第 32 卷 第 17 期 计 算 机 工 程 2006 年 9 月
Vol.32 № 17 Computer Engineering September 2006
·安全技术· 文章编号：1000—3428(2006)17—0230—03 文献标识码：A 中图分类号：TP309

一种新的基于 PDF 文档结构的信息隐藏算法

刘友继[1]，孙星明[2]，罗 纲[2]
(1. 湖南大学软件学院，长沙 410082；2. 湖南大学计算机与通讯学院，长沙 410082)

摘 要：通过分析格式化文件 PDF(Portable Document Format)文档的数据结构，提出了一种新的基于 PDF 文档结构的大容量信息隐藏算法。将秘密信息预处理后伪装成合法 PDF 对象的形式，以文件流的操作方式嵌入到载体文件中，并满足嵌入的信息不影响文件在阅读器、编辑器与打印机中的输出。实验实现了线性化 PDF 文档的信息隐藏与检测。理论分析与实验结果均表明，该算法具有较大的信息隐藏容量、很快的隐藏与检测速度及依赖于加密算法和密钥的安全性。

关键词：信息隐藏；PDF 文档；伪装数据

A Novel Information Hiding Algorithm Based on Structure of PDF Document

LIU Youji[1], SUN Xingming[2], LUO Gang[2]
(1. Software School, Hunan University, Changsha 410082; 2. School of Computer and Communication, Hunan University, Changsha 410082)

【Abstract】After analyzing the data structures of PDF file, a novel information-hiding method based on the structures of PDF file is proposed. First, the secret data is camouflaged to form the legal PDF object, and then the data will be embedded in the carrier PDF file by operating the document flows. The embedding won't affect the output of the readers, the editors and the printers. The information hiding and detecting of linearized PDF documents have been implemented. Theoretical analysis and experimental results show that the algorithm can achieve large capability, high speed of hiding and detecting, and security which depends on the encrypted algorithm and the key.

【Key words】 Information hiding; PDF document; Camouflaged data

1 概述

随着 Internet 的迅速发展和应用，信息的安全问题日益突出。信息隐藏[1~3]作为隐秘通信的一种实现方式，因其没有表明重要信息的存在、不易被破解者注意、为重要信息提供了另一层安全保障，故成为了近几年研究热点之一[3]。而格式化文档因具有丰富的文本格式、图文并茂、支持交互浏览的特点，如 Word 文档、PDF 文档、PostScript 文档、HTML 文档等，成为了网络信息传输的主要载体，其中 PDF 格式以其开放性、便携性、安全性、高效性与跨平台的特性，成为了网络中资料传输的主要方式之一[4]，因此基于 PDF 格式化文档结构的信息隐藏具有广阔的应用前景。

目前的文本信息隐藏领域中，基于文本格式和文本内容的信息隐藏研究较多[5,6]，而针对格式化文档结构的信息隐藏算法很少。基于文本格式的隐藏算法的特征是以调整文本某种格式信息：如各种间距微调、相似字体、相近颜色或亮度等进行信息隐藏[7,8]，因为其格式统计特征明显，很容易被破解者注意，加之容量较小，故而隐蔽性较弱。随着研究的深入，基于语法或语义[5,9,10]等文本内容的相关算法相继提出，解决了对格式的依赖性。基于语法的信息隐藏的隐蔽载体主要是以句子为单位的自然语言[5]，很难形成结构、意义连贯的段落或章节，隐蔽效果较差。基于语义的信息隐藏能够抵抗去格式攻击和文本转换攻击，但因其是通过同义词替换、句型变换、标点处理等实现信息隐藏[5]，故容量受到载体的同义词、句型及标点等限制，较难实现大容量的信息隐藏。

基于以上情况，本文提出了一种新的基于 PDF 文件数据结构的大容量、不可见的信息隐藏算法。首先将隐秘信息加密，再伪装成 PDF 合法数据结构的形式随机地嵌入到文件数据序列中，然后重新构造文件及相关信息，最后依次写入文件。因伪装的结构化数据不进行页面描述，故不影响文件输出。实验结果表明，本算法突破了常规的信息隐藏容量限制，能进行大容量的信息隐藏，具有较好的安全性和很好的读写性能。

