



信息管理
与信息系统
专业规划教材

XML技术教程

李禹生 孙平 张琳 贾瑜 编著

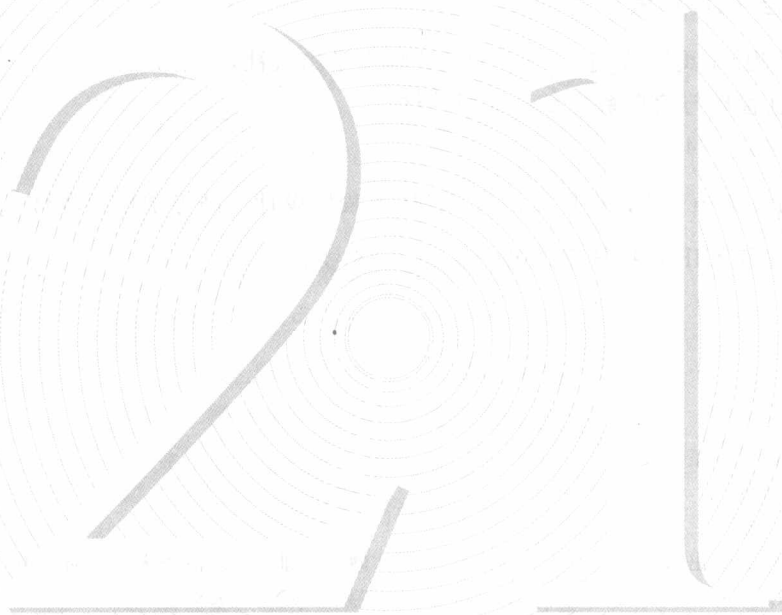


清华大学出版社

21世纪信息管理与信息系统专业规划教材

XML技术教程

李禹生 孙平 张琳 贾瑜 编著



清华大学出版社
北京

内 容 简 介

本书系统地介绍了扩展标记语言(XML)文法及其应用技术,其主要内容有:XML标记的核心概念、XML文档类型描述(DTD)及其XML模式(Schema)、XML文档显示样式(CSS/XSL)、XML应用标准及其XML文档编辑与浏览工具、XML链接(XLink/XPointer)与接口(DOM/SAX)技术、XML数据库与数据交换。在每一章节中,均采用相关实例深入讲解。最后,介绍XML技术的应用领域,包括资源描述框架(RDF)、频道定义格式(CDF)、基于XML的电子商务标准以及XML与跨媒体出版等应用领域。

本教材可作为信息管理专业“XML技术”及其相关课程的教材,也可作为XML技术应用工程师的技术参考书。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

XML技术教程/李禹生等编著. —北京:清华大学出版社,2009.5

(21世纪信息管理与信息系统专业规划教材)

