北京科海培训中心

新概念

XML

教程

范国平 编著

两张多媒体光盘：包括作者近10小时的实景讲座录像，内容生动丰富，通俗易懂。采用最新流媒体技术制作，操作简单，能有效的提高学习兴趣和效率。

北京科海集团公司 出品

北京科海培训中心

新概念 XML 教程

范国平　　编著

北京科海集团公司　出品

2001.9

内 容 提 要

XML 语言（扩展标记语言）是目前和将来最有前途的、在国际互联网上保存和传递信息的语言。它的语法十分灵活，使用起来特别方便。由于它能够直接描述信息内容，这一特点使得 XML 文档具备了构造和标识信息的能力，进而使文档处理程序可以轻松地对各类信息进行抽取、排列和过滤。

本书按照 W3C 的成熟标准和发展趋势，带领读者循序渐进地学习 XML 的相关知识，并向读者展示了如何在实际的网络应用中编写、维护和处理 XML 文档的示例。

全书共分为 11 章，主要包括：HTML 语言速成、初识 XML、使用工具、XML 文档的框架结构、格式良好的 XML 文档、用 DTD 定义文档类型、名域的使用、CSS 样式表、XSL 样式表、其他技术和 XML 典型示例。

本书语言简洁，内容丰富，涉及面广。书后所附的多媒体光盘采用了流媒体技术，邀请作者本人亲自讲解，并录制了其讲解过程，使讲解幻灯片、实验演示、课程目录索引与录像同步，并可随时跳转切换、反复播放、进退自如。具有在课堂上学习无法比拟的优越性。这样，不但图、文、声、像并茂，形式生动，内容翔实，而且有助于读者快速掌握本书内容。

品　　名：新概念 XML 教程
作　　者：范国平
责任编辑：成 洁
出　　品：北京科海集团公司
印 刷 者：北京市艺辉印刷有限公司
发　　行：新华书店总店北京科技发行所
开　　本：787×1092　1/16　　印张：17.375　　字数：400 千字
版　　次：2001 年 9 月第 1 版　　2001 年 9 月第 1 次印刷
印　　数：0001~5000
盘　　号：ISBN 7-89999-333-4
定　　价：35.00 元（2 张多媒体光盘）

前　言

随着XML被正式批准为W3C的建议标准，XML已成为Internet上最热门的话题。在技术上，各软件厂商纷纷推出支持XML的应用软件；在商业上，各行各业正在抓紧制定本行业的XML应用标准。这一潮流充分说明，在未来的互联网上，XML将成为信息的重要载体。所以，无论是网络内容供应商、网页设计者还是普通的上网用户，都有必要对XML有所了解。

相信对很多读者来说，XML还是一个很陌生的名词。它能干什么？它应该怎么使用？为什么现在XML如此引人瞩目？读者朋友一定会有很多疑问。请不要着急，当读完本书，掌握了书中提出的主要概念，看懂了作者给出的文档实例后，就会完全了解XML的。在进入一个全新的XML世界之前，首先请确定，读者朋友已经具备了如下的预备知识：

- 能够用文本编辑器编写基本的HTML页面，包括超链接的建立、图像的显示、文本规格的控制等。
- 能够使用主流浏览器（如IE、Netscape）浏览网页，加载页面。

这些要求是不是太低了？是的！学习XML以前，我们只要求这些。

如果读者熟悉HTML语言，有一点编写经验，就可以跳过第1章的内容，直接进入XML的学习。没有学过HTML的朋友们，请跟作者一起进入第1章，在那里将解答“为什么要先学习HTML？”

本书主要面向XML的初、中级用户。为了帮助初学者全面掌握XML的基本概念，并能快速编写XML应用程序，笔者将自己的实践经验整理成册，用生动的实例、大众化的语言将这一全新的技术奉献给国内的读者朋友。

全书共11章，内容包括：HTML语言速成、初识XML、使用工具、XML文档的框架结构、格式良好的XML文档、用DTD定义文档类型、名域的使用、CSS样式表、XSL样式表和其他技术等。最后安排了两个大型的实例，使读者能够应用和检验自己所学的知识。

由于时间仓促，本书难免有不足之处，恳请读者朋友批评指正。

编　者

2001年8月

目 录

第 1 章　HTML 语言速成

HTML（Hyper Text Markup Language）译为“超文本标记语言”。它不是一种编程语言，但却比编程语言简单，而且是Internet上最重要的语言。在Internet里，用户的浏览器通过URL向远程的服务器提出浏览页面的请求，于是，服务器将用户所需的Web页面传送到本地的计算机上，但这时用户还不能看到生动活泼、色彩斑斓的网页，因为网页的内容还是以一种特殊的代码，即HTML语言的方式存储在文件中。但是，很快浏览器就会自动将这些HTML语言翻译成最终可以展现在显示器上的生动页面。事实上，当读者看完本章后，就能看懂这个对现在来说还十分神秘的语言了。如果读者曾学过编程语言，就会感到HTML十分好学。HTML实际上就是一系列的规则，从形式上看，大多都是相应英语单词的缩写，如果学过英语那就非常容易了。

1.1　为什么要先学习HTML语言

之所以先要求读者学习HTML语言，主要是为了便于读者在后续内容中能更快地接受XML。为什么这么说呢？有两个理由：

1. 两种语言有可比性。

XML的完整名称是：Extensible Markup Language（可扩展的标记语言），也就是说它是一种标记语言，在这一点上，它和HTML（超文本标记语言）是一样的。请看下面的这个例子：

```
<题目> 这是一个“题目”标记 </题目>        ——用XML标记
<title> 这是一个“题目”标记 </title>      ——用HTML标记
```

可以看到，无论是XML还是HTML都可以用标签来标识文档中的文本元素，二者在形式上相差不是很大。当然，它们的含义和作用是有区别的，这一点将在以后说明。

事实上，XML和HTML都是SGML（标准通用标记语言）的子集。HTML是当今网络世界的主导语言，目前在Web上看到的缤纷世界，基本上都是用HTML描述出来的。如果很熟悉HTML，那么再来学习XML就会很轻松了。

2. 现在编写的XML文档，很大一部分都是面向网络应用的。

也就是说，这些文档最终都需要使用支持HTML的浏览器来把内容显示给用户。由于XML文档只注重内容，而不包括显示方式的控制，因此不能直接在浏览器上显示。这就要求XML文档的作者在编写完文档后，需要设计一个样式表，把文档内容按浏览器可以识别的方式显示出来。方法之一就是把XML文档按照设想转化成HTML文档的形式，然后再显示出来。显然，这里又对XML文档作者的HTML语言能力提出了要求。

所以，如果读者要想学好XML语言，最好先了解一下HTML的一些情况。当然，学习这部分内容不会占用读者太多的时间，作者会用尽量简洁的篇幅，把HTML语言最主要而

且最有学习价值的部分展现在读者面前。

1.2 HTML中关于文本显示的标记

HTML的很多标记都是用来控制文本显示的，例如，段落的对齐，文字的粗细等等。这一节将择要向读者介绍这些内容。

1.2.1 制作第一个HTML文档

先来看一个HTML文档，了解一下这种文件的主体框架，如文档清单1.1所示。

文档清单1.1

```
框架.htm（这是文件名，不包括在文件内容中）
<html>
<head>
“此处填写head的内容”
</head>
<body>
“此处填写body的内容——一般是HTML的正文”
</body>
</html>
```

注意：读者在做文档清单中的示例时，不要将首行的文件名也输入到文件内容中，否则会发生错误。在本书中，所有的HTML文件内容都从<html>标记开始。

由文档清单1.1可以看到，整个HTML文档由两两成对的“<>……</>”标签逐层嵌套而成，正文内容出现在一对“<body></body>”标签内，虽然只有几行文字，但已经是一个完整的HTML文件了。“原来做网页这么简单！”读者可能已经迫不及待地想尝试编写一个文档了。好的，下面马上来制作一个网页，看看效果如何。

编写网页的工具很多，但是在这里，建议大家使用最简单的文本编辑器，如Windows的记事本，因为这会有助于读者更加熟练地掌握HTML文档的编写。

现在，用苏轼的词来替换<body>标签中的内容，写成如文档清单1.2所示的形式，并存为Page1.htm。

文档清单1.2

```
Page1.htm
<html>
<head>
</head>
<body>
江城子
密州出猎
老夫聊发少年狂，左牵黄，右擎苍。
```

```
锦帽貂裘，千骑卷平冈。
欲报倾城随太守，亲射虎，看孙郎。
酒酣胸胆尚开张，鬓微霜，又何妨！
持节云中，何日遣冯唐？
会挽雕弓如满月，西北望，射天狼。
</body>
</html>
```

一定要把文档存为*.htm，即存为HTML格式的文件。读者在操作时可以有两种方法：一种方法是在保存时直接选择存为Web页；另一种方法是先将文件保存为.txt类型，然后通过对文件的重命名，将文件转换为HTML格式。

注意：*的内容可以任意，但一定要使用.htm的扩展名，这样刚刚制作的网页就可以直接用IE浏览器来打开了。

双击该文件，或者在IE中用“文件/打开”来显示刚才制作的网页。呵呵，是不是有些紧张？如图1.1所示的就是读者的第一个HTML作品在浏览器中的输出效果。

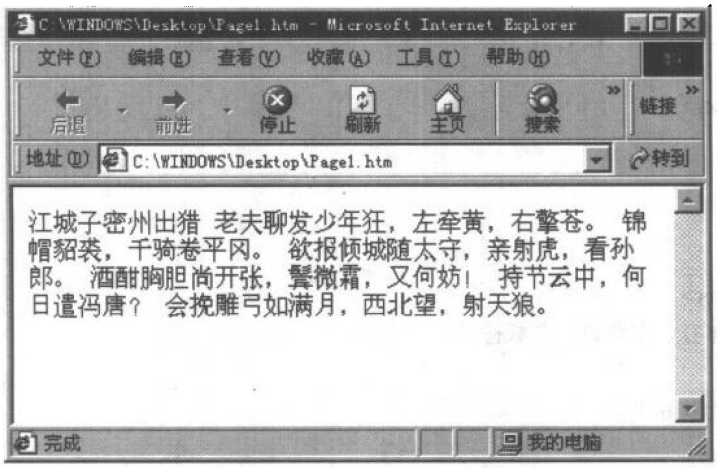

图 1.1　第一个 HTML 作品在浏览器中的输出效果

1.2.2　控制段落的显示

图1.1就是读者的第一个HTML文件的显示结果，是不是有些失望？因为文件的输出格式和预想的完全不一样。读者明明在编辑的时候分好了段落，也换了行，输出的文本却全混在了一起，没法分段，这是什么原因呢？

实际上，浏览器是很笨拙的，读者必须用标签来提醒它如何显示，例如，要分段就要使用<p>标签来标识，要显示红颜色，就得在某处注明color=red。读者在编辑时用的回车换行，浏览器是无法理解的，它不知道如何处理，只是简单地加以忽略，所以也不能期待浏览器会在输出页面时，做回车换行的处理。

注意：要想让浏览器输出预想的格式，就必须用HTML指定的标签来告诉它该怎么做。

现在介绍3个控制段落显示的标记。

- <p > 表示Paragraph（段落），作用相当于插入一个空行。
-
 表示Breakline，作用相当于换行。
- <hr> 表示Horizontal Rules，作用相当于插入一个水平线。

另外为了控制文本居中显示（或居左、居右显示），在<p>标签里可设置align（排列）属性：

- <p align=#> #=left, center, right

意思很好明白，现在就请读者利用这些标记和属性把先前撰写的文档作如下修改，如文档清单1.3所示。

文档清单1.3

Page2.htm

```
<html>
<head>
</head>
<body>
<p align=center>
江城子
<br>
密州出猎
<hr>
<p align=center>
老夫聊发少年狂，左牵黄，右擎苍。
<br>
锦帽貂裘，千骑卷平冈。
<br>
欲报倾城随太守，亲射虎，看孙郎。
<br>
酒酣胸胆尚开张，鬓微霜，又何妨！
<br>
持节云中，何日遣冯唐？
<br>
会挽雕弓如满月，西北望，射天狼。
</body>
</html>
```

文档清单1.3保存后，用IE打开，这个有段落控制的HTML文件的浏览效果，如图1.2所示。

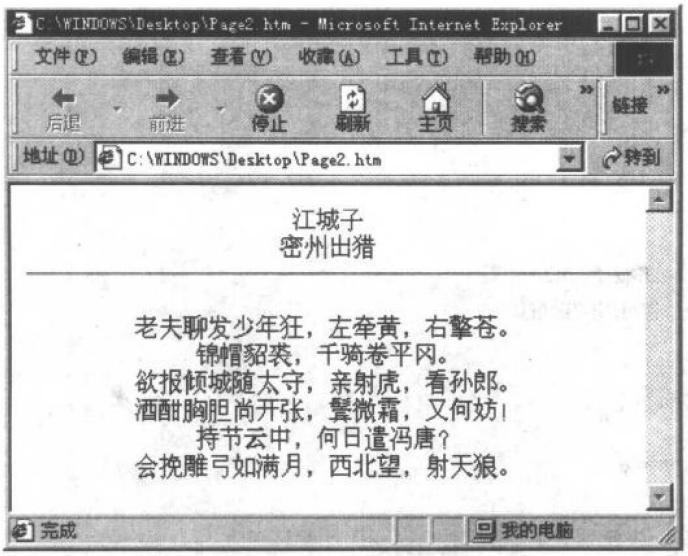

图 1.2　有段落控制的 HTML 文件的浏览效果

现在的效果比刚才好多了，但是读者可能想要使显示结果更加丰富多彩一些。下一步，就来学习如何控制文字的显示。

1.2.3　控制文字的显示

控制文字的显示，主要是向浏览器说明将采取多大的字号、是否加粗、是否采用斜体，以及文字用何种颜色显示等。首先介绍一下将要用到的标签及属性。

- <h#>……</h#>　　#=1, 2, 3, 4, 5, 6
  表示标题的大小，#代表的数目越大，字体越小。
  例如，<h1>标题</h1>，表示标签内的文字用一级标题显示出来。
- ……　　#=1, 2, 3, 4, 5, 6, 7
  表示字体的大小。
  例如，今天天气真好，表示用5号字体显示文本内容。
- ……
  表示用粗体显示文本。
- ……
  表示用斜体显示文本。
- ……
  表示用何种颜色显示文本，#=rrggbb 16进制数码，或者是下列预定义的色彩：Black，Olive，Teal，Red，Blue，Maroon，Navy，Gray，Lime，Fuchsia，White，Green，Purple，Silver，Yellow，Aqua。

以上介绍了几种常用的控制文字显示格式的标签，现在利用这些标签尝试着对源文件再做修改，如文档清单1.4所示。

文档清单1.4

```
Page3.htm
<html>
<head>
</head>
<body>
<h2 align=center>江城子</h2>
<h3 align=center>密州出猎</h3>
<hr>
<b>
<i>
<font size=3 color=red>
<p align=center>
老夫聊发少年狂，左牵黄，右擎苍。
<br>
锦帽貂裘，千骑卷平冈。
<br>
欲报倾城随太守，亲射虎，看孙郎。
<br>
酒酣胸胆尚开张，鬓微霜，又何妨！
<br>
持节云中，何日遣冯唐？
<br>
会挽雕弓如满月，西北望，射天狼。
</font>
</i>
</b>
</body>
</html>
```

文档清单1.4指定了“江城子”和“密州出猎”以2级和3级标题的形式输出。文中的具体内容以粗体、斜体和红色输出（本书在出版时，用的是黑白色，读者可能无法从书上区分出颜色的变化）。现在把文件保存，再用IE打开，会看到如图1.3所示的效果。

图1.3所示的结果和预期的完全一致。所以，只要按照HTML的要求设置标签，就可以随心所欲地显示文档内容了。

1.3 HTML中关于表格的标记

在实际应用中，经常要使用表格来汇总数据。例如，要得到一个食堂菜价的比较结果，并用表格的形式显示出来。应该怎么做呢？

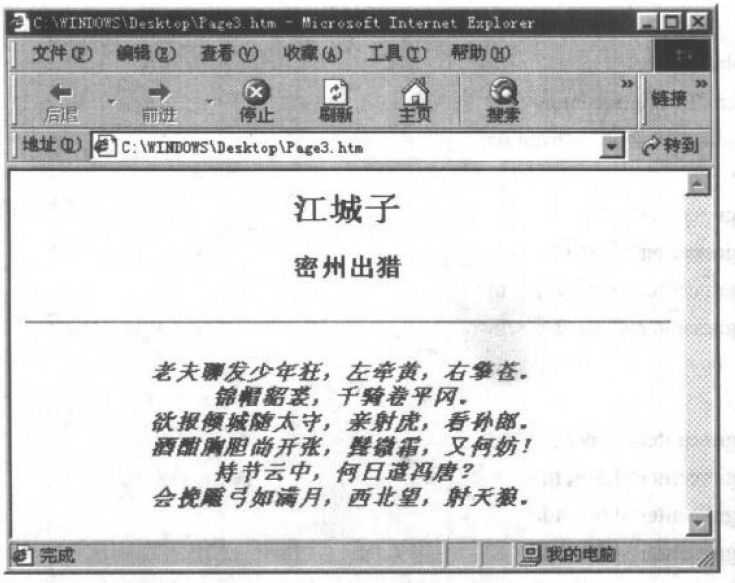

图 1.3　文字显示控制后的 HTML 文件的浏览效果

先介绍4个关于表格的标签。

- <table>……</table>　定义表格。
- <tr>……</tr>　定义表行。
- <th>……</th>　定义表头。
- <td>……</td>　定义表元（表格的具体数据）。
- <caption>……</caption>　定义表格标题。

再介绍5个附带在标签中的属性，通过设置这些属性，表格可以获得十分丰富的表现形式。

- <table border=#>　用于设置表格的尺寸。
- <tr align=#>　#=left, center, right　用于设置表行内文字的对齐。
- <th align=#>　#=left, center, right　用于设置表头内文字的对齐。
- <td align=#>　#=left, center, right　用于设置表元内文字的对齐。
- <th bgcolor=#>　用于设置表元背景颜色。

现在，利用这些标签来实现上文提到的表格的制作。这里把表格命名为“清华大学菜价表”，表中给出3个食堂3道相同热菜的价格，此外还设置了表元的背景色，并使表格数据居中对齐。

按照上面的设计计划，写出HTML文档，如文档清单1.5所示。

文档清单1.5

Page4.htm

```
<html>
<head>
```