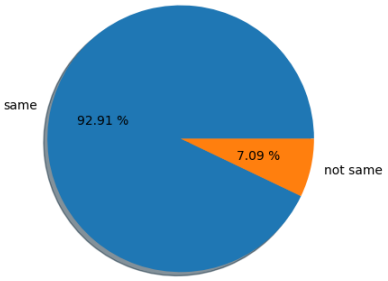
第二种等待是暂停整个脚本进行等待，在等待用户登录过程中采用这种等待类型，在获取链接的过程中额外等待 5s 时间便于浏览器对获取到网页进行渲染。

3.2.8 多次实验对比结果

由于网络状态随时间变化，因此进行多次试验进行结果对比，根据分析实验中分别采集的数据结果基本一致，故只分析最后一次采集到的测试数据。

四 实验结果

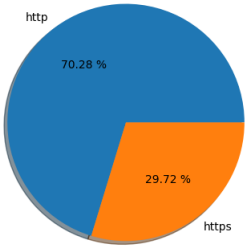
Chrome 和 Firefox 两浏览器分别分析 757 个最热门链接（百度搜索引擎爬虫仅提供 757 个热门链接），然而两个浏览器结果存在差异，经过统计一共收集 783 个网页，其中搜索结果一致的共有 635 个链接，其余均存在一定差异。两个浏览器收集到的 635 个相同链接中有 45 个存在差异，差异点均为 ERROR0 结果不一致，深究原因可能有 1）浏览器对相同代码信息渲染结果存在差异，使得 selenium 包获得的结果产生不同。2）网络状态随时间变化，难以控制，网络状态不同导致加载时间存在差异。如图所示：

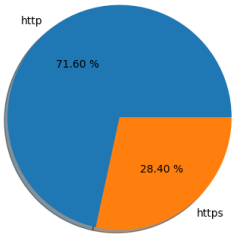


下面分别进行讨论：

4.1 Chrome 浏览器和 Firefox 浏览器 https 使用情况

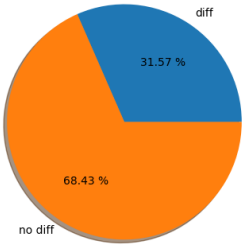
二者类似，仅有约 30%的网站使用 https 进行保护。

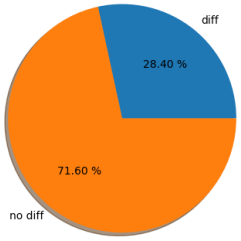




4.2 Chrome 浏览器和 Firefox 浏览器 title 信息与搜索结果页面提供信息不符情况

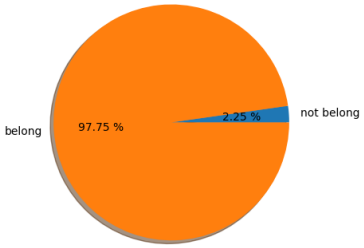
均有 30%左右的数据与搜索结果页面不符，可能是百度搜索引擎的爬虫更新信息不及时。

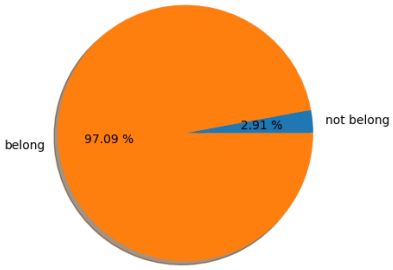




4.3 Chrome 浏览器和 Firefox 浏览器搜索引擎获取的链接不属于 nju.edu.cn 顶级域名的情况

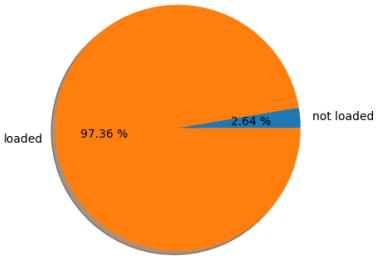
约有 2%-3%的网页链接不属于 nju.edu.cn，原因可能是加载速度太慢，脚本测试分析到的是百度的链接重定向页面。

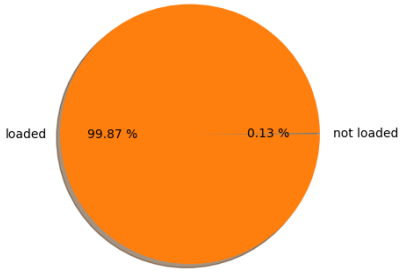




4.4 Chrome 浏览器和 Firefox 浏览器链接指向网页无法加载的情况

Chrome 浏览器的无法加载情况大于 Firefox 浏览器无法加载情况，存在该版本的 Firefox 浏览器比 Chrome 浏览器兼容性更好的可能。





五 讨论、工作亮点

1. 南京大学部分机构和部门的网站负责人员安全意识较为淡薄，约 70%的网页都没有使用 https 加密，有很高的安全风险，建议统一设置相同证书，一方面可以增强安全性，另一方面也可以防止三级域名被劫持伪造。甚至缴费系统的网站也没有使用 https，例如：南大缴费系统,http://net.nju.edu.cn/,WARNING0,,,

2. 部分网页，例如 vpn2.nju.edu.cn 等网站不对外网用户开放（测试环境为外网服务器），但被百度爬虫获取，因此检测被判定为无法加载，这部分数据价值不大。

3. 百度爬虫在获取 title 信息时会对 title 文本进行处理，或之前爬取的文本未作更新，因此可能出现 title 与搜索结果不吻合的情况，应该主动向搜索引擎提交更改信息，及时维护相关信息。同时也有南京大学研究生网页 title 直接被改为研究生，与南京大学相关性被限制，建议学校及时整改。同时，校内很多网站的 title 只写了具体部门或具体部门的简称，而没有标注南京大学，例如保卫处，人武部等，相同风格的网站有很高的安全风险，易于被攻击者伪造。

4. 两种浏览器精确搜索的数量一致，但最终结果并不完全相似，极有可能是浏览器随访问数量实时调整的结果。

5. 部分网页例如 PDF 版学生手册，可以直接使用浏览器阅读 PDF 文件，该网页没有 title。

六 结论

南京大学机构和部门的网页作为向公众开放的窗口基本能做到全面覆盖，然而安全意识较为淡薄，不仅绝大多数链接没有使用 https 进行保护，而且部分财务和缴费网页可以被外网访问的同时也没有使用 https 保护，一旦攻击者利用安全漏洞进行攻击，后果将非常严重。此外，很多网页的 title 与百度搜索引擎结果页面提供的信息严重不符，其中很有可能是百度搜索引擎爬虫更新数据不及时或 CDN 数据更新不及时，不过网站的相关工作人员在对网页进行较大改动的时候也应该主动向搜索引擎提交变更信息，便于信息及时更新。

最后，部分南京大学的机构或部门网页的 title 仅仅写了部门的名称，甚至部门的简称，没有包括南京大学这一关键信息，很容易造成误导，希望学校的相关部门能够提升对学校信息安全建设的重视程度。