设计研究与应用

基于 HTML 结构的水印方法研究

杨方兴[1] 卢秋如[2]
（1.广州汇智通信技术有限公司南京分公司，江苏南京 210019；
2. 江苏省电子信息产品质量监督检验研究院（江苏省信息安全测评中心），江苏无锡 214073）

摘 要：网页水印是网页版权保护中的一种重要方式，现有的网页水印技术大多数是基于网页代码中的格式变换达到隐藏水印信息的目的，水印信息的隐蔽性差，抗攻击能力不强。提出一种新的基于 HTML 结构的网页水印嵌入方法，预处理后的水印信息与定义好的 HTML 结构映射，实现水印的嵌入。对网页代码的格式和内容不做任何修改，具有良好的隐蔽性，鲁棒性强，同时有较高的水印容量。

关键词：网页水印；版权保护；预处理；映射
中图分类号：TP309.2 **文献标识码：**A **DOI：**10.3969/j.issn.1003-6970.2022.09.040
本文著录格式：杨方兴,卢秋如.基于HTML结构的水印方法研究[J].软件,2022,43(09):139-141+149

Research on Watermarking Method Based on Div Structure of HTML

YANG Fangxing[1], LU Qiuru[2]
(1.Guangzhou Huizhi Communication Technology Co., Ltd. Nanjing Branch, Nanjing Jiangsu 210019;
2.Jiangsu Electronic Information Product Quality Supervision and Inspection Research Institute (Jiangsu Information Security Evaluation Center), Wuxi Jiangsu 214073)

【Abstract】：Web page watermarking is an important way to protect web page copyright. Most of the existing web page watermarking technologies are based on the format transformation in web page code to embed the watermark information. The concealment of watermark information is poor and the anti-attack ability is not strong. A new web page watermark embedding method based on HTML structure is proposed. The preprocessed watermark information is mapped with the defined HTML structure of web page to achieve the watermark embedding. The format and content of the web page code are not modified, which has good concealment, strong robustness and good watermark capacity.

【Key words】：web page watermarking;copyright protection;preprocess;mapping

0 引言

随着互联网的普及发展，网页作为互联网的信息载体，逐渐被广泛应用。互联网信息以明文形式在互联网中传输，网页中重要信息被非法复制、分发、篡改和认证，知识产权保护变得非常困难。网页水印是通过某种规则把水印信息隐藏在网页中，隐蔽性较强，在网页遭到非法复制时可以提取水印验证网页版权归属。因此，研究网页水印对网页版权保护具有重要意义 [1]。

1 网页水印研究现状

网页水印技术研究较少，网页中冗余度少，常见的图像水印技术和音频视频水印技术不适用于网页水印，因此在网页中嵌入水印信息有较大的难度 [2]。目前已知的网页水印技术存在明显的不足。改变大小写 [3] 或者空格数 [4] 的方法都会改变 HTML 代码的内容，很容易被察觉，水印的隐蔽性差，简单的格式变换或者大小写转换就可以去除水印信息，鲁棒性较差；定义不存在的标签 [5] 容易被识别，隐蔽性和抗攻击能力不好；使用不同的代码格式 [6] 虽然隐蔽性较好，但是水印容量小，缺乏足够的水印嵌入点。

为了解决上述问题，本文提出一种新的基于 HTML 结构的网页水印嵌入方法，预处理后的水印信息与定义好的 HTML 结构映射，实现水印的嵌入。网页代码的

作者简介：杨方兴（1990—），男，山东济宁人，硕士研究生，工程师，从事信息系统安全方面的研究工作；卢秋如（1990—），女，江苏徐州人，硕士研究生，工程师，从事信息安全测评方面的研究工作。

格式和内容不做任何修改，具有良好的隐蔽性，鲁棒性强，同时有较高的水印容量 [7]。

2 方法的描述

2.1 HTML 结构

HTML 的结构包括头部（Head）、主体（Body）及多种属性标签、样式标签组成。Head、Body 等部分标签是固定不变的，不适合作为水印嵌入点。而像 、 与 class、style 等属性组合变化较多，因此，我们考虑提取属性与标签的结构组合，通过一定的映射规则将水印信息隐藏在属性和标签的变化结构中 [8]。

2.2 水印信息预处理

在水印算法中，如果直接嵌入原始的中文水印信息很容易被识别，传统的水印技术将中文转换为由 0 和 1 表示的二进制字符串。为进一步提高水印容量，本文采用四元 Huffman 编码对水印信息压缩，比如“版权保护”四个中文汉字，其二进制有 60 位，压缩后的编码缩短为 42 位，有效提高了水印容量。

2.3 HTML 结构位置提取

HTML 标签 div、span 和属性 class、style 的搭配使用频率非常高，因此考虑提取四种组合结构的位置信息分别对应压缩后的四元 Huffman 编码，如表 1 所示。

表 1 码元映射表

Tab.1 Symbol mapping table

码元	映射结构
0	
1	
2	
3	

以如图 1 所示为例，讲解 HTML 结构的位置信息提取过程。

```
<body>
<div class='s-skin-container s-isindex-wrap' ></div>----1层
<span class="c-font-nornmal">查看天气信息</span>
<div id="head"class="">
    <div id="s_top_wrap"class="s-top-wrap s-isindex-wrap">----2层
        <div class="s-top-nav"><ldiv>----3层
        <div class=" s-center-box"></div>
    </div>
    <span id="show-city"><l span>
</div>
</body>
```

图 1 HTML 结构示例

Fig.1 HTML structure

图 1HTML 结构示例中展示了标签的层级结构，以提取 结构为例，最外层定义为第 1 层，依次往下为第 2 层等，将第 1 层所有的 DIV 取出，记录 class 是当前第 i 个 DIV 中第 j 个属性， 为第 1 层结构中第 1 个 div，div 中第 1 个属性为 class，位置信息记录为 D111， 为第 1 层结构中第 1 个 span，位置信息记录为 S111。重复上述过程，取出 HTML 中所有的四种结构的位置信息，分别存入四个集合，用类似表达式（1）来记录位置信息，其中 C_n(n=0, 1, 2, 3) 表示码元 n 对应的位置信息集合，D 表示每个标签属性的位置 [9]。

$$C_0=\{D_{111},\ldots,D_{1ij}\} \quad (1)$$

3 实现步骤

3.1 水印的嵌入

通过水印信息和 HTML 结构的位置信息映射完成水印的嵌入。首先提取四种 HTML 映射结构的位置信息，选取的四种位置信息存储在四个集合 C_0、C_1、C_2、

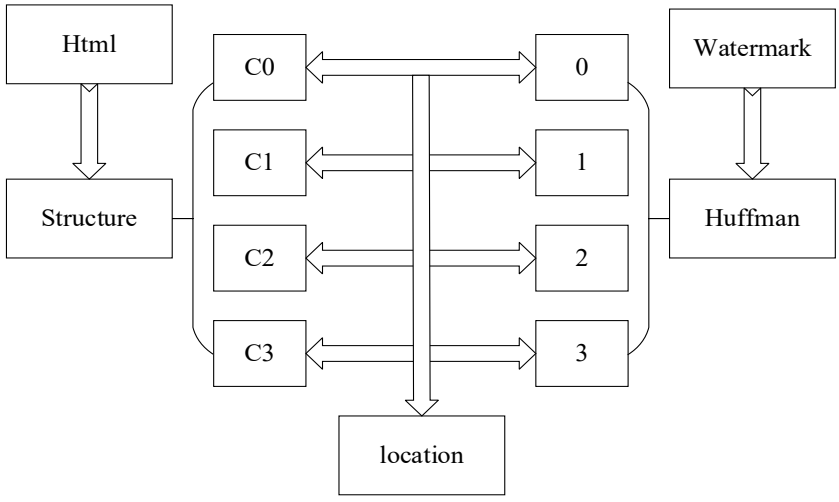

图 2 水印的嵌入

Fig.2 Watermark embedding

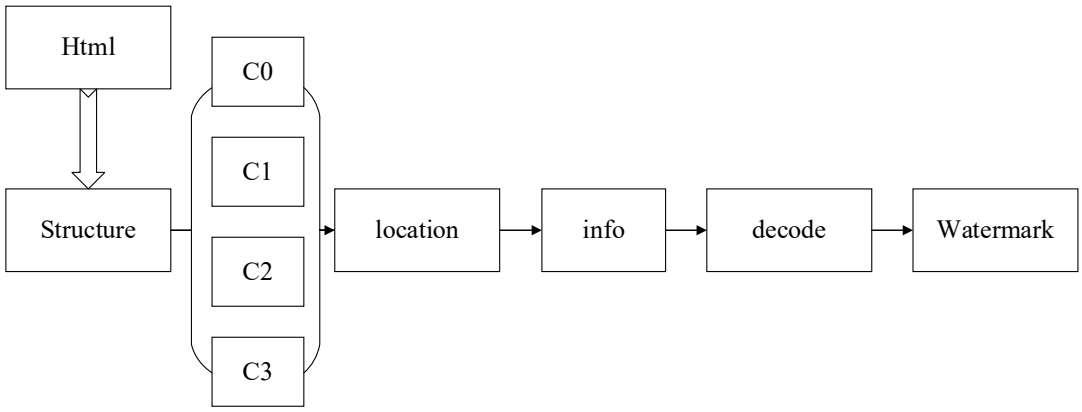

图 3 水印的提取

Fig.3 Watermark extraction

C_3 中。然后对水印信息进行 Huffman 压缩，水印信息预处理后得到由 0、1、2、3 构成的信息串。最后将得到的四种码元 0、1、2、3 与四种映射结构类型一一对应。逐个取出水印信息串中的码元，如取出的码元为 0，则将其与集合 C_0 中的一个位置信息相映射；如取出的码元为 1，则将其与集合 C_1 中的一个位置信息相映射；以此类推。将每个码元映射的位置信息读出，依次存放于一文件中。重复上述过程，将水印信息串中的所有码元均映射为四个集合中的一个位置信息。映射完成即实现了水印的嵌入，最终得到一个存储了四种 HTML 结构位置信息的文件。为了增加水印映射位置的全局分布性，四种 HTML 结构集合中的位置信息采用随机选取的方式。如图 2 所示直观展示了水印信息的映射过程。

3.2 水印的提取

提取水印是根据嵌入水印时得到的位置信息，找到 HTML 中对应位置的结构信息，根据四种映射结构与水印信息码元（0、1、2、3）的对应关系，反向映射（或称译码）得到水印信息的过程。如果结构为 ，则译码为 0；如果结构为 ，则译码为 1；如果结构为 ，则译码为 2；如果结构为 ，则译码为 3。直到所有的位置信息全部反向映射完毕，即得到嵌入的水印信息串，过程如图 3 所示，最后进行 Huffman 解码，得到原始的水印信息。

4 算法的性能

上述水印方法对 HTML 内容不做任何修改，对网页信息的样式不会产生任何影响。水印信息隐藏在 HTML 结构中，没有任何痕迹，不易被察觉，具有很好的隐蔽性，增强了算法的鲁棒性。对于大小写转换、空格变化的等格式上的攻击，不会对水印的正常提取产生影响。水印信息隐藏在 HTML 的结构中，插入、删除等非法操作只要不涉及针对样式的改动，也不会对水印的完整提取产生影响 [10,11]。

5 结语

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。本文基于对 HTML 结构的研究，提出一种 HTML 结构映射嵌入水印的方法，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，进一步研究提高网页水印的鲁棒性。

参考文献

[1] 张鑫,闪永强.一种新型网页防篡改策略的研究与部署[J].河南师范大学学报(自然科学版),2011,39(5):157-160.

[2] 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.

[3] 万唯一.基于数字水印的网页防篡改技术研究[D].成都:西南交通大学,2012.

[4] ZHANG Z,PENG H,LONG X.A Fragile Watermarking Scheme Based on Hash Function for Web Pages[C]// