爬虫概述

学习目标：

- 了解 爬虫的概念
- 了解 爬虫的作用
- 了解 爬虫的分类
- 掌握 爬虫的流程(原理)

1.爬虫的概念

> **模拟浏览器，发送请求，获取响应**

网络爬虫（又被称为网页蜘蛛，网络机器人）就是模拟客户端(主要指浏览器)发送网络请求，接收请求响应，一种按照一定的规则，自动地抓取互联网信息的程序。

- **原则上,只要是客户端(浏览器)能做的事情，爬虫都能够做**
- **爬虫也只能获取客户端(浏览器)所展示出来的数据**

1.1.举例说明（百度--搜索引擎爬虫--根据访问权重排序结果）

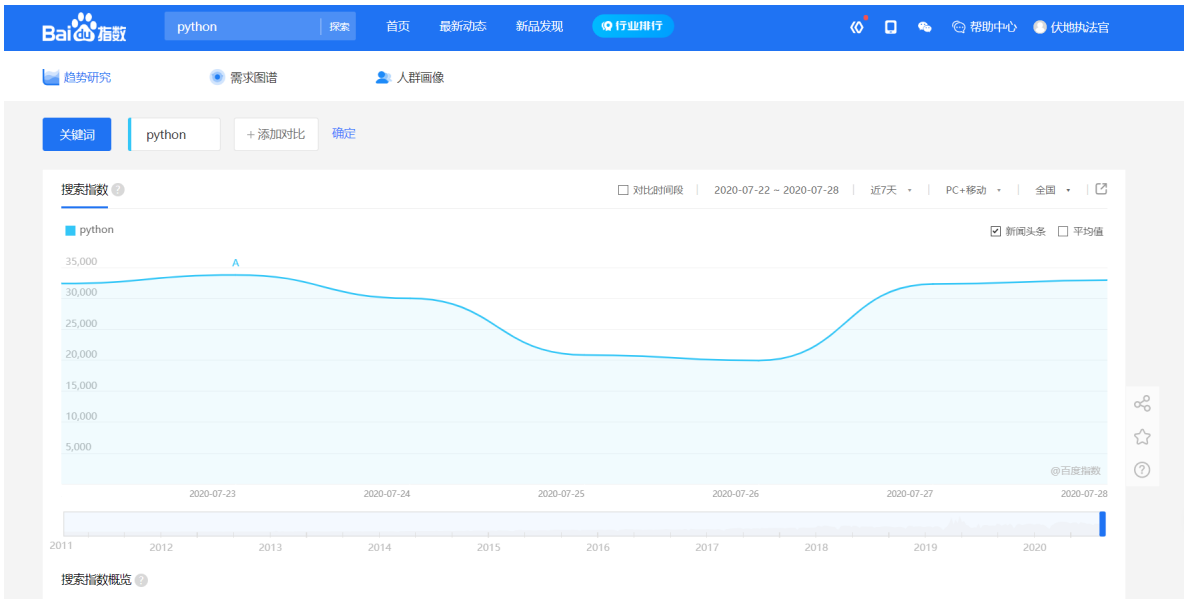
> 用户的访问次数

当我们去百度一个词条的时候，百度这个搜索引擎的爬虫机制，会在互联网中，根据自己的算法收录符合词条的结果，反馈给你，当然，这个结果，会根据百度自身的用户访问统计----(这种统计我们称为访问权重)----进行排序。

所以，我们在研究爬虫的时候，不仅要了解爬虫如何实现，还需要知道一些常见爬虫的算法，如果有必要，我们还需要自己去制定相应的算法，在此，我们仅需要对爬虫的概念有一个基本的了解。

2.爬虫的作用(领域)

> 思考：如今，人工智能，大数据离我们越来越近，很多公司在开展相关的业务但是人工智能和大数据中有一个东西非常重要，那就是数据，但是数据从哪里来呢?



这是百度的百度指数的一个截图，它把用户在百度上的搜索关键词做了一个统计，然后根据统计结果得出一个流行趋势，之后进行了简单的展示。

像微博上的热搜，就是这么一个原理，类似的指数网站还有很多，比如阿里指数，360指数等等，而这些网站有非常大的用户量，他们能够获取自己用户的数据进行统计和分析

那么，对于一些中小型的公司，没有如此大的用户量的时候，他们该怎么办呢?

2.1.数据来源

1.去第三方的公司购买数据（比如：企查查）

2.去免费的数据网站下载数据（比如：国家统计局或者公开的数据源网站）

3.通过爬虫爬取数据

4.人工收集数据（比如：问卷调查）

在上面的数据来源中，人工的方式费时费力，效率低下，免费的数据网站上面的数据质量不佳，很多第三方的数据公司他们的数据往往也是爬虫获取的，所以获取数据最有效的途径就是通过爬虫爬取

```
# 作用领域
1. 数据采集
    1. 抓取微博评论(机器学习舆情监控)
    2. 抓取招聘网站的招聘信息(数据分析、挖掘)
    3. 新浪滚动新闻
    4. 百度新闻网站
2. 软件测试
    1. 爬虫之自动化测试
    2. 虫师
3. 12306抢票
4. 网站上的投票
    1. 投票网
5. 网络安全
    1. 短信轰炸
        1. 注册页面1
        2. 注册页面2
        3. 注册页面3
```

