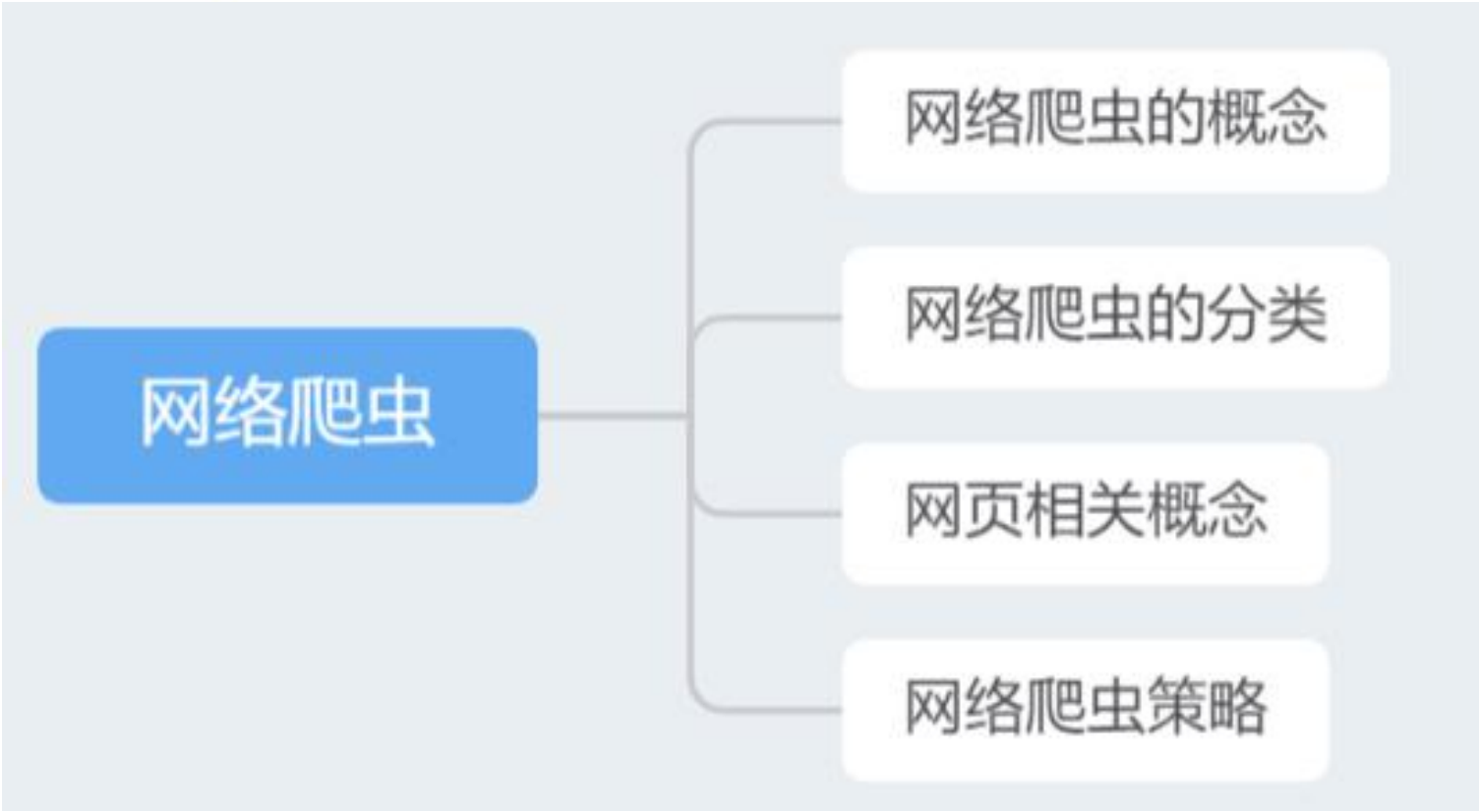
爬虫应用开发

网络爬虫

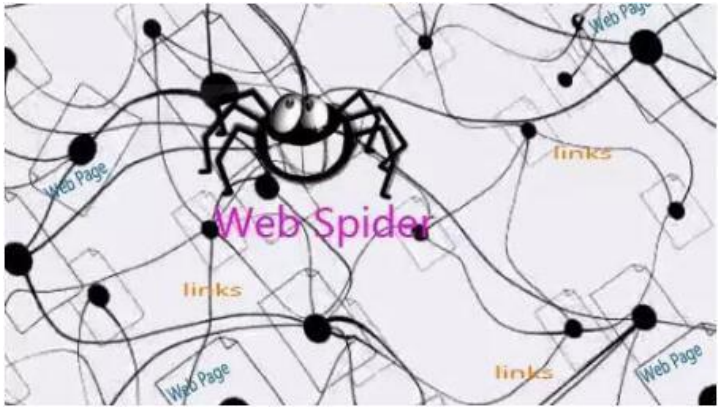


网络爬虫的概念

■ 1、网络爬虫的定义

✓网络爬虫（ Crawler，又被称为网页蜘蛛--Spider、网络机器人），是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。

✓网络爬虫经典的应用案例，如Google、百度、Bing（必应）。



网络爬虫的概念

■2、网络爬虫的作用

疑问？ **网络爬虫有什么用？**

【案例1】电影评价网站记录着观影者对电影的喜好程度和评价信息，通过对相关网站用户评价信息的收集，可以为电影相关的数据的分析和挖掘做支撑，常见的后期应用包括：

- 对电影针对的用户群体做分析；
- 获得大众娱乐/舆情热点；
- 电影推荐（广告推送）；

网络爬虫的概念

■2、网络爬虫的作用

疑问？ **网络爬虫有什么用？**

【案例2】有一个销售电子设备的店铺，想要及时了解竞争对手的价格。他们可以通过每天访问电商平台或相关产品网站，与店铺中出售的电子设备进行价格对比。但是，由于店铺中的电子产品种类繁多，而且希望能够更加频繁的查看价格变化的动态，采用传统手工查询的方式需要花费大量的时间。

网络爬虫的概念

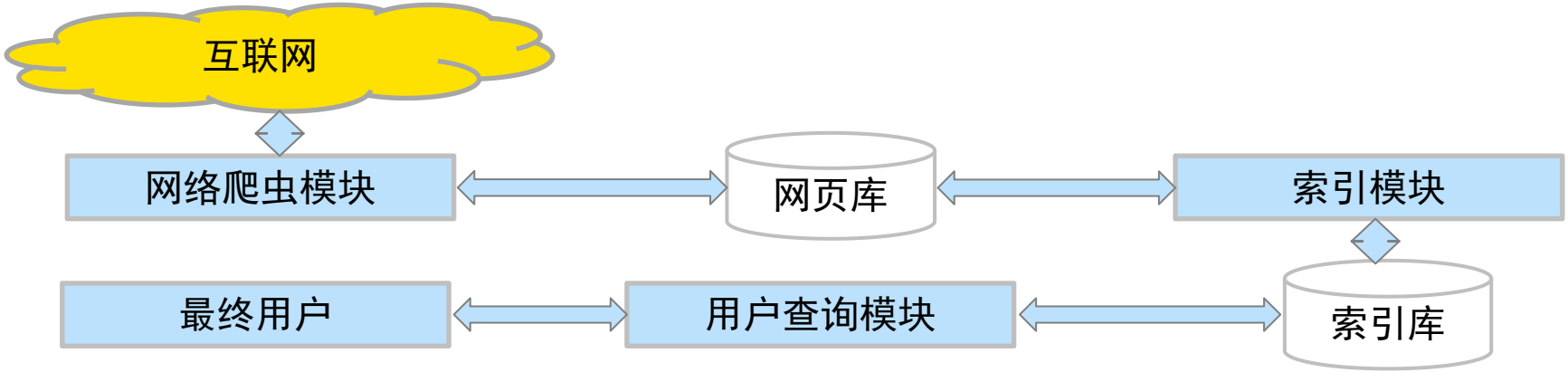
■2、网络爬虫的作用

疑问？ **网络爬虫有什么用？**

【案例3】福州的一家上市公司开会做出股权转让决定，刚在官网公告频道发布消息，很多内部员工还没及时了解到变更信息，国税局就立刻找到该上市公司。国税稽查人员透露说上市公司股权转让问题越来越成为税收征管的热点和难点，国税“种植”的网络“爬虫”第一时间监测到该上市公司转让股权的消息，于是引起高度重视。

