大数据编程

4-1

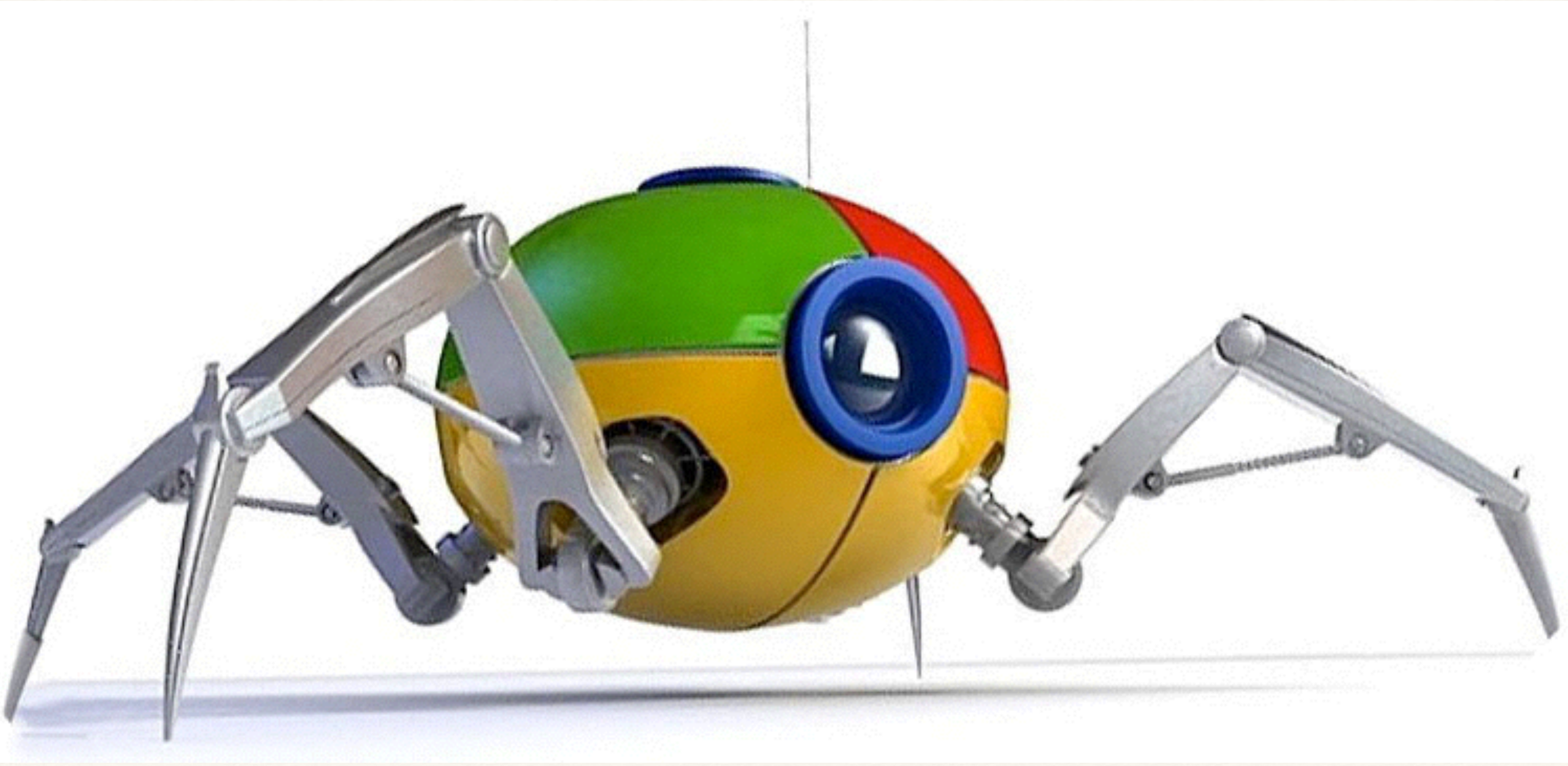
数据获取方法

中央财经大学 商学院
姚凯
2016

主要内容

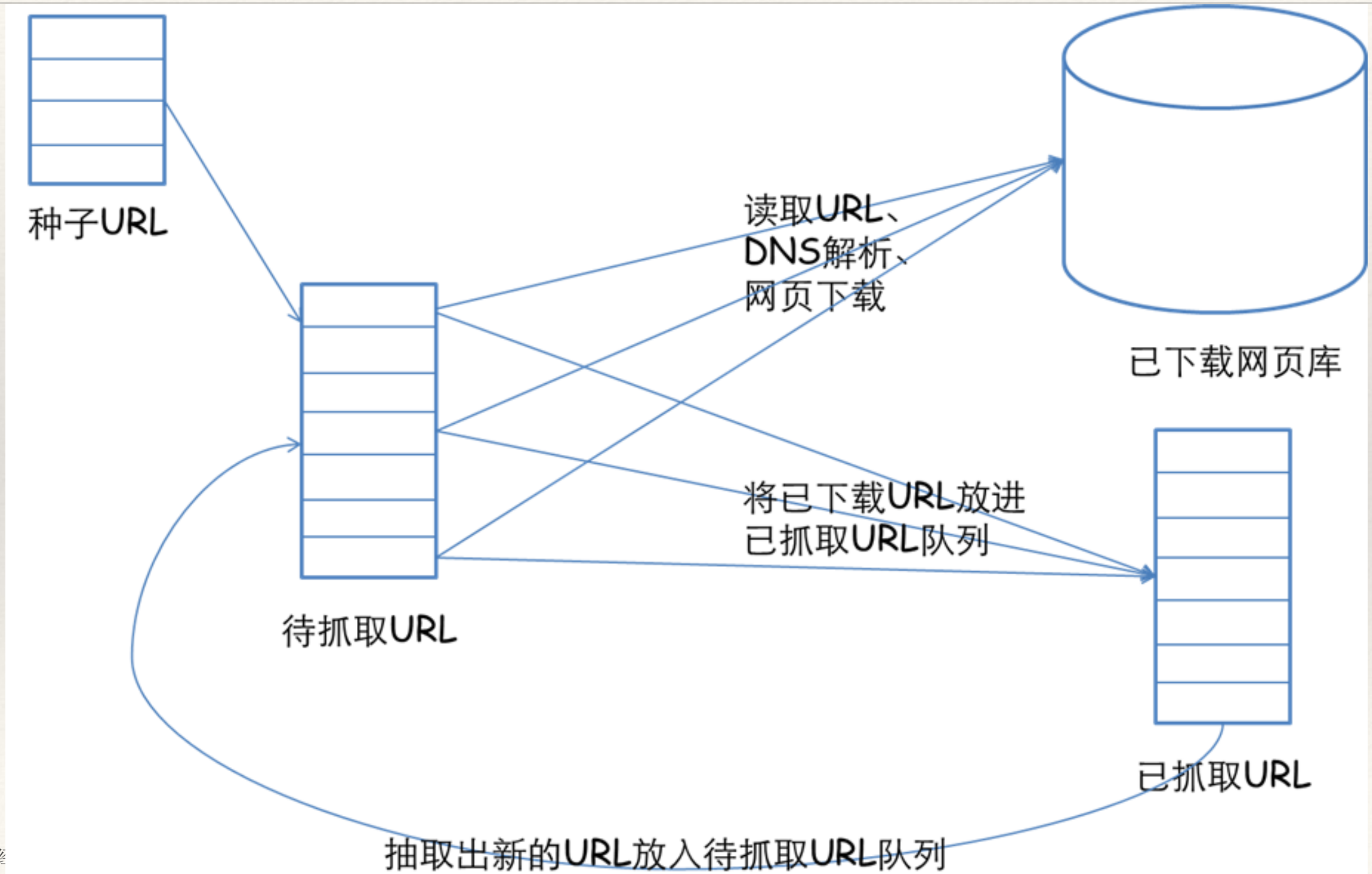
- 知识回顾
- 网页基础
- R语言编写爬虫
- 爬虫例子

网页爬取数据



中央财经大学　商学院　姚凯

网页爬虫的基本原理



数据爬取流程

1.获取URL网页内容

2.对网页内容进行解析

3.将解析后的数据存储到本地文件或数据库中

HTML基本标志

1. 文档标志

<HTML></HTML>。<HTML>标志用于HTML文档的最前面，用来标识HTML文档的开始。而</HTML>标志恰恰相反，它放在HTML文档的最后边，又来标识HTML文档的结束，两个标志必须成对使用。

2. 文件头标志

<head></head>。<head>和</head>构成HTML文档的开头部分，在此标志之间可以使用<title></title>、<script></script>等标志对。<head></head>标志对之间的内容是不会在浏览器的框内显示出来。两个标志必须成对使用。

HTML基本标志

3. 文件主体标志

<body></body>. <body></body>是HTML文档的主体部分，在此标志对之间可以包含 、<h1></h1>、
、<hr>等众多标志。它们所定义的文本、图像等将会在浏览器的框内显示出来。两个标志必须成对使用。<body>标志中还可以有如表1-1所示的属性。

