# **[亚马逊跨境电商大数据智能热卖商品选款平台(美国，日本，英国，德国)](http://ran.hotaws.lenggirl.com/)**

# **使用说明手册**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 修改内容 | 责任人 | 修改时间 | 版本 | 联系方式 |
| 初次制定 | hunterhug | 20180527 | V1.0 | 见知乎 |
| 新增报告数据 | hunterhug | 20180527 | V1.1 | 见知乎 |
| 优化Web端界面，新增监控功能等 | hunterhug | 20180603 | V1.2 | 见知乎 |
|  |  |  |  |  |
|  |  |  |  | 版权所有 |

**https://www.zhihu.com/people/hunterhug**

**目录**

[序言 1](#_Toc15549)

[授权相关 1](#_Toc3850)

[正文 2](#_Toc30713)

[一．登录 2](#_Toc28821)

[二．用户管理 2](#_Toc13241)

[2.1.新增用户 3](#_Toc19924)

[2.2.修改密码 3](#_Toc21915)

[三．角色管理 4](#_Toc31219)

[3.1.新增角色 4](#_Toc24599)

[3.2.角色授权 4](#_Toc24982)

[3.3.用户授权 6](#_Toc30986)

[四．类目数据 6](#_Toc30664)

[4.1.查询类目 7](#_Toc8556)

[4.2.调整类目 7](#_Toc4952)

[五．小类数据 8](#_Toc4029)

[5.1.筛选查询 8](#_Toc2052)

[5.2.精确查询 8](#_Toc27386)

[六．大类数据 9](#_Toc25130)

[6.1.筛选查询 9](#_Toc21011)

[6.2.精确查找 9](#_Toc8408)

[6.3.数据导出 10](#_Toc15419)

[七．ASIN数据 11](#_Toc26985)

[7.1.筛选查询 11](#_Toc24714)

[7.2.精确查找 11](#_Toc18715)

[7.3.历史趋势 12](#_Toc1982)

[八． 报告数据 14](#_Toc31807)

[附录 16](#_Toc305)

[历史 16](#_Toc12269)

[知乎 17](#_Toc31794)

[V1.2补充 21](#_Toc6761)

[类目数据 21](#_Toc20832)

[监控数据 22](#_Toc26522)

# 序言

用途： 选款，特别适合亚马逊跨境电子商务运营公司(不支持中国亚马逊)

原理： 首先通过获取亚马逊热卖所有类目的URL，即从第一层大类，一直获取到第六层小类。通过这些类目URL可以依次抓取到这些类目某段时间的Top100的商品（类目下的爆款），这些Top100的商品排名我们称为小类排名，每个小时会变一次，但是由于变化基本不会太频繁以及抓取的商品数量很多，基本能覆盖。比如：有一个大类，下面有某一个三层类目，这个三层类目下面有几十个四层，四层下面又有五层，很多个Top100组在一起构成了三层我们需要的商品。通过这些小类商品数据，我们再进详情页获取更多的字段（评论数，星数，是否FBA，价格等），包括每件商品的最顶层排名，我们称大类排名。通过商品去重，分布式代理以及数据的一些预处理设计，加大马力，运用IT采集技术，我们能得到亚马逊大部分卖得好的商品，通过筛选，排序，我们可以从不同角度观察商品趋势。对于卖家来选款的话是极好的。

关于选款： 亚马逊和国内天猫的差别在于店铺概念弱化，亚马逊以单品为为单位，基本一个ASIN就是一个商品类型，卖得好的商品很多人可以跟卖。不同的商家会有一样ASIN的商品，如果谁的商品好（省略...）。步骤一般是：通过该平台Web端查看某大类排名前一万名，进行一些筛选，比如价格在20刀的，FBA的商品，然后可以再点进去商品，看这件商品十几天的排名和价格变化等，然后我决定跟卖，先去阿里巴巴批发看看有没有这个东西，有！价格利润很多。好，我们卖！然后每天可以上来平台搜我们这件商品的ASIN，查看最近的变化。

亚马逊选款平台主要用来商品选款，目前有四个站点，美国站，英国站，德国站和日本站。购买该平台的服务后，你可以获取平台的入口网址，登录后进行操作。通过该平台，你可以从多维度数据进行分析，以便选款，正文将会从平台的各个功能进行详细介绍。

# 授权相关

服务声明：

此平台服务包括采集端和网站端，需要的最基本环境：2G内存，另挂200G硬盘的ubuntu服务器（自费或托管），代理IP机构提供的可用API地址（自费）

本团队提供独立产品的部署服务，和次低版本源码的提供。版权所有，侵权必究|署名-非商业性使用-禁止演绎 4.0 国际，商业授权请联系邮箱：gdccmcm14@live.com

关于源码授权有几个原则:

1. 源码是IT工作中最重要的知识产权，除非商业授权，或者该源码开源协议允许，否则不允许他人进行拷贝，修改，删除，添加等二次开发。授权源码不得扩散和商业传播，只能供自己内部使用。