网络爬虫的概念

3、网络爬虫的流程



注意

网络爬虫是**搜索引擎**的重要组成部分，它作为一个功能强大的自动提取网页程序，为搜索引擎从万维网上下载网页。

网络爬虫的分类

- 网络爬虫按照系统结构和实现技术，可以分为以下几种类型。
  - 通用网络爬虫（ General Purpose Web Crawler ）
  - 聚焦网络爬虫（ Focused Web Crawler ）
  - 增量式网络爬虫（ Incremental Web Crawler ）
  - 深层网络爬虫（ Deep Web Crawler ）
- 实际的网络爬虫系统通常是几种爬虫技术相结合实现的。

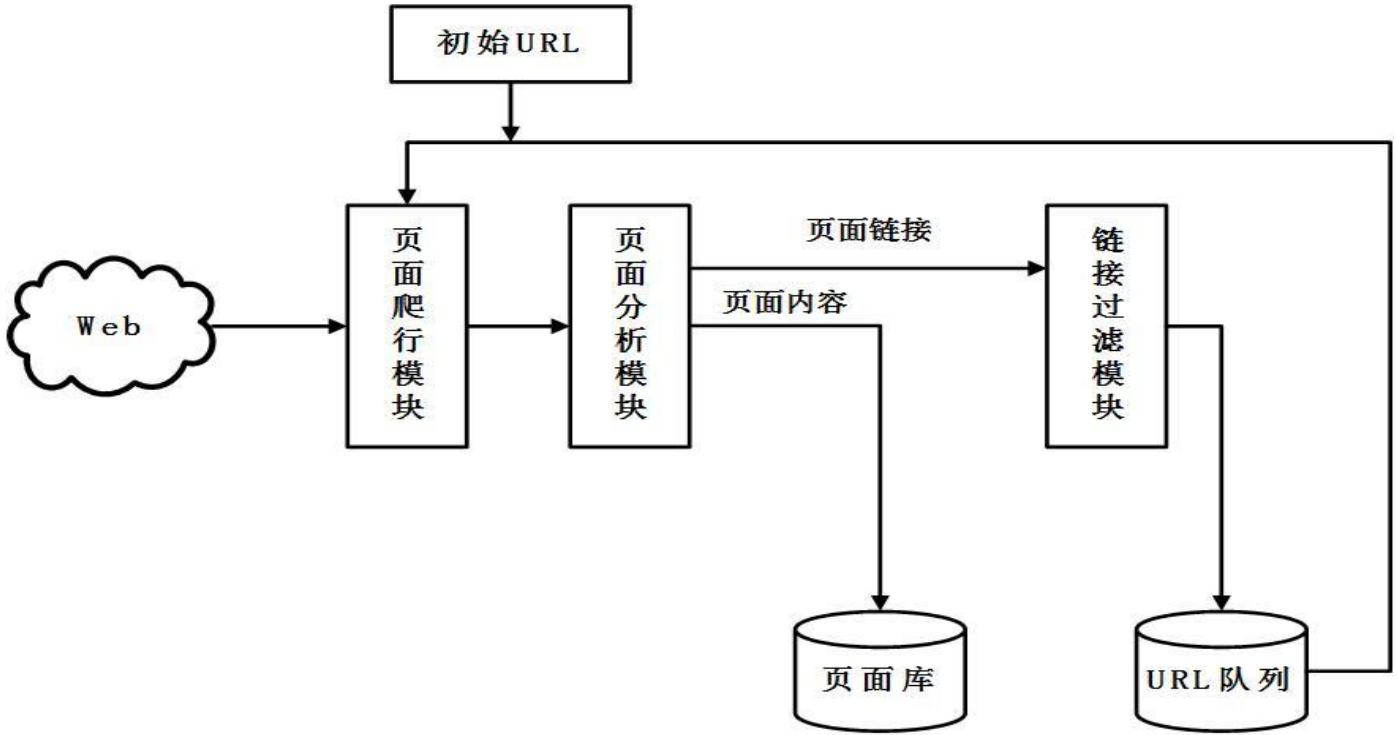
网络爬虫的分类

- 1、通用网络爬虫（General Purpose Web Crawler）
  - 原理：通用网络爬虫又称全网爬虫（Scalable Web Crawler），爬行对象从一些种子 URL 扩充到整个 Web，主要为门户站点搜索引擎和大型 Web 服务提供商采集数据。
  - 结构：分为页面爬行模块 、页面分析模块、链接过滤模块、页面数据库、URL 队列、初始 URL 集合几个部分。
  - 爬行策略：深度优先策略、广度优先策略。

网络爬虫的分类

- 1、通用网络爬虫（General Purpose Web Crawler）

通用网络爬虫体系图



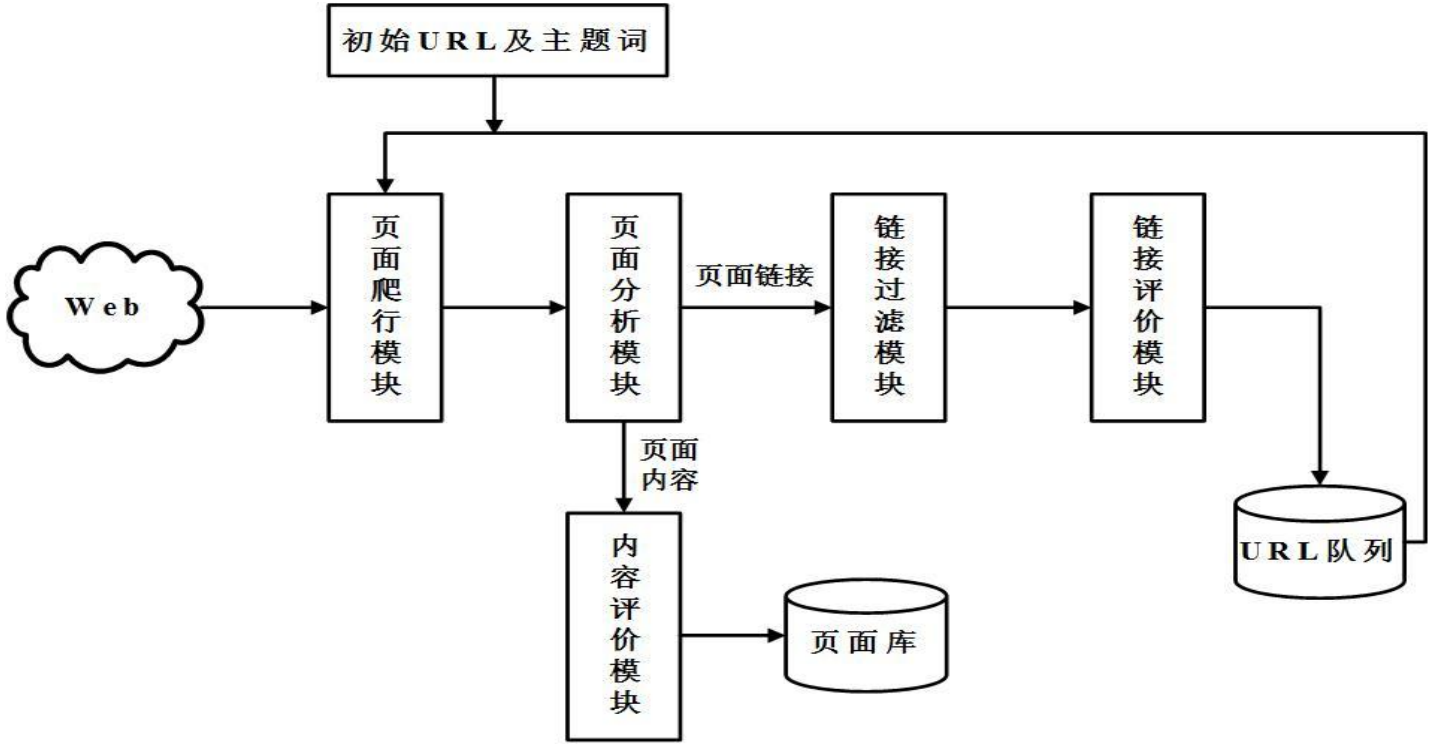
网络爬虫的分类

- 2、聚焦网络爬虫（ Focused Web Crawler ）
  - 原理：聚焦网络爬虫（ Focused Crawler ），又称主题网络爬虫（ Topical Crawler ），是指选择性地爬行那些与预先定义好的主题相关页面的网络爬虫。
  - 结构：分为页面爬行模块 、页面分析模块、链接过滤模块、页面数据库、URL 队列、初始 URL 集合、链接评价模块以及内容评价模块几个部分。
  - 爬行策略：基于内容评价的爬行策略、
    基于链接结构评价的爬行策略 、
    基于增强学习的爬行策略、
    基于语境图的爬行策略。

