HZ BOOKS
华章教育

高等院校计算机教材系列

PRINCIPLES AND APPLICATIONS OF XML

XML原理与应用

夏天 编著

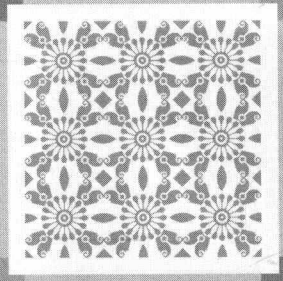
机械工业出版社
China Machine Press

高等院校计算机教材系列

Principles and Applications of XML

XML原理与应用

夏天 编著



机械工业出版社
China Machine Press

图书在版编目（CIP）数据

XML 原理与应用 / 夏天编著 . —北京：机械工业出版社，2015.3
（高等院校计算机教材系列）

ISBN 978-7-111-49378-5

Ⅰ. X… Ⅱ. 夏… Ⅲ. 可扩充语言 – 程序设计 – 高等学校 – 教材 Ⅳ. TP312

中国版本图书馆 CIP 数据核字（2015）第 031625 号

本书全面系统地介绍了 XML 的基本原理和关联技术，注重 XML 技术体系的整体性和连续性，便于读者快速把握各组成技术在 XML 中所起的主要作用。

全书共 10 章，介绍了 XML 的基础知识、XML 的验证处理、XML 的呈现与转换、XML 的编程接口和应用等。各章最后均提供了习题，便于读者巩固和拓展所学知识。

本书不仅可供信息管理与信息系统、计算机专业的本科生 / 专科生使用，也可供信息技术领域的研究生和工程技术人员参考使用。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号，邮政编码：100037）
责任编辑：李　燕
印　　刷：三河市宏图印务有限公司　　版　　次：2015 年 3 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：14.25
书　　号：ISBN 978-7-111-49378-5　　定　　价：35.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：(010) 88378991　88361066　　投稿热线：(010) 88379604
购书热线：(010) 68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

前　言

XML（eXtensible Markup Language，可扩展标记语言）是由 W3C 于 1998 年 2 月发布的 Web 标准之一。经过十余年的发展与完善，XML 相关技术已经对当今社会的技术发展产生了深刻的影响，从网页编写的标准化到 WebService，从异构系统之间的数据交换到行业标准的制定，从开放办公文档的表示到矢量图形的表示，都能发现 XML 在其中的应用。不仅程序开发人员应该掌握 XML 技术的基本原理和组成体系，系统架构人员、管理人员同样需要对 XML 有深入的理解。

本书从 XML 技术体系本身出发，结合具体实例、历史背景和解决问题的思路，逐一介绍 XML 最为重要的核心技术：DTD、Schema、CSS、XPath、XSLT、DOM，同时结合技术的发展变化，介绍 XML 在可缩放矢量图形方面的应用 SVG 和 XML 在数据表示领域的竞争技术 JSON。XML 作为信息技术的核心标准之一，是发展新的 Web 应用、制定新的信息规范、描述新的行业应用必不可少的基础理论，是 Web 标准体系的核心组成部分。

本书在写作过程中参考了大量国内外相关资料，并结合作者多年的应用开发实践和教学经验，重新设计了内容体系和实例，所有章节均由作者独自完成，保证了内容的原创性和连贯性。

在写作过程中，作者力图加强对基本概念、基本理论和基本技术的介绍，既强调 XML 技术体系的整体性，也注重各个部分的相对独立性，以适合教学使用。本书的特色如下：

- **问题驱动展开讲解**。对 XML 技术体系中的每一种特定技术，首先介绍引入该技术的基本目的，再进一步讲解其详细技术，力求使读者能了解该技术的来龙去脉。
- **适用面广**。对 XML 技术的讲解尽量不涉及特定的操作系统、开发商或实现方式，需要借助程序语言讲解的部分采用易于获取和测试的 JavaScript 语言，适用于所有对 XML 技术感兴趣的人群。
- **实例丰富，适合教学**。教材体系完整，大量的实例和习题有助于读者对 XML 技术的理解和学习。
- **注重技术变化**。结合新流行的 Web 技术，介绍与 XML 相关的 SVG 和 JSON 技术。

