PostScript 文件与 PDF 文件间数据转换

孙 殷[1],王 鹏[2]
(1.浙江商业职业技术学院,浙江 杭州 310053;
2.浙江工业大学,浙江 杭州 310014)

摘 要:针对可变数字印刷中常用的页面描述语言及其文件格式,通过研究 PostScript 文件结构和 PDF 文件结构,介绍了如何实现 PostScript 文件与 PDF 文件间的数据转换,给出两种文件间转换算法流程图,并利用 PostScript 解释器 Ghostscript 提供的 API 接口,实现了 PostScript 文件和 PDF 文件间的转换。转换结果表明,该转换算法转换效果比较好,实现了所见即所得。

关键词:PostScript;PDF;数据转换
中图分类号:TP311 **文献标识码**:A **文章编号**:1674-7720(2013)11-0019-03

The data conversion of PostScript files and PDF files

Sun Yin[1], Wang Peng[2]
(1.Zhejiang Business College, Hangzhou 310053, China;
2.Zhejiang University of Technology, Hangzhou 310014, China)

Abstract: Considering the usual page description language and file format in variable digital printing, this paper introduces how to realize the conversion between PostScript files and the PDF files base on their structure. Then gives the conversion flow chart and realize the conversion by using the API interfaces of PS interpreter Ghostscript . The conversion results show that this conversion effect is better, realize what you see is what you get.

Key words: PostScript; PDF; data conversion

在可变数据印刷系统中,每个电子页面的内容(如文字、图形、图像)经过页面描述语言合成后产生相应的可变数据印刷页面,最终形成特定的数据文件。PostScript 文件(简称 PS)和 PDF 文件是目前存储可变数据印刷页面常用的文件格式,因此,研究这两种文件格式之间的数据转换对可变数据印刷的发展具有一定的意义。而当前 PS 文件与 PDF 文件之间的转换主要有两种方式[1]:(1)通过专业的软件(如 Acrobat Distiller)转换;(2)通过虚拟打印机来实现。

但是这两种方式都不利于自主研究数据印刷系统的开发。本文根据 PS 文件格式和 PDF 文件格式的特点,利用 PS 解释器 Ghostscript 提供的 API 接口,通过 VC6.0 实现了 PS 文件和 PDF 文件的相互转换,并给出相应的代码。

1 PostScript 和 PDF 文件结构

1.1 PostScript 文件

PostScript 是由 Adobe 公司开发的页面描述语言,其最大特点是可以将印刷品中包含的文字、图形、图像、字体和颜色等各种元素用一种计算机数据来表现和描述,然后经过 RIP(光栅图像处理器)快速地解释为可控制打印设备输出用的点阵信息。用 PostScript 语言所描述的页面文件称为 PS 文件,其后缀名为.ps 或.eps。

PostScript 程序可访问的所有数据都是以对象(Object)形式存在的[2],对象由操作符产生、管理和撤销。常用的对象有整数和实数、布尔型、数组、压缩数组、串、名字、字典等[3-4]。程序中的数据存储在堆栈中并通过堆栈被操作符管理执行。

PostScript 语言解释器对语句的执行是逐句解释执行,控制比较灵活,一切操作均通过堆栈进行。例如用粗线画一个圆:

```
%%Title:用粗线画一个圆
/inch{72 mul} def
4.25 inch 5.5 inch
```

```
2.5 inch
0 360 arc
1.75 inch setlinewidth
stroke
showpage
```

以‘%’开头的第一段语句表示注释，第二段语句定义了名字对象‘inch’，接着定义圆的中心位置(4.25，5.5)，半径2.5；然后画角度为360°的圆；最后定义粗线的宽度1.75，画出粗线的路径进行显示。

1.2 PDF文件

便携式文件格式PDF(Portable Document Format)是Adobe公司继PostScript后于1993年推出的一种电子文件格式[5]。它具有能够完整地保存任何原文档中的文字、格式、颜色、图形、可加密等优秀特性，广泛应用于数据印刷系统中。

PDF文件主要由四部分组成：文件头(Header)、文件体(Body)、交叉引用表(Cross-reference table)和文件尾(Trailer)。PDF作为一种结构化的文件格式，它是由一些具有特定数字标号的“对象”的模块所组成。其文档结构是一种树形结构，通过文件尾(Trailer)可以找到文件体的根对象Catalog[6]，根对象包含PDF文档的大纲(Outline)、页面组对象(Pages Tree)等。文档结构具体层次关系如图1所示。