网络爬虫的分类

- 2、聚焦网络爬虫（ Focused Web Crawler ）

聚焦网络爬虫体系图



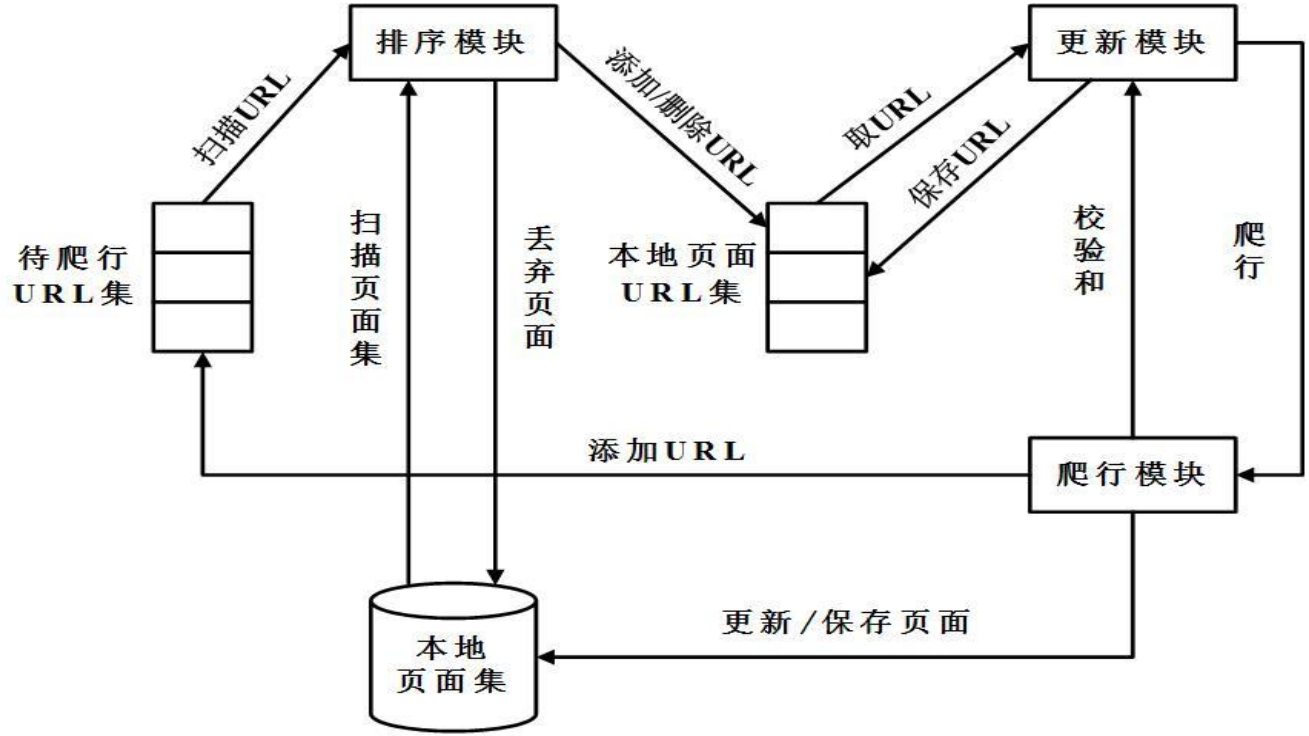
网络爬虫的分类

- 3、增量式网络爬虫（ Incremental Web Crawler ）
  - 原理：增量式网络爬虫（ Incremental Web Crawler ）是指对已下载网页采取增量式更新和只爬行新产生的或者已经发生变化网页的爬虫，它能够在一定程度上保证所爬行的页面是尽可能新的页面。
  - 结构：包含爬行模块、排序模块、更新模块、本地页面集、待爬行 URL 集以及本地页面URL 集。
  - 爬行策略：统一更新法、个体更新法、基于分类的更新法（保持本地页面集中存储的页面为最新页面）；广度优先策略、PageRank 优先策略等（提高本地页面集中页面的质量）。

网络爬虫的分类

- 3、增量式网络爬虫（ Incremental Web Crawler ）

增量式爬虫体系图



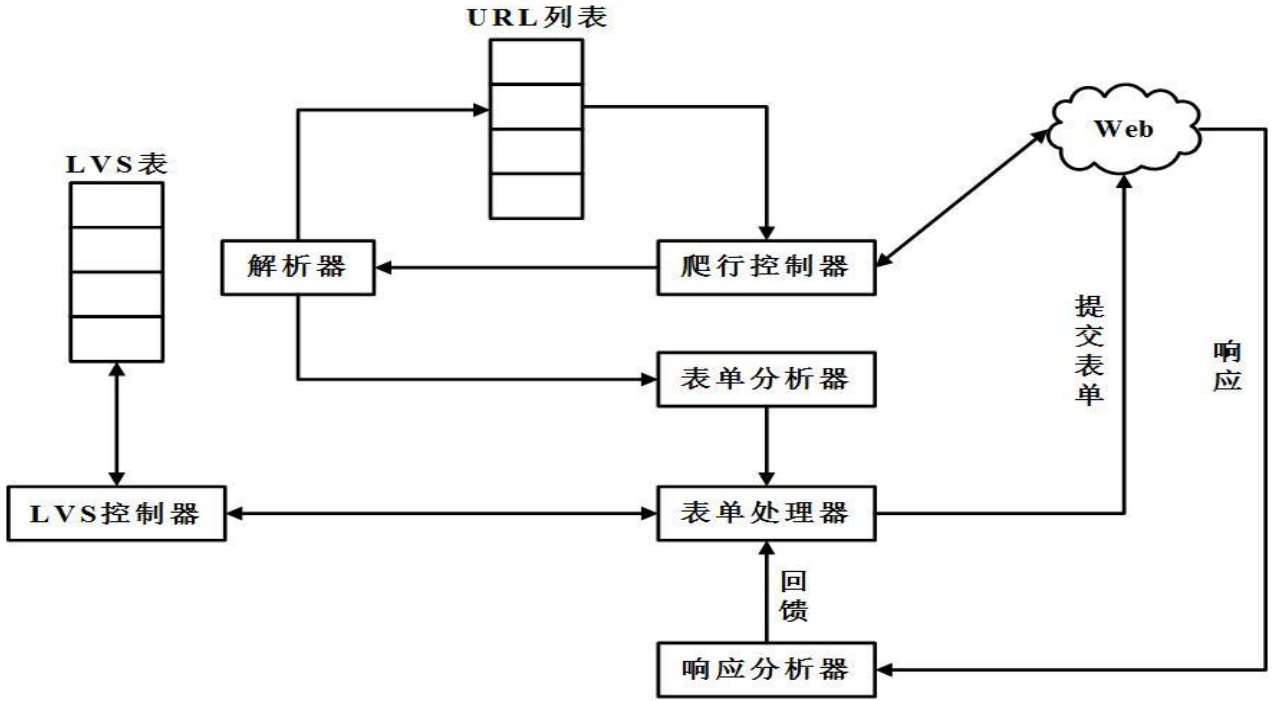
网络爬虫的分类

- 4、深层网络爬虫（ Deep Web Crawler ）
  - 原理：Deep Web 是那些大部分内容不能通过静态链接获取的、隐藏在搜索表单后的，只有用户提交一些关键词才能获得的 Web 页面。
  - 结构：包含六个基本功能模块 （爬行控制器、解析器、表单分析器、表单处理器、响应分析器、LVS 控制器）和两个爬虫内部数据结构（ URL 列表、LVS 表）。
  - 爬行策略：Deep Web 爬虫爬行过程中最重要部分就是表单填写，包含两种类型：基于领域知识的表单填写、基于网页结构分析的表单填写。

