就业班系列课程

爬虫基础1

网络爬虫基础

网络爬虫定义

- ✓ 网络爬虫（Crawler）又被称为网页蜘蛛（Spider），网络机器人，网页追逐者，它是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本
- ✓ 狭义与广义定义：狭义上指遵循标准的http协议，利用超链接和Web文档检索方法遍历万维网的软件程序；而广义的定义则是能遵循http协议，检索Web文档的软件都称之为网络爬虫

网络爬虫基础

网络爬虫用途

- ✓ 主要用途：数据采集

  金融，金融新闻/数据，制定投资策略，进行量化交易

  旅游，各类信息，优化出行策略

  电商，商品信息，比价系统

  游戏，游戏论坛，调整游戏运营

  银行，个人交易信息，征信系统/贷款评级

  招聘，职位信息，岗位信息

  舆情，各大论坛，社会群体感知，舆论导向

- ✓ 其他用途：12306抢票、各种抢购、投票、刷票、短信轰炸、网络攻击、Web漏洞扫描器

网络爬虫基础

爬虫原理

- ✓ 网络蜘蛛从一组要访问的URL链接开始（种子URL），爬虫访问这些链接，它辨认出这些页面的所有超链接，然后添加到这个URL列表并按照一定的策略反复访问这些URL

网络爬虫基础

网络爬虫是否合法？

- ✓ 从目前的实践来看，如果抓取数据的行为用于个人使用，则不存在问题：而如果数据用于转载或者商业用途，那么抓取的数据类型就非常关键
- ✓ 从很多历史案件来看，当抓取的数据是现实生活中的真实数据（比如，营业地址、电话清单）时，是允许转载的。但是，如果是原创数据（比如，意见和评论），通常就会受到版权限制
- ✓ 无论如何，当你抓取某个网站的数据时，请记住自己是该网站的访客，应当约束自己的抓取行为，否则他们可能会封禁你的IP，甚至采取更进一步的法律行动。这就要求下载请求的速度需要限定在一个合理值之内，并且还需要设定一个专属的用户代理来标识自己

网络爬虫基础

- 反爬虫
  - ✓ 初学者写的爬虫：简单粗暴，不管对端服务器的压力，甚至会把网站爬挂掉了
  - ✓ 数据保护：很多的数据对某些公司网站来说是比较重要的不希望被别人爬取
  - ✓ 商业竞争问题：这里举个例子是关于京东和天猫，假如京东内部通过程序爬取天猫所有的商品信息，从而做对应策略这样对天猫来说就造成了非常大的竞争

网络爬虫基础

网络爬虫分类

- ✓ 根据使用场景，网络爬虫可分为通用爬虫和聚焦爬虫两种
- ✓ 通用爬虫，搜索引擎和web服务商用的爬虫系统。通用网络爬虫是搜索引擎抓取系统（Baidu、Google、Yahoo等）的重要组成部分。主要目的是将互联网上的网页下载到本地，形成一个互联网内容的镜像备份
- ✓ 聚焦爬虫，是"面向特定主题需求"的一种网络爬虫程序，它与通用搜索引擎爬虫的区别在于：聚焦爬虫在实施网页抓取时会对内容进行处理筛选，尽量保证只抓取与需求相关的网页信息

网络爬虫基础

通用搜索引擎工作原理

- ✓ 尽可能的把互联网上的所有的网页下载下来，放到本地服务器里形成备份，再对这些网页做相关处理(提取关键字、去掉广告)，最后提供一个用户检索接口
- ✓ 通用网络爬虫从互联网中搜集网页，采集信息，这些网页信息用于为搜索引擎建立索引从而提供支持，它决定着整个引擎系统的内容是否丰富，信息是否即时，因此其性能的优劣直接影响着搜索引擎的效果

网络爬虫基础

- 通用搜索引擎工作原理
  - ✓ 第一步：抓取网页
  - ✓ 首先选取一部分种子URL，把这些URL放到待爬取队列
  - ✓ 从队列取出URL，然后解析DNS得到主机IP，然后保存这个IP对应的服务器里下载HTML页面，保存到搜索引擎的本级服务器，之后把这个爬过的URL放入已爬过的队列
  - ✓ 分析这些网页内容，找出网页里其他的URL链接，继续执行第二步，直到爬取结束

