提高 Word 文本文档信息隐藏容量的方法研究

李向辉,钟 诚
(广西大学 计算机与电子信息学院,广西 南宁 530004)

摘 要:介绍文本文件信息隐藏的几种典型编码方法,并比较各种方法的信息隐藏量;分析 Word 文档的文件结构,提出一种通过字符缩放编码、字体 RGB 灰度编码、改变 Word 文本文档中字符下划线 RGB 灰度值来实现隐藏秘密信息的方法。理论分析和实验结果表明该方法能提高信息隐藏量。

关键词:信息隐藏;Word 文本文档;RGB
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2006)09-0097-03

Research on Improving Information Hiding Capacity for Word Text Document

LI Xiang-hui, ZHONG Cheng
(School of Computer and Electronics and Information, Guangxi University, Nanning 530004, China)

Abstract: In this paper, some typical coding methods for information hiding are introduced and their information hiding capacity is analyzed. And a new method of hiding secret information is presented by analyzing the structure of the word text document, zooming characters, coding RGB value of font gray and modifying the RGB value of the character underline in the word text document. The theoretical analysis and experiment show that it can obtain high information hiding capacity.

Key words: information hiding; word text document; RGB

0 引 言

信息隐藏技术研究如何将某一信息隐藏于另一公开的信息中,然后通过公开信息的传输来传递隐藏信息[1]。信息隐藏技术不同于传统的密码学技术。为了增加破译的难度,可将加密技术与隐藏技术结合起来应用,即先对待嵌入对象进行加密得到密文,再把密文隐藏到载体对象中。信息隐藏技术在军事、电子政务、电子商务、网络出版等方面发挥重要作用。目前,关于信息隐藏的研究大多集中在图像和视频、音频方面,基于文本的方面的研究较少。这是因为信息隐藏的前提条件有两个:① 载体信息本身存在冗余性;②人的感观对某些信息有一定的屏蔽效应。由于图像、视频、音频等载体的信息冗余性较大,人的感官对这些信息的掩蔽效应明显,可隐藏的信息量也就相对较大。相比之下,文本中信息冗余较小,嵌入不可感知的信息较难。文中将研究 Word 文本文档信息隐藏容量的问题。

收稿日期:2006-01-05
基金项目:广西科学基金(桂科自 0339008);广西大学博士科研基金(B0309031)
作者简介:李向辉(1976-),男,广西桂林人,硕士研究生,主要研究网络信息安全;钟 诚,博士,教授,主要研究网络信息安全、并行分布计算等。

1 文本信息隐藏方法比较

适用于 Word 文本信息隐藏的有词义替换法、空间特征替换法和颜色特征替换法三大类。

1.1 词义替换法

Bender 等人提出了对文本特定的单词进行同义词替换的方法[2]。例如用“big”替换“large”、“smart”替换“clever”等。需替换的单词表示“0”,无需替换的单词表示“1”。在提取信息时还需要同义词替换表作为参考。文本中隐藏的信息量与文本中同义词组出现的频率一致。因此,这种方法可隐藏的信息量较小,而且可隐藏的信息量不固定。

1.2 空间特征替换法

这类方法不修改文本内容,只对文本行、字和词在页面上做不易被识别的轻微调整,最具代表性的是 Brassil 等人提出的行间距编码、字间距编码和特征编码三种编码方法[3~5]。

(1) 行间距编码。

此方法通过垂直移动文本行的位置实现,通常当一行文本被上移或下移时,与其相邻的两行或其中的一行文本保持不动,不动的相邻行被看作是解码过程中的参考位置。可以规定行上移表示“0”,下移表示“1”。根据要嵌入文本中信息的二进制位内容,编码器将文本中若干行上移或下移来隐藏信息。解码器同样根据文档中相邻行的行

间距离的差别进行信息的提取。此方法可嵌入的隐藏信息量较小,16 行的文本才能嵌入 1 个字节的信息量。

(2)字间距编码。

与行间距编码方法类似,字间距编码方法通过使文本行内字符发生平移,即利用字间距离的变化嵌入需要隐藏的信息。采用这种方式时,相邻字之间的距离各不相同。此方法比行间距编码隐藏的信息量大,8 个字符可隐藏 1 个字节的信息量。

(3)特征编码。

在特征编码方法中,观察文本文档并选择一些特征量,再根据要嵌入的数据修改这些特征。特征可以是 b,d,h,k 等字母中的垂直线,其长度可稍作修改以使得一般人不易发觉。相对某种给定的字体可以改变其字符高度,总有一些字母特征未作改变以帮助解码。此方法可隐藏的信息量大小与字间距编码方法差不多,最多 8 个字符可隐藏 1 个字节的信息量。