属性	用途	示例
<body bgcolor=”# rrggbb”>	设置背景颜色	<body bgcolor=”red”>红色背景
<body textr=”# rrggbb”>	设置文本颜色	<body textr=”#0000ff”>蓝色文本
<body link=”# rrggbb”>	设置链接颜色	<body link=”blue”>链接为蓝色
<body vlink=”# rrggbb”>	设置已经使用的链接的颜色	<body vlink=” #ff0000”>

HTML基本标志

4. 文件标题标志

<title></title>。使用过浏览器的人可能都会注意到浏览器窗口最上边的蓝色部分显示的文本信息，那些信息一般是网页的“主题”。要将网页的主题显示到浏览器的顶部其实很简单，只要在<title></title>标志对之间加如显示的文本即可。

注意：<title></title>标志对只能放在<head></head>标志对之间。

下面是一个综合的例子，说明了HTML文档中最基本标志的使用。

```
<HTML>
<head>
<title>显示在浏览器最上边蓝色条中的文本</title>
</head>
<body bgcolor=”red”text=”blue”>
<p>红色背景、蓝色的文本</p>
</body>
</HTML>
```

页面格式标志

1. 段落标志

**(1). **

标志对是用来创建一个段落,在此标志对之间加入的文本将按照段落的格式显示在浏览器上。另外， 标志还可以使用align属性，它用来说明对齐方式，语法是 。align可以是Left（左对齐）、Center（居中）和 Right（右对齐）三个值中间的一个。

如： 表示标志对中的文本使用居中对齐方式。

(2).<per></per>

<per></per>标志队有来对文本进行预处理操作。

页面格式标志

2.换行标志

是一个很简单的标志，它没有结束标志，因为它是用来创建一个回车换行的。在
的使用上面还有一定的技巧，如果把
加在 标志对的外边,将创建一个很大的回车换行,即
前面和后面的文本的行与行之间的距离很大,若放在 的里面,则
前面和后面的文本行与行之间的距离比较小.

页面格式标志

3.列表标志

(1) <dl></dl>、<dt></dt>、<dd></dd>

<dl></dl>用来创建一个普通的列表,<dt></dt>用来创建列表中的上层项目,<dd></dd>用来创建列表中最下层项目, <dt></dt>和<dd></dd>都必须放在<dl></dl>标志对之间。

下面是一个创建普通列表的例子

```
<html>
<head>
<title>一个普通的列表</title>
</head>
<body tetx="blue">
<dl>
<dt>中国城市</dt>
<dd>北京<dd>
<dd>上海<dd>
<dd>广州<dd>
<dt>美国城市</dt>
<dd>华盛顿<dd>
<dd>芝加哥dd>
<dd>纽约<dd>
</dl>
</body>
</html>
```

页面格式标志

标志	含义
<table>	最外层，创建一个表格
<tr>	创建一行
<td>要输出的文本只能放在此处**</td>** **<td>**要输出的文本只能放在此处**</td>** **<td>**要输出的文本只能放在此处**</td>**	创建一个单元个（这里总共创建了三个单元格）
</tr>	行末尾
</table>	最外层

表格标志

<th></th>

<th></th>标志对用来设置表格头，文字通常是黑体、居中。

表格标志

```
<html>
<head>
<title>表格标志的综合示例</title>
</head>
<body>
<table  border="1"width="80%"bgcolor="#e8e8e8" cellpadding="2" bordercolor="30000ff">
<tr>
    <th width="33%" colspan="2" valign="bottom">意大利</th>
    <th width="36%" colspan="2" valign="bottom">英格兰</th>
    <th width="36%" colspan="2" valign="bottom">西班牙</th>
    <tr>
    <td width="16%" align="center">AC米兰</td>
    <td width="16%" align="center">拂罗伦莎</td>
    <td width="17%" align="center">曼联</td>
    <td width="17%" align="center">纽卡斯尔</td>
    <td width="17%" align="center">巴塞罗那</td>
    <td width="17%" align="center">皇家社会</td>
```

表格标志

```
<tr>
<td width="16%" align="center">尤文图斯</td>
<td width="16%" align="center">桑普多利亚</td>
<td width="17%" align="center">利物普</td>
<td width="17%" align="center">阿申纳</td>
<td width="17%" align="center">皇家马德里</td>
        <td width="17%" align="center">.....</td>
        <tr>
        <td width="16%" align="center">拉奇奥</td>
        <td width="16%" align="center">国际米兰</td>
        <td width="17%" align="center">切尔西</td>
        <td width="17%" align="center">米德尔斯堡</td>
        <td width="17%" align="center">马德里竞技</td>
        <td width="17%" align="center">.....</td>
        </table>
        </body>
        </html>
```