2. web漏洞扫描：
https://jsjxy.xsyu.edu.cn/info/1085/1953.htm

- 人脸识别：您做人工智能是需要大数据的，举个例子您想做一个自动识别人脸的人工智能机器。您首先需要根据人脸生物特征建立AI模型，然后需要几千万或者几十亿张人脸图片进行不断的训练这个模型，最后才得到精准的人脸识别AI。几十亿的人脸图片数据哪里来呢？ 公安局给你？不可能的！一张张去拍照？更不现实啦！ 那就是通过网络爬虫技术建立人脸图像库，比如我们可以通过爬虫技术对facebook、qq头像、微信头像等进行爬取，来实现建立十几亿的人脸图像库。
- 市场分析：电商分析、商圈分析、一二级市场分析等
- 市场监控：电商、新闻、房源监控等
- 商机发现：招投标情报发现、客户资料发掘、企业客户发现等

3.爬虫的分类

根据被爬取网站的数量不同，可以分为：

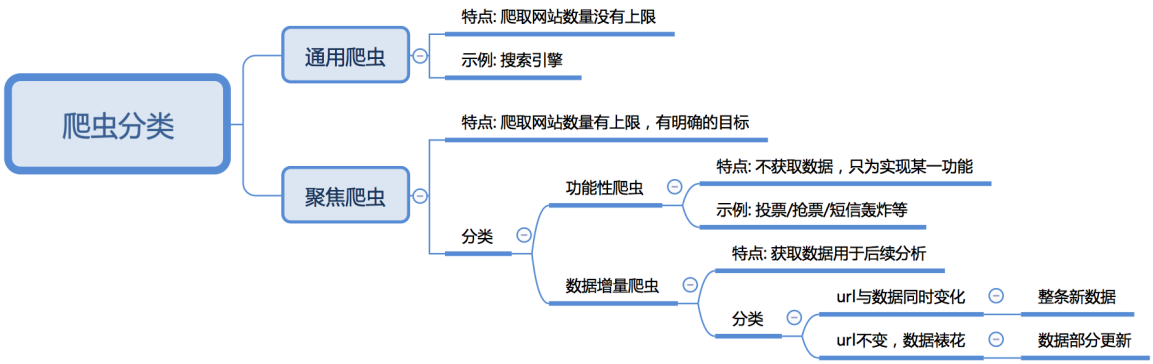
- 通用爬虫，如 搜索引擎
- 聚焦爬虫，如12306抢票，或专门抓取某一个（某一类）网站数据

根据是否以获取数据为目的，可以分为：

- 功能性爬虫，给你喜欢的明星投票、点赞（不抓取数据，只为实现某一个功能）
- 数据增量爬虫，比如招聘信息（获取数据，用于后续分析）

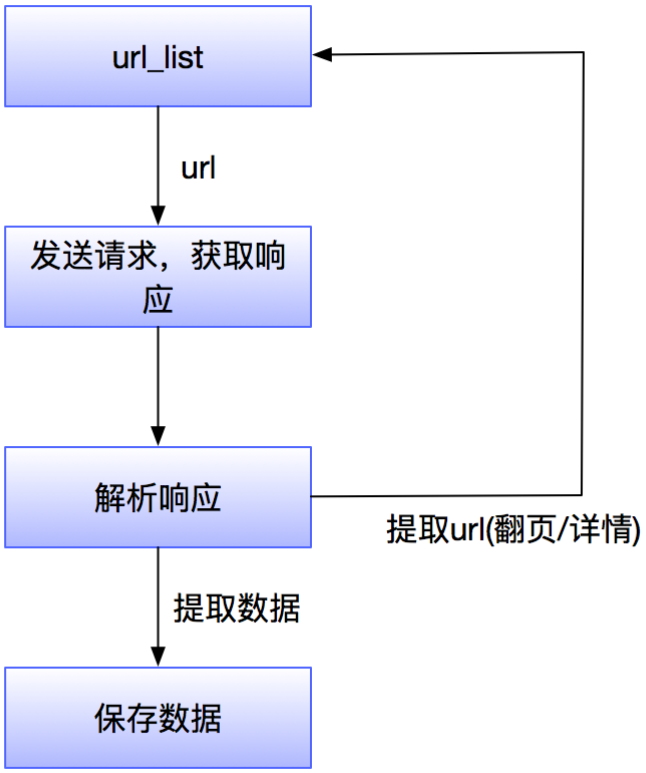
根据url地址和对应的页面内容是否改变，数据增量爬虫可以分为：

- 基于url地址变化、内容也随之变化的数据增量爬虫
- url地址不变、内容变化的数据增量爬虫



4.爬虫的流程(原理)

> 如图所示



#原理

#开发流程

1. 获取一个url

----->1.准备数据(请求的地址，headers，请求参数，用户输入等等)

2. 向url发送请求，并获取响应（需要http协议）

----->2.发送请求，获取响应

3. 如果从响应中提取url,则继续发送请求获取响应

----->3.解析响应，数据提取(url--继续请求，数据--执行第4步)

4. 如果从响应中提取数据，则将数据进行保存

----->4.保存数据

1. 确定目标网站（确定网址。爬取哪一个网站)
2. 程序模拟浏览器向服务器发请求，(要数据)
3. 服务器给你响应(响应对象)

4. 数据的精确提取
5. 数据保存