文本伪装算法研究

钮心忻,杨义先
(北京邮电大学信息安全中心,北京 100876)

摘 要: 本文提出了文本伪装的一种新的算法,它是将不具有冗余度的文本信号通过信号处理的变换后,得到具有冗余度的信号,再对冗余信号进行文本的伪装. 其效果是,该算法可以实现用一段普通文本来掩饰机密文本的传输,并且该算法具有一定的抵抗干扰的能力.

关键词: 文本伪装; 信息隐藏; 冗余度
中图分类号: TP335⁺.2 **文献标识码**: A **文章编号**: 0372-2112 (2003) 03-0402-04

Research on the Algorithm of Text Steganography

NIU Xin-xin, YANG Yi-xian
(Beijing Unviersity of Posts & Telecommunications, Beijing 100876, China)

Abstract: A new algorithm of text steganography is proposed. The main idea of the algorithm is to transform the text signal without redundancy into redundant signal. The text steganography is then implemented on the redundant signals. By using this algorithm, a secret text can be replaced by a pubilic text. The algorithm can survive from communication noise.

Key words: text steganography; information hiding; redundancy

1 引言

在信息伪装研究领域中,研究的是如何对机密信息增加一层伪装色,使得机密信息的传输不会引起注意,从而实现隐蔽通信. 在信息伪装中,通常使用的载体一般是图像、声音、视频和文本,文献[1]中全面论述了信息伪装和数字水印的发展历史和研究现状,对目前在图像、声音、文本等方面的研究进行了论述和总结. 从目前发表的大量研究论文看,研究得最多和最深入的是在图像载体中隐藏信息和嵌入数字水印,这一方面是由于图像处理的直观性,另一方面是由于图像中存在大量的冗余信息,由于这些冗余信息的存在,使得我们可以在其中隐藏一些信息,而不致引起观察者的怀疑. 同样,对于声音信号,它也存在大量的冗余信息,因此在声音中也可以进行信息的隐藏或者水印的隐藏. 但是对于文本信号就不同了,文本信号中不存在冗余,文本的一个比特发生变换,文本就发生错误,因此在文本中进行信息伪装的方法就不同于在图像和声音信号中的方法. 文献[1]中总结了在目前发表的论文中,关于文本伪装的一些方法,它们主要是利用文本字符的字间距、行间距、标点符号等位置隐藏几个比特的信息,而这些信息当对文本进行文字编辑,或者重新读取并存盘后,就会消失. 另外文献[1]中还提到了另外一种方法,称为自由上下文语法,它是用一些常用主、谓、宾单词根据要隐藏比特进行组合,组合出一些具有正常含义的句子. 这种方法的适用面很局限,而且组出的句子即使具有正常含义,其上下文也无法形成一段正常的文字,容易引起怀疑.

本文从一个全新的角度来研究文本信息的伪装. 本文提出将不具有冗余度的文本信号经过变换后,得到具有冗余度的信号,再在冗余空间中进行文本的伪装. 首先,把文字以其编码方式读入,为一串编码数字,当然这些数字与文字一一对应,数字发生微小的变化,将引起文字的错乱. 而将这串数字信号进行某种变换,在变换域的信号就可以允许有误差,这点微小的冗余可以考虑用来进行文本的伪装.

本文第二部分对文本伪装的算法进行了介绍,分析了算法特性和抵抗攻击的能力,第二部分给出了仿真结果,最后对文本伪装进行了总结.

