环球网与超媒体信息检索

沈艺 （南京师范大学）

摘要：本文分析了环球网的运行机制及超媒体信息的检索方法，介绍了环球网的发展和使用，最后讨论了目前环球网存在的问题。

一、引言

环球网 World Wide Web（简写 WWW 或 W3 或 Web）是当前 Internet 上最受欢迎、最为流行、最新的信息检索服务系统。它把 Internet 上现有信息资源全部连接起来，使用户能够在 Internet 上查找已经建立了 WWW 服务器的站点（Site）所提供的超媒体或超文本信息资源。WWW 把各种类型的信息（静止图象、文本、声音和影像）天衣无缝地集成起来，并提供图形界面方式下快速查找，使用同样的图形界面还可与 Internet 上其他服务器对接。

WWW 为全世界提供了查找和共享知识的手段，是世界上各种组织机构、科研机关、大专院校、公司厂商甚至个人热衷于研究开发、共享的知识集合。同时也是人们进行交互多媒体通讯的一种动态格式。它为网络上的用户提供一种兼容的手段，以简单的方式去访问各种媒体，是第一个真正全球性超媒体网络。

二、WWW 的发展

WWW 产生的历史并不长，至今也不过五、六年的时间。早在 1989 年 3 月，欧洲粒子物理实验室（CERN）的科学家 Tim Bemers－Lee 首先提出环球网 WWW 这一概念，并把它作为高能物理学界科学家传输新想法、新成果的工具。到了 1990 年末，第一个环球网软件在 NeXT 计算机上实现，该软件能够让用户在 Internet 上查阅、传输超文本文档，并具有编辑超文本的功能。1990 年，CERN 公开发表了环球网 WWW。

WWW 一经推出，就引起了广泛的注意。一些人在自己的主机上创建自己的 WWW 服务程序，以便他们的信息能够用于 Internet，并开始研制 WWW 客户机，设计易于使用 WWW 的图形界面。到 1993 年末，针对不同类型的计算机系统（X－Windows、PC/Windows、Apple Macintosh）的浏览程序（browser）相继开发出来。1994 年夏天，WWW 已成为访问 Internet 资源的最流行的手段。现在，WWW 服务器正以每年 2000％以上的速度增长，总数已超过 1.5 万台。我国于 1994 年春已正式建立了 Internet 上的 WWW 服务器，它们主要集中在北京。

三、WWW 运行机制

WWW 利用好几种协议去传输和显示驻留在世界各地计算机上的多媒体信息源，它与 WWW 服务器一起工作，为 Internet 提供“分布式客户机/服务器”的运行环境。WWW 的客户机是指在 Inernet 的一个站点上请求 WWW 文档的用户计算机。WWW 服务器则是指 Internet 上保存 WWW 信息的计算机，它利用超文本传输协议 HTTP（Hyper Text Transport Protocol）允许用户在客户机上发出请求，访问超文本或超媒体信息。

WWW 采用分布式运行，客户程序可以在与服务器完全分开的计算机上运行，服务器可能在其他房间，也可能在其他国家。客户机和服务器有所分工，其中文档存储的任务交给了服务器，文档显示的任务就留给了客户机。WWW 客户机与服务器之间进行通信所用的共同语言是超文本传输协议（HTTP）。

用户要访问 WWW，就必须在他的客户机上运行 WWW 程序，它知道如何去解释和显示在 WWW 上找到的超文本文档。这是由于超文本包含一些借用标题、章节本身等构造文本的命令，从而允许浏览程序格式化每一种文本类型。例如：在 PC 机上可以采用浏览程序 Mosaic for Windows。此外，浏览程序还可以访问超媒体，只要在 PC 机上装有声音卡及驱动软件，就能听到包含在 WWW 超媒体里的声音片断。

有的浏览程序还可以自动调用其他应用程序，以显示特殊类型的文档。例如，若 WWW 文档包含对以 Mi-

crosoft－Word for Windows 格式的文档访问时，就可以让浏览程序自动调用 Word for Windows，以显示相应格式的文档。有的浏览程序（如 Mosaic）还具备访问 Internet 上其他类型服务器的功能，如无记名 FTP、Gopher、WAIS 以及 Usenet news 等。

四、超文本与超媒体

WWW 一般可用超文本作为和用户交互的基本手段，超文本指的是计算机内的一种文档。用户在阅读这种文档时，从其中的一个地点移向另一个地点，或从一个文档移向另一个文档都是按非线性或者说非顺序方式进行的。也就是说，用户不是按从头到尾顺章逐句的传统方式去获取信息。这是由于超文本包含着可用作连接（Link）的一些字或短语（一般用下画线或不同的颜色标明，或图标，用户只需用鼠标在其上轻轻一点，就能立即跳至与当前正在阅读的文档相关的新地点或新文档。

超媒体是超文本的自然扩展，是超文本在内容形式上的一种进步，是超文本与多媒体的组合。在超媒体里连接不只是连到文本文档，还可以连到其他形式的媒体，如图形图像、声音或影视动画等。这样，超媒体就把死板的文档变成活生生的文档，把个人计算机变成了多媒体设备，比音响、电视更加生动。