网络爬虫基础

- 通用搜索引擎工作原理

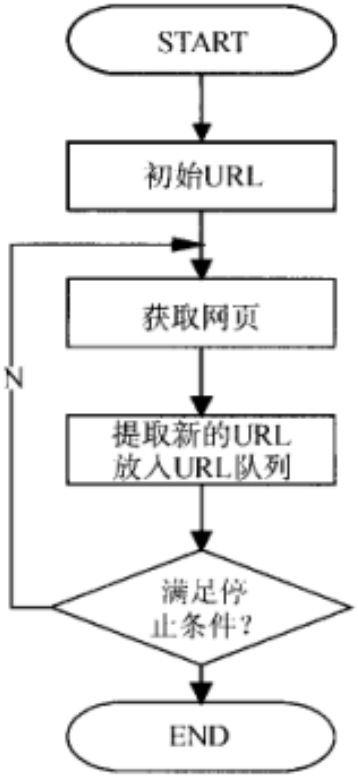


图 1　通用网络爬虫工作流程图

网络爬虫基础

- 通用搜索引擎工作原理
  - ✓ 第二步：数据存储
  - ✓ 搜索引擎通过爬虫爬取到的网页，将数据存入原始页面数据库，其中的页面数据与用户浏览器得到的HTML是完全一样的
  - ✓ 搜索引擎蜘蛛在抓取页面时，也做一定的重复内容检测，一旦遇到访问权重很低的网站上有大量抄袭、采集或者复制的内容，很可能就不再爬行

网络爬虫基础

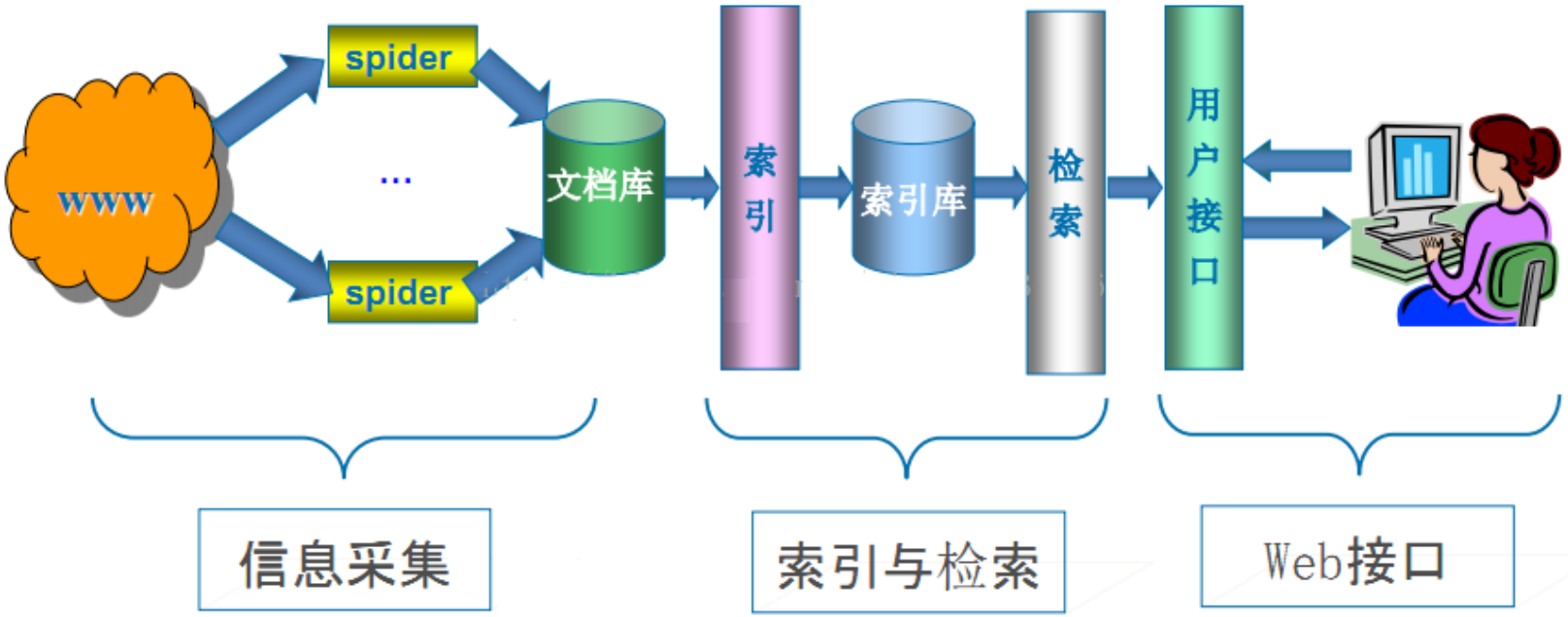
- 通用搜索引擎工作原理
  - ✓ 第三步：数据预处理，搜索引擎将爬虫抓取回来的页面，进行各种步骤的预处理
  - ✓ 提取文字
  - ✓ 中文分词
  - ✓ 消除噪音（比如版权声明文字、导航条、广告等……）
  - ✓ 索引处理
  - ✓ 链接关系计算
  - ✓ 特殊文件处理
  - ✓ ....

