一个利用 Chrome 插件实现微信公众号采集实现分类阅读、RSS 的思路

airyland · 112 天前 · 1192 次点击

今天看到这个帖子 #452686 把之前想到的列一下。

以前写过一个 Chrome 插件用以搜索搜狗微信某关键词匹配到的所有文章，扩展开来可以实现一个多节点爬虫系统。 Chrome 插件没有 CORS 的问题可以采集所有网站，这是采集的核心基础。

核心构成

一个简单的数据保存服务端

- 分发待进行的采集任务
- 支持保存、返回特定公众号内容

一个 Chrome 插件

- 用以采集搜狗微信公众号最新 10 条群发
- 数据复用，相同公众号的内容直接从服务端获取，每个公众号只要当天已经获取到就不需要重新获取(特殊权限的暂不处理)
- 当有足够多的人(比如>200)使用该插件时，将能实现以极低的频率去获取内容并且不会触及防采集限制，也不会影响到浏览器性能
- 当不幸还是被搜狗判定机器嫌疑时提醒输入验证码

本质上就是一个爬虫系统，和使用 ip 代理来采集只是实现上的差别，但是省去了购买代理 ip 或者维护 ip pool 的麻烦。

可以实现的功能

- 本地分类阅读
- 本地未读统计
- RSS

既然任务从服务端分发，使用者如何确保不会被用来干坏事(采集其他站或者攻击)?

- 后端代码及插件代码开源。
- 上传 Chrome 应用商店代码不混淆，因此也可以被用来审查是否有恶意代码。
- 插件能访问的域受 manifest 控制，配置里只支持搜狗微信域。

其他

还没想到，欢迎补充。

第 1 条附言 · 91 天前

补充一下进展，目前用 chrome 写了个爬虫，抓取首页的每日热门帖子，以及用任意关键词为搜索词无目的进行采集，现在大概有百万文章及 3 万个自动发现采集的公众号。

2 回复 | 直到 2018-05-22 08:22:34 +08:00 采集 Chrome 搜狗 插件

yuanfnadi 101 天前 via iPhone 1

如何防止用户胡乱 post 污染数据?

airyland 101 天前 2

@yuanfnadi 好问题，需要关注公众号(或者其他方式)获取一个用户 client_id，chrome 插件在提交时生成一个 token(逻辑代码混淆提高伪造难度)，后台也会有频率监控，超过某个设定频率的也必定是错误数据，另外后端也会校验提交的数据，因为前端不清楚逻辑不好伪造，多次尝试失败后端就可以拉黑这个 client 了。