

网络数据采集

2021年11月11日 16:14

大数据的海量、多样、高速和易变特性带来冲击
对传统数据采集、存储、管理方法提出新的挑战（传统的数据采集来源单一，一般通过关系型数据库和并行数据仓库进行处理，数据采集、存储和分析工具已无法满足大数据处理分析的基本需求）

网络数据采集来源（目的、场合、需求者）
以互联网为载体存在的、用户容易直接获取的数据
采集：网络爬虫或API（应用程序编程接口）

信息获取的重要方式：搜索引擎、数据库
搜索引擎：通过关键词和检索技巧，能够满足用户的一般性需求
数据库：提供极具价值、全面的数据信息

研究型用户或商业型用户：要想获得感兴趣的足量的数据，各大数据开放平台

网络爬虫

控制框架：控制器、解析器、资源库

1. 通用网络爬虫（用于搜索广泛主题，在搜索引擎中具有重要的应用价值）

2. 聚焦网络爬虫（“主题网络爬虫”，程序过滤与主题无关的链接，只抓取与主题相关的页面信息）

3. 增量式网络爬虫（增量式地抓取新产生的页面或更新已变化的页面，并不重新遍历未发生变化的页面）

4. 深层网络爬虫（深层页面：无法通过静态链接获取、存储在后台数据库、通过接口才能获得信息资源的Web页面，与可通过超链接获取信息、静态的表层页面相区别）

从一个URL集合开始运行，分析页面内容并提取新的URL到待提取的URL队列，爬虫如此往复，遍历整个Web，直到URL队列为空或满足爬虫终止条件（搜索时长、数量）

节省了大量的资源和开销，页面更新速度相比通用网络爬虫更快，能够满足特定领域的信息需求


极大地降低了爬虫在时间和空间上的开销，并保证资源池中的页面尽可能的新

- 考虑的问题：
- 1. 网站是否提供API接口
 - 2. 论文作者是否提供科研数据
 - 3. 数据交易平台是否提供免费数据

网络爬虫（Web Spider/Crawler）是一个自动获取网页信息的计算机程序，是搜索引擎的重要组成部分。如果把互联网比作蜘蛛网，那网络爬虫就是在这张网络游走爬行的蜘蛛（Spider），其像爬行者（Crawler）一样在网络空间按照一定的规则获取信息。

- 网络日志的采集
- 日志数据采集是企业获取海量数据的重要方式（通过分布式架构，满足海量数据采集及传输需求）
- 网络日志：在服务器或Web应用程序上有关网络访问等用户行为的各种日志文件（访问日志、引用日志、代理日志、错误日志），包含了大量的用户访问信息
- 产生不受人为因素的影响
- 1. 客户端网络日志
 - 2. 代理服务器端网络日志
 - 3. Web服务器端网络日志

个性化的数据分析，需要依据业务情况进行开发和定制，不能依赖于开源日志工具。



大数据驱动产业发展，是企业的核心资产和竞争力，全面、准确的数据量有助于企业研判。一些企业也针对业务需要建立了系统日志采集平台。