2 文本伪装算法

首先,根据编码方式,可以把所有文字以它的编码方式读入(如 0 至 127 是 ASCII,128 以上是汉字编码),这些编码数字是以整数形式存在的,它们不存在任何的冗余,数字发生微小的变化,将引起相应文字的错误. 为了在没有冗余的文字编码中引入冗余,我们现将这串数字以它的比特流表示,将这串 0 和 1 组成的比特流进行某种变换,如小波变换、FFT 变换、DCT 变换等,在变换域中的这串数字就具有了一些冗余度,比如,变换域中的数字产生了微小的变化,而进行相应的逆变换,数

收稿日期:2001-08-27;修回日期:2002-02-26
基金项目:国家自然科学基金项目(No. 60073049);国家重点基础研究发展规划项目(No. TG1999035804);国防科技保密通信重点实验室基金项目(No. 51436060101DZ0801)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

据取整后仍然变为原来的 01 比特串,那么在变换域中冗余范围之内的微小变化,就没有影响原来的文本信号.

然后,考虑在冗余的信号中进行信息的伪装. 假设有一个普通文本 p 和一个机密文本 s,机密文本的传输需要以普通文本做掩护. 首先将普通文本和机密文本都变为具有冗余的变换域内的信号 p_w 和 s_w,然后将这两个信号进行归一化,变为 $[0,1]$ 内的信号 p_{wn} 和 s_{wn},然后对 p_{wn} 进行压缩编码,我们采用的编码方式是,根据精度要求,选用一个具有 2^n 个等级的码本,将 p_{wn} 的每一个值与这个码本进行比较,每一个值用它在码本中的序号来代替,这样就得到了一个具有误差的对 p_{wn} 的编码. 这种编码方式类似于图像的编码,即图像的象素值用与其对应的调色板的序号来代替. 同样,对机密文本的归一化信号用同一个码本进行编码,对这两个信号的编码值进行运算(如相加或异或等),运算后的值作为密钥发送给接收方.

在这个算法中,需要秘密传给接收方的信息有:密钥,码本的选择,机密信号归一化时的最大值和最小值.

在接收方,接收者收到公开发来的文本 p,以及秘密发来的密钥、码本的定义、机密信号归一化时的最大值和最小值. 首先对公开的文本 p 进行冗余化处理,变为 p_w,再进行归一化,变为 p_{wn},用约定的码本进行编码,得到 p_{wn} 信号的编码序号,将这个编码序号与密钥进行与发送方相反的运算(如相减或异或等),就得到了秘密文本相对于码本的编码序号,根据这个序号和码本可以得到 s_{wn} 的信号值,将它进行反归一化,再进行冗余化的逆向处理,就可以得到原始文本的 01 比特流. 但是由于编码的误差,以及传输过程中密钥有可受到微小噪声的干扰,因此恢复的信号不是单纯的 01 比特流,而是一些实数,将这些实数取整,大于 0.5 的判为 1,小于 0.5 的判为 0,得到一个 01 比特流. 当误差以及干扰在一定范围之内时,01 比特流可以精确恢复,也可以精确恢复机密文本.

在这个算法中,需要讨论以下几点:

(1) 密钥的发送过程. 在这个算法中,我们将两个文本数据冗余化,再进行编码,并将两个文件的编码序号进行运算,产生密钥. 这个密钥是从两个文本的冗余化数据中得来的,它与两个文本文件密切相关. 另一方面,尽管进行了编码,密钥的数据量还是远远大于原始机密文本的数据量. 在这里我们考虑的是,首先,用一个普通文本的传输来掩盖机密文本的传输,以达到不引起攻击者怀疑的目的. 第二,这样产生的密钥,尽管数据量增大,但是它存在部分冗余,就是说,密钥在传输过程中如果受到一定的人为破坏或噪声干扰,仍然不影响恢复机密文本的正确性. 因此密钥传输时可以考虑在一个公开的图像或者声音文件中进行隐藏,接收者收到图像或者声音文件后,提取出隐藏的密钥,再进行文本的恢复. 这个算法的一种应用方式是,传输一幅公开的图像(或声音),并附带这幅图像(或声音)的文字说明,这个文字说明做为可公开的文本,将需要隐蔽传输的秘密文本与这个公开文本用文本伪装算法进行处理,得到的密钥可以隐藏在这幅图像(或声音)中,这样秘密信息就包含在这个图像(或声音)及其文字说明中,在接收端用文本恢复算法可以恢复出秘密文本.