ISBN 978-7-302-19440-8

I. X… II. 李… III. 可扩充语言,XML—程序设计—高等学校—教材 IV. TP312

中国版本图书馆CIP数据核字(2009)第015988号

责任编辑:闫红梅 王冰飞

责任校对:焦丽丽

责任印制:杨 艳

出版发行:清华大学出版社

地 址:北京清华大学学研大厦A座

<http://www.tup.com.cn>

邮 编:100084

社 总 机:010-62770175

邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:北京市清华园胶印厂

经 销:全国新华书店

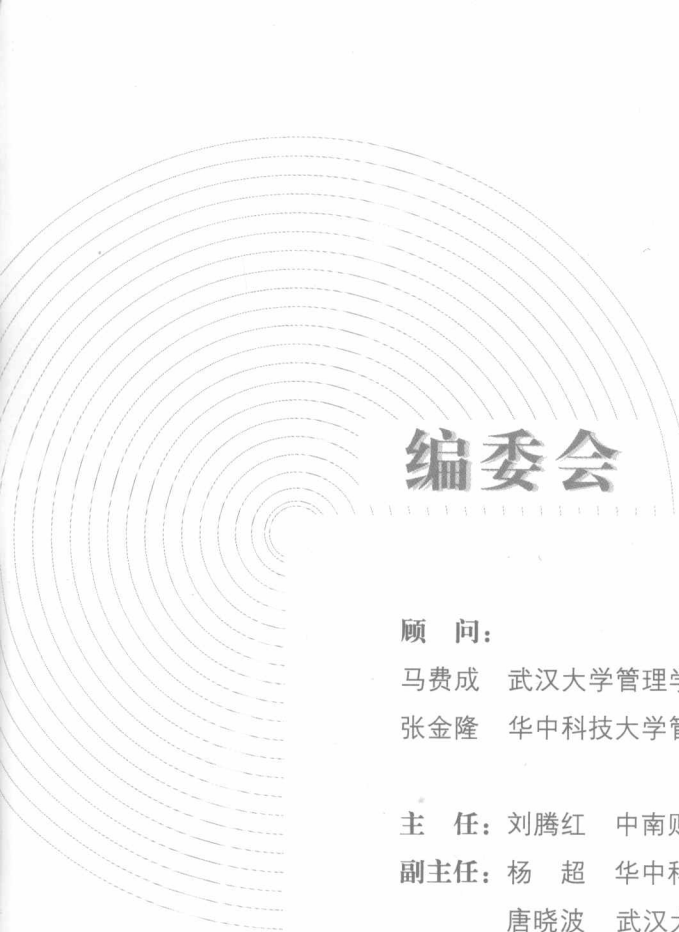
开 本:185×260 印 张:18 字 数:438千字

版 次:2009年5月第1版 印 次:2009年5月第1次印刷

印 数:1~3000

定 价:27.00元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题,请与清华大学出版社出版部联系调换。联系电话:(010)62770177 转 3103 产品编号:027001-01



编委会

顾问:

马费成 武汉大学管理学院

张金隆 华中科技大学管理学院

主任: 刘腾红 中南财经政法大学信息学院

副主任: 杨超 华中科技大学管理学院

唐晓波 武汉大学管理学院

编委: (按姓氏笔画为序)

马学才 华中农业大学信息管理系

王伟军 华中师范大学信息管理系

王虎 武汉理工大学管理学院

邓仲华 武汉大学管理学院

史金平 湖北大学商学院

孙细明 武汉工程大学管理学院

宋克振 中南财经政法大学信息学院

张志清 武汉科技大学管理学院

张新兰 中国地质大学管理学院

沈先钊 湖北工业大学信息管理系

李禹生 武汉工业学院计算机与信息工程系

郑双怡 中南民族大学管理学院

柯林 江汉大学商学院

鲁耀斌 华中科技大学管理学院



序

根据国家教育部 1998 年 7 月 6 日公布的《普通高等学校本科专业目录》的内容,将原经济信息管理、图书情报学、科技信息管理、林业信息管理和**管理信息系统**等专业合并为**管理学科门类中的信息管理与信息系统专业**。目前,我国已有二百多所高等院校设置了**信息管理与信息系统专业**。该专业的发展伴随着世界信息化的发展而发展,为我国培养了大量的**信息化专门人才**。

网络化、信息化、全球经济一体化是当今世界的主要特征。20 世纪 90 年代,信息技术不断创新,信息产业持续发展,信息网络广泛普及,信息化成为全球经济社会发展的显著特征,并逐步向一场全方位的社会变革演变。21 世纪,信息化对经济社会发展的影响更加深刻,信息资源日益成为重要的生产要素、无形资产和社会财富。我国信息化发展的进展十分迅速。

基于此,信息管理与信息系统专业人才培养的任务十分艰巨。首先要定位,再定向,还要定措施。不同的高校要根据自己的特色来定位,如:以经、法、管理为主的综合性人文社科大学,其信息管理与信息系统专业就要定位在和经济、法律、管理的结合上,培养的人才主要适合在经济管理部门、司法部门、企事业单位等从事信息系统建设和管理以及科学研究等工作。定向的具体内容由培养目标来确定,本专业直接以信息化建设的人才需求为培养目标与标准,培养熟练掌握现代信息技术手段和方法,具有坚实的现代管理科学理论知识,具备较强的计算机应用能力的综合型、实用型的高级专门人才。定措施则是要确定对培养目标的具体实施过程和方法,包括师资要求、全程教学计划和教材建设等。

