PDF文档的一种数字水印算法

谭国律
TAN Guolv
上饶师范学院 数学与计算机科学学院,江西 上饶 334001
School of Mathematics & Computer Science, Shangrao Normal University, Shangrao, Jiangxi 334001, China

TAN Guolv. Digital watermarking algorithm for PDF document. Computer Engineering and Applications, 2012, 48(32):85-88.

Abstract: Targeted at structural features of PDF documents, a new text digital watermarking algorithm is presented. According to itself syntax and structure of binary PDF file, the watermark is embedded without changing the original any visible attributes. Experimental results show that the digital watermark embedding method has good imperceptibility, and can effectively achieve document content integrity and copyright protection.
Key words: digital watermarking; PDF; appreciable

摘 要: 针对PDF文档的结构特点,研究并提出一种新的文本数字水印算法。在不改变PDF文档原有任何可见属性的前提下,利用PDF二进制文件自身的语法和结构特征嵌入水印信息。实验结果表明,该数字水印嵌入方法具有很好的不可感知性,能有效地实现文档版权和内容完整性保护。
关键词: 数字水印; PDF; 可觉察性
文献标识码: A **中图分类号:** TP309.2 doi:10.3778/j.issn.1002-8331.1105-0066

1 引言

数字水印技术主要解决网络环境中数字媒体的信息安全和版权保护等问题。嵌入水印之后的宿主媒体不能在视觉上有所降质或可察觉性,但又要求在无意或恶意攻击之后,水印难以被去除,且能检测出水印;或者,只要非授权者对媒体作任何改动,都能被觉察,以达到媒体完整性保护。当前,数字水印许多研究成果主要还是基于图像、视频、音频等具有大量冗余空间的载体水印算法。由于文本字符所固有的原子性问题,冗余空间非常小,这就大大增加了文本数字水印的鲁棒性和有效载荷问题的解决难度。当前基于文本的水印技术主要可以分为基于格式的文本水印算法、基于扩展空间的水印算法和基于自然语言处理的文本水印算法。文献[1-2]就是属于格式算法,它们通过改变文本的行间距或字间距或字的颜色、字体等来嵌入水印。这些方法完全依赖于文档格式,只要对文本的格式进行修改,嵌入的水印信息便会消失。文献[3-4]是扩展水印算法,它们把文档转换为图像格式,大大增加了载体信息冗余,这使得文档即使经过打印或扫描,水印信息仍能识别出来。但问题是文档图像化后,文档的体积显著增加,而且水印的检测难度很大。文献[5-6]是语意水印技术,这类技术是在不改变文本原意的前提下,通过调整句子的结构以嵌入水印信息到文档内容中的一种信息隐藏技术,具有良好的抗攻击性。但是,载体文本容易出现语义的改变,特别是中文文档,稍微进行一下词语替换、语义或句法的修改就不能表达作者的原意,甚至让人难以理解或产生歧义。

PDF格式是由Adobe公司制作的一种电子文件格式,已经逐渐成为国际标准,其跨平台、便携、安全、高保真等特性使得它的应用遍及各个行业,在网络出版、电子出版和印刷出版中得到广泛的应用。许多的政府机关、档案馆、图书馆等机构都将PDF作为电子文档长期保存的格式和电子文档交换的标准

基金项目: 国家自然科学基金(No.61165003)。
作者简介: 谭国律(1957—),男,教授,硕导,主要研究领域为信息安全。E-mail:tan-gl@163.com
收稿日期: 2011-05-05 **修回日期:** 2011-07-12 **文章编号:** 1002-8331(2012)32-0085-04
CNKI出版日期:2011-10-13 http://www.cnki.net/kcms/detail/11.2127.TP.20111013.0959.106.html

模式。因此,研究在PDF文档中嵌入水印就显得非常重要,具有广泛的应用前景。目前,关于此方面的研究已经有一些,比如文献[7-10]就属于此类。文献[7]利用PDF文档页面对象和结构特征嵌入水印信息,同时结合纠错编码和公钥密码算法以提高水印的鲁棒性和安全性;文献[8-10]都是利用PDF与PS之间的转换,调整字符间距或行间距来实现水印的嵌入。本文在分析PDF文档相关特性的基础上,利用PDF二进制文件自身的语法和结构特征,提出了一种新的在PDF文档中嵌入水印的方法,在不影响PDF文档原有任何可见属性的前提下,能有效地实现PDF文档的版权保护和内容完整性保护。

2 PDF文件格式简介

PDF是一种文件格式,它不依赖于产生或显示它的程序、硬件及操作系统。一个PDF文件由一系列的对象构成,这些对象一起定义各个页面的外观,还可能包含一些相互作用的对象及更高级的应用数据。PDF文档表示为8位二进制字节流,在PDF文档页面中显示的任何内容都是作为图形的。关于PDF的更详细的内容,参考文献[11],在此仅就本文中涉及到的作一简要介绍。

一个规范的PDF文档由如下四个元素构成。

文件头(Header):只有一行,用以说明该PDF文档对应PDF规范的哪个版本。

文件体(Body):包含组成该文档的对象。

交叉引用表(Cross-reference table):包含该文档中间接对象的信息。

文件追踪体(Trailer):指明交叉引用表及某些特殊对象在文件中的位置。

一个PDF文件的内容可以逐步更新而无需重写整个文件,更改被附加到文件的末尾,留下了原来的内容不变,即所谓的增量更新。因此,整个PDF文件的结构如图1所示。

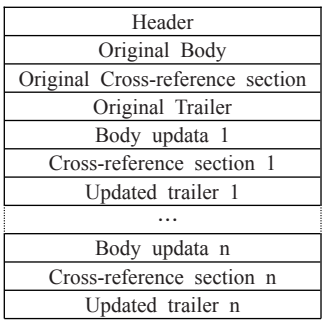

图1 PDF文件结构

从图1看出,一个从未被修改过PDF文档仅由图中的前四部分组成,而修改却是追加在后面。

在PDF文件中,句法结构是用空白符来分隔的,除了在注释、字符串和流中,所有的空白符都是等价的,而且PDF格式视一串连续的空白字符为一个空白字符。

交叉引用表是PDF文件中唯一具有固定格式的部分。交叉引用表中所包含的信息,允许在文件中随机访问任何的间接对象,使整个文件不需要为任何特定的对象定位。交叉引用表包括一个或多个的交叉引用节,每个交叉引用节都以关键字“xref”开头,后跟一个或多个交叉引用小节。每个交叉引用小节包含若干个对象号在一个连续范围内的间接对象,每个间接对象都有相应的一行来指示该对象在文件中的位置等信息,具体的格式为:

```
nnnnnnnnnn ggggg n eol
```

或

```
nnnnnnnnnn ggggg f eol
```

其中nnnnnnnnnn是一个10位字节的偏移,ggggg是一个5位数的世代号,n表示该对象已在使用,f表示该对象是自由的,eol是一个2字节的结束符,nnnnnnnnnn和ggggg及n或f严格地用一个空白符隔开。例如:

```
xref
0 1
0000000000 65535 f
3 1
0000025325 00000 n
23 2
0000025518 00002 n
0000025635 00000 n
30 1
0000025777 00000 n
```

就是一个交叉引用节,含有4个交叉引用小节。比如第3个小节中的“23 2”表示该小节有两个间接对象,编号从23开始到24结束。编号为23的那个间接对象相对于文件开头的偏移地址为0000025518,世代号为00002,并且是在用的。

在PDF-1.4版及之前,交叉引用表的偏移地址是放在文件追踪体中关键字startxref之后。例如,从如下文件追踪体

```
trailer
    <</Info 104 0 R
      /Root 103 0 R
      /Size 105
      /ID [<f3443f968999c70a071445021b601633>
          <02e767b2bdd7bc67f831de7a953526be>
        ]
    >>
startxref
287950
%%EOF
```

可看出,交叉引用表的偏移地址是287950。

在PDF 1.5以后,交叉引用信息既可以放在交叉引用表中,也可存放在交叉引用流(cross-reference stream)中,或两者兼用的称为混合引用(hybrid-reference)。每个交叉引用流都包含了一个交叉引用节中交叉引用和追踪体的全部信息,追踪体信息存放在交叉引用流的流字典(stream dictionary)中,而交叉引用信息则存放在流数据(stream data)中,如下所示:

```
...objects...
12 0 obj                %Cross-reference stream
    <</Type/Xref   %Cross-reference stream dictionary
      /Size...
      /Root...
      /W...
      /Index...
    >>
    stream
       ... %Stream data containing cross-reference information
    endstream
    endobj
    ...more objects...
    startxref
    byte_offset_of_cross-reference_stream %Points to object 12
    %%EOF
```

隐藏在交叉引用流中的每个对象在流数据(stream data)中都有类似于如下形式的一项:

```
01 0E8A 00
00 0002 01
```

其中格式由“W...”规定,如W[1 2 1]就表示格式中的三个域分别占1、2和1个字节,域之间严格地用一个空白符相隔。第一个域指明类型,用0、1和2分别表示自由对象链(相当于交叉引用表中的f)、非压缩和压缩在用对象。0类的第二个域表示下一个自由对象号,第三个域表示该项对象重新使用的世代号;1类的第二个域表示对象相对于文件头的偏移地址,第三个域表示该项对象世代号;2类的第二个域表示该对所在对象流的对象号,第三个域表示该对象在对象流中的索引号。比如,上面第一项中的“01”是指该对象是在用的非压缩对象,“0E8A”是它的偏移地址。另外,在对象流中的许多信息都是编码压缩存储的,且通常都用FlateDecode进行编码压缩。因此,为了取得交叉引用流中的交叉引用信息,需要进行解码。

3 水印嵌入方法

由前面关于PDF二进制文件格式,可设计新的数字水印嵌入方法。

3.1 水印嵌入方案1

正如上章所述,在PDF文件中,交叉引用表是唯一具有固定格式的部分,可以利用每个交叉引用条目

```
nnnnnnnnnn ggggg n eol 或nnnnnnnnnn ggggg f eol
```

中nnnnnnnnnn与ggggg及n(或f)之间的两个空白符来嵌入水印信息,这是由于把这两个空白符随便改写成什么都不影响该PDF文档的任何信息,特别不影响PDF阅读器对它的阅读显示。由于PDF文件是由各类对象一起定义的,每个间接对象在交叉引用表中都有相应的格式如上的一行交叉引用条目。一般地,在PDF文件中,间接对象的个数有许多,少则几十个,多则成千上万个,故嵌入信息的空间是很大的。比如,文件追踪体中的“/Size 105”表示有105个间接对象。而在PDF 1.5以后,流数据中形如

```
01 0E8A00 00
```

的交叉引用信息中的两个空白符也具有同样的作用。

利用这种思想嵌入水印信息的算法如下:

(1)以二进制形式打开PDF文件。

(2)通过关键字“startxref”找到交叉引用表的最后一个交叉引用节的偏移地址,如果有的话,通过键/Prev的值找出所有交叉引用节的偏移地址。

(3)统计交叉引用条目中的空白符字节数,并根据待嵌入水印信息的字节数,采用某种策略以决定水印信息嵌入的位置。

(4)把相应的交叉引用条目中的空白符改写成所需的内容。

(5)重复(4),直到水印嵌入完毕。

(6)保存文件,结束。

3.2 水印嵌入方案2

PDF一共有6个空白字符,具体见表1。

表1 PDF空白字符集

DECIMAL	HEXADECIMAL	OCTAL	NAME
0	00	000	Null(NUL)
9	09	011	Tab(HT)
10	0A	012	Line feed(LF)
12	0C	014	Form feed(FF)
13	0D	015	Carriage return(CR)
32	20	040	Space(SP)

由于在PDF文件中,除了在注释、字符串、流和编码压缩及加密字符中,所有的空白符都是等价的,而且它视一串连续的空白字符为一个空白字符。为了方便,把除了在注释、字符串、流和编码压缩及加密字符中以外的空白符称之为正常的空白符。利用PDF文件对待空白符的这种行为,又可得出另一类水印嵌入方法。

方法1 统计PDF文件中5种正常的空白字符,采用替换技术嵌入水印。比如,SP用HT替换代表一位

信息“1”,HT用LF替换代表一位信息“0”,等等。

方法2 统计PDF文件中5种正常的空白字符,采用添加技术嵌入水印。比如,在一个SP后面再插入一个HT代表一位信息“1”,在HT后面再插入一个SP代表一位信息“0”,等等。

方法3 方法1与方法2的结合,即,采用替换和插入相结合的技术嵌入水印。

4 实验结果及分析

根据上面的水印嵌入方案1,用VC++6.0编写了一个PDF文件交叉引用信息的提取和水印嵌入程序。并取一个PDF文件,用此程序提取的嵌入水印信息前后的交叉引用信息如表2。用PDF浏览器打开嵌入水印后的文件,显示的效果与嵌入水印前的完全一样,没有任何的区别,打开时也没有任何的障碍。当然,上面嵌入的水印简单地用“A”和“B”表示,实际使用时需要用真正的水印信息取代。当然,根据需要还可对水印信息事先进行加密,并最好加密成不可见字符。

表2 水印嵌入前后的交叉引用信息

原交叉引用信息	嵌入水印后的交叉引用信息
xref 0 142	xref 0 142
0000000000 65535 f	0000000000A65535Bf
0000116473 00000 n	0000116473A00000Bn
0000116638 00000 n	0000116638A00000Bn
0000000016 00000 n	0000000016A00000Bn
0000109126 00000 n	0000109126A00000Bn
0000071256 00000 n	0000071256A00000Bn
0000000180 00000 n	0000000180A00000Bn
…	…
trailer	trailer
<</Size 142	<</Size 142
.....>>	.....>>
startxref	startxref
116766	116766
%%EOF	%%EOF

根据上面的水印嵌入方案2,并取一个PDF文件,在正常空白符中,把所有原来的字符SP替换成字符HT,把所有原来的字符HT替换成字符SP,用PDF浏览器打开经这样替换以后的PDF文件,显示的效果与原先的完全一样,没有任何的区别,打开时也没有任何的障碍。当然,在实际使用时,由于PDF文档中的空白字符很多,可根据某种策略(如统计方面)决定对那些空白符进行处理,以增强随机性和均匀性。

分析这两种水印嵌入方法,容易看出,水印信息的嵌入和提取是很方便的。由于PDF文件的二进制存储方式,很明显它能防拷贝复制、防另存,这里的另存指的是在PDF浏览器下另存为PDF文件。这样,此方法能有效地抵抗非法复制和肆意传播。

再对方案1进行实验。取一个用方案1嵌入水印(比如签名)后的PDF文档,用PDF编辑器(比如Adobe Acrobat 7.0 Professional)打开进行编辑(包括缩小或放大)并保存,发现水印信息被破坏。这说明,利用方案1可实现文档的版权和文档内容的完整性保护。

5 结束语

本文提出了一种新的文本水印嵌入方案,为文本数字水印技术提供了一种新的思想。该方案水印信息的嵌入和提取简便实用,而且,由于是利用PDF文档本身的语法和结构特点,从而具有完全的不可见性。在PDF文档中用此方法嵌入文本的签名,可很好地实现对文档版权和完整性的保护。在实际使用时,采用两个方案相结合的方式会取得更好的效果。

参考文献:

[1] Huang Ding,Yan Hong.Interword distance changes represented by sine waves for watermarking text images[J].IEEE Trans on Circuits and Systems for Video Technology,2001,11(12):1237-1245.

[2] 周新民,孙星明,刘超.基于汉字结构知识的鲁棒性公开文本水印[J].计算机工程与应用,2006,42(8):165-167.

[3] 弋英民,李人厚,梅时春,等.一种基于文本行和对角侧面特性的数字水印方法[J].小型微型计算机系统,2005,26(2):293-296.

[4] Shirali-Shahreza M H,Shirali-Shahreza M.A new approach to Persian/Arabic text steganography[C]//Proc of the 5th IEEE/ACIS International Conference on Computer and Information Science,2006:310-315.

[5] Atallah M J,Raskin V,Crogan M,et al.Natural language watermarking:design,analysis,and a proof-of-concept implementation[C]//Proc of IH'01.Berlin:Springer,2001.

[6] 张宇,刘挺,陈毅恒,等.自然语言文本水印[J].中文信息学报,2005,19(1):56-62.

[7] 王强,刘星彤.基于纠错码的PDF文档数字水印算法[J].计算技术与自动化,2009,28(3):137-141.

[8] 张秋余,余冬梅,管伟.中文PDF文档数字水印算法[J].计算机工程与设计,2007,28(24):5983-5984.

[9] 张静,张春田.用于PDF文档认证的数字水印算法[J].天津大学学报,2003,36(2):216-219.

[10] 廖柯宇,李炳法,马增辉,等.一种基于PDF文档的数字水印算法[J].现代计算机:专业版,2005(5):4-8.

[11] Adobe Systems Incorporated.PDF reference(sixth edition)[Z].2006.