一种基于中文字符编码的文本水印算法研究

陈 翔
(长沙师范高等专科学校 教育技术中心,湖南 长沙 410100)

摘 要:随着计算机网络技术的迅速发展,数字文本的取得已经变得非常方便,因为数字文本易于复制和更改的特性,使得版权保护的问题长期阻碍着数字出版行业的发展,然而文本水印技术的出现为解决这一问题提供了可行的方法。针对中文字符的特性,给出了一种利用人眼生理特性和中文字符的书写变化的文本水印方案。通过调整中文字符偏旁部首的位置关系,结合密钥,嵌入水印信息。经过实验分析:该算法具有鲁棒性强、隐蔽性高和抗攻击性强的特点。

关键词:文本水印;中文字符;字符结构;特征编码
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2013)02-0237-04
doi:10.3969/j.issn.1673-629X.2013.02.061

Research of a Text Watermarking Algorithm Based on Chinese Character Coding

CHEN Xiang
(Education Technology Center, Changsha Normal College, Changsha 410100, China)

Abstract: Along with rapid development of the computer network technology, the digital text obtained has become very convenient, because the digital text is easy to copy and change, the copyright protection problem is hindering the digital publishing industry development, however the text watermarking technology to solve this problem provides a feasible method. According to the Chinese characters, present a method based on human physiological characteristics and the Chinese character writing changes to the text watermarking scheme. Through the adjustment of Chinese character radical position relations, combined with the key, embedded watermark information. After experimental analysis: the algorithm has strong robustness, high concealment and resistance to aggressive characteristics.

Key words: text watermarking; Chinese character; character structure; feature coding

0 引 言

随着计算机网络技术的迅速发展,数字文本的取得已经变得非常方便,因为数字文本易于复制和更改的特性,使得版权保护的问题长期阻碍着数字出版行业的发展[1]。文本水印作为一种保护数字文本的技术逐渐成为人们关注的对象。文本水印作为一种数字文本版权保护技术主要是利用文本文档,如:PDF,WORD,CAJ 等文档的冗余空间将秘密信息嵌入到文本文档中,能够对文本数字作品的版权进行有效保护,而且不会影响原始文本的使用,有比较好的隐蔽性[2]。目前相当多的成功算法主要还应用于音频、视频、图像等具有大量可嵌入空间的文件中。而文本字符具有的原子特性,可嵌入的空间比较小,因此增加了数字文本核心技术的解决难度[3]。

例如:文献[4,5]提出在数字文本中嵌入水印的算法有两种:第一种是在数字文本语义中嵌入水印数据;第二种是在数字文本格式中嵌入水印数据。将水印数据嵌入到语义的算法类似于情报学中的密码电文,两者相比较,在数字文本的格式中嵌入水印的算法并不修改原始文本中的具体内容,并能够得到比较好的水印不可见性,获得了数字水印领域的关注。文献[6,7]提出在数字文本中嵌入编码的算法:在空字符位置插入空格的方法、基于中文标点的编码算法、基于字符本身调整编码的语义算法。利用行末空白字符区域嵌入空格的水印算法,此类算法可以在非格式化数字文本中利用。数字文本中随机一行行末嵌入若干空格与没有嵌入空格编码在人眼上很难分辨。文献[7,8]在格式化数字文本中的水印技术提出:文本水印嵌入数据方法,而且对水印检测相关方法做出探讨。行移方法的水印嵌入是利用数字文本的任意一行垂直方向调整,基于某一行被向上调整或向下调整,再通过上

收稿日期:2012-05-26;**修回日期**:2012-08-28
基金项目:湖南省教育科学研究项目(09C123);长沙师范科研基金项目(KYYB201007)
作者简介:陈 翔(1980-),男,湖南长沙人,硕士,实验师,研究方向为信息与网络安全。

下的两行的位置保持不调整。不做调整的上下两行被视为提取水印的相对位置。通过字移方法嵌入水印是利用文本的任意一行中某个字符做水平方向上的调整。特征编码方法是基于修改某个字符的某个字符特性来嵌入水印信息的技术。文献[9]针对文本水印技术进行了探讨,并提出了文本水印研究方向的问题。文献[10]提出基于汉字数学表达式的方法,该方法的关键思路是中文能够描述成汉字偏旁部首作为运算数、偏旁部件之间的内部关系作为表达式。文献[11]提出基于文本水印的冗余空间编码算法。利用字符及字符串内部关系,提出字形不同而语义上相同的字符,设计出针对字符内部关系的特性的合适编码,基于字符特性来插入水印编码。文献[12]提出的水印方案是将文本转化为二值图像或灰度图像,利用图像信息的冗余来嵌入水印信息。文献[13]提出的基于文本格式的水印技术是通过改变文本中的空格符、字符颜色、字符尺寸等嵌入水印信息。

在文本水印技术领域不管针对行移,还是字移的算法,或者是修改字体字号等格式的数字文本的算法,都普遍存在嵌入信息的数据量较小的问题,这势必会在身份认证、版权说明等方面的使用上比较困难;针对恶意的攻击防御能力不强,目前大部分的文本水印算法都不具有防止恶意攻击者用自识别的途径破坏文本水印环境的攻击,而且破坏时付出的代价一般来说都比较小。根据文本语义的算法在数字文本中隐藏信息容易出现基本语义修改后让阅读者很困难分析的语义环境,而且可嵌入的数据量都比较小,在现实环境中应用比较困难。

文中提出一种新的基于人眼视觉特性的算法,运用文本字符的结构变化来嵌入秘密信息,它的基本思想是在于修改文本字符结构,算法在自然语义上相同的字符形体不一样,然而也不影响视觉效果,并对文本字符修改的内部关系特征实施编码,利用中文书写的细微变化来插入水印数据,最终生成水印文档。这种水印文档针对现有文本水印技术中暴露出的例如:鲁棒性不强、文档视觉效果不理想、文本水印容量小等缺陷,对于电子文本信息的文件都适用。

1 算法的基本设计思想

关键的算法思路是调整单个字符或多个连续字符字形结构,构建出自然语义上不变化的然而字符结构不一样的文本。

多种调整字符字形结构的应用都是可以利用的,只要不影响人眼视觉的分辨能力,不影响自然语义上的歧义。例如:比较合适的中文字符形态设计方法是针对修改汉字内部的偏旁部首之间的空间关系来修改结构,因为没有统一规定的字符的结构,对字符的内部结构的定义是确定的,在对文本字符自然语义识别的过程期间,对汉字大小及外形的调整具有自识别修改功能。相同单个中文或者句子、段落之间会因字符形态的差异和书写格式的不同,汉字结构也有巨大的区别,但这种情况的不同是可以允许的,并不会对语义产生影响。

基于人眼视觉理论,嵌入水印数据,增强水印不可见性和抗攻击性。如图 1 所示,可以对汉字各笔划之间的位置关系做调整和修改。

图 1 中文字符结构示例

如图,同一宽带、高度,但书写风格略有区别的字体,都是同一字体,在部分笔画上做了调整。然而针对汉字的语义来判读,小空间的偏旁部首的修改很难影响对汉字语义环境的识别。不过却引起各偏旁部首笔划之间的位置关系的改变,从而引起字体结构外观的很大的变化。

对单个中文字符内部位置的调整,从而改变整个中文字符串结构。把字符串当成一个比较复杂的单个字符,针对字符之间的笔画位置关系,来改变字符串的整体结构。

文中针对字符串整体结构进行讨论,结合字符形态实施的思路综合考虑以下几个方面的内容:

1)字符串结构编码。目前的文本数字水印中,字符串编码的设计应与单个字符和规定字形的编码方式融合起来,针对特定的中文字符规则来构建相应的字体内部及外延结构。应用的构建规则是:字符串内部结构的调整可以有不同的编码方式。在形式多样的字符串的应用中,应充分考虑合适的编码的规则,定义合适的字符串内部结构,取得较大的水印空间,进而提高数字文本水印容量。

2)字符串书写方式编码。在需要嵌入水印信息的数字文本中,如果嵌入水印的字符串之间书写方式越相似,那么带来的视觉上的差异就越小,然而这需要文本中的字符串具有相似的书写方式及字符大小。另一方面,有可能原始文本文件的字符串之间具有不同的书写方式,因此,还应根据不同的书写方式,设计具有相同内部结构,但具有不同书写格式的形态字符串,这种方法的优点是不改变水印检测方法而扩充了字符串之间的内部结构的设计。

2　水印的嵌入与提取算法

2.1　嵌入算法

根据以上算法的基本思想,考虑实施以下的水印嵌入算法。初始算法时需要输入若干参数信息:如:原始文本位置 $W1$、水印密钥 E、秘密信息 Q、水印文本生成位置 $W2$。对该算法的相关步骤引入定义。

定义 1 嵌入水印字符串,是指在生成的水印文本中,已经嵌入秘密信息的字符串队列 $T(n)$,嵌入水印信息前需要水印密钥 Q。

算法相关的步骤如下:

步骤 1 使用哈希函数和密钥随机选取文本字符串,使用密钥合并水印信息嵌入选取的字符串队列 Ti 。

步骤 2 $i=1$。

步骤 3 遍历选取字符串队列:

①Text= $T[i]$。

②根据相似度在特征编码集合中查找适应 i 位置的特征编码(集合为 W)。

③遍历 $i+n$ 个位置的特征编码集合中的 W 集合:

a)计算 W 特征编码的相似度 H 及其字符串融合编码。根据 $i+n$ 个位置特征编码相似度 ptring,字符及字符串内部相似度的大小,对编码内的词语升序排序,调整阈值 a 的大小和去除小于阈值 b 的融合编码序列。

b)计算 WI 编码。合并密钥得到 WI。编码表示为 FF。

c)嵌入秘密信息编码。第二次遍历 W。表示 W 编码和 T 一致,则用 WI 替换原始文件中的字符串, i 加 1。

④如果嵌入字符串为空,结束遍历。

2.2　提取算法

水印提取算法的步骤为:

步骤 1 根据哈希函数和密钥以及水印文本,确定水印字符串队列 T。

步骤 2 $i=l$。

步骤 3 遍历水印字符串队列 Ti:

1. 根据相似度在特征编码集合中查找适应 i 位置的特征编码(集合为 W)。

2. 遍历 $i+n$ 个位置的特征编码集合中的 W 集合:

a)计算 W 特征编码的相似度 H 及其字符串融合编码。根据 $i+n$ 个位置特征编码相似度 ptring,字符及字符串内部相似度的大小,对编码内的词语升序排序,调整阈值 a 的大小和去除小于阈值 b 的融合编码序列。

b) 计算 WI 编码。合并密钥得到 WI。编码表示为 FF。

c) 提取秘密信息。第二次遍历 W。表示 W 编码和 T 一致,则用 WI 替换原始文件中的字符串, i 加 1。

3. 如果嵌入字符串为空,结束遍历。

3　实验及性能分析

笔者设计了一个可行的实验过程:将一些秘密信息嵌入到该文章中构成嵌入字符串,用单项散列函数获取样本字符串的修改值,再通过 AES 算法对修改值加密,最后加密数据直接嵌入到样本字符串中,生成水印文本。实验的硬件环境为:CPU:英特尔双核 E5300,内存 1G。软件环境为:Windows XP,算法实现语言为 Visual C++6.0。

3.1　鲁棒性实验

对于该算法来说,用户的普通编辑在一定范围内是许可的,并且不会因这些编辑而丢失重要信息。而对于恶意攻击文本内容的操作也具有相应的抵抗力,与基于行移、字移等特征编码的文本水印算法很难实现相比,普通的编辑操作很容易移除水印信息。选取该文章一段验证该水印算法。中文文本格式为楷体小四。嵌入水印信息为 11001100。为分析该水印算法的鲁棒性,对选取的文章为攻击对象,对其实施字体大小放缩、加噪、滤波、剪切等操作后,最后提取水印信息并比较结果。表 1 中含经过各种操作后提取的水印信息码、误码率等。

表 1　中文水印文本经过各种操作后提取的水印码及误码率

操作方式	1 行	2 行	3 行	提取水印码	误码率
放大字体 1/2 倍	1.652	0.818	1.410	11001100	0
放大字体 1 倍	1.054	1.211	0.620	11001100	0
加噪 6%	1.059	0.792	1.212	11001100	0
加噪 12%	1.055	0.770	0.744	10001100	12.5%
滤波	0.879	0.686	1.122	00001100	12.5%
剪切 1/2	0.845	0.775	1.030	11001101	12.5%
无操作	1.046	0.770	1.151	11001100	0

3.2　性能分析

对该算法从隐蔽性、鲁棒性等方面进行探讨。

隐蔽性:由于该算法根据中文字符及字符串特征向原始文本嵌入信息,视觉影响很小,故不会对原始文本造成视觉上的影响,具有不可见性,是比较理想的水印处理。图 2 为嵌入前后的文本。

对于文中的算法鲁棒性水印对用户的普通操作是许可的;另一面,水印针对恶意攻击(如文本恶意修改)也有一定的抵抗能力。以往基于特征编码和纯格式的水印算法不能实现这一点,普通排版和修改都会丢失水印信息。该算法是从中文字符的角度考虑的水印,不受到中文字符字体、间距等的限制。通过实验可以知道该算法在鲁棒性和抗攻击性两方面都有提高。

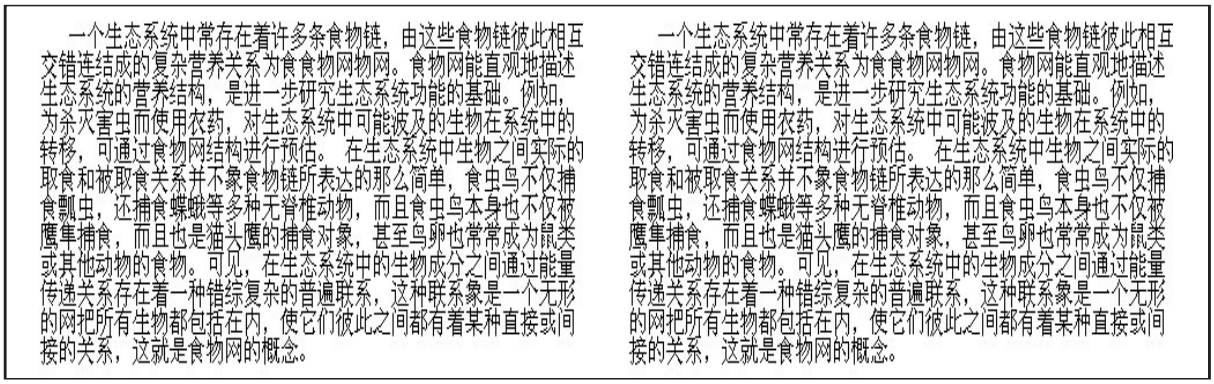

图 2 水印嵌入前后文本比较

4 结束语

在现实活动过程中大部分数字文本文件比图像等多媒体文本的数据更具有商业价值。目前在计算机网络技术的广泛应用下，相当部分的纸质出版物都向电子文本文件转化，最终可以预料到在电子商务市场中会占据比较大的市场份额。文本数字水印技术作为数字水印领域的一个重要分支，它的出现为解决电子文本被恶意拷贝、销售或伪造提供了办法。因此文本水印技术能够很好地解决有价值数字文本的版权保护问题。

文本数字水印领域是在人类视觉理论科学、网络技术、密码学原理等学科基础上发展起来的。可以说是涉及多个计算机领域的分支科学，技术含量和理论深度都比较高，在现实生活中广泛应用的问题还需要深入研究，是目前具有极高价值的研究新方向之一，也必定能得到相应的社会和经济效益。

文中从汉字文本的内部关系出发，基于人眼生理特征，提出基于中文字符的算法具有视觉隐蔽性，鲁棒性较高、抗攻击性强的特点，对于电子中文出版读物有版权保护作用，可推广到电子阅读器等产品中使用。

参考文献：

[1] 张 宇，刘 挺，陈毅恒，等. 自然语言文本水印[J]. 中文信息学报，2005，19(1)：56-62.

[2] 刘曼昊，孙堡垒，郭云彪. 文本数字水印技术研究综述[J]. 东南大学学报，2007，37(1)：225-230.

[3] 刘 超，孙星明，周新民. 基于模糊聚类方法的盲文本水印算法研究[J]. 计算机应用研究，2007，24(2)：148-150.

[4] Brassil J T，Low S，Maxemchuk N F. Copyright protection for the electronic distribution of text documents[J]. Proceedings of IEEE，1999，87(7)：1181-1196.

[5] Huang D，Yan H. Interword distance changes represented by sine waves for watermarking text images[J]. IEEE Transactions on Circuits and Systems for Video Technology，2001，11(12)：1237-1245.

[6] Wong P W，Memon N. Secret and Public Key Image Watermarking Schemes for Image Authentication and Ownership Verification[J]. IEEE Transaction on Image Processing，2001，10(10)：1593-1601.

[7] Kim Young-Won，Moon Kyung-Ae，Oh Il-Seok. A Text Watermarking Algorithm Based on Word Classification and Interword[C]//Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDR). [s. l.]：[s. n.]，2003：78-90.

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

[9] 黄 华，齐 春，李 俊，等. 文本数字水印[J]. 中文信息学报，2001，15(5)：52-56.

[10] 孙星明，殷建平，陈火旺，等. 汉字的数学表达式研究[J]. 计算机研究与发展，2002，39(6)：701-711.

[11] 刘 东，周明天. 一种文本数字水印系统解决方案[J]. 计算机应用，2006，26(1)：84-86.

[12] 李 刚，杨 杰. 一种基于二值印刷图像的数字水印方案[J]. 上海交通大学学报，2005，39(4)：570-573.

[13] 唐承亮，肖海青，向华政. 基于文字 RGB 颜色变化的脆弱型文本数字水印技术[J]. 计算机工程与应用，2005(36)：6-8.

一种基于中文字符编码的文本水印算法研究

作者：陈翔
作者单位：长沙师范高等专科学校 教育技术中心,湖南 长沙410100
刊名：计算机技术与发展
英文刊名：Computer Technology and Development
年，卷(期)：2013(2)