21 世纪高职高专规划教材·计算机系列

XML 语言及应用

华铨平　张玉宝　编著

清 华 大 学 出 版 社
北京交通大学出版社
·北京·

内 容 简 介

本书是面向高等职业教育、高等专科学校和成人高等教育编写的计算机类教材，全书共分 9 章，内容包括 XML 概述、XML 语法、实体的定义与使用、DTD 与 Schema、使用 CSS 和 XSL 格式化 XML 文档、使用 DOM 和数据岛显示 XML 文档，以及 XML 技术的应用前景等。每章配有一定数量的例题和习题，从多个方面介绍 XML 语言及应用技术。

本书内容丰富、结构合理，技术阐述与实验指导相结合，由浅入深，由简到繁安排整个教学内容。本书不仅适合作为高职高专院校相关专业的教材，也可以供广大的 XML 技术爱好者参考。

图书在版编目（CIP）数据

XML 语言及应用 / 华铨平，张玉宝编著. — 北京: 清华大学出版社；北京交通大学出版社，2005.9

（21 世纪高职高专规划教材·计算机系列）

ISBN 7-81082-558-5

Ⅰ. X…　Ⅱ. ①华…　②张…　Ⅲ. 可扩充语言，XML-程序设计-高等学校：技术学校-教材　Ⅳ. TP312

中国版本图书馆 CIP 数据核字（2005）第 065030 号

责任编辑：谭文芳

出 版 者：清 华 大 学 出 版 社　　邮编：100084　　电话：010-62776969

　　北京交通大学出版社　　邮编：100044　　电话：010-51686414

印 刷 者：

发 行 者：新华书店总店北京发行所

开　　本：185×260　　印张：13.75　　字数：352 千字

版　　次：2005 年 9 月第 1 版　　2005 年 9 月第 1 次印刷

书　　号：ISBN 7-81082-558-5 / TP·206

印　　数：1～5 000 册　　定价：19.00 元

本书如有质量问题，请向北京交通大学出版社质监组反映。对您的意见和批评，我们表示欢迎和感谢。

投诉电话：010-51686043，51686008；传真：010-62225406；E-mail：press@center.bjtu.edu.cn。

出 版 说 明

高职高专教育是我国高等教育的重要组成部分，它的根本任务是培养生产、建设、管理和服务第一线需要的德、智、体、美全面发展的高等技术应用型专门人才，所培养的学生在掌握必要的基础理论和专业知识的基础上，应重点掌握从事本专业领域实际工作的基本知识和职业技能，因而与其对应的教材也必须有自己的体系和特色。

为了适应我国高职高专教育发展及其对教学改革和教材建设的需要，在教育部的指导下，我们在全国范围内组织并成立了“21世纪高职高专教育教材研究与编审委员会”（以下简称“教材研究与编审委员会”）。“教材研究与编审委员会”的成员单位皆为教学改革成效较大、办学特色鲜明、办学实力强的高等专科学校、高等职业学校、成人高等学校及高等院校主办的二级职业技术学院，其中一些学校是国家重点建设的示范性职业技术学院。

为了保证规划教材的出版质量，“教材研究与编审委员会”在全国范围内选聘“21世纪高职高专规划教材编审委员会”（以下简称“教材编审委员会”）成员和征集教材，并要求“教材编审委员会”成员和规划教材的编著者必须是从事高职高专教学第一线的优秀教师或生产第一线的专家。“教材编审委员会”组织各专业的专家、教授对所征集的教材进行评选，对列选教材进行审定。

目前，“教材研究与编审委员会”计划用2～3年的时间出版各类高职高专教材200种，范围覆盖计算机应用、电子电气、财会与管理、商务英语等专业的主要课程。此次规划教材全部按教育部制定的“高职高专教育基础课程教学基本要求”编写，其中部分教材是教育部《新世纪高职高专教育人才培养模式和教学内容体系改革与建设项目计划》的研究成果。此次规划教材编写按照突出应用性、实践性和针对性的原则编写并重组系列课程教材结构，力求反映高职高专课程和教学内容体系改革方向；反映当前教学的新内容，突出基础理论知识的应用和实践技能的培养；适应“实践的要求和岗位的需要”，不依照“学科”体系，即贴近岗位，淡化学科；在兼顾理论和实践内容的同时，避免“全”而“深”的面面俱到，基础理论以应用为目的，以必要、够用为度；尽量体现新知识、新技术、新工艺、新方法，以利于学生综合素质的形成和科学思维方式与创新能力的培养。

此外，为了使规划教材更具广泛性、科学性、先进性和代表性，我们希望全国从事高职高专教育的院校能够积极加入到“教材研究与编审委员会”中来，推荐“教材编审委员会”成员和有特色、有创新的教材。同时，希望将教学实践中的意见与建议及时反馈给我们，以便对已出版的教材不断修订、完善，不断提高教材质量，完善教材体系，为社会奉献更多更新的与高职高专教育配套的高质量教材。

此次所有规划教材由全国重点大学出版社——清华大学出版社与北京交通大学出版社联合出版，适合于各类高等专科学校、高等职业学校、成人高等学校及高等院校主办的二级职业技术学院使用。

21世纪高职高专教育教材研究与编审委员会

2005年7月

21 世纪高职高专规划教材·计算机系列
编审委员会成员名单

主 任 委 员　李兰友　边奠英

副主任委员　周学毛　崔世钢　王学彬　丁桂芝　赵　伟
韩瑞功　汪志达

委　　员　（按姓名笔画排序）

马　辉	万志平	万振凯	王永平	王建明
尤晓昳	丰继林	尹绍宏	左文忠	叶　华
叶　伟	叶建波	付晓光	付慧生	冯平安
江　中	佟立本	刘　炜	刘建民	刘　晶
曲建民	孙培民	邢素萍	华铨平	吕新平
陈国震	陈小东	陈月波	李长明	李　可
李志奎	李　琳	李源生	李群明	李静东
邱希春	沈才梁	宋维堂	汪　繁	吴学毅
张文明	张权范	张宝忠	张家超	张　琦
金忠伟	林长春	林文信	罗春红	苗长云
竺士蒙	周智仁	孟德欣	柏万里	宫国顺
柳　炜	钮　静	胡敬佩	姚　策	赵英杰
高福成	贾建军	徐建俊	殷兆麟	唐　健
黄　斌	章春军	曹豫莪	程　琪	韩广峰
韩其睿	韩　劼	裘旭光	童爱红	谢　婷
曾瑶辉	管致锦	熊锡义	潘玫玫	薛永三
操静涛	鞠洪尧			

前　言

可扩展标记语言（XML）是标准通用标记语言（SGML）的一个子集。它兼有超文本标记语言（HTML）和 SGML 二者之长，既全面通用，又简洁明晰，非常适合各种网络应用的需要。自 1998 年 XML 的标准推出以来，XML 技术受到行业的广泛关注和认同。基于 XML 的应用、支持 XML 软件和开发的工具越来越多，功能也越来越强，由于 XML 本身的开放性，它是连接不同系统、不同平台数据的桥梁，许多工具都提供了开放的源码，便于在此基础上进行二次开发。可以肯定，XML 将会在网络世界占有重要的地位。

XML 技术是当前炙手可热的网络技术之一，且 XML 技术的很多方面还处于开发和标准化过程中，本书将着重介绍 XML 技术中已成熟的标准和应用技术，包括什么是 XML、XML 的语法，以及 XML 的数据表示、数据存储、数据显示等数据处理技术。

本书在编写过程中始终体现“理论够用、讲清操作、注重应用”的原则。针对初学者的需求，由浅入深地介绍 XML 语言及应用技术。为便于读者的理解和实际应用，每章节还安排有大量的应用实例和练习。本书可以作为各类大专院校学生，尤其是高职高专类学生学习 XML 技术的专用教材，也可供广大程序员学习 XML 技术时参考。

本书第 1、5、6、9 章由华铨平编写，第 2、3、4、7、8 章由张玉宝编写，全书由华铨平统稿。本书实例源代码可在北京交通大学出版社网站 http://press.bjtu.edu.cn 上下载。

由于编者水平有限，编写时间仓促，书中难免有不妥之处，欢迎广大读者批评指正。

编者联系邮件地址：watchping@tom.com。

编　者

2005 年 8 月

目　录

第 1 章　XML 语言简介

本章要点：

- ☑ 什么是置标语言
- ☑ 什么是 XML
- ☑ HTML 的缺点和不足
- ☑ XML 的优点
- ☑ XML 的主要用途
- ☑ XML 实例

XML（eXtensible Markup Language）是一种可扩展的元置标语言，它的设计动机是要克服超文本置标语言（Hyper Text Markup Language，HTML）的种种不足，将网络上传输的文档规范化，并赋予标记一定的含义，与此同时，还要保留 HTML 所具有的简捷、适于网上传输和浏览的优点。因此，XML 最终以标准通用置标语言（Standard Generalized Markup Language，SGML）子集的形式出现，它集 SGML 和 HTML 的优势于一身，具有易于编辑、便于管理、适于存档、容易查询等诸多优势，已经成为网络发展的又一个亮点。

1.1　XML 的产生

1.1.1　什么是置标语言

在介绍 XML 之前，先介绍几个相关的名词概念。XML、SGML、HTML 中的“ML”（Markup Language）翻译成中文的含义就是“置标语言”，那么什么是置标语言？XML 与 SGML、HTML 有什么渊源呢？

“置标”的定义是：为了处理的目的，在数据中加入附加信息，这种附加信息称为置标。“置标语言”的定义则是：运用置标方法描述的形式语言。这两个定义都有些抽象，其实，“置标”的概念在现实生活中还是比较常见的。例如，在一段书面语言中，为了标注某一语句的重要，在这条语句下面画上下划线，这就是一种“置标”，这种通过下划线的置标方法是一种图形化的置标。下划线称为置标的标记，如例 1-1 所示。

【例 1-1】　图形化置标举例。

运用置标方法描述的形式语言，就称为置标语言。

用图形标记置标有一个缺点，就是标记的含义不明确，带有二义性。下划线标记，既可以理解为重点强调标记，也可以理解为名词标记。为了更准确地标识置标的含义，直接

用文字作为标记是一种很有效的办法。针对前面的例子，改用文字作为标记后，见例 1-2。

【例 1-2】 文字置标举例。

运用置标方法描述的形式语言，就称为<重要>置标语言</重要>。

例 1-2 中的<重要>称为起始标记，</重要>称为结束标记。这种用文字给出的标记不仅含义明确，而且便于计算机处理。

“置标”在计算机世界中的应用甚为广泛。文字编辑器借助置标来定义格式与外观；通信程序依靠置标来理解线路上所传输信息的语义；数据库通过置标来将数据字段与一定的含义相连，并表明字段之间的关系；多媒体应用中，置标则用来标识什么是图像的源数据、什么是声音的源数据。

从形式语言的角度讲，含有标记的书面自然语言，还不能算是置标语言。学习过程序设计语言的读者都知道，程序设计语言对语言中出现的语句、变量、表达式等都有严格的形式化的定义。置标语言虽然不是一般意义上的程序设计语言，但它与程序设计语言一样，是一种有严格语法定义的形式语言。

介绍置标语言，不能不提标准通用置标语言（SGML）。SGML 的前身是 IBM 公司为解决公司内部大量文档的交换和存储，于 1969 年发明的通用置标语言 GML（Generalized Markup Language）。经过十几年的完善和改进，由 GML 发展成为 SGML，并在 1986 年被国际标准化组织公布为国际标准——ISO8879。

SGML 是一个可以定义其他置标语言的元置标语言。通过 SGML 定义出来的置标语言实例有很多，但最知名、最流行的是在互联网上描述数据表现的 HTML。HTML 定义了一系列的标记，每个标记表明数据的一种显示格式。被置标后的文档（即同时包含纯文本和关于文本显示格式标记的文档）由一个 HTML 处理工具（最常见的是浏览器）进行读取，然后再根据标记所代表的显示规则来加以显示。

下面，通过一个例子来介绍 HTML 中的置标是如何发挥作用的。在 HTML 中，标记<B>的含义是要求 HTML 浏览器将一段文本以加粗的形式表示，而标记<CENTER>的含义是告诉浏览器将这段文本在一行的中间显示。所以，在浏览器中会将以下带有置标的数据以粗体居中显示。

```
<CENTER><B>浙江纺织服装职业技术学院</B></CENTER>
```

同样，例 1-3 这一段 HTML 代码显示了一个学生的信息列表。

【例 1-3】 HTML 代码举例。

```
<ul> 200120101</ul>
<li>季慧奇</li>
<li>女</li>
<li> 01 信管 1 班</li>
<li> 1985-1-2</li>
```

这段 HTML 置标数据在浏览器中的显示效果如图 1-1 所示。

- 200120101
- 季慧奇
- 女
- 01信管1班
- 1985-1-2

图 1-l HTML 置标数据显示结果

1.1.2　XML 的来源

XML 有两个先驱——SGML 和 HTML，这两个语言都是非常成功的置标语言，但是它们都在某些方面存在着与生俱来的缺陷。XML 正是为了解决它们的不足而诞生的。

SGML 的全称是标准通用置标语言，它从 20 世纪 80 年代初开始使用。正如 XML 一样，SGML 也可用于创建成千上万的置标语言，它为语法置标提供了异常强大的工具，同时具有极好的扩展性，因此在分类和索引数据中非常有用。目前，SGML 多用于科技文献和政府办公文件中。

但是，SGML 非常复杂，其复杂程度对于网络上的日常应用简直不可思议。不仅如此，SGML 非常昂贵。目前比较便宜的 SGML 软件之一是 Adobe FrameMaker，其标准版本价格为 850 美元，而 Adobe FrameMaker+SGML 是以 1995 美元售出的。还有最关键的一点，几个主要的浏览器厂商都明确拒绝支持 SGML，这无疑是 SGML 在网上传播遇到的最大障碍。

相反，HTML 免费、简单，而且它获得了广泛的支持。HTML 最初于 1990 年由 CERN 设计，它是一个非常简单的 SGML 语言，可以方便普通人的使用。而正如设计之初所构想的那样，HTML 目前在世界范围内得到了广泛的应用。

正因为如此，1996 年人们开始致力于描述一个新的置标语言，它既具有 SGML 的强大功能和可扩展性，同时又具有 HTML 的简单性。万维网联盟 W3C 决定专门成立一个 SGML 专家小组来从事此项工作，由 Sun 公司大名鼎鼎的 Jon Bosak 担任小组的指挥。

事实上，Bosak 和他领导的专家小组对 SGML 所做的贡献就像 Java 研究组对 C++做出的贡献一样。SGML 中所有非核心的、未被使用的和含义模糊的部分都被删除，剩下的就成为短小精干的置标工具——XML。对于 XML 的描述只有 26 页，而当初 SGML 的描述却长达 500 页之多。而值得一提的是，对于 XML 的描述尽管篇幅只是 SGML 的二十分之一，但 SGML 中所有的精华都被保留了下来。

这以后，XML 不断发展演化，并且从化学置标语言（Chemistry Markup Language，CML）和数学置标语言（Mathematical Markup Language，MathML）中汲取了大量的经验。1997 年春天，可扩展链接语言（eXtensible Link Language，XLL）草案已被拟定，到了 1997 年夏天，微软也开始了关于频道描述格式（Channel Definition Format，CDF）的定义工作，这应该算是 XML 的第一个真正的应用。

最后，XML 于 1998 年修成正果。W3C 于 1998 年 2 月批准了 XML 的 1.0 版本，一个崭新而大有前途的置标语言诞生了。

1.1.3　什么是 XML

XML 不但是置标语言，而且是可扩展的（extensible）置标语言，并非像 HTML 那样，提供了一组事先已经定义的标记，而是提供了一个标准，利用这个标准，可以根据实际需要，自定义新的置标语言，并为这个置标语言规定它特有的一套标记。因此准确地说，XML 是一种元置标语言，它允许程序开发人员根据它所提供的规则制定各种各样适合实际问题需要的置标语言。这也正是 XML 语言制定之初的目标所在。

XML1.0 标准中描述的制定 XML 的目标如下。

- XML 应该可以在互联网上直接使用。

- XML 应该支持各种不同的应用方式。
- XML 应该与 SGML 兼容。
- 处理 XML 文档的应用程序应该容易编写。
- XML 中的可选特性的数量应该减到最小，最好减至没有。
- XML 文档应该具有良好的可读性，并且比较清晰。
- 用 XML 设计新的置标语言应该方便快捷。
- XML 设计的置标语言应该正式、简洁。
- XML 文档应该容易编制。
- XML 标记的简洁性并不重要。

下面介绍一个非常简单的例子。如果需要定义一个新的置标语言，以便这个语言定义一些标记来描述学生及相关信息。这组标记很简单，如例 1-4（ch4-1.xml，该例对应的文件名，下同）所示，它们的优点是代表了一定的语意。与在 HTML 中用标记<UL>和<LI>表示这些信息相比，这种表示方法显然更加清晰易读。

【例 1-4】 XML 标记举例。

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE 学生列表 SYSTEM "xuesheng.dtd">
<学生列表>
    <学生>
        <学号>200120101</学号>
        <姓名>季慧奇</姓名>
        <性别>女</性别>
        <班级>01 信管 1 班</班级>
        <出生年月>1985-1-2</出生年月>
    </学生>
    <学生>
        <学号>200120102</学号>
        <姓名>凌怡</姓名>
        <性别>女</性别>
        <班级>01 信管 1 班</班级>
        <出生年月>1983-6-20</出生年月>
    </学生>
</学生列表>
```

这一段代码是一个非常简单的 XML 文档（document）。看上去它和 HTML 非常相像，但细心的人会发现这里的标记所代表的不再是显示格式，而是对于学生信息数据的语义解释。事实上，用 XML 定义的置标语言可以根据标记描述的侧重点不同而分为两大类。一类偏重于语意描述，如例 1-4 所示。还有一类偏重于显示方式的描述，像现在已经出炉的扩展文本置标语言（eXtensible Hyper Text Markup Language，XHTML）、可缩放矢量图形语言（Scalable Vector Graphic，SVG）和同步多媒体综合语言（Synchronized Multimedia Integration Language，SMIL）。值得一提的是，这里对于显示方式的描述不仅限于对文本的描述，还可以包括矢量图形、图像和声音。比如，一个形如<强调>的标记在描述文本时可能是要求将文本加粗，而在描述声音时则要求将音量加大。

不过，仅仅将数据置标还不够。为了让别人读懂这些数据，置标语言中的置标标准还需包括置标的语法和每个置标的含义。

换句话说，如果想让计算机应用程序读懂并能处理这段数据，它还必须知道什么是一个有效的置标（如标记），如何处理一个有效的置标。具体地说，浏览器如何知道怎样显示上面的这段 XML 文档？标记<学号>是什么含义？它究竟是不是一个合法的标记？它又应该以什么方式表现？因此，置标语言必须能够告诉应用程序它所采用的置标的语法，以便应用程序能够对其进行正确的处理。

在 XML 中，置标的语法是通过文档类型定义（Document Type Definition，DTD）或 Schema 来描述的。也就是说，通过 DTD 或模式（Schema）来描述什么是有效的标记，从而进一步定义置标语言的结构。第 3 章将详细讨论 DTD 和 Schema 的定义方法，这里先通过例 1-5（xuesheng.dtd）了解一下关于例 1-4 中用到的 DTD 文件“xuesheng.dtd”中的内容。

【例 1-5】　DTD 文件“xuesheng.dtd”。

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT 学生列表 (学生*)>
<!ELEMENT 学生 (学号, 姓名, 性别, 班级, 出生年月)>
<!ELEMENT 学号 (#PCDATA)>
<!ELEMENT 姓名 (#PCDATA)>
<!ELEMENT 性别 (#PCDATA)>
<!ELEMENT 班级 (#PCDATA)>
<!ELEMENT 出生年月 (#PCDATA)>
```

除了定义置标的语法外，还需定义置标的含义，以便正确地加以显示。为了明确各个标记的含义，XML 使用与之相连的样式单（Style Sheet），由它来向应用程序（如浏览器）提供如何处理显示的指示说明。样式单的使用在第 4 章具体描述，现在只简单介绍样式单所做的规定可能是如下所示的。

（1）每当看到一个<学号>标记，用一个<UL>标记显示它。同样，</学号>转换为一个</UL>标记。

（2）所有的<姓名>、<性别>、<班级>、<出生年月>标记被转换为<LI>标记加以显示。同样，</姓名>、</性别>、</班级>、</出生年月>转换为一个</LI>标记。

例 1-5 中，使用 HTML 的标记功能来定义该 XML 数据文档的显示格式。但如果 XML 文档不是由浏览器，而是由其他应用程序来进行处理，可能会采用其他相应的标记来输出。这样，应用处理程序就需要综合 DTD、样式单及 XML 文档数据三方面要素，根据这些数据和规定来显示它。

表面上看，原先只要一个 HTML 文档就能把数据和显示方式都包括进去，现在却需要 XML 文档、DTD 和样式单三个文件来共同完成。同时，我们也知道，浏览器只是用来处理一种特定的置标语言（比如 HTML）的，而不是用来处理所有置标语言的，这说明可能还需要制作或购买一个额外的程序来处理 DTD、样式单和 XML 文档，这显得更复杂了。的确，对于初学者而言，在使用 XML 时确实会感到一些困难，不过在下面一节中学过 XML 的诸多优点之后，读者将看到这是非常值得的。

1.2 为什么要发展 XML

1.2.1 HTML 的缺点和不足

HTML 是最早应用于网络信息传输的置标语言，也是近几年互联网上最普及的一种网页制作通用语言。它侧重于主页表现形式的描述，大大丰富了主页的视觉和听觉效果，为推动 WWW 的蓬勃发展、推动信息和知识的网上交流发挥了不可取代的作用。可是，HTML 自身的特点使它蕴藏了许多危机，随着它不断地发展，这些危机不但没有减弱，反而越来越突出，甚至已然成为 HTML 继续发展应用的障碍。

HTML 制定之初的本意在于根据信息的含义来为它们置标，而没有具体规定它们应该如何在浏览器中显示。在 HTML 的早期版本中，<title>代表题目，<h1>代表第一层的大标题，<h2>代表第二层的大标题，、代表强调的文本，<address>代表作者的联系信息。至于这些题目、各层大标题究竟如何显示，应该由浏览器决定，因为 HTML 标准的制定者相信，比起网页的制作人员，浏览器更了解用户的偏好和使用的浏览环境。显然网页的制作者事先并不知道哪些用户决定不显示图片，又有哪些用户喜欢大一些的字体，只有浏览器才能保证为这些特殊用户提供良好的支持。

但是，浏览器的开发者同样也无从了解这些特殊用户的偏好。他们引入了自己定义的一些标记和属性，用这些新的标记来专门描述显示格式，比如标记<font>、 、<bgcolor>，等等。浏览器厂商还开发了自己的网页制作软件，如 Netscape 开发的 Netscape Composer，微软开发的 Frontpage，等等。这些所见即所得的网页制作工具自动生成 HTML 文件，而这些 HTML 文件更是忽略了标记的语义信息，而几乎完全将它们作为格式表现的工具。比如说，现在关于表格的标记（如<table>、<tr>、<td>等）不仅可以代表表格中不同行、列的信息，还可能专门用于网页布局。这样一来，HTML 越来越侧重于信息的表现方式，标记中原本就很微弱的对信息含义的描述也被削弱了。最后，HTML 终于演变为专门用于 Netscape 和 Microsoft IE 两大浏览器的页面显示语言。

可能读者会觉得虽然某些有特殊癖好的用户的要求得不到满足，但毕竟对大多数人而言，浏览页面最基本的显示问题还是解决了；而且，有了这些专门的显示标记，这个问题可能还解决得不错，其实不然。浏览器生产厂家在激烈的市场竞争中，为了显示自己的独特性，给 HTML 加入了一些特殊的标记，以便为自己的浏览器增加一些特殊的显示效果。日益增多的标记不但使 HTML 越来越庞大，浏览器的开发越来越复杂，还降低了不同浏览器之间的兼容性。比如说网页是针对 IE5 浏览器 800×600 像素的屏幕分辨率来制作的，那么在 640×480 像素的屏幕上观看的效果就会大打折扣，而如果放到 Netscape 浏览器中，显示效果与最初的设计构想甚至会大相径庭。

不仅如此，尽管 HTML 的标记越来越多，其显示力却还远远不够。如果希望非常精确地表现一些自己的数据，可能需要一些现在 HTML 中尚不存在的标记。一个化学家，可能需要表现化学分子式中的一些特别的符号。一个飞机设计师，可能希望能够表现飞机引擎的三维曲线造型。可对于这些，HTML 都望尘莫及。要想满足各行各业对显示的不同要求，显然需要大量的标记，这无疑使当今日益臃肿的 HTML 雪上加霜。

问题还不止这些，现在 HTML 内部结构的条理性越来越差。程序员写的 HTML 文件，甚至是那些专门的所见即所得工具自动生成的 HTML 文件，可能在语法上会错误百出，但浏览器照样能阅读它。HTML 中的标记可以不满足嵌套关系的层次完整性，比如<h1><h2></h1></h2>，也可以不配对出现，只有<h1>而没有</h1>，更不会要求在使用标记二级标题<h2></h2>的外面一定要保证有一级标题<h1></h1>。乍一看，这仿佛对网页制作者而言是个福音，可浏览器的开发者就不得不把大量的精力耗费在文法错误的容错上，相应地，浏览器的程序也变得复杂，甚至牺牲浏览时的时间效率和空间效率。

另外，HTML 也无助于搜索引擎的开发。因为从 HTML 的标记本身，搜索引擎几乎得不到任何有用的信息。如果需要到网络中找出世界上所有关于 XML 书籍的价钱，搜索引擎将不得不分辨网络中哪些“XML”字段对应的是书名，同时要知道这些书名所对应的价钱。如果根据数据库进行搜寻，数据库中的各个字段都有着明确的含义，问题比较好解决。但搜索引擎在网络中是根据 HTML 文件来进行搜索的，那些原本条理清晰，层次分明的数据库内容在 HTML 文件中已经被各种各样的标记所拆散，搜索引擎的任务就艰巨多了。

概括起来说，HTML 有以下几个固有弱点。

- HTML 是专门为描述网页的表现形式而设计的，它疏于对信息语义及其内部结构的描述，不能适应日益增多的信息检索要求和存档要求。
- HTML 对表现形式的描述能力实际上也还是非常不够的，它无法描述矢量图形、科技符号和一些其他的特殊显示效果。
- HTML 的标记集日益臃肿，而其松散的语法要求使得文档结构混乱而缺乏条理，导致浏览器的设计越来越复杂，降低了浏览的时间效率与空间效率。

1.2.2　XML 的优点

1．XML 良好的可扩展性

在 XML 产生之前，要想定义一个置标语言并推广利用它非常困难。一方面，如果制定了一个新的语言而期望它能生效，需要把这个标准提交给相关的组织（如 W3C），等待它接受并正式公布这个标准，经过几轮的评定和修改，到这个置标语言终于成为一个正式推荐标准时，可能已经过了几年的时间。另一方面，为了让这套标记得到广泛应用，制订者必须为它配备浏览工具。这样，就不得不去游说各个浏览器厂商接受并支持新制定的标记，或者索性自己开发一个新的浏览器去与现有的浏览器竞争，无论哪个办法，都需要耗费大量的时间和工作。现在借助 XML 的帮助，制定新的置标语言要简单易行得多了，这也正是 XML 的优势所在。

大家都知道，各个不同的行业可能会有一些独特的要求。比如说，化学家需要化学公式中的一些特殊符号，建筑家需要设计图纸中的某些特殊的标记，音乐家需要音符，这些都需要单独的标记。但是，其他网页设计者一般不会用这些记号，也不需要这些标记。XML 的优点就在于它允许各个组织、个人建立适合他们自己需要的标记库，并且这个标记库可以迅速地投入使用。

不仅如此，随着当今世界越来越多元化，要想定义一套各行各业能够普遍应用的标记既困难，也没有必要。XML 允许各个不同的行业根据自己独特的需要制定自己的一套标记，同时，它并不要求所有浏览器都能处理这成千上万个标记，同样也不要求置标语言的