其中页面对象(Page)作为PDF中最重要的对象，包含了该页面的文字、图片、页面大小等信息。页面中包含的信息是包含在一个称为流(stream)[7-8]的对象里，这个流的长度(字节数)必须直接给出或指向另外一个对象。

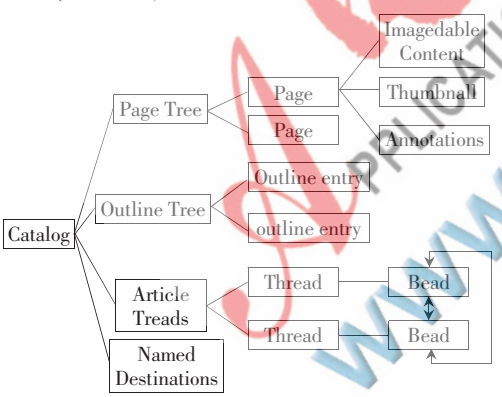

图1 PDF文档结构

2 文件格式转换算法的实现

根据前两节的PostScript文件结构可知，要实现PS文件到PDF文件的转换，首先需要对PS文件中的不同信息(文字、图形和图像等)进行解析，然后转换为对应的PDF对象。两者间转换的总流程主要为：

(1) 导入要转换的PostScript文件，初始化PS解释器；

(2) 开始扫描PS文件，记录当前代码段所在页。判断页信息是否已经到末尾，是则退出，否则继续向下扫描；

(3)读取PS页面描述信息，对其中的文本信息、图形信息和图像信息分别进行提取并处理；

(4)将第三步中处理的文本、图形和图像信息进行重构，然后分别转换为相对应的PDF对象；

(5)判断PS文件是否扫描结束，是则回到第(2)步，否则回到第(3)步；

(6)转换算法结束。

图2为PS文件转换为PDF文件的总流程图。

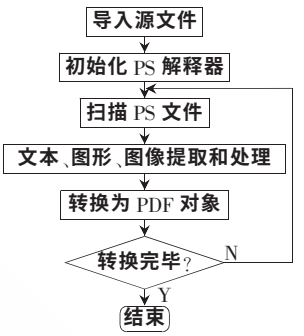

图2 转换总流程图

由于PostScript文件对文本、图形和图像的描述机制各不相同，因此，其相应的信息提取算法也不尽相同。在PostScript文件中，图像是取样值的矩形数值，每个取样值表示某种彩色。按行或者列扫描图像矩形所得的一串取样数据定义了一个图像。除了矩形数组之外，PS程序中还包括一些图像参数：源图像的格式、图像取样数据的数据源、图像空间坐标等。因此，对PS文件中图像的提取主要是将图像的参数和图像取样数据进行提取，其对应的提取算法流程图如图3所示。

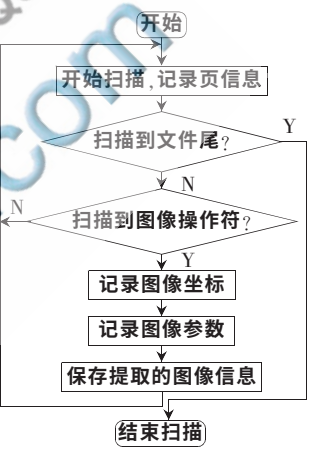

图3 图像提取流程图

在标准的PostScript文件中，文本大都是以字库[9]的形式进行保存，因此在PostScript中的字符可以由PS解释器根据字库提取出对应的字符。对PS文件中文本信息的提取流程如图4所示，主要过程如下：