网络爬虫的分类

- 4、深层网络爬虫（ Deep Web Crawler ）

DEEP WEP网络爬虫体系图



网页相关概念

1、 HTTP基本原理

HTTP的基本原理，主要包括以下内容：

- URI和URL
- 超文本
- HTTP和HTTPS
- HTTP请求过程
- 请求
- 响应

网页相关概念

1、HTTP基本原理

- URI概述

统一资源标识符（Uniform Resource Identifier，URI）是一个用于标识某一互联网资源名称的字符串。 该种标识允许用户对任何（包括本地和互联网）的资源通过特定的协议进行交互操作。

统一资源定位符（Uniform Resource Locator，URL）、统一资源名称（Uniform Resource Name，URN）是URI的子集。

网页相关概念

1、HTTP基本原理

URL概述

URL（Uniform Resource Locator，统一资源定位符）是对可以从互联网上得到的资源的位置和访问方法的一种简洁的表示，是互联网上标准资源的地址。

互联网上的每个文件都有一个唯一的URL，它包含的信息指出文件的位置以及浏览器应该怎么处理它。

注意

现在互联网中，URN用的非常少，几乎所有的URI都是URL，一般的网页链接既可以称为URL，也可以称为URI。

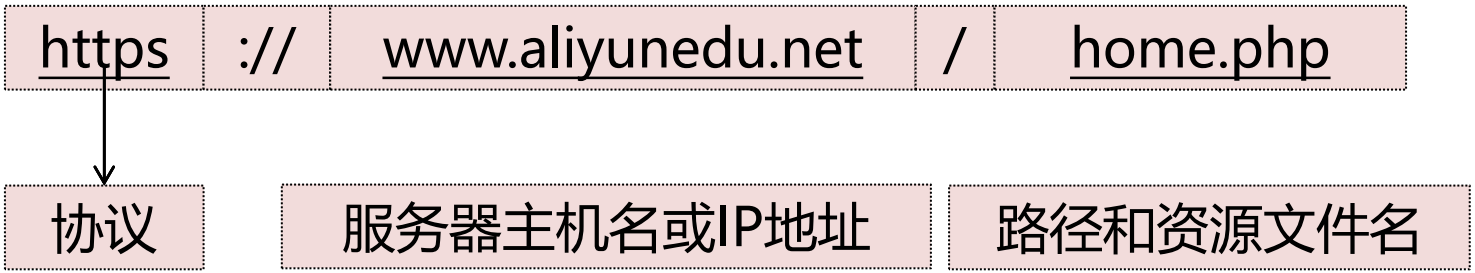
网页相关概念

1、HTTP基本原理

URL概述

URL采用一种统一的格式来描述各种信息资源，包括文件、服务器的地址和目录等。URL的格式由三部分组成：

- 第一部分是协议；
- 第二部分是存有该资源的主机IP地址（有时也包括端口号）；
- 第三部分是主机资源的具体地址，如目录和文件名等。



网页相关概念

1、HTTP基本原理

- URL概述

依据URL定义，给出了常用的两种URL协议的示例：

示例：HTTPS协议的URL

```
http://yundaxue.org/home/index.mooc
```

解释：其计算机域名为 yundaxue.org。超文本文件是在目录/home下的 index.mooc。

网页相关概念

1、HTTP基本原理

- URL概述

依据URL定义，给出了常用的两种URL协议的示例：

示例：文件的URL

```
file://ftp.yoyodyne.com/pub/files/foobar.txt
```

解释：上面这个URL代表存放在主机 ftp.yoyodyne.com上的 pub/files/目录下的一个文件， 文件名是 foobar.txt。

网页相关概念

1、HTTP基本原理

URL概述

完整的带有授权的普通URL语法描述如下：

协议://用户名:密码@子域名.域名.顶级域名:端口号/目录/文件名.文件后缀?参数=值#标志

URL中的端口号用来区分在同一个主机上的不同服务，其编号范围从0到65535。HTTP默认端口号为80，指定URL时可以省略。

如前述【示例】HTTP协议的URL带端口号的表达形式：

```
http://yundaxue.org:80/home/index.mooc
```

协议或服务	默认端口号
HTTP	80
FTP	21
telnet	23
SMTP	25
MySQL	3306

网页相关概念

1、HTTP基本原理

超文本

超文本（Hyper Text，HT)是由信息结点和表示信息结点间相关性的链构成的一个具有一定逻辑结构和语义的网络。

超文本是由节点(Node)和链(Link)构成的信息网络。

- 节点是表达信息的单位，通常表示一个单一的概念或围绕一个特殊主题组织起来的数据集合。节点的内容可是文本、图形、图像、动画、音频、视频等，也可以是一般计算机程序。
- 链是固定节点间的信息联系，它以某种形式将一个节点与其他节点连接起来。由于超文本没有规定链的规范与形式，因此，超文本与超媒体系统的链也是各异的，信息间的联系丰富多彩引起链的种类复杂多样。但最终达到效果却是一致的，即建立起节点之间的联系。

网页相关概念

1、HTTP基本原理

- HTTP和HTTPS

HTTP协议（HyperText Transfer Protocol，超文本传输协议）是因特网上应用最为广泛的一种网络传输协议，所有的WWW文件都必须遵守这个标准。HTTP是一个基于TCP/IP通信协议来传递数据（HTML文件、图片文件、查询结果等）。

网页相关概念

1、HTTP基本原理

- HTTP和HTTPS

HTTPS（Hyper Text Transfer Protocol over SecureSocket Layer），是以安全为目标的HTTP通道，在HTTP的基础上通过传输加密和身份认证保证了传输过程的安全性。

HTTPS在HTTP的基础下加入SSL层，HTTPS的安全基础是SSL，因此加密的详细内容就需要SSL。HTTPS 存在不同于HTTP的默认端口（端口为443）及一个加密/身份验证层（在HTTP与TCP之间）。

网页相关概念

1、HTTP基本原理

- HTTP和HTTPS

HTTPS的安全基础是SSL，通过它传输的内容都是经过SSL加密，它的主要作用可以分为以下两种：

- 建立一个信息安全通道来保证数据传输的安全；
- 确认网站的真实性，凡是使用了HTTPS的网站，都可以通过点击浏览器地址栏的锁头标志来查看网站认证之后的真实信息，也可以通过CA机构颁发的安全签章来查询。

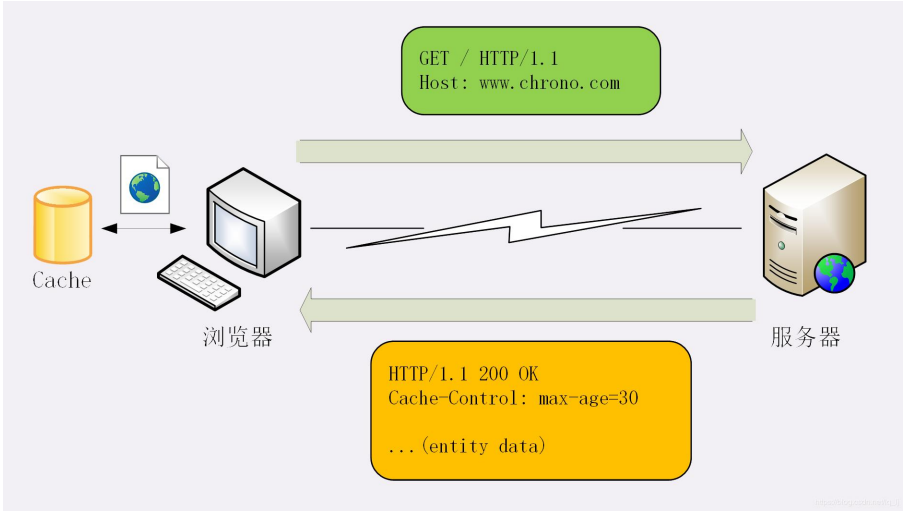
网页相关概念

1、HTTP基本原理

- HTTP和HTTPS

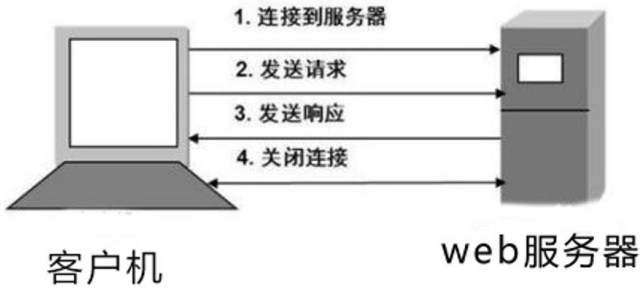
HTTP请求过程可分为以下几个部分：

1. 建立TCP连接;
2. 浏览器向服务器发送请求命令；
3. 服务器应答；
4. 服务器关闭TCP连接；
5. 浏览器接受到服务器响应的数据。



网页相关概念

1、HTTP基本原理



- HTTP和HTTPS

1. 建立TCP连接

在 HTTP工作开始之前，Web 浏览器首先要通过网络与Web服务器建立连接，该连接是通过TCP来完成的，该协议与 IP 协议共同构建 Internet，即著名的 TCP/IP 协议，因此 Internet 又被称作是TCP/IP网络。

