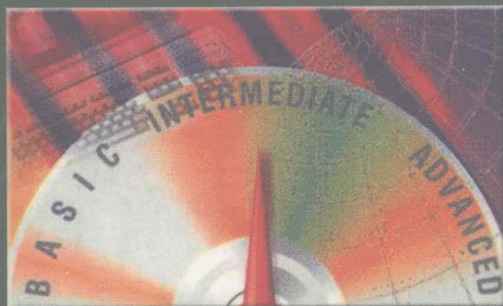


印度优质 IT 职业教育教学用书



XML 核心技术

Comp-U-Learn Tech India Ltd.

贾素玲 王 强 许 珂 罗 昌 编译



高等教育出版社
HIGHER EDUCATION PRESS

印度优质 IT 职业教育教学用书

XML 核心技术

Comp-U-Learn Tech India Ltd.

贾素玲 王 强 许 珂 罗 昌 编译

高等教育出版社

内容提要

本书是引进的印度优质 IT 职业教育教学用书, 是 Comp-U-Learn Tech India Ltd. 的培训课程用书, 与 COMP-U-LEARN 软件开发职业资格证书相配套。

本书共八章, 主要内容包括: XML 简介, XML 文档与 DTD, 使用样式表显示 XML, XSL 转换, XML 的 API, 模式、命名空间与 XHTML, XLink 与 XPointer 等。同时本书配有相应的指导手册。

本书适用于高等职业学校、高等专科学校、成人高校、本科院校举办的二级技术学院, 也可供示范性软件职业技术学院、继续教育学院、民办高校、技能型紧缺人才培养培训使用, 还可供本科院校师生、计算机从业人员和爱好者参考使用。

图书在版编目(CIP)数据

XML 核心技术 / Comp-U-Learn Tech India Ltd.; 贾素玲等编译. —北京: 高等教育出版社, 2005.4
ISBN 7-04-017181-3

I. X... II. ①C...②贾... 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 018860 号

策划编辑	冯 英	责任编辑	焦建虹	封面设计	张 楠	责任绘图	杜晓丹
版式设计	王艳红	责任校对	王效珍	责任印制	韩 刚		

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100011
总 机 010-58581000
经 销 北京蓝色畅想图书发行有限公司
印 刷 北京市鑫霸印务有限公司

购书热线 010-58581118
免费咨询 800-810-0598
网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landaco.com>
<http://www.landaco.com.cn>

开 本 787×1092 1/16
印 张 26.5
字 数 600 000

版 次 2005 年 4 月第 1 版
印 次 2005 年 4 月第 1 次印刷
定 价 42.00 元

本书如有缺页、倒页、脱页等质量问题, 请到所购图书销售部门联系调换。

版权所有 侵权必究
物料号 17181-00

The Team

Production/Series Supervision

Jambu Krishnamurthy

Vice President-Product Development

The Head

Sridhar Subramaniam

Head-Training Solutions

编审委员会

主任 陈 明

副主任 董 葛 尹 洪 贾素玲

委 员 (中文姓氏按拼音顺序排序)

Vikram Gupta (印度)

Roxanne Ang (加拿大)

冯 英 贾俊义 刘 超

倪金生 任树森 万世基

王克勤 王 强 王人骅

王树英 许 珂 严 亮

杨小平 张学群 郑晋梅

朱永庆

前 言

这是一套难得的 IT 教材，因为它是我们师徒几个千辛万苦从西天取回的“真经”。

众所周知，印度是当今世界计算机软件出口王国，但上帝为何如此偏爱，把这个光环戴在了她的头上，人们并不十分清楚。一位印度朋友告诉我，印度软件业的发展之所以令世人赞叹，其原因除政府长期实施一系列扶持政策外，主要是得益于行之有效的培训教育。这一点，在我们对印度考察时得到充分的印证。为贯彻全国职教会关于“积极引进国外优质职业教育资源”的精神，教育部所属中国华育发展总公司和北京华育发国际技术培训有限公司，于 2003 年组织专家对印度进行了为期两周的考察学习，先后走访了印度软件行业协会、4 所大学以及 7 家软件培训机构。通过实地考察、分析和比较，结合中国软件产业发展情况及软件教育培训现状，专家认为 Comp-U-Learn Tech 印度有限公司 (Comp-U-Learn Tech India Ltd.) 的培训课程比较适合我国职业技术学院计算机软件专业教学，经研究最终决定引进 Comp-U-Learn 系列教材。

