PDF文件结构（一）

原创 2009年07月07日 15:51:00

标签：stream (http://so.csdn.net/so/search/s.do?q=stream&t=blog) / filter (http://so.csdn.net/so/search/s.do?q=filter&t=blog) / dictionary (http://so.csdn.net/so/search/s.do?q=dictionary&t=blog) / string (http://so.csdn.net/so/search/s.do?q=string&t=blog) / postscript (http://so.csdn.net/so/search/s.do?q=postscript&t=blog) / 文档 (http://so.csdn.net/so/search/s.do?q=文档&t=blog)

24806

PDF文件结构（一）

————物理结构

作者：bobob
邮件：zxbbobob@hotmail.com
(mailto:zxbbobob@hotmail.com)

PDF(Portable Document Format,便携式文档结构)是一种很有用的文件格式,其最大的特点是平台无关而且功能强大(支持文字/图象/表单/链接/音乐/视频等).做PDF的解析，首先要熟悉PDF文件的物理结构和逻辑结构。PDF文件物理结构可分为以下几块:

1. 文件头

文件头是PDF文件的第一行,格式如下:

```
%PDF-1.4
```

这是个固定格式，表示这个PDF文件遵循的PDF规范版本，目前PDF的生成工具，除了官方的acrobat，其他生成的以1.4版本的居多。对于做PDF开发来说，一个最简单的原则就是生成PDF的时候尽量符合低版本规范，以保证大多数解析器能支持；解析PDF的时候尽量支持高版本的规范，以保证支持大多数工具生成的PDF文件。

从1.4版本以后，PDF文件的版本并不唯一的只是在这里表示了，可能后面会改写（catalog的Version词条），所以解析PDF的时候，如果这里的版本大于等于1.4，应该再比较一下catalog里面的version，取其中高一点的版本。

2. 对象集合

这是一个PDF文件最重要的部分，文件中用到的所有对象,包括文本/图象/音乐/视频/字体/超连接/加密信息/文档结构信息等等,都在这里定义。格式如下:

```
2 0 obj
   ...
endobj
```

一个对象的定义包含4个部分：

前面的2是对象序号，其用来唯一标记一个对象；0是生成号，按照PDF规范，如果一个PDF文件被修改，那这个数字是累加的，它和对象序号一起标记是原始对象还是修改后的对象，但是实际开发中，很少有用这种方式修改PDF的，都是重新编排对象号；obj和endobj是对象的定义范围，可以抽象的理解为这就是一个左括号和右括号；省略号部分是PDF规定的任意合法对象(一共8种，见后面附A)。

可以通过R关键字来引用任何一个对象，比如要引用上面的对象，可以使用2 0 R，需要主意的是，R关键字不仅可以引用一个已经定义的对象，还可以引用一个并不存在的对象，而且效果就和引用了一个空对象一样。

3. 交叉引用表

交叉引用表是PDf文件内部一种特殊的文件组织方式，可以很方便的根据对象号随机访问一个对象。其格式如下:

```
xref
0 1
0000000000    65535    f
4 1
0000000009    00000    n
8 3
0000000074    00000    n
0000000120    00000    n
0000000179    00000    n
```

其中,xref是开始标志,表示以下为一个交叉引用表的内容;每个交叉引用表又可以分为若干个子段，每个子段的第一行是两个数字，第一个是对象起始号，后面是连续的对象个数，接着每行是这个子段的每个对象的具体信息——每行的前10个数字代表这个这个对象相对文件头的偏移地址,后面的5位数字是生成号（用于标记PDF的更新信息，和对象的生成号作用类似），最后一位f或n表示对象是否被使用(n表示使用,f表示被删除或没有用)。上面这个交叉引用表一共有3个子段，分别有1个，1个，3个对象，第一个子段的对象不可用，其余子段对象可用。

4. trailer:

通过trailer可以快速的找到交叉引用表的位置，进而可以精确定位每一个对象；还可以通过它本身的字典还可以获取文件的一些全局信息（作者，关键字，标题等），加密信息，等等。具体形式如下:

```
trailer
<<
  key1    value1
  key2    value2
  key3    value3
…
>>
startxref
553
%%EOF
```

trailer后面紧跟一个字典，包含若干键-值对。具体含义如下：

键	值类型	值说明
Size	整形数字	所有间接对象的个数。一个PDF文件，如果被更新过，则会有多个对象集合、交叉引用表、trailer，最后一个trailer的这个字段记录了之前所有对象的个数。这个值必须是直接对象。

Prev	整形数字	当文件有多个对象集合、交叉引用表和trailer时，才会有这个键，它表示前一个相对于文件头的偏移位置。这个值必须是直接对象。
Root	字典	Catalog字典（文件的逻辑入口点）的对象号。必须是间接对象。
Encrypt	字典	文档被保护时，会有这个字段，加密字典的对象号。
Info	字典	存放文档信息的字典，必须是间接对象。
ID	数组	文件的ID

startxref: 后面的数字表示最后一个交叉引用表相对于文件起始位置的偏移量。

%%EOF :文件结束符.

一个PDF文件，都会有上面这样的结构（线性化优化的PDF例外，这个后面单独说）。实际一个pdf文件是很复杂的,但是上面几个部分是确定的,只能多不能少.了解了PDF文件的物理结构，就可以提取出一个一个的对象了.PDF中的对象有8种：

1.booleam

用关键字true或false表示,可以是array对象的一个元素,或dictionary对象的一个条目.也可以用在PostScript计算函数里面，做为if或ifesle的一个条件。

2.numeric

包括整形和实型,不支持非十进制数字,不支持指数形式的数字.

例:

1)整数 123 4567 +111 -2

范围:正2的31次方-1到负的2的31次方

2)实数 12.3 0.8 +6.3 -4.01 -3. +.03

范围:±3.403 × 10的38次方 ±1.175 × 10的-38次方

注意:如果整数超过表示范围将转化成实数,如果实数超过范围就出错了

3.string

由一系列0-255之间的字节组成,一个string总长度不能超过65535.string有以下两种方式:

1) 直接字串

由()包含起来的一个字串,中间可以使用转义符"/".

例:

(abc) 表示abc

(a//) 表示a/

转义符的定义如下：

转义字符	含义
/n	换行
/r	回车
/t	水平制表符
/b	退格
/f	换页（Form feed (FF)）
/(	左括号
/)	右括号
//	反斜杠
/ddd	八进制形式的字符

2) 十六进制字串

由<>包含起来的一个16进制串,两位表示一个字符,不足两位用0补齐

例:

<bob>表示bob的ascii码

〈AAB〉 表示AA和BO两个字符

4.name

由一个前导/和后面一系列字符组成,最大长度为127.和string不同的是,name是不可分割的和唯一的,不可分割就是说一个name对象就是一个原子,比如/name,不能说n就是这个name的一个元素;唯一就是指两个相同的name一定代表同一个对象.从pdf1.2开始,除了ascii的0,别的都可以用一个#加两个十六进制的数字表示.

例:

/name 表示name

/name#20is 表示name is

/name#200 表示name 0

5.array

用[]包含的一组对象,可以是任何pdf对象(包括array).虽然pdf只支持一维array,但可以通过array的嵌套实现任意维数的array(但是一个array的元素不能超过8191)

例:

[549 3.14 false (Ralph) /SomeName]

6.Dictionary

用"<<"和">>"包含的若干组条目,每组条目都由key和value组成,其中key必须是name对象,并且一个dictionary内的key是唯一的;value可以是任何pdf的合法对象(包括dictionary对象).

例:

```
<< /IntegerItem 12
   /StringItem (a string)
   /Subdictionary
    << /Item1 0.4
       /Item2 true
       /LastItem (not!)
       /VeryLastItem (OK)
    >>
>>
```

7.stream

由一个字典，和紧跟其后面的一组关键字stream和endstream以及这组关键字中间包含一系列字节组成.内容和string很相似,但有区别:stream可以分几次读取,分开使用不同的部分,string必须作为一个整体一次全部读取使用;string有长度限制,但stream却没有这个限制.一般较大的数据都用stream表示. 需要注意的是，Stream必须是间接对象，并且stream的字典必须是直接对象。从1.2规范以后，stream可以以外部文件形式存在，这种情况下，解析PDF的时候stream和endstream之间的内容就被忽略掉。

例:

dictionary

stream

… data …

endstream

stream字典中常用的字段如下：

字段名	类型	值
Length	整形	（必须）关键字stream和endstream之间的数据长度，endstream之前可能会有一个多余的EOL标记，这个不计算在数据的长度中。
Filter	名字 或 数组	（可选）Stream的编码算法名称（列表）。如果有多个，则数组中的顺序表示数据被编码的顺序。

DecodeParms	字典 或 数组	（可选)一个参数字典或由参数字典组成的一个数组，供Filter使用。如果仅有一个Filter并且这个Filter需要参数，除非这个Filter的所有参数都已经给了默认值，否则的话DecodeParms必须设置给Filter。如果有多个Filter，并且任意一个Filter使用了非默认的参数， DecodeParms 必须是个数组，每个元素对应一个Filter的参数列表（如果某个Filter无需参数或所有参数都有了默认值，就用空对象代替）。如果没有Filter需要参数，或者所有Filter的参数都有默认值，DecodeParms 就被忽略了。
F	文件标识	（可选）保存stream数据的文件。如果有这个字段， stream和endstream就被忽略，FFilter将会代替Filter, FDecodeParms将代替DecodeParms。Length字段还是表示stream和endstream之间数据的长度，但是通常此刻已经没有数据了，长度是0.
FFilter	名字 或 字典	(可选)和filter类似，针对外部文件。
FDecodeParms	字典 或 数组	(可选)和DecodeParams类似，针对外部文件。

8. NULL

用null表示，代表空. 如果一个key的值为null，则这个key可以被忽略；如果引用一个不存在的object则等价于引用一个空对象.

例:(略)

以上八种对象是按照对象内涵来分的，如果按照对象的使用规则来说，对象又分为间接对象和直接对象。间接对象是PDF中最常用的对象，如前面对象集合里面的，所有对象都是间接对象，在其他位置通过R关键字来引用，在交叉引用表里面都是通过间接对象来引用的。直接对象就更好理解了，上面的8种对象单独出现的时候就叫直接对象。

一个简单PDF文件的结构分析 adolphfend 2014年04月29日 17:24 1550

原文链接：http://blog.csdn.net/pdfMaker/article/details/573990 非常感谢原

(http://blog.csdn.net/adolphfend/article/details/24729025)

PDF格式详解 bobob 2006年05月23日 16:03 31267

pdf(Portable Document Format,便携式文档结构)是一种很有用的文件格式,其最大的特点是 平台无关而且功能强大(支持文字/图象/音乐/视频).今天先讲一下pd...

(http://blog.csdn.net/bobob/article/details/751381)

0

PDF文件格式分析 wzyzzu 2016年01月05日 10:01 849

一、概述： 结构化的文档格式PDF(Portable Document Format)是由美国排版与图像处理软件公司Adobe于1993年首次提出的。Adobe Reader这款pdf阅读器...

(http://blog.csdn.net/wzyzzu/article/details/50460423)

一个简单的PDF文件结构的分析 nncrystal 2013年09月29日 11:34 2160

Adobe的PDF参考告诉我们一个PDF文件可以通过下面4个方面来理解： 1. 对象, 一个PDF文档是由一组基本数据类型组成的数据结构。 2. 文件（物理结构）,...

(http://blog.csdn.net/nncrystal/article/details/12157581)

PDF文件结构（二） bobob 2009年07月07日 15:55 12289

PDF文件结构（二） ————逻辑结构 作者：bobob ...

(http://blog.csdn.net/bobob/article/details/4328450)

PDF文件格式 hetoby 2016年05月10日 18:18 681

PDF 文件格式

(http://blog.csdn.net/hetoby/article/details/51365460)

一个简单PDF文件的结构分析 pdfMaker 2006年01月09日 00:00 33662

一个简单的PDF文件结构的分析Adobe的PDF参考告诉我们一个PDF文件可以通过下面4个方面来理解：1. 对象, 一个PDF文档是由一组基本数据类型组成的数据结构。2. ...

(http://blog.csdn.net/pdfMaker/article/details/573990)

PDF文件结构查看器

2012年12月23日 18:10 1.21MB 下载

PDF文档结构说明 lamb7758 2017年02月20日 16:57 322

目录 一、PDF文件格式---------- 2 1.标准的pdf文档格式------...

(http://blog.csdn.net/lamb7758/article/details/56015876)

pdf reference 格式详细说明 jinshixie 2016年04月08日 13:20 4536

1. PDF概要 1.1. 图像模型 PDF能以平台无关、高效率的方式描叙复杂的文字、图形、排版。 PDF 用图像模型来实现设备无关。图像模型允许应用程序以抽象对象描叙文字、图像、图标，而不是通过具体...

(http://blog.csdn.net/jinshixie/article/details/51095771)

iOS PDF之旅（一）创建PDF文件 u010962810 2013年11月05日 22:02 7320

直接用iOS程序和Quartz 2D创建PDF文件，并在上面添加网络URL和本地文件URL链接。...

(http://blog.csdn.net/u010962810/article/details/14209855)

Android加载pdf格式文件 xiaoqiang_0719 2016年09月28日 14:16 3097

pdf格式在安卓界面是上无法直接展示的，但是遇到了要加载pdf格式的操作并且在界面展示出来，所以必须要想解决的方法查找资源的过程中找到了AndroidPDFview的第三方控件，使用方法： 1、...

(http://blog.csdn.net/xiaoqiang_0719/article/details/52689871)

入门级 PDF 文件格式分析

CreatedSign 2014年01月22日 11:26 2309

一、概述： 结构化的文档格式PDF(Portable Document Format)是由美国排版与图像处理软件公司Adobe于1993年首次提出的。Adobe Reader这款pdf阅读器...

0

(http://blog.csdn.net/CreatedSign/article/details/18657301)

一个简单PDF文件的结构分析

jssyy123 2014年02月14日 14:50 485

一个简单的PDF文件结构的分析 Adobe的PDF参考告诉我们一个PDF文件可以通过下面4个方面来理解： 1. 对象, 一个PDF文档是由一组基本数据类型组成的数据结构。 2. ...

(http://blog.csdn.net/jssyy123/article/details/19199969)

一个简单的PDF文件结构的分析

wzyzzu 2015年12月27日 15:10 417

Adobe的PDF参考告诉我们一个PDF文件可以通过下面4个方面来理解： 1. 对象, 一个PDF文档是由一个由基本数据类型组成的数据结构。 2. 文件（物...

(http://blog.csdn.net/wzyzzu/article/details/50412458)

APache PDFbox API使用（3）----如何得到一个带表单的PDF文件的表单结构

我们知道，在PDF文件中不但可以保存图片和文字，而且我们还可以在PDF文件里面建立表单。比如，下面的图1就是一个PDF文件里面建立了一些表单。 其实PDF文件是一个有特殊结构的文件，那么，如果我们需要...

chancein007 2015年05月28日 22:36 2332

(http://blog.csdn.net/chancein007/article/details/46136219)

PDF文件结构（一）物理结构

lx111000lx0 2012年11月11日 20:17 932

PDF文件结构（一） ————物理结构 作者：bobob 邮件：zxbbobob@hotmail.com 原文：ht t...

(http://blog.csdn.net/lx111000lx0/article/details/8171843)

PDF文件结构（二）逻辑结构

lx111000lx0 2012年11月11日 20:20 909

PDF文件结构（二） ————逻辑结构 作者：bobob 邮件：zxbbobob@hotmail.com 原文：...

(http://blog.csdn.net/lx111000lx0/article/details/8171853)

JavaClass文件的结构分析及其校验.pdf

2008年05月04日 10:59 919KB 下载

PDF文件主结构解析

2010年04月18日 19:14 1.71MB 下载