现各个高校在信息管理与信息系统专业的教材使用上五花八门,教材主要由任课教师自己选定。计算机方面的教材主要选用计算机科学与技术专业的教材,管理方面的教材主要采用管理学科的教材。尽管近年来一些出版社陆续出版了几套信息管理与信息系统专业的教材,但仍然不能满足教学的需要。根据教育部 1998 年信息管理与信息系统专业课程要求,结合中国高等院校信息系统学科课程体系课题组撰写的《中国高等院校信息系统学科课程体系 2005》(征求意见稿)(清华大学出版社,2005 年 11 月)的内容,我们组织长期从事信息管理与信息系统专业教学和研究的教师,在清华大学出版社的大力支持下,经过多次讨论和研究,组织编委会,制定教材编写规划,审定编写大纲,并采取主编负责制,层层把关,力争使本套教材成为具有系统性、完备性的高水平、高质量的信息管理与信息系统专业教材。

本套教材的主要特点是:

1. 系统性。教材自成体系,系统地体现本专业的知识体系和结构。
2. 完整性。教材能完整、准确地反映本专业的教学内容,满足培养高层次人才的需要。
3. 新颖性。教材要反映本学科的最新发展动态和研究成果。

4. 理论性。教材注重理论基础的培养,使学生具备扎实的理论知识。

5. 实用性。教材注重理论与实践结合,把培养学生分析问题、解决问题和实际动手能力作为一项重要的内容予以体现。

本套教材的成功出版,凝聚了众多长期从事信息管理与信息系统专业建设的专家、学者及相关人员的心血。我们殷切希望从事信息管理与信息系统专业的教育工作者对本套规划教材提出宝贵建议,使教材质量不断得到提高。让我们共同为培养高素质的信息化人才而努力。

刘腾红 教授

本规划教材编委会主任

2007年8月



前言

进入 21 世纪以来,Web 信息结构化以至 Web 信息语义的识别问题已经成为学术界、企业界所共同关注的问题。对此,人们正在探索多种形式的解决方案及其相关技术。在所有这些技术方案中,XML 都是作为一项基础标准为大家所认同。因此,XML 技术应用是非常值得学习和研究的。

扩展标记语言 XML(eXtensible Markup Language)是 SGML 的一个简化子集,其标准化工作由 W3C 的 XML 工作组负责,其倡导的公共格式,既不专属于某一家公司,也不依赖于特定的浏览器。XML 具有良好的结构特性,构成 Internet 环境中跨平台的、专注于内容及其结构标记的一项重要技术,成为当前处理结构化文档信息的有力工具。XML 使用一系列简单的标记描述数据,易于掌握和使用,易于在任何应用程序中读写数据,进而成为当前数据交换的唯一公共语言。这就意味着程序可以更容易地与 Windows、Mac OS、Linux 以及其他平台下产生的信息结合,进而很容易地加载 XML 数据到程序中进行分析,并以 XML 格式输出结果。

作为 SGML 子集的 XML 重新定义了 SGML 的一些内部值和参数,去掉了大量的、很少用到的功能。XML 保留了 SGML 的结构化功能,这样就使得网站设计者可以定义自己的文档类型,XML 同时也推出一种新型文档类型,使得开发者也可以不必定义文档类型。

本书系统地介绍扩展标记语言(XML)文法及其应用技术,共分 10 章,其主要内容有:XML 标记的核心概念、XML 文档类型描述(DTD)及其 XML 模式(Schema)、XML 文档显示样式(CSS/XSL)、XML 应用标准及其 XML 文档编辑与浏览工具、XML 链接(XLink/XPointer)与接口(DOM/SAX)技术、XML 数据库与数据交换。在每一章节中,均采用相关实例深入讲解。最后,介绍 XML 技术的应用领域,包括资源描述框架(RDF)、频道定义格式(CDF)、基于 XML 的电子商务标准以及 XML 与跨媒体出版等应用领域。

