信息科学与技术丛书

XHTML 系列语言网页设计

吴云标　编著

机械工业出版社

本书全面系统地介绍模块化 XHTML 系列语言的网页设计，第 1 章至第 3 章为必要的基本概念和 XML 有关知识，以及 XHTML 模块化知识。第 4 章至第 6 章为 XHTML 系列网页的核心构件，包括文档结构、文本、超链接与清单。第 7 章至第 10 章扩展到 XHTML Basic，包括了 XHTML 系列网页的多功能多媒体基础。第 11 章介绍 CSS 样式语言来呈现 XHTML 网页的方法，第 12 章介绍 IE 的静态滤镜，第 13 章与第 14 章通过文本、表格、表单与图像的扩展，完成 XHTML1.1，第 15 章介绍 XHTML 帧网页及 XHTML+SMIL 文档，包括 IE 动态滤镜的使用。

本书各章均有小结及习题，并配有 XHTML 元素与属性索引，以方便学习和查询参考。

本书适用于计算机网页专业设计人员。

图书在版编目（CIP）数据

XHTML 系列语言网页设计/吴云标编著．
—北京：机械工业出版社，2002.9
（信息科学与技术丛书）
ISBN 7-111-10665-2

Ⅰ．X…　Ⅱ．吴…　Ⅲ．超文本标记语言，XHTML—程序设计
Ⅳ．TP312

中国版本图书馆 CIP 数据核字（2002）第 053757 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）
策　　划：
责任编辑：田　梅
责任印制：

·新华书店北京发行所发行

2002 年 9 月第 1 版·第 1 次印刷
1000mm×1400mm　B5·14.375 印张·557 千字
0001—5000 册
定价：34.00 元

前　言

Internet 在世界范围的迅速崛起，HTML 功不可没。但是最近几年，随着各种信息资源登陆国际互联网，网上信息服务的层出不穷，已使 HTML 无法招架，为了国际互联网的广阔前景，人们开发出了 XML，它指明了解决 HTML 所面临困境的必由之路。但 XML 是元语言，并不取代 HTML，所以需要一种符合 XML 标准，并具备 XML 特性和 HTML 功能的标记语言。满足该条件的 XHTML 由此而生，它将完全取代 HTML，成为新一代的网页设计语言，而 HTML 也将随着时间的推移逐渐成为历史。

XHTML 是一种灵活、基本、可拼装的内容标记语言，可以说，它是一系列语言的统称。本书根据模块化思想，以 XHTML Basic 为起点，全面系统地介绍应用 XHTML 系列语言设计网页的概念与方法。全书共 15 章。第 1 章至第 7 章为 XHTML 网页设计基础知识，介绍必要的基本概念和 XML 有关知识、XHTML 模块化知识。第 8 章至第 10 章为 XHTML Basic，包括 XHTML 系列网页的基本要件以及 XHTML 系列网页的多功能设计。第 11 章至第 15 章为 XHTML 扩展，主要包括通过样式语言来呈现 XHTML 网页的方法；从 XHTML 基本网页通过增加模块扩展到其他严格的 XHTML 系列网页，也包括了 XHTML1.1 以外的帧网页，并对 XHTML 发展趋势之一的 XHTML+SMIL 动态网页文档作了简介。

本书特色：

- 假设读者没有学过 HTML，没有做过网页，你完全可以掌握新一代网页标记语言 XHTML。
- 遵循 XHTML Basic，设计基本网页；遵循 XHTML 模块化，设计复杂网页，为各行各业创作可适合多种平台的网上信息内容提供标准方法。
- XHTML 使用 XML 语法，本书深入介绍 XHTML，所以能使读者在学会 XHTML 的同时，了解 XML 的语言结构和应用。从而如读者想学其他符合 XML 的标记语言，就会得心应手。也有益于读者进一步学习研究 XML。
- XHTML 与 HTML 的重要区别之一是 XHTML 使得内容结构与显现样式相互分离。所以本书也将详细解说 CSS 中各种排版属性的功用，如何结合 XHTML 进行网页设计。

读者不必参考其他 HTML 书籍。如果你使用过 HTML 编制网页，为编制 XHTML 网页，你可能得用心去克服一些编制 HTML 文档的习惯。

如果没有注明，书中范例所用的浏览器均为中文 Internet Explorer 6。目前它对 XHTML 规范的支持较好。

由于笔者水平与时间有限，书中错误之处在所难免，恳请读者批评指正。

出 版 说 明

随着信息科学与技术的迅速发展，人类每时每刻都会面对层出不穷的新技术、新概念。毫无疑问，在节奏越来越快的工作和生活中，人们需要通过阅读和学习大量信息丰富、具备实践指导意义的图书，来获取新知识和新技能，从而不断提高自身素质，紧跟信息化时代发展的步伐。

众所周知，在计算机硬件方面，高性价比的解决方案和新型技术的应用一直倍受青睐；在软件技术方面，随着计算机软件的规模和复杂性与日俱增，软件技术受到不断挑战，人们一直在为寻求更先进的软件技术而奋斗不止。目前，计算机在社会生活中日益普及，随着因特网延伸到人类世界的层层面面，掌握计算机网络技术和理论已成为大众的文化需求。正是这种在社会各领域的全方位渗透，信息科学与技术正在电工、电子、通信、工业控制、智能建筑、工业产品设计与制造等专业领域中得到充分、广泛的应用。相应地，这些专业领域中的研究人员和工程技术人员将越来越迫切需要汲取自身领域信息化所带来的新理念和新方法。

