基于双文本段的信息隐藏算法

陈志立[①]　黄刘生[①②]　余振山[①]　杨　威[①②]　陈国良[①]
[①](中国科学技术大学计算机科学与技术系国家高性能计算中心　合肥　230026)
[②](中国科学技术大学苏州研究院　苏州　215123)

摘　要：信息隐藏是一种在传输或存储过程中将隐秘信息隐藏在特定载体中，以保证隐秘信息安全性的技术。常用的载体有图像、音频、视频、文本等类型文档。由于文本文档特别是纯文本文档中的冗余信息非常少，基于纯文本文档的信息隐藏具有很大的挑战性。现存的基于纯文本文档的算法都是基于单文本段的，在安全性方面还存在许多难以克服的缺陷。该文提出了一种新的基于双文本段的信息隐藏算法，通过在多种隐藏形式中选择适当的隐藏形式和信息分散存储，大大地提高信息隐藏的隐蔽性、安全性。另外，算法具有很高的灵活度，可以根据具体的应用情景进行适当的变形或调整，以便更好地适用于实际需求。

关键词：信息隐藏；同义词替换；双文本段；异或分解
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2009)11-2725-06

An Information Hiding Algorithm Based on Double Text Segments

Chen Zhi-li[①]　Huang Liu-sheng[①②]　Yu Zhen-shan[①]　Yang Wei[①②]　Chen Guo-liang[①]
[①](*National High Performance Computing Center*, *Department of Computer Science and Technology*, *University of Science and Technology of China*, *Hefei* 230026, *China*)
[②](*Suzhou Institute for Advanced Study*, *University of Science and Technology of China*, *Suzhou* 215123, *China*)

Abstract: Information hiding is a technique that hides secret messages in some carrier during transmission or storage. Carriers in common use include image, audio, video, text documents and so on. Since there is little redundancy in text documents, particularly in plain text documents, information hiding based on plain text documents is much more challenging. Previous algorithms based on plain text documents are all based on a single plain text segment. Therefore, there are many inherent limitations in security of them. In this paper, a novel information hiding algorithm based on double text segments is proposed. The algorithm can enhance the concealment and security of information hiding greatly by choosing a proper hiding form out of many ones and scattering information to places. In addition, the algorithm is so flexible that it can be modified or adjusted according to the certain application scene to fit the practical requirements better.

Key words: Information hiding; Synonym replacement; Double text segments; XOR decomposition

1　引言

近年来，信息隐藏已经成为信息技术领域一个新兴的研究方向，它在版权保护、秘密传输与存储、隐蔽信道和匿名技术等方面有广泛的应用。目前，基于图像、音频、视频载体的信息隐藏的研究已经比较成熟。由于文本中存在很少的冗余信息，把隐秘信息嵌入其中比较困难，文本信息隐藏的成果并不多。然而，由于文本特别是纯文本具有占用空间很少，处理方便直观等优点，同时文本文档在信息的存储和传输中仍然占据着统治地位，基于文本的信息隐藏仍然很有吸引力。本文主要研究基于纯文本的信息隐藏算法。目前，基于纯文本的信息隐藏算法大体上可以分为三大类，分别是：基于排版、基于语法和基于语义。

基于排版的算法实质是利用纯文本排版产生的冗余信息，来进行信息隐藏。这类算法不会改变载体文本的文本信息，对载体文本所表达的意义不会有影响[1−3]。这类算法通常利用行尾空格个数变化，标点符号前后是否加空格以及中英文标点符号替换等来嵌入信息。基于排版的信息隐藏算法的隐藏容量很小；鲁棒性也很差；隐藏信息后的文本比较容易检测出来[4]，甚至有可能还原出的隐秘信息，从而算法的安全性也很值得担忧，限于它的缺陷，很难走向实用。