Comp-U-Learn Comp-U-Learn 于 1993 年在美国密歇根州成立，是北美开展计算机软件职业培训的先导者，在培训中形成了独特的 ICAM 教学理念，已有上万名的学员在财富 500 强、IT 行业和其他公司从事相应的工作。Comp-U-Learn Tech 印度有限公司也有超过 2 万 5 千多名学生受益于 Comp-U-Learn 的 IT 培训计划。Comp-U-Learn 的教材和课件由美国统一研发，并在全球推广。Comp-U-Learn 的软件开发职业资格证书在美国、印度、加拿大等国家业内得到广泛认可。

课程设置 本套教材的形成，一是遵循教育部《两年制高等职业教育计算机应用与软件技术专业领域技能型紧缺人才培养指导方案》的精神，以培养技术应用人才为目标，力求在课程和课时设置上与高职院校的教学计划无缝衔接；二是以企业对人才的需求为依据，把软件工程的思想完全融入教材体系中，将基本技能培养和主流技术相结合，课程设置中重点突出、主次分明、结构合理、衔接紧凑。据此，我们对 Comp-U-Learn 的数本教材进行了精选和组合，最终完成以 Java 开发为主线的职业教育课程设置。

按照循序渐进的方式，课程设置分为 6 个部分，其中计算机基础类有《计算机基础》，语言基础类有《C 语言编程》、《Java 编程基础》、《HTML 设计》，数据库基础类有《基于 Oracle 的数据库系统》，分析工具有《UML 系统分析设计》，Java 进阶类有《高级 Java 编程》、《JSP 开发技术》、《JavaScript 程序开发》、《XML 核心技术》，企业级开发类有《EJB 企业级应用开发》、《J2EE 技术实践》，共计 12 本教材。这样的课程设置，构成了一个完整的 Java 培训体系结构，它从最基本的《计算机基础》入手，渐进到当今最流行的《J2EE 技术实践》，可使初学计算机的人在学完本套教材后，成为熟练使用各种 Java 工具的企业级开发专业人才。

教材特点 凡阅读过 Comp-U-Learn 教材的朋友,都会有相似的感觉,无不佩服作者面对复杂技术的实践体系,运用巧妙的方式而设计的编排模式。每章开始均列有学习目标和本章内容,学习目标可作为评估学生掌握本章知识的参考标准;本章内容实际上为教学重点,既有助于教师备课,又有助于学生掌握要点。每章的结尾都配备有复习题,使得学生能够在学习的基础上进一步思考,巩固前一阶段所学的知识,以便得到知识的升华。每本教材最后还添加了附录和词汇表,以便学生扩展相关主题的知识面,帮助学生梳理知识结构,同时也为学生展示了新的探索和发展空间。

侧重培养学生的实战操作能力是该系列教材的另一特点。在“学”和“思”的基础上,本教材还增加“练”的环节,除《J2EE 技术实践》外,每册教材都配有单独的案例指导手册,并附有完整的解决方案,旨在通过项目实践,增强学生的编程能力,使知识从书本中释放出来,转化为专业技能。

此教材还有一个特点是课件配套齐全。不仅有学生用书、案例指导手册,还配备有教师授课用 PPT、考试题库和对应的职业资格认证。PPT 可用来帮助教师教学和引导学生深刻理解每章主题、轻松完成相关实例、准确掌握概念定义。职业资格认证对应教学的内容,分为 CUL 软件编程助理工程师、软件编程工程师和软件开发工程师。学生考试合格后,一旦取得证书,就等于获得了迈入 IT 行业的绿色通行证(详细信息可登录 www.culnet.net.cn 或 www.ceditt.com)。

译者 为使这套引进的教材能原汁原味地呈现给读者并更好地服务于职业教育,我们组建 CED-CUL 学术委员会,制订了针对性的编译方案。本套系列教材是由北京航空航天大学的贾素玲教授、王强副教授主持编译。他们从事多年的信息系统开发,在信息技术领域有着广泛的涉猎和独到的见解。他们严谨的学术态度和精深的专业造诣,保证了本套系列教材的翻译质量。从译稿到审稿,无不凝聚着几位教授的心血。

参加该系列教材翻译工作的老师还有:许珂、陈当阳、罗昌、张成、张剑、韩小汀、卢根、姚琪琳、张国强、王卫星、柴庆慧、朱磊、姜浩、张小强、孙海涛、武心清、孔令鹏、傅玉、郝娟君、姚冠扬、潘雪梅、赵宏伟等。首都师范大学计算机教育硕士赵俊莉对全套教材进行了校阅,对可能存在的问题进行了精心的修饰。他们一丝不苟的工作精神和相互间的密切配合,特别值得赞扬。在此,对他们卓有成效的工作表示感谢!