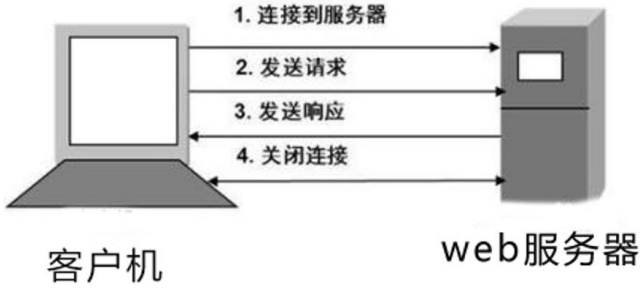
2.Web浏览器向Web服务器发送请求命令

建立TCP连接后，Web浏览器会向Web服务器发送请求命令。

网页相关概念

1、HTTP基本原理

- HTTP和HTTPS



3.Web服务器应答

客户机向服务器发出请求后，服务器会向客户机进行应答，应答内容包括：协议的版本号和应答状态码（HTTP/1.1 200 OK），响应头信息来记录服务器自己的数据，被请求的文档内容。

网页相关概念

1、HTTP基本原理

4.Web服务器关闭TCP链接

一旦Web服务器向浏览器发送了请求的数据，它就要关闭 TCP连接；但如果浏览器或者服务器在其头信息加入了这行代码：Connection:keep-alive，TCP连接在发送后将仍然保持打开状态，浏览器可以继续通过相同的连接发送请求。

5.浏览器接受到服务器响应的数据

浏览器接受服务器应答回来的HTML代码、CSS、JS代码,再进行页面的渲染或者接受到应答的文件进行保存等操作。

网页相关概念

1、HTTP基本原理

- 请求方式

在客户机和服务器之间进行请求响应时，两种最常被用到的方式是GET和POST：

- GET - 从指定的资源请求数据。
- POST - 向指定的资源提交要被处理的数据。

网页相关概念

1、HTTP基本原理

GET 方式

查询字符串（名称/值对）是在 GET 请求的URL中发送的：

```
/test/demo_form.asp?name1=value1&name2=value2
```

POST 方式

查询字符串（名称/值对）是在 POST 请求的 HTTP 消息主体中发送的：

```
POST /test/demo_form.asp HTTP/1.1
Host: w3schools.com
name1=value1&name2=value2
```

网页相关概念

1、HTTP基本原理

- 请求方式

  除了GET和POST方式，还有其他的请求方式：

  - HEAD - 类似于 GET 请求，只不过返回的响应中没有具体的内容，用于获取报头；
  - PUT - 从客户端向服务器传送的数据取代指定的文档的内容；
  - DELETE - 请求服务器删除指定的页面；
  - CONNECT - HTTP/1.1 协议中预留给能够将连接改为管道方式的代理服务器；
  - OPTIONS - 允许客户端查看服务器的性能；
  - TRACE - 回显服务器收到的请求，主要用于测试或诊断；
  - PATCH - 是对PUT方法的补充，用来对已知资源进行局部更新。

网页相关概念

1、HTTP基本原理

- 响应

响应，由服务端返回给客户端，可分为四部分：

- 响应状态码
- 响应头
- 空行
- 响应体

网页相关概念

1、HTTP基本原理

- 响应

响应状态码

HTTP状态码表示 HTTP协议所返回的响应的状态。

HTTP状态码通常分为5种类型， 分别以1～5五个数字开头， 由3位整数组成：

- 1XX表示消息；
- 2XX表示成功；
- 3XX表示重定向；
- 4XX表示请求错误；
- 5XX表示服务器错误；

网页相关概念

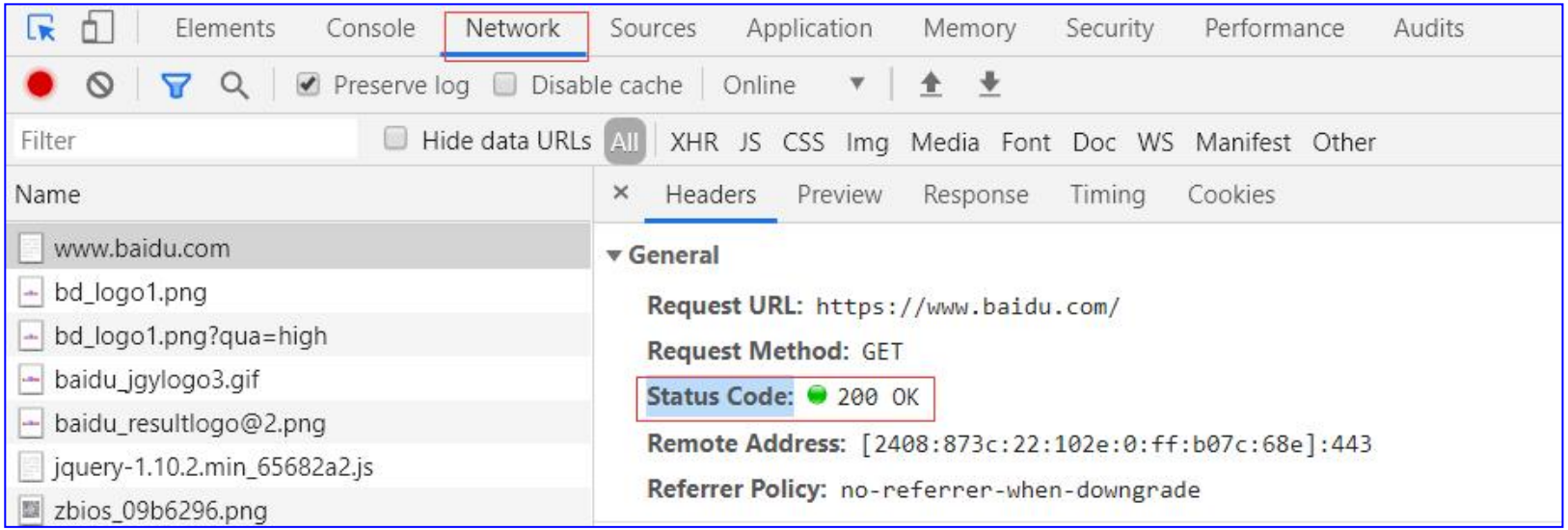
命令	功能	爬虫处理方式
200	请求成功	获得响应内容，进行处理
201	请求完成，结果是创建了新资源	爬虫中不会遇到
202	请求被接受，但处理尚未完成	阻塞等待
204	服务器端已经实现了请求，但是没有返回新的信息	丢弃
300	存在多个可用的被请求资源	若程序中能够处理，则进行进一步处理，如果程序中不能处理，则丢弃
301	请求到的资源都会分配一个永久的URL，这样就可以在将来通过该 URL来访问此资源	重定向到分配的URL
302	请求到资源在一个不同的URL处临时保存	重定向到临时的URL

网页相关概念

1、HTTP基本原理

- 响应

响应状态码



网页相关概念

1、HTTP基本原理

- 响应

响应头

响应头包含了服务器对请求的应答信息，常见的响应头如下：

- Allow：服务器支持哪些请求方法（如GET、POST等）；
- Content-Type：表示后面的文档属于什么类型。Servlet默认为text/plain，但通常需要显式地指定为text/html；
- Date：当前的GMT时间；

