第 36 卷 第 7 期 计 算 机 工 程 2010 年 4 月
Vol.36 No.7 Computer Engineering April 2010
·安全技术· 文章编号：1000—3428(2010)07—0150—03 文献标识码：A 中图分类号：TP301.6

基于词平台汉字编码的文本信息隐藏算法

张洪礼，刘 丹，温学谦，何新宇
(燕山大学信息科学与工程学院，秦皇岛 066004)

摘 要：文本信息隐藏是版权维护的一种重要手段，针对现有算法存在信息隐藏量不足、鲁棒性不高及多数仅适用于英文文本等问题，提出一种基于词平台汉字编码的文本信息隐藏算法，运用标志位和编码变换规则实现密文信息的嵌入，在算法的信息隐藏量和鲁棒性上有较大提高，增强密文信息的安全性。理论分析和实验结果验证了该算法具有一定实用性。

关键词：信息隐藏；文本；词扩展；顺位编码；逆位编码

Text Information Hiding Algorithm Based on Chinese Characters Coding in Words Platform

ZHANG Hong-li, LIU Dan, WEN Xue-qian, HE Xin-yu
(School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

【Abstract】Information hiding is an important way of copyright maintenance. But the existed algorithms have many problems in hiding capacity, robustness, and most of these are just suitable for English documents. Aiming at those problems, a novel information hiding algorithm is presented, which is based on Chinese characters coding method in words platform. The algorithm implements text embedding by using coding-transform rules and sign bits. The algorithm is improved greatly on information hiding capacity and robustness, and is enhanced in security of secret information. Practicability is proved by theoretical analysis and experimental results.

【Key words】information hiding; text; words expansion; sequenced coding; inverted sequence coding

1 概述

随着多媒体技术的飞速发展和计算机网络的成熟，各种信息媒体通信越来越频繁，这必然带来通信信息安全的问题。信息隐藏技术是近几年来信息安全领域出现的一种新技术，载体形式可为图像、声音、视频或一般的文档等。其中，文本信息隐藏技术是将代表著作权人身份的特定信息，按照某种方式嵌入电子出版物中，在产生版权纠纷时，通过合法的发行者、运营者相应的算法提取出该隐藏的信息，从而验证版权的归属，确定泄漏与泛滥渠道，确保数字产品著作权人的合法利益，避免非法盗版的威胁。

2 相关研究

按嵌入信息的策略不同，文本信息隐藏算法主要可分为 2 大类：(1)基于文档结构；(2)基于自然语言处理技术[1-2]。文献[3]提出 3 种著名的基于文档结构文本信息隐藏算法：(1)行间距编码；(2)字间距编码；(3)特征编码。其中，行间距编码具有较强的鲁棒性，但其不可见性较差，而且隐藏容量非常小。与其相比基于字间距编码和特征编码算法的隐藏容量较大，但特征编码算法不可见性仍旧不理想，而且在解码时要求原始的文档，给解码造成不便。文献[4]提出一种基于 TMR(Text Meaning Representation)树的自然语言文本水印方法。该方法具有良好的鲁棒性和一定的抗攻击性，但是受限于自然语言处理技术，嵌入水印后的载体文本容易发生语义改变和难以理解的情况，不可见性不够理想。所有自然语言文本隐藏的方法相对在鲁棒性上提高了系统的灵活性和承受攻击的能力，但其不适用于文本内容不宜更改的情况。同时，由于要借助、依赖于自然语言处理技术的发展，该技术还存在很多尚待解决的问题。文献[5]提出基于词归类文本信息隐藏算法，将英文文本进行词归类(Word Classification)处理，这是文档结构算法中的一个新算法，但该算法的鲁棒性没有明显提高，且仅限于英文。

目前大多数文本水印嵌入和检测方法都是面向英文文本，对中文文本效果并不理想，需要更多的基于中文文本的数字水印算法。本文从基于词平台汉字编码入手，提出一种新的适用中文的文本信息隐藏算法。该算法不改变任何汉字编码，即不改变文档输出，以标志位为桥梁将密文信息隐藏到载体文档中，因此，完全可以通过盲检，不可见效果很好。并且算法通过词扩展规则增大了信息隐藏的容量，而且算法较大程度地提高了信息隐藏的鲁棒性，具有较强的安全性。

3 基于词平台汉字编码的文档信息隐藏算法

根据中文文本文档的特点，本文提出一种基于词平台汉字编码的文本信息隐藏算法。利用词平台汉字编码方法得到相应的二进制编码，通过分词和词扩展将相应词典词的二进制编码进行处理。

实验结果表明，这一算法具有较好的隐藏效果，有一定

基金项目：国家火炬计划基金资助项目“基于 SAAS 的信息化共性技术服务平台”(国科发计[2008]658 号)
作者简介：张洪礼(1962－)，男，副教授，主研方向：网络与信息安全，密码学及其应用；刘 丹、温学谦、何新宇，硕士研究生
收稿日期：2009-11-30 **E-mail：**lfdandanhappy@163.com

的抗攻击能力，从而使其鲁棒性有所提高，并且词扩展规则提高了密文信息的隐藏量。

3.1 词典词和词典码

假设一段汉语文字，将文字中词汇按它们最常用的词性如名词、动词、形容词等进行分类。称按以上方法得到的词为词典词(W_i)。对每个词典词进行编码，称为词典码(DC_i)，所有词典码都是由 4 个字节二进制编码构成，具体形式为

[1010xxxx xxxxxxxx xxxxxxxx xxxxxxxxB]

其中，B 是二进制的表示符号，下同，以下词典码均用二进制表示。第 1 个字节的高 4 位必须是 1010。第 1 字节的低 4 位 xxxx 的范围是 0001~1111，用来表示这个词的词性，1110 和 1111 保留，以待将来扩充功能。具体对应关系见表 l。

表 1 词性表

0001	0010	0011	0100	0101	0110
名词	形容词	副词	代词	数词	量词
0111~1000	1001		1010-1101		
介词、连词	助词、语气词、象声词、叹词		动词和动词短语		

第 2 字节的高 4 位为保留位，第 2 字节低 4 位用来表示该词所包含的字数(1 个~15 个)；将剩余的第 3 字节和第 4 字节组成一个顺序码，范围是 l~65 535，用来将词汇按拼音顺序进行排列。按此方式编码至少能容下的词条数是：14×65 535=917 490 条。

经过处理的每个词的新编码与机内码词典码表(即一个数据库)对应。经过编码后的整个文档由一系列码字组成，其中，西文字符采用国际标准的表示西文字符和符号的 ASCII 码表示(本算法主要介绍基于中文编码的信息隐藏)。

3.2 分词和词扩展

为了将词平台汉字编码技术与文本信息隐藏技术更好地结合，定义了一些相关概念及适用规则。其中，词扩展规则提高了信息隐藏量，顺位编码和逆位编码规则增大了词编码信息与密文信息的匹配量，编码不匹配处理规则主要是词编码信息与密文信息不匹配时相应的解决措施。

3.2.1 分词

中文文档经过上述汉字编码得到词典码(DC_i)，由于词典码是二进制编码，可以计算出每个词典码中含有“1”的奇、偶数，通过含有“1”的奇、偶数对词典码进行分类，因此经过分词处理的每一个词典词对应一位二进制编码。就此可以制定一个规则(规则 1)，将词典词分类，即实现分词(C_i)。如表 2 所示。

表 2 规则 1

Conditions	C_i
DC_i 中含有奇数个 1	0
DC_i 中含有偶数个 1	1

3.2.2 词扩展

由于对词典码进行分词使得每一个词典词只对应一位二进制编码，这样可以实现密文的隐藏，但隐藏密文的信息量不是很大。考虑到隐藏密文的信息量，本算法对分词进行词扩展，得到词扩展编码(E_i)，由原来的一位二进制编码扩展为 2 位二进制编码，增大信息的隐藏量。

词扩展规则(规则 2)如下：

If ($m \geqslant h$ and $p \geqslant q$) E(i)=00;

If ($m \geqslant h$ and $p<q$) E(i)=01;

If ($m<h$ and $p \geqslant q$) E(i)=10;

If ($m<h$ and $p<q$) E(i)=11;

其中，$m=C_{i-2}+C_{i-1}$; $h=C_{i+1}+C_{i+2}$; $p=C_{i-1}+C_{i+1}$; $q=C_{i-2}+C_{i+2}$。

为了确定一段中的第 1 个词典词和最后一个词典词的词扩展编码，可以将一段中的词典词序列看作是循环的，假定某段中最后一个词典词是 W_j，那么 E_j 的值是由 C_{j-2}, C_{j-1} 和 C_1, C_2 决定的。

3.2.3 顺位编码和逆位编码

算法是通过载体文档编码和密文编码的匹配来实现信息隐藏的，为了增加信息的匹配量，提高算法的鲁棒性，提出了顺位编码和逆位编码的规则(规则 3)，如表 3 所示。

表 3 规则 3

E_i	R_i
01	10
10	01
00	11
11	00

其中，E_i 为原词扩展编码，亦即此处提到的顺位编码。将顺位编码按位取反即得到逆位编码 R_i。

顺位编码在标记位中标记为 1，逆位编码标记为 0，标志位作为密钥的一部分，以此来告知信息接受者隐藏信息是顺位编码还是逆位编码。

3.2.4 编码不匹配处理规则

经过规则 3 处理后的载体文档能够与大部分的密文相匹配。以下 2 种情况则不能完成编码匹配：(1)载体为 00/11，密文为 01/10；(2)载体为 01/10，密文为 00/11。此时按照规则 4，将词典词编码与密文编码进行异或运算，得出标志位信息。具体规则(规则 4)如表 4 所示。

表 4 规则 4

词典词编码	运算符	密文编码	运算结果	标志位
00	⊙	01	01	a(01)
11		10		
01		00	10	b(10)
10		11		

规则 4 是针对密文信息的嵌入，而从载体文档提取密文信息时，需要遵循规则 5 进行提取。规则 5 如表 5 所示。

表 5 规则 5

词典词编码	运算符	标志位	运算结果(密文编码)
00	⊙	a(01)	01
			10
11		b(10)	10
			01
01		a(01)	00
			11
10		b(10)	11
			00

4 秘密信息嵌入和提取算法描述

基于以上提出的各个规则，本文对该算法进行具体描述。算法分为 2 个部分：秘密信息嵌入算法和秘密信息提取算法。

4.1 秘密信息的嵌入算法

秘密信息的嵌入算法流程如下：

(1)将密文信息进行编码、加密，得到密文二进制编码流 S_j；

(2)运用词平台汉字编码技术将文本载体 T 进行编码处理，得到词典码 DC_i；

(3)根据规则 1，将词典词分类，即实现分词 C_i；

(4)应用词扩展规则将分词码 C_i 词扩展，得到词扩展编码 E_i，即每个词典码 DC_i 对应 2 位二进制码。

(5)将密文二进制编码流 S_j 与词扩展编码 E_i 每 2 位进行对比，依据规则 3 判断两者是否匹配，如果匹配，则记下相应的标志位信息，否则转到(6)；

(6)密文二进制编码流 S_j 与词扩展编码 E_i 不匹配，应用编码不匹配处理规则 4 进行处理，记下相应的标志位信息 a(01)或 b(10)；

(7)判断密文信息 S_j 是否结束，如果结束则算法结束，否则转到(5)继续执行。

4.2 秘密信息的提取算法

秘密信息的提取算法流程如下：

(1)运用词平台汉字编码技术将文本载体 T 进行编码处理，得到词典码 DC_i；

(2)根据规则 1，将词典词分类，即实现分词 C_i；

(3)应用词扩展规则将分词码 C_i 词扩展，得到词扩展编码 E_i，即每个词典码 DC_i 对应 2 位二进制码。

(4)读取一位密文标志位信息，判断是否是 0/1 代码，如果是 0/1 代码则转到(5)执行，否则转到(6)；

(5)根据读取的标志位信息，运用规则 3，结合词扩展编码 E_i，得到 2 位密文编码；

(6)根据读取的标志位信息，运用编码不匹配处理规则 5，结合词扩展编码 E_i，得到 2 位密文编码；

(7)判断标志位信息 Sn_m 是否结束，如果结束则算法结束，否则转到(4)继续执行。

5 实验与分析

根据文本信息隐藏算法的特性，对本文提出的算法进行了理论分析和实验验证，并与其他经典的文本信息隐藏算法进行比较，并得出结论。

5.1 特性分析

(1)隐藏量

本文提出的算法由于采用了词扩展规则，由原来的每个词对应一位二进制码，扩展到 2 位，有效地增大了信息的隐藏量。理论上隐藏量是可以根据载体大小而任意设定，但考虑到隐蔽性，防止被恶意攻击者怀疑，可将容量限制在一个范围，一般限制在载体文件大小的 5%内是很难被注意的。密文信息量较大时，可以对应选择较大的载体文件。

(2)鲁棒性

基于文本特征的信息隐藏算法基本是通过文本格式的微调或修改文本的一些特征来嵌入密文信息。但文本载体使用这些嵌入算法嵌入密文信息后不能通过一些检测，如盲检，算法鲁棒性不佳。本文提出的算法由于没有改变文本载体的输出，因此完全可以通过盲检及其他检测系统，鲁棒性较以往的算法有明显的提高。

(3)隐蔽性

实验证明，使用本文提出的算法没有改变隐蔽载体文档的输出，无论是在内容上，还是在格式上，都具有很好的隐蔽性，嵌入隐藏信息前后的文档对照如图 1 所示。

5.2 实验

以本文概述中的一段文字为文本载体，以“张洪礼”为密文信息，应用本文提出的算法将其嵌入到文本载体中，实验效果如图 1 所示，其中，图 1(a)为原始文本载体，图 1(b)为嵌入密文信息后的文本载体。

随着计算机技术和网络的飞速发展，以往基于文本格式、语法、语义的文本信息隐藏算法，在秘密信息隐藏的安全方面将面临很大的挑战。文中提出了基于词平台汉字编码的文本信息隐藏算法，该算法基于词平台汉字编码技术，运用编码变换规则和标志位实现了秘密信息的文本嵌入，在算法的鲁棒性上有很大提高，增加了秘密信息的安全性，理论分析和实验结果验证的了本算法的有效性。

(a)原始文本载体

随着计算机技术和网络的飞速发展，以往基于文本格式、语法、语义的文本信息隐藏算法，在秘密信息隐藏的安全方面将面临很大的挑战。文中提出了基于词平台汉字编码的文本信息隐藏算法，该算法基于词平台汉字编码技术，运用编码变换规则和标志位实现了秘密信息的文本嵌入，在算法的鲁棒性上有很大提高，增加了秘密信息的安全性，理论分析和实验结果验证的了本算法的有效性。

(b)嵌入水印后

图 1 隐藏信息嵌入前后文档对照

为更好地测试该算法的特性，随机选择 500 篇中文文本文档，使用本文算法进行信息隐藏，同时对其他经典信息隐藏算法仿真，将其特性进行对比，从隐藏算法的隐藏量、隐蔽性、鲁棒性 3 大特性进行实验验证，统计数据如表 6 所示。

表 6 实验结果比较

指标	本文算法	基于格式的算法	基于语法、语义的算法
隐藏容量	文档大小的 3.003%	小于文档大小的 0.655%	小于文档大小的 2.5%
隐蔽性	很好，完全不可见	较好，肉眼难以察觉	较好，完全不可见，但对语义有破坏
鲁棒性	抗攻击能力强	能抵抗一些篡改攻击	能抵抗各种格式攻击

从统计结果可以看出，本文算法无论是在隐蔽性、隐藏容量还是鲁棒性上，较优于其他信息隐藏算法，具有一定的应用价值。

6 结束语

本文提出一种基于词平台汉字编码的文本信息隐藏算法，在中文文档中取得很好的隐藏效果。通过实验也可以看出，此方法的透明性是非常好的，而且文本里面的任何一个字符，包括西文字符，均可以进行信息隐藏，信息隐藏量比较大。并且运用词扩展规则，进一步增大了隐藏的数据量。该方法技术简单、容易实施，在实际的信息安全保护中的应用前景广泛。

参考文献

[1] Gupta G, Pieprzyk J, Wang Huaxiong. An Attack-localizing Watermarking Scheme for Natural Language Documents[C]// Proceedings of 2006 ACM Symposium on Information, Computer and Communications Security. Taipei, China: [s. n.], 2006.

[2] 袁树雄，孙星明. 英文文本多重数字水印算法设计与实现[J]. 计算机工程, 2006, 32(15): 146-148.

[3] Brassil J, Low S, Maxemchuk N F, et a1. Electronic Marking and Identification Techniques to Discourage Document Copying[J]. IEEE Journal on Selected Areasin Communications, 1995, 13(8): 1495-1504.

[4] Mikhail A J, Victor R, Christian F H. Natural Language Watermarking and Tamperproofing[C]//Proceedings of the 5th International Information Hiding Workshop. Berlin, Germany: [s. n.], 2002.

[5] Kim Young-Won, Moon Kyung-Ae. A Text Watermarking Algorithm Based on Word Classification and Inter-word Space Statistics[J]. Proceedings of the IEEE, 2005, 87(7): 1181-1196.

编辑 陈 文