基于语法的算法是利用自然语言的语法结构来嵌入隐秘信息的。这类算法与其它的两类一个明显的不同是，它的载体文本是算法在隐秘信息的控制

2008-11-14 收到，2009-04-06 改回
国家自然科学基金重大研究计划(90818005)，国家自然科学基金(60773032，60703071)，教育部博士点基金(2006CB303006)和江苏省自然科学基金(BK2007060)资助课题

下产生的，载体文本本身并没有完整的意义，而仅仅是由符合语法但意义上却杂乱无章的句子构成。这类算法的例子有基于 Markov 链的隐藏方法[5]，基于句子模板的隐藏方法[6]和基于文章样式的隐藏方法[7,8]等。基于语法的信息隐藏算法生成的文本没有完整的意义，通过人眼视觉很容易察觉载体文本的异常；可以通过语义分析、统计分析等方法，来对载体文本实现自动化的检测[9−12]。

基于语义的算法是通过同义替换来隐藏隐秘信息的，它希望在把隐秘信息嵌入载体文本的同时，尽可能地维持载体文本的语义不变。这类方法根据替换成份可分为基于同义词或者同义短语的替换[13−15]、基于缩写的替换和基于同义句子的替换[16]等。它们的算法思想类似，都是把表达相同或相近意思的多种表达形式进行编码，然后根据要隐藏的隐秘信息选取某种表达形式，来替换载体文本中相应的部分。基于语义的算法可以比较好地维持载体文本语法正确和语义不变，检测比较困难。但是这类算法嵌入率不高，同时实现起来需要涉及到自然语言处理方面的一些难题，实现难度比较大。

据我们所知，现存的基于文本的信息隐藏算法都是基于单文本段的，它们把隐秘信息嵌入单一的文本段中。由上述可知，这些算法在给定隐秘信息的情况下，隐藏后的文本形式一般就确定下来。这一点对于同义替换尤为不利。比如，对同义词替换来说，很难保证两个词在意义上真的就完全等价，即使意义上完全等价的两个词，也很难保证它们在用法以及表达风格等其它方面上也完全一样。因此，即使是用像文献[13]那样严格筛选出来的同义词库进行同义替换，也不能消除替换后在习惯用法，表达风格等方面造成的不良影响[17−19]。

本文提出的基于双文本段的信息隐藏算法，截取了特定比特数目的隐秘信息，对其进行异或分解，并使用双载体文本段分别隐藏分解后的信息，把隐秘信息进行空间分散。由于每一次异或分解的形式有多种，隐藏同样的隐秘信息，隐藏后的文本多种多样，故算法可以灵活地选择比较安全的文本形式进行隐藏，以提高隐藏算法的隐蔽性和安全性。故算法的优点可以总结如下。(1)对同一隐藏信息，隐藏形式有多种变化，可以根据需要选择安全性高的隐藏方式，提高安全性和隐蔽性。比如可以根据某种检测算法或攻击方式来选择合适的隐藏形式。(2)把隐藏信息进行分解继而空间分散，通过任一部分的信息都无法还原出原信息，提高了安全性；(3)算法具有很高的灵活性，既可以对算法本身进行修改变形(详见第 3，4 节)，又可以修改(1)中的选择标准，抵御不同的检测算法和攻击方式。(4)与其它基于语义的算法类似，算法能抵抗对载体文本进行的重新排版，甚至修改或删除非关键的文本内容(不影响文本中的同义词的分布)等变换，因而具有一定的鲁棒性。

文章的结构安排如下：第 2 节详细描述了新的隐藏算法，分析了算法的安全性；第 3 节给出了算法的两种应用情景；第 4 节给出算法的 3 种扩展；第 5 节是文章总结。

2 新算法描述

2.1 算法的基本思想

目前存在的基于文本的信息隐藏算法一般都是基于某个文本段的，隐藏算法的所有操作都只作用于该文本段，也就是说如果知道足够多的信息，肯定可以通过这个文本段还原出被隐藏的隐秘信息。那么，如果把隐秘信息隐藏至多个文本段，结果将如何呢？可以把一个隐秘信息隐藏在两个文本段(甚至多个文本段)中，而通过其中任何一文本段，都无法还原出任何原来的隐秘信息。把这两个文本段的对应关系作为隐藏算法密钥的一部分，那么在一般情况下只有隐秘信息的发送者和接收者知道这个对应关系，而对此毫不知情的攻击者来说，想要从浩如烟海的互联网信息海洋中获得此对应关系，极其困难。