网页相关概念

1、HTTP基本原理

- 响应

响应头

响应头包含了服务器对请求的应答信息，常见的响应头如下：

- Location：表示客户应当到哪里去提取文档；
- Refresh：表示浏览器应该在多少时间之后刷新文档，以秒计；
- Server：服务器名字，Servlet一般不设置这个值，而是由Web服务器自己设置；
- Set-Cookie：设置和页面关联的Cookie；

网页相关概念

1、HTTP基本原理

- 响应

空行

作为内容分割，表示以下不再是响应头的内容。

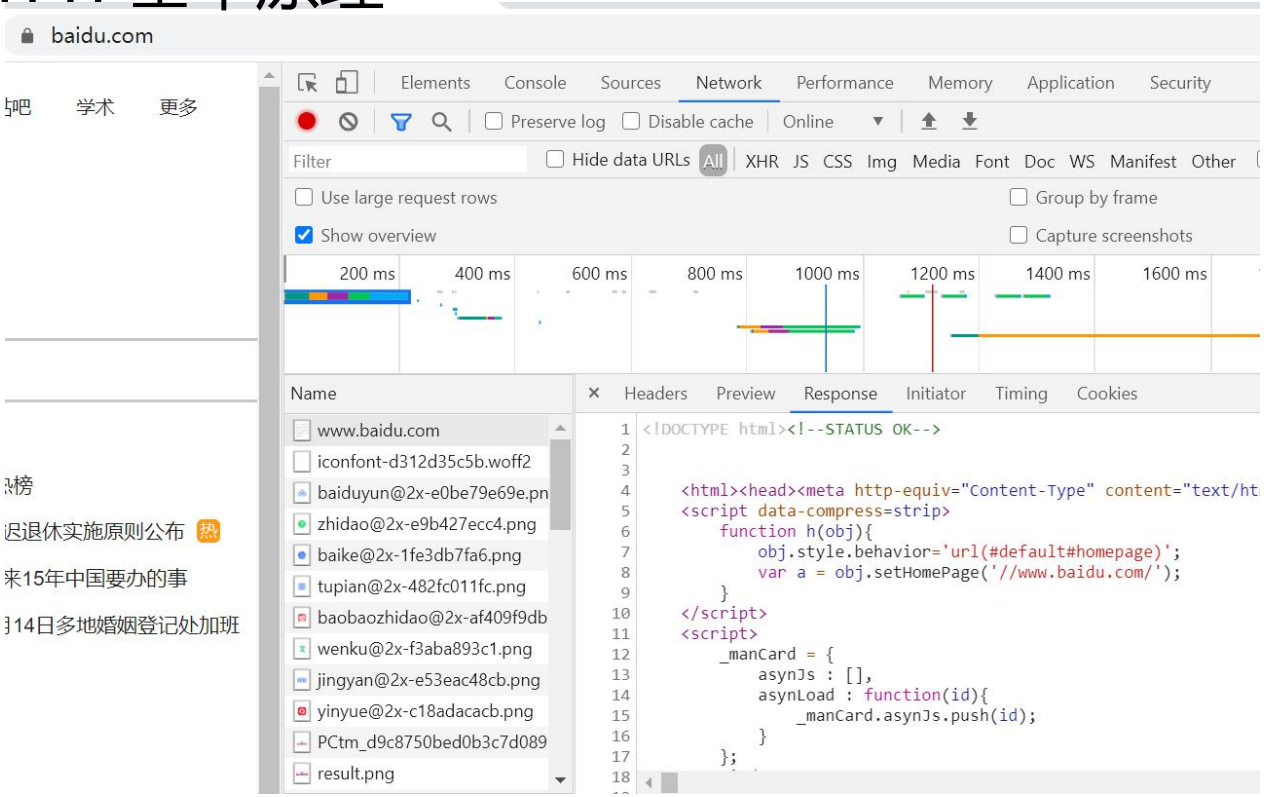
响应体

响应头是服务器返回给客户端的文本信息，响应的正文数据在响应体中；比如请求网页时，它的响应体就是网页的HTML代码；爬虫请求网页后，要解析的内容就是响应体。

网页相关概念

1、HTTP基本原理

- 响应



网页相关概念

2、网页基础

浏览器访问网站时，页面各不相同，现在介绍一下网页的基本组成、结构和节点等内容，具体内容如下：

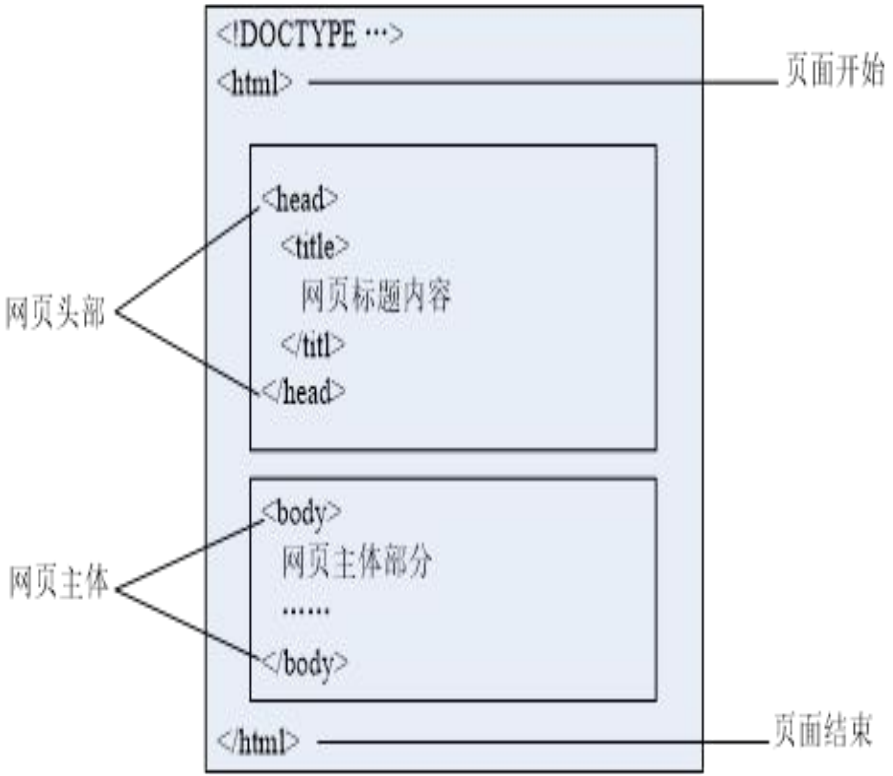
- 网页的组成
- 节点数及节点的关系
- 选择器

网页相关概念

2、网页基础

网页组成

网页的组成是以HTML文档、JavaScript脚本和CSS样式单的方式组织。在网页上查看源码可以了解其内部结构。

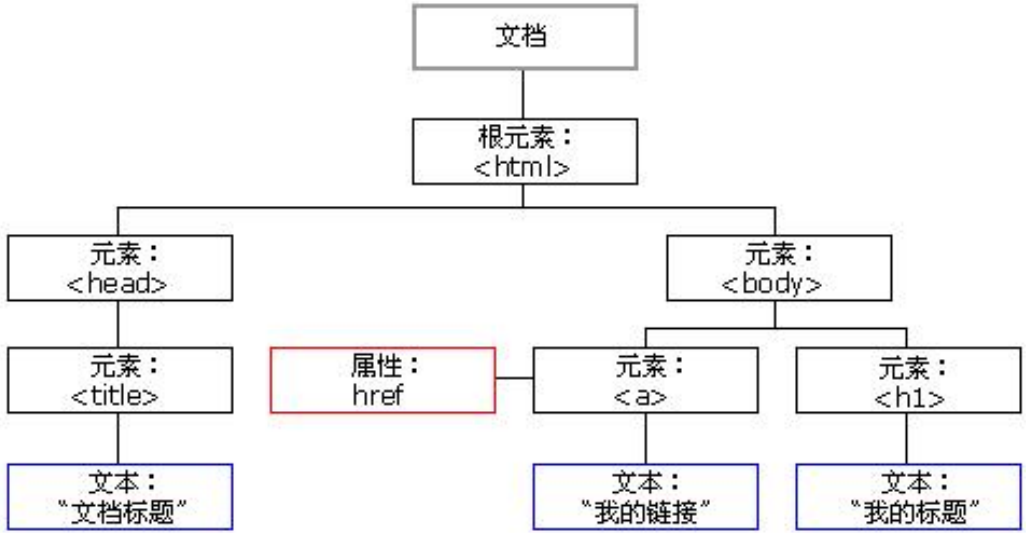


网页相关概念

2、网页基础

- 节点树及节点间的关系

HTML DOM将HTML文档视作树结构，这种结构被称为节点树，如下图所示：



通过HTML DOM，树中的所有节点均可通过JavaScript进行访问。所有HTML元素（节点）均可被修改，也可以创建或删除节点。

网页相关概念

2、网页基础

●节点树及节点间的关系

节点树中的节点彼此拥有层级关系。父（parent）、子（child）和同胞（sibling）等术语用于描述这些关系。父节点拥有子节点，同级的子节点被称为同胞（兄弟或姐妹）：

