PDF文档在Web上出版技术研究

琼州大学计算机中心　林　天

本文详细讨论了PDF文档的特点，并对Web服务器对PDF文档的支撑环境、PDF文档与应用程序的参数传递、表单数据格式FDF的基本组成结构等方面进行探讨。最后通过实例给出应用ASP技术实现PDF与Web数据库集成的过程，该项技术对PDF的网络出版、电子商务票据生成有很好的应用前景。

PDF文档概述

1.HTML与PDF

HTML是SGML(Standard Generalized Markup Language)的一个应用，是目前Internet上主要的Web出版形式。它通过各种标记描述出Web页面基本的样式，图文并茂，并有交互及超级链接功能。它可结合JavaScript、CSS等脚本语言增强客户端页面的表达及交互能力，可配合CGI、NSAPI、ISAPI、ASP、Java JDBC等各种接口技术实现与服务器的交互和Web数据库的访问。但HTML文本有一个明显的缺点是在不同的浏览器、不同的显示器尺寸上显示或打印常常会出现版面不一致的情况。

PDF(Portable Document Format)是Adobe(R)公司公布用于进行全球电子文档分发的开放式标准，最新版本为1.2版。PDF是从在印刷领域具有统治地位的页面描述语言PS（PostScript）发展而来，具有与PS一样精美印刷版面的描述能力和相似的描述方法。PDF可通过打印的形式、附加在电子邮件中、位于Web服务器上、或刻录到CD-ROM上等多种方式发行。任何人都可利用免费提供的Adobe Acrobat Reader软件或安装了Acrobat Reader插件(Plug-ins)的网络浏览器随意地共享、查看、导航和打印PDF文档。从而真正实现了一次制作多处使用，纸张印刷与电子出版的统一的思想。PDF具有如下特点:

- PDF是一种文件结构，而PS是一种编程语言。PDF具有比PS更高的处理效率。
- PDF文件象HTML一样可包含超级链接、表单、JavaScript等交互特性，也支持声音、动画。
- PDF支持与Windows目录树窗口极为相似的多级链接菜单控制方式，即书签。而HTML页面实现该控制方式较困难。
- 支持对页面的随机存取。
- 支持多种压缩、编码方式，文件更紧凑。压缩、编码方式有：ASCIIHex、LZW、RunLength、CCITT Group3、CCITT Group4、JPEG、flate等。
- 支持各种不同级别的安全特性。如数字签名；可阅读可打印但不能修改；可阅读不可打印等。这种安全性控制对保护电子出版物版权非常重要。
- PDF具有软硬件平台的无关性。即用户在不同的操作系统、不同的浏览器、不同的显示器尺寸上显示或打印的版面均保持一致。

PDF不仅具有HTML的优点，还克服了它的缺点。可以预见，随着目前各种宽带网技术在Internet上的应用，PDF将在Web出版中占有重要位置。

2.PDF的应用流程

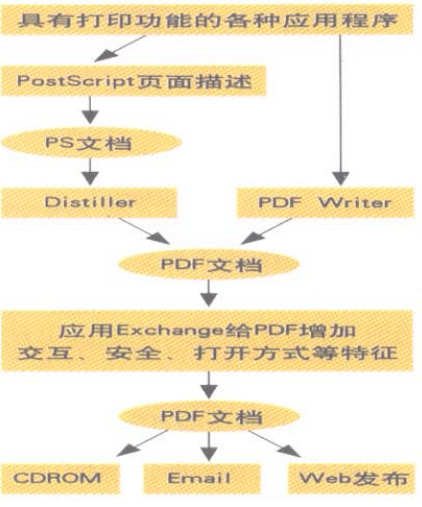

图 1

对于具有打印功能的各种应用程序，使用 Acrobat Distiller 或Acrobat PDF Writer软件可将任何文档（甚至是所扫描的纸面内容）转换为 精美的PDF 文档。无论创建源文档的应用程序和平台如何，PDF 均保留原文档的字体、格式、 颜色和图形等显示与打印的一致性。

使用 Acrobat Exchange 3.0、 Adobe Acrobar4.0 软件，可将超级链接、书签、活动表单等交互元素嵌入 PDF 文档，以增强阅览的灵活性，并方便实现独立或与 HTML 混合在 Web 上发布信息。应用流程如图 1 所示。

PDF 在 Web 上出版

1.Web 对 PDF 文档的支持

对于存放在 Web 上的 PDF 文档，使用含 Acrobat Reader 插件(Plug-ins)的网络浏览器都可阅读。但是，如果 Web 支持 PDF 文档一次下载一页(page-at-a-time)的字节流服务(byte range serving)，且浏览器端 Acrobat Reader 设置了允许后台下载选项(缺省值)，显示 PDF 文档首页面的速度将大大提高。否则，将等待整个 PDF 文档下载后才显示其内容。

目前支持 PDF 文档一次下载一页(page-at-a-time)的 Web 服务器产品有:

- Microsoft IIS 3.0
- Enterprise Server 2.0
- FastTrack Server 2.0
- Website 1.1
- WebSTAR 2.0; 等。

例如，设置 Window NT4.0 注册表:

HKEY_LOCAL_MACHINE
\ SYSTEM
\ CurrentControlSet
\ Services
\ W3SVC
\ Parameters

AcceptByteRanges 1

Microsoft IIS 3.0 将支持字节流服务。如果是不支持这种服务的 Web 服务器，也可以通过 CGI 应用程序来实现。

2. 在浏览器中调用 Web 上的 PDF 文档

方法一，通过 URL 直接访问。例如:

http://www.qzu.edu.cn/pdf/lzfq.pdf

方法二，在HTML网页中通过超级链接来调用。例如:

