计算机网络 爬虫 (计算机网络) 网页爬虫 修改

现在的网络爬虫的研究成果和存在的问题有哪些? 修改

现在的网络爬虫的研究成果和存在的问题有哪些?

修改

关注问题 写回答 邀请回答 添加评论 分享 邀请回答 举报 ...

关注者 283 被浏览 7,922

他们也关注了该问题

查看全部 5 个回答

Felis sapiens
函数式编程、编程语言、编程 话题的优秀回答者

开源哥等 9 人赞同了该回答

visual scraper，不用写代码也能通过图形界面快速定义出一个爬虫来用，比如Portia。

发布于 2016-04-20

赞同 9 添加评论 分享 收藏 感谢 ...

更多回答

xlzd
Python 话题的优秀回答者

20 人赞同了该回答

首先有三个最需要解决的问题：

- **法律和道德风险**：爬虫抓取其它网站数据，虽然抓取的内容大部分是公开的，但是商用或者有损源网站利益，于法于理都说不过去。目前我国（或者说大部分国家）针对互联网的方方面面法律覆盖度还远远不够。
- **访问速度与瓶颈**：爬虫的访问速度依赖于网速（尤其是服务器出口带宽以及用户入口带宽）和开发者的水平，而大部分商业网站都会有反爬虫机制，其中最简单就是通过频率限制，复杂的则会加上很多维度的判断。如何高效抓取？如果数据量不大，则可以通过在两次请求间休息一段时间，如果数据量很大，则需要考虑有一套高效、可用的代理 IP 机制。
- **验证码**：现在的验证码已经从简单的输入几个字母，变得复杂了很多，比如拖动滑块甚至是 Google 的 reCAPTCHA 这样基于机器学习的验证码模块。在识别验证码的开销与数据所能获得的收益之间，要找到一个平衡点。

剩下可能存在的问题：

- **如何不基于规则地解析数据**：大部分网上的爬虫教程，都是讲如何发请求、如何抽取数据。对于特定网站这是可行的，但是对于几百上千个网站，这样的做法就实在太慢了，如何不基于规则而解析数据，才能达到高效获取数据（高效指的是开发效率，因为不需要针对特定网站单独实现规则）。
- **通用性与易用性**：现有的所谓现成的采集工具，大多是不够通用易用的。那些采集工具，专业的看不上，小白依然不会，用户估计（没有调查，纯脑洞）大部分都是半吊子水平，代码写不出，但是又多少知道点。
- **数据变现**：整体来讲，虽然抓取数据有很多门槛，但是其实想要从互联网抓数据还是非常容易的，如何让你抓下来的数据产生价值，这是一个难题。
- **其他**：欢迎补充~

发布于 2016-06-18

关于作者

Felis sapiens

函数式编程、编程语言、编程 话题的优秀回答者

电影旅行敲代码、Antokha Yuuki、暮无井见铃也关注了她

回答 624 文章 40 关注者 14,871

已关注 发私信

被收藏 10 次

python 35 人关注
每天进步一点点 创建

编程学习 2 人关注
淘小黑 创建

没用的 1 人关注
Tim 创建

Python 0 人关注
steven chen 创建

▲ 赞同 20 ▼ 3 条评论 分享 收藏 感谢 ···

王小平
网络智能，机器学习，大踏步的人工智能

13 人赞同了该回答

在工程中有这样一些问题，
1，快速频繁访问会被封IP，一般可通过代理和增加等待时间解决；
2，需要登录信息，例如微博，可通过携带cookie解决；
3，国内下载国外网站可以用国外代理；
4，网页解析，有比较成熟的各种库，常用的有python语言；

展开阅读全文 ∨

▲ 赞同 13 ▼ 2 条评论 分享 收藏 感谢 ···

查看全部 5 个回答

计算机 0 人关
李垚圣 创建

相关问题

为什么网络爬虫好难，涉及到的知识我不会？ 15 个回答

Python 3 网络爬虫学习建议？ 30 个回答

通俗的讲，网络爬虫到底是什么？ 24 个回答

App中的数据可以用网络爬虫抓取么？ 13 个回答

从目前的就业形势来看，是从事python web后端开发好呢，还是从事网络爬虫比较好呢？ 4 个回答

相关推荐

100 个 iOS 11 实用技巧
少数派
228,412 人读过 阅读