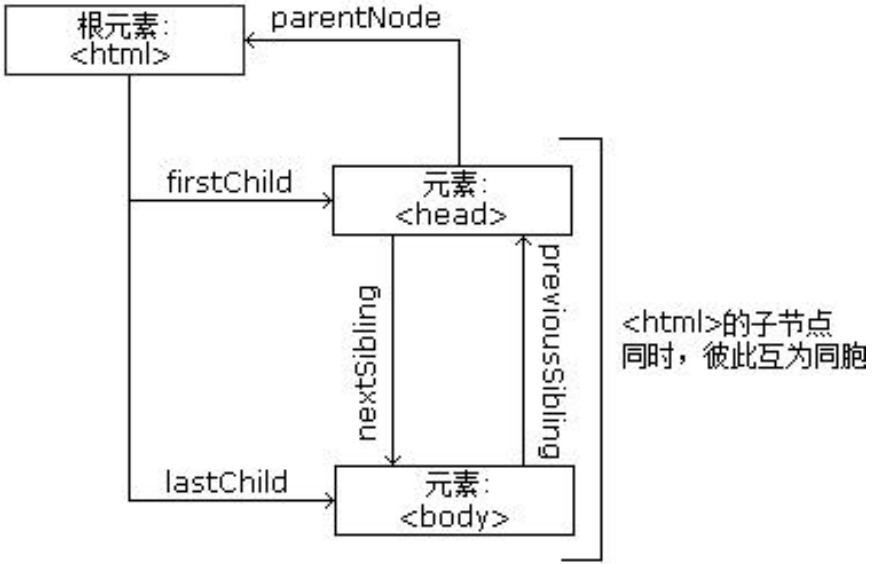
- 在节点树中，顶端节点被称为根（root）；
- 每个节点都有父节点、除了根（它没有父节点）；
- 一个节点可拥有任意数量的子；
- 同胞是拥有相同父节点的节点。

网页相关概念

2、网页基础

- 节点树及节点间的关系

下面的图片展示了节点树的一部分，以及节点之间的关系：



网页相关概念

2、网页基础

选择器

网页由一个个节点组成，CSS选择器依据不同的节点设置不同的样式规则。

在CSS中，使用选择器来定位节点。

网页相关概念

选择器	示例	示例说明
.class	.intro	选择所有class="intro"的元素
#id	#firstname	选择所有id="firstname"的元素
*	*	选择所有元素
element	p	选择所有<p>元素
element,element	div,p	选择所有 元素和<p>元素
element element	div p	选择 元素内的所有<p>元素
element>element	div>p	选择所有父级是 元素的 <p> 元素
element+element	div+p	选择所有紧接着 元素之后的<p>元素
[attribute]	[target]	选择所有带有target属性元素
:link	a:link	选择所有未访问链接
:visited	a:visited	选择所有访问过的链接
:first-letter	p:first-letter	选择每一个<p>元素的第一个字母

网络爬虫策略

互联网可以看成一个超级大的“图”，而每个页面可以看做是一个“节点”，页面中的链接可以看成是图的“有向边”，能够通过图的遍历方式对互联网这个超级大“图”进行访问。

依据互联网的特性，互联网遍历有以下几种：

- 宽（广）度优先遍历
- 深度优先遍历
- 带偏好的爬虫

网络爬虫策略

- 1、 图的宽度优先遍历

图的宽度优先遍历（BFS）算法是一个分层搜索的过程，和树的层序遍历算法相同。图中一个节点，作为起始节点，然后按照层次遍历的方式，一层一层地进行访问。

依据互联网的特性，互联网遍历过程如下：

1.顶点 V入队列；

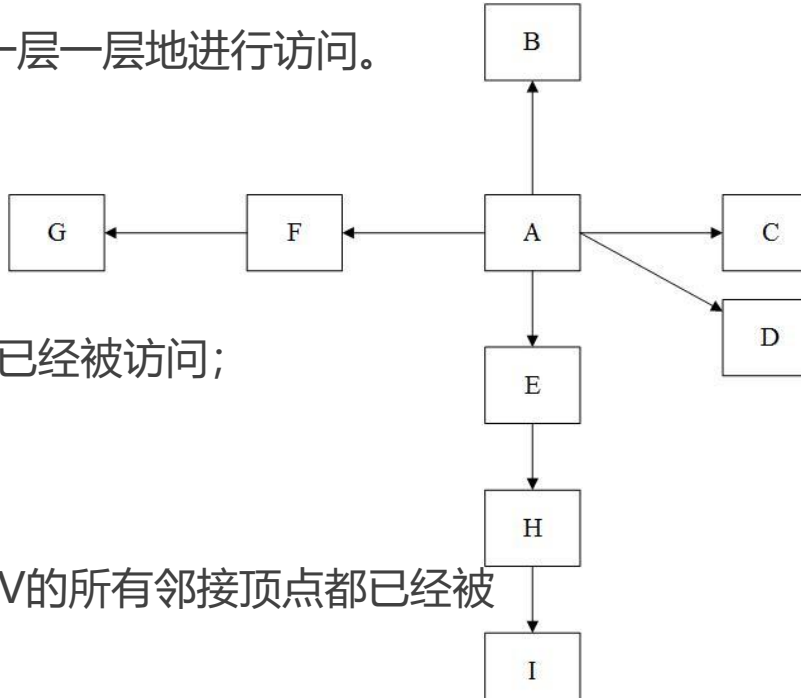
2.当队列非空时继续执行，否则算法为空；

3.出队列， 获得队头节点 V， 访问顶点 V并标记 V已经被访问；

4.查找顶点 V的第一个邻接顶点 col；

5.若 V的邻接顶点col未被访问过，则col进队列；

6.继续查找 V的其他邻接顶点 col，转到步骤5，若 V的所有邻接顶点都已经被访问过，则转到步骤2。



网络爬虫策略

- 1、 图的宽度优先遍历

爬虫项目是从一系列的种子链接开始的。 所谓种子链接， 就好比宽度优先遍历中的种子节点。实际的网络爬虫项目中种子链接可以有多个，而宽度优先遍历的种子节点只有一个。

宽度优先遍历流程：

1.将初始的种子URL放入TODO表；

2.TODO表中取得一条链接，和 Visited表中的链接进行比较， 若Visited表中存在此链接（表示已被访问过），跳过不做处理；不存在（表示未被访问过），继续进行。

