爬虫技术

1.概念

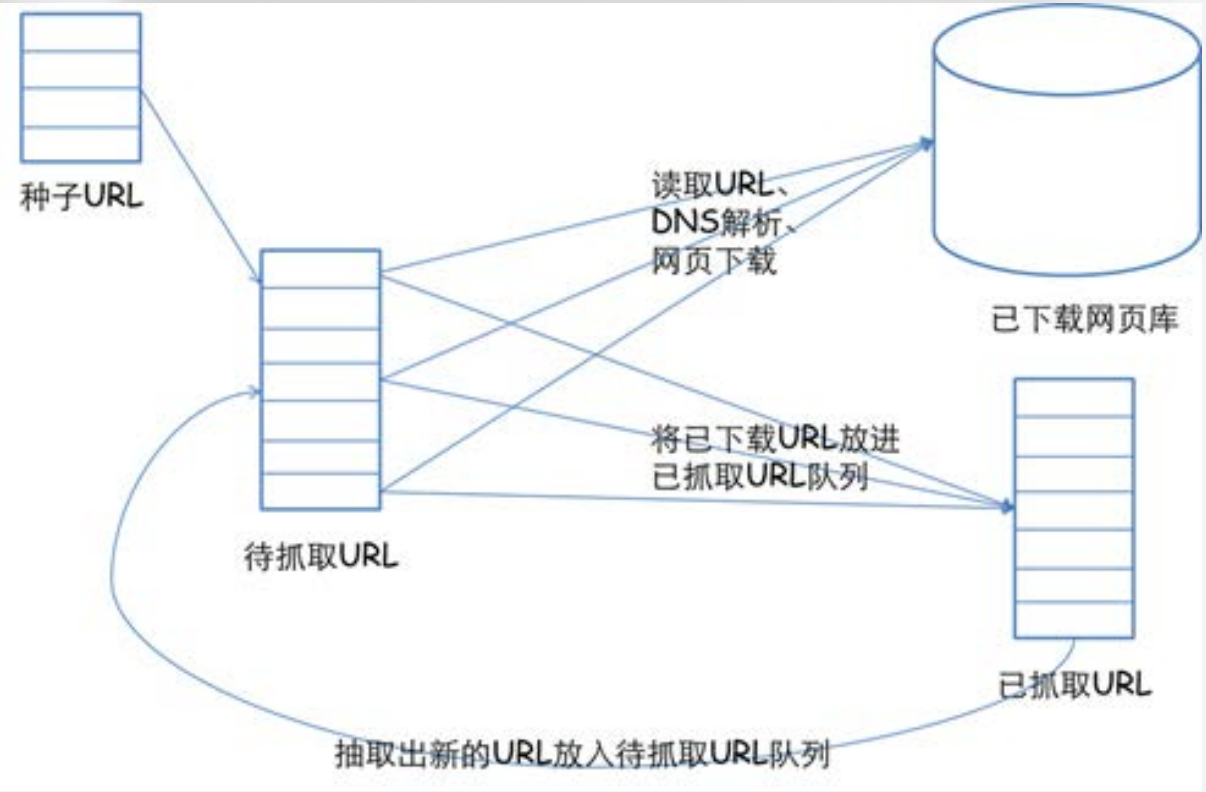
- 网络爬虫（又被称为网页蜘蛛，网络机器人，在FOAF社区中间，更经常的称为网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁，自动索引，模拟程序或者蠕虫。
- 网络爬虫是搜索引擎抓取系统的重要组成部分。爬虫的主要目的是将互联网上的网页下载到本地形成一个或联网内容的镜像备份。

2.网络爬虫的基本结构

- 在网络爬虫的系统框架中，主过程由控制器，解析器，资源库三部分组成。
- 1.控制器的主要工作是负责给多线程中的各个爬虫线程分配工作任务。
- 2.解析器的主要工作是下载网页，进行页面的处理，主要是将一些JS脚本标签、CSS代码内容、空格字符、HTML标签等内容处理掉，爬虫的基本工作是由解析器完成。
- 3.资源库是用来存放下载到的网页资源，一般都采用大型的数据库存储，如Oracle数据库，并对其建立索引。

