中文文本信息隐藏研究进展

吴国华，龚礼春，袁理锋，姚晔
（杭州电子科技大学网络空间安全学院，浙江 杭州 310018）

摘 要：文本信息隐藏是保护文本内容安全性与完整性的重要技术。综述了中文文本信息隐藏的研究进展，根据中文文本信息隐藏的线索，将已有的算法分为 3 类：基于文本图像的算法、基于文本格式的算法和基于文本内容的算法，分别阐述了每类算法的实现过程，分析其优势与不足，并且对比分析了它们的原理、嵌入容量和抵抗攻击能力等。此外，总结了中文文本信息隐藏技术存在的问题，并且对其研究趋势进行展望，期望为该领域的研究提供参考。

关键词：信息隐藏；文本隐写；中文文本信息隐藏；文本特征
中图分类号：TP309
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019208

Review of information hiding on Chinese text

WU Guohua, GONG Lichun, YUAN Lifeng, YAO Ye
School of Cyber Space Security, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: Text information hiding is an important technology to protect the security and integrity of text content. The research progress of Chinese text information hiding was reviewed. According to the clue of Chinese text information hiding, the existing algorithms were divided into three categories, algorithms based on text images, algorithms based on text format and algorithms based on text content. The implementation process of each type of algorithm was elaborated and the advantages and disadvantages of the algorithm were analyzed. At the same time, their principles, embedded capacity and anti-attack capability was compared and analyzed. In addition, the problems existing in Chinese text information hiding technology was summarized, and the research trends were discussed in the future. It is expected to provide reference for research in this field.

Key words: information hiding, text steganography, Chinese text information hiding, text feature

1 引言

信息隐藏技术是信息安全领域的一个热门研究方向。国际上信息隐藏研究起步较早，而国内的信息隐藏研究是在 1999 年全国信息隐藏暨多媒体信息安全学术大会（CIHW, China Information Hiding Workshop）之后兴起的[1]。经过多年的发展，在国内外众多学者的努力下，以视频和图像为载体的信息隐藏研究取得了不少的成果[2]。目前，以图像、视频和音频为载体的信息隐藏研究成果在数量上大大超过以文本为载体的研究成果[3]。文本的数据量较小，存在的冗余信息也较少，较难将秘密信息嵌入其中[4]。在文本信息隐藏研究的初级阶段，大多数方法是将文本视为文本图像，通过图像信息隐藏方法嵌入待隐藏信息[5]。然而，将文本数据当成图像来处理，没有利用文本数据具有的属性[6]，不能取得较好的信息隐藏效果。

文本是信息交流与信息传递的重要载体。由于

收稿日期：2019−06−15；修回日期：2019−08−30
通信作者：姚晔，yaoye@hdu.edu.cn
基金项目：教育部人文社科基金资助项目（No.17YJC870021）
Foundation Item: Humanities and Social Sciences Foundation of Ministry of Education of China (No.17YJC870021)

互联网的开放性及信息传播的不确定性，文本被恶意伪造和非法篡改的事件时常发生[7-8]。通过文本信息隐藏技术保障文本内容安全是信息安全领域亟待解决的重要任务[9]。本文总结已有的中文文本信息隐藏的研究成果，根据算法的线索，将已有的算法归为 3 类，分别对其进行分析和总结，并给出展望，以便该领域的研究者了解其研究进展。

2 现有算法分类与分析

文本文档包含文本内容和文本格式两部分。文本内容是文档中需要传递的明文信息。文本格式是对文本内容进行组织的方式。文本内容经过字符编码，然后保存成 wrod、pdf、xml 等文档格式。

针对文本信息隐藏，研究人员通过借鉴视频图像的隐写算法，或基于文本的格式与内容，提出了多种适用于文本信息隐藏的算法[10]。本文在广泛调研现有算法的基础上，根据隐藏信息算法的线索，将文本信息隐藏算法划分为 3 类，如图 1 所示。

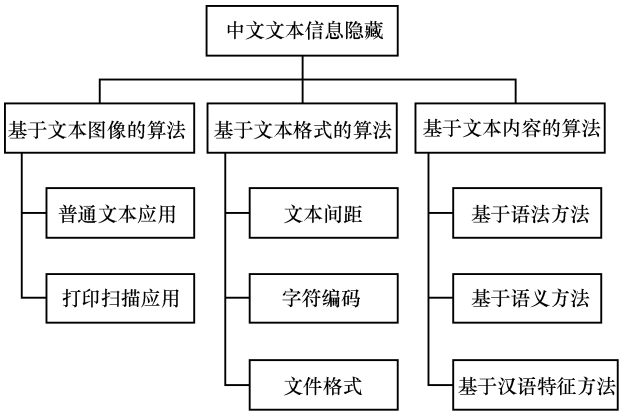

图 1 文本信息隐藏分类

2.1 基于文本图像的算法

通过扫描仪、数码相机、截屏等方式，纸质文件或电子文档被存储为图像格式，称为文本图像。图像文件在内容与格式上有较多的冗余空间，便于将秘密信息嵌入其中。文本图像与人物、景物图像相比，其图像纹理分布更均匀、区域边缘特征更明显[11]。因此，可以借助文本图像的特征隐藏信息。图像隐写技术为文本图像信息隐藏提供了可靠的技术支持。基于文本图像的信息隐藏算法，可以应用于普通文本的信息隐藏场景中，或者纸质文本的抗打印扫描场景中，具有较强的实用性。

2.1.1 普通文本应用

Ding 等[12]通过调整单词之间的细微间距，使文本图像中行与行之间的平均单词间距表现出正弦曲线的特征，以此把水印信息编码在正弦曲线内。该算法的稳健性较高，在非盲检测与盲检测中效果较佳。但是，该算法仅在英文文本中具有较好的效果，且嵌入容量较小。

Kim 等[13]提出适用于中文、英文、韩文这 3 种语言的文本图像信息隐藏算法。算法将文本图像转换为灰度图像，然后用 Sobel 算子提取文本图像边缘，统计图像中每行文本在 16 个方向上的直方图。实验表明，3 种语言各自构成的文本图像在 16 个方向上拥有不同的统计特征，并且每种语言构成的文本图像每行的统计特征基本相同，如图 2 所示。文中将此现象命名为“sub-image consistency”。根据行直方图特征一致性规律，嵌入时将前三行直方图特征作为参考，改变后面行直方柱的长短来嵌入 0 bit 和 1 bit。该算法稳健性不强，如果参考行文本图像特征被破坏，隐藏信息将全部无法提取。同时，算法嵌入容量低，每行仅能嵌入 1 bit。

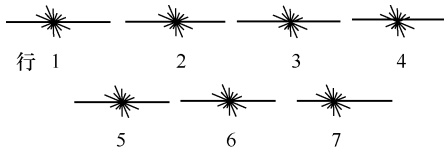

图 2 行子图特征一致性

赵星阳等[14]通过调整中文、英文二值文本图像中单个字符除水平、竖直笔画外的阶梯边沿，改变每行文本图像中上下两半部分黑色像素之和的比值，利用比值之间的关系嵌入水印信息。水印检测时，只需要根据算法识别出比值关系就能成功恢复水印信息。与文献[13]比较，该算法在水印提取时不依赖参考行，增强了水印的稳健性，且能较好地抵抗文本图像缩放攻击。然而，该算法嵌入容量受限于文本行数量，若文本图像受到噪声污染，水印信息将无法准确地提取。

Behrooz 等[15]提出了一种针对两端对齐文本的隐写算法。该算法使用光学字符识别（OCR, optical character recognition）技术获取文本图像中的主行（HL, host line）（包含至少 3 个 add space 与 9 个 normal space），然后通过事先设定的嵌入规则，在 HL 先嵌入密钥信息，再嵌入秘密信息中的字符频数，最后以将 normal space 替换为 add space 的方式嵌入经过 Huffman 编码的码字信息。该算法在每个 HL 嵌入 4 bit 信息，单行嵌入容量比文献[14]高，

但是 HL 具有不确定性，使算法只有对长文本才具有较好的嵌入效果。

2.1.2　打印扫描应用

与普通文本水印嵌入算法比较，抗打印扫描的文本信息隐藏算法不仅需要较高的嵌入容量，而且需要具备较高的稳健性，才能抵抗针对打印扫描的纸质文本的攻击。

亓文法等[16]针对打印扫描文本水印无法在保持理想的视觉效果下实现盲提取这一缺陷，提出一种隐蔽性较高的文本图像水印算法。该算法切割文本图像中字符，统计每个字符图像黑色点的个数与文本图像中所有字符图像黑色点总数，发现两者的比值在打印前后基本保持不变。基于此发现，该算法在嵌入时翻转字符图像的黑色点数，翻转部分黑色点来补偿嵌入部分的修改，从而保证整篇文档字符的平均黑色点数不变，达到嵌入水印的目的。该算法具有较高的不可见性，但嵌入容量低，且使用的特征在打印扫描后不稳定，使算法针对二次复印件的检测不够理想。

Tan 等[17]从单个字符图像的笔画着手，提出一种大容量的抗打印扫描的文本水印算法。算法提取出字符图像笔画及笔画之间的交叉点，并筛选出适合旋转的笔画（如撇、捺），通过笔画的相对旋转嵌入水印信息。为了使笔画绕交叉点旋转后尽可能不被肉眼发现，笔画可旋转的角度被限定在某个区间内。水印提取时通过检测单个字符笔画旋转方向判断嵌入的比特信息。该算法以单个字符笔画为基本嵌入单位，相较于文献[16]水印提取时不受其他部分或整体因素的影响，算法嵌入容量较大，但是实现起来相对困难。

雷敏等[18]经过大量的实验研究发现，文本中相邻字符高度并不相同，并且字符之间高度的相对关系经过打印扫描处理后基本保持不变。因此，在秘密信息嵌入的过程中，通过定义的复杂度函数调整字符的高度，使相邻字符图像的高度关系在打印扫描前后不发生变化，从而根据 2 个相邻字符图像的高度相对关系来嵌入水印信息。该算法在 2 个字符图像中嵌入 1 bit 信息，嵌入容量较文献[17]低，但算法利用的字符特征较为稳定，能够较好地抵抗打印扫描攻击，适用于对稳健性要求较高的应用场合。

将文本转换为文本图像嵌入信息，是文本信息隐藏算法的常见处理方式。无论算法使用文本图像行表现出的整体特征，或是字符图像的局部特征，这些特征均是人眼无法直接发现的，必须经过大量的实验与统计分析得出。与此同时，当载体受到如文本污损、图像噪声污染、OCR 攻击等外部因素干扰时，算法的稳健性和可行性都会受到一定程度的影响。

2.2　基于文本格式的算法

文本格式是组织文本内容的方式。文本内容可以被不同的文本格式组织、封装、存储，并通过终端屏幕呈现在人们眼前。基于文本格式的信息隐藏算法利用文本的排版方式、字符编码特征、文本封装格式等属性隐藏信息。

2.2.1　文本间距

为了方便人们阅读文本，不论是何种语言形成的文本，都是由字组成行、经行排成段、段落构成文章。这种按照顺序组织文本的方式能作为文本中隐藏信息的自然属性。因此，通过修改字、行和段在文本中的排版间距是一种常用的信息嵌入方法。

Low 等[19]提出了字移编码和行移编码算法。字移编码是指以某字符的相邻字符为参照，在人眼不可感知的前提下，将该字符向左或向右移动一定的距离，达到嵌入信息的目的。行移编码选取待嵌入信息行的相邻两行作为参考行。在不移动参考行前提下，用待嵌入信息行上移或下移来表示嵌入 0 或 1。行移编码算法嵌入容量小，字移编码算法虽能提高嵌入容量，但其稳健性较低。2 种算法提取水印时均是采用非盲方式，需要原始文本作为隐藏信息提取的参照。

文献[20]改进了上述算法，提出将行移编码与字移编码方法相互结合的信息嵌入算法。算法在文本行水平方向使用字移编码，垂直方向使用行移编码。该算法结合 2 种方法的优势，既保证水印的稳健性，又能提高嵌入容量，并且实现了水印的盲提取，简化了秘密通信的成本，具有较好的适用性。

调整文本间距的算法在对抗一定强度的伸缩攻击上有较好的效果。但是需要严格遵守人眼视觉特点对文本字符间距、行间距和段间距进行调整。如果间距调整程度过大，会导致可视性严重降低，秘密信息的嵌入位置也易被人眼识别出来。

2.2.2　字符编码

字符通过编码形成二进制数据并存储于计算机中。已有算法通过字符编码的奇偶性嵌入秘密信息，或者将字符编码表中的不可见字符插入文本中

隐藏信息，甚至通过改变字符颜色的编码值来嵌入信息。

Unicode 编码集集成了大多数语言的字符编码，是基于字符编码隐藏算法的首选编码方式。文献[21]提出用 Unicode 编码的奇偶性隐藏秘密信息。该算法将文本字符用十进制的 Unicode 编码表示，把隐藏信息视为二进制比特串，然后将每比特隐藏信息与单个字符 Unicode 编码的奇偶性对比：若两者均为奇数则嵌入 1；两者均为偶数则嵌入 0；两者奇偶性不一致，则修改字符蓝色分量或字符下划线颜色中一种属性的最低位值，使算法能够对嵌入位置进行标记。该算法的嵌入位置默认从文本第一个字符开始，若提取隐藏信息从其他字符位置开始，将无法提取完整的秘密信息。

陆绿等[22]为了解决上述非顺序提取水印信息导致秘密信息不能复原的问题，提出一种扩展水印信息，并在组间插入分隔符作为区分标志的嵌入算法。在水印提取时，即使某部分载体被破坏，在分隔符的作用下，依然可以从未被破坏的载体中获得完整的水印信息，具有较好的稳健性。

文献[21-22]中提出的算法均是利用单个字符编码隐藏信息，没有利用词语之间的相关性嵌入信息。文献[23]将字符编码扩展到词编码，提出基于词平台汉字编码的文本信息隐藏算法。文本中的词根据其词性被划分为不同的词典词，每个词典词使用 4 个字节编码成词典码，然后通过分词与词扩展方式嵌入信息。该算法不对字符的属性进行修改，因此具有较好的隐蔽性。但是，用 4 个字节编码一个词，会造成较多的冗余。

字符属性还包括字符颜色。基于人眼锥状细胞对蓝色不敏感这一特点，刘豪等[24]通过修改字符的蓝色分量编码值嵌入水印信息，该算法能在一个字符中嵌入 1 bit 信息。Tang 等[25]改进上述算法，通过修改每个字符颜色属性 RGB 的 3 个通道最低位值的嵌入信息，使改进后的算法嵌入容量是文献[24]的 3 倍。修改字符颜色属性编码的算法不适用于对隐蔽性要求较高的应用场合，过度修改字符颜色编码值会降低隐蔽通信的安全性。

通过研究文本字符编码表，领域内的学者发现虽然字符编码表中有不少字符被赋予编码，但是将它们插入文本中却不能被人眼感知。此类型的字符被称为“不可见字符”，ASCII 编码表中部分不可见字符举例如表 1 所示。

表 1 ASCII 码中部分不可见字符

编码	含义
0000	NUL
0001	SOH
0010	STX
0011	ETX

利用上述不可见字符，Liu 等[26]提出一种基于 Hash 函数与不可见 ASCII 字符替换的信息隐藏算法。该算法使用“SOH”这一不可见字符，替换文本分段中的空格。对替换后的分段文本进行 Hash 运算，将 Hash 值与隐藏信息比较，根据设定规则嵌入信息。文献[27]根据约束函数确定嵌入位置。若秘密信息为 0，载体对应位置加入空格，否则，添加“SOH”字符。该算法的嵌入能力取决于约束函数在文本中获取嵌入位置的数量。

相比于文献[26-27]，文献[28]扩展了不可见字符编码的方式，提出基于 Unicode 编码的不可见字符水印嵌入算法。该算法将 Unicode 不可见字符编码两两组合，形成一组映射规则插入文本每个句子的句号前。为了提高安全性，算法中使用了 16 位循环冗余校验。在水印未嵌入载体前对每个句子做散列计算，并将计算的结果根据映射规则转为不可见水印编码，附加到嵌入的水印编码句子的尾部，使算法的稳健性得到进一步提升。

利用人眼视觉的不敏感性，文献[29]提出一种同形字符替换的文本水印算法。为区分 Unicode 编码不同而人眼看似相同的字符，作者整理出一份同形字符表。算法对同形字进行编码，从而通过文本中同形字的编码替换来嵌入信息。提取水印时检测相应字符的编码，与编码映射表作对比，进而获得水印信息。该算法的隐蔽性与嵌入容量都较好，但是水印的稳健性差，嵌入位置一旦被其他字符替换，水印不能被完整提取。

目前，出于对安全性与隐蔽性的考虑，在实际应用中使用不可见字符和视觉不可区分算法嵌入信息的应用较多。然而，这些应用中大多数被用于结构化文本中。非结构化的文本中基本不存在冗余空间，无法实现信息嵌入。基于字符编码属性的信息隐藏算法，绝大多数以字符为基本嵌入单位，因而该类算法的嵌入容量普遍要高于基于文本间距的算法。

2.2.3 文件格式

文本文件的格式多样（常见文档格式包括

word、pdf、xml 等），利用文本格式隐藏信息的算法对载体文本的文件格式具有很强的针对性。已有算法使用文件格式内部未使用空间嵌入信息，或者利用文件格式自身的特殊性隐藏信息。

文献[30]分析 word 文档的数据结构，提出一种使用 word 文档空间中控制结构数据、嵌入式对象等属性隐藏信息的算法。该算法利用 word 文档格式中未使用的空间嵌入数据，从而在抵抗文本复制攻击方面具有较高的稳健性。然而不同版本的 word 文档内部数据结构不相同，因此提取隐藏信息时使用的 word 文档格式必须和嵌入信息时一致。

文献[31]提出一种使用 pdf 文档结构嵌入水印的算法。该算法根据 pdf 文档行末标识符不显示的特殊性，获取交叉引用表中每行的行末标识符，通过水印信息控制行末标识符的修改方式，从而间接嵌入水印信息。该算法能抵抗文本复制攻击，且文档中的标识符被替换后不改变文档的大小，因而能较好抵抗统计攻击。

文献[32]提出一种在 xml 文档中嵌入水印的算法。该算法根据 xml 文档结构中属于同一层次标签的先后排列顺序不影响文档内容展示的特点，将不同层次与同一层次的标签进行组合与排列，构建秘密信息与标签组合排列之间的对应关系，通过映射函数实现秘密信息的嵌入与提取，使算法具有较好的隐蔽性和抵抗复制攻击的能力。

由此可见，基于文本格式的算法使用 word、pdf、xml 文档格式的内部结构或特殊属性来隐藏信息，能较好地抵抗文档的复制攻击，并且具有较高的隐蔽性。然而，该类算法仅针对某一指定文件格式或者特定版本的文档来设计，算法不具有通用性。此外，该类算法嵌入信息之后，可能引起文件大小的改变，易引起攻击者怀疑。在对文件大小敏感的基于文本格式的信息隐藏应用中，可以研究文件格式自身的特点来使得隐藏信息后的文件尺寸保持不变。

2.3　基于文本内容的算法

基于文本内容的信息隐藏算法重点分析文本内容，挖掘文本内容特征，构造合适的算法将秘密信息嵌入其中。该类方法近些年来备受学者关注，这是因为自然语言处理技术的成熟是文本内容研究的重要基础，基于文本内容的信息隐藏方法在不改变文本语义（或不修改载体内容）的前提下，通过等价信息替换（或从文本中提取特征）能够较好地隐藏秘密信息。根据所选文本内容的差异，基于文本内容的文本信息隐藏算法分为基于语法的方法、基于语义的方法、基于汉语特征的方法。其中，基于语法的方法和基于语义的方法以自然语言处理技术为支撑，而基于汉语特征的方法利用汉语言特点嵌入信息。为了提高基于文本内容算法的隐蔽性，在嵌入过程中可以使用零水印与无载体的嵌入方法。

2.3.1　基于语法方法

基于语法的信息隐藏技术以自然语言语法结构为依据，利用句中词语的依赖关系，或者句式变换（如主动变被动、移动附加语）等语法规则，构造算法嵌入秘密信息。

文献[33]以汉语助词“的”为典型代表，提出一种基于虚词变换的隐写算法。算法从文本中找出含有“的”字的句子，在不影响文本原意的前提下，以增加或删除句中“的”字方式嵌入 1 bit 信息。该算法通过定义模板作为增删“的”字的依据，具有较好的灵活性。但是，其嵌入容量不高，修改后的文本易被察觉。

文献[34]提出了一种基于句子长度的文本信息隐藏算法。该算法以不改变句子原意为前提，对句子进行句式变换，通过改变句子的长度嵌入水印信息。为了能够抵抗增加、删除、句子变换对文本的攻击，算法对嵌入位置进行选取。在提取水印时引入了投票机制，使算法的稳健性被提高。

文献[35]给文本中的每个句子分配序号，用整篇文本中词语出现的频率，定量的计算每个句子的熵。将句子熵大于设定阈值的句子当成是文本的重要句子。把筛选出句子的序号作为零水印信息，发送到第三方认证机构，实现文本内容的版权保护。

2.3.2　基于语义方法

基于语法的文本信息隐藏方法一般是在句子级别嵌入秘密信息，因此，隐藏信息的容量相对较低。基于语义信息隐藏方法细化了文本内容研究粒度，从字词层面嵌入以提高文本的隐藏容量[36]。语言学中广泛存在的同义词被基于同义词替换的隐写方法作为嵌入依据。该类算法需要构造同义词库，对同义词集内的同义词进行编码，在不改变句子原意下，通过同义词的相互替换隐藏信息。

Chiang 等[37]将基于同义词替换的方法用于中文文本。水印在嵌入的过程中用二次剩余理论先选取适合替换的句子，再选择句中需要被替换的同义

词。该算法运行复杂，且一个同义词只能嵌入 1 bit 信息。经过该算法替换生成的句子不可避免地产生同义词替换不当导致句子出现歧义的问题。

文献[38]提出一种改进的中文同义词替换信息隐藏算法。该算法依据《同义词词林》等对同义词分类。若同义词属于完全可替换类，则直接替换；若属于不完全可替换的类，则需要根据词性来判断是否替换；若属于歧义词类，则根据上下文搭配词计算同义词被替换的概率。经过划分后，算法筛选出适合替换的同义词，降低替换后的语义失真程度。

为了进一步量化同义词替换评价标准，姜传贤等[39]定义了同义词替换评价模型，提出基于文本重要内容的稳健水印算法。算法对文本分词处理后，统计主题词（文本中权重较高的一些词），提取包含主题词的句子以及句子中的同义词集。在完全知道句中同义词上下文搭配词的条件下，通过依存句法分析，选取同义词集中与上下文搭配最合适的同义词替换原来的词。

Chang 等[40]针对同义词替换后的隐蔽性问题及同义词集交叉现象，对同义词替换算法提出两点改进。其中，通过机器学习算法，利用大型语料库训练好的 *n*-gram 模型剔除模棱两可的同义词，之后对同义词集进行评估，从而提高同义词替换后的隐蔽性。再者，利用图论的思想，将同义词作为图的顶点，词与词之间用边连接。使用顶点着色算法，使出现在不同同义词集内的某个词的编码一致，从而解决了同义词编码不一致在解码时导致的歧义问题。

文献[41-42]从信息论角度出发，利用信息编码方法提高同义词嵌入效率。文献[41]设计了一种基于矩阵编码的同义词替换方法。文献[42]在将秘密信息嵌入前，先用串表压缩（LZW, lempel ziv welch）算法对秘密信息压缩。秘密信息经过压缩后其长度减小。相对不压缩而言，同样载体文本能够隐藏更多的信息。

目前，同义词替换隐写算法是词语级别嵌入信息较成熟的方法。然而，中文的语义环境较为复杂，算法需要耗费大量的时间通过语义分析来消除同义词替换的歧义。

文献[43]使用词阶（文本集/文本中词语按词频排序的序号）图实现文本无载体信息隐藏。算法需要构建一个包含大量文本的数据集，信息隐藏者与接收者选择一个文本子集计算整个文本子集、子集中单个文本、子集中每个词语的词阶图。待传递的秘密信息以词为基本单位，在转换协议和密钥共同作用下，秘密信息被转为文本子集中的常见词，通过设计好的标签定位协议为转换后的每个词分配一个词阶。依据词语的词阶图与文本的映射关系，用文本表示秘密信息中的词，将文本载体按照秘密信息中词语出现的先后顺序发送给接收方。

文本无载体以秘密信息为驱动，且不对载体修改，因而在信息隐藏容量与隐蔽性方面，其比一般的嵌入式文本信息隐藏方法占据更大的优势。但当隐藏的信息量较少时，使用无载体的方法可能会带来较大的通信开销。少量秘密信息需要多篇文本来传递，使通信效率大幅度降低。

2.3.3 基于汉语特征方法

汉语文字具有中国特色，具有几千年的演变历史。面向汉语的文本信息隐藏算法利用汉字偏旁、汉语拼音、汉字笔画、汉字字体等特点在中文文本内嵌入秘密信息。

Sun 等[44]提出基于汉字偏旁的水印嵌入算法。该算法依据《信息处理用 GB13000.1 字符集汉字部件规范》定义了构成汉字的 6 种空间结构，如图 3 所示。其中，A 和 B 表示汉字的基本偏旁，lr 表示左右结构，ud 表示上下结构，we 表示包围结构，lu 表示左上结构，ld 表示左下结构，ru 表示右上结构，将汉字基本偏旁用唯一数字标识。每个汉字均可以由一个或多个基本偏旁按照 6 种空间结构组合而成。算法筛选出具有左右结构的汉字，作为信息嵌入的载体。将水印信息转为二进制流，在满足能够完全嵌入水印信息的前提下，依次从文本中读取一个汉字：若获取的汉字非左右结构，那么将其完整输出；若获取的汉字具有左右结构且待嵌入为 0 bit，同样也完整输出该汉字；如果当前汉字是左右结构且待嵌入为 1 bit，那么将该汉字拆成最基本偏旁输出；重复上述过程直至水印嵌完为止。该算法仅用了 6 种空间结构中的一种，文本整体嵌入能力受限于文本中左右结构汉字数量。

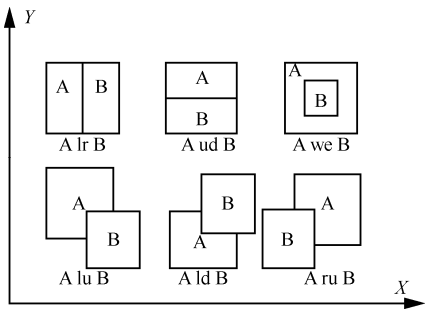

图 3 汉字 6 种空间结构

Wang 等[45]基于上述算法，在原来仅使用左右结构汉字嵌入信息的基础上又增加了上下结构的汉字，相比于文献[44]同样的一篇文本，水印的嵌入容量得到提高。不仅如此，该算法还引入了可逆的嵌入思想，即在水印信息被提取后，载体文本依旧可以保持原样，当水印信息被提取出后，载体能够被重复利用。

Fei 等[46]根据汉字一字多音的现象，提出一种基于多音字的中文文本水印算法。算法统计相邻 2 个多音字间汉字的个数与汉语拼音字母数记为十进制，然后将每位十进制转为 8 位二进制，同时还统计前一个多音字读音的个数与其拼音字母个数，同样转成 8 位二进制数。将两组 16 bit 数相与生成一组新的 16 bit 数，将该数作为水印信息保存。虽然该算法在 2 个多音字之间可以嵌入 16 bit 数据，嵌入容量大，但是其在抵抗文本多音字插入、删除、替换攻击上表现较弱。

文献[47]提出一种使用汉语拼音声韵母特征构造零水印的算法。该算法根据声韵母编号，统计文本中所有汉字的音数值（声母、韵母编号值之和）作为零水印，并按照设定的阈值选取出部分音数值作为用户提取水印时的密钥，将零水印与密钥一起保存。由于算法提取的特征不依赖于某一汉字或者拼音，因此算法的稳健性较高。

文献[48]提出一种基于汉字笔画的文本信息隐藏算法。对中文汉字的笔画数统计发现所有汉字中 8 笔画汉字最多。算法以 8 笔画汉字为嵌入点，统计其左右相邻汉字的笔画数。将 2 位十进制的笔画数转化为 16 位二进制，从而将一个用 Unicode 编码的汉字嵌入其中。该算法的嵌入容量取决于文本中 8 笔画汉字的数量，且很难抵抗对文本中 8 笔画汉字及其左右汉字的攻击。

文献[49]提出一种基于笔画的中文文本零水印算法。该算法统计文本中每个汉字出现的频率，筛选出频度较高的汉字。将这些汉字的笔画数作为文本的特征与水印信息进行位运算。运算结果作为用户的注册码，发送到第三方认证中心保存。

孙新梅等[50]针对汉字中的繁体字与简体字混用现象，提出一种基于字体的中文信息隐藏算法。算法执行前需要构建一个简/繁体字对应的字典，执行中可选用 3 种替换方法中的一种实现信息隐藏。1) 简单替换算法：若文本中的字在字典中，则根据待隐藏信息的二进制数来替换，待隐藏信息为 0 则保持简体字不变，为 1 则用相应的繁体字替换简体字；若文本中的字不在字典中，则保持不变。2) 高效替换算法：将待隐藏信息分为长度相同的组，将每组二进制数转为对应的十进制数 D，替换 D 个字后的简体字。该方法一个字符可以嵌入多个比特，且隐蔽性较好。3) 基于模板替换算法：其能够在多个字符中嵌入多个比特信息。文献[43]算法可选用 3 种替换方法中的一种，能够较好地平衡载体嵌入能力与隐蔽性之间的矛盾。

基于汉语特征的信息隐藏算法，无论是利用汉字的偏旁部首、汉语拼音和字体笔画嵌入水印信息，还是使用这些特征构造零水印实现信息内容安全，都能充分获得较为稳定的汉语特征。因此，该类算法能够较好地适用于中文文本信息隐藏，具有一定的研究价值和应用前景。

3　现有算法对比

本文在详细介绍文本信息隐藏算法的基础之上，选取了其中部分论文，分类整理成表 2 所示的结果。从表 2 可以看出，文本信息隐藏算法隐写过程中依赖的文本属性及算法的嵌入能力具有显著差异，具体分析如下。

1) 基于文本图像的算法

该类算法多以文本图像行处理后表现的整体特性出发嵌入信息[12-15]，或者利用单个字符图像的统计特征隐藏信息[16-18]。从算法的应用场合上看，抗打印扫描攻击的嵌入算法的稳健性通常要高于普通文本嵌入算法。该类算法大多对图像噪声敏感，一旦受到噪声攻击，信息提取的精度将受到严重的挑战。

2) 基于文本格式的算法

该类算法目前研究较为成熟，相应的成果较多，在信息隐藏的过程中可依赖的载体属性较多，不同算法的嵌入能力也有较大的差异，且多数算法只针对某一属性提出。例如，基于字符编码的算法以每字符嵌入 1 bit 居多[21-22,24,26]。如果将字编码扩展到词编码[23]，单颜色编码扩展到多颜色编码[25]，或者使用规则映射方法[29]可以提高嵌入容量。但是，该类算法只对某种字符编码方式生效。基于文本格式的算法[30-32]仅适用于某一特定格式的文本，文本格式的改变会导致算法失效。此外，修改文本间距的算法与修改字符编码的算法均存在隐蔽性与嵌入容量之间的矛盾。然而，基于文本格式的算法在文件格式内寻找冗余空间，能够较好地避开这一问题。

表 2　　**文本信息隐藏算法对比**

分类	子类	文献	原理	嵌入能力	局限
基于文本图像的算法	普通文本应用	Ding 等[12]	微调单词间距，在文本行间表现的正弦曲线中编码水印	取决于对波的采样方式与文本行数	受噪声干扰很难精确提取，嵌入容量低，稳健性差
		Kim 等[13]	文本图像行直方图的子图一致性特征	一行嵌入 1 bit	高强度噪声破坏隐藏信息，嵌入容量低，稳健性差
		赵星阳等[14]	调整二值文本图像字符阶梯边缘	一行嵌入 1 bit	受噪声干扰水印稳健性差
		Behrooz 等[15]	两端对齐文本的 HL 条件	一个 HL 嵌入 4 bit	HL 特征不稳定，适用长文本
	打印扫描应用	亓文法等[16]	打印扫描比值不变量，翻转字符图像边界上的像素点	一个嵌入字符隐藏 1 bit	易受噪声干扰，字符切割严格有序
		Tan 等[17]	调整字符图像的可旋转笔画	旋转一个笔画嵌入 1 bit	易受图像旋转，噪声攻击
		雷敏等[18]	相邻字符图像的相对高度	满足相对高度关系的 2 个字符嵌入 1 bit	相对高度的相邻汉字较为随机，无法抵抗 OCR 攻击
基于文本格式的算法	文本间距	Low 等[19]	基于字移编码、行移编码	调整的一个单词或一行嵌入 1 bit	仅用于非盲水印
		杨洁等[20]	字移编码与行移动编码结合	调整的一个单词或一行嵌入 1 bit	降低隐蔽性提高嵌入容量
	字符编码	付兵等[21]	基于 Unicode 编码奇偶性与字符属性修改	每个字符嵌入 1 bit	严格按照初始位置提取，水印稳健性差
		陆绿等[22]	基于 Unicode 编码奇偶性	每个字嵌入 1 bit	只适用于 Unicode 编码字符
		张洪礼等[23]	基于词平台汉字编码	每个词嵌入 2 bit	编码资源冗余
		刘豪等[24]	修改字符蓝色分量编码值	每个字嵌入 1 bit	隐蔽性低，嵌入容量低
		Tang 等[25]	基于字符颜色属性 RGB 的 3 个通道的最低位值	每个字嵌入 3 bit	隐蔽性较低
		Liu 等[26]	基于不可见 ASCII 字符替换	每个段落嵌入 1 bit	嵌入能力低
		崔光明等[27]	基于不可见 ASCII 字符替换	与文本中空格数量与约束函数选择相关	受限于字符的编码方式
		张震宇等[28]	基于 Unicode 编码的不可见字符嵌入	每个句子嵌入 2 bit	受限于字符的编码方式
		Stefano 等[29]	基于 Unicode 编码同形字	同形字符嵌入 1 bit	稳健性较差
	文件格式	杨德明等[30]	word 文档中未使用属性隐藏信息	与其未使用属性空间大小相关	仅用于特定版本的 word 文档
		Zhong 等[31]	pdf 文档行末标识符不显示的特性	每行嵌入 1 bit	仅用于 pdf 文档
		Yang 等[32]	xml 文档标记的排列组合	与其标签排列组合数相关	仅用于 xml 文档
基于文本内容的算法	基于语法方法	赵敏之等[33]	增删句中助词“的”字	修改一个“的”字嵌入 1 bit	分析句子过程较复杂
		Meng 等[34]	利用句式变换改变句长	修改一个句子长度嵌入 1 bit	计算复杂度高，嵌入率低
		Meng 等[35]	计算句子熵提取零水印	零水印与阈值的设定有关	未考虑停用词的影响
	基于语义方法	Chiang 等[37]	基于同义词替换	每替换一个同义词嵌入 1 bit	算法复杂，且无法抵抗同义词替换攻击
		甘灿等[38]	筛选出适合替换中文同义词	取决于同义词编码位数	无法抵抗同义词替换攻击
		姜传贤等[39]	筛选出中文文本主题句并用句内同义词替换	取决于主题句数量与同义词编码方式	无法抵抗同义词替换攻击
		Chang 等[40]	英文同义词筛选与顶点着色编码	取决于同义词编码方式	无法抵抗同义词替换攻击
		杨潇等[41]	基于矩阵编码的同义词替换	取决于矩阵编码效率	无法抵抗同义词替换攻击
		Wu 等[42]	基于 LZW 同义词替换	取决 LZW 效率	无法抵抗同义词替换攻击
		Zhang 等[43]	基于词阶图无载体信息隐藏	以词为嵌入单位，容量不受限	顺序提取，且通信过程开销较大
	基于汉语特征方法	Sun 等[44]	基于左右结构汉字偏旁拆分	每个左右结构汉字嵌入 1 bit	无法抵抗对左右结构汉字的添加、删除、替换攻击
		Wang 等[45]	将左右、上下结构汉字偏旁拆分结合可逆的方法	每个上下、左右结构汉字嵌入 1 bit	无法抵抗对上下、左右结构汉字的添加、删除、替换攻击
		Fei 等[46]	基于汉字多音字	相邻 2 个多音字间嵌入 16 bit	无法抵抗对多音字的添加、删除、替换攻击
		Zhu 等[47]	基于汉字声韵母的零水印方法	取决于音数值阈值的选取	对文本内容的添加、删除、替换攻击水印的稳健性较弱
		Tang 等[48]	以 8 笔画汉字定位嵌入点	每个 8 笔画汉字嵌入 16 bit	无法抵抗对 8 笔画汉字的添加、删除、替换攻击
		Liu 等[49]	基于汉字笔画数的零水印方法	取决于选取笔画数设定的阈值	无法抵抗对汉字的添加、删除、替换攻击
		孙新梅等[50]	基于简体字与繁体字替换	取决于其算法选取的嵌入方式	无法抵抗对嵌入位置上汉字的添加、删除、替换攻击

3) 基于文本内容的算法

该类算法以修改文本内容和提取文本特征为主，研究成果也相对较多。对于修改文本内容的语法方法[33-35]与语义方法[37-43]而言，虽然在嵌入级别上有差别，但是绝大多数算法在嵌入信息前均需要对文本进行分词，使用自然语言处理技术分析句子，因此会产生较为复杂的计算过程。然而，基于汉语特征的方法[44-50]以字符为处理对象，可以省去较为烦琐的自然语言分析，并且获取的文本特征能够用于构造零水印[46-49]，提高隐蔽性。无论是基于语义、语法，还是汉语特征的算法，嵌入信息都容易受到对文本内容的增加、删除、替换攻击。这些攻击一旦发生必定严重影响隐藏信息提取的精确性。因此，为了提高算法的安全性与隐蔽性，此类算法在改进的过程中，要么尽量保持内容修改过后与原文的一致性，要么尽可能地挖掘文本中的特征，使其嵌入之后不易被人察觉。

表 2 还进一步给出了不同类文本信息隐藏算法在抵抗攻击能力上的表现。基于文本图像的算法抗图像噪声攻击能力弱，但其能够抵抗文本格式变换攻击。基于字符编码与文本格式的算法对文本内容的复制攻击具有较好的稳健性，但很难抵抗文本格式攻击。基于文本内容的算法具有很强的抵抗噪声与文本格式攻击的能力，但是文本内容上的增删、替换都会破坏隐藏信息。

4　现阶段的中文文本信息隐藏的主要问题及解决办法

归纳起来，现阶段的中文文本信息隐藏仍然存在以下几个主要问题。

1) 嵌入容量较小

文本的嵌入容量主要受到两方面因素的制约：文本载体冗余空间较少；研究者设计的任何嵌入算法，只要其使用修改载体的方式以达到嵌入水印的目的，那么在水印嵌入的过程中都要考虑嵌入容量过大是否会造成载体被修改后的隐蔽性降低。考虑到这些问题将会使嵌入算法具有嵌入容量上限，从而导致信息嵌入容量较小。

2) 稳健性较差

已有算法的局限中，稳健性不高是多数算法所具有的一致问题。该问题表现在当文本图像局部受噪声污染、文件格式整体被替换、文本内容部分被篡改等情况下，隐藏信息无法完整提取或者全部被破坏。

3) 汉语结构复杂

英文字母书写简单，26 个字母可组成所有单词。而汉字字库庞大，汉字笔画构词汉字的方式多样。利用字符图像与汉语特征嵌入信息的算法，需要对汉字的笔画结构、偏旁结构、字体结构进行较为复杂的分析后，才能发现合适的嵌入条件。基于汉语语法和语义的隐藏算法，也需要分析汉语句子结构与词语依赖等问题。

4) 算法通用性不高

从已有算法的分类对比来看，算法的通用性较弱不仅表现在不同语言的信息隐藏算法不能共用外，还表现在大部分论文中提出的算法，要么以某一特点的文件格式嵌入，要么针对文本中的某一特定属性，或者着眼于文本中某一内容特征上。将文本图像、文本格式、文本内容算法相互融合的方法不多，不同类别方法相互迁移的研究成果也较少。因此，在中文文本信息隐藏算法的通用性研究上还有更大的提升空间。

相应地，未来本领域的研究预计将重点围绕以下几个方向展开。

1) 提高信息隐藏容量

信息隐藏容量是隐写算法始终关注的问题。如何在文本冗余空间十分有限、嵌入尽量减少失真的条件下，尽可能嵌入更多的信息是研究的重点。对于文本图像而言，除行特征与统计特征外，可以进一步挖掘字符图像的连通域、孔洞数、骨架等特征[51-52]，利用这些特征来隐藏信息。针对结构化与非结构化文本冗余空间不足的问题，可以在文本中寻找更多潜在的冗余空间，或者将同样内容的文本用其他文件格式存储，从而提高嵌入的相对容量。另外，在文本信息隐藏过程中引入压缩编码的方法对嵌入信息处理，减少嵌入信息需要的相对空间，提高嵌入效率。这些都是未来提高信息隐藏容量值得深入思考和研究的方法。

2) 增强稳健性

算法的稳健性直接关系到嵌入信息的完整性，一个较好的信息隐藏算法必定具有较好的稳健性。在提高信息嵌入算法的稳健性上，对于局部被攻击产生的信息缺失问题，可以引入冗余嵌入或者数据编码校验机制。针对整体被攻击信息完全无法提取问题，应该加强隐蔽通信研究。例如，将深度学习方法与无载体的信息隐藏方法结合[53]，从而不修改文本载体嵌入信

息；或者对嵌入信息后的文本、文本图像、文本内容进行加密处理，防止未经授权的用户对嵌入信息后的载体进行修改，从而在保证载体信息安全性的前提下提高稳健性。

3) 挖掘汉语特点

汉语在语法结构、语义表达上丰富多彩，且汉字在音、形、意方面各有特色。针对嵌入式信息隐藏方法在语言分析方面较为复杂，将来可以建立在对汉语言科学研究的基础上，更加充分地挖掘汉语语法、语义、汉字中的特征，将这些特征用于零水印的构造。对于汉字编码复杂问题，通过研究编码映射方法，用该方法降低算法处理汉字的复杂程度。此外，修改汉字字体、构建汉字字库并对汉字进行微小的变形[54]，均是中文信息隐藏未来在应用上的研究方向。

4) 设计通用性算法

算法通用性弱不仅会制约算法嵌入容量，使算法的抗攻击能力较差，还会导致不同类算法之间较难融合的问题。未来的研究可能会更加关注多方法融合的问题，包括将多种算法结合起来，在各自的优势上设计新的算法；或者使用多种载体的混合隐写方法；或者将文本、文本图像载体两者有机结合，嵌入信息在文本载体中完成，检测隐藏信息用图像识别的方法。此外，还可以借鉴其他语言的隐写方法，将方法与中文信息隐藏融会贯通；或者借助其他方向的信息隐藏方法，以及其他领域的知识来推动文本信息隐藏发展。将来如果能够在通用性算法研究上取得一定进展，那么文本信息隐藏研究领域将会向前迈进一大步。

5　结束语

随着人们的信息安全意识不断增强，企业、家庭和个人在网络中传递文本信息时均会关注信息内容的安全性。文本信息隐藏技术必定会引起广泛的关注，其相关的理论研究也会趋于完善与成熟，研究成果也将投入实际的应用中[55]，从而更好地应对中文文本的信息安全问题。

参考文献：

[1] 刘智涛. 基于信息隐藏技术研究综述[J]. 工业仪表与自动化装置, 2015(3): 13-15.
LIU Z T. Summary of the research based on information hiding technique[J]. Industrial Instrumentation & Automation, 2015(3): 13-15.

[2] 陈威兵, 杨高波, 陈日超, 等. 数字视频真实性和来源的被动取证[J]. 通信学报, 2011, 32(6): 177-183.
CHEN W B, YANG G B, CHEN R C, et al. Digital video passive forensics for its authenticity and source[J]. Journal on Communications, 2011, 32(6): 177-183.

[3] 舒后, 杨潮, 何薇. 基于文本内容的数字水印算法的设计与实现[J]. 计算机工程与设计, 2008, 29(5): 1299-1302.
SHU H, YANG C, HE W. Design and implementation of digital watermarking algorithm based on text[J]. Computer Engineering and Design, 2008, 29(5): 1299-1302.

[4] FENG B, WANG Z H, WANG D, et al. A novel, reversible, Chinese text information hiding scheme based on lookalike traditional and simplified Chinese characters[J]. KSII Transactions on Internet & Information Systems, 2014, 8(1): 269-281.

[5] KAMARUDDIN N S, KAMSIN A, POR L Y, et al. A review of text watermarking: theory, methods and applications[J]. IEEE Access, 2018, 6(1): 8011-8028.

[6] 林新建, 唐向宏, 王静. 编码与同义词替换结合的可逆文本水印算法[J]. 中文信息学报, 2015, 29(4): 151-158.
LIN X J, TANG X H, WANG J. A reversible text watermarking algorithm based on coding and synonymy substitution[J]. Journal of Chinese Information Processing, 2015, 29(4): 151-158.

[7] 冯春晖, 徐正全, 郑兴辉, 等. 数字可视媒体取证[J]. 通信学报, 2014, 35(4): 155-165.
FENG C H, XU Z Q, ZHENG X H, et al. Digital visual media forensics[J]. Journal on Communications, 2014, 35(4): 155-165.

[8] JALIL Z, MIRZA A M, JABEEN H. Word length based zero-watermarking algorithm for tamper detection in text documents[C]// International Conference on Computer Engineering & Technology. IEEE, 2010: 378-382.

[9] 周继军, 杨著, 钮心忻. 文本信息隐藏检测算法研究[J]. 通信学报, 2004, 25(12): 97-101.
ZHOU J J, YANG Z, NIU X X. Research on the detecting algorithm of text document information hiding[J]. Journal on Communications, 2004, 25(12): 97-101.

[10] MILAD T A , LI Q , HOU J. Modern text hiding, text steganalysis, and applications: a comparative analysis[J]. Entropy, 2019, 21(4): 355.

[11] 褚勇俊, 张云华. 面向文本图像的地纹数字水印研究[J]. 工业控制计算机, 2013, 26(3): 65-66.
CHU Y J, ZHANG Y H. Study on document watermarking using block-patterns for text images[J]. Industrial Control Computer, 2013, 26(3): 65-66.

[12] DING H, HONG Y. Interword distance changes represented by sine waves for watermarking text images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 11(12): 1237-1245.

[13] KIM Y W, OH I S. Watermarking text document images using edge direction histograms[J]. Pattern Recognition Letters, 2004, 25(11): 1243-1251.

[14] 赵星阳, 孙继银, 李琳琳, 等. 基于字符阶梯边沿调整的文本水印算法[J]. 计算机应用, 2008, 28(12): 3175-3178.
ZHAO X Y, SUN J Y, LI L L, et al. Watermarking of text images using character step edge adjustment[J]. Computer Application, 2008, 28(12): 3175-3178.

[15] BEHROOZ K, BEHNAM K, BAHMAN K, et al. A new method for pdf steganography in justified texts[J]. Journal of Information Security and Applications, 2019, (45): 61-70.

[16] 亓文法, 李晓龙, 杨斌,等. 用于信息追踪的文本水印算法[J]. 通信

学报, 2008, 29(10): 183-190.
QI W F, LI X L, YANG B, et al. Document watermarking scheme for information tracking[J]. Journal on Communications, 2008, 29(10): 183-190.

[17] TAN L N, SUN X, SUN G. Print-scan resilient text image watermarking based on stroke direction modulation for Chinese document authentication[J]. Radio engineering, 2012, 21(1): 170-181.

[18] 雷敏, 杨榆, 胡若翔. 基于中文汉字复杂度的抗打印扫描文本水印算法[J]. 北京邮电大学学报, 2015, 38(s1): 58-62.
LEI M, YANG Y, HU R X. A print resilient watermark scheme based on character complexity[J]. Journal of Beijing University of Posts and Telecommunications, 2015, 38(s1): 58-62.

[19] LOW S H, MAXEMCHUK N F, BRASSIL J T, et al. Document marking and identification using both line and word shifting[C]// Infocom 95 Fourteenth Joint Conference of the IEEE Computer & Communications Societies Bringing Information to People. IEEE, 1995: 853-860.

[20] 杨洁, 张敏瑞. 基于行移和字移编码的二值文本数字水印技术[J]. 西安邮电学院学报, 2006, 11(3): 101-105.
YANG J, ZHANG M R. Document marking technique based on both line and word shifting[J]. Journal of Xi'an University of Posts and Telecommunications, 2006, 11(3): 101-105.

[21] 付兵. 基于字符 Unicode 编码奇偶性的文本信息隐藏算法研究[J]. 福建电脑, 2008, 24(12): 66.
FU B. Research on text information hiding algorithms based on Unicode coding parity[J]. Fujian Computer, 2008, 24(12): 66.

[22] 陆绿, 方勇. 基于字符 Unicode 奇偶性的数字水印设计与实现[J]. 计算机技术与发展, 2010, 20(8): 176-179.
LU L, FANG Y. Design and implementation of digital watermark based on parity of Unicode[J]. Computer Technology & Development, 2010, 20(8): 176-179.

[23] 张洪礼, 刘丹, 温学谦, 等. 基于词平台汉字编码的文本信息隐藏算法[J]. 计算机工程, 2010, 36(7): 150-152.
ZHANG H L, LIU D, WEN X Q, et al. Text information hiding algorithm based on Chinese characters coding in words platform[J]. Computer Engineering, 2010, 36(7): 150-152.

[24] 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.
LIU H, SUN X M, LIU J B. Color-based watermarking algorithm for text documents[J]. Computer Engineering, 2005, 31(15): 129-131.

[25] TANG X, CHEN M. Design and implementation of information hiding system based on RGB[C]// International Conference on Consumer Electronics. IEEE, 2013: 217-220.

[26] LIU F, LUO P, MA Z, et al. Security secret information hiding based on hash function and invisible ASCII characters replacement[C]// Trustcom/bigdatase/ispa. IEEE, 2016: 1963-1969.

[27] 崔光明, 洪星, 袁翔. 基于不可见字符替换的信息隐藏方法研究[J]. 计算机应用与软件, 2016, 33(4): 277-280.
CUI G M, HONG X, YUAN X, et al. Research on information hiding based on invisible characters replacement [J]. Computer Applications & Software, 2016, 33(4): 277-280.

[28] 张震宇, 李千目, 戚湧. 基于不可见字符的文本水印设计[J]. 南京理工大学学报（自然科学版）, 2017, 41(4): 405-411.
ZHANG Z Y, LI Q M, QI Y. Text watermarking design based on invisible characters[J]. Journal of Nanjing University of Science and Technology, 2017, 41(4): 405-411.

[29] STEFANO G R, FLAVIO B, DANILO M. Content-preserving text watermarking through Unicode homoglyph substitution[C]//. International Database Engineering & Applications Symposium. ACM, 2016: 97-104.

[30] 杨德明, 郭盛. 基于 Word 文档的数据隐藏方法[J]. 计算机应用与软件, 2015, 32(5): 314-318.
YANG D M, GUO S. Data hiding method based on word document[J]. Computer Applications and Software, 2015, 32(5): 314-318.

[31] ZHONG Z Y, GUO Y H, XU G A. Digital watermarking algorithm based on structure of PDF document[J]. Journal of Computer Applications, 2012, 32(10): 2776-2778.

[32] YANG J. Algorithm of XML document information hiding based on equal element[C]// International Conference on Computer Science & Information Technology. IEEE, 2010: 250-253.

[33] 赵敏之, 孙星明, 向华政. 基于虚词变换的自然语言信息隐藏算法研究[J]. 计算机工程与应用, 2006, 42(3): 158-160.
ZHAO M Z, SUN X M, XIANG H Z. Research on the Chinese text steganography based on the modification of the empty word[J]. Computer Engineering and Application, 2006, 42(3): 158-160.

[34] MENG Y J, GUO X P, ZHANG W, et al. Text watermarking algorithm based on sentence length[J]. Computer Engineering and Applications, 2007, 43(32): 52-54.

[35] MENG Y J, GUO T, GUO Z H. Chinese text zero-watermark based on sentence's entropy[C]// International Conference on Multimedia Technology. 2010: 1-4.

[36] 徐迎晖, 杨榆, 钮心忻. 基于语义的文本隐藏方法[J]. 计算机系统应用, 2006, 15(6): 91-94.
XU Y H, YANG Y, NIU X X. Text steganography based on sementic[J]. Application of Computer System, 2006, 15(6):91-94.

[37] CHIANG Y L, CHANG L P, HSIEH W T, et al. Natural language watermarking using semantic substitution for Chinese text[C]// Digital Watermarking, Second International Workshop. DBLP, 2003: 192-140.

[38] 甘灿, 孙星明, 刘玉玲. 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(s1): 137-140.
GAN C, SUN X M, LIU Y L. Improved steganographic algorithm based on synonymy substitution for Chinese text[J]. Journal of Southeast University (Natural Science Edition), 2007, 37(s1): 137-140.

[39] 姜传贤, 陈孝威, 李智. 基于文本重要内容的稳健水印算法[J]. 自动化学报, 2010, 36(9): 1250-1256.
JIANG C X, CHEN X W, LI Z. Robust text watermarking based on significant components[J]. Acta Automatica Sinica, 2010, 36(9): 1250-1256.

[40] CHANG C Y, CLARK S. Practical linguistic steganography using contextual synonym substitution and vertex color coding[C]// The 2010 Conference on Empirical Methods in Natural Language Processing. ACM, 2010: 1194-1203.

[41] 杨潇, 李峰, 向凌云. 基于矩阵编码的同义词替换隐写算法[J]. 小型微型计算机系统, 2015, 36(6): 1296-1300.
YANG X, LI F, XIANG L Y. Synonym substitution-based steganographic algorithm with matrix coding[J]. Journal of Chinese Mini-Micro Computer Systems, 2015, 36(6): 1296-1300.

[42] WU W S, XIANG L Y, WANG W. A synonym substitution-based steganography using LZW compression algorithm[J]. Revista De La Facultad De Ingenieria, 2017, 32(1): 763-770.

[43] ZHANG J J, WANG L C, LIN H J. Coverless text information hiding method based on the rank map[J]. Journal of Internet Technology, 2017, 18(2): 127-434.

[44] SUN X M, LUO G, HUANG H J. Component-based digital watermarking of Chinese texts[C]// International Conference on Information Security. ACM, 2004: 76-81.

[45] WANG Z H, CHANG C C, LIN C C, et al. A reversible information hiding scheme using left-right and up-down Chinese character representation[J]. Journal of Systems and Software, 2009, 82(8): 1362-1369.

[46] FEI W B, TANG X H. A Chinese text watermark algorithm based on polyphone[C]// Cross Strait Quad-regional Radio Science & Wireless Technology Conference. IEEE, 2011: 1215-1218.

[47] ZHU P, XIANG G, SONG W, et al. A text zero-watermarking algorithm based on Chinese phonetic alphabets[J]. Wuhan University Journal of Natural Sciences, 2016, 21(4): 277-282.

[48] TANG X H, WANG L N. Text watermarking algorithm based on the stroke of Chinese characters[C]// International Conference on Multimedia Technology. IEEE, 2011: 794-796.

[49] LIU J, PAN J S. Text zero-watermark based on using strokes of Chinese characters[J]. Computer Engineering & Applications, 2013, 49(9): 99-101.

[50] 孙新梅, 孟朋, 黄刘生. 基于字体的中文信息隐藏算法[J]. 计算机工程与设计, 2013, 34(9): 3063-3067.
SUN X M, MENG P, HUANG L S. Chinese text steganography based on character forms[J]. Computer Engineering & Design, 2013, 34(9) : 3063-3067.

[51] TAN L N, HU K, ZHOU X M, et al. Print-scan invariant text image watermarking for hardcopy document authentication[J]. Multimedia Tools and Applications, 2018(10): 1-23.

[52] KHADIJA G, HASSAN D, RACHID H, et al. A zero-bit Fourier image watermarking for print-cam process[J]. Multimedia Tools and Applications, 2019, 78(2): 2621-2638.

[53] YANG Z L, GUO X Q, CHEN Z M, et al. RNN-Stega: linguistic steganography based on recurrent neural networks[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1280-1259.

[54] XIAO C, ZHANG C, ZHENG X X. Fontcode: Embeding information in text documents using glyph perturbation[J]. ACM Transactions on Graphics, 2018, 37(2): 15.

[55] FANG H, ZHANG W M, ZHOU H, et al. Screen-shooting resilient watermarking[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(6): 1403-1418.

[作者简介]

吴国华（1970− ），男，山东济南人，博士，杭州电子科技大学教授、博士生导师，主要研究方向为保密信息化、定密理论与实务。

龚礼春（1995− ），男，福建南平人，杭州电子科技大学硕士生，主要研究方向为信息内容安全、文本信息隐藏。

袁理锋（1983− ），男，浙江诸暨人，博士，杭州电子科技大学讲师，主要研究方向为图像内容安全、视觉秘密分享。

姚晔（1978− ），男，湖北随州人，博士，杭州电子科技大学讲师，主要研究方向为多媒体内容安全、视频图像智能分析。