网络爬虫基础

- 通用搜索引擎工作原理
  - ✓ 第四步：提供检索服务，网站排名
  - ✓ 搜索引擎在对信息进行组织和处理后，为用户提供关键字检索服务，将用户检索相关的信息展示给用户
  - ✓ 同时会根据页面的PageRank值（链接的访问量排名）来进行网站排名，这样Rank值高的网站在搜索结果中会排名较前，当然也可以直接使用Money购买搜索引擎网站排名，简单粗暴

网络爬虫基础

- 通用搜索引擎工作原理



网络爬虫基础

通用搜索引擎的局限性

- ✓ 通用搜索引擎所返回的结果都是网页，而大多情况下，网页里90%的内容对用户来说都是无用的
- ✓ 不同领域、不同背景的用户往往具有不同的检索目的和需求，搜索引擎无法提供针对具体某个用户的搜索结果
- ✓ 万维网数据形式的丰富和网络技术的不断发展，图片、数据库、音频、视频多媒体等不同数据大量出现，通用搜索引擎对这些文件无能为力，不能很好地发现和获取
- ✓ 通用搜索引擎大多提供基于关键字的检索，难以支持根据语义信息提出的查询，无法准确理解用户的具体需求

网络爬虫基础

- 解决通用爬虫的缺点，聚焦爬虫出现了
  - ✓ 聚焦爬虫，爬虫程序员写的针对某种特定内容爬虫
  - ✓ 面向主题爬虫、面向需求爬虫：会针对某种特定的容去爬取信息，而且保证内容需求尽可能相关
- 聚焦爬虫流程
  - ✓ 聚焦爬虫根据一定的网页分析算法过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止

网络爬虫基础

- 两种爬虫比较

	通用网络爬虫	聚焦爬虫
目标	通用网络爬虫的目标是尽可能多的采集信息页面，而在这一过程中它并不太在意页面采集的顺序和被采集页面的相关主题。这需要消耗很多的系统资源和网络带宽，并且对这些资源的消耗并没有换来采集页面的较高利用率	聚焦爬虫的目标是尽可能快地爬行、采集尽可能多的与预先定义好的主题相关的网页。聚焦爬虫可以通过对整个Web按主题分块采集，并将不同块的采集结果整合到一起，以提高整个Web的采集覆盖率和页面利用率

