第 32 卷 第 3 期 计 算 机 工 程 2006 年 2 月
Vol.32 № 3 Computer Engineering February 2006
·安全技术· 文章编号：1000—3428(2006)03—0161—03 文献标识码：A 中图分类号：TP391

基于 PDF 文档作为掩体的信息隐写方法

钟尚平[1,2]，陈铁睿[1]
(1. 中国科学院计算技术研究所软件室，北京 100080；2. 福州大学数学与计算机科学学院，福州 350002)

摘 要：目前应用极为广泛的 PDF 文档，发现了其中存在可以用作信息隐写的隐密信道。通过采用以一定的冗余换取安全性的策略，并使用基于混沌模型的随机选择隐写单元的方法，使隐写系统满足 Kerckhoffs 原理。分析和实验结果表明，该文隐写方法可嵌入任意大小的信息，并保持在 PDF 阅读器中显示的透明性，具有较好的简单实用性。

关键词：PDF 文档；隐写术；隐密信道；Kerckhoffs 原理

Information Steganography Algorithm Based on PDF Documents

ZHONG Shangping[1,2], CHEN Tierui[1]
(1. Software Division, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350002)

【Abstract】 Aims at the prevalent PDF documents, this paper proposes a novel steganography algorithm. Firstly, it points out the secret channels in the PDF documents. Then, it describes the algorithm which is integrated with several strategies that are applied to improve security. The algorithm makes use of redundancy to complement security, and constitutes a chaotic map to meet the Kerckhoffs principle. The analysis and experimental results show that this algorithm is efficient that can embed data of unlimited length into PDF documents and the embedded PDF documents keep transparency when being displayed in PDF readers.

【Key words】 PDF Documents; Steganography algorithm; Secret channel; Kerckhoffs principle

1 概述

隐写技术是信息隐藏技术的重要分支。与密码术保护信息内容的目的不同，隐写技术是为了隐蔽信息存在的事实或信息存在的位置，即将重要信息隐藏在其它信息之中，使人们觉察不到它的存在，或者知道它的存在，但未经授权无法确定它的位置。一个通用的、在其它数据中隐藏数据的模型可以描述如下[1]：嵌入信息是希望秘密地发送的消息，它隐藏于一个掩体消息中，从而产生隐写文本或者其它隐写对象。一个隐秘密钥用于控制隐藏过程，使检测或恢复过程仅限于那些知道密钥的人(或者那些知道密钥的起源的人)。

PDF 是 Portable Document Format 的缩写，是由 Adobe 公司在 1992 年发表的。PDF 文件的页面元素不仅包括正文文本图元，还包括图形图元和图像图元等，甚至还可包含超文本链接，视音频等[2]。PDF 格式文件在某种程度上是独立于软硬件的文件代表。现在它已经成为跨平台的通用格式。

现在，PDF 文档及其应用已经越来越流行，其应用的广泛性决定了对以 PDF 文档作为掩体的信息隐写算法进行研究的重要性，但是目前对此类算法的研究只吸引了很少的关注。对以 PDF 文档图像图元(或其它媒体)作为掩体的信息隐写算法的研究可参考已有的大量的图像隐写算法[3]。与音、视频等数据不同，文本数据中含有很少的可用来进行秘密通信的冗余信息，以文本作为掩体与以音、视频作为掩体的信息隐写算法不尽相同。现有在格式文档中隐写信息的算法不多[4]等，典型的做法是将信息存储在行间距或列间距中以及采用字形特征编码等隐写单位为 1bit 的方法。针对以 PDF 文档作为掩体的信息隐写算法的研究，我们仅在网上找到了一个隐写工具 wbStego4[5]，该工具支持将数据隐写到 Adobe PDF 文档中。但 wbStego4 的可隐写的信息量很小；按 wbStego4 的 FAQs，wbStego4 可隐写的信息量是难以估计的。