2 PDF 文件结构特点与分析

2.1 PDF 文档结构

(1)逻辑结构

PDF 文档由一个以目录对象为根的对象树结构组成[4]。目录对象包含了对页面树、大纲树、指定的外部文件等其它对象的引用。页面对象作为文档主体，通过一个页面树进行描述，页面树即一个页面集，页面集由若干子页面集和若干页面组成，形成一个递归定义。页面又包含了另外的对象或引用。其它对象组织形式类似。PDF 文档的所有使用中的对象都属于目录对象为根的树，通过这棵树便可以轻易访问属于文档的任何对象。即这棵树把 PDF 文档中的信息联成了一个逻辑整体。

基金项目：国家自然科学基金资助项目(60373062,60573045)；高校博士点基金资助项目(20050532007)
作者简介：刘友继(1978—)，男，工程师、硕士生，主研方向：文本信息隐藏，图像水印；孙星明，教授、博士、博导；罗 纲，博士
收稿日期：2006-02-12 **E-mail：**liuyouji@163.com

对象是组成 PDF 的基本元素。PDF 支持的对象包括：逻辑值，数值，字符串，名字，数组，字典，流，以及空对象等。较复杂的如字典、流对象定义如下[4]：

<字典>:=<<
{</属性名 属性值>}* //若干对
>>

字典由若干(属性名，属性值)对组成，属性的值可以是字典或其它对象。

<流>:=<字典>
stream
{<若干行的字符>}*
endstream

(2)物理结构

一个简单的PDF文件物理结构可分为文件头、文件主体、交叉引用表、文件尾 4 个组成部分[4]，我们关心的 PDF 对象依次存储在文件主体中，而在交叉引用表中存储了对应对象的偏移地址、产生号和使用标记。为了提高执行效率，每对 PDF 文件进行一次标准更新，包括内容的增加、删除和修改等，不对文件进行重构，只在文件的尾部追加一段<文件主体、交叉引用表、文件尾>实现。故原始数据具有标准更新稳定不变性。

2.2 PDF 结构特性分析

首先，对象作为 PDF 基本组成元素，在物理结构上是相互独立的，使得以对象为基础的隐藏机制成为可能，即可以通过在对象(如字典对象、流对象)中嵌入信息实现信息的隐藏。其次，文件更新后，增加了部分对象，也有一些对象的引用被删除，但对象依然存在文件中，还有一些重定义的对象。故有对象增加，也有对象被丢弃。因此，可以利用被丢弃对象的废弃空间获取有限的的隐藏空间，或者利用增加对象获取理论上任意容量的隐藏空间，从而保证了信息隐藏的容量需求。

3 基于 PDF 文档结构的信息隐藏与检测算法

根据上述 PDF 的结构特性， 本文信息隐藏算法的基本思想如下：首先获取 PDF 文件的所有对象(包括没有被引用的对象和重定义对象的历次定义)和对象树结构，然后联合待隐藏信息的容量要求，以增加对象与重新利用废弃对象空间的方法相结合，将隐秘信息伪装成若干个 PDF 的合法对象插入到文件的对象树中，形成一个新的对象树，再按照原始结构写入文件。新插入的对象不进行页面输出或进行无效输出，从而使隐藏的信息完全不可见。隐藏算法框图如图 1，具体见算法 2。

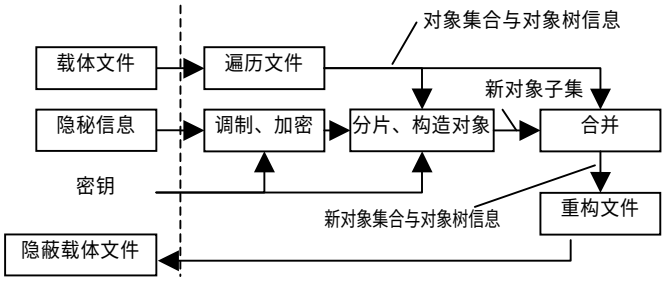


图 1 信息隐藏框图

检测算法是隐藏算法的逆过程，如图 2 所示，先获取 PDF 文件的所有对象和对象树结构，然后找出隐蔽对象、从中提取隐秘信息，再进行合并与解密操作，得到隐秘信息。详见算法 3。