2.防止行业竞争者的模仿和剽窃，团队只提供服务，不提供与服务相关的代码粒度上的技术细节咨询，不会提供架构图，部署流程，逻辑流程等，不提供源码任何层面的技术指导和咨询，开源代码除外，。

4.商业授权允许的源码版本要与目前相应版本后退两个版本，以保护从业者的经济效益。

# 正文

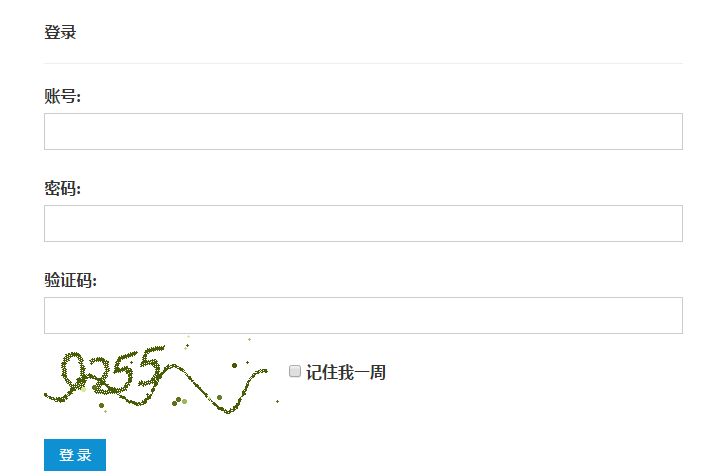
平台主要有以下几部分构成：权限系统，商品系统。

权限系统主要有：分配子账户，分配角色，分配权限，登录等

商品系统主要有：小类数据，大类数据，类目数据，ASIN数据等

## 一．登录

浏览器输入域名会自动进入登录界面，输入账号密码即可，记住一周可一周内免登录。



## 二．用户管理

进入平台后，你将会看到登录的时间，和相应功能的导航图。



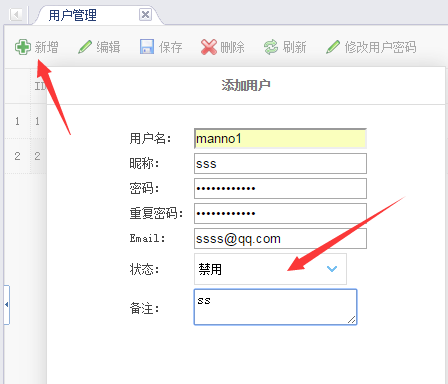
您可以点击权限中心-用户管理，进行用户的增加和删除，激活等，如果子员工的密码忘记了，可以由有权限的管理员修改密码。

