基于空样式的网页水印方法

陈韦旭 陈建平 文万志 蔡 亮
(南通大学计算机科学与技术学院 江苏 南通 226019)

摘 要 网页水印在网页版权保护和防仿冒篡改等方面具有重要应用,现有的网页水印技术主要是利用 HTML 对一些格式变化的不敏感性进行水印信息的嵌入,这类方法嵌入的水印信息与网页内容严重分离,水印的隐蔽性不强,容易受到攻击。文中提出一种新的基于空样式的网页水印方法,利用 HTML 对于没有定义的样式不做任何操作的性质,将水印信息转换为没有内容定义的空样式,并嵌入到网页的 HTML 代码中,使得水印信息与 HTML 代码联系紧密,隐蔽性强,不易被检测和攻击,同时拥有较大的水印信息容量。与现有方法相比,所提方法具有较大的优越性。

关键词 网页水印,信息隐藏,空样式,版权保护
中图法分类号 TP309.2 **文献标识码** A

New Method for Webpage Watermarking Based on Empty Styles

CHEN Wei-xu CHEN Jian-ping WEN Wan-zhi CAI Liang
(School of Computer Science and Technology, Nantong University, Nantong, Jiangsu 226019, China)

Abstract Webpage watermarking has important applications in webpage copyright protection and anti-tampering. The existing webpage watermarking methods mainly embed watermark information by using the insensitivity of HTML to some format changes. Such methods have the problems that watermark information is separated from the webpage contents, which makes the watermark less concealed and vulnerable to attacks. This paper proposed a new webpage watermarking method based on empty styles. Making use of HTML's feature that it does no operation to a style without content definition, the method transforms watermark information into empty styles and embeds them into the HTML code. The embedded watermark information is closely linked with the HTML code and is of strong concealment. It is not easy to detect and attack, and has a large watermark capacity as well. Compared with the existing methods, the proposed method is more superior.

Keywords Webpage watermarking, Information hiding, Empty style, Copyright protection

1 引言

随着信息技术和互联网的普及发展,网站和网页得到了广泛应用,它给政府部门和企事业单位的生产、经营、管理以及人们的工作和生活带来极大的便利。与此同时,网页的非法复制、网页的仿冒现象也日益严重,是一个亟待解决的问题。网页水印是近年来出现的网页保护的一项技术,它通过某种方式在网页中嵌入版权标识信息或身份认证信息(水印),当发现网页遭到非法复制或仿冒时,可以提取这些信息来证明网页的版权归属,鉴别网页的真伪,从而确认非法复制和仿冒行为。除此之外,网页水印技术还可用于网页防篡改、在网页中隐藏和传递秘密信息等方面。

与传统的图像、视频和音频水印技术的研究相比,网页水印技术的研究目前还很少。网页的构造与图像、音频文件不同,常用的图像和音频水印技术不适用于网页水印。表示网页的 HTML 代码结构简单、冗余度少,在网页中嵌入水印难度较大。现有的网页水印嵌入方法主要是基于 HTML 语言对于一些信息的改变不敏感的原理,通过改变这些不敏感的信息来实现水印信息的嵌入。比如,通过改变标签名和属性名的大小写实现水印信息的嵌入[1],通过改变空格的个数实现水印信息的嵌入[2],通过定义不存在的标签实现水印信息的嵌入[3],以及使用不同的代码格式实现水印信息的嵌入[4]等。

以上这些网页水印的嵌入方法存在明显的不足。改变大小写或改变空格数的方法会使得 HTML 代码的大小写或空格数不断变换,很容易被识别,水印的隐蔽性不好,抗检测能力很弱,只要用一个简单的过滤系统将 HTML 代码中所有字母改成小写或所有空格数改为一个,便能轻松去除水印。对于使用不存在的标签的方法,由于 HTML 代码中的标签是固定的,定义不存在的标签很容易被识别,隐蔽性和抗攻击能力也不强。最后一种方法虽然隐藏性好一些,但是代码嵌入点很少,水印信息容量小。