本书可作为高等院校电子政务、电子商务、信息管理与信息系统、计算机应用等专业本科生的教材，也可作为相关标准制定人员、信息技术管理人员、应用系统设计和开发人员的参考书。

本书附带的电子教案和实例源代码均可从华章网站（www.hzbook.com）免费下载，与本书有关的问题和建议也可以直接通过 https://github.com/iamxiatian/xml/ 网站在线反馈。同时，由于 XML 的相关标准和应用还在不断发展，我们将紧跟变化，对本书内容不断优化和完善，而每一位读者的反馈都将是我们前进的最大动力。

作者

2015 年 1 月于中国人民大学

教学建议

教学章节	考核要求	课　时
第 1 章 XML 引论	• 了解标记语言及其发展历史； • 理解 XML 的目标与特点； • 掌握 XML 的整体技术体系； • 了解 XML 的常见应用场景； • 掌握 XML 的编辑器、解析器和浏览器的特点，学会基本的 XML 编辑器的安装。	4
第 2 章 XML 的基本语法	• 掌握 XML 的文档组成结构； • 掌握 XML 声明、处理指令、注释、元素、属性、命名空间等基本概念； • 能够编写格式良好的 XML 文档。	4 ~ 6
第 3 章 文档类型定义 DTD	• 了解 DTD 的作用； • 掌握 DTD 的语法规则和使用方法； • 能够利用 XML 编辑工具实现 DTD 的编辑和验证。	6
第 4 章 XML Schema	• 理解 XML Schema 的含义及用途； • 掌握 XML Schema 的元素、属性的作用及使用方式； • 掌握 XML Schema 的数据类型； • 理解 XML Schema 的命名空间概念。	6 ~ 8
第 5 章 利用 CSS 格式化 XML	• 掌握 CSS 与 XML、HTML 的关联方式； • 掌握 CSS 的常见选择器，理解 CSS 的属性继承与覆盖原则； • 掌握 CSS 的颜色、字体和文本属性以及 CSS 的盒状模型。	4 ~ 6
第 6 章 XML 路径语言 XPath	• 了解 XPath 的工作原理； • 掌握 XPath 的定位路径表达式； • 熟悉 XPath 的常用函数和数据类型； • 掌握 XPath 2.0 的部分新特性。	6
第 7 章 可扩展样式语言转换 XSLT	• 了解 XSLT 的发展历史、XSLT 与 XSL 的关系； • 掌握 XSLT 的转换处理过程； • 掌握 XSLT 的基本语法； • 能够使用 Saxon 进行 XSLT 测试。	6 ~ 8
第 8 章 JavaScript（选读）	• 了解 JavaScript 的起源； • 掌握 JavaScript 的测试方法； • 熟悉 JavaScript 的基本语法，能够借助于文档进行基本的脚本编写； • 了解浏览器对象模型 BOM； • 理解 JavaScript 的定时器操作。	6 ~ 8

（续）

教学章节	考核要求	课　时
第 9 章 文档对象模型 DOM	• 了解 DOM 的特点及规范级别； • 掌握常见的 DOM 对象； • 熟悉浏览器操纵 HTML 文档和 XML 文档的基本方法。	6 ~ 8
第 10 章 XML 的应用与挑战	• 了解 SVG 的特点； • 能够基于 SVG 进行简单的图形表示； • 了解 d3.js 的功能特点； • 掌握 JSON 的数据结构和类型； • 能够通过 JavaScript 解析 JSON 字符串。	6 ~ 8

说明：

1）建议总课时为 54~68。

2）建议课堂教学全部在多媒体机房内完成，实现“讲 - 练”结合。

3）本书配有实验指导书，可通过访问 http://www.github.com/iamxiatian/xml/ 获取，任课教师在讲授完每一章的基本理论后，可以安排学生练习实验指导书中的内容，并根据学生的练习情况对重点内容做进一步讲解。