用户管理看出用户在哪个地址进行登录，登录的次数。



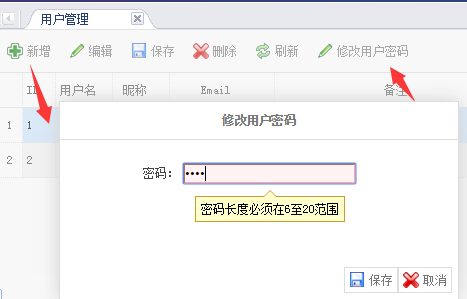
### 2.1.新增用户

新增用户你只需点击新增，填相应的信息保存即可，如果状态为禁用，账号未激活，不可登录。



### 2.2.修改密码

鼠标单击表格的行，然后点击修改用户密码，即可修改相应行账号的密码。

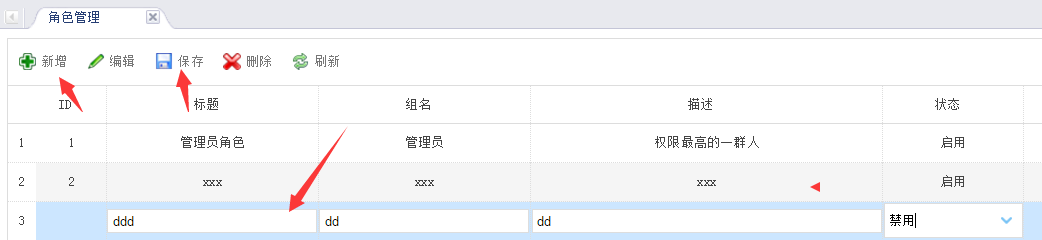


## 三．角色管理

角色管理是指建立一个群组，该群组有某些应用功能的使用权限，可以将批量的用户归类到这些群组，这些用户即可访问相应的功能，否则提示无权限。

### 3.1.新增角色

点击新增按钮，出现编辑框，进行编辑后，按保存按钮即可。

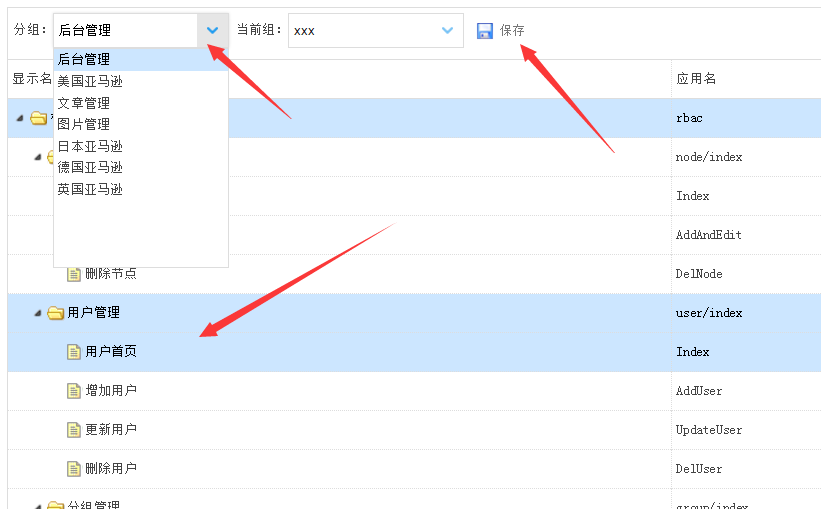


### 3.2.角色授权

新建角色后，为角色授权一定的权限，请点击授权。



进入授权页面后可以点击相应的分组，给相应的组，即角色，赋予一定的权限，单击下方的应用名，然后保存即可。如果删除用户应用名没被勾选，那么该角色下方的用户不可以执行删除用户操作。



比如美国亚马逊，美国类目数据更新未被勾选，那么该角色下的用户不可以更改类目的采集策略。



### 3.3.用户授权

点击用户列表，进入页面后，可以为将用户归到某个角色里。



选择当前组，即角色，单击下方的用户，保存即可



## 四．类目数据

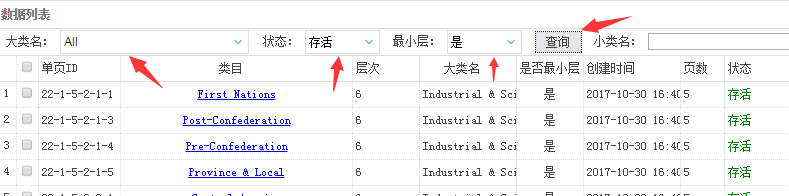
四个站点我们以美国站为例，选款的基础数据主要包括以下部分。

1. 类目数据，主要是调整采集的策略，表示那哪些类目抓取，哪些不抓，抓几页等。
2. 小类数据，主要指从Top100抓取到的第一丛商品数据，有精确的小类排名，商品基本价格，标题，图片等。
3. 大类数据，从小类数据进入商品的详情页，获取得到的其他数字字段，比如在整个大类的排名，商品的评论数等。
4. ASIN数据，主要是我们维护的一个商品库，可以看出某件商品在某段时间的价格变化，排名变化，出现在Top100的次数等。

类目数据可以调整哪个类目需要采集，哪个不用。

### 4.1.查询类目

通过选择大类名，和其状态和层数，点击查询即可。状态为死亡的表明没有进入采集计划，最小层指的是类目是否是最底层的类目，比如下方查询到了类目信息。

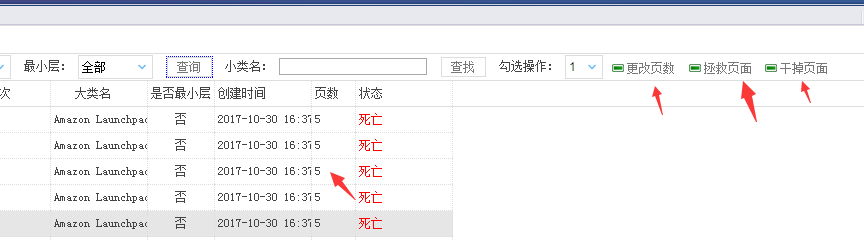


查询某一个分类所属大类的信息和级别，可以在小类名框填入，然后点击查找即可。



### 4.2.调整类目

通过单击表格下的分类，点击拯救和干掉页面，可以让该分类死亡或存活，调整后第二天相应的类目将不会被采集到。



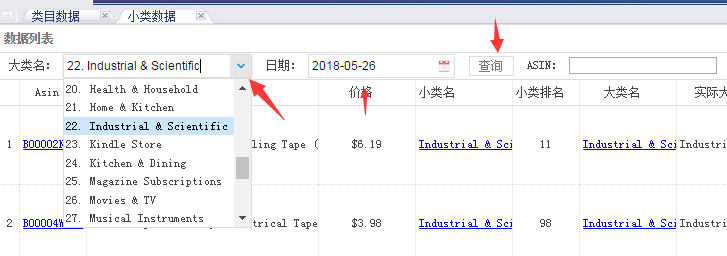
更改页数可以使采集的时候只采集Top100的前几页，每页20个。

## 五．小类数据

小类数据是指从TOP100页面上采集到的数据，比如该类目的地址：<https://www.amazon.com/Best-Sellers-Industrial-Scientific-3D-Printing-Filament/zgbs/industrial/6066129011>