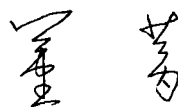
本套教材的翻译工作得到了教育部高教司、教育部职成司、北京航空航天大学、中国软件行业协会教育与培训委员会、北京软件产业促进中心等单位领导和相关人士的指导,同时还得到了 Comp-U-Learn Tech 印度有限公司亚洲首席代表 Vikram Gupta 先生的大力配合和支持,在此一同对他们表示真诚的谢意!我要特别感谢中国社会科学院王树英教授、教育部教育发展研究中心培训中心朱永庆教授对引进这套系列教材所付出的辛苦。感谢北京大学计算机工程硕士许珂和太原理工大学高级工程师郑晋梅在课程设置工作中做出的贡献。感谢中国地质大学陈明教授、北京航空航天大学王人骅教授、刘超教授、首都经济贸易大学张学群教授、中国人民大学杨小平教授对编译工作的指导,感谢张璠、邱钦伦两位硕士真诚的关心和支持。这里我

们还要特别感谢高等教育出版社为这套系列教材的出版给予的大力支持。

在编译过程中，我们虽然力求准确，尽量保持原文的特色，但限于时间仓促，难免存在失当，恳请各位同仁和读者不吝赐教。

希望每位学员能从本套教材中汲取丰富的营养，预祝每位学员能尽快加入令人羡慕的 IT 行业中！一旦进入了这个五彩缤纷的世界，将会感谢本套系列教材给你带来的一切。

同学们，这里是信息化高速公路起跑线，这里是人生旅途的转折点，让我们共同点击“开始”吧！



北京华育发国际技术培训中心

二〇〇五年元月

I 开始之前

万维网正在经历剧烈的变化，这种变化不仅为用户引入了许多新颖的服务，而且还为 Web 站点开发者和企业提供了更多的机会。HTML 使 Web 变成世界级的图书馆。现在，它的同胞 XML，即可扩展标记语言（eXtensible Markup Language），正在使 Web 成为全世界商业和金融业的枢纽。XML 已经成为 W3C 的推荐标准，并且每天有数以百万计的 XML 文档在线传输。

今天，XML 作为 W3C 的推荐标准，就意味着它已经通过了 W3C 成员的审核，并且得到一致认可。因此，XML 被认为是“稳定的”，并且将被广泛应用。推荐标准是 W3C 组织认可的最高级别的文档。XML 是在 1998 年 10 月成为推荐标准的。

XML 是一种文档标记语言，它包含了文档的结构化信息。所谓结构化信息，既包含了信息的内容（例如，文字、图片等），也包含了用于说明信息内容在文档中角色的标识符（例如，页眉区域的内容与页脚区域的内容是不同的，也就是说结构化信息既不同于图形注解，也不同于一般的数据库表）。几乎所有的文档都具有某种结构。

II 如何使用本书

本书适合于熟悉 Java 语言编程并且具有 HTML 实际操作经验的读者，同时也要求读者熟悉 Windows 98/NT/2000 系列操作系统。只要熟悉 Java 和 HTML 的基本知识，就可以非常容易地理解本书所讲述的内容。为了帮助读者理解每章的内容，在必要的地方都有大量的例子。

对于熟悉 HTML、Windows 环境和 Java 语言的读者来说，将能够很容易地掌握书中的内容。本书条理清晰，易于理解，并且书中所有的例子都非常容易实现。

III 本书编写目标

编写本书主要有以下目的：

- ◆ 使学生熟悉 XML 语言。
- ◆ 使学生能够配合 DTD、层叠样式表（Cascading Style Sheet）和 XSL 创建 XML 文档。
- ◆ 使学生能够将 Java 和 XML 结合使用，并且能够使用 SAX 和 DOM 解析器分析数据。
- ◆ 向学生介绍其他新技术，如 Schema、Namespace、Xlink、XPointer 和 XHTML。