4）各章最后都附有习题，任课教师可根据情况给学生留一些基本和中等难度的习题作为课外作业。

目　录

第 1 章　XML 引论

本章主要讨论与 XML 相关的三个重要问题：①为什么会出现 XML，它的目标是什么？② XML 的技术体系是如何构成的？③ XML 有哪些典型应用场景？在此，我们只是对 XML 的体系结构做一个整体描述，其详细内容将在后续章节中依次展开讲解。

1.1　XML 的起源

XML（eXtensible Markup Language），即可扩展标记语言，自从 1998 年 2 月成为 W3C 的推荐标准之后，便得到了人们的广泛关注。各个软件厂商纷纷宣布或者已经在自己的产品中增加对 XML 的支持，例如，微软公司在其新一代的办公软件中，采用 XML 格式代替原先的二进制格式对文档进行描述；Oracle 公司则将 XML 数据模型完全集成到其数据库中，并提供浏览和查询 XML 的新的标准访问方法，通过 Oracle XML DB，人们可以获得相关数据库技术的所有优势和 XML 的优势；原 Sun 公司更是在其开发语言 Java 当中内置了对 XML 的支持；与此同时，W3C 也在积极地致力于完善 XML 的标准体系。

近几年来，众多与 XML 相关的新名词、新技术如潮水般不断涌现，WebService、XHTML、手机 WAP、RSS、开放式电子公文格式、XML 数据交互接口，无一不与 XML 密切相关。即便是不参加具体开发的管理人员，甚至是与软件过程无直接关联的人员，也开始认识到 XML 的重要性，开始想方设法获取对 XML 的一个完整的认识。那么，到底什么是 XML，它从何方来，要往何方去？要弄清这个问题，我们就不得不回顾标记语言以及它的发展历史。

1.1.1　标记简介

标记语言（Markup Language）起源于传统印刷。文字在印刷之前必须先进行排版，在电子出版业出现以前，人们需要以手工或打字方式复制手稿副本，并对副本以人工方式进行标记，通过这些标记说明告诉排版印刷人员如何处理版面，如印刷选用的字体、行间距、对齐方式等格式信息，最终完成印刷前的制版工作。这些人工增加的注释就被称为“标记”（Mark-up）。

类似的，在计算机中，人们也需要对输入的文字指定其外观样式。例如，用户可以对写好的文本选择字体，设置字符间距、段落等格式信息，这些信息被称为电子标记（电子标记是“ Markup”，以区别于传统的手写标记“ Mark-up”），它们以特殊代码形式存储在文本中。计算机处理程序在读取文本的同时还要读出以电子形式标注的格式信息，最终告知计算机安装规定的格式展现文本，例如，大家熟悉的 Word、WPS、OpenOffice 等文本编辑器都是借助于标注代码来定义格式与外观的。与直接手工标注格式不同，在计算机中，用户通常通过在菜单中选择命令的方式来向文本中添加格式指令，进行标注。

通俗地说，标记语言就是一种用来给文本添加“格式标注”以指明文档中文本编排格式的语言，一般由定义文档格式的一些规定代码和控制标记组成。

1.1.2 过程标记

RTF（Rich Text Format，富文本格式）就是这样一种简单的过程标记语言，它是由微软公司于 1987 年开始开发的一种跨平台文档格式，大多数的文字处理软件都能读取和保存 RTF 文档。作为微软公司的标准文件，在早期人们需要向微软付费才能购买一本薄薄的 RTF 标准文件。后来，随着 RTF 的普及，微软公司把标准文件公开，放在网上供开发者下载，无需支付任何费用，目前 RTF 的最新版本是 RTF1.9.1[⊖]。下面，我们就通过一个示例来了解一下 RTF 文档的过程标记。

【例 1-1】创建、查看 RTF 文档的标记。

1）打开 Word 软件，输入“XML and RTF!”，选择字体为 Times New Roman，字号为 28，并把“XML”的颜色设置为红色，如图 1-1 所示。

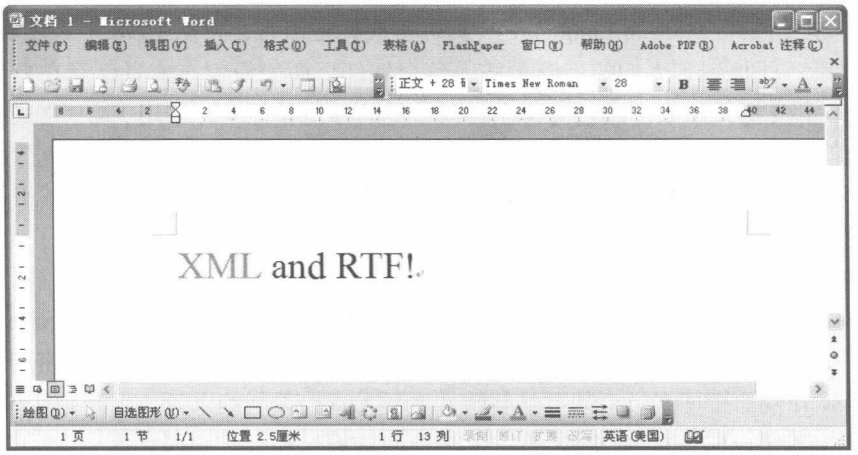


图 1-1 在 Word 中输入一段文字

2）单击“保存”按钮，在弹出的“另存为”对话框中输入文件名“1-1.rtf”，注意必须把保存类型选择为“RTF 格式 (*.rtf)”，如图 1-2 所示，然后单击“保存”按钮。

图 1-2 保存为 RTF 文档

⊖ http://www.microsoft.com/downloads/details.aspx?FamilyId=DD422B8D-FF06-4207-B476-6B5396A18A2B&displaylang=en。

3）在记事本中打开刚刚建立的“1-1.rtf”文档，如图 1-3 所示是加上标记后的文本格式。

图 1-3 用记事本打开“1-1.rtf”

从图 1-3 中可以看出，其中的内容全部是文本字符，并且比开始输入的“XML and RTF!”要多很多，这些多出来的字符正是用于描述字符串“XML and RTF!”的格式的标记信息。RTF 格式能够得到大多数文字编辑软件的支持，与它的优点是分不开的。首先，RTF 是一种得到广泛支持的文件格式，绝大多数文字处理程序都能够创建和读取 RTF 文档，如 Word、WPS 等；其次，RTF 可以存储复杂的格式信息，如图片、表格、超链接等。然而，RTF 格式的缺点也比较明显，如：

1）通过 RTF 格式保存的文件通常都比较大，RTF 中对相同文本段落会有不同版本的格式标注，另外 RTF 本质上是一种文本格式，对于内嵌的图片，也只能用文本的方式对图片进行等价表示，于是导致文件体积变大。

2）RTF 格式不能保存一些特殊的信息，例如，对图片文件只能通过转换成位图的方式进行保存。我们在把一个复杂的 Word 文档转换为 RTF 时，经常会发生一些信息丢失的情况，正是这一原因造成的。

3）RTF 格式不够灵活，对格式规则的任何变更意味着需要对文档进行手工修改。还有，标记的多少是和系统有关的，这就增加了移植的困难性。

1.1.3 通用编码

1967 年，William W. Tunnicliffe 在加拿大政府印刷局的一次会议上提议出版商应该采用通用编码 (Generic Markup) 对文本进行编码，标识文本中不同部分的功能，例如哪段文字是标题，哪段文字属于引用等，而不是仅为最终的打印样式进行编码控制。相对于设置文本出版格式的传统标记，通用编码更侧重于文本不同部分的功能标记，力图实现文本的文字信息和呈现格式的分离。根据这一思想，以 Tunnicliffe 领导的小组在 1970 年开发出了出版业的通用编码标准“GenCode”，成为标准通用标记语言 SGML 的基础。