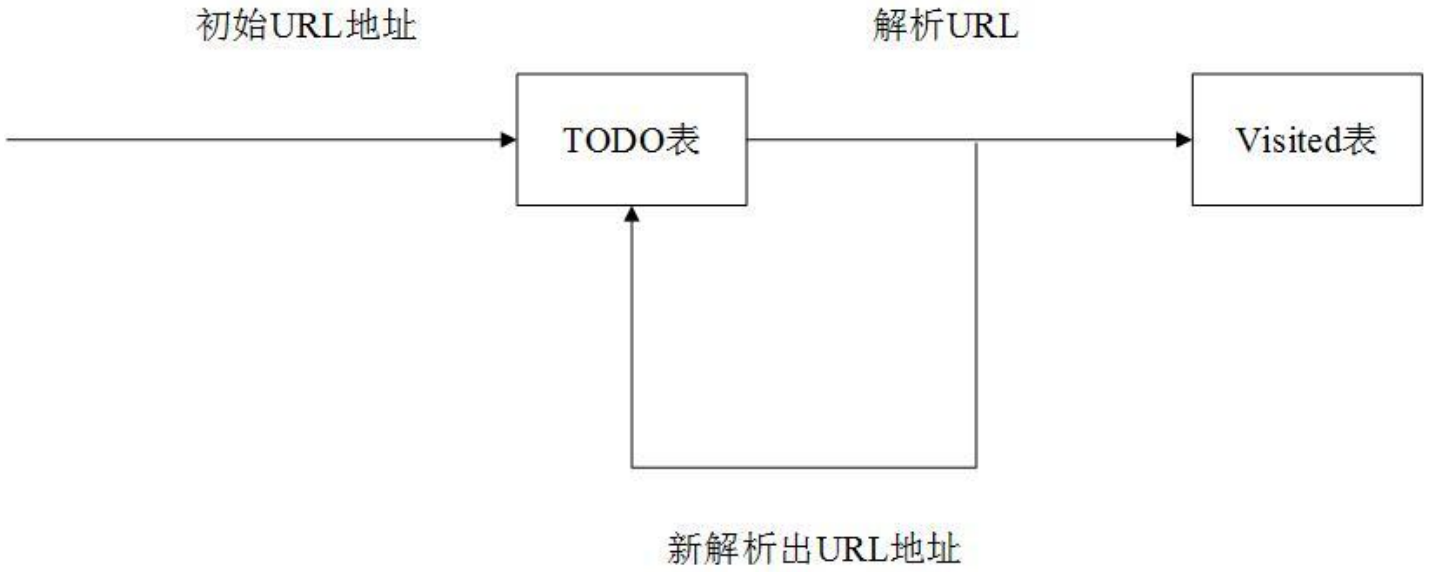
3.对链接进行解析，把页面中新解析出的URL放入TODO表中

4.处理完毕后， 将本页面的链接地址直接存入 Visited表中；

5.继续步骤2，循环往复。

网络爬虫策略

- 1、 图的宽度优先遍历



网络爬虫策略

- 1、 图的宽度优先遍历

宽度优先遍历爬虫优势：

- 重要的网页往往离种子比较近，距离越远重要性越低，宽度优先遍历能最先抓取重要页面；
- 万维网的最大深度能达到17层，到达指定网页总存在一条最短路径，宽度优先遍历会以最快的速度达到指定网页；
- 宽度优先有利于多爬虫的合作抓取， 多爬虫合作通常先抓取站内链接，抓取的封闭性很强。

网络爬虫策略

2、深度优先遍历

深度优先遍历类似于树的前序遍历。采用的搜索方法是尽可能先对纵深方向进行遍历。具体流程如下：

1.访问出发点V，并将其标记为已访问过；

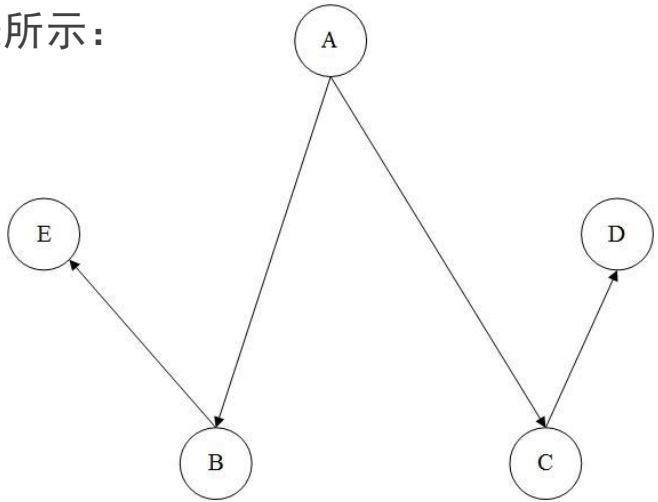
2.依次从V出发搜索V的每个邻接点W。若W未曾访问过，则以W为新的出发点继续进行深度优先遍历，直至所有和源点V有路径相通的顶点（亦称为从源点可达的顶点）均已被访问为止；

3.若此时仍有未访问的顶点，则另选一个尚未访问的顶点作为新的源点重复上述过程，直至所有顶点均已被访问为止。

网络爬虫策略

2、深度优先遍历

示例：如图选择A作为种子节点，则深度优先遍历的过程如表所示：



操作	队列中元素
初始	空
A入队列	A
A出队列	空
B入队列	B
B出队列	空
E入队列	E
C入队列	EC
E出队列	C
C出队列	空
D入队列	D
D出队列	空

遍历过程中出队列的节点顺序就是图的深度优先遍历的访问顺序：A->B->E->C->D

网络爬虫策略

宽度优先VS深度优先

- 广度优先遍历是以层为顺序，将某一层上的所有节点都搜索到了之后才向下一层搜索；
- 深度优先遍历是将某一条枝干上的所有节点都搜索到了之后，才转向搜索另一条枝干上的所有节点。

网络爬虫策略

3、 带偏好的爬虫

抓取URL时，给待遍历的网页赋予一定的优先级，根据这种优先级进行遍历，这种方法称为带偏好的遍历。

网络爬虫策略

3、 带偏好的爬虫

- 优先级的依据

判断网页的重要性的因素很多，主要有以下因素：

- 链接欢迎度 。 主要是由反向链接（ backlinks， 即指向当前 URL 的链接）的数量和质量决定的，定义为IB(P)。
- 链接的重要度 。 一个关于 URL字符串的函数，考察字符串本身（如 .com，.cc ），定义为IL(P)。
- 平均链接深度 。根据宽度优先的原则计算出全站的平均链接深度，然后认为距离种子站点越近的重要性越高， 定义为 ID(P)。

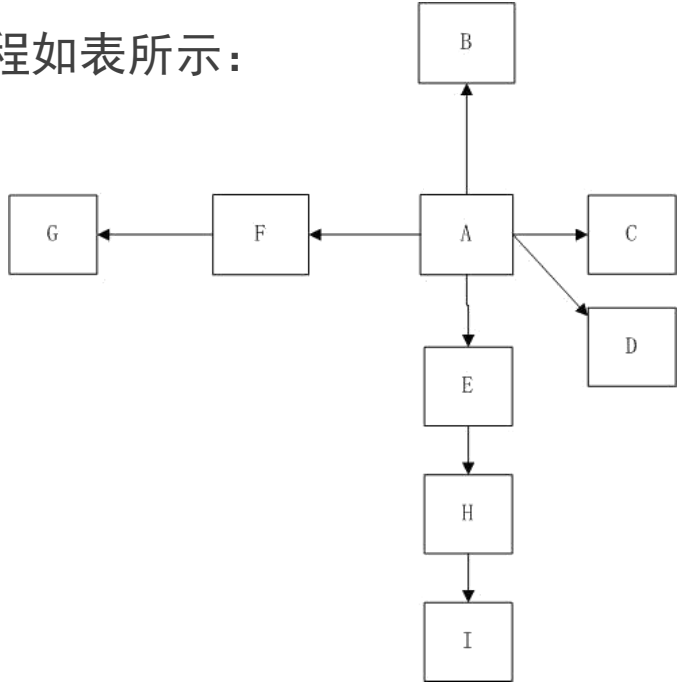
网络爬虫策略

- 最佳优先爬虫

实现最佳优先爬虫最简单的方式可以使用优先级队列来实现 TODO表，并且把每个URL的重要性作为队列元素的优先级。这样，每次选出来扩展的 URL就是具有最高重要性的网页。

网络爬虫策略

- 最佳优先爬虫示例：若图节点的重要性为D>B>C>A>E>F>I>G>H，则遍历的过程如表所示：



TODO表	Visited表
A	空
BCDEF	A
B,C,E,F	A,D
C,E,F	A,D,B
E,F	A,D,B,C
F,H	A,D,B,C,E
G,H	A,D,B,C,E,F
H	A,D,B,C,E,F,G
I	A,D,B,C,E,F,G,H
空	A,D,B,C,E,F,G,H,I

课堂小结

网络爬虫流程、分类
网页相关概率：http协议、请求与响应
网页结构（html，js，css）

谢谢聆听