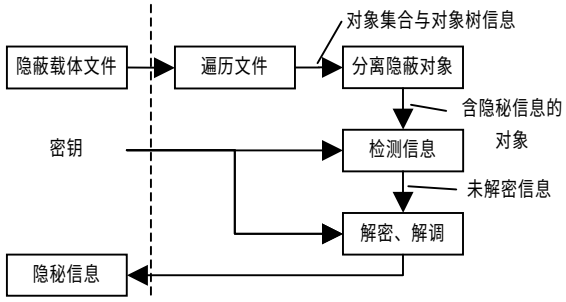


图 2 信息检测框图

下面是 3 个主要算法的描述，其中算法 1 用于获取 PDF 文档的对象信息。

算法 1 PDF 文件对象信息读取算法

输入 原始载体 PDF 文件

输出 PDF 对象集合、对象树

Step1 打开 PDF 文件，取得文件指针。

Step2 将文件访问偏移坐标定位到文件尾部。

Step3 向文件头方向搜寻到最后一个文件尾标记。

Step4 读文件尾部信息，取得根对象号和交叉索引表偏移地址。

Step5 读文件尾部之前的对应交叉索引表，取得当次更新的所有对象的对象号与偏移地址。

Step6 根据对象偏移地址循环读取当次更新的所有对象，得到对象的数据及对其它对象的引用信息。

Step7 向文件头方向继续搜寻前一个文件尾标记。若找到，则跳到 Step4 执行；否则，执行 Step8。

Step8 关闭原 PDF 文件。

Step9 根据对象间的引用关系生成对象树。

算法 2 信息隐藏算法

输入 原始载体文件、隐秘信息、密钥

输出 隐蔽载体文件

Step1 根据文件尾部信息和交叉索引表，使用算法 1 遍历载体文件获取所有对象(对象集合)和对象树的信息。

Step2 将隐秘信息的长度及信息本身通过密钥进行混沌调制与加密，得到加密信息。

Step3 根据待隐藏的加密信息容量和 Step1 的结果，智能地使用增加对象与重新利用废弃对象空间方法相结合，通过密钥控制，将加密信息进行分段、伪装构造对象，得含有隐秘信息的对象子集。

伪装构造的对象应该是数据结构合法的 PDF 文件对象，具有良好的统计特性，从而保证不被格式分析软件察觉；且具有一个确定的标记，以便能够被正确检测。

Step4 将 Step3 的结果随机插入，合并到 Step1 的结果中，得新的对象集和对象树。

Step5 保持原始文件的物理结构基本不变、即不增加新更新段的前提下，使用新的对象集和对象树重构文件，并更新交叉索引表的对应信息，生成含隐秘信息的隐蔽载体文件。

算法 3 信息检测算法

输入 含隐秘信息的隐蔽载体文件、密钥

输出 隐秘信息

Step1 同信息隐藏算法的 Step1，使用算法 1 遍历含隐秘信息的载体文件，获取对象集合和对象树的信息。

Step2 根据隐蔽对象的约定标记，从 Step1 的对象集合和对象树中分离出含隐秘信息的对象。

Step3 根据 Step2 的结果，使用密钥检测出被隐藏的加密信息，进行整合。

Step4 利用密钥，对 Step3 的整合结果使用对应的算法依次解密、解调，得到可能的原始隐秘信息。

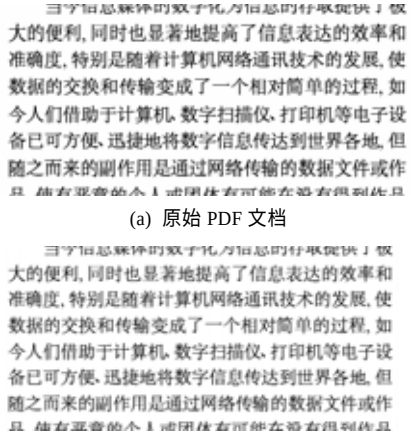
Step5 根据隐秘信息及其长度匹配结果，判断信息的可靠性。

4 算法的特性比较与分析

实验相关数据结果得自实验环境为：CPU，P4，2.4GHz；MEM，512MB；HD，40GB；OS，WindowsXP；开发平台，MS VC6.0。

4.1 隐蔽性测试

实验证明，使用本文提出的算法没有改变隐蔽载体文档的输出，具有很好的隐蔽性，嵌入隐藏信息前后的 PDF 文档对照如图 3 所示。



(a) 原始 PDF 文档

(b) 含隐秘信息的 PDF 文档

图 3 PDF 信息隐藏前后的截图

4.2 与其它信息隐藏算法的实验结果比较

随机选择了 50 个线性 PDF 文档使用了本文算法进行信息隐藏、同时对其它类型算法[7~10]仿真，进行相关数据统计。统计结果如表 1 所示。

由表 1 中数据可以得出，本文算法对载体文件的要求较低，其隐蔽性、隐藏容量、隐藏与检测速度等均优于其它算法，但鲁棒性较之要差。和基于语法、语义的算法相比，该算法不能以纯文本文档作为隐蔽载体，但考虑到目前的大部分资料的传播都是格式化文档的形式，这一点是可以接受的。