由著名计算机科学家 Donald E. Knuth 在 20 世纪 70 年代发明的排版系统 Tex，是另一种较有代表性的通用编码。TeX 引入了宏的概念，可以生成高质量的 DVI 格式文件并打印输出，还可以利用 dvips、dvipdfmx、pdflatex 等程序生成 PDF 或 PS 格式文件，也可以利用 latexhtml 生成 HTML 网页文件。同时，TeX 能够处理非常复杂的数学公式，排版出精美的打印效果，也可以借助于宏程序包制作幻灯片、定义模板等，因此，它在数学、物理学和计算

机科学等学术界十分流行。

【例 1-2】TeX 示例清单 Test.tex。

```
\noindent Xiatian\par
\noindent School of Information Resource Management\par
\noindent Renmin University of China\par
\smallskip

This is a book about XML. If you have some problems, contact us please.\par
\bye
```

在上述清单中，\noindent、\par、\bye 等都是一些具有意义的标识符，例如 \par 代表的是段落标记，可见，TeX 采用了中性标记来标识文档的逻辑结构。

相对于 RTF 这类过程标记语言，通用编码具有如下优点：

1）通用编码具有更好的可移植性，灵活性高。为改变文档的外观，人们可以不修改文档本身，而只对宏的定义进行修改即可实现，编辑一个宏，其变化会自动应用于整篇文档。特别是我们不需要对标记进行重新编码，避免了这个既费时间又容易出错的过程。

2）标记本身更倾向于描述文档的逻辑结构。例如，通过自定义 \title 宏，声明这是一篇文章的标题，\abstract 宏表示文档的摘要等。对标题和摘要的格式信息则交给宏的具体定义予以描述。用户在编写自定义的宏时，一般都会为标记起一个比较有意义的名字，如 title、abstract 等，清楚地体现了结构对于格式的优越性。

1.1.4 SGML

SGML（Standard Generalized Markup Language），即标准通用标记语言，是一种定义电子文档结构和描述其内容的国际标准语言，早在 Web 发明之前 SGML 就已存在。

注意：SGML 中的 G、M 和 L 分别代表 SGML 的三位创始人：Goldfarb、Mosher 和 Lorie。

20 世纪 60 年代，在 IBM 计算机之间存在大量格式不同的电子文件，为了方便数据交换，IBM 公司的研究人员经过研究认为：要提高系统的移植性，必须采用一种通用的文档格式，这种文档的格式必须遵守特定的规则，并对文档的内容和样式实现分离。1969 年，IBM 研究人员开发出了通用标记语言（Generalized Markup Language，GML），GML 是一种自参考的语言，可以描述任何其他语言的语法和词汇，能够标识出任何数据集合的结构。在标记语言的概念基本达成共识的基础上，IBM 研究人员 Charles Goldfarb 带领的开发团队完善了 GML，并将其称为 SGML，SGML 成为了 IBM 内部格式化和维护合法化文件的手段。通过对 SGML 不断地拓展和修改，SGML 在 1986 年被国际标准化组织（ISO）采纳成为正式标准。

与通用编码不同，SGML 不仅能够对文档本身进行标注，还能够描述文档的层次结构信息。SGML 实际上是一个通用的文档结构描述符号化语言，用来定义文档模型的逻辑和物理类结构。一个 SGML 文件由三部分组成，即语法定义、文档类型定义（Document Type Definition，DTD）和文档实例。语法定义部分定义了文档类型定义和文档实例的语法结构；文档类型定义部分定义了文档实例的结构和组成结构的元素类型；文档实例是 SGML 的文档

主体部分。

在很多大型组织中 SGML 被用来创建、处理和发布大量的文本信息，得到了广泛应用。虽然 SGML 并没有给文档强加一种结构，但标准委员会、企业联盟及其他组织却开发了基于 SGML 的结构应用，它们为特定的应用程序定义了一些标准的 DTD。一些著名的应用包括：

