基于文本字体的信息隐藏算法

陈 芳，王 冰
（西北大学 计算机科学系，陕西 西安 710069）

摘 要：研究了文本信息隐藏技术，提出一种改进的改变文本载体字符字体的隐藏算法。该算法首先把秘密文本信息中的字符转换为ASCII码，再把ASCII码转化16位二进制码，然后把16位二进制码的高八位和低八位转换为十进制数，用RSA加密法将十进制数加密，最后把十进制码再转换成16位二进制码并利用伪随机置换法把秘密信息代码嵌入到载体文本中。嵌入时选取两种字体，若代码为"1"，载体文本的字体不变，若为"0"则改为一种与原字体相近的字体。通过这些处理手段大大提高了单纯的基于特征编码的信息隐藏技术的安全性。

关键词：文本；信息隐藏；RSA；字体
中图分类号：TP309 **文献标识码**：A **文章编号**：1005－3751(2006)01－0020－03

An Algorithm of Text Information Hiding Based on Font

CHEN Fang, WANG Bing
(Dept. of Computer Sci., Northwest University, Xi'an 710069, China)

Abstract: An improved algorithm of text information hiding based on font is proposed in this paper. First of all, the characters in the secret text are changed into ACSII codes. Secondly, ASCII codes are changed into 16 bit－binary codes. Then, the higher 8 bits and lower 8 bits in 16－binary codes are transformed into decimal codes, which are encrypted by RSA algorithm. Finally, switch the encrypted decimal to 16－bit binary again, and hide them into the cover text with faking random replacing rule. Chose two kinds of similar font, if the code is 1, the font of character keeps the original style, and if the code is 0, the font of character switches to other one, which is quite similar to the original one. As the result, the algorithm is safer than just using characteristic－based information hiding algorithm.

Key words: text; information hinding; RSA; font

0 引 言

随着信息技术不断发展，互联网已被广泛应用到社会的各个领域，融入到人们生活的方方面面。随之而来的网络安全问题也倍受人们的关注，对诸如图像、文本、音频、视频等数字媒体的保护问题就愈显迫切。从1994年第一次国际信息隐藏会议的召开，到现在已经连续召开了四届国际信息隐藏大会。国内也举办了三次信息隐藏会议，这些举措极大地推动了信息隐藏技术的发展。然而，现在大多数信息隐藏算法[1]主要关注于在灰度/彩色图像、音频和视频信息中嵌入秘密信息，但是电子商务的长足发展让人们不得不面对要对网络中日益增加的票据、合同、证明、ID等一些有价值的文本文档加以保护的问题。文本文档不像其他数字信息含有大量的冗余信息，在文本文档中每一个字符都含有确切的信息，不允许噪声的出现，即使是微小的改动都有可能是可见的，甚至会改动文本文档包含的原始信息。

目前对文本信息隐藏的研究还比较少，以文本为载体的信息隐藏技术主要基于文本载体本身的特性，目前较为常用的以文本为载体的信息隐藏方法有以下3种[2]：

1) 行移编码。

该方法通过垂直移动整个文本行来实现信息隐藏。当某行被上下移动时，与其相邻的文本行保持不动，以作为译码时的参考位置。根据人眼的辨别能力，文本行上下移动不超过1/300英寸，人眼是察觉不出来的。由于文本中行间距离是确定的，所以在恢复秘密信息时只需以不动的文本行参照，而不需要参考原始文本。

2) 字移编码。

该方法通过移动某字的水平位置嵌入秘密信息。隐藏过程中与被移动字符左右相邻的字符保持位置不动。由于在对文档进行格式化的时候字间距离也会发生改变，所以字与字之间的距离不是一个固定不变的值，因而在恢复嵌入信息时要参考原始文档。字水平移动量不超过1/150英寸时，人眼是不易察觉的。

3) 特征编码。

该方法通过改变文本文档中某些字符的一些特征量，

收稿日期：2005－04－29
基金项目：国家自然科学基金(60372072)；陕西省科技攻关基金(2004K05－G25)
作者简介：陈 芳(1979—)，女，甘肃镇原人，硕士研究生，研究方向为图像处理、信息隐藏、数字水印；王 冰，副教授，硕士研究生导师，研究方向为图像处理、信息隐藏、数字水印。

如字体、字的颜色、字高或字母(b,d,h,k等)中垂线的长度,而保持其他字符的这些特征不变来实现编码。

在实际应用中为了提高隐藏的鲁棒性,通常将3种方式结合起来使用。

以上隐藏算法都是基于空间域的改变,一旦恶意攻击者使用某些手段对文本行距、字间距或字体特征等进行检测,很容易发现秘密信息。为了进一步提高安全性,笔者提出一种把改变文本文件中的字体与用RSA加密法将秘密信息加密相结合的隐藏方法。这样即使攻击者发现了秘密信息也无法将其破解。

1　改进的信息隐藏算法

1.1　嵌入秘密信息

秘密信息的嵌入通过以下步骤完成(如图1所示):

(1)读入文本秘密信息,把每个字转化为相应的ASCII码;

(2)将ASCII码再转化为与之对应的16位二进制码;

(3)取出16位二进制码中的高八位和低八位,分别转化成十进制数;

(4)用RSA法[3~6]加密这些十进制数;

RSA是第一个既能用于数据加密也能用于数字签名的算法。算法的名字以发明者Ron Rivest, Adi Shamir和Leonard Adleman的名字命名。RSA的数学基础是欧拉定理。对任意小于n且与n互质的正整数a,总有

$$a^{\varphi(n)} \bmod n = 1$$

其中,a和n互质,$\varphi(n)$是比n小但与n互质的正整数个数。

* 密钥的产生。

a.首先取两个质数p和q,p和q都必须保密。

b.计算$n = p \cdot q$,$\varphi(n) = (p-1)(q-1)$,其中n公开,$\varphi(n)$保密。

c.随机选取整数e,满足$\gcd(e, \varphi(n)) = 1$,e公开。其中gcd( )函数用来求两个整数的最大公约数。

d.计算d,满足$(d \cdot e) \bmod \varphi(n) = 1$,在这儿$e$和$\varphi(n)$互质,所以$d$一定存在,使用辗转相除法就可以求得$d$。$d$保密。其中$(e,n)$是公共密钥,$(d,n)$是私有密钥。

* 加密过程。

若明文为a,$a < n$时,将其看成是一个大整数,$a \geqslant n$时,就将a表示成s位二进制数($s \leqslant n$,通常取$s = 2^t$,t是正整数),然后将a分成L段$a[1],a[2],\cdots,a[L]$(L是正整数),每一段表示的数都小于n,最后分段加密。接下来,计算

$$b[i] = a[i]^e \bmod n,\ (0 \leqslant m[i] < n, 0 < i < L+1, e \text{和} n \text{是公共密钥})$$

$b[i]$就是加密后的密文。如果$a < n$,直接进行加密:

$$b = a^e \bmod n$$

b就是加密后的密文。

(5)将加密后的每个十进制数转化为16位的二进制码;

(6)确定载体比特数,用伪随机置换法把秘密信息随机地嵌入伪装载体中;

(7)数据隐藏[7]。

隐藏信息时,用两种相似的字体来表示要隐藏的信息,在一般的字处理器中(如word、WPS等)都存在极为相似的几组字体,如宋体与新宋体,它们在视觉上很难分辨出来。例如,用宋体代表秘密信息中的"1",用新宋体代表秘密信息中的"0"。

原始文本:数字水印是信息隐藏技术的一种具体应用。

秘密信息:100100。

隐藏后的文本:数字水印是信息隐藏技术的一种具体应用。

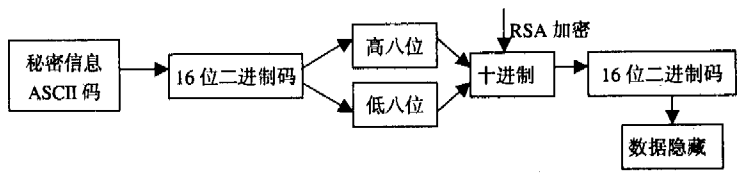

图1　秘密信息嵌入过程

1.2　提取秘密信息

秘密信息的提取过程与嵌入过程相反。

(1)读入伪装载体,根据预先约定,把相应的字体翻译为"0""1"代码;

(2)根据伪随机置换的伪装密钥k,重构"0""1"代码的位置;

(3)读入16位二进制数,将其转化为十进制数;

(4)用RSA解密算法解密,产生ASCII码的高八位h_i;

* RSA解密过程。

$$c[i] = b[i]^d \bmod n$$

($0 < i < L+1$,d和n是公共密钥)

$c[i]$就是解密后的明文,将$c[1],c[2],\cdots,c[L]$都计算出来,由于加密时若明文$a \geqslant n$时将明文进行分段编码,将这L段再合并起来就是解密后的明文c。如果$a < n$,则直接进行计算:

$$c = b^d \bmod n$$

c就是解密后的明文。解密后的明文c和原来的明文a是相等的。

如果第三者进行窃听,有可能会得到3个数:e,n,c(e和n是公共密钥,$n = p \cdot q$,c是密文)。想要进一步解密的话,必须想办法得到$\varphi(n)$,而$\varphi(n) = (p-1)(q-1)$,所以,先得对n作质因数分解,而在大数范围内做合数分解是十分困难的,因此窃密者很难成功。

(5)将读入的16位二进制数转化为十进制数;

(6)用RSA解密算法解密,形成ASXCII码的低八位l_i;

(7) 计算 ASCII 码 c_i:

$c_i = h_i * 28 + l_i$

(8)还原 ASCII 码对应的字符,获得秘密信息。

2 结果和算法分析

2.1 隐藏结果

根据文中所提出的信息隐藏算法,做了大量的实验。图 2 是载体文本,有 6360 个字符(含标点),图 3 为秘密信息,图 4 为嵌入秘密信息后的文本,图 5 为提取出的秘密信息。RSA 加密算法的公共密钥 e 是 983,n 为 1144;私有密钥 d 为 135,n 为 1144。在该算法中以连续出现 32 位"0"作为结束标志,因此该载体文本最多隐藏 197 个字符。

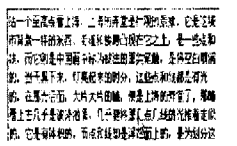

图 2 载体文本

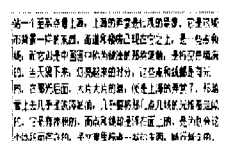

图 4 嵌入后的文本

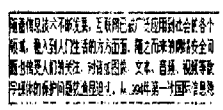

图 3 秘密信息

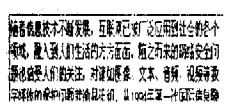

图 5 提取出的秘密信息

2.2 隐藏信息量分析

由于每一个要隐藏的字符先将被转换为 16 位二进制,然后又将 16 位中的高八位和低八位单独取出来,分别进行 RSA 加密,加密之后的 8 位二进制数字就变成了 16 位,这样的话一个秘密字符最终被转换为 32 位二进制的"0"、"1"序列,所以每 32 个文本载体字符能够隐藏一个秘密字符,同时在载体文本中还要设置一个结束标志,表示所隐藏信息内容结束。所以载体文本的字符个数要大于 32 倍秘密信息字符个数。

2.3 安全性分析

隐藏信息所选用的两种字体相当接近,人们在视觉上很难发现。即使攻击者发现有秘密信息隐藏在伪装载体文本中,攻击者试图提取出隐藏的秘密信息时,由于隐藏信息用 RSA 加密算法进行了加密,攻击者即便提取了相应的"0"、"1"序列,提取出来的也只是加密后的密文,再加上密文又通过伪随机置换处理,攻击者没有私有密钥很难破解截获的信息。

2.4 鲁棒性分析

通过改变载体文本字体的方法隐藏信息,对载体文本进行字体的编辑,或进行其他编辑(如删除、修改、替换等)之后,有可能会使隐藏的秘密信息全部或部分丢失。在隐藏的信息量不很大时,可以采用多次隐藏的方法来提高鲁棒性。

3 总 结

提出了一种基于文本字体的信息隐藏方法,通过改变载体文本的字体而将秘密信息文本嵌入到载体文本中,这种方法原理简单,易于理解,实现起来也很容易。但还是存在一些问题,抗干扰性和鲁棒性比较差,例如对载体文本进行修改、删除或对文本字体进行编辑之后文本水印会很容易地被破坏掉,对这一方面还需要改进。

参考文献:

[1] 王慧琴,李人厚.二值文本数字水印技术的研究与仿真[J].系统仿真学报,2004,16(3):521-524.

[2] 黄 华,齐 春,李 俊,等.文本数字水印[J].中文信息学报,2001,15(5):52-57.

[3] Rabin M O. Digitalized signatures and public key function as factorization[R]. Technical Report MIT/LCS/TR212, Cambridge, MA, USA: MIT Lab., 1979.1-16.

[4] Williams H C. A modification of the RSA public-key encryption procedure[J]. IEEE transactions on information theory, 1980, VIT-26(6):726-729.

[5] 张焕国,戴大为,覃中平.FA 公开密钥密码体制的软件实现[A].第二届中国密码学学术会议论文集[C].北京:科学出版社,1992.110-114.

[6] 周长飞,朱根标,张晓丰.PE 可执行文件 RSA 验证加密机制的分析[J].微电子学与计算机,2004,21(8):96-98.

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

(上接第 19 页)

Machine Vision[M]. 艾海舟, 武 勃, 等译. 北京:人民邮电出版社,2003.83-127.

[2] Kim Jong-Bae, Kim Hang-Joon. Multiresolution-based watersheds for efficient image segmentation [J]. Pattern Recognition Letters, 2003, 24:473-488.

[3] Gonzalez R C, Woods R E. Digital Image Processing[M]. 阮秋琦, 阮宇智, 等译. 北京:电子工业出版社,2004.500-507.

[4] Prasad L, Iyengar S S. Wavelet Analysis with Applications to Image Processing[M]. New York: CRC Press, 1997.

[5] Zhao Jianwei, Wang Peng, Liu Chongqing. Watershed Image Segmentation Based on Wavelet Transform[J]. 光子学报, 2003, 32(5):601-604.

[6] 杨福生.小波变换的工程分析与应用[M].北京:科学出版社,2001.

[7] 马丽红,张 宇,邓健平. 基于形态开闭滤波二值标记和纹理特征合并的分水岭算法[J].中国图象图形学报,2003, 8A(1):77-83.

[8] 谢凤英, 姜志国, 周付根. 基于数学形态学的免疫细胞图象分割[J].中国图象图形学报,2002, 7A(11):1119-1122.

[9] 杜啸晓,杨 新,施鹏飞. 一种新的基于区域和边界的图象分割方法[J].中国图象图形学报,2001,6A(8):755-759.