Xpath入门教程2

目录

CONTENTS

Xpath、Html概念

Xpath:是一种路径查询语言，简单的说就是利用一个路径表达式找到我们需要的数据位置。

Html:超文本标记语言，是用来描述网页的一种语言。主要用于控制数据的显示和外观。HTML文档也被称为网页。

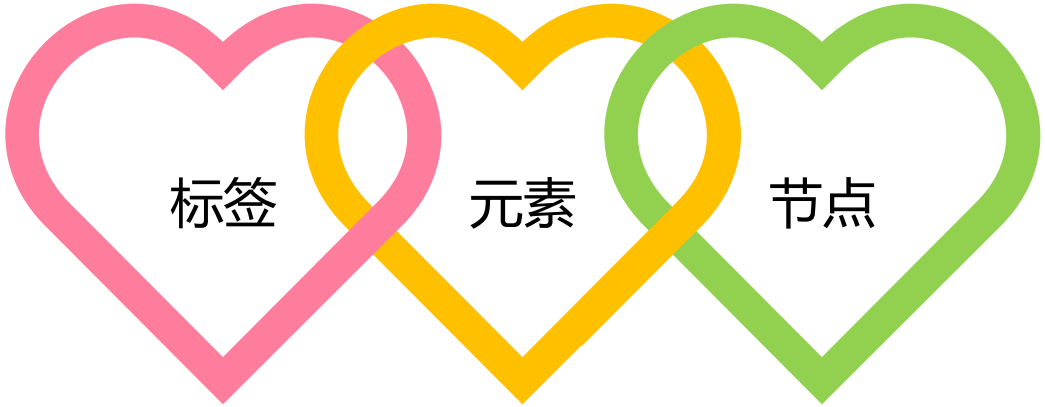
Xpath专用于xml中沿着路径查找数据用的，但是八爪鱼采集器内部有一套针对Html的Xpath引擎，使得直接用Xpath就能精准的查找定位网页里面的数据。

Html结构

```
<html>
<head>
页面的头部信息
</head>
<body>
页面的主体内容
</body>
</html>
```

完整的HTML文件至少包括<HTML>标签、<HEAD>标签、<TITLE>标签和<BODY>标签，并且这些标签都是成对出现的，开头标签为<>，结束标签为</>，在这两个标签之间添加内容。通过这些标签中的相关属性可以设置页面的背景色、背景图像等。

Html标签、元素、节点



作为开始和结束的标记
由尖括号包围的关键词，比如 <html>
标签对中的第一个标签是开始标签，第二个标签是结束标签

HTML的网页内容是由元素组成的，从开始标签到结束标签的所有代码。
元素的开始和结束都使用标签作为开始和结束的标记

所有事物都是节点
整个文档是一个文档节点
每个 HTML 元素是元素节点
HTML元素内的文本是文本节点
每个 HTML 属性是属性节点
注释是注释节点

Html常见标签

<a>

定义超链接，用于从一张页面链接到另一张页面

<h1></h1>

文本标题标签，最大的标签。从1到6，有6层选择

段落标记标签

可定义文档中的区域或节、可以把文档分割为不同的部分，是一个块级元素

创建一个列表

创建列表内容项

Html常见标签

<input>

用于搜集用户信息
可以是文本字段、复选框、按钮等等

<img></img>

向网页中嵌入一幅图像，从网页中链接图像

<table></table>

创建一个表格

<tr></tr>

表格中的每一行

<th></th>

设置表格头，通常是黑体居中文字

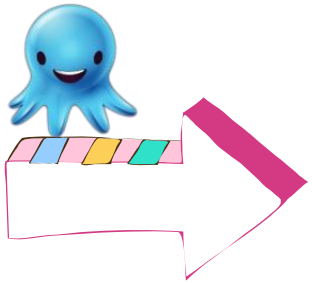
<option></option>

设置每个表单项的内容，选项

Html常见属性

class	id	href	src
规定元素的类名，大多数时候用于指定样式表中的类	唯一标识一个元素的属性，在html里面必须是唯一的	指定超链接目标的url	图像文件的url
1	2	3	4

属性是用来修饰标签的，放在开始标签里面



Xpath常见写法

text()

文本定位位置

contains()

用来判断字符串的一部分

contains(text(),'')

contains(@class, '')

positon()

表示节点的序号

last()

//div[last()]

following-sibling

当前元素的兄弟元素

and\or\not

and　并且与关系

or　并且或关系

not　不是