1）HTML。用于描述 Web 文档的著名的超文本标记语言，尽管很少有 HTML 的开发人员知道 SGML，但 HTML 确实被定义为一种 SGML 的 DTD。

2）DocBook[⊖]。由结构化信息标准促进组织 OASIS 开发，用于技术文章和书籍的结构化表示。

1.1.5 HTML

毫无疑问，HTML 是 SGML 最流行的一种应用。这就是说，HTML 是遵守 SGML 规则的一组标记集合，这组标记适应于对互联网文档的结构描述。最初的 HTML 是于 1989 年由欧洲量子实验室的研究人员 Tim Berners Lee 在 SGML 的基础上开发的一个简化子集。HTML 继承了 SGML 许多重要的特点，比如结构化、独立于具体实现、可描述性，但是 HTML 也存在很多缺陷。HTML 只能使用固定的有限的标记，最初设计标记的重点在于如何表示网页的展示样式，随着网络的普及，人们开始对简单的 HTML 有所不满，因此，又开始了对 HTML 的改造过程。HTML 从最开始的 1.0 版本一直到最近发布的 HTML 5.0，扩展了许多命令，同时抛弃了一些用于单纯描述显示格式的标记，使得 HTML 得到了更为广泛的应用。

HTML 本质上是通过标记的方式，以纯文本形式对网页进行描述，而对标记的解释则由浏览器执行，如现在流行的 IE（Internet Explorer）浏览器、FireFox 浏览器、Opera 浏览器等。浏览器读取网页源代码，即 HTML 标记文本，并通过渲染呈现给用户，就形成了用户最终看到的网页。

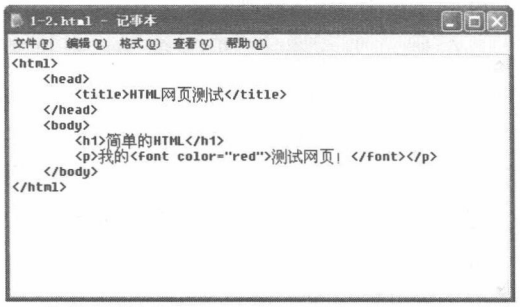


图 1-4 用记事本编辑 HTML 代码

【例 1-3】HTML 编写方式及显示效果。

1）在记事本中输入一段 HTML 代码，如图 1-4 所示，其中通过尖括号包括起来的就是用于网页描述的 HTML 标记。

2）选择“文件”菜单中的“保存”菜单项，在“另存为”对话框中，选择保存位置和保存文件名称，此处将文件命名为“1-2.html”，保存类型选择为“所有文件”，如图 1-5 所示。

注意：保存文件时需将“保存类型”选择为“所有文件”，且文件名必须包含“.html”。正常保存完毕，生成的新文件的图标类型为网页类型，如果是文本图标，一般情况下都表示刚才保存的是文本类型。

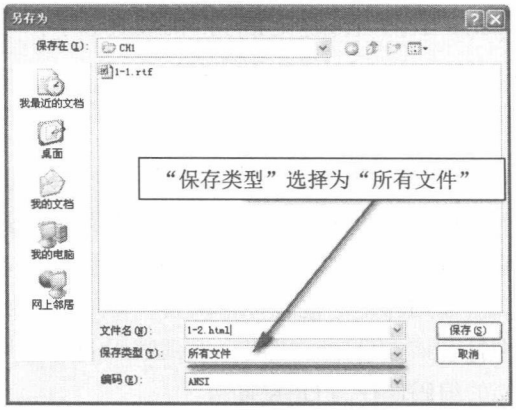


图 1-5 保存 HTML 文档

⊖ http://www.docbook.org。

需要补充说明的是，Windows 系统中默认会隐藏已知文件类型的扩展名，比如创建了一个“1-2.html.txt”文件，本质上是一个文本文件，但由于文本文件是 Windows 中默认可以识别的文件类型，如果不更改 Windows 的默认选项，则在资源管理器当中看到的文件名称是“1-2.html”，为避免出现混淆，请去掉 Windows 中的“隐藏已知文件类型的扩展名”选项。Windows XP、Windows 7、Windows 8 等不同 Windows 版本的取消方式相同，下面以 Windows XP 为例来具体说明取消方式。

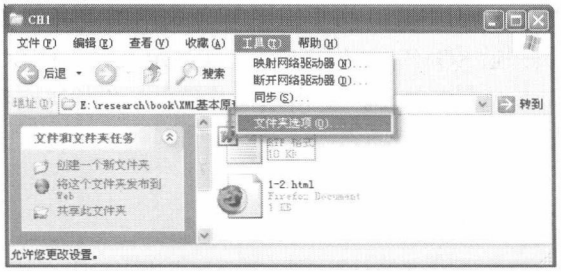


图 1-6 更改文件夹选项

如图 1-6 所示，打开资源管理器，进入任意一个目录，单击“工具”菜单中的“文件夹选项”命令，随即弹出如图 1-7 所示的窗口。

选择“查看”选项卡，取消“高级设置”区域内的“隐藏已知文件类型的扩展名”复选框，然后单击“确定”按钮即可。

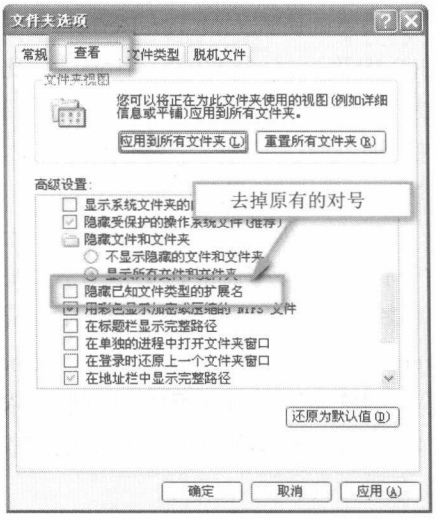


图 1-7 显示已知文件类型的扩展名

3）双击刚刚创建的“1-2.html”文件，默认会由浏览器打开该文件，并解析 HTML 标记，以网页的形式呈现最终的结果，在 FireFox 浏览器中的显示效果如图 1-8 所示。如果个人计算机上没有安装 FireFox 浏览器，也可以通过 IE 浏览器浏览。

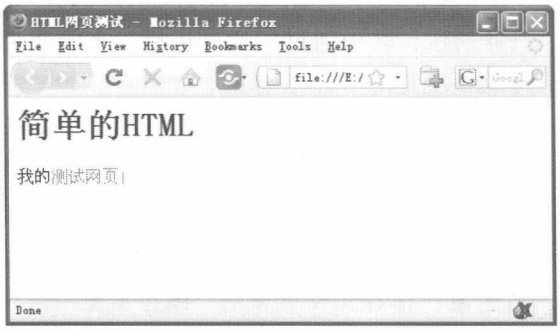


图 1-8 在 FireFox 浏览器中显示“1-2.html”文件

在浏览器中我们可以看到“测试网页”文字是红色的，对应于源代码中的标记即为“测试网页”，类似于这种标记就是一种过程标记。在 HTML 中，这类过程标记也可以通过通用标记予以实现，我们可以把“测试网页”更改为“测试网页”，然后再定义 highlight 样式如下：

```
.highlight {
     color:red;
}
```

这种描述样式在有关 CSS 的后续内容中（1.3.2 节）还会详细介绍。这个例子则表明现在的 HTML 是一种既支持过程标记又支持通用标记的混合标记语言。由于该示例文件没有明确指定网页的编码，读者在自己系统中打开该文件时，可能会遇到乱码问题，此时，只需要选择浏览器中的编码菜单，把当前文档的编码选择为 GBK 即可。