### 5.1.筛选查询

选择大类名，和指定的日期，点击查询，可以查询到当天某时段TOP100的排名和相应的数据信息。

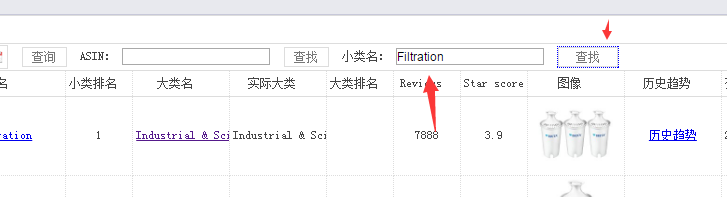


### 5.2.精确查询

点击ASIN框，点击查找，可以查找到某一ASIN的具体信息，如小类排名，评论数，星数等。



填写小类名，点击查找可以查询精确小类下排名情况，列表时间表示采集到的数据，点击历史趋势会查看商品变化趋势。

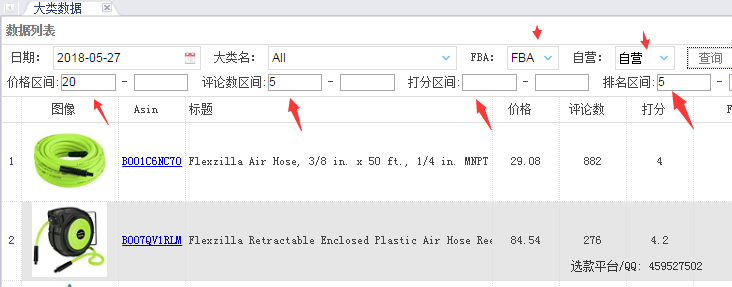


## 六．大类数据

大类数据字段更全，筛选条件更多，支持导出数据等。

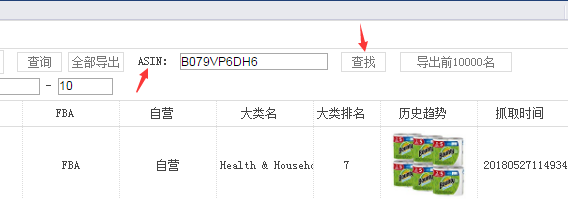
### 6.1.筛选查询

可筛选FBA，自营，价格区间，评论数区间等，查询某天的数据信息，如下筛选出20180527当天某时期FBA自营，价格20美元以上，评论数5以上，大类排名5-10名的商品。点击查询即可。



### 6.2.精确查找

某一商品在小类数据查找的数据字段不全，可以在大类数据进一步精确查找。如下图。



### 6.3.数据导出

导出数据有两种，一种是导出前10000名的商品数据，一种是导出条件筛选查询出列表中的数据。比如点击导出前10000名按钮，将会下载一个文件，比如big-Electronics-20180527.csv



对于10000名以后的，或者其他的数据，我们可以点击查找，显示出列表数据后，再点全部导出按钮。



## 七．ASIN数据

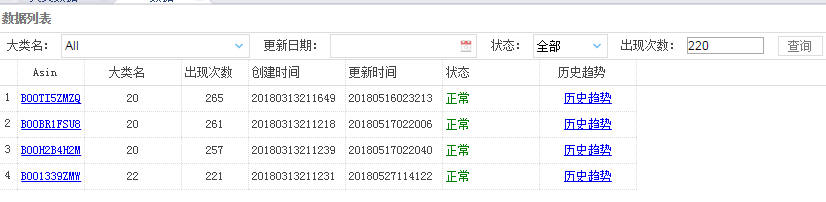
Asin数据是一个商品库，我们以往采集过的商品数据将会在这个页面显示，你可以在此查找任意一段时期某ASIN的情况。大类数据和小类数据都只能查询某一天的数据，而ASIN数据可以查找某段时间的数据。

### 7.1.筛选查询

可选选择大类名，可选选择更新日期，如果选择更新日期，只筛选出在该日期内更新的数据（时间段筛选还未开发）。数据列表按出现次数降序，出现次数指该商品出现在热卖榜的数量。状态一般为正常，某些ASIN会下架后，会出现注销的状态。