本书凝聚了作者多年教学及 XML 技术应用的经验,也是应用型本科院校“十一五”国家课题“我国高校应用型人才培养模式研究——信息管理与信息系统专业理论与实践双主体教学模式研究(FIB070335-A8-14)”的研究成果之一。其内容丰富,结构完整,概念清楚,深入浅出,通俗易懂,可读性、可操作性强。不仅适合作为信息管理专业“XML 技术”及其相关课程的教材,也适合作为 XML 技术应用工程师的参考书,具有很强的实用性和可操作性。

参与本书编写的主要人员有李禹生、孙平、张琳、贾瑜、刘昌华等,全书由李禹生教授统

由于 XML 技术的灵活与精深,本书在内容的取舍上必有值得商榷之处,甚至存在遗漏与不妥,恭请同行和读者不吝赐教,以求趋于完善。作者 E-mail: lee@whpu.edu.cn。

2009 年 1 月于汉口常青花园



21世纪 目录

第1章 XML 概述	1
本章学习要点	1
1.1 什么是 XML	1
1.1.1 XML 是一种元标记语言	1
1.1.2 XML 是一种信息的容器	4
1.1.3 XML 是一种开放的标准	5
1.2 XML 的历史及目标	6
1.2.1 XML 的历史	6
1.2.2 XML 的目标	7
1.3 XML 的主要特性	9
1.3.1 存取数据	9
1.3.2 格式化文档	11
1.3.3 保证数据完整	16
1.4 习题	17
第2章 标记和核心概念	18
本章学习要点	18
2.1 XML 标记	18
2.2 XML 文档	20
2.3 XML 文档序言	20
2.3.1 XML 声明	21
2.3.2 文档类型声明	22
2.4 XML 元素	24
2.4.1 语法	24
2.4.2 属性	26
2.4.3 名字空间	29
2.4.4 空白字符	34
2.4.5 XML 文档树	35
2.5 实体	37
2.5.1 字符实体	38
2.5.2 混合内容实体	40

2.5.3 不可解析实体	44
2.6 其他标签	46
2.6.1 注释	46
2.6.2 CDATA 段	48
2.7 习题	50

第 3 章 XML 有效性与模式

53

本章学习要点	53
3.1 模式信息的基本概念	53
3.2 DTD(Document Type Definition)	55
3.2.1 DTD 初步	55
3.2.2 关联 DTD 与 XML 文档	56
3.2.3 声明元素	59
3.2.4 声明元素的属性	63
3.2.5 定义实体	70
3.3 XML Schema	72
3.3.1 XML Schema 的提出	72
3.3.2 XML Schema 的基本结构	75
3.3.3 声明元素	75
3.3.4 全局声明与 ref 引用	88
3.4 小结	90
3.5 习题	91

第 4 章 XML 文档的显示

92

本章学习要点	92
4.1 样式单基础	92
4.2 CSS 样式单	93
4.2.1 CSS 的结构及规则	93
4.2.2 关联 CSS 与 XML 文档	96
4.3 XSL 样式单	98
4.3.1 XSL 概述	98
4.3.2 XSLT 初步	99
4.3.3 XSLT 元素	101
4.3.4 XPath 简介	112
4.3.5 FO 概览	119
4.4 两种样式单比较	120
4.5 小结	121
4.6 习题	121

第 5 章 有关 XML 的其他协议	122
本章学习要点	122
5.1 SGML 与 XML 标准体系	122
5.1.1 XML 基础标准及其相互关系	122
5.1.2 SGML 标准体系框架	126
5.1.3 XML 标准体系框架	128
5.2 主要国际组织简介	129
5.2.1 OASIS	129
5.2.2 W3C	129
5.3 主要 XML 应用标准简介	131
5.3.1 面向 Web 网页: XHTML	131
5.3.2 面向科技领域: MathML	133
5.3.3 面向图形图像: SVG	136
5.3.4 面向多媒体: SMIL	143
5.3.5 面向电子商务: cXML 等	145
5.3.6 面向无线网: WML	146
5.3.7 面向电子书: OPS 和 OPF	147
5.4 习题	148
第 6 章 XML 的编辑工具与浏览工具	149
本章学习要点	149
6.1 XML 分析工具	149
6.1.1 XML4J	150
6.1.2 Xerces	150
6.1.3 Oracle 的 XML Parser	151
6.1.4 Microsoft 的 MSXML	151
6.1.5 JDOM	151
6.1.6 DOM4J	152
6.1.7 其他 XML 分析器	153
6.2 XML 浏览工具	154
6.2.1 Internet Explorer	154
6.2.2 Mozilla Firefox	156
6.2.3 W3C 的试验基地——Amaya	158
6.3 相关编辑工具	159
6.3.1 XML 编辑工具	159
6.3.2 XSLT 编辑工具	162
6.4 其他相关协议的编辑浏览工具	165
6.4.1 图形应用工具	165
6.4.2 WAP 应用工具	167