IV 本书使用惯例

- ◆ 每章结尾部分有一个总结，用斜体字列在特殊的方框中。
- ◆ 每章以“复习题”结束。这将有助于学生巩固本章所学内容。
- ◆ 图片、图表和截屏图像下面都有编号和说明，有助于指导学习。
- ◆ 每章都有配套的案例练习（见《XML 核心技术指导手册》），提高读者的实际分析能力和动手能力。

目 录

I 开始之前	I
II 如何使用本书	I
III 本书编写目标	II
IV 本书使用惯例	II

第一章 XML 简介	1
1.1 标记语言的发展历史	3
1.2 超文本标记语言的局限性	6
1.3 进入 XML	7
1.4 XML 文档	11
1.5 XML 编辑器和处理器	16
1.6 附加参考	27
总结	27
复习题	29

第二章 XML 文档与 DTD	31
2.1 创建 XML 文档	33
2.2 创建 DTD	40
总结	64
复习题	65

第三章 使用样式表显示 XML	67
3.1 样式表简介	69
3.2 将样式表附加到 XML 文档	70
3.3 进一步了解层叠样式表	71
3.4 使用选择符	72
3.5 块级元素、行内元素和不可见元素	78
3.6 STYLE 属性	81

3.7 继承	82
3.8 在样式表中包含注释	83
3.9 长度值	83
3.10 字体属性	86
3.11 颜色和背景属性	91
3.12 文本属性	100
3.13 方框属性	109
3.14 定位属性	114
总结	116
复习题	117

第四章 XSL 转换	119
4.1 XSL 的定义	121
4.2 XSL 模板	123
4.3 输出	134
总结	144
复习题	145

第五章 XML 的 API (一)	147
5.1 整合 Java 和 XML 的必要性	149
5.2 XML、Java、数据库和 Web	153
5.3 解析 XML	159
5.4 API 总览	161
5.5 SAX	162
总结	181
复习题	181

第六章 XML 的 API (二)	183
6.1 文档对象模型 (DOM)	185

XML 核心技术

6.2 DOM 核心接口概览	191
6.3 DOM HTML 接口简介	192
6.4 DOM API 的定义	194
6.5 DOM 和 Java	200
6.6 使用 DOM API 进行树操作	202
6.7 通过 Servlet 访问 XML 文档	216
6.8 DOM 和 SAX 之间的关系	225
总结	226
复习题	227

第七章 模式、命名空间

与 XHTML	229
7.1 模式	231
7.2 命名空间	242
7.3 XHTML	251
总结	261

复习题	262
-----------	-----

第八章 XLink 与 XPointer

8.1 简介	265
8.2 XLink	265
8.3 XPointer	270
8.4 应用程序示例	275
总结	333
复习题	334

附录

附录 A: XML、Java 和 JavaScript	335
附录 B: 所有包的层次结构	344
附录 C: COMP-U-LEARN 证书样本	358

英中文对照词汇表

第一章

XML 简介

完成本章学习，你将能了解：

-  标记语言的发展历史
-  XML 的优点
-  格式规范的和有效的 XML 文档
-  了解 XML 编辑器和处理器

本章内容

- 标记语言的发展历史
- 超文本标记语言的局限性
- 进入 XML
- XML 文档
- XML 编辑器和处理器
- 附加参考

1.1 标记语言的发展历史

在瞬息万变的今天，没有事物是永恒的，相反的是，所有事物的存在都是十分短暂的。创新总是在迫不得已的时候出现，并且人们总希望通过不断地创造新的工具和事物来消除日常生活中所遇到的障碍。以知识为基础的 IT 业正伴随着这股变化潮流来改变人们每天的生活。

作为一种标记语言，尽管 HTML (HyperText Markup Language, 超文本标记语言) 的应用十分普遍和流行，但它实际上受到许多因素的制约。HTML 非常适用于传播非正式的文档，但在 HTML 中不能将数据表示格式从具有确切含义的上下文数据中分离出来。正是在这一点上，XML (eXtensible Markup Language, 可扩展标记语言) 超越了 HTML，并且赢得了世界范围内的认可。本章将对 XML 的发展历史及其超越 HTML 的优点进行回顾，最后将介绍 XML——这种功能强大的文档创建工具——的编辑器和处理器。