针对人们对了解和掌握新知识、新技能的热切期待，以及由此促成的人们对语言简洁、内容充实、融合实践经验的图书迫切需要的现状，机械工业出版社适时推出了“信息科学与技术丛书”。这套丛书涉及计算机软件、硬件、网络、工程应用等内容，注重理论与实践相结合，内容实用，层次分明，语言流畅，是信息科学与技术领域专业人员学习和参考不可或缺的图书。

现今，信息科学与技术的发展可谓一日千里，机械工业出版社欢迎从事信息技术方面工作的科研人员、工程技术人员积极参与我们的工作，为推进我国的信息化建设作出贡献。

机械工业出版社

第 1 章　国际互联网与标记语言

本章为导引，第一部分介绍 Internet 和 WWW 的一些基础知识；第二部分介绍标记语言，SGML，HTML 及其问题，最后是 XML 与 XHTML 产生的必然性。

1.1　国际互联网基础知识

国际互联网是目前世界上使用人数最多、连接国家最多、信息资源最多的大型计算机网络，在国际上被称为“21 世纪的信息高速公路和信息平台”，它允许世界上数以亿计的人们进行通信和共享信息。国际互联网仍在迅猛发展，并在发展中不断地更新。

使用国际互联网其实包含了使用 Internet 和使用 WWW（World Wide Web）。本章简单地介绍与网页设计语言有关的 Internet 和 WWW 知识。

1.1.1　Internet 的基本工作原理简介

Internet 的工作原理的细节是很具技术性的。在此简单地介绍一些基本概念。Internet 的基本原理主要有分组交换、网际协议（Internet Protocol）和传输控制协议（Transmission Control Protocol）。

1．分组交换原理

计算机网络连结通常都采用共享方式，即多台计算机共享一条传输线路。这样做经济，但缺点是在传输时间上产生延迟。这类似于许多车辆同时来到同一路口，不可能让所有的车都通过，只能允许几辆车同时通过，这时其他车辆就只好排队等候。在网络上，当大量数据流通过时，排在前面的数据流可以马上通过，而其他数据流就只能等待，这种现象叫“延迟”。

为使网络中每一台计算机都不需要等候很多时间，方法是让网络中每一台计算机每次只传送一定的数据量，对连接的每台计算机实行“轮流服务”。

每次所能传送的数据量的单位称为一个数据包（或数据段），这种将信息的数据总量分割传送，设备轮流服务的原则，叫分组交换。计算机网络用来保证每台计算机平等地使用网络资源的技术，称为分组交换技术。这种技术给 Internet 带来许多好处。

分组交换系统可以使多台计算机在一个共享网络上进行通信时，具有最小的延迟。因为分组交换将每个要传送的信息分成小的数据包，并且使共享网络的计算机

轮流发送数据包。分段传输很快，常常达到每秒一千个以上的数据包，当几个人同时发送信息到一个共享网络时，千分之几秒的时间间隔是感觉不到的，所以可以认为延迟是不存在的。

分组交换允许任何一台计算机在任何时候都能发送信息。分组交换技术能够在有计算机准备发送信息和有计算机停止发送时立即进行自动调整，重新分配网络资源，使每台计算机在任何时刻都能够公平地分享网络资源。

分组交换技术让很多用户在同一时间使用同一通信线路。同一时刻在网络上流动着来自各个方向的多台计算机的数据包。就如一条高速公路上，各种汽车（即使它们开向不同的地方）都在公共道路上行驶。

分组交换技术使网络具有很大的灵活性。在信息传输中，各个数据包完全可能沿不同的路由传输。当条件改变时，控制数据流动的计算机通常可以找到当时最好的路由。例如，当网络的某一特定部分过载，数据包可以改变路线去走那些比较空闲的线路。用数据包传输的又一个好处是，当某处出错，只须重新传送单个数据包，而不是整个信息。这就从整体上提高了网络的传输速度和效率。

这些灵活性保证了高度可靠性。虽然所有数据包都必须通过很多计算机，但 Internet 运行非常好，它只需几秒钟就可把一个文件从一主机传输到另一主机，哪怕它们远隔重洋。

2．TCP/IP

计算机网络是由许多计算机组成的，为了确保不同类型的计算机能在一起工作，就要有共同遵守的通信协议。就如两人交流思想，需要有双方都懂的语言，才能进行交流一样。通信协议正像两台计算机交换信息所使用的共同语言。这里所谓协议，是指通信双方在通信中所共同遵守的约定，是一套用技术术语描述某些事应如何做的规则。

计算机的通信协议精确地规定了计算机在彼此通信时的所有细节。它定义计算机发送每条信息的格式和含义，规定哪些情况下应发送哪些特殊的信息，以及接收方的计算机应做出什么反应等等。

（1）网际协议 IP

Internet 上使用的一个关键的底层协议是网际协议（Internet Protocol），通常称 IP 协议。IP 非常详细地规定了计算机在通信时应遵循的规则。例如，数据包必须怎样组成，路由器必须怎样将每一个数据包递交到它的目的地，如何传送、如何接受等问题。IP 协议给每个数据包写上发送主机和接收主机的地址，一旦写上源地址和目的地址，数据包就可以在网上传送数据了。