表 1 实验结果比较

	本文算法	基于格式的算法[7,8]	基于语法、语义的算法[9,10]
载体要求	格式化文档	格式化文档，文档中要有一定 数量的文本数据	纯文本文档，或带有文本数据的格式化文档
隐藏容量	文档大小的 5% (理论上可任意设定)	容量<文档大小的 0.655%	格式文档：容量<0.5%；纯文本文档：容量<2.5%
蔽性	很好，完全不可见，对语义无任何影响	较好，肉眼难以察觉，对语义无任何影响	较好，完全不可见，但对有些语句的语义有破坏而导致该语句或段落难以理解
速度	0.490 KB/ms	约 0.143 KB/ms	<0.048 KB/ms
稳定性	易碎的	能抵抗一些篡改攻击	能抵抗各种格式攻击，文档类型转换攻击

4.3 特性分析

(1)嵌入容量

文中给出两种方式提供嵌入容量，一是通过使用没有被最终引用的对象和重定义的对象提供，不会改变载体文件的大小，其容量是不确定的，依赖于文件的固有特性，对一个指定的文件，这种容量可能是零，也可能很大；二是通过新建对象提供，其容量不依赖于文件本身，理论上可以是任意大的，但会使隐蔽载体文件相应变大。考虑到隐蔽性，防止被破解者嫌疑，可将容量限制在一个范围，如限制在载体文件大小的 5%内是很难被注意的。隐秘信息量较大时，可以对应选择较大的载体文件。

本文算法使用二者的结合，既保证了隐蔽性，又获得了相当的容量。和以往算法[7~10]相比，容量优势明显。

(2)安全性

本文算法因没有改变文档的输出，因而更加隐蔽。算法的安全性通过以下两个方面支持：1)隐秘信息是以伪合法数据的形式嵌入到文件中的，且没有有效输出，故破解者很难发现其存在。2)隐秘信息经过了密钥的调制、加密，且嵌入过程受密钥控制，即使被嫌疑，也无法被检测和解密。

而且本文算法嵌入的数据在 PDF 线性化操作和非追加方式的更新操作下是易碎的，因此可以较好的保证隐藏信息的安全性，并可作为是否受到攻击的依据之一。

(3)快速性

算法通过数据流的操作方式对载体文档进行信息隐藏和检测操作，保证了执行效率。和以往算法相比，本算法没有对原始文档的数据流进行解密与解析，只需将隐秘信息直接加密，再分段隐藏，因此隐藏相同量数据所需的时间更短，速度更快。

5 结束语

提出了一种新的基于 PDF 文档的大容量信息隐藏方法，将隐秘信息加密后以伪装的数据对象形式嵌入到文件数据流序列中，再重构文件。并使用 VC6.0 实现了以 PDF 线性文件为载体的信息隐藏和信息检测系统。理论分析与实验结果均表明，该算法具有隐蔽安全性较好、执行速度快、隐藏容量大的特点，适用于较大规模的隐秘信息传输，有较广阔的应用前景。

参考文献

1 Jack T B, Steven L, Nicholas F M. Copyright Protection for the Electronic Distribution of Text Documents[J]. Proceedings of the IEEE, 1999, 87(7): 1181-1196.

2 Stefan K, Fabien A P P. Information Hiding Techniques for Steganography and Digital Watermarking[M]. Artech House Publishers, 2000.

3 杨义先，钮心忻．多媒体信息伪装综论[J]. 通信学报，2002, 23(5):32-38.

4 Adobe Systems Incorporated. PDF Reference Fifth Edition[Z]. http://partners.adobe.com/public/developer/en/pdf/PDFReference16.pdf, 2005/10/27.

5 曹卫兵, 戴冠中, 夏 煜等. 基于文本的信息隐藏技术[J]. 计算机应用研究, 2003, 20(10): 39-41.

6 周继军, 杨 著, 钮心忻. 文本信息隐藏检测算法研究[J]. 通讯学报, 2004, 25(12): 97-101.

7 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): 215-219.

8 白 剑, 徐迎晖, 杨 榆. 利用文本载体的信息隐藏算法研究[J]. 计算机应用研究, 2004, 21(12): 147-148.

9 张 宇, 刘 挺, 陈毅恒等. 自然语言文本水印[J]. 中文信息学报, 2004, 19(1): 56-70.

10 Mohan S K, Hau K F. Watermarking of Electronic Text Documents[J]. Electronic Commerce Research, 2002, 2(1/2): 169-187.