(2) 码本的选择和约定. 在算法中,发送方和接收方需要事先选择或者约定一个共同的码本,比如最简单的是一个线性函数($y = i/N$, $N = 2^n$, $i = 0,1,\cdots,N$),或者是一些单调上升或单调下降的曲线. 当然,选择码本函数越复杂,伪装的安全性就能提高.

(3) 算法抵抗干扰的能力. 如果直接在图像或者声音文件中隐藏原始机密文本,由于其不存在冗余度,图像或声音等载体受到少许破坏或干扰,都会造成错误的文本恢复. 而我们提出的算法中,采用了对文本信号进行冗余化处理的技术,使得密钥存在一定的冗余度,因此可以抵抗一定的人为破坏和干扰.

(4) 冗余化变换的选择. 在算法中,我们采用了一个冗余化处理技术,通过大量试验发现,对文本文件的 01 比特流进行小波变换,其冗余化的效果最好. 它使得在密钥的传输时,叠加了 1.5 倍的噪声仍然能够精确恢复原始机密文本.

3　仿真结果

我们对算法做了大量的仿真试验. 在这个算法中,注意到机密文本和普通文本要求数据量是一样大的. 我们取这样两个文本,图 1(a) 作为普通文本,图 1(b) 作为机密文本.

```
figure                                              (a)  普通文本
hold on
plot(m4_100_db101v4(:,1),m4_100_db101v4(:,2),m4_100_
db101v4(:,1),m4_100_db101v4(:,3),'*')
plot(m4_100_db41v1(:,1),m4_100_db41v1(:,2),m4_100_db41v1
  (:,1),m4_100_db41v1(:,3),'*')
plot(m4_100_db41v4(:,1),m4_100_db41v4(:,2),m4_100_db41v4
  (:,1)
```

信息隐藏与检测算法的特性分析　　(b)　机密文本

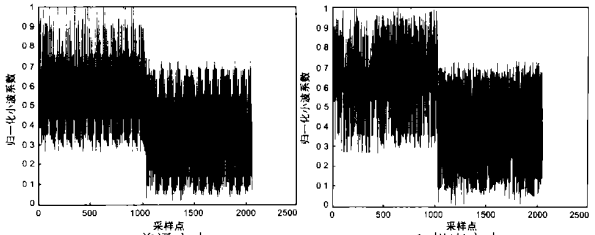
钮心忻,杨义先

(北京邮电大学信息安全中心,北京 100876)

摘要:本文从系统模型的角度研究了信息隐藏与检测的问题,提出了用参数估计理论来衡量信息提取算法的优劣. 同时,用信息隐藏中常用的两种算法和四种应用环境为例,推导了参数估

图 1

这两个文本的大小均为 256 字节. 首先分别对它们的 01 比特流信号进行冗余化,采用一级分解的小波变换,小波基采用"Daubechies5". 在变换域中的信号归一化后,其数据如图 2 所示. 然后,归一化的数据用一个 16 级的线性码本来编码,其



(a) 普通文本　　(b) 机密文本

图 2　两个文本文件在小波变换域中的波形

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

码本见图 3 所示. 编码后的数据见图 4,在这里,密钥取为两个文本序号的和,见图 5(*a*). 密钥传输时,考虑受到噪声的干扰,我们用一个[- 1,1]内均匀分布的白噪声来模拟,将这个白噪声叠加到密钥上并取整,密钥会发生 + 1 和 - 1 的变化,叠加了噪声的密钥见图 5(*b*). 接收端用此密钥减去公开文本的序号,得到秘密文本的序号(图 6),这个编码的序号已受到 + 1 和 - 1 噪声的干扰. 再用它通过码本还原,通过逆向小波变换,得到恢复的秘密文本见图 7 所示,得到了精确的恢复.

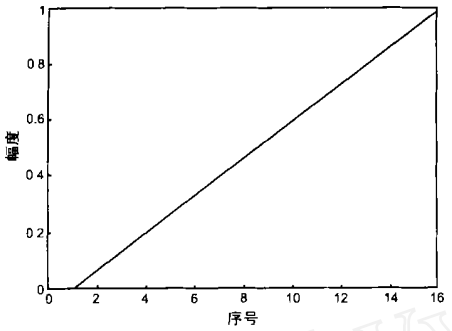


图 3 线性码本

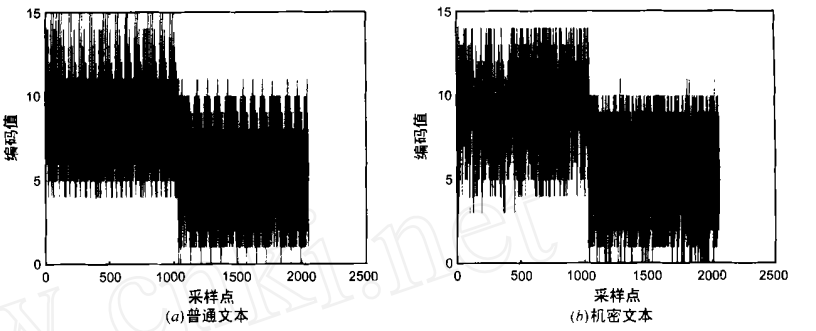


图 4 两个文本文件在小波变换域中编码后的波形

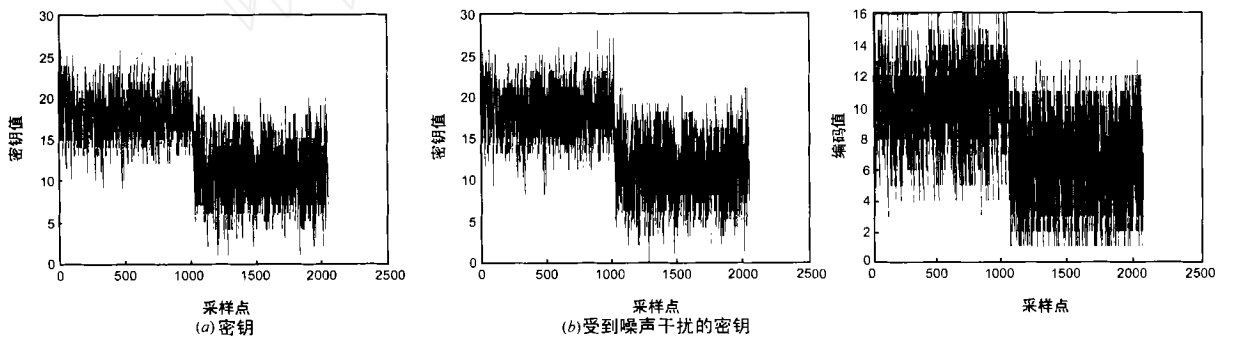


图 5 密钥及受到干扰的密钥

图 6 恢复的受到干扰的机密文件的编码值

如果受到的噪声强度加大,如噪声在[- 1.5,1.5]内均匀分布,恢复的文本会产生少量的误差,如图 8 所示,在这 256 个字节中,有 2 个字节发生变化,产生两个中文字符的错误.

信息隐藏与检测算法的特性分析

钮心忻,杨义先
(北京邮电大学信息安全中心,北京 100876)

摘要:本文从系统模型的角度研究了信息隐藏与检测的问题,提出了用参数估计理论来衡量信息提取算法的优劣. 同时,用信息隐藏中常用的两种算法和四种应用环境为例,推导了参数估

图 7 精确恢复的秘密文本

信息隐藏与检测算法的特性分析

钮心忻,杨义先
(北京邮电大学 息安全中心,北京 100876)

摘要:本文从系统模型的角度研究了信息隐藏与检测的问题,提出了用参数估计理论来衡量信息提取穗法的优劣. 同时,用信息隐藏中常用的两种算法和四种应用环境为例,推导了参数估

图 8 噪声强度提高时,恢复的秘密文本

为了说明问题,对这 256 字节的文本做了大量的试验,其中,每一次叠加的噪声是随机的. [- 0.5,0.5]内均匀分布白噪声时,200 次仿真中没有错误;如噪声在[- 1,1]内均匀分布,其出现错误的概率为 256 字节中有 0.385 个字节发生错误;当噪声为[- 1.5,1.5]时,平均错误率为 1.27 个字节;当噪声强度提高到[- 1.6,1.6]时,平均误差率为 6.58 个字节. 图 9 给出了噪声强度与字节错误率的关系.

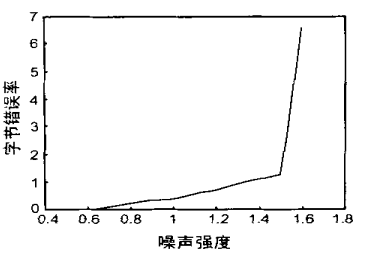


图 9 字节错误率与噪声强度的关系

我们注意到,当均匀分布白噪声的幅度大于[- 1.5,1.5]范围时,算法恢复的机密文本的字节错误率大幅度提高. 而[- 1.5,1.5]的噪声意味着密钥受到 - 2 到 + 2 的干扰.

本算法的缺点是密钥的数据量大,交换密钥时需要利用信息隐藏技术将密钥隐藏在具有较大冗余空间的图像或者声音文件中.

而此算法的优点有二,首先,算法具有抵抗干扰的能力. 因为在算法中,我们采用了对文本信号进行冗余化处理的技术,使得密钥存在一定的冗余度,因此密钥在传输过种中可以抵抗一定的人为破坏和干扰. 第二,采用了密钥与明文分离的传输方式,提高了安全性. 如果直接在图像或者声音文件中隐藏原始机密文本,攻击者可以直接在隐藏了信息的图像或声音文件上进行破坏或者破译. 而使用此算法隐藏的只是密钥,

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

即使攻击者提取出了密钥,但是找不到相应的明文的话,仍然无法恢复出密文.因此在应用时甚至不需要传输明文,在双方共同约定某一段选定的明文后,只需要传输密钥就可以进行秘密的文本传输了.

4 结论

本文提出了一种全新的文本伪装算法,它不同于以往的文本伪装的思路,而是将不具有冗余度的文本信号变换为具有一定冗余度的信号,在此冗余信号中进行信息的伪装.运用该算法可以将一段机密文本变为一段普通的文本,从而掩盖了机密文本传输的事实.而密钥的传输可以隐藏在普通的图像或者声音中,并且该算法产生的密钥可以抵抗干扰和噪声.

参考文献:

[1] Stefan Katzenbeisser, Fabien A P Petitcolas. Infomation Hiding Techniques for Steganography and Digital Watermarking [M]. Artech House Publishers, 2000.

[2] Low S H, N F Maxemchuk, A M Lapone. Dcument identification for copyright protection using centroid detection [J]. IEEE Transactions on Communication, 1998, 46(3): 372 - 383.

[3] Stefan Katzenbeisser, et al(美). 信息隐藏技术 ——隐写术与数字水印 [M]. 吴秋新, 钮心忻, 杨义先, 罗守山, 杨晓兵, 译. 北京: 人民邮电出版社, 2001.

作者简介:

钮心忻　女,1963 年 10 月出生于北京,北京邮电大学信息安全中心,博士,副教授,主要研究方向为信息伪装与数字水印、网络与信息安全、数字信号处理、软件无线电等.

杨义先　男,1961 年 3 月出生于四川省盐亭县,北京邮电大学信息安全中心,博士,教授,博士生导师,首批长江学者特聘教授,全国政协委员。主要研究领域包括网络信息安全、编码密码学、伪装式信息安全、应用数学等.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net