设计 WWW 的一个目的是为了能很容易地检索到 Internet 上的文档，而不管这些文档是在什么地方。当决定超文本作为 WWW 文档的标准格式后，人们制定了能够快速查找这些超文本文档的协议，该协议所检索的文档包含用户进一步检索的连接。当然，在查找 WWW 文档时，没有必要知道 HTTP 的具体内容。但是，如果您有兴趣的话，可以在下述的 URL 地址上找到 IETF http 技术说明的副本。

http://info. cern. ch/hypertext/WWW/Protocols/HTTP2.html

查阅 WWW 文档时，在屏幕上显示的文本都是非常漂亮的格式化文本。之所以能做到这一点，就是因为在书写这类文档时，利用了超文本标记语言 HTML。HTML 是一组非常简单的命令，这组命令描述了 WWW 文档是如何构造的。它只是定义了文档的各个组成部分，并未真正进行文档的格式化，格式化过程是在运行了浏览程序时才完成。

五、信息检索

当用户把自己的计算机与远程计算机连接时，若采用 FTP 协议，则连接是持续的，即一旦线路接通并且使用，那么直到任务全部结束之前，别人是不可能使用这条线路的。若采用 WWW 的浏览程序 Mosaic，则连接是非持续的，最初它只检出原始信息，然后就很快撤销连接，让出线路给他人所用。仅当需要把更详细的信息传送到客户时，才重新打开连接。这种方式的连接只占用几分之一秒的时间，对 Internet 上的有限资源带来的影响最少。

为了访问 Internet 上各种类型文档，WWW 设计者开发了一种工具叫统一资源定位器 URL（Uniform Resource Locator），URL 完整地描述了 Internet 上超媒体文档的地址。这种地址可以是本地磁盘，也可以是 Internet 上的站点。地址访问可以是相对的，也可以是绝对的。在相对方式下，假定主机名和路径名就是当前正在使用的名字，只要指出子目录名和文件名即可。绝对方式则应包括完整的主机名，路径名和文件名。URL 不仅限于描述 WWW 文档的地址，还可以描述其他服务器（无记名 FTP、Gopher、WAIS、Usenet news 和 Telnet）的地址，典型的 URL 地址格式如下：

http://WWW. net. edu. cn/inet/inder/html + 其中，“http”代表是检索文档的 HTTP 协议，它规定如何使用 Internet 上特定的服务器。

“//”表明其后跟的是 Internet 上的有效宿主机名。

跟在宿主机名之后的是用户要查找文档文件的 UNIX 风格的路径名和文件名。

上面 URL 地址的意思是：告诉 Mosaic，利用 HTTP 协议，在 Internet 的宿主机 WWW. net. edu. cn 上的 inet 目录下查找文件 index. html（注：该宿主机是清华大学的 WWW 站点服务器）。

再如，要想查阅 1995 年 4 月在德国召开的第三次 WWW 国际大会信息，就可以键入它的 URL 地址：

http://WWW.igd.fhg.de/WWW95.html

其中主机名 WWW. igd. fhg. de 的意思是德国弗郎霍费学会计算机图形研究所 WWW 服务器，文档 WWW95. html 是用超文本标记语言写成的超文本文档，内容是关于 WWW'95 大会的。

当启动 Internet 上某个 URL 地址上的文档时，告诉 Mosaic 首先要显示的那个文档，叫起始页 Home Page。使用 WWW 的每个用户都可以建立自己的起始文档。在该文档中，可以加入表征用户特点的图形或图象，列出最常用的一些连接。对于经常使用相同资源的集体和单位，如一家公司、研究所或大学，都可以设计自己的起始

页。.

下面介绍一下检索我国 WWW 服务器上信息的几种方法：

1.在 Mosaic 的 File 菜单下，选取 Open URL，在所出现的对话框中键入：

hHP://WWW. ihep. ac. cn/china-WWW. html 就可以显示出中国的起始页。

2.也可以在 URL 地址。

hHp://WWW. W3. org/hypertext/Data Source/WWW/servers. html

所显示的起始页上读取"China"，获得同样的结果。

3.还可以在 URL 地址：

hHp://Wings. buffalo. deu/World

显示的世界地图上读取"Asia"，接着读取"China"，来显示起始页。目前，在中国起始页上列出的 WWW 服务器包括：

(1)北京化工大学

http://www.bat.edu.cn/

(2)中国互连网络 China Net

http://www.cnc.ac.cn/

(3)中国教育和研究网络 CERNET

http://www.cernet.edu.cn/

(4)中国科学院高能物理研究所 IHEP

http://utkvxl.utk.edu/~xurs/ihep.html

(5)中国微生物信息网络

http://www.imnet.ac.cn/

(6)北京大学

http://www.net.edu.cn/

(7)清华大学

http://www.net.edu.cn/

(8)电子杂志《神州学人》

http://www.china.edu.cn/

六、WWW 存在的问题

环球网自产生以来，经过短短几年的发展，已有相当的规模。但在使用中仍存在以下一些问题：

1.世界上没有一个人、公司或机构声称他拥有 WWW，因此环球网缺乏一个监督机构。

2.多媒体信息——声音、图片、影视需要巨大的空间开销，如果使用传输速率低于 14.4K bps 的调制解调器，把多媒体信息从宿主机传到客户机，等待时间是无法忍受的。

3.由于许多编写 WWW 文档的人员并不是专业的多媒体开发人员，因此他们开发的许多文档都包含着大量的图象连接。这就造成用户把全部图象装入到客户机之前，用户就不可能在起始页上访问任何连接。

4.WWW 的用户依赖文档或服务器的提供者去修改他们的材料时，连接到的信息可能就是过时的，或者是连接到一台脱机的服务器上，从而造成出错。

七、结束语

环球网使得信息检索变得快速、高效、直观，给用户带来了巨大的效益。但是，为了使环球网更加完善，还需要环球网的开