6.4.3 电子出版工具 Microsoft Reader	170
6.5 习题	171
第 7 章 XML 链接	172
本章学习要点	172
7.1 XML 链接与 HTML 链接	172
7.2 XLink 语言	174
7.2.1 XML 链接术语	174
7.2.2 简单 XML 链接	175
7.2.3 扩展 XML 链接	177
7.3 XPointer 语言	185
7.3.1 XPointer 概述	185
7.3.2 XPointer 的三种形式	186
7.3.3 位置路径和位置集合	188
7.3.4 XPointer 对 XPath 的扩展	189
7.4 XML Base	192
7.5 小结	193
7.6 习题	193
第 8 章 XML DOM	194
本章学习要点	194
8.1 应用程序接口综述	194
8.1.1 接口的由来	194
8.1.2 DOM 与 SAX 并存	195
8.2 文档对象模型(DOM)	196
8.2.1 DOM 的组成	196
8.2.2 DOM 树	197
8.2.3 DOM 的 4 个基本接口	198
8.2.4 DOM 的应用	200
8.3 XML 简单应用程序接口(SAX)	204
8.3.1 SAX 分析器接口简介	204
8.3.2 SAX 的应用	205
8.3.3 应用程序的编译与执行	209
8.4 小结	209
8.5 习题	210

第 9 章 XML 与数据交换	211
本章学习要点	211
9.1 XML 的数据交换机制与数据存取机制	211
9.1.1 XML 数据交换机制	211
9.1.2 XML 数据存取机制	213
9.2 XML 与数据库	215
9.2.1 XML 与关系数据库	215
9.2.2 XML 与面向对象数据库	216
9.2.3 基于 XML 数据库的模型分析	217
9.2.4 XML 在数据库中的应用模式	218
9.3 XML 数据交换技术	220
9.3.1 ADO 控件技术	220
9.3.2 HTTPXML 对象技术	224
9.3.3 DB2XML 转换工具	226
9.3.4 ODBC2XML 转换工具	227
9.3.5 XOSL 转换工具	227
9.3.6 ASP2XML 组件	228
9.3.7 InterAccess 软件包	229
9.3.8 XML Servlet 软件包	230
9.3.9 Web 分布式数据交换(WDDX)	231
9.4 数据库对 XML 的支持	233
9.4.1 传统关系数据库对 XML 的支持	233
9.4.2 面向对象数据库对 XML 的支持	237
9.4.3 Software AG 的 Tamino Server	240
9.5 小结	242
9.6 习题	243
第 10 章 XML 的应用及展望	244
本章学习要点	244
10.1 XML 应用概述	244
10.2 XML 与语义网	246
10.2.1 语义网	246
10.2.2 资源描述框架	247
10.2.3 本体描述语言	248
10.3 XML 与电子商务	249
10.3.1 XML 是现代电子商务的基石	249
10.3.2 电子商务发展中的 XML“构件”	251
10.3.3 基于 XML 的电子商务标准	252
10.4 XML 与 Web 应用程序开发	255

10.4.1	当前 Web 应用程序存在的问题	255
10.4.2	Ajax	257
10.4.3	Silverlight	259
10.5	XML 与移动通信	261
10.5.1	无线应用协议 WAP	261
10.5.2	WAP 设计思想	262
10.5.3	WAP 应用环境	262
10.5.4	WAP 通信协议	263
10.5.5	WAP 的影响	265
10.5.6	WAP 的局限性	265
10.6	XML 技术展望	265
10.6.1	XQuery	265
10.6.2	Atom Publishing Protocol	266
10.6.3	XForms	267
10.6.4	XProc	268
10.7	习题	269
参考文献		270