本文提出的基于双文本段的算法，先把隐秘信息异或分解成两部分，然后分别使用同义词替换算法，嵌入两个载体文本段中。由于对同样的隐秘信息存在多种异或分解形式，隐藏同样的隐秘信息后所得到的载体文本具有多种形式，可以从中选取隐蔽性比较高的形式使用。同时，分解后的隐秘信息分别嵌入不同的载体文本段中，这就把隐秘信息进行分散，使得单个的载体文本段都无法还原出原始的隐秘信息，很大程度地提高了隐藏的安全性。由于同义替换算法具有比较高的安全性，同义词替换具有同义替换的高安全性的优势，实现起来也相对比较容易，对每个部分的隐藏算法选用同义词替换，这样既考虑了隐藏算法本身的安全性，也考虑了实现难度。

2.2 同义词词典和编码

在本文算法中，指定的同义词词典满足如下条件：同义词组中每个词对特定的某个替换都能维持原来的词性；同义词组中大部分词对特定的某个替换都能维持原来的语义；每个词只能属于一个同义词组，也就是说一个同义词组与其它的同义词组都是不相交的；每个词都仅包含一个单词。对满足上

述条件的每个同义词组进行编码。例如，已经知道同义词组 $S=\{s_0,s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$，那么可以用 3 位编码来表示组里的每个同义词，即 000 表示 s_0，001 表示 s_1，…，111 表示 s_7。在同义词词典中，为了编码方便，同义词组的大小都是 2 的若干次方。

2.3 基本算法

隐藏算法主要包含两个动作：异或分解隐秘信息以及把分解得到的各部分隐秘信息分别隐藏在载体文本中。异或分解的方法是：任意长度为 n 的比特串，都可以分解成两个或者两个以上长度也为 n 的有序比特串，满足对所有分解得到的比特串求异或，结果等于原来的比特串。记异或运算符为“$\oplus$”，则有：$1=1\oplus 0$ 或者 $1=0\oplus 1$；$01=00\oplus 01$ 或者 $01=01\oplus 00$ 或者 $01=10\oplus 11$ 或者 $01=11\oplus 10$；…。

需要注意的是，把一个长度为 n 的比特串进行异或分解可以得到 2^n 种分解形式，那么如何选取其中最合适的形式进行分解是算法的关键。本文提出的选择思想是：对每次隐藏，跟据上下文信息评估每个文本段所对应的同义词组中所有词对上下文的不匹配程度，称作每个词的危险系数，然后对所有的异或分解形式计算其所对应的词的危险系数之和，选取危险系数和最小的分解形式作为最佳分解。

隐藏算法首先同步扫描两个载体文本，直至每个文本都有一个词与同义词词典 D 中的某个同义词组中的词语成功匹配，然后在词典中分别查询这两个单词所在同义词组的大小，设分别为 n_1 和 n_2，其中大者为 $n=\max(n_1,n_2)$ 算法从隐秘信息里面取出 $\log_2 n$ 位，把取出的 $\log_2 n$ 位的初始信息按最佳分解异或分解成两部分 s_1 和 s_2，最后通过同义词替换分别把 s_1 和 s_2 隐藏在文本段 T_1 和 T_2 中。重复上面的过程直至隐藏信息取完为止。

还原过程大体相反，这里不再详述。如下所示，算法 1 和算法 2 分别是隐藏算法和还原算法的描述。图 1 和图 2 分别是隐藏算法和还原算法的流程图。其中的密钥为两个文本段的对应关系和编过码的同义词词典。

算法 1　隐藏算法描述

步骤 1　扫描文本段 T_1 和 T_2，直至分别找到词 w_1 和 w_2 匹配词典 D 中的词；

步骤 2　求得词 w_1 和 w_2 在词典 D 中的同义词组的大小分别为 n_1 和 n_2，取 $n=\max(n_1,n_2)$，从隐秘信息 I 中取 $\log_2 n$ 位的比特串 s (不足补 0)；

步骤 3　按照最佳异或分解构造长度为 $\log_2 n$ 的比特串 s_1 和 s_2，使得 $s=s_1\oplus s_2$，并且 s_1 的后 $(\log_2 n-\log_2 n_1)$ 位和 s_2 的后 $(\log_2 n-\log_2 n_2)$ 位都为 0。

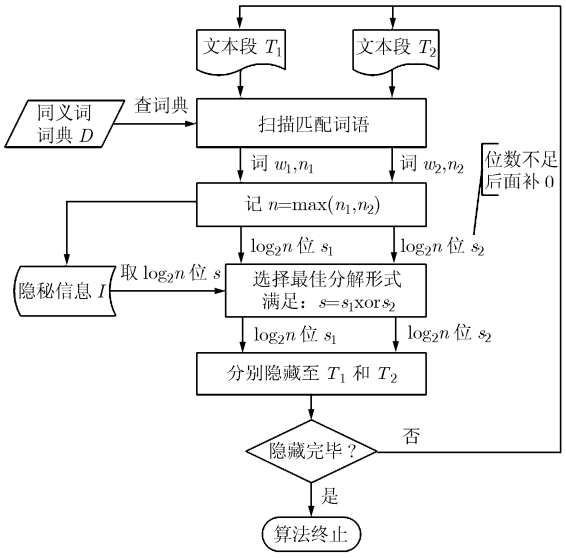

图 1 基于双文本段的隐藏算法流程图

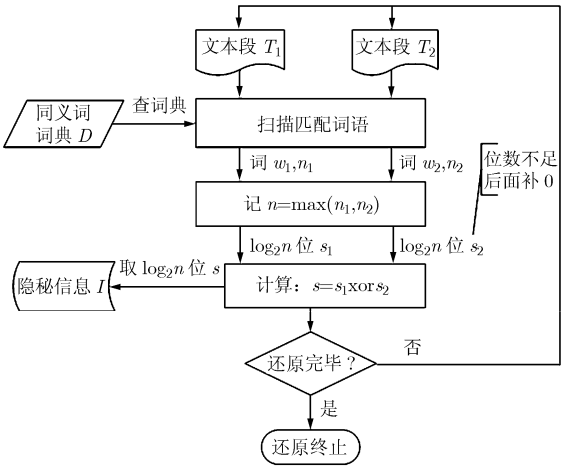

图 2 基于双文本段的还原算法流程图

步骤 4　分别用 s_1 的前 $\log_2 n_1$ 位和 s_2 的前 $\log_2 n_2$ 位作为索引，分别在词典 D 中词 w_1 和 w_2 所在的同义词组里查询对应的单词 w_1' 和 w_2'，并把文本段 T_1 中当前位置的 w_1 替换成 w_1'，把文本段 T_2 中当前位置的 w_2 替换成 w_2'

步骤 5　判断隐秘信息 I 是否已经取尽。若是，算法终止，隐藏信息成功；若到达文本段 T_1 或者 T_2 的末尾，则算法结束，隐藏信息失败；否则，转向步骤 1。

算法 2　还原算法描述

步骤 1　扫描文本段 T_1 和 T_2，直至分别找到词 w_1 和 w_2 匹配词典 D 中的词；

步骤 2　求得词 w_1 和 w_2 在词典 D 中的同义词组的大小分别为 n_1 和 n_2，取 $n=\max(n_1,n_2)$，求 w_1 和 w_2 在各自的同义词组中的编号 c_1 和 c_2，分别取 c_1

的低 $\log_2 n_1$ 位和 c_2 的 $\log_2 n_2$ 位，并分别右移($\log_2 n - \log_2 n_1$)位和($\log_2 n - \log_2 n_2$)位得到 $\log_2 n$ 位的比特串 s_1 和 s_2；

步骤 3　求 $\log_2 n$ 位的隐秘信息 $s = s_1 \oplus s_2$；

步骤 4　判断隐秘信息 I 是否已经还原完成。若是，算法终止，还原信息成功；若到达文本 T_1 或者 T_2 的文件末尾，则算法结束，还原信息失败；否则，转向步骤 1。

2.4 安全性讨论

从直观上来看，对于单文本段信息隐藏，比如单文本段同义词替换，有时候不得不把原词语替换成一个很不合适的词语，因为它的选择是唯一的且由隐秘信息决定的；而对于双文本段则不然，由于存在多个替换的选择，可以不必把原词替换成一个很不合适的词语。

现在对本文提出的算法的安全性进行论证，主要从抗击信息隐藏检测算法的能力来讨论算法的安全性。

现行的检测手段主要包括统计方法[10−12,17−19]或者语义分析的方法[9,10,19]。它们都是利用隐藏算法对载体文本的修改引起统计特性和语义特性的变化来进行检测的。本文的算法对同一个隐秘信息可以有多种隐藏形式，因而可以选取对统计特性和语义特性影响较小的隐藏形式进行隐藏，以提高隐蔽性。下面证明只要适当地选取隐藏形式，基于双文本段的隐藏算法的隐蔽性必然会比基于单文本段的算法高或者与之相等。假设嵌入 n 比特隐秘信息时，文本段 T_1 和 T_2 所对应的同义词组(假设同义词组大小相等，不相等的情况同理可证)分别为

$$\left.\begin{aligned} W_1 &= \{w_{11}, w_{12}, \ldots, w_{1N}\} \\ W_2 &= \{w_{21}, w_{22}, \ldots, w_{2N}\} \end{aligned}\right\} \tag{1}$$

这里 $N = 2^n$。用同义词组中的词替换载体文本段的原词所产生的对检测算法有利的因素，可以用危险系数来表示。设式(1)表示的同义词组所对应的危险系数集合分别为

$$\left.\begin{aligned} R_1 &= \{r_{11}, r_{12}, \ldots, r_{1N}\} \\ R_2 &= \{r_{21}, r_{22}, \ldots, r_{2N}\} \end{aligned}\right\} \tag{2}$$

式(2)表示的集合中，原词对应的危险系数为 0，其它危险系数为非负数。设它们的平均值分别为 r_1 和 r_2。由算法描述知，只能有 N 对可替换的不相交的词(w_{1i}, w_{2j})，对应的危险系数为(r_{1i}, r_{2j})。其中 $1 \le i \le N$，$1 \le j \le N$。假设所有可替换词对被选中的概率相等。用基于单文本段算法嵌入时，两个文本段分别产生的平均危险系数在各个词被选中概率相等的情况下，为它们的所对应危险系数集的均值，即 r_1 和 r_2。那么只要找到一个词对，它们的危险系数之和不大于 $r_1 + r_2$，所证命题就得证。用反证法证。假设找不到上述的词对，那么对任意(r_{1i}, r_{2j})，有

$$r_{1i} + r_{2j} > r_1 + r_2 \tag{3}$$

把所有的可选词对相应的危险系数相加得

$$\sum_{i=1}^{N} r_{1i} + \sum_{i=1}^{N} r_{2i} > N(r_1 + r_2) = \sum_{i=1}^{N} r_{1i} + \sum_{i=1}^{N} r_{2i} \tag{4}$$

由式(3)的假设得到式(4)自相矛盾。故基于双文本段的隐藏算法的隐蔽性必然会比基于单文本段的算法高或者与之相等，而且仅在集合 R_1 和 R_2 中所有的危险系数均为 0 的情况下相等，而这在实际情况下几乎不可能。因此，正常情况下基于双文本段的算法隐蔽性都会明显比基于单文本段的高。　　证毕

在上述的证明中，基于单文本段的算法的隐藏容量约为基于双文本段的两倍，因此后者的隐蔽性比前者好似乎是理所当然的。实际上仍然可以证明，当隐藏同样多的信息时，后者的隐蔽性仍然比前者好。

假设在某次隐藏中，双文本段所对应的同义词组的大小同样都为 N，且两个集合分别有 p 和 q ($0 \le p, q < N$)个词不适合替换。为了隐藏同样多的信息，对基于单文本段算法仅随机选取两个文本段中的一个来隐藏信息，那么它选到不合适替换的词语的概率总是为 $\frac{1}{2}\left(\frac{p}{N} + \frac{q}{N}\right) = \frac{p+q}{2N}$，而对基于双文本段的算法，当 $p + q < N$ 时，由抽屉原理可知总是可以找到一个分解形式，使得替换的词语都是两个集合中适合替换的词语；而当 $N \le p + q < 2N - 1$，仍然可以有一定的概率找到一个分解形式使得两个替换的词语都适合替换，剩余的情况也都能找到一个分解形式使得一个词语适合替换，另一个不适合替换，因此综合起来基于双文本段算法的词语替换效果仍然比基于单文本段的好。由此可知，基于双文本段的算法，由于引进了选择机制，能更好地提高隐藏算法的隐蔽性，进而提高算法的安全性。　　证毕

此外，算法在空间上对隐秘信息进行分散，使得单纯知道各部分的信息都无意义，这也提高了算法的安全性。同时，算法具有一般基于语义的算法的优点，即能抵抗对载体文本进行的重新排版。又由于还原算法仅使用了同义词词典，在不引进，修改或删除文本中所含有同义词词典中的词的情况下，即在不影响文本中的同义词的分布的情况下，

算法可以抵抗修改或删除部分文本内容等变换，因而具有一定的鲁棒性。

3　应用情景

基本算法可以根据实际情况进行调整或者变化，下面举两个应用情景为例：

3.1　应用情景 1：基于互联网上的两个文本文件的算法

假设基本算法中的两个文本段分别来自互联网上的两个不同的文本文件，那么可以把两个文本文件在互联网上对应的 URL 以及同义词词典作为隐藏及还原算法的密钥。隐秘信息的发送者和接收者在通信之前必须进行密钥协商，指定共同使用的两个文本文件对应的 URL 以及同义词词典，然后发送者把隐藏过信息的两个文本文件分别放在互联网上对应的 URL 上，并通知接收者接收；接收者则分别从两个 URL 下载两个文本文件，并还原出隐秘信息。

这个情景比较适用于通信者掌握了比较多的互联网资源的情况。在此情景中，可以很容易把两个文本文件推广为多个文本文件(相应地把基本算法改成基于多文本段的算法，详见第 4 节算法的扩展)。在这个应用情景下，算法把隐秘信息在空间上进行分散隐藏，并可以根据需要调整每次替换时的分解形式，例如可以把分解形式适当调整使得每次替换仅修改多个文本文件中的一个文本文件或者使得隐藏后每个文本文件尽可能地保持原来的语义不变。

3.2 应用情景 2：基于单文本文件的隐藏密度自适应算法

当载体文本的隐藏容量远大于隐秘信息大小时，例如载体文本的隐藏容量为隐秘信息大小的若干倍时，一般更希望隐藏密度也能根据两者的大小关系进行适当的调整。在这个情景中，本文的算法可以很好地做到这一点。

假设载体文本的隐藏容量为隐秘信息大小的 2 倍，那么可以根据隐秘信息的大小，把载体文本分解成两个文本段，使得每个文本段都可以完整地隐藏隐秘信息。这个时候对这两段载体文本运用基本算法(即基于双文本段的隐藏算法)，并使每次替换时采用仅改变一个载体文本段的分解形式，那么将得到隐藏密度为一般同义词替换算法一半的算法。同样的道理，当载体文本的隐藏容量为隐秘信息大小的 n 倍时，可以设计出隐藏密度为原来的 $1/n$ 的隐藏算法。这个情景下的算法具有隐藏密度自适应性。

4　算法的扩展

基本算法可以做多方面的扩展，以下列出几种扩展。

4.1 基于多文本段

可以很容易地把基本算法扩展成基于多文本段的算法。对于隐秘信息的异或分解，可以从分解成两部分扩展到分解成多个部分。因此，可以把分解得的几部分信息分别嵌入对应的载体文本段中。

这种分解有利于更大程度地分散隐秘信息，降低隐藏密度等，可以比较大地提高隐蔽性和安全性，但是隐藏容量相应地降低了。

4.2 基于混合隐藏算法

根据上述基于多文本的扩展，可以把隐秘信息分解得到的各个部分分别用各种隐藏算法隐藏至对应的载体文本中，即得到基于混合算法的扩展。例如可以把隐秘信息分解得到两部分中的一部分用同义词替换来隐藏，另外一部分用 NiceText 来隐藏。

这种扩展有利于各种载体文本段根据各自的具体情况采用适合的算法进行隐藏，具有更大的灵活性，因而算法的隐蔽性和安全性也可以得到有效地提高。

4.3 基于混合数字载体

这种扩展的范围更广泛，不再局限于纯文本了。隐藏算法可以把分解得到的各部分隐秘信息，采用各种嵌入方式，分别隐藏至各种数字载体中。例如，对于图文并茂的文档，可以把隐秘信息分解成两部分，分别隐藏至图像和文本之中。对于电影文件，也可以把隐秘信息分解成两部分，分别隐藏至视频流和音频流之中。

考虑到文本隐藏的困难，一种极端的情况，甚至可以在文本和其它数字载体混合隐藏的时候，采用适当的分解形式，不改变文本的任何信息，而仅仅修改另一种冗余信息比较多，隐藏信息比较容易的数字载体。

这种扩展，可以尽可能地利用各种数字载体的特点，并把它们结合起来，发挥出更佳的信息隐藏效果。

5　结束语

本文提出了一种新的基于双文本段的信息隐藏算法，它首先把待隐藏的隐秘信息采用异或分解的方式进行多种分解，接着根据某种选择标准选择一种最佳的分解，把原信息分解成两部分，然后采用同义词替换的方法，分别将各部分信息隐藏至载体文本段中。算法通过对隐秘信息的多种异或分解形式进行最优选择，提高了隐藏信息后文本的隐蔽性；算法通过对隐秘信息在空间上进行分散，提高了自身的安全性；算法可以制定不同的分解选择标准，

抵抗不同的攻击，也可以进行多种变形，具有很高的灵活性。同时，与一般基于语义的算法类似，本文的算法也可以抵抗对载体文本的重新排版，修改删除非关键文本(不影响文本中的同义词的分布)等操作，具有一定的鲁棒性。

参 考 文 献

[1] Bender W, Gruhl D, and Morlmoto N, *et al*.. Techniques for data hiding [J]. *IBM System Journal*, 1996, 35(3&4): 313–336.

[2] 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.
Cao Wei-bing, Dai Guan-zhong, and Xia Yu, *et al*.. Technology of information hiding based on text document [J]. *Application Research of Computers*, 2003, 20(10): 39–41.

[3] 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.
Bai Jian, Xu Ying-hui, and Yang Yu. An algorithm of text steganography [J]. *Application Research of Computers*, 2004, 21(12): 147–148.

[4] 眭新光, 沈蕾, 燕继坤等. 基于AdaBoost的文本隐写分析[J]. 通信学报, 2007, 28(12): 136–146.
Sui Xin-guang, Shen Lei, and Yan Ji-kun, *et al*.. Text steganalysis using AdaBoost [J]. *Journal on Communications*, 2007, 28(12): 136–146.

[5] 吴树峰. 信息隐藏技术研究[D]. [硕士论文], 中国科学技术大学, 2003.
Wu Shu-feng. Research on information hiding [D]. [Master dissertation], University of Science and Technology of China, 2003.

[6] Maher K. TEXTO. URL:ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz.

[7] Mark C. Hiding the hidden: A software system for concealing ciphertext as innocuous text [D]. [Master dissertation], University of Wisconsin-Milwaukee. http://www.NICETEXT. dissertationcom /NICETEXT/doc/thesis.pdf. 1997.

[8] Mark C, Davida G, and Rennhard M. A practical and effective approach to large-scale automated linguistic steganography [C]. Lecture Notes in Computer Science, 2001, Vol. 2200: 156–167.

[9] 周继军, 杨著, 钮心忻等. 文本信息隐藏检测算法研究[J]. 通信学报, 2004, 25(12): 97–101.
Zhou Ji-jun, Yang Zhu, and Niu Xin-xin, *et al*.. Research on the detecting algorithm of text document information hiding [J]. *Journal on Communications*, 2004, 25(12): 97–101.

[10] Chen Zhi-li, Huang Liu-sheng, and Yu Zhen-shan, *et al*.. Linguistic steganography detection using statistical characteristics of correlations between words [C]. Information Hiding 2008, USA, May 2008, LNCS 5284: 224–235.

[11] Chen Zhi-li, Huang Liu-sheng, and Yu Zhen-shan, *et al*.. A statistical algorithm for linguistic steganography detection based on distribution of words [C]. ARES2008, Spain, Mar 2008: 558–563.

[12] Chen Zhi-li, Huang Liu-sheng, and Yu Zhen-shan, *et al*.. Effective Linguistic Steganography Detection [C]. CIT Workshops, Australia, Jul 2008: 224–229.

[13] Keith W. Lexical steganography through adaptive modulation of the word choice hash. http://alumni.imsa.edu/~keithw/tlex/lsteg.ps.Ms.

[14] Liu Yu-ling, Sun Xing-ming, and Gan Can, *et al*.. An efficient linguistic steganography for chinese text [C]. Proc. IEEE Int. Conf. on Multimedia & Expro (ICME), China, July 2007: 2094–2097.

[15] 甘灿, 孙星明, 刘玉玲等. 一种改进的基于同义词替换的中文文本信息隐藏方法[J]. 东南大学学报(自然科学版), 2007, 37(1S): 137–140.
Gan Can, Sun Xing-ming, Liu Yu-ling, *et al*.. An improved steganographic algorithm based on synonymy substitution for chinese text [J]. *Journal of Southeast University* (*Natural Science Edition*), 2007, 37(1S): 137–140.

[16] Murphy B. Syntactic information hiding in plain text [D]. [Master dissertation], CLCS, Trinity College Dublin. https://www.cs.tcd.ie/Brian.Murphy/publications/murphy01hiding Masters.

[17] Taskiran C M, Topkara U, and Topkara M, *et al*.. Attacks on lexical natural language steganography systems [C]. SPIE, USA, Jan 2006, Vol. 6072: 97–105.

[18] 眭新光, 朱中梁. 字典隐藏法的脆弱性分析与改进. 计算机工程, 2008, 34(8): 144–149.
Sui Xin-guang and Zhu Zhong-liang. Frangibility analysis and improvement of dictionary steganography method. *Computer Engineering*, 2008, 34(8): 144–149.

[19] Yu Zhen-shan, Huang Liu-sheng, and Chen Zhi-li, *et al*.. Detection of synonym-substitution modified articles using context information [C]. FGCN 2008, China, Dec 2008, Vol. 1: 134–139.

陈志立： 男，1980年生，博士生，研究方向为信息安全、信息隐藏.

黄刘生： 男，1957 年生，教授，博士生导师，研究领域为信息安全、高性能算法、分布式计算等.

余振山： 男，1982年生，博士生，研究方向为信息安全、自然语言数字水印.