网络爬虫基础

- URL的搜索策略
  - ✓ 基于IP地址搜索策略
  - ✓ 广度优先
  - ✓ 深度优先
  - ✓ 最佳优先

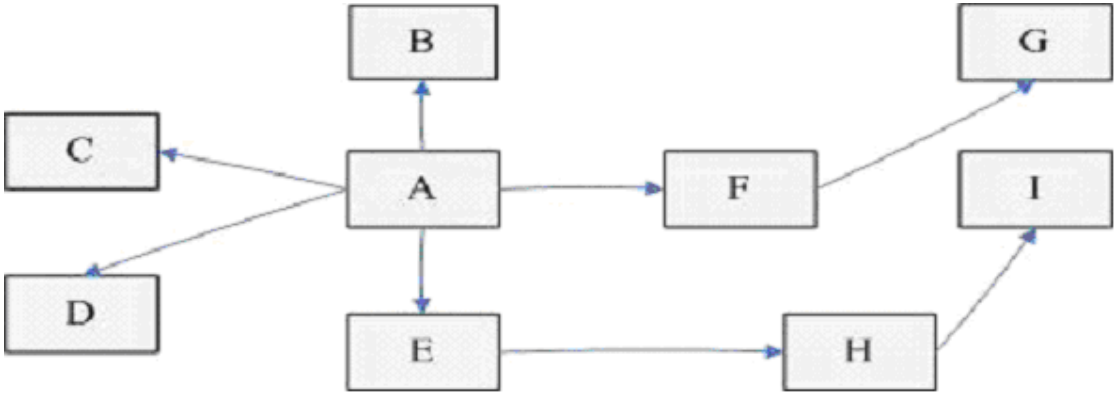
网络爬虫基础

- 基于IP地址搜索策略
  - ✓ 先赋予爬虫一个起始的IP地址，然后根据IP地址递增的方式搜索本口地址段后的每一个WWW地址中的文档，它完全不考虑各文档中指向其它Web站点的超级链接地址
  - ✓ 优点是搜索全面，能够发现那些没被其它文档引用的新文档的信息源
  - ✓ 缺点是不适合大规模搜索

网络爬虫基础

广度优先搜索策略

- ✓ 广度优先搜索策略是指在抓取过程中，在完成当前层次的搜索后，才进行下一层次的搜索。这样逐层搜索，依此类推
- ✓ 该算法的设计和实现相对简单。为覆盖尽可能多的网页，一般使用广度优先搜索方法



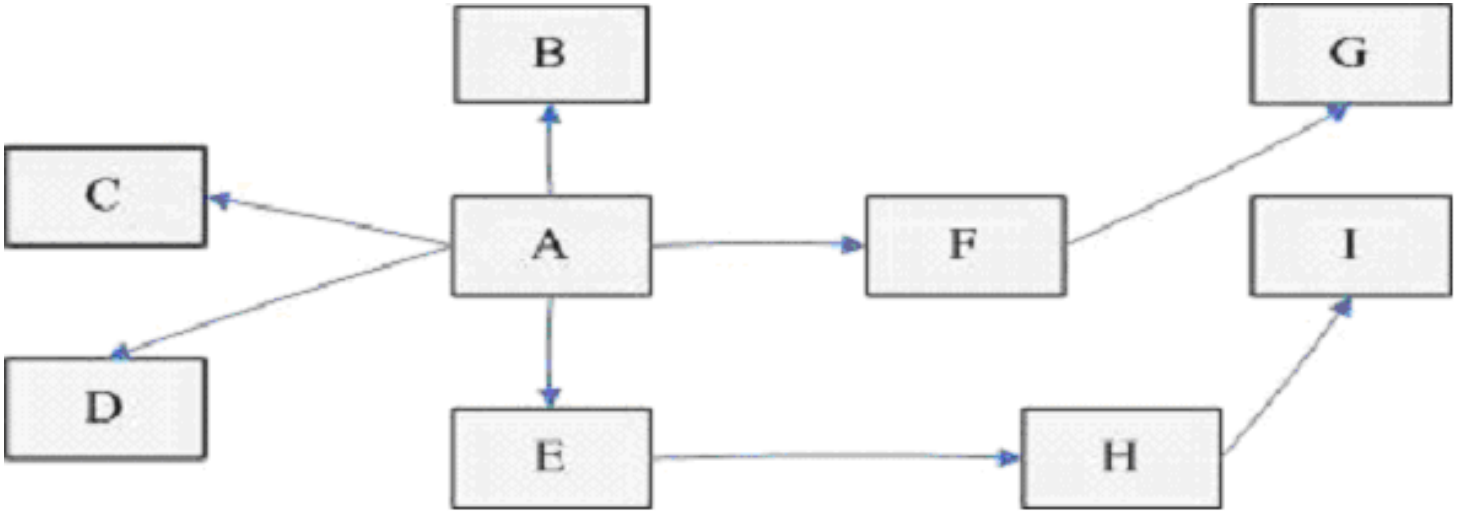
使用广度优先策略抓取的顺序为：A-B、C、D、E、F-G、H-I

网络爬虫基础

- 深度优先搜索策略
  - ✓ 深度优先搜索在开发网络爬虫早期使用较多的方法之一，目的是要达到叶结点，即那些不包含任何超链接的页面文件
  - ✓ 从起始页开始在当前HTML文件中，当一个超链被选择后，被链接的HTML文件将执行深度优先搜索，一个链接一个链接跟踪下去，处理完这条线路之后再转入下一个起始页，继续跟踪链接，即在搜索其余的超链结果之前必须先完整地搜索单独的一条链。当不再有其他超链可选择时，说明搜索已经结束。

网络爬虫基础

- 深度优先搜索策略



使用深度优先策略抓取的顺序为：A-F-G、E-H-I、B、C、D

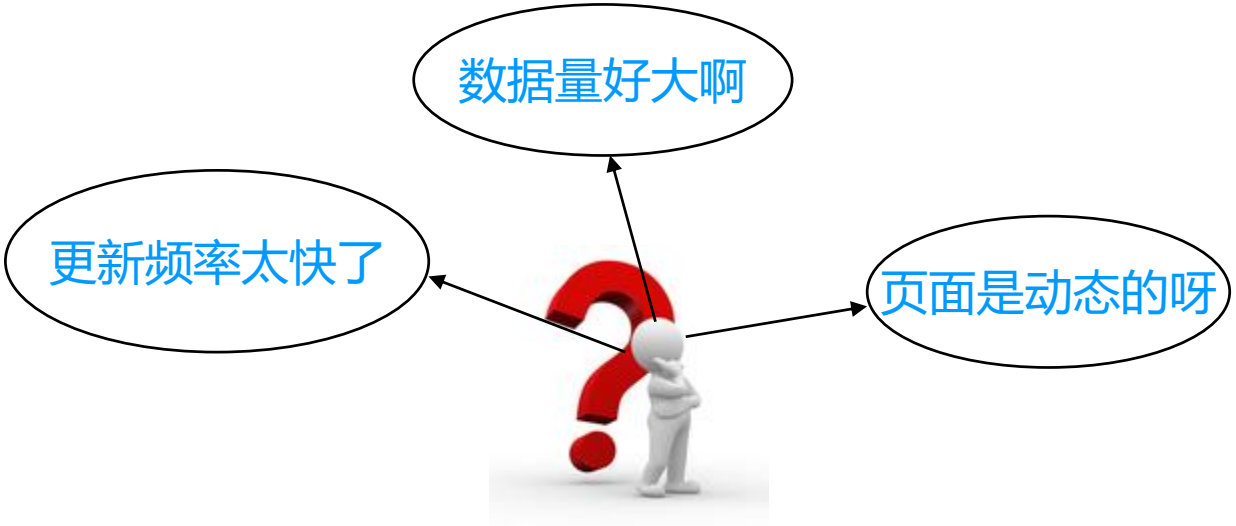
网络爬虫基础

最佳优先搜索策略

- ✓ 最佳优先搜索策略按照一定的网页分析算法，先计算出 URL描述文本的目标网页的相似度，设定一个值，并选取评价得分超过该值的一个或几个URL进行抓取。它只访问经过网页分析算法计算出的相关度大于给定的值的网页
- ✓ 存在的一个问题是，在爬虫抓取路径上的很多相关网页可能被忽略，因为最佳优先策略是一种局部最优搜索算法。因此需要将最佳优先结合具体的应用进行改进，以跳出局部最优点
- ✓ 有研究表明，这样的闭环调整可以将无关网页数量降低30%-90%

网络爬虫基础

- 爬行策略



三种网络特征使得设计网
页爬虫抓取策略变得很难

Talk is cheap

Show me the

CODE