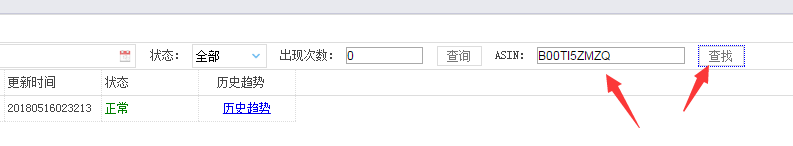


次数筛选，如果出现次数的框填写不为0，那么筛选出大于该数字的商品，比如下方只筛选出出现大于220次的商品。



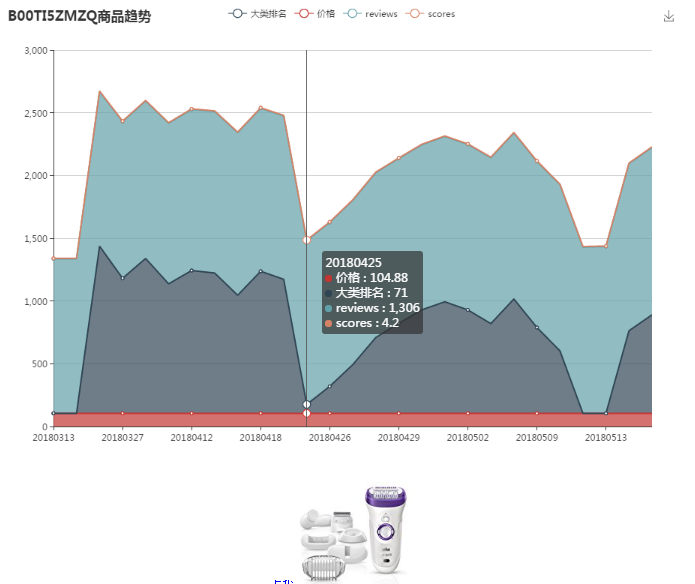
### 7.2.精确查找

在大类数据和小类数据查找ASIN需要指定日期，在此不需要指定日期。只需在ASIN框填写后，点击查找即可。



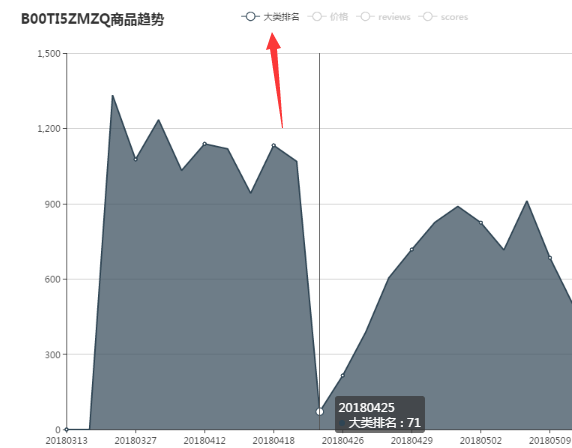
### 7.3.历史趋势

该功能是可视化最友好的功能，通过此功能，你可以查看一个商品近段时间的采集状况，排名，价格等的变化。

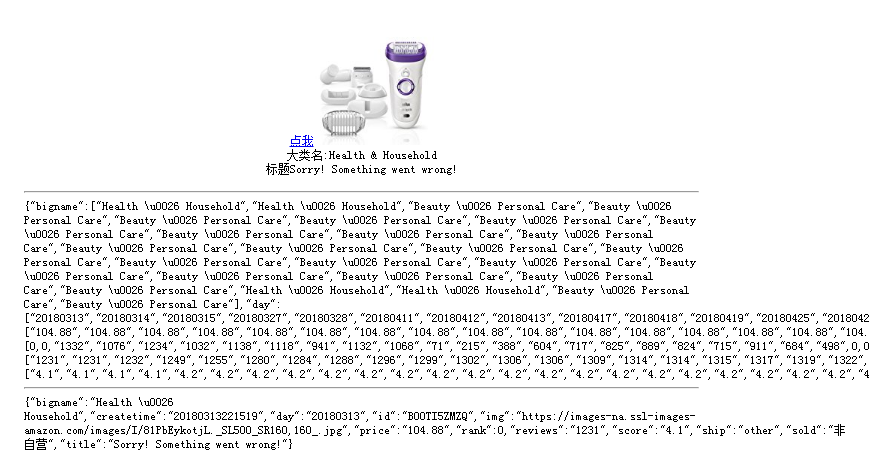


比如，以上该商品在20180313到20180526接近出现了30多次，大类排名从前到后波动，表明该商品在总类中销售情况不如其他商品。或者因为周末原因，导致在周末大类排名靠前。