(4)行尾附加空格编码。

此编码方法的一种典型方式是在每行的行尾插入空格,每行后最多有几个空格是事先约好的[6]。例如每行后最多有 16 个空格,则编码为 4 位。每 2 行可隐藏 8 位(1 个字节)的信息量。

1.3 颜色特征替换法

根据人眼对蓝色最不敏感的特征,刘豪等人提出了通过修改文本字符的蓝色成分使其嵌入隐藏信息的方法[7]。刘豪等人称该方法的信息隐藏空间可提高到文本的字符数,即 1 个字符可嵌入 1 个字母的信息量。而要达到此信息量,只能嵌入 26 个英文小写字母,不能嵌入其他字符,因此应用有很大局限性。

2 提高 Word 文本信息隐藏量的方法及算法实现

2.1 字符缩放编码

在文本文档中,轻微改变字符的大小,人的肉眼是不易察觉的。实现信息隐藏的方法是改变文本中字符的缩放比例,在一篇正常的 Word 文本文档中字符的缩放比例通常是标准形,即 100%,可以采用缩放的比例分别为 100%、101%、102% 和 103%,实现四位二进制码的隐藏而不易被发觉。

例如,若要在文本中嵌入"Gxdx"这 4 个字母,则通过改变文本中的"如果我有一千万,我就能买一栋房子"这 16 个文字的缩放比例来实现:

秘密信息:Gxdx

十六进制 ASCII 码: 47 78 64 78

二进制 ASCII 码: 01000111 01111000 01100100 01111000

缩放比例 100% + :1 0 1 3 1 3 2 0 1 2 1 0 1 3 2 0

载体信息:如果我有一千万,我就能买一栋房子。我有一千万吗?……

通道信息:如果我有一千万,我就能买一栋房子。我有一千万吗?……

与原文相比,对这 16 个字的缩放比例作了轻微改变

使用此编码方式,4 个字即可嵌入 1 个字节的信息量。

2.2 字体 RGB 灰度编码

若同时轻微改变字体 R,G,B 灰度值,则既可提高信息隐藏量又不易被发觉。方法是同时置换设置 R,G,B 低 4 位的值,则可实现 12 位二进制数的隐藏,如图 1 阴影部分所示。

显然,此方法使得 1 个字可嵌入 1.5 个字节的信息量。

2.3 字符下划线 RGB 灰度编码

在绝大多数 Word 文本文档中,字符有下划线的情况是很少的,如果通过改变没有下划线的字体的下划线颜色 RGB 灰度值来嵌入信息(如图 2 阴影部分),那么可进一步提高信息隐藏量。而下划线的可见性仍设为"无",嵌入隐藏信息后的文本视觉上与原文完全一样,因此更不易被察觉。

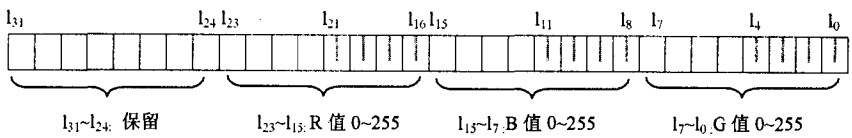


图 1 Word 文档中 RGB 值的长整型数据结构

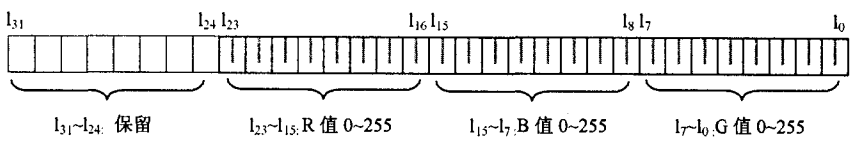


图 2 Word 文档中文本下划线 RGB 值的长整型数据结构

此方法使得 1 个字最多可嵌入 3 个字节的信息量。表 1 给出了各种文本隐藏信息方法信息隐藏量的比较结果。

表 1 各种隐藏方法信息隐藏量的比较

方法种类	编码方式	隐藏 1 字节信息需要的文本量
词义替换法		很多,且没有确定值
空间特征替换法	行间距编码	≈16 行文本
	字间距编码	≈8 个字符
	特征编码	>8 个字符
	行尾附加空格编码	≈2 行文本
	字符大小编码	≈4 个字符
颜色特征替换法	字体蓝色灰度编码	>1 个字符
	字体 RGB 灰度编码	≈0.66 个字符
	下划线 RGB 灰度编码	≈0.33 个字符

字符下划线 RGB 灰度编码算法描述如下:

(1) 嵌入算法的实现过程:

① 判断隐藏信息能否嵌入载体文本中。

② 顺序取隐藏信息的每个字节的 ASCII 码值。

③ 顺序寻找载体文本中适合嵌入信息的字符,根据隐藏信息的每个字节的 ASCII 码值,分别替换该字符下划线颜色的 R,G,B 值。为提取秘密信息时方便,可先把秘密信息的文件长度、文件名等信息嵌入到载体文本文件中。

(2) 提取算法的实现过程:

① 查找隐藏信息在载体文本中的位置。

② 顺序取出字符下划线颜色的 R,G,B 值,转换成相应字符 ASCII 码值,恢复秘密信息的长度及文件名,生成秘密信息文件。

例如,将秘密信息"Gxdx 广西大学"进行隐藏的过程如下:

秘密信息:Gxdx 广西大学

十六进制 ASCII 码:47 78 64 78 B9E3 CEF7 B4F3 D1A7

下划线 RGB 值:477864 78B9E3 CEF7B4 F3D1A7

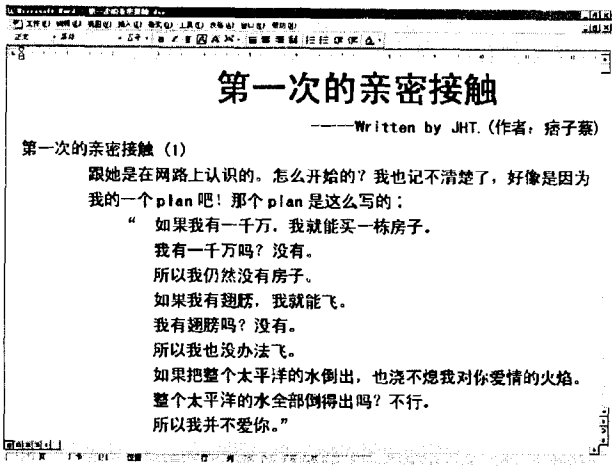
RGB 值(十进制):(71,120,100)(120,185,227)(203,247,180)(243,209,167)

载体信息:如果我有一千万,我就能买一栋房子。我有一千万吗? ……

通道信息:如果我有一千万,我就能买一栋房子。我有一千万吗? ……

与原文相比,对这 4 个字符的下划线 RGB 值作了改变,而下划线的可见性仍设为"无"

为了检验采用字符下划线 RGB 灰度编码算法进行 Word 文本信息隐藏的效果,进行如下实验:选择的秘密信息文本共 576 个字符、采用的载体文本为 Word2000 文档 404 个字符、文件长度 19456 字节,嵌入秘密信息后生成的通道信息文本仍为 Word2000 文档 404 个字符、文件长度变为 23552 字节。图 3 和图 4 分别是嵌入秘密信息文本和未嵌入秘密信息文本的对比,图 5 则是检测到的隐藏信息文本。



第一次的亲密接触

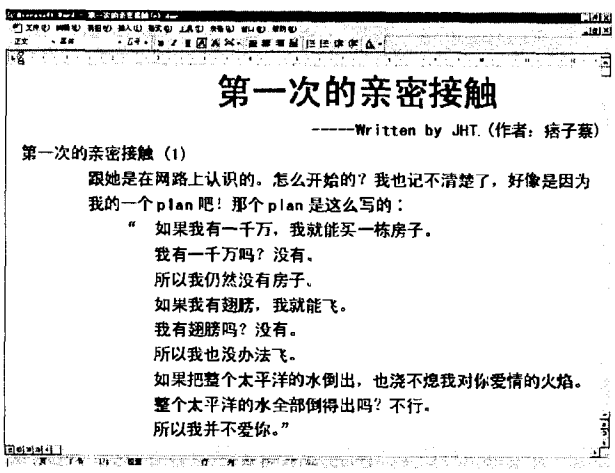
——Written by JHT. (作者:痞子蔡)

第一次的亲密接触 (1)

跟她是在网路上认识的。怎么开始的?我也记不清楚了,好像是因为我的一个 plan 吧!那个 plan 是这么写的:

" 如果我有一千万,我就能买一栋房子。
我有一千万吗?没有。
所以我仍然没有房子。
如果我有翅膀,我就能飞。
我有翅膀吗?没有。
所以我也没办法飞。
如果把整个太平洋的水倒出,也浇不熄我对你爱情的火焰。
整个太平洋的水全部倒得出吗?不行。
所以我并不爱你。"

图 3　未嵌入秘密信息文本的载体文件



第一次的亲密接触

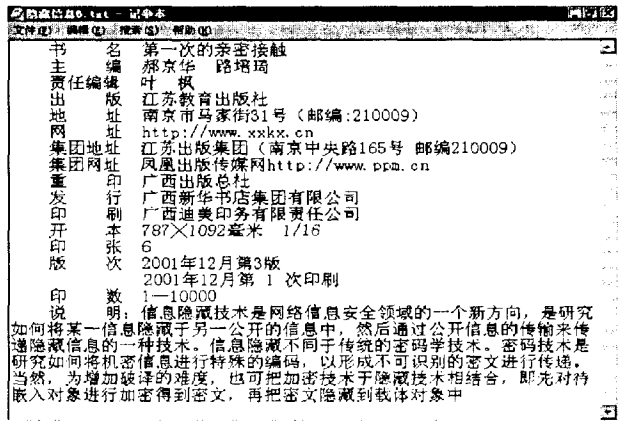
——Written by JHT. (作者:痞子蔡)

第一次的亲密接触 (1)

跟她是在网路上认识的。怎么开始的?我也记不清楚了,好像是因为我的一个 plan 吧!那个 plan 是这么写的:

" 如果我有一千万,我就能买一栋房子。
我有一千万吗?没有。
所以我仍然没有房子。
如果我有翅膀,我就能飞。
我有翅膀吗?没有。
所以我也没办法飞。
如果把整个太平洋的水倒出,也浇不熄我对你爱情的火焰。
整个太平洋的水全部倒得出吗?不行。
所以我并不爱你。"

图 4　嵌入秘密信息文本后的通道文件



书　　名　第一次的亲密接触
主　　编　郝京华　路塔琦
责任编辑　叶　枫
出　　版　江苏教育出版社
地　　址　南京市马家街31号 (邮编:210009)
网　　址　http://www.xxkx.cn
集团地址　江苏出版集团 (南京中央路165号 邮编210009)
集团网址　凤凰出版传媒网http://www.ppm.cn
重　　印　广西出版总社
发　　行　广西新华书店集团有限公司
印　　刷　广西迪美印务有限责任公司
开　　本　787×1092毫米　1/16
印　　张　6
版　　次　2001年12月第3版
　　2001年12月第 1 次印刷
印　　数　1—10000
说　　明:信息隐藏技术是网络信息安全领域的一个新方向,是研究如何将某一信息隐藏于另一公开的信息中,然后通过公开信息的传输来传递隐藏信息的一种技术。信息隐藏不同于传统的密码学技术。密码技术是研究如何将机密信息进行特殊的编码,以形成不可识别的密文进行传递。当然,为增加破译的难度,也可把加密技术于隐藏技术相结合,即先对待嵌入对象进行加密得到密文,再把密文隐藏到载体对象中

图 5　检测到的隐藏信息文本

从实验的结果看,打开后的文档人眼是无法察觉所隐藏的秘密信息的。

3　结束语

与其他文本信息隐藏方法相比,文中提出的信息隐藏方法具有如下优点:可加载的信息量大、不可察觉性较好、具有完全解码独立性、算法实现简单等。虽然与其他文本信息隐藏算法一样,所给出的方法的鲁棒性也比较脆弱,攻击者如果重设文本字符下划线的颜色或重新复制文件内容,则可使隐藏信息改变或者消失。但是利用这一特性,可以将信息隐藏方法应到数字认证领域以检验文本是否被修改或者是否被非法拷贝。

参考文献:

[1] 王炳锡,陈　琦,邓峰森. 数字水印技术[M]. 西安:西安电子科技大学出版社,2003.

[2] Bender W. Techniques for data hiding[J]. IBM Systems Journal,1996,35(3-4):313-336.

[3] Brassil J,Low S,Maxemchuk N,et al. Electronic marking and identification techniques to discourage document copying[A]. In Proc Inforcom'94[C]. [s.l.]:[s.n.],1994. 1278-1287.

[4] Brassil J,Low S,Maxemchuk N,et al. Electronic marking and identification techniques to discourage document copying[J]. IEEE J Select Areas Commmon,1995,13:1495-1504.

[5] Brassil J,Low S,Maxemchuk N,et al. Copyright protection for the electronic distribution of text document[J]. Proc IEEE,1999(7):1181-1196.

[6] 傅　瑜,王保保. 文本水印附加空格编码方法的实现及其性能[J]. 长安大学学报(自然科学版),2005,22(3):85-87.

[7] 刘　豪,孙星明,刘晋飚. 基于字体颜色的文本数字水印算法[J]. 计算机工程,2005,31(15):129-131.