 黎族风情

方法三，在 HTML 网页中通过嵌入对象标签 <EMBED> 来调用。例如: <embed src=lzfq.pdf width=60% height=200>

在IE3.0、Netscape Navigator3.0以上版本浏览器中，方法一与方法二调用的 PDF 文档均在整个窗口显示，而方法三将限定在一定范围中显示。但是，方法三在 Netscape Navigator3.0 中仅可显示首页，且无PDF 浏览工具条出现。

如果希望随机访问某 PDF 页面并控制其显示方式，可在链接调用中加入 page、view、pagemode 等属性。例如:

 黎族风情

此外，PDF 文档本身也可以通过交互元素调用其他 PDF 文档和 HTML 网页。

3.PDF 文档与应用程序的参数传递

在 PDF 文档中，通过表单域(field)收集的信息，可以通过事件驱动行为 Submit form 实现向服务器端应用程序传递参数信息。在定义 Submit form 行为时，需设置服务器端应用程序的 URL，设置输出格式为 HTML form (URL encoded)，并可随意选择输出的域。

在服务器端CGI应用程序中，将通过PDF 文档表单域的域名来获取参数信息。例如，ASP 程序中获取 PDF 文档表单域信息的语句如下:

Request("field_name")

PDF 与 Web 数据库的集成

FDF(Form Data Format)是 PDF 文档表单域数据输出(Export)和导入 (Import) 的一种存储格式，它也是实现 PDF 与 Web 数据库信息交流的桥梁。下面详细讨论 FDF 文件的基本组成结构，并通过实例给出结合 Windows NT 4.0 IIS3.0的ASP (Active Server Pages) 技术实现将浏览器端PDF表单数据提交给 Web数据库及从 Web数据库读取记录动态生成 PDF 文档的过程。该项技术对 PDF 的网络出版、电子商务票据生成有很好的应用前景。

1.FDF 文件的基本组成结构

FDF 是一种纯文本文件。FDF 文件由文件头(Header)、文件主体(Body)、交叉引用表(Cross Reference Table)、文件尾(Trailer)四个部分组成。

FDF文件头的首行表明了当前文件所使用的FDF规范版本。例如%FDF-1.2 。

FDF文件主体主要由Catalog对象组成。Catalog对象仅含一个关键字FDF。关键字FDF 的值包含下列项目:

- Fields 包含一组域属性的定义;
- F 指定Form输出或导入数据的PDF文件名; 等。

其中，Fields的属性定义包含下列描述:

- T 域名;
- V 域值;
- Opt 选择项域的各项值; 等。

FDF文件尾由关键字trailer、间接引用FDF文件体的Catalog对象关键字Root的一个值对及文件结束标记%%EOF组成。

FDF文件书写格式如下所示:

```
%FDF-1.2
1 0 obj
<< /FDF
<< /Fields [
    << /T (Field Name) /V (Value) >>
    << /T (Field Name) /V / Value /Opt [(Item1)
,(Item2),…] >>
      ……
     ]
/F (PDF file Name)
>>
>>
endobj
trailer << /Root 1 0 R >>
%%EOF
```

2. 提交PDF表单数据给Web数据库

假设在Web上建立Access数据库test_fdf.mdb,它包含Name、Sex、Birthday、Nationality、Marry、Salary字段。

Submit Form to Web

Name	lintian	Sex	Male ◉ Female ○
Birthday	1962/09/15	Na.	Nationality
Marry	✓	Salary	1277.50

Reset Snbmit

图 2

定义图2所示含有表单的PDF文档，文件名为Form.pdf ，表单中各域名与数据库字段名相对应。 其中，按钮Submit通过鼠标事件Mouse Up驱动行为Submit form实现向服务器端应用程序传递参数信息。在定义Submit form行为时，设置服务器端应用程序的URL为 Submit.asp，设置输出格式为HTML form (URL encoded)，并选择输出的域名。

在服务器端ASP应用程序中，将通过PDF文档表单域的域名来获取参数信息，VBScript语句格式为

Request("field_name")

接收PDF表单提交的数据并写入基于ODBC的Web数据库的Submit.asp程序代码略。

3. 从Web数据库读取记录生成PDF文档

假设定义图3所示含有表单的PDF文档，文件名为Card.pdf ，表单中各域名与数据库字段名相对应。希望表单向Web提交Name值时，能从Web数据库中获取其它相关信息。其中，按钮Submit通过鼠标事件Mouse Up驱动行为Submit form实现向服务器端应用程序传递参数信息。在定义Submit form行为时，设置服务器端应用程序的URL为 getData.asp#FDF。这里字符串'#FDF'是必须的，以确保服务器端将返回FDF格式。需设置输出格式为HTML form (URL encoded)，并选择输出的域名。

Get Data From Web

Name	lintian	Sex	Male ○ Female ○
Birthday		Na.	Nation
Marry		Salary	0

Reset Snbmit

图 3

在服务器端ASP应用程序中，从Web数据库读取记录后，将自动生成FDF格式文本返回浏览器端接收FDF数据的PDF文档。在ASP程序中，在生成FDF格式文本代码之前必须含定义服务器端的MIME类型为application/vnd.fdf的语句。

从Web数据库读取所需记录并动态生成PDF文档的getData.asp程序代码略。■

参考文献

1 Adobe System Inc. Exchange online Guide

2 Adobe FDF ToolKit Overview Technical Note#5194 , August 10,1999