(1)扫描文档，记录当前页面信息。判断是否已扫描到文件尾，是则跳转到第(5)步，否则跳到第(2)步；

(2)继续扫描，判断是否扫描到文本提示符，是则跳到第(3)步，否则跳回第(1)步；

(3)根据文本提示符获取相应的字库词典，查找字库获取PS所描述的字符；

(4)保存第(3)步提取的文本信息；

(5)结束文本扫描。

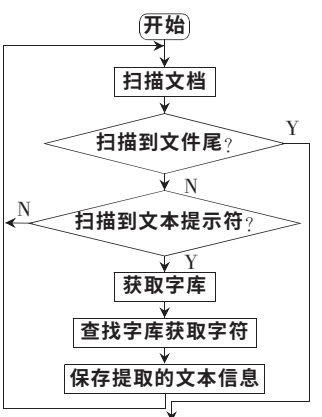

图4 PS文件中文本提取算法流程图

扫描文件时通过寻找关键操作符函数FindStringInBuffer()，可以得到需要的操作符位置。源程序如下：

```
size_t FindStringInBuffer (char*buffer,char*search,size_t buffersize)
{
    char*buffer0=buffer;
    size_t len=strlen(search);
    bool fnd=false;
    while (! fnd)
    {
      fnd=true;
      for(size_t i=0;i<len;i++)
      {
        if(buffer[i]! =search[i])
        {
          fnd=false;
          break;
        }
      }
      if (fnd) return buffer-buffer0;
      buffer=buffer+1;
    if (buffer-buffer0+len>=buffersize)
    return-1;
    }
    return-1;
}
```

由于PS解释器的实现比较复杂，本文通过利用PS解释器Ghostscript提供的API接口函数对PS文件进行解释转换。首先通过接口函数gsapi_new_instance()新建一个要转换的实例。转换过程主要在接口函数gsapi_init_with_args()中进行，即初始化PS解释器并将输入PS对象转换为对应的PDF对象。主要转换程序如下：

```
if(code=gsapi_new_instance(&minst, NULL))
{
    printf("Can't create Ghostscript instance\n");
    return 1;
}
code=gsapi_init_with_args(minst,gsargc,(char**)gsargv);
code1=gsapi_exit(minst);
if((code==0)||(code==e_Quit))
code=code1;
gsapi_delete_instance(minst);
if((code==0)||(code==e_Quit))
  return 0;
```

3 效果及结论

本文的程序在VC6.0上编译通过，并可以将输入的PS文件(my.ps)转换为PDF文件(my.pdf)。转换前后的结果如下：

由图5和图6可以看出，转换前后的图像几乎完全一样。由图7中文本的转换结果可以看出，转换的字符内容一样。通过修改输入文件名和输出文件名类型，例如输入文件为.pdf格式文件，输出文件为.ps文件，也可以实现PDF文件到PS文件的转换。因此，在设计开发可变数据印刷系统时，可以将该程序作为数据转换的一部分嵌入到印刷软件系统中，具有一定的实用性。但是，由于PS解释器并非独立设计，受到其接口函数的限制，程序的延伸性不是很好。

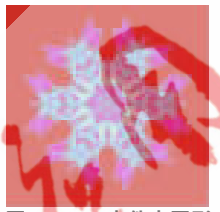

图5 my.ps文件中图形

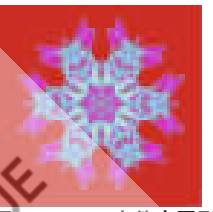

图6 my.pdf文件中图形

图7 文本转换结果

参考文献

[1] 张志伟，孔凡让，吴欣.PostScript格式文献中数学表达式的提取方法[J].计算机应用与软件，2008，25(11)：157-159.

[2] USA Adobe Systems Inc. PostScript language reference mannal. Addison Wesley Protessional，1990.

[3] 何明，匡燕玲，李小龙，等.页面描述语言PostScript及其转换程序[J]. 北京工业大学学报，2004，20(4)：102-104.

[4] 徐福培，潘志庚.页面描述语言及其程序设计[M].南京：南京大学出版社，1994.

[5] USA Adobe Systems Inc. Document management-Portable document format-Part 1:PDF 1.7 [J]. Adobe Systems Incorporated，2008，PDF 32000-1.

[6] 李珍，田学东.PDF文件信息的抽取与分析[J].计算机应用，2003，23(12)：145-148.

[7] 王婉，韩逸秋，徐福培.PDF文件格式及其向PS文件转换的研究[J].计算机科学，2001，28(9)：123-127.

[8] 吴一民，朱檬，罗绵川.基于.NET平台PostScript文件解析标引系统设计与实现[J].微计算机应用，2009(10)：5863.

[9] 段华伟，黄灵阁.计算机文字处理技术现状[J].印刷质量与标准化，2004(5)：39-41.

(收稿日期：2013-03-14)

作者简介：

孙殷，女，1984年生，硕士，专任教师，主要研究方向：机电一体化。

王鹏，男，1986年生，硕士，主要研究方向：可变数字印刷。