连接到 Internet 上的每台计算机都必须遵守 IP 协议，所产生的数据包必须使用 IP 规定的格式。为此使用 Internet 的每台计算机都必须运行 IP 软件，以便时刻准备发送或接收。一旦 Internet 上的每台计算机都安装了 IP 协议软件，任何网上计算机

都能产生 IP 协议数据包并将其发送给其他计算机，使全世界千千万万各种类型的计算机彼此能够通信。

（2）传输控制协议

传输控制协议（Transmission Control Protocol，TCP）负责提供可靠无差错通信服务。

线路容量总是有限的，大量到达的数据包会引起超载而发生堵塞，这时 Internet 上的设备将不得不丢弃一部分数据包，直到堵塞解除。TCP 将自动检测丢失的数据包并能恢复丢失的数据包。

信息被分成数据包传送，同一信息的不同数据包可能沿不同路径到达目的地，因此有可能使先发送的数据包因绕行而迟到，后发送的数据包反而早到。TCP 协议会自动检测到达目的地的数据包的顺序，并进行调整。

有时因硬件故障也可能导致重复发送同一数据包，结果在目的地会得到同一数据包的多个副本。TCP 协议自动检测发来的数据包有没有重复，如果有，只接受最先到达的数据包。

当一台计算机准备与另一台远程计算机连接时，TCP 协议会以类似于人打电话的方式向对方发出呼叫信号，请求连接。被呼叫的计算机接受呼叫，发出应答信号。一旦连接建立，双方就能相互发送数据，直到发送、接收完毕，终止连接。在接收端，TCP 接收到数据包并核查错误。如果有错误发生，TCP 可以要求重发这个特定的数据包。只要所有的数据包都被正确地接收到，TCP 将用序号来重构原始信息。

IP 协议只保证计算机能发送和接收分组数据，但 IP 并不解决数据传输中可能出现的问题。这个问题由 TCP 协议来解决。换句话说，IP 的工作是把信息（数据包）从一地传送到另一地；TCP 的工作是管理这种传送并确保其数据是正确的。IP 协议负责数据的传输，TCP 协议负责数据的可靠传输。

（3）TCP/IP 协议的数据传输过程

这两个协议可以分开使用，各自完成自己的功能。但它们是作为一个系统来设计的，在功能上是相互配合、相互补充的。连接 Internet 的计算机必须同时使用这两个协议。因此在实际中常把这两个协议称作 TCP/IP 协议。

TCP/IP 协议所采用的通信方式是分组交换方式。TCP/IP 协议的基本传输单位是数据包，它们在数据传输过程中主要完成以下功能：首先由 TCP 协议把数据分成若干数据包，给每个数据包写上序号，以便接收端把数据还原成原来的格式。IP 协议给每个数据包写上发送主机和接收主机的地址，一旦写上了源地址和目的地址，数据包就可以在物理网上传送数据了。IP 协议还具有利用路由算法进行路由选择的功能。这些数据包可以通过不同的传输途径（路由）进行传输，由于路径不同，加上其他的原因，可能出现顺序颠倒、数据丢失、数据失真甚至重复的现象。这些问题都由 TCP 协议来处理，它具有检查和处理错误的功能，必要时还可以请求发送端重发。

需指出的是，虽然 TCP／IP 的实际名字是来自最重要的两个协议，TCP 和 IP。但是 TCP／IP 实际上是许许多多用来连接计算机和网络的协议合起来的共有名字，是一个很完整的协议组。除了 TCP 和 IP 两个协议之外，TCP/IP 还包括其他协议，其中有工具性协议、管理协议和应用协议等，譬如，HTTP、SMTP、FTP 等都是 TCP/IP 协议族中的协议。TCP／IP 是把计算机和通信设备组织成网络的协议大家庭。

3．Internet 的地址结构

Internet 采用一种惟一、通用的地址格式，为 Internet 中的每一个网络和几乎每一台主机都分配了一个地址。IP 地址是网上计算机惟一的标识号，Internet 上的其他计算机都认识这个标识号。有了地址，信息才可以传到那里，这正如日常生活中发送邮件一样，它用的是邮政编码，如 310020 中 31 表示浙江省，00 表示杭州市，20 表示杭州市区中的一块。Internet 中的地址类型有 IP 地址和域名地址两种。

（1）IP 地址

IP 地址（IP address）是 32 位（4 字节）二进制数值，用于惟一标识与 Internet 或与另一台 Internet 主机相连接的一台主机。

根据 TCP／IP 协议规定，IP 地址用二进制来表示，每个 IP 地址长 32 位。例如：

11000010.11000111.00011000.00000011-->202.199.24.3

IP 地址是 Internet 主机的一种数字型标识。由于二进制不容易记忆，通常将 32 位分为 4 个字节，每个字节转换成十进制，中间用小数点分开，每组十进制数代表 8 位二进制数，其范围为 0～255，但是 0 和 255 这两个地址在 Internet 有特殊用途（用于广播），因此实际上每组数字可真正使用的范围是 1～254。IP 地址的这种表示法叫做“点分十进制表示法”。

（2）域名地址

域名（domain name）是连接到 Internet 或其他 TCP/IP 网络上的计算机设备的地址，以层次结构“服务器.组织.类型”确定地址的拥有者。

数字型标识对计算机网络来讲自然是最有效的，但是对使用网络的人来说有不便记忆的缺点，为此人们研究出一种字符型标识，这就是域名，或域名地址。域名地址命名采用层次型命名系统，更直接体现出层次型的管理方法，其域名结构的通用格式如下：

第 n 级子域名.…….第二级子域名.第一级子域名

这里一般 $2 \leq n \leq 5$。

域名可以用一个字母或数字开头和结尾，并且中间的字符只能是字母、数字和连字符，标号必须小于 255。为了简便并容易记住名字，每个标号小于或等于 8 个

字符，但这不是必须的。

第一级域名往往是国家或地区的代码，第二级域名往往表示主机所属的网络性质，比如属于教育界还是政府部门等。

所以对于美国之外的国家的域名形式为：主机名.机构名.机构类型.国家名。

美国的域名形式为：主机名.机构名.机构类型。

常见国家或地区域名包括 cn：中国大陆，tw：中国台湾，hk：中国香港，uk：英国，jp：日本，ca：加拿大，fr：法国，de：德国，au：澳大利亚。机构类型包括 com：商业公司，edu：教育，gov：政府部门，mil：军事部门，net：网络支持中心，org：非赢利组织团体。

chinanet 的用户一级域名为 cn，各省则用其拼音缩写如 zj，bj 等。

注意几点：1）域名在整个 Internet 中必须是惟一的，当高级子域名相同时，低级子域名不允许重复。2）大小写字母在域名中没有区别。3）一台计算机可以有多个域名（通常用于不同的目的），但只能有一个 IP 地址。4）主机的 IP 地址和主机的域名对通信协议来说具有相同的作用，使用没有什么区别。但是，当用户所使用的系统没有域名服务器，就只能使用 IP 地址而不能使用域名。5）为主机确定域名时应尽量使用有意义的符号。

（3）统一资源定位器

统一资源定位器（Uniform Resource Locator，URL）是 Internet 上资源的地址。统一资源定位器指定了在访问资源时所使用的协议、资源所在的服务器名以及资源的路径，其中资源的路径是可选项。Web 浏览器或用户代理器通过使用统一资源定位符（URL）来对 Internet 上的资源进行定位。它反映了一个基本思想，那就是：通过 URL，用户应能在 Internet 上的任何一台机器上访问任何可用的公共数据。URL 的标准格式如下：

<Protocol>://<HostName:Port>/<Path>/<Filename>

其中，Protocol：所使用的访问协议；HostName：文档和服务所在的主机名，一般表示为服务器域名，少数情况下是十进制 IP 地址；Port：服务端口号，各种协议都有自己相应的端口号，如果采用标准端口号，则可以忽略此处的端口号；Path：通向数据的全部路径；Filename：包含了所需数据的文件的名称。

例如：http://www.computerworld.com.cn/2000/news/03/0316_17.asp。其中 http 是协议，www.computerworld.com.cn 是主机域名，WWW 是主机类型，表明站点是 Web 主机；computerworld 是服务器名；com 是商业；cn 是中国。2000/news/03/是路径。0316_17.asp 是文件名，使用 URL 可以访问 Internet 上许多类型的资源。注意：在 URL 中，目录分隔符用斜杠“/”，而不是反斜杠“\”。具体分析如图 1-1 所示。

协议和主机总是必需的，但路径和文件名并非对所有网页来说都是必需的。

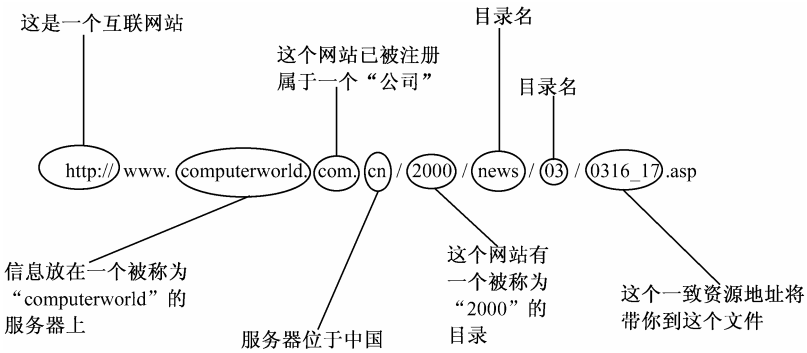

图 1-1　分析 URL 的一个示例

常见的协议有 HTTP：**超文本传输协议（**HyperText Transfer Protocol**）；**FTP：**文件传输协议（**File Transfer Protocol**）；**GOPHER：GOPHER **协议；**TELNET：**远程登录协议；**NEWS：**新闻组协议；**WAIS：**广域信息检索协议；**IGMP：Internet **组管理协议；**IMAP4：Internet **消息访问协议；**MAILTO：**发送电子邮件协议；**POP：**邮局协议。**

● URL **片断标识符**

Web **通信接受一种被叫做** URL**“片断标识符”的约定来支持** URL **指向一个资源内的一个地址，譬如网页内的某个段落。这样的** URL **以井字号“#”加一个锚标识（称之为片段标识符）跟着。例如，下面是一个指向名为** section4 **的一个锚的** URL**：**

http://www.somecorp.com/xhtml/top.html#section4

● **相对** URL

一个相对 URL **不包括协议或主机域名信息。它的路径通常指向与当前文档位于同一机器的一种资源。相对** URL **可能包括相对路径的组成成分（“..”意味着父目录），也可能包含** URL **片断标识符。**

根据相对 URL **可以决定完整的** URL**。简单地说，一个完整** URL **可以通过附加一个“基础”部分（基** URL**）在相对** URL **上引申而来。如果一个基** URL **用斜线结束，则完整** URL **通过加上相对** URL **获得。例如，基** URL **是** http://www.nocorp.com/dir1/dir2/**，并且相对** URL **是** dog.html**，那么最后得到的** URL **是** http://www.nocorp.com/dir1/dir2/dog.html**。如果基** URL **不以斜线结束，基** URL **的最后部分被认为是资源，而完整** URL **通过把相对** URL **加到基** URL **的上层而得。如，基** URL **是** http://www.nocorp.com/dir1/dir2**，而相对** URL **是** dog.html**，那么最后得到的完整** URL **为** http:// www.nocorp.com/dir1 /dog.html**。**

（4）统一资源标识符

统一资源标识符（Uniform Resource Identifier**，**URI**）是通过类型和位置来标识**

Internet 上任何资源的字符串。它提供了一种简单而且可以扩展的方式，用来标识抽象的物理资源。由 RFC2396 规定为国际 Internet 标准。URI 可以是定位器（locator）、名称或兼有两者。它的含义比较广泛，泛指所有以字串标示的网络资源，其范围涵盖了 URL 和 URN。

统一资源命名（Uniform Resource Name，URN）是一种通过使用名字来惟一标识 Internet 上可用资源的方案，该方案并不考虑资源所在的具体位置，用来标识由专门机构负责的稳定且全球惟一的资源（如图书馆的图书总目）。它包括那些不属于 URL 的 URI。

1.1.2 WWW 中的技术

1．WWW 的含义及其组成

World Wide Web，缩写为 WWW，也写作 w3，W3，Web，常译为万维网，是 Internet 上的一个基于客户/服务器体系结构的分布式多平台的超文本超媒体信息服务系统。它是 Internet 的最主要的信息服务，允许用户在一台计算机上通过 Internet 存取另一台计算机上的信息，存取世界各地的超媒体文件，内容包括文字、图形、声音、动画、资料库，以及各式各样的软件。

WWW 由千千万万个网站、网页组成。网页（Web page）是位于 WWW 上的文档。网页一般由一个超文本文件以及相关的图形和脚本文件组成，这些文件被保存在特定计算机的特定目录中（因而可用 URL 区别）。通常，一个网页中也包含着对其他网页的链接。网站（Web site）是一组相关网页以及有关的文件、脚本和数据库等，它们通过 WWW 上的 HTTP 服务器提供服务。网站上的网页集合以及其他文件的入口页被称为主页（home page）。

WWW 上的信息通过超文本或超媒体链接（即超链接）被相互连接起来。超文本允许用户选择来自文本的一个词汇，从而访问包含与那个词有关的其他信息的其他文献；超媒体文献以链接到图像、声音、动画和电影为特色。

网页用 XHTML（或 HTML、XML 等）语言编写，并使用 URL 进行标识，URL 指明了特定的计算机和路径名，用户通过它可对信息资源进行访问，并在 HTTP 协议下将资源在节点间进行传输，直至传输到最终用户。

WWW 是当今全世界最大的电子资料世界，它包含整个 Internet 上的所有 Web 站点、Gopher 信息站、FTP 档案库、Telnet 公共存取账号、News 新闻讨论区以及 Wais 资料库。也是目前 Internet 上最流行的一种工具。人们日常一般所说的“上网”或者说“上 Internet”，其实指的就是连上 World Wide Web。

2．客户机与服务器

WWW 服务采用早已成熟的客户机/服务器模式，其体系结构如图 1-2 所示。

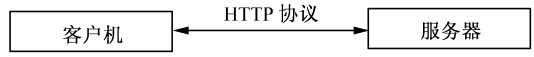

图 1-2　Web 的客户机/服务器结构模型

一般来说，客户机（Client）是指在 Internet 或其他网络中，当一个用户需要服务时向服务器提出请求的程序，也指运行该程序的计算机。通常情况下，一个客户机启动与某个服务器的对话。而服务器（Server）一般是指在 Internet 或其他网络中，用以对客户的命令作出反应的程序，也指运行该程序的计算机。服务器存放着要发布的信息，并提供相应服务，服务程序是等待客户机请求的一个自动程序。一个客户机可以向许多不同的服务器请求。一个服务器也可以向多个不同的客户机提供服务。客户机与用户交互，提供一个界面允许用户请求服务器的服务，并且显示由服务器返回的结果。客户机通常做一些解释或翻译，将由用户输入的命令转化成为服务器要求的格式。客户机也可以通过在提交用户的命令前验证用户的身份和授权来提供系统安全性。客户机还可以检查用户命令的有效性和完整性；例如，它们可以把银行账户传输限制在某一个最大量。另一方面，服务器是被动的，它们从来不先开始通信，而是等待客户机请求的到达，然后响应那些请求。客户机在单个工作站或个人计算机上可得，而服务器一般放在网络的其他地方，通常是放在更有力的机器上。协议是客户机请求服务器和服务器如何应答请求的各种方法的定义。WWW 使用的协议为超文本传输协议（HyperText Transfer Protocol，缩写为 HTTP）。所以 WWW 的基本结构是由一个服务器和客户机组成，它们之间使用 HTTP 相互通信。

WWW 客户机又可称为浏览器（Browser）或用户代理器（User Agent），作为一种客户端应用程序，它允许用户查看位于 WWW、另一网络以及用户计算机上的超文本文档；允许用户沿着文档中的超链接进行浏览或传输文件。

浏览器主要包括：Lynx、Mosaic、Netscape、Internet Explorer（IE）等。目前最流行的是微软的 IE 和网景（现在美国在线的一个附属公司）的 Netscape。对许多人而言，使用 IE 或 Netscape 几乎成了使用 Web 的同义词。但无论是 IE 还是 Netscape，只是网上浏览器的一种，作为网页作者，在设计网页时，不能只考虑网页在 IE 或者 Netscape 中的浏览效果，还要考虑到其他浏览器，否则将会影响甚至妨害某些用户阅读这些网页。

WWW 客户机并不限于向 Web 服务器发出请求，还可以向其他服务器（例如 GOPHER、FTP、NEWS、MAIL）发出请求，以获得相关服务。

为了在 WWW 上发布信息，必须有一个存放信息的 Web 服务器（Webserver），即 HTTP 服务器。它是指使用 HTTP 协议的服务器程序。当客户程序（如网络浏览器）发出请求时，服务程序会提供相应的超文本文档和相关文件及脚本。一旦提供完所要求的文档或文件，客户机和服务器之间的连接就会被切断。HTTP 服务程序用于 WWW 和局域网的站点上。Web 服务器也指任何正在运行 HTTP 服务程序的机器。

一般而言，Web 服务器的信息是允许所有人访问的，但是也有例外，Web 服务器可以建立用户，设定访问权限，限制用户对某些内容的访问。

通常的服务器来自于 CERN、NCSA、Netscape、Microsoft。

3．HTTP 协议

超文本传输协议（Hypertext Transfer Protocol，HTTP）用于在 WWW 上访问信息的客户机/服务器协议。Internet 的基本协议是 TCP/IP 协议，FTP、Archie Gopher 等是建立在 TCP/IP 协议之上的应用层协议，不同的协议对应着不同的应用。WWW 服务器使用的主要协议是 HTTP 协议。HTTP 协议也是建立在 TCP/IP 基础之上的应用层的通信协议，它提供 WWW 浏览器和 WWW 服务器之间的通信。它不仅需要保证能够正确地传送文档，还要能够确定传送文档的哪一部分，以及哪一部分将优先显示给用户。

由于 HTTP 协议支持的服务不限于 WWW，还可以是其他服务，因而 HTTP 协议允许用户在统一的界面下，采用不同的协议访问不同的服务，如 FTP、Archie、SMTP、NNTP 等。另外，HTTP 协议还可用于名字服务器和分布式对象管理。

基于 HTTP 协议的客户/服务器模式的信息交换过程包括四个步骤。建立连接：客户与服务器建立 TCP 链接；发送请求：打开一个连接后，客户机把请求消息送到服务器的停留端口上，完成提出请求动作；发送响应：服务器在处理完客户的请求之后，要向客户机发送响应消息；关闭连接：客户和服务器双方都可以通过关闭套接字来结束 TCP/IP 对话。

简言之，用户上网的工作机制可描述如下：客户端使用浏览器与服务器建立连接，用户在客户机上所下达的指令通过浏览器分析后，将连往服务器，并将用户的要求传送给服务器。当服务器接收到用户的请求后，就会作出响应，回送应答数据，把存放在服务器上的信息（网页形式）传回给用户，然后再由浏览器显示在屏幕上。当客户端发出断开连接的请求后，服务器关闭连接，一次会话结束。

在 WWW 中，“客户”与“服务器”是一个相对的概念，只存在于一个特定的连接期间，即在某个连接中的客户在另一个连接中可能作为服务器。WWW 服务器运行时，一直在监听，等待连接的出现。

1.2 标记语言简介

1.2.1 什么是标记语言

标记语言（Markup Language）作为一种语言，它具有语言的一般特征，即它是一种符号系统，具有自己的词汇（符号）和语法（规则）。所谓标记，就是作记号。譬如在写文章时，为了强调某些用语，在某些词下加点；在读书时，为了表示一句

话的重要性，在这句话的下面画一条下划线，等等，这就是作标记。标记语言如果简单地从字面来理解，就是一套用来作标记的符号系统。

语言的功能是用来交流，标记语言也不例外。像上面的例子，用点或下划线可说是人们之间的约定俗成。这个例子说明，作为交流的标记语言至少要确定三点：1）什么是有效的标记，或者说这套标记语言的词汇有哪些。在上面例子中，前者标记被定义为在词语下的点，后者是下划线。2）每个标记的具体含义，即词汇所代表的意义。上面例子中的点或下划线标记意味着被加点的词语或加下划线的句子很重要。3）每个标记的具体用法，即这套标记语言的语法。上面例子中的用法为将标记加在要强调的词语或句子下面。

本书所讲的标记语言，实际上是一种为了让计算机看的语言。譬如，有一篇文章，标题是《什么是标记语言》，作者是伍兹仁，包括摘要和参考文献。现在想把这篇文章输入计算机，存作电子文件，并且这个电子文件能让某种应用软件解读。让它“知道”这篇文章的标题是什么，文章的作者是谁，摘要是什么，哪些是正文，参考文献有哪些。为此不妨设计这样一套标记：

〈标题〉　〈/标题〉;〈作者〉　〈/作者〉;〈摘要〉　〈/摘要〉;〈正文〉　〈/正文〉;〈参考文献〉　〈/参考文献〉。

利用这套标记和这篇文章结合起来，即将这套标记添加到这篇文章的适当位置中去，把相关内容夹在标记之间，便形成一份有用的电子文件。如下所示：

〈标题〉什么是标记语言〈/标题〉

〈作者〉伍兹仁〈/作者〉

〈摘要〉本文着重论述了标记语言的功能与类型。……〈/摘要〉

〈正文〉标记语言作为一种语言，它具有语言的一般特征，……〈/正文〉

〈参考文献〉……〈/参考文献〉

所谓有用的电子文件是针对应用软件而言的，也就是说这种应用软件能够解读这份电子文件中的标记语言，并且根据标记语言的意思来对这篇电子文件作特定的处理。譬如将文件内容在屏幕上显示，凡标题和作者用黑体显示，摘要用楷体显示，并且左右两边采用悬挂缩进排，等等。或者应用程序将文件内容输出到语音设备上，按照某种规则读出来。例如，文章的标题重读，在开始读作者姓名前，先加作者一词，即读作：作者，伍兹仁，如此等等。

所以标记语言是指在文本文件中使用一个代码集，用于指示计算机在打印机或显示器上编排文件的格式，以及文件中的索引和链接内容等。标记语言是为了方便计算机处理而设计的，其中所用的标记，往往是代表一定含义的文字或数字。不过对计算机来说，它并不像人那样理解这些标记的含义，对它这些标记无非是一些代码，在应用程序中起相应的控制作用，所以它们被称为标记代码（codes）或控制标签（tags）。这些代码或控制标签脱离相应的应用程序是无用的，而且如果没有标记内容，控制标签也是毫无意义的。所以一种标记语言不只是一套标记符号，而是指

标记语言所定义的整体，包括标记内容。习惯上，把用这种标记语言编写的电子文件，称为某某语言文档。例如用 XHTML（标记）语言标识的文档被称为“ XHTML（标记语言）文档 ”。

标记语言主要是为了方便计算机处理文档（documents）的。如果没有标记语言，一篇文章在计算机中，不像在屏幕上或者书上看到的那样，能感觉到它的结构，什么是标题，作者是谁，哪些是摘要，哪些是正文，正文的结构如何，等等。存在计算机中，对程序而言，只是字符流，或者说文字流。标记语言基本上可以被看作是一种文档结构描述语言，它不同于一般的程序设计语言，在计算机处理过程中，标记语言的标记既可以作为控制语句来使用，也可以作为数据来使用。标记语言可使电子文件更具结构性，应用软件通过标记语言来解读这种结构性，并加以应用，对文档作特定处理。譬如，按规定的结构进行排版，或者进行数据交换、整合、搜寻检索等等。

目前，标记语言可分为两类：特殊用途的标记语言和一般通用的标记语言。一般来说，标记语言的复杂程度与它的应用范围宽广以及应用的深度或者说涉及问题的详细程度有很大关系，应用范围越广，内容越详细，所要表达的结构就越多，越复杂，从而标记语言也就越复杂，越难学。

特殊用途的标记语言是针对某种特殊应用的，有的甚至是专门针对某种特殊的应用软件而特别制定的。如 XHTML 语言就是一种特殊用途的标记语言。XHTML 语言是应用于网页设计的语言，它是由 HTML 语言发展而来的。

同类应用也可能使用不同的标记语言，有些是专门针对特定软件制定的。譬如，现在常用的多种文本编辑软件，像 Word、WPS 等等。这些软件的功能各有千秋，但大同小异。不过它们都有自己的标记语言。只是当用户在使用这些软件来编辑文本时，只需知道有哪些功能就行。而把相应的标记语言与被编辑的文本结合起来，即在文本的相应地方插入标记的工作是由软件自身完成的，所以用户并不感到这些标记语言的存在。但这类软件的标记语言的不同，也往往使得用一种软件编辑的文档，在另一种软件中无法使用或者出错。而且文本编辑软件主要使用格式化标记，即用于控制显示样式的标记。

一般通用的标记语言它只描述较大应用范围的某些结构，多种类型文档所共有的内容结构或逻辑结构，而不管文档的呈现样式。也就是说，它不是针对某一特殊应用领域，更不是为某一种特别的软件来量身定做的。这样的标记语言，也需要得到多种应用领域的软件开发者的认可，有很多应用软件都支持它，都可以解读它，也可以基于这种标记语言来开发新软件。这种标记语言本身也应相当地成熟和稳定。如果用一般通用的标记语言来标记文档，其制作出来的电子文件的可携性（portable）就会很高。像 SGML 就属于一般通用的标记语言。

从标记语言历史看，XHTML 只是一系列密切相关的标记语言的最新成员，它们以 SGML 开始，历经若干代 HTML 的演化，延续到 XML。

1.2.2 标准通用标记语言 SGML

标准通用标记语言（Standard Generalized Markup Language，SGML）是一种表示文档的通用标记语言。它是描述一篇文档的内容与它的结构之间的关系的一个国际标准。SGML 允许基于文档的信息以一种开放和中立的格式跨平台被共享和重用。

SGML 是一种描述语言，它定义了以电子形式表示文本的方法。它的优点是：高稳定性、高可携性和高完整性。SGML 作为国际标准，其规范结构相当严谨，可信度相当高。自 1996 年以来 SGML 规范几乎未曾变更过，是一种相当成熟高度稳定的通用标记语言。SGML 的设计目标就是要制作跨平台的电子文档，可以在不同的计算机硬件或操作系统上使用，甚至可以被不同的应用软件使用。SGML 作为国际标准已经使用了几十年，已有许多支持其格式的应用软件和相关数据转换技术，因而 SGML 文件可以在各个应用领域中被广泛采用，表现出 SGML 的高度可携性。SGML 规范考虑得相当全面，与 SGML 配套使用的家族如 HyTime 和 DSSSL 也都是国际标准，反映了 SGML 的高完整性。

然而，SGML 的稳定和完整也带来 SGML 的致命弱点，即高度复杂和费用昂贵。为了保证稳定和完整，其复杂性大大增加。从了解 SGML 的整个规范，到根据它来制定应用领域的 DTD，都是十分费时费力的事情，从而使开发 DTD 费用昂贵，也使开发 SGML 相关软件变得复杂。

标准通用标记语言 SGML，是 HTML、XML 和 XHTML 的老祖宗，是一个在 1986 年就已经制定的 ISO（国际标准化组织）的标准，ISO8879。

在 20 世纪 80 年代末，SGML 引起了包括 CERN 在内的一些组织的注意，1990 年，World Wide Web 的发明者 Tim Berners Lee 选择了 CERN 使用的一组 SGML 的 DTD 标记标签，在最早的 Web 浏览器和编辑器 NEXUS 中他使用了这些标签和样式表进行排版，并增加了最重要的功能——链接，这就是 HTML 的前身，基于 SGML 的 HTML 是让 SGML 走向 World Wide Web 的第一步。

1.2.3 超文本标记语言 HTML

1．HTML

超文本（hypertext）和超媒体（hypermedia）是信息管理的一种方法，用这种方法数据被存储在由链接联接的节点构成的网络。节点可以是文本、图形、音频、视频以及源代码或其他格式的数据。如果信息主要是文本形式的，则被视为超文本；如果还包括视频、音乐、动画以及其他元素，则被视为超媒体。超文本的基本特征，是机器支持的链接的概念（包括在文档内部和文档之间）。正是这种链接能力，允许文本的非线性组织（参见第 6 章）。

超文本标记语言（Hypertext Markup Language，HTML），是一种建立超文本/超

媒体文档的标记语言，它用标签标记文档中的文本及图像等元素，指示浏览器如何显示这些元素，以及如何响应用户的行为。HTML 是 SGML 的一种应用。

HTML 最初是由 Tim Berner Lee 在 CERN 时开发，90 年代随着 World Wide Web 的爆炸性增长和 NCSA 开发的 Mosaic 浏览器而流行于世。HTML 在它短暂的生命期间已经过多次反复。第一版本，HTML 1.0，出现在 1990 年。而一个非正式版本，HTML+，在 1993 年后期被引进，包含了 78 个元素，其中许多不再留在 HTML。许多废弃的元素例如 abstracts（摘要）、notes（笔记）和 bylines（标题下署名之行），定义了文档部件。HTML 2.0 在 1994 年发布，它有一个正式说明，并且是官方标准的第一个版本（参见 RFC1866）。这个版本包含了 49 个元素。在 1995 年 3 月，出现了 HTML 3.0，增加了一些新特色。元素 NOTE 是 HTML+的一个特色，但是从 HTML 2.0 中消失了，在 HTML 3.0 中重新出现。表单 FORM 元素的属性出现在这个版本中。在 1996 年 5 月问世的 HTML3.2 被认为是 HTML 2.0 的真正继承者（换句话说，绕过 HTML 3.0）。HTML 3.2，它增加了 19 个新元素，保持了来自 HTML 3.0 的表格和文本流属性并且综合了许多 Netscape 扩展（见下述）。HTML 4.0 版本，在它的开发期间被冠于代号 Cougar（美洲狮），增加了对 object 元素的支持，这是一个有力的图像和多媒体嵌入元素。另外，HTML 4.0 支持层叠样式单，改进了填充表单和表格、客户方脚本、国际化（即识别那些由特殊字符字母表组成的语言，可以从右到左方向读），以及用于数学和高级出版的其他特别字符。HTML 的最后版本是 4.01。

2．微软、网景和 WebTV 的扩展

从 HTML2.0 开始，微软公司和网景公司已经开发了被他们的浏览器支持的专有元素。这些被称之为微软和网景的扩展。微软开发者创造了 object 元素、帧和一些表格元素。网景创造了 font、center、big、small、sub 和 sup 元素，帧，以及客户方图像地图。所有这些扩展已经被加到 HTML。另外，微软和网景给已有元素添加了新属性。

WebTV 网络服务是一种基于电视的浏览器，用这种浏览器能浏览互联网，访问电子邮件服务，并且随意地与某些电视频道和节目交互。WebTV 浏览器支持 HTML 3.2 的元素和属性，微软扩展的 bgsound、embed 和 marquee，以及网景扩展的 embed、nobr 和 noembed。另外，WebTV 开发者创造了两个独有的 WebTV 扩展：audioscope 和 blackface。

3．HTML 存在的问题

（1）规则不严。HTML 所存在的一个问题是它让网页制作者在使用标记时有太多的自由。例如，网页作者可以使用段落标识符来标明段落的开始，但是却并不一定需要在段落的末尾使用段落结束符。HTML 文档中的标签甚至交叉使用也不报错，比如。子标题标签<hn></hn>（n=1-6）往往被当作显示格式使用，如