小结

- HTML是按照标签对构成
- 标签对之间可以按照内容进行嵌套或平行排列
- 标签对中不仅可以设定显示的内容，还能设置显示的格式
- 网页代码中还可以动态执行程序

主要内容

- 知识回顾
- 网页基础
- R语言编写爬虫
- 爬虫例子

第一种方式

1.从服务器获取网页内容

2.将网页内容视为字符串进行处理

3.利用正则表达式获取自己想要的内容

4.将结果保存至文件

第二种方式

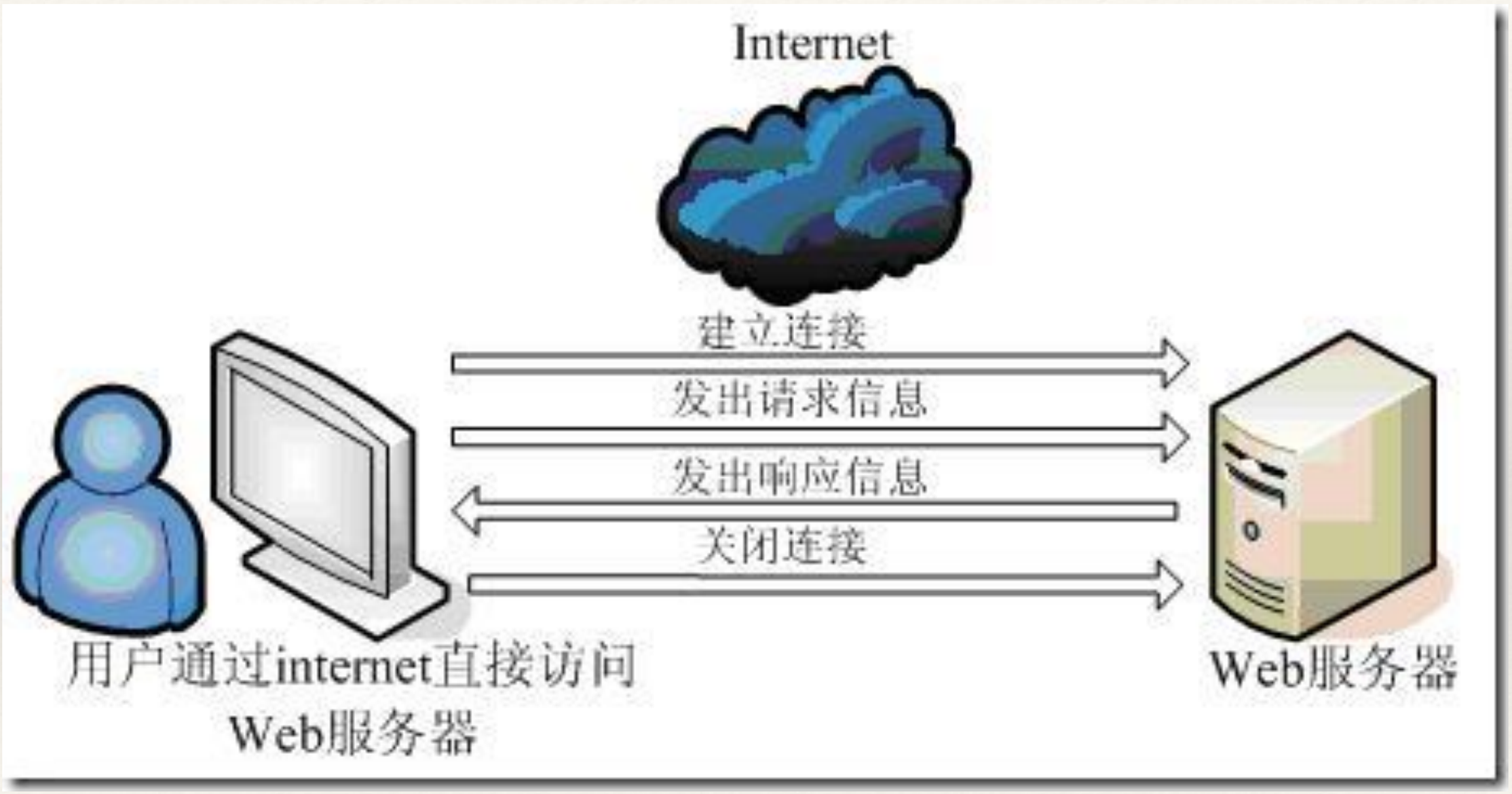
1.从服务器获取网页内容，保存为字符串对象

2.将网页内容转为XML数据进行处理

3.利用XML包获取自己想要的节点内容

4.将爬取结果保存至文件

用户访问网站过程


Internet
建立连接
发出请求信息
发出响应信息
关闭连接
用户通过internet直接访问
Web服务器
Web服务器

URL详解

基本格式:schema://host[:port#]/path/.../[?query-string][#anchor]

- scheme 指定低层使用的协议(例如:http, https, ftp)
- host HTTP服务器的IP地址或者域名
- port# HTTP服务器的默认端口是80,这种情况下端口号可以省略。
- path 访问资源的路径
- query-string 发送给http服务器的数据
- anchor- 锚

请求 （request）

METHOD /path - to - resource HTTP/Version-number
Header-Name-1: value Header-Name-2: value
Optional request body

- Method 表示请求方法,比如“GET”,“POST”,““HEAD”,”PUT“等
- Path-to-resource 表示请求的资源
- Http/version-number 表示HTTP协议的版本号

请求（request）

- 请求报头 **(header)**
  - Host 服务器地址
  - Accept 浏览器端可以接受的媒体类型，text/html
  - Accept-encoding 浏览器接收的编码方法，通常所指的是压缩方法
  - Accept-language 浏览器声明自己接收的语言
  - User-agent 告诉服务器客户端的操作系统、浏览器版本
  - Cookie 最重要的请求报头的成分，为了辨别用户身份、进行session跟踪而储存在用户本地终端上的数据（通常经过加密）
  - Referer 跳转页
  - Connection 客户端与服务器的连接状态

响应（request）

- HTTP/version-number表示HTTP协议的版本号
- status-code(状态码)： 状态码用来告诉HTTP客户端,HTTP服务器是否产生了预期的Response.
- ? HTTP/1.1中定义了5类状态码, 状态码由三位数字组成,第一个数字定义了响应的类别
- – 1XX 提示信息 - 表示请求已被成功接收,继续处理
- – 2XX 成功 - 表示请求已被成功接收,理解,接受
- – 3XX 重定向 - 要完成请求必须进行更进一步的处理
- – 4XX 客户端错误 - 请求有语法错误或请求无法实现
- – 5XX 服务器端错误 - 服务器未能实现合法的请求

RCurl包

- RCurl作者
- Duncan Temple Lang
- 现任加州大学 U.C. Davis分校副教授
- 致力于借助统计整合进行信息技术的探索

RCurl包

- RCurl三大函数
  - getURL()
  - getForm()
  - postForm()

XML

- Extensible Markup Language：可扩展标记语言
- 设计用来传输和存储数据
- 超文本标记语言（HTML）被设计用来显示数据

XML包

- 可以方便载入XML文件并提取有用信息、转换成R对象
- 在使用XML时候，需要使用XPath来获得需要的节点
- 主要函数用getNodeSet

XML包

- 解析xml文件的XPath设置
- 斜杠(/)作为路径内部的分割符
- /:表示选择根节点
- //:表示选择任意位置的某个节点
- @: 表示选择某个属性
- *表示匹配任何元素节点
- @*表示匹配任何属性值
- node()表示匹配任何类型的节点

大数据编程

4-2

数据获取方法

中央财经大学 商学院
姚凯
2016

第二种方式

1.从服务器获取网页内容，保存为字符串对象

2.将网页内容转为XML数据进行处理

3.利用XML包获取自己想要的节点内容

4.将爬取结果保存至文件

从服务器获取数据

- RCurl包
  - 判断URL是否存在用： url.exists()
  - 获取URL对应网页的内容用： getURL()

将网页内容转为XML数据

- XML包
  - 将网页转xml用： htmlParse()
  - 获取节点的内容用: getNodeSet()
  - 关键因素是设定获取节点内容的路径值: xpath

解析xml文件的XPath设置

- 斜杠(/)作为路径内部的分割符
- /:表示选择根节点
- //:表示选择任意位置的某个节点
- @: 表示选择某个属性
- *表示匹配任何元素节点
- @*表示匹配任何属性值
- node()表示匹配任何类型的节点

作业

- 阅读RCurl包文档，了解核心函数功能
- 阅读XML包文档
- 爬取一个电商网站数据

- 截止时间：2016.10.10
- 小组形式：小组名+作业2.pdf
- 附抓取代码
- 作业以报告形式呈现，说明成员工作内容，对核心步骤和心得进行说明