第1章

XML 概述

本章学习要点

本章主要介绍标记语言的相关基础知识以及 XML 文档基本结构,着重介绍了 XML 的设计目标及主要特性。通过本章内容的学习,读者应该掌握以下主要内容:

- 了解 XML 的基本知识及其历史
- 掌握 XML 的设计目标及主要特性

XML 代表扩展标记语言(eXtensible Markup Language),是由互联网联盟(World Wide Web Consortium,W3C)的 XML 工作组定义的。这个工作组是这样描述该语言的:

“扩展标记语言(XML)是 SGML 的子集,其目标是允许普通的 SGML 在 Web 上以目前 HTML 的方式被服务、接收和处理。XML 被设计成易于实现,且可在 SGML 和 HTML 之间互相操作。”

XML 是一种专门在 World Wide Web(WWW)上传递信息的语言,就像 HTML(超文本标记语言)一样(自从 Web 出现以来,HTML 已经成为了创建 Web 页的标准语言)。既然已经有了 HTML,而且它已发展成一种似乎可以满足任何需要的语言,那么为什么还需要在 Web 上采用一种全新的语言呢?XML 有什么新特性和不同之处呢?它有什么特有的优点和长处呢?它与 HTML 有什么关系?它是要替换还是增强 HTML?

最后,这个以 XML 作为其子集的 SGML 是什么?为什么不用 SGML 来创建 Web 页呢?本章将回答所有这些问题。

1.1 什么是 XML

1.1.1 XML 是一种元标记语言

当前在 Internet 上大行其道的是超文本标记语言(Hyper Text Markup Language,HTML),XML 和 HTML 同属一个大家族——标记语言家族,因此,为了更好地理解 XML,需要首先来了解一下 HTML。

超文本标记语言(HTML)的精髓在于“标记”,然而“标记”的含义究竟是什么呢?“标记”的一个精确定义是:就数据本身的信息对数据进行编码的方法。

比方,一名学生正在复习准备考试,需要阅读大量的课程书籍,但并不是所有书中的信息都是同样重要的,通常情况下为了多次阅读过程中快速找到重点部分复习,该学生会使用一种荧光笔将一些重要语句或段落的文字加亮。在上述例子中,该学生其实就是对课程书籍里面的重要信息进行了标记,以便和一些不那么重要的信息进行区分。而且这种阅读书

籍的方式,对于大多数人来说都是相通的,也就是说,人们阅读课程书籍时都会遵循相类似的标记方式,那么某位同学借阅另外同学的课本的时候,他也会对该课本的精华部分略知一二。

由上述例子可以看出,当需要通过标记将有用的信息告知一组用户时:

(1) 首先,必须有一个标准,用它来描述什么是有效的标记。在上面例子中,标记可以被定义为在文字上的黄色荧光墨水印迹。而在 HTML 中,标记就是所谓的“标签”(tag)。

(2) 其次,还要有一个标准描述每个标记的具体含义。上面例子中的黄色荧光加亮标记意味着被加亮的句子很重要。而在 HTML 中,每一个标签都表明了一种显示的格式。

同样,“标记”的应用在计算机世界中也甚为广泛。文字编辑器借助标记来定义格式与外观,通信程序依靠标记来理解线路上所传输的信息的语意,数据库通过标记来将数据字段与一定的含义相连,并表明字段之间的关系,多媒体应用中标记则用来标示什么是图像和声音的源数据,等等。

当这些数据被传送给计算机或应用程序时,它必须自身携带一些必要的信息,以表明这些数据的含义,以及接收者应该如何处理这些数据。

可以想象,到了考试期间,如果学生的课本上没有任何重点标记,那么这位学生只能对着它望洋兴叹。同样,如果数据中不带有任何背景信息,应用程序也只能对它望洋兴叹!

HTML 就是众多著名的计算机标记系统中的一个。它描述了一系列标签,每个标签表明了一定的显示格式。被标记后的文件(即同时包含了纯文本和关于文本显示格式的标签的文件)由一个 HTML 文档处理工具,譬如一个浏览器,进行读取,然后再根据上述标记规则来加以显示。

例 1-1 一段简单的 HTML 代码,以列表方式显示客户信息。

```
<UL>
<LI>张三</LI>
<UL>
<LI>用户 ID: 001</LI>
<LI>公司: A 公司</LI>
<LI>EMAIL: zhang@aaa.com</LI>
<LI>电话: (010)62345678</LI>
<LI>地址: 五街 1234 号</LI>
<LI>城市: 北京市</LI>
<LI>省份: 北京</LI>
<LI>ZIP: 100001</LI>
</UL>
<LI>李四</LI>
<UL>
<LI>ID: 002</LI>
<LI>公司: B 公司</LI>
<LI>EMAIL: li@bbb.org</LI>
<LI>电话: (021)87654321</LI>
<LI>地址: 南京路 9876 号</LI>
<LI>城市: 上海市</LI>
<LI>省份: 上海</LI>
<LI>ZIP: 200002</LI>
</UL>
</UL>
```

上述 HTML 代码通过 IE 浏览器的显示结果如图 1-1 所示。

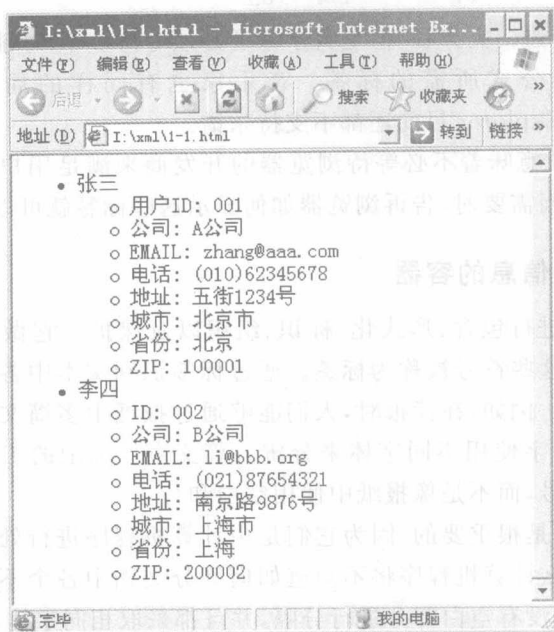


图 1-1 客户信息在 IE 浏览器中的列表显示

XML 和 HTML 一样,也是一种标记语言,在形式上有很多相似的地方。HTML 定义了一套固定的标签,用来描述一定数目的元素。如果标记语言中没有所需的标签,用户也就没有办法了。这时只好等待标记语言的下一个版本,希望在新版本中能够包括所需的标签,但是这样一来就得依赖于软件开发商的选择了,在这方面 XML 与 HTML 有所不同。

XML 是一种元标记语言。用户可以定义自己需要的标签。这些标签必须根据某些通用的原理来创建,但是在标记的意义上,也具有相当的灵活性。新创建的标签可在文档类型定义(Document Type Definition,在以后的篇幅中常简称为 DTD)中加以描述。DTD 可以看做是一本词汇表和某类文档的句法。例如,在 Peter Murray-Rust 的 Chemical Markup Language (化学标记语言,CML)中的 MOL.DTD 文件中描述了词汇表和分子科学的句法,其中包括 chemistry(化学)、crystallography(结晶学)、solid state physics(固体物理)等词汇,包括用于 atoms(原子)、molecules(分子)、bonds(化学键)、spectra(光谱)等的标签。这个 DTD 可与分子科学领域中的许多不同的人共享。对于其他领域也有其他的 DTD,用户还可以创建自己的 DTD。

XML 定义了一套元句法,与特定领域有关的标记语言(如 MusicML、MathML 和 CML)都必须遵守。如果一个应用程序可以理解这一套元句法,那么它也就自动地能够理解所有的由此元句法建立起来的语句。在实际应用中,浏览器不必事先了解多种不同的标记语言使用的每个标签,而只需在读入文档或是它的 DTD 时才建立给定文档使用的标签含义。

因此,DTD 的作用是定义标记语言标签的含义,而将这些标签所标识内容的显示方式设定交由另外一份附加在文档上的样式单完成。例如,考虑 Schrodinger 方程: