

# 第一章：初识网络爬虫

---

- 网络爬虫的产生背景（大数据时代）
  - 大（数据量大）
  - 多（类型繁多）
  - 低（价值密度低）
  - 快（处理速度快）
- 网络爬虫的定义
  - 搜索引擎的工作流程
  - 互联网上运行的爬取数据的程序，遵循互联网规则（协议TCP/IP）
- 网络爬虫的用途
  - 爬取图片
  - 去除广告
  - 金融分析
  - 学情预警
  - .....
- 网络爬虫的分类
  - 爬取的场景
    - 全网（通用）爬虫
    - 主题（聚焦）爬虫
  - 爬取的形式
    - 累计式爬虫
    - 增量式爬虫
  - 爬虫的存放形式
    - 表层爬虫
      - 类似广度优先算法
    - 深层爬虫
      - 类似深度优先算法