针对上述问题,本文提出一种基于空样式的网页水印的嵌入方法,将水印信息转换为 HTML 代码中的样式,嵌入到网页的 HTML 代码中,该水印具有很好的隐蔽性和抗攻击性。

2 方法的描述

2.1 样式与空样式

网页是一种存储在 Web 服务器上,通过网络进行传输,被

本文受国家自然科学基金项目(61602267),南通市应用基础研究项目(GY2015012)资助。
陈韦旭(1992-),男,硕士生,主要研究方向为信息安全;**陈建平**(1960-),男,硕士,教授,主要研究方向为信息安全,E-mail:chen.jp@ntu.edu.cn(通信作者);**文万志**(1983-),男,博士,副教授,主要研究方向为软件安全;**蔡 亮**(1992-),男,硕士生,主要研究方向为信息安全。

浏览器解析和显示的文档类型，其内容由 HTML 语言构成。

网页的样式是指使得网页页面显示达到一定效果的辅助代码或文件，它可以改变网页中元素的外观，如改变按钮的颜色、表格的大小等。网页的样式一般会用层叠样式表 CSS (Cascading Style Sheets)进行封装，将一系列样式封装在一起并存储在 CSS 文件中，HTML 使用样式名调用封装的样式，以显示样式所定义的外观。例如样式调用〈div class＝“suspend susp_nav”〉content〈/div〉，其中 suspend 和 susp_nav 是两个封装好的样式，div 层中的 content 显示效果就是这两个样式表效果的叠加效果。

样式通常由样式名和表示样式内容的定义构成。如果一个样式只有样式名，没有内容定义，则可将其称为空样式。例如样式调用〈div class＝“susp”〉content〈/div〉，如果 CSS 文件中没有样式名为 susp 的样式定义，HTML 代码中也没有定义 susp 的样式，那么 susp 为空样式。根据 CSS 规范，HTML 中的元素不会匹配错误的或者不存在的样式名，因此上述调用不会进行任何操作。同时，上述调用不违反 HTML 的语法规范，对网页的运行不会产生任何影响。本方法利用 HTML 的这种特性，将水印信息转换为自定义的空样式嵌入到 HTML 代码中，实现水印的嵌入与提取。

2.2　水印信息预处理

水印信息通常为一组带有版权、归属权等信息的字符串，由英文、中文或其他字符组成。为了便于水印信息的嵌入，需要把水印字符串转换成一定形式的编码，如常用的 ASCII 编码。ASCII 码对于嵌入英文水印信息比较方便，为了便于嵌入中文和其他字符，使方法更具通用性，本文采用 UNICODE 编码。UNICODE 编码把包括中文、英文在内的各种语言和符号用 4 位十六进制数表示。如字符串“copyright 南通大学”，转换成 UNICODE 码为：

\u0063\u006f\u0070\u0079\u0072\u0069\u0067\u0068\u0074\u5357\u901a\u5927\u5b66

去除每个字符编码的码头\u，得到处理后的水印信息为：

0063006f00700079007200690067006800745357901a59275b66

2.3　空样式的设计

经过预处理之后，原始水印信息被转换为由 UNICODE 编码表示的十六进制的码串。下一步任务是将十六进制的水印信息码串转换成 HTML 的空样式并嵌入到网页中。水印信息码串包含数字 0～9 和字母 a～f 共 16 种码元，我们设计 16 个不同的空样式一一对应这 16 种码元。空样式的设计遵循以下两个方面的原则：一方面，样式的名字与 HTML 代码的内容要有一定的联系，具有合理性和真实性，让人觉察不到它是空样式，从而使水印具有良好的隐蔽性；另一方面，空样式的名字不能与现有样式的名字相同，若命名相同，则会导致水印信息提取出错。

以中国建设银行官网信用卡页面为例，其 HTML 代码中使用了一个名为“aright”的样式，由此可以设计一个名为“bright”的空样式，这个空样式名与真实的样式名很相似，让人难以分辨。类似地，通过分析该 HTML 代码中样式命名的特点，可以设计出 16 个空样式，分别对应水印信息的 16 种码元，如表 1 所列。

表 1　空样式设计

码元	空样式	码元	空样式
0	susp_nave	8	zz_top2_rig
1	susp_hind	9	showt_del
2	user_det	a	commited
3	hover_sus	b	box_white
4	sroclcom	c	hind_sit
5	bright	d	card_info
6	busine_menu	e	icon_get
7	zxkf_radio	f	index_right

2.4　空样式的嵌入

在设计好水印信息码元对应的空样式后，下一步任务是将空样式嵌入到 HTML 代码中。

在 HTML 代码中，网页上要显示的内容通常包含在一个个的层级元素中。例如，样式调用〈div class＝“suspend”〉content〈/div〉，页面上显示的内容 content 被包含在这个 div 层之中。其中的 class＝“suspend”设置了 content 的样式，即外观。可以单独增加一条这样的语句来嵌入空样式，例如嵌入上表中的第一个空样式 susp_nave，可在原 HTML 代码中增加一条语句〈div class＝“susp_nave”〉content〈/div〉。这种方式不是很好，一方面会增加 div，class 这些与水印信息无关的字符串，增大了网页文件的大小；另一方面，嵌入的语句有些明显，隐蔽性不够好，可能会被察觉出来。我们利用 HTML 的样式可以叠加的特点，将空样式作为一个叠加的样式嵌入到已存在的层级之中，而不添加新的语句。比如，将空样式 susp_nave 嵌入到原有的语句＜div class＝“suspend”〉content〈/div〉中，嵌入之后原代码变为〈div class＝“suspend susp_nave”〉content〈/div〉，这种嵌入方式增强了隐蔽性，对原 HTML 代码的影响很小。

3　实现步骤

算法的整体框架如图 1 所示，其包含两部分，一部分为水印信息的嵌入，另一部分为水印信息的提取，其中虚线框部分为可选的。

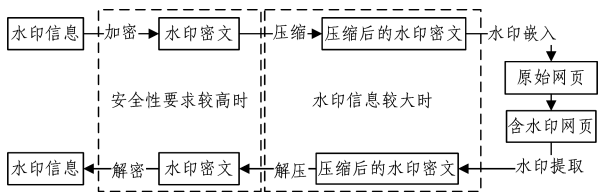

图 1　算法整体框架

3.1　水印的嵌入

第 1 步　水印信息预处理。将水印信息的每个字符转换为对应的 UNICODE 码，形成十六进制的水印信息码串。

第 2 步　设计空样式。根据网页 HTML 代码中样式命名的特点，设计 16 个空样式，每个空样式对应一种十六进制的水印信息码元，形成一个空样式表。

第 3 步　将水印信息转换成空样式。根据空样式表，将十六进制的水印信息码串的每一个码元转换成对应的空样式。

第 4 步　空样式的嵌入。查找网页 HTML 代码中的样式，依次将每个空样式叠加插入到原有的样式之后。

3.2　水印的提取

第 1 步　提取空样式。根据空样式表中的样式名，从网页的 HTML 代码中依次提取出每个空样式。

第 2 步　生成十六进制水印信息码串。对照空样式表，将提取的每个空样式转换成对应的十六进制的水印信息码元，从而形成十六进制的水印信息码串。

第 3 步 生成原始水印信息。根据 UNICODE 编码，将十六进制的水印信息码串转换成对应的水印信息字符串。

4 实验测试

选取中国建设银行官网信用卡页面进行实验测试，页面的截图如图 2 所示。通过对该页面的 HTML 代码进行分析，设计出如表 1 所列的空样式表。

图 2 实验网页截图

使用包含中英文在内的字符串“copyright 南通大学”作为水印信息，将水印信息转换成十六进制的水印信息码串，过程如图 3 所示。

```
水印信息：copyright 南通大学
unicode 码：\u0063\u006f\u0070\u0079\u0072\u0069\u0067\u0068\u0074
\u5357\u901a\u5927\u5b66
十六进制水印信息：0063006f007000790072006900670068007453579 01a59
275b66
```

图 3 水印信息预处理

将水印信息码串中的每个码元转换为空样式，并嵌入到 HTML 代码中。以嵌入前 3 个码元 006 为例，嵌入水印前相关的 HTML 代码为：

〈div id=“fadee” class=“shade”〉〈div〉

〈div class=“sidebar”〉

〈div class=“suspend susp_nav” id=“suspend”〉

嵌入水印后的 HTML 代码为：

〈div id=“fadee” class=“shade susp_nave”〉〈div〉

〈div class=“sidebar susp_nave”〉

〈div class=“suspend susp_nav busine_menu” id=“suspend”〉

如上述代码所示，这 3 个码元对应的空样式 susp_nave，susp_nave 和 busine_menu 分别嵌在原有的样式 shade，sidebar 和 susp_nav 后面。

嵌入水印后网页的页面没有发生任何变化，仍如图 1 所示。水印的提取过程如图 4 所示，最终成功提取出网页中嵌入的水印信息“copyright 南通大学”。

```
十六进制水印信息：0063006f007000790072006900670068007453579 01a592
75b66
unicode 码：\u0063\u006f\u0070\u0079\u0072\u0069\u0067\u0068\u0074
\u5357\u901a\u5927\u5b66
水印信息：copyright 南通大学
```

图 4 水印提取过程

5 性能评价

网页水印的性能主要有隐蔽性、抗攻击性以及水印嵌入容量等。隐蔽性是指阻碍人识别出网页中存在水印信息的能力；抗攻击性是指嵌入的水印在经受攻击后仍能被正常提取的能力，网页水印的攻击手段主要有格式变换、水印检测等；水印容量则是可以嵌入水印信息量的大小。

如前所述，现有的几种网页水印方法将水印信息转换为 HTML 代码中的不同格式，这种做法使得水印信息和 HTML 代码严重分离，水印的痕迹比较明显，容易被识别，一些简单的攻击方式就能破解。与这些方法相比，基于空样式的网页水印方法具有更好的隐蔽性和抗攻击性。一方面，空样式名与原有样式名非常相似，不易引起注意和怀疑；另一方面，样式内容的定义通常存放在 CSS 文件中，水印的攻击者根据 HTML 代码难以区分其中的真实样式和空样式，无法检测到空样式的存在。如果试探性地修改样式名，会有很大的风险，若改动了真实样式名，会对页面的显示产生影响，因此较难用过滤系统去除网页中的水印信息。

在以上分析的基础上，我们在实验中进一步对嵌入的水印进行了抗攻击测试，包括格式变换和水印检测等。

(1)格式变换

对上述嵌入水印的中国建设银行官网网页文件进行各种格式变换，包括改变 HTML 代码的空格数、标签大小写、调换标签属性的顺序等，以及去除代码中不符合代码规范的多余字符。实验结果显示，网页中的水印信息“copyright 南通大学”均能被正确并完整地提取出来。

(2)水印检测

基于空样式的网页水印信息是网页样式的一种，与 HTML 代码紧密联系，完全符合 HTML 代码规范，根据常规语法难以检测出水印信息的存在。我们使用市面上常用的网页水印检测系统检测含水印网页代码，结果表明未能检测出网页中的水印信息。

本文提出的方法也具有较大的水印信息容量，因为样式在网页的 HTML 代码中大量存在。除了建行官网信用卡页面外，我们还对淘宝、新浪、携程、优酷等主流网站的网页进行了应用测试，表 2 给出了这些网页水印信息的嵌入点数量和可嵌入的信息量。它们中最小的嵌入量为 122 个字符，最大的为 1177 个字符，平均为 432 个字符，这对于各种版权信息或身份信息的嵌入以及一定篇幅的秘密信息隐藏是完全足够的。

表 2 水印信息容量

网页地址	嵌入点	字符数
creditcard. ccb. com	489	122
www. taobao. com	602	150
www. sina. com	1665	416
www. ctrip. com	1192	298
www. youku. com	4708	1177

结束语 互联网时代网页承担着传播信息的重要责任，网页水印技术作为一种保护网页版权、防止网页被仿冒篡改的重要手段，具有重要的研究和应用价值。本文基于空样式的概念，提出一种基于空样式的网页水印方法，它将水印信息转换为有含义的网页代码，使得水印具有很好的隐蔽性和抗攻击性，同时具有较大的水印嵌入空间。与现有的网页水印方法相比，本文提出的方法具有较大的优越性。下一步将在本文所提方法的基础上，对将空样式水印应用于网页防篡改做进一步的研究。

参 考 文 献

[1] SUN P，LU H T. An efficient web page watermarking scheme [C]// 2009 2nd IEEE International Conference on Computer Science and Information Technology. Beijing，2009：163-167.

[2] ZHANG Z，PENG H，LONG X. A Fragile Watermarking Scheme Based on Hash Function for Web Pages[C]// 2011 International Conference on Network Computing and Information Secu-

rity. Guilin，2011：417-420.
[3] LI D J，ZHANG B. DWTC：A Dual Watermarking Scheme Based on Threshold Cryptography for Web Document[C]// 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). Taiyuan，2010：510-514.
[4] CHOU Y C，LIAO H C. A Webpage Data Hiding Method by Using Tag and CSS Attribute Setting[C]// 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Kitakyushu，2014：122-125.
[5] SAINI S. A survey on watermarking web contents for protecting copyright[C]// 2015 International Conference on Innovations in Information，Embedded and Communication Systems (ICIIECS). Coimbatore，2015：1-4.
[6] SONNLEITNER E. A user-tracable watermarking scheme for dynamic web-page content[C]// The First International Conference on Future Generation Communication Technologies. London，2012：121-125.
[7] HE C. Research of Web Resources Protection Based on Digital Watermarking and Digital Signature[C]// 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS). Ostrawva，2016：294-297.
[8] 黄华军，王保卫，孙星明. 基于 CSS 类选择符重复引入的网页信息隐藏算法[J]. 计算机研究与发展，2009，46(z1)：138-142.
[9] 杨旭光，唐文龙. 基于 Shamir 门限和 html 标签 id 的网页水印方法[J]. 计算机系统应用，2013，22(8)：98-102.
[10] 沈勇. 一种基于 HTML 文档的信息隐藏方案[C]// 第一届中国可信计算与信息安全学术会议. 武汉：中国计算机学会，2004：217-220.
[11] NECHTA I. Robustness analysis for dynamic watermarks[C]// 2017 International Multi-Conference on Engineering，Computer and Information Sciences (SIBIRCON). Novosibirsk，Russia，2017：298-300.
[12] YU Z，X J Y，L Y X，et al. A spectrum watermark embedding and extracting method based on spread spectrum technique[C]// 2016 IEEE International Conference on Electronic Information and Communication Technology (ICEICT). Harbin，2016：16-22.
[13] SOBHA R，VSUCHARITHA M. Secure transmission of data using audio watermarking with protection on synchronization attack[C]// 2015 Global Conference on Communication Technologies (GCCT). Thuckalay，2015：592-597.
[14] SAINI S. A survey on watermarking web contents for protecting copyright[C]// 2015 International Conference on Innovations in Information，Embedded and Communication Systems (ICIIECS). Coimbatore，2015：1-4.
[15] 张立和，杨义先，钮心忻，等. 软件水印综述[J]. 软件学报，2003，14(2)：298-300.
[16] 陈真勇，唐龙，唐泽圣，等. 以鲁棒性为目标的数字多水印研究[J]. 计算机学报，2006，29(11)：2037-2043.
[17] 张冠男，王树勋，温泉. 一种嵌入可读水印的自适应盲水印算法[J]. 电子学报，2005，33(2)：308-312.
[18] HARTUNG F，RAMME F. Digital rights management and watermarking of multimedia content for m-commerce applications[J]. IEEE Communications Magazine，2000，38(11)：78-84.
[19] 丁伟. 基于 Web 网页的文本水印技术的研究[D]. 武汉：武汉理工大学，2012.

(上接第 337 页)

4.3 Matlab 仿真

仿真平台为 Matlab R2012a。目前，神经网络的隐层节点数为：

$$m=\sqrt{n+l}+a$$
$$m=\log_2 n \tag{10}$$

其中，m 是隐含层节点数，n 是输入层节点数，l 代表输出层节点，a 是 1～10 之间的常数。本节建立的 BP 神经网络结构模型为 3 层(见图 2)，根据式(5)－式(10)计算出初始隐层节点数在 5～14 之间(其中网络输入层有 23 个，输出层有 1 个)。综合考察网络精度及泛化能力，根据凑试法，可以得出结果，最后确定隐层节点数为 6，因此网络结构为 23-6-1。

4.4 结果分析

通过 $\Delta x_i=(x_i-x_{\min})/(x_{\max}-x_{\min})$ 预处理输入数据，使数据量化至区间[0，1]。其中 x_i 为代表数据，$x_{\min}$ 为最小值，$x_{\max}$ 为最大值。图 8 是改进 BP 神经网络评估态势值与实际态势值的对比结果。

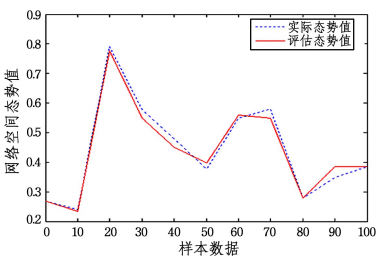

图 8　改进 BP 神经网络评估系统态势值与实际值的比较结果

实验结果表明，不仅基于 BP 神经网络模型的 80 个训练集满足了专家和教授的评价，而且 20 个测试集也与实际情况相符。可知 BP 神经网络算法可评估网络空间态势感知情况，其削弱了人的因素并提高了结果的客观性和权威性，能够非常有效地处理非线性问题。实验结果表明，改进的 BP 神经网络可以准确地评价网络空间态势感知系统。

结束语　本文利用改进的 BP 神经网络算法定量地评估网络空间态势感知水平，首先分析网络空间态势感知评估的需求，接着介绍了传统 BP 神经网络的不足之处，并利用退火优化法进行改进，然后阐述改进 BP 神经网络算法的流程，最后基于虚拟 HoneyNet 模拟网络环境，在 Matlab 软件进行仿真测试，实验结果验证了本文方法的合理性。

参 考 文 献

[1] 罗守山. 入侵检测[M]. 北京：北京邮电大学出版社，2004：82-87.
[2] 于德江. 灰色系统建模方法的探讨[J]. 系统工程，1991，9(5)：9-12.
[3] 谢丽霞，王亚超，于巾博. 基于神经网络的网络安全态势感知[J]. 清华大学学报，2013，53(12)：1750-1760.
[4] 梁颖，王慧强，赖积保. 一种基于粗糙集理论的网络安全态势感知方法[J]. 计算机科学，2007，34(8)：95-97.
[5] 张云涛，龚玲. 数据挖掘原理与技术[M]. 北京：电子工业出版社，2004：1-57.
[6] 孙德衡. 基于指标融合的网络安全态势评估模型研究[D]. 西安：西北大学，2012.
[7] 郑皆亮. 基于灰色理论的网络信息安全评估模型研究[D]. 南京：南京信息工程大学，2005.
[8] 飞思科技. 神经网络理论和 matlab7 实现[M]. 北京：电子工业出版社，2006：15-30.
[9] 林蔚天. 改进的粒子群优化算法研究及其若干应用[D]. 上海：华东理工大学，2012.