1.1.1 通用标记语言的演变

在 20 世纪 60 年代，IBM 的 3 名研究人员 Charles Goldfarb、Ed Mosher 和 Ray Lorie 开始深入研究一个普遍存在于各种法律文件中的问题，文件是由不同的系统按照各自不同的格式生成的，由于文件格式各异，系统的可移植性实现起来就非常困难。由这 3 位杰出研究人员所进行的研究得出的结论是：要提高系统的可移植性，必须采用一种通用的文档格式，这种文档格式化必须遵循特定的规则。这样，创建 GML 的指导原则就建立起来了，这些原则在标记语言的历史上是第一次提出。

1.1.2 标准通用标记语言

任何一个发明者对自己发明的作品都有一种偏爱。和其他的发明者一样，Charles Goldfarb 带领着一个开发团队继续完善他的作品。最后，他们就这种标记语言的概念达成了共识，把它称为 SGML (Standard Generalized Markup Language, 标准通用标记语言)。SGML 最终于 1986 年被国际标准化组织 (ISO) 所采纳。

顺便提供一些附加信息，ISO 是一个由世界上 130 个国家的标准化机构组成的国际性联盟。这个组织建立于 1947 年，致力于自然科学和学术领域的协作，促进全球的商业和信息交流。

SGML 是一个非常强大的工具，它为描述网络电子文档提供了一套必要的通用框架，并且允许建立新的一致性规则。但为什么 SGML 没有被广泛使用呢？主要是因为它非常复杂。可以用以下 3 个突出特点来描述 SGML：

- ◆ 不提倡使用特定的文档结构。
- ◆ 不阻碍建立新的文档标准。
- ◆ 与 HTML 不同，SGML 没有必须使用的限定性标记集。

由 SGML 所提供的功能已经被美国的一些大型企业所采用，这些企业包括汽车公司和电信

公司。SGML 同时也作为一种文档标准而被美国国防部和美国国内税务局所采用。

SGML 本质上是一种“元语言”，可以用它来建立其他的语言。例如，空运协会使用 SGML 创建了一套标记集来维护飞行器文档。那么，确切地说什么是“元语言”呢？看一个例子读者便会很透彻地了解它。

首先用 HTML 写一个使人产生食欲的巧克力蛋糕配方。有一点要注意，不要试着自己去做它。

列表 1.1 一些 HTML 代码

```
<HTML>
  <HEAD>
    <TITLE>MY RECIPE FOR A CAKE</TITLE>
  </HEAD>
  <BODY>
    <marquee>TO MAKE A YUMMMY CHOCOLATE CAKE YOU'LL NEED</marquee>
    <P>COCOA - 2 tsp <BR>
    EGGS - 2 <BR>
    FLOUR - 1 small bowl <BR>
    SUGAR - 1/2 small bowl </P>
  </BODY>
</HTML>
```

在浏览器（IE）中可以看到以下的内容，如图 1.1 所示。

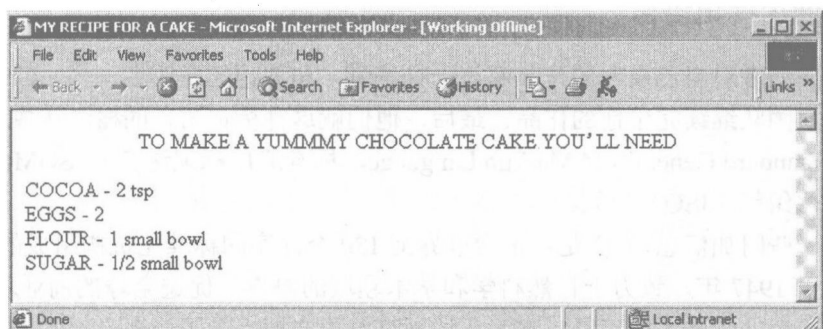


图 1.1 HTML 代码的输出结果

仔细地观察一下 HTML 代码。它非常容易编写，但是对于一个外行来说，几乎不可能解释出其中一些标记如<marquee>的意思。顺便提一下，这个标记的意思是使其中的文字按照水平方向在屏幕上滚动。更进一步看，蛋糕的配料包含在标记<P>当中。由于这是一个很小的文档，它仅仅是想让一个不懂 HTML 的人浏览一下代码就能明白文档的内容。但是设想一下，当配料

表有 40 个条目，并且包含要用到的数量和烤制过程中应该遵循的步骤时，会发生怎样的情况呢？是否有可能存在这样一个人（即使他懂 HTML），他能够理解文档中每一条数据信息的含义？更糟糕的是，有些人可能会对<TITLE>和<marquee>标记中的内容置之不理。在这种情况下，仅看配料，也就是<P>标记中的内容，可以让一个优秀的厨师辨别出文档中所说的是巧克力蛋糕配料，但有些人可能会以为文档中提及的是采购单。

想像一下，读者可以创建自己定义的语言，使用它来描述菜谱，这样将获得在菜谱中扩展信息结构的自由，即信息之间是如何彼此联系的。这就是元语言带给人们的自由——创建自定义标记的自由。

现在用自定义标记来写菜谱，以增强可读性。

用记事本程序建立下面这个文档并且将其保存为 createcake.xml，然后可以用 IE 来浏览。

列表 1.2 用自定义标记写代码

```
<?xml version = "1.0"?>
<Recipeforachocolatecake>
  <Ingredients>
    <Ingredient>
      <Item>Cocoa</Item>
      <Qty>2 tsp</Qty>
    </Ingredient>
    <Ingredient>
      <Item>Eggs</Item>
      <Qty>2</Qty>
    </Ingredient>
    <Ingredient>
      <Item>Flour</Item>
      <Qty>1 small bowl</Qty>
    </Ingredient>
    <Ingredient>
      <Item>Sugar</Item>
      <Qty>1/2 small bowl</Qty>
    </Ingredient>
  </Ingredients>
</Recipeforachocolatecake>
```

能够自定义标记还有许多其他优点，后面讨论到 XML 的这些优点时将会特别指出。下面先介绍 HTML 的局限性。

1.2 超文本标记语言的局限性

由于 HTML 不是一种元语言，所以它不能创建某一特定领域的标记集。而对于类似 SGML 这样的元语言来说（当然，XML 是 SGML 的一个子集），可以将标记的表示符号由尖括号（<tag>）变为花括号（{tag}）。同样也可以将标记名称的长度由 8 个字符改变为 20 个字符。作为 SGML 的一个子集，XML 是 SGML 的一种简洁的、与 Web 兼容的版本。二者的差异都是技术层次的，例如，对标记的生成和缩写的处理风格、注释结构以及其他一些定界符，本章仅提及其中的一部分。

标记语言的诞生使建立需要存储大量电子文档的数据仓库变得更加方便。通常，这样的一个电子文档由 3 个部分组成。

- ◆ 文档结构。提供如何书写文档的基本框架。
- ◆ 文档内容。
- ◆ 文档格式。指定文档显示时的排列样式。

超文本标记语言可以指定一个文档的内容和格式，但不能指定文档的结构。回想一下，在列表 1.1 中可以识别出文档的内容。格式也可以通过使用<marquee>、<P>和
标记来指定，但是关于结构呢？全部的内容都封装在<P>标记中，因此文档中不存在任何的框架或者构架。

再看一下列表 1.2，能看出差别吗？当然可以。该列表用结构化的方法来显示文档内容，但这里没有指定具体的格式。与此同时，如果理解了列表 1.2，那么就已经理解了 XML 的基本知识。列表 1.2 中的代码实际就是一段 XML 代码。简而言之，XML 指定了文档的内容和结构，但没有指定文档的格式。这就是它超越 HTML 的一个优势。由于没有指定格式，人们可以采用不同的方式来显示相同的信息，而不像 HTML 那样需要在文档中使用<P>、
和其他的格式标记来指定格式。

HTML 实际上是一种用于描述文档的语言，包含标题、名称和说明等。事实上，它明显是面向表示的。下面是 HTML 的主要不足：

- ◆ HTML 不具有可扩展性。

可扩展性标记语言允许应用程序开发者为具体的应用环境定义自定义标记。而在 HTML 的环境中，程序开发者被浏览器开发者和 W3C 组织定义的标记集所限制，用户期望有一种语言允许创建自定义的标记，而不被某种特定的标记所限制。

- ◆ HTML 是以显示为中心的，而不是面向结构的。

HTML 确实为网页设计者显示数据提供了很大的灵活性。它可以使文本以加粗、斜体显示，并且允许设置图片的各种排列特性。但它几乎没有语意结构。它显示数据是按照布局而不是语意。然而，当通过搜索引擎来进行数据搜索时，按照语意而不是布局来显示数据具有很多优点。