豆瓣图书 Locke 推荐系统文档

Locke

2012 年 9 月 28 日

1 数据来源

作为一个推荐系统，必然需要数据源，而且得是一个足够大的数据源，所以如何获得这些数据就是第一个问题。

按照传统思路，逐个抓取数据是最直接的办法，但是很遗憾，豆瓣并不是第一天存在了，也和各种爬虫斗争了这么长时间，相信它的“反数据挖据型爬虫”已经做得很好了，所以要想获得足够多的数据往往需要相当长的时间，这是很不利的。另一个问题是，豆瓣上有相当多的不活跃用户，很多用户的评分信息为 0，这进一步降低了直接爬虫的收益。所以直接从豆瓣上爬数据只能作为补充方案，作为主要数据来源并不适合。

无奈之余偶然在网络上发现了一个豆瓣评分的数据包：`http://www.datatang.com/data/42832/`，里面大概有 383K 个用户的 3.6M 条评分数据，这里要向共享这个数据的童鞋表示衷心的感谢。这个作为主要数据来源应该够了。但其中的数据就只有评分，忽略了是否有评论、tag、无评分记录、时间等信息。

然后才意识到搜索引擎这个爬虫的老大哥那也应该有页面的数据，经过尝试，Google 上有大量豆瓣图书页面的缓存，缓存的 URL 格式如

http://webcache.googleusercontent.com/search?q=
cache:book.douban.com/subject/2152385/

但对于用户信息页面的则相对较少，可以作为补充。百度的快照暂时还没有看出规律，不过可以通过在 baidu 搜索 URL 来获得快照地址。

所以总的数据来源是：主要从数据包获取，书籍信息通过搜索引擎快照，其他信息通过从豆瓣网站直接抓取。