点击相应的按钮，可以分开不同的维度。



因为某些时候出现的大类排名采集不到，出现的排名为0，此等情况请复制相应的数据，在Excel自行分析。

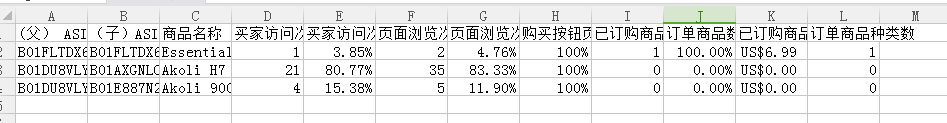


## 报告数据

报告数据是内部功能，可以进行内部KPI考核，比如做广告前后订单数目的变化。你需要从亚马逊后台导出每天的报表，手动导进系统。



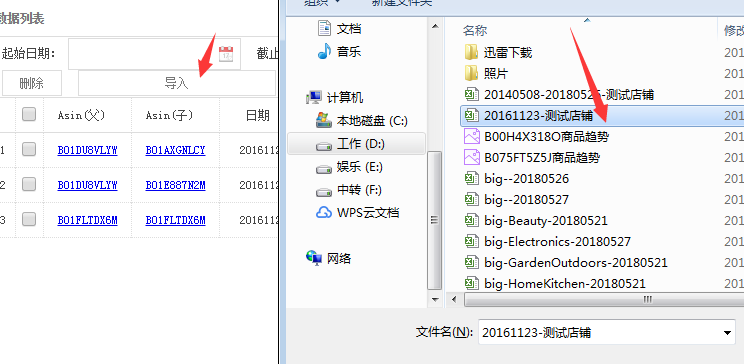
从亚马逊后台导出的报表格式如下：



你需要按照“日期-店铺名“的形式重命名。



点击导入按钮后导入。



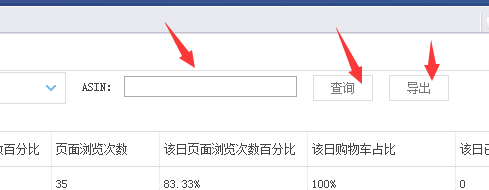
导入后你可以按指定日期，店铺查看：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 买家访问次数 | 买家访问次数百分比 | 页面浏览次数 | 页面浏览次数百分比 | 购买按钮页面浏览率 | 已订购商品数量 | 已订购商品销售额 | 订单商品种类数 |

点击筛选条件，点击查找可以进行筛选。



您还可以导出数据和精确查找某一Asin数据。



# 附录

## 历史

开发这个产品从2016年10月就开始了, 迭代从1.0.0到2.5.0

采集端支持：

1. 列表页和详情页可选择代理方式

2. 多浏览器保存cookie机制

3. 机器人检测达到阈值自动换代理

4. 检测日期过期自动停止程序

5. IP池扫描周期填充代理IP

6. 支持分布式跨平台抓取

7. 高并发进程设置抓取

8. 默认网页爬取去重

9. 日志记录功能

10. 配套可视化网站，支持多角度查看数据，小类数据，大类数据，Asin数据和类目数据，支持查看每件Asin商品的历史记录，如排名，价格，打分，reviews变化。部分数据支持导出，且网站支持RBAC权限，可分配每部分数据的查看和使用权限。

11. 网络端监控爬虫，可查看爬虫当前时段数据抓取状态，爬取的进度，IP的消耗程度。 \*\*可支持网络端启动和停止爬虫，彻底成为Saas\*\*（待做）

12. 可自定义填入IP，如塞入其他代理IP网站API获取的IP

13. 可选择HTML文件保存本地

分布式，高并发，跨平台，多站点，多种自定义配置，极强的容错能力是这个爬虫的特点。机器数量和IP代理足够情况下，每天每个站点可满足抓取几百万的商品数据。

此项目可以持续优化成功一个更好的平台, 因为国内目前还没有像淘宝数据参谋一样的亚马逊数据参谋. 由于高并发百万级每天导致的数据抓取速度问题, 和数据获取后的清洗和挖掘问题, 我们可以在以下方面做得更好。

1. 首先数据抓取速度保证和爬虫部署问题, 可以采用`Docker`自动构建, 构建`kubernetes`集群进行`deployments`部署, 自动跨容和缩容爬虫服务, 分布式爬虫不再需要手工上去跑任务.

2. 其次数据保存在`MYSQL`产生的分表问题, 因为`MYSQL`是非分布式的集中式关系型数据库, 大量数据导致数据查找困难, 多表间数据`union`和`join`困难, 所以可以采用`ElasticSearch`来替换`MYSQL`, 著名的`JAVA Nutch搜索引擎框架`使用的就是`ES`.

3. 最后, 关于数据获取后的清洗和挖掘问题, 是属于离线操作问题, 保存在`ES`的数据本身支持各种搜索,`ES`的文本搜索能力超出你的想象, 一般需求可以满足, 不能满足的需求则要从`ES`抽取数据, 构建不同主题的数据仓库进行定制化挖掘. 此部分, 需要开发另外的项目.

4. 配套的`UI`网站端可以有更好的用户体验, 目前基本可以满足选款的需求, 商品的各种数据优美的显示出来.

## 知乎

亚马逊爬虫比较严格，我采用自己的爬虫框架，高度伪装，高并发ip池轮转获取亚马逊海量商品数据，已经形成稳定选款产品，带界面

开源地址:https://github.com/hunterhug/AmazonBigSpider

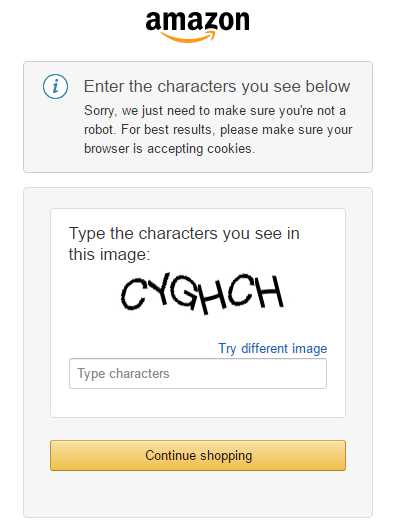
正文:

以下只支持国外 亚马逊(主要是美国，其次是日本，英国和德国)

印度好像也有亚马逊，不知道有没有做印度市场的

一个IP加上头部，加上cookie保持机制，无论你暂不暂停，五百次左右就会被机器人，机器人后，继续爬，30-100次会抓到一张正确的详情页。所以，可以不暂停，直接代理IP去买

机器人如下:



被机器人后，这个IP歇歇4-8个小时，又可以继续爬接近300-500个详情页。

还有，列表页ajax页面不反爬虫(日本反)，厉害了。

动态IP池控制抓数据可以，也是每个IP五百次左右被反，机器人页7kb，详情页最大接近1.4mb，所以机器人也疯狂抓吧，也是可以的。

亚马逊API有限制，某些商品还不对外开放。利用API和爬虫一起搞吧。API目前有两种，广告和卖家API，都可以搞商品数据，一个卖家API一个小时可以得到10万Asin的商品信息。

2016.12.13补充

列表页日本站会反爬虫，德国，日本，英国亚马逊列表页页数不足会出现404，但是美国不会。日本站反爬虫更严格。

出现机器人后，如果你尝试破解验证码，获取超过四百次验证码，以后只会返回一堆乱码给你，所以验证码样本收集也是危险，搞不好，这个IP今天就不能用了。所以，不要试图破验证码，老老实实机器人都不会有问题。

尝试破解验证码:

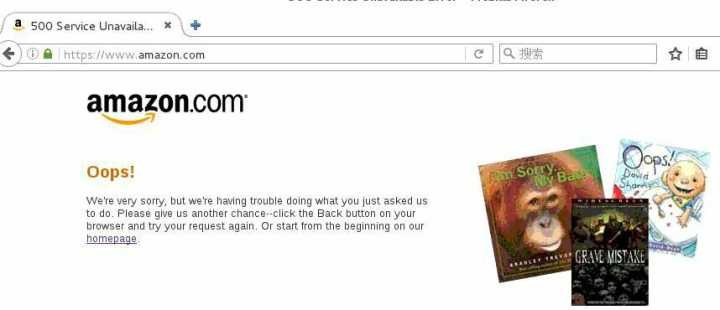


列表页会有几十种样子，大类排名的大类名可能有别名，而且一个商品可能上架又被k掉了，会突然404，然后隔几天又恢复了，厉害了！

每个站的Asin商品在小类列表Top100维持在180万个左右。

补充:

机器人太多可能出现以下: Oops



抓到的数据差不多是这样:

类目链接抓取

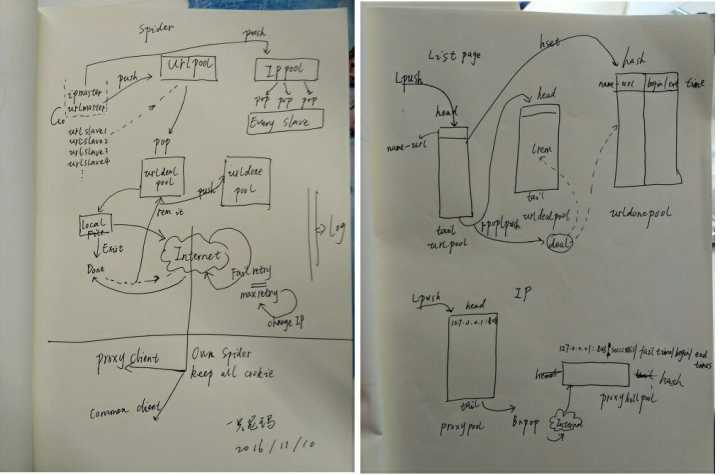


形成类目链接后抓取列表页，再进详情页提取字段。

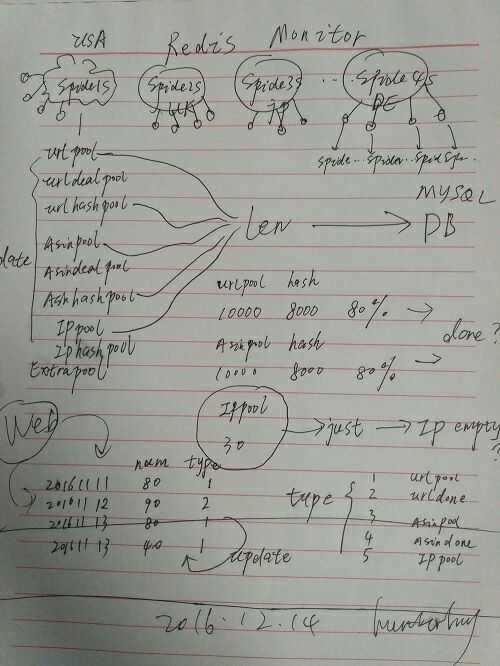
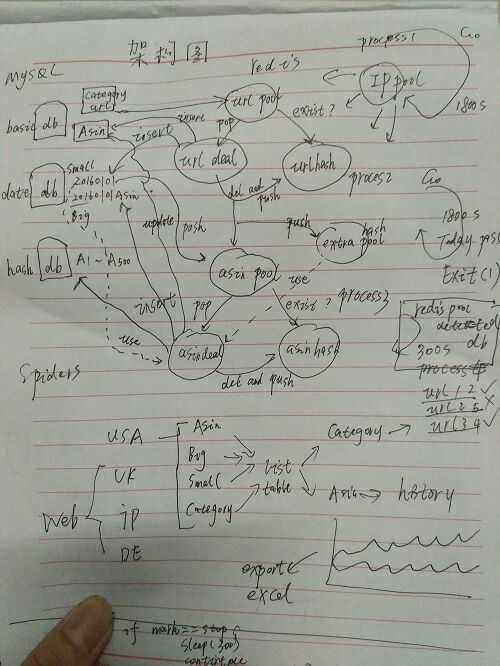


按天数分表不足以满足需求，所以按hash再分表，冗余数据一份，统计商品趋势。

架构图(老的):



(新的)

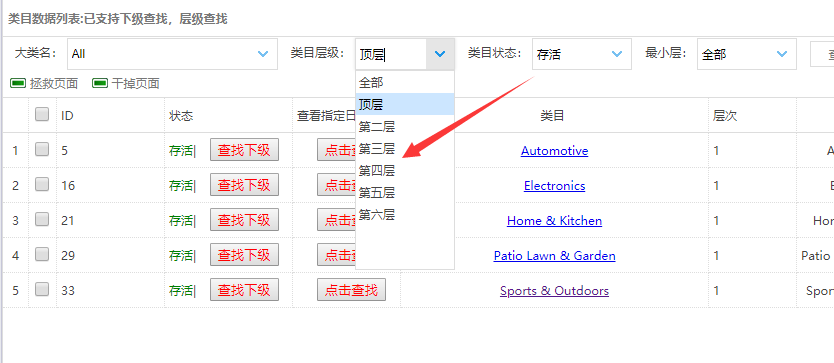


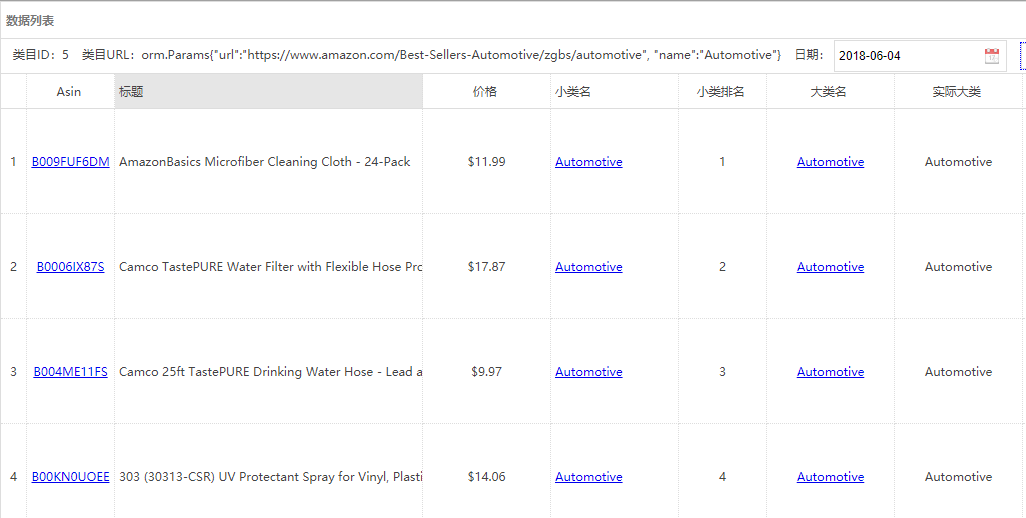
# V1.2补充

## 类目数据

类目数据已经支持查找下级，以及更复杂的查询，可点击查找，查看指定日期排名。







## 监控数据

现在可以查看每天采集的情况



## 导航调整

新增前台功能，前台可定制化作为企业网站。并且导航重新归类，业务数据移到顶层。