2.网络爬虫的基本结构

- 一个通用的网络爬虫的框架



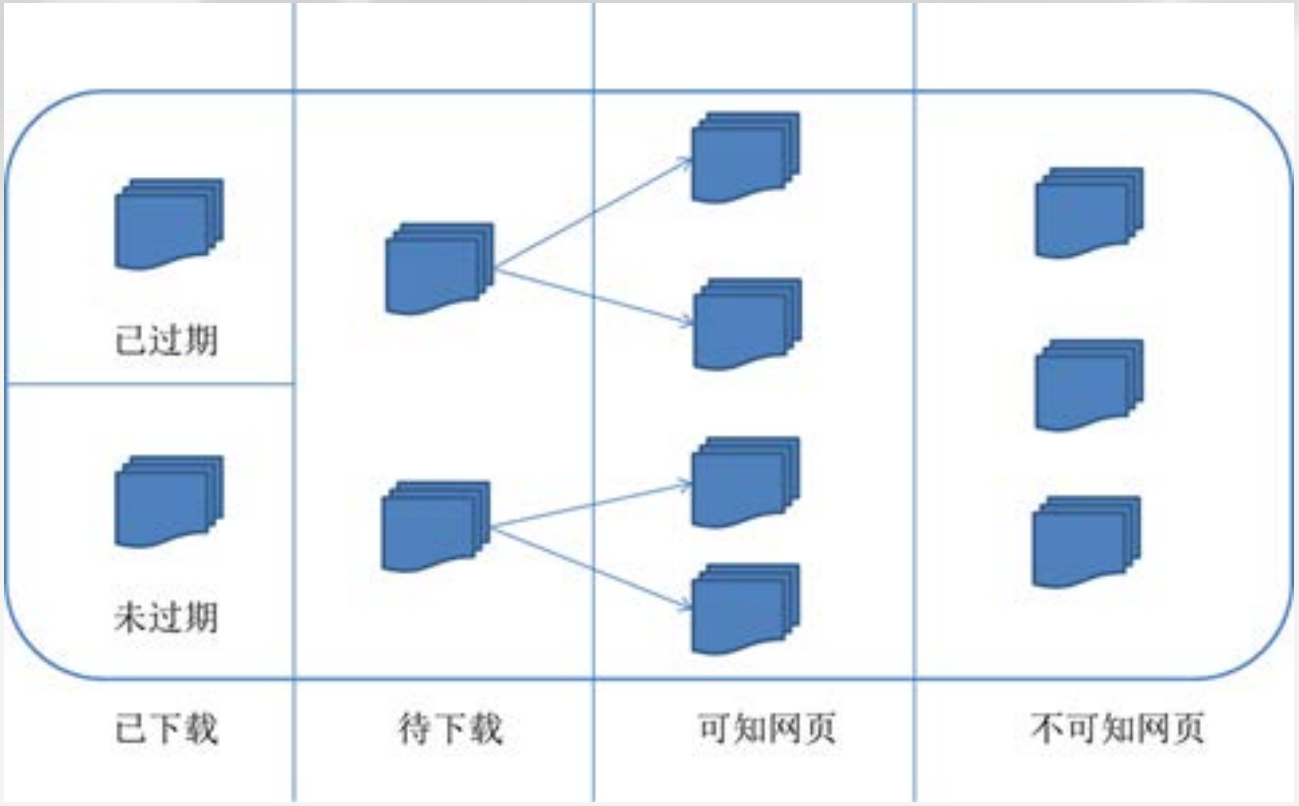
3.网络爬虫的工作流程

- 1.首先选取一部分精心挑选的种子URL；
- 2.将这些URL放入待抓取URL队列；
- 3.从待抓取URL队列中取出待抓取在URL，解析DNS，并且得到主机的ip，并将URL对应的网页下载下来，存储进已下载网页库中。此外，将这些URL放进已抓取URL队列。
- 4.分析已抓取URL队列中的URL，分析其中的其他URL，并且将URL放入待抓取URL队列，从而进入下一个循环。

4.从爬虫的角度对互联网进行划分

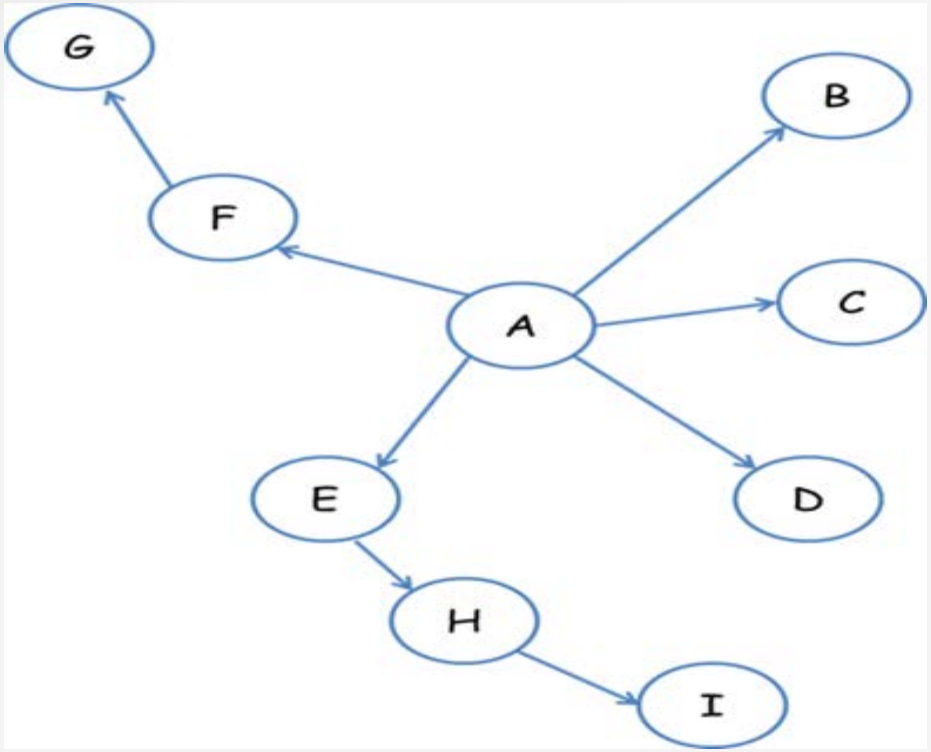
- 主要可以分为以下5部分:
- 1.已下载未过期网页
- 2.已下载已过期网页：抓取到的网页实际上是互联网内容的一个镜像与备份，互联网是动态变化的，一部分互联网上的内容已经发生了变化，这时，这部分抓取到的网页就已经过期了。
- 3.待下载网页：也就是待抓取URL队列中的那些页面
- 4.可知网页：还没有抓取下来，也没有在待抓取URL队列中，但是可以通过对已抓取页面或者待抓取URL对应页面进行分析获取到的URL，认为是可知网页。
- 5.还有一部分网页，爬虫是无法直接抓取下载的。称为不可知网页。

4.从爬虫的角度对互联网进行划分



5.抓取策略

- 在爬虫系统中，待抓取URL队列是很重要的一部分。待抓取URL队列中的URL以什么样的顺序排列也是一个很重要的问题，因为这涉及到先抓取那个页面，后抓取哪个页面。而决定这些URL排列顺序的方法，叫做抓取策略。以下图为例：



5.1.深度优先遍历策略

- 深度优先遍历策略是指网络爬虫会从起始页开始，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。

- 遍历的路径：A-F-G　E-H-I　B　C　D

5.2.宽度优先遍历策略

- 宽度优先遍历策略的基本思路是，将新下载网页中发现的链接直接插入待抓取URL队列的末尾。也就是指网络爬虫会先抓取起始网页中链接的所有网页，然后再选择其中的一个链接网页，继续抓取在此网页中链接的所有网页。还是以上面的图为例：

- 遍历路径：A-B-C-D-E-F G H I

5.3.反向链接数策略

- 反向链接数是指一个网页被其他网页链接指向的数量。反向链接数表示的是一个网页的内容受到其他人的推荐的程度。因此，很多时候搜索引擎的抓取系统会使用这个指标来评价网页的重要程度，从而决定不同网页的抓取先后顺序。

- 在真实的网络环境中，由于广告链接、作弊链接的存在，反向链接数不能完全等他我那个也的重要程度。因此，搜索引擎往往考虑一些可靠的反向链接数。

5.4.Partial PageRank策略

- Partial PageRank算法借鉴了PageRank算法的思想：对于已经下载的网页，连同待抓取URL队列中的URL，形成网页集合，计算每个页面的PageRank值，计算完之后，将待抓取URL队列中的URL按照PageRank值的大小排列，并按照该顺序抓取页面。

- 如果每次抓取一个页面，就重新计算PageRank值，一种折中方案是：每抓取K个页面后，重新计算一次PageRank值。但是这种情况还会有一个问题：对于已经下载下来的页面中分析出的链接，也就是我们之前提到的未知网页那一部分，暂时是没有PageRank值的。为了解决这个问题，会给这些页面一个临时的PageRank值：将这个网页所有入链传递进来的PageRank值进行汇总，这样就形成了该未知页面的PageRank值，从而参与排序。

5.5.OPIC策略策略

- 该算法实际上也是对页面进行一个重要性打分。在算法开始前，给所有页面一个相同的初始现金（cash）。当下载了某个页面P之后，将P的现金分摊给所有从P中分析出的链接，并且将P的现金清空。对于待抓取URL队列中的所有页面按照现金数进行排序。

5.6.大站优先策略

- 对于待抓取URL队列中的所有网页，根据所属的网站进行分类。对于待下载页面数多的网站，优先下载。这个策略也因此叫做大站优先策略。

6.网站与网络蜘蛛

- 网络蜘蛛需要抓取网页，不同于一般的访问，如果控制不好，则会引起网站服务器负担过重。去年4月，淘宝 就因为雅虎搜索引擎的网络蜘蛛抓取其数据引起淘宝网服务器的不稳定。
- 网站是否就无法和网络蜘蛛交流呢？其实不然，有多种方法可以让网站和网络蜘蛛进行交流。一方面让网站管理员了解网络蜘蛛都来自哪儿，做了些什么，另一方面也告诉网络蜘蛛哪些网页不应该抓取，哪些网页应该更新。

6.网站与网络蜘蛛

- 每个网络蜘蛛都有自己的名字，在抓取网页的时候，都会向网站标明自己的身份。网络蜘蛛在抓取网页的时候会发送一个请求，这个请求中就有一个字段为User－agent，用于标识此网络蜘蛛的身份。例如Google网络蜘蛛的标识为GoogleBot，Baidu网络蜘蛛的标识为BaiDuSpider，Yahoo网络蜘蛛的标识为Inktomi Slurp。如果在网站上有访问日志记录，网站管理员就能知道，哪些搜索引擎的网络蜘蛛过来过，什么时候过来的，以及读了多少数据等等。如果网站管理员发现某个蜘蛛有问题，就通过其标识来和其所有者联系。

6.网站与网络蜘蛛

- 网络蜘蛛进入一个网站，一般会访问一个特殊的文本文件Robots.txt，这个文件一般放在网站服务器的根目录下，网站管理员可以通过robots.txt来定义哪些目录网络蜘蛛不能访问，或者哪些目录对于某些特定的网络蜘蛛不能访问。例如有些网站的可执行文件目录和临时文件目录不希望被搜索引擎搜索到，那么网站管理员就可以把这些目录定义为拒绝访问目录。
- Robots.txt语法很简单，例如如果对目录没有任何限制，可以用以下两行来描述：
- User-agent: *
- Disallow:

6.网站与网络蜘蛛

- Robots.txt只是一个协议，如果网络蜘蛛的设计者不遵循这个协议，网站管理员也无法阻止网络蜘蛛对于某些页面的访问，但一般的网络蜘蛛都会遵循这些协议，而且网站管理员还可以通过其它方式来拒绝网络蜘蛛对某些网页的抓取。

6.网站与网络蜘蛛

- 网络蜘蛛在下载网页的时候，会去识别网页的HTML代码，在其代码的部分，会有META标识。通过这些标识，可以告诉网络蜘蛛本网页是否需要被抓取，还可以告诉网络蜘蛛本网页中的链接是否需要被继续跟踪。
- 例如：表示本网页不需要被抓取，但是网页内的链接需要被跟踪。

6.网站与网络蜘蛛

- 现在一般的网站都希望搜索引擎能更全面的抓取自己网站的网页，因为这样可以让更多的访问者能通过搜索引擎找到此网站。为了让本网站的网页更全面被抓取到，网站管理员可以建立一个网站地图，即SiteMap。许多网络蜘蛛会把sitemap.htm文件作为一个网站网页爬取的入口，网站管理员可以把网站内部所有网页的链接放在这个文件里面，那么网络蜘蛛可以很方便的把整个网站抓取下来，避免遗漏某些网页，也会减小对网站服务器的负担。

7.内容提取

- 搜索引擎建立网页索引，处理的对象是文本文件。对于网络蜘蛛来说，抓取下来网页包括各种格式，包括html、图片、doc、pdf、多媒体、动态网页及其它格式等。这些文件抓取下来后，需要把这些文件中的文本信息提取出来。准确提取这些文档的信息，一方面对搜索引擎的搜索准确性有重要作用，另一方面对于网络蜘蛛正确跟踪其它链接有一定影响。

- 对于doc、pdf等文档，这种由专业厂商提供的软件生成的文档，厂商都会提供相应的文本提取接口。网络蜘蛛只需要调用这些插件的接口，就可以轻松的提取文档中的文本信息和文件其它相关的信息。

7.内容提取

- HTML等文档不一样，HTML有一套自己的语法，通过不同的命令标识符来表示不同的字体、颜色、位置等版式，如：、、等，提取文本信息时需要把这些标识符都过滤掉。过滤标识符并非难事，因为这些标识符都有一定的规则，只要按照不同的标识符取得相应的信息即可。但在识别这些信息的时候，需要同步记录许多版式信息，例如文字的字体大小、是否是标题、是否是加粗显示、是否是页面的关键词等，这些信息有助于计算单词在网页中的重要程度。同时，对于 HTML网页来说，除了标题和正文以外，会有许多广告链接以及公共的频道链接，这些链接和文本正文一点关系也没有，在提取网页内容的时候，也需要过滤这些 无用的链接。例如某个网站有“产品介绍”频道，因为导航条在网站内每个网页都有，若不过滤导航条链接，在搜索“产品介绍”的时候，则网站内每个网页都会搜索到，无疑会带来大量垃圾信息。过滤这些无效链接需要统计大量的网页结构规律，抽取一些共性，统一过滤；对于一些重要而结果特殊的网站，还需要个别处理。这就需要网络蜘蛛的设计有一定的扩展性。

7.内容提取

- 对于多媒体、图片等文件，一般是通过链接的锚文本（即，链接文本）和相关的文件注释来判断这些文件的内容。例如有一个链接文字为“张曼玉照片 ”，其链接指向一张bmp格式的图片，那么网络蜘蛛就知道这张图片的内容是“张曼玉的照片”。这样，在搜索“张曼玉”和“照片”的时候都能让搜索引擎找到这张图片。另外，许多多媒体文件中有文件属性，考虑这些属性也可以更好的了解文件的内容。

8.更新周期

- 由于网站的内容经常在变化，因此网络蜘蛛也需不断的更新其抓取网页的内容，这就需要网络蜘蛛按照一定的周期去扫描网站，查看哪些页面是需要更新的页面，哪些页面是新增页面，哪些页面是已经过期的死链接。
- 搜索引擎的更新周期对搜索引擎搜索的查全率有很大影响。如果更新周期太长，则总会有一部分新生成的网页搜索不到；周期过短，技术实现会有一定难度，而且会对带宽、服务器的资源都有浪费。搜索引擎的网络蜘蛛并不是所有的网站都采用同一个周期进行更新，对于一些重要的更新量大的网站，更新的周期短，如有些新闻网站，几个小时就更新一次；相反对于一些不重要的网站，更新的周期就长，可能一两个月才更新一次。

8.更新周期

- 一般来说，网络蜘蛛在更新网站内容的时候，不用把网站网页重新抓取一遍，对于大部分的网页，只需要判断网页的属性（主要是日期），把得到的属性和上次抓取的属性相比较，如果一样则不用更新。

小结

压缩，超时设置

多代理自动选择

使用代理服务器

验证码的处理

伪装成浏览器访问

多线程并发抓取

反”反盗链”

代码简单，使用方便，性能也不俗，可谓居家旅行，杀人放火（黑网站），咳咳，之必备神器。