就目前应用极为广泛的 PDF 文档，本文在深入分析其结构的基础上，发现了 PDF 文档中存在的隐密信道。通过采用以一定的冗余换取安全性的策略，并使用基于混沌模型的随机选择隐写单元的方法，使隐写系统满足 Kerckhoffs 原理：通信的安全性不依赖于对所使用方法本身的保密性，而仅依赖于隐写密钥。本文隐写方法可嵌入任意大小的信息，并保持在 PDF 阅读器中的显示的透明性，具有较好的简单实用性。

2 PDF 文档的结构

PDF 文档的结构可以从文件结构和逻辑结构两个方面来理解。PDF 文档的文件结构指的是其文件物理组织方式，逻辑结构则指的是其内容的逻辑组织方式。

2.1 数据对象类型

组成 PDF 文档的基本元素是 PDF 对象(PDF Object)。PDF 对象包括直接对象(Direct Object)和间接对象(Indirect Object)。直接对象有如下几种基本类型：Boolean，Number，String，Name，Array，Dictionary，Null，Stream 等。间接对象是一种标识了的 PDF 对象，标识的目的是为了让别的 PDF 对象引用。任何 PDF 对象标识后都变成了间接对象。这个标识称为间接对象的 ID。

2.2 PDF 文档的物理结构

PDF 文档的文件结构(即物理结构)包括 4 个部分：文件头，文件体，交叉引用表和文件尾。

基金项目：国家自然科学基金资助项目(60273016)
作者简介：钟尚平(1969—)，男，博士，主研方向：网络信息安全，算法设计等；陈铁睿，硕士生
收稿日期：2005-07-05 **E-mail：**zhongshangping@software.ict.ac.cn

文件头指明了该文件所遵从 PDF 规范的版本号，它出现在 PDF 文件的第 1 行。如%PDF—1.2 表示该文件格式符合 PDF1.2 规范。

文件体由一系列的 PDF 间接对象(Indirect Object)组成。这些间接对象构成了 PDF 文件的具体内容如字体、页面、图像等。交叉引用表则是为了能对间接对象进行随机存取而设立的一个间接对象地址索引表。

文件尾声明了交叉引用表的地址，指明了文件体的根对象(Catalog)，还保存了加密等安全信息。根据文件尾提供的信息，PDF 的应用程序可以找到交叉引用表和整个 PDF 文件的根对象，从而控制整个 PDF 文件。

2.3 PDF 文档的逻辑结构

PDF 文档的逻辑结构反映了文件体中间接对象间的等级层次关系。PDF 文档的逻辑结构是一种树型结构。树的根节点就是 PDF 文档的根对象(Catalog)。根节点下有 4 个子树：页面树(Pages Tree)，书签树(Outline Tree)，线索树(Article Threads)，名字树(Named Destination)。

3 PDF 文档中存在的隐密信道及基于此的隐写算法

3.1 PDF 文档中存在的隐密信道

PDF 文件是二进制码和 ASCII 码的混合字节序列。按上述 PDF 文档的物理结构，根据 PDF 文件尾提供的信息找到交叉引用表，据交叉引用表提供的间接对象地址索引表，可以在 PDF 文档的各个间接对象的字节之间嵌入任意数据，通过正确修改交叉引用表及交叉引用表开始地址等参数的值，可完全不影响 PDF 阅读器的输出。在此称 PDF 文档的每两个间接对象的字节之间的单元为信息隐写单元，可用于嵌入数据块。

图 1 表示了嵌入数据后的 PDF 隐藏文件的物理结构。

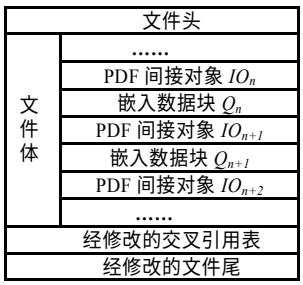

图 1 嵌入数据后的 PDF 隐藏文件的物理结构

3.2 以 PDF 文档作为掩体的信息隐写算法

3.2.1 基于混沌模型的随机选择隐写单元的方法

美国数学生态学家梅（May R）指出，在生态学中一些非常简单的确定性的数学模型却能产生随机的行为。如著名的 Logistic 混沌模型：$x_{n+1}=f(x_n,\mu)=\mu x_n(1-x_n)$。式中 μ 为控制参数，且 $0\le\mu\le 4$，$x\in[0,1]$。当 $\mu>3.57$ 时，序列 $x_0, x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_n, \ldots$像是分布在区间[0，1]上的随机数。

本文利用 Logistic 混沌模型的随机特性，只选择部分隐写单元嵌入信息,以一定的冗余换取隐写算法的安全性。在此，取参数 $\mu=3.8$，初始值 x0 为由隐写密钥经过单向 HASH 函数 SHA-1 产生的[0，1]间的数值，不妨设为 ChaoKey。

这里可以随意取所需要的隐写信道利用率 $\varphi\in[0,1]$，是否采用第 n 个隐写单元嵌入信息(1 表示采用，0 表示不采用，用 y_n 表示)，由 x_n 及 ϕ 的大小关系决定：

$$y_n=\begin{cases}1, x_n\le\varphi\\ 0, x_n>\varphi\end{cases}\qquad(1)$$

因为序列 $x_0, x_1, x_2, x_3, x_4, x_5, x_6, \ldots, x_n\ldots$是伪随机的，根据式(1)选择隐写单元是否用于嵌入信息也是伪随机的，且满足隐藏信道利用率为 ϕ。嵌入信息的提取决定于隐写密钥。

3.2.2 嵌入算法

Step1 生成包括如下内容的“嵌入数据库”：

(1)嵌入信息经加密算法(如:三重 DES 算法等，加密密钥可以采用隐写密钥)运算生成二进制码，设为 QEFile；

(2)准备一份二进制码的“伪装”信息，设为 CamFile；

Step2 隐写密钥经 SHA 及 MOD(16)运算生成的 20 个“0～15 数字”，设为标识串 FlagStr；在标识串 FlagStr 前加“0.”，生成(0 1)间的实数作为混沌序列密码生成器 Chao 的迭代初值 ChaoKey，设迭代值为 ChaoNum；

Step3 设 PDF 掩体文件中的信息隐写单元总数为 PDFCUC。遍历 PDF 掩体文件中的信息隐写单元，如果 ChaoNum 小于等于隐写信道利用率 φ (如:90%)，则取 QEFile 中的一个分块信息(把 QEFile 大约分为“$PDFCUC\times\varphi$”块，分块的字节大小为：除最后一块外，字节数均分，剩下的字节均归为最后一块)，嵌入在 PDF 隐写单元中，同时修改 PDF 交叉引用表中的间接对象在文件中的字节偏移量；否则，如果 ChaoNum 大于 φ，则取 CamFile 中的一个分块信息(把 CamFile 分为“PDFCUC×φ”块)，嵌入在 PDF 隐写单元中，同时修改 PDF 交叉引用表中的间接对象在文件中的字节偏移量；

Step4 修改 PDF 交叉引用表开始地址参数的值,最终把已完全嵌入信息的 PDF 掩体文件转成 PDF 隐写文件。

3.2.3 提取算法

Step1 把提取密钥经过 SHA 及 MOD 运算生成 20 个“0～15 数字”，各自再加“1”得到 20 个 1～16 数字，设为标识串 FlagStr；

Step2 在标识串 FlagStr 前加“0.”，生成(0 1)间的实数作为下述混沌序列密码生成器的迭代初值 ChaoKey；

Step3 构造 Logistic 混沌序列密码生成器 Chao，以 ChaoKey 为迭代初值，设迭代值为 ChaoNum；

Step4 设 PDF 隐写文件中的信息隐写单元总数为 PDFSUC。遍历 PDF 隐写文件中的信息隐写单元，如果 ChaoNum 小于等于隐写信道利 φ (如:90%)，则提取出信息隐写单元中的隐写信息；否则，如果 ChaoNum 大于 φ，则不提取信息隐写单元中的隐写信息；

Step5 合并生成嵌入信息。

4 实验与性能分析

4.1 实验

图 2 为选取 PDF 参考手册[2]的 Chapter1.1～2.3.1 的内容生成的 MS WORD 文档，使用 Adobe PDF Writer 转化成的 PDF 掩体文件 PDFCoverFile.pdf，共 6 页；图 3 为经过本文算法处理的隐写了 Lena 图片(49 206 Bytes)的隐写文件 PDFStegoFile.pdf；明显地，隐写文件 PDFStegoFile.pdf 保持了显示的透明性。

4.2 性能分析

以图 2 的掩体文件(包含 24 026 个字符)为例，使用 wbStego4 的结果是：隐写容量只有 107Bytes，显然此方法无法把 Lena 图片(49 206 Bytes)隐写其中。而将信息存储在英文格式文档的字符间距中的方法(在此，不妨称为“其它方法”)最多能隐写 24 026 bits 的信息量，显然此方法也无法把 Lena 图片(49 206 Bytes)隐写其中。对于将信息存储在格式文档的行间距或列间距中以及采用字形特征编码等方法，其隐写容

量更小。

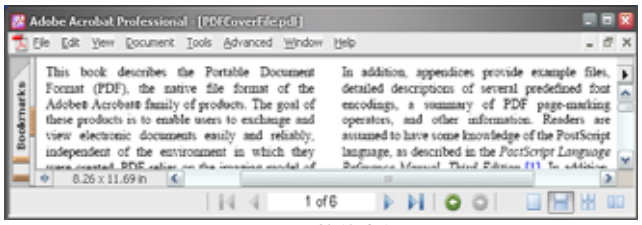

图 2　PDF 掩体文件

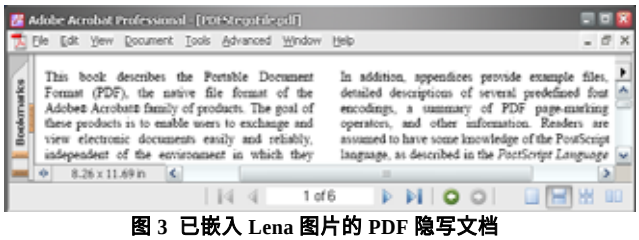

图 3 已嵌入 Lena 图片的 PDF 隐写文档

表 1 为几种基于格式文档的隐写算法的性能比较结果。因为目前还没有很好的对视觉透明性的量化标准，在表中仅用“透明”表示视觉透明性。对性能“隐写文件字节数是否大于掩体文件的字节数“，用“隐写文件是否长个”来表示。

表 1 几种隐写算法的性能比较结果

性能	本文算法	wbStego4	其它方法
视觉透明性	透 明	透 明	透 明
隐写容量大小	任 意	856 bits	24 026bits
容量是否可估计	是	否	是
隐写文件是否长个	是	是	否

由表 1，本文隐写方法可嵌入任意大小的信息，并保持在 PDF 阅读器中的显示的透明性。通过采用以一定的冗余换取安全性的策略，并使用基于混沌模型的随机选择隐写单元的方法，使隐写系统满足 Kerckhoffs 原理，具有较好的简单实用性。另外，本文算法无论对何种排版格式的 PDF 文件(单列/双列，包含图形、图像或数学公式等)，都具有上述隐写性能。但是，把数据隐写到 Adobe PDF 文档中将增大 PDF 文档的文件字节数，也就是说，与 wbStego4 一样，本文隐写方法难以抵抗统计等攻击。

我们知道，说一个隐写系统是安全的是相对而言的。在具体的安全应用中，不仅要考虑隐写算法的抗攻击性能，也要考虑算法的实用性。另外，综合算法的应用成本等因素，还要充分考虑受保护信息的安全等级及可能遭受的攻击强度来选择不同的隐写算法。简单实用但安全性稍差的隐写算法的应用成本较低，在安全等级要求低和可能遭受的攻击强度弱的场合也有广泛应用。

5 结论

就目前应用极为广泛的 PDF 文档，本文发现了其中存在的隐密信道。通过采用以一定的冗余换取安全性的策略，并使用基于混沌模型的随机选择隐写单元的方法，使隐写系统满足 Kerckhoffs 原理。本文隐写方法可嵌入任意大小的信息，并保持在 PDF 阅读器中的显示的透明性，具有较好的简单实用性。本文隐写算法可在安全等级要求低和可能遭受的攻击强度弱的场合得到应用。

参考文献

1 Petitcolas P F A P, Anderson R J, Kuhn M G. Information Hiding-A Survey[J]. Proceedings of the IEEE, Special Issue on Protection of Multimedia Content, 1999, 87(7): 1062-1078.

2 Adobe Systems Incorporated. Portable Document Format Reference Manual(Version1.3) [Z]. http://www.adobe.com, 1999-03.

3 Springer-Verlag. Lecture Notes in Computer Science[C]. Proc. of the 6th Information Hiding Workshop, Berlin Heidelberg, 2004.

4 Low S H, Maxemchuk N F, Lapone A M. Document Identification for Copyright Protection Using Centroid Detection[J]. IEEE Transactions on Communications, 1998, 46(3): 372-383.

5 wbStego Studio. The Steganography Tool wbStego4[EB/OL]. http://www.wbailer.com/wbstego, 2005.

（上接第 155 页）

无法使第 3 个等号成立，故不能安装软件。当需废除多个盗版者的电子证书解密密钥的解密能力时，只要利用上述方法不更新其码向量即可。

4.3 方案的其它性能分析

(1)直接不可否认性：由 OPE 协议的安全性[5]可知，用户 i 的电子证书解密密钥中的$(d_i, f(d_i))$只有其本人知道，其它用户(包括 DS)无法获取该秘密钥，从而当一个包含 $ELDK_i=(d_i, f(d_i), r^i, t_i)$的盗版解码器被没收后，DS 从中提取出$(d_i, f(d_i), \gamma^{(i)}, t_i)$连同用户 i 的购买记录 Δ_i 向 TA 起诉用户 i 的盗版行为，用户 i 无法否认。

(2)自身强化性：由于用户 i 的 $ELDK_i=(d_i, f(d_i), \gamma^{(i)}, t_i)$ 中包含有其签字密钥 d_i，这样用户 i 必然不会让别人获取其签字密钥 d_i，有利于阻止盗版行为的发生。

5 小结

本文利用线性空间码和 OPE 协议给出了一种动态数字版权保护方案。在方案中，数据发行商可以动态地废除盗版者的电子证书及电子证书解密密，维护自己的版权，保护自己的合法权益，本方案还具备自身强化性、直接不可否认性、可追踪性等性能。

参考文献

1 Pfitzman B, Waidner M. Anonymous Fingerprinting Advances[C]. Proc. of Cryptology-eurocrypt'97, LNCS1233, 1997: 88-102.

2 Chor B, Fiat A, Nor M. Tracing Traitor Advance[C]. Proc. of Cryptology-Crypto'94. Berlin: Springer-verlag, 1994: 257-270.

3 Tzeng W G, Tzeng Z J. A Public-key Traitor Tracing Scheme with Revocation Using Dynamic Shares[C]. Proc. of PKC2001, LNCS, Berlin: Springer-verlag, 2001: 207-224.

4 Choi E Y, Hwang J Y, Lee D H. An Anonymous Asymmetric Public Key Traitor Tracing Scheme[C]. Proc. of EC_Web 2003, LNCS 2738. Berlin: Springer-verlag, 2003: 104-114.

5 Naor M, Pinkas B. Oblivious Transfer and Polynomial Evaluation[C]. Proc. of CRYPTO'99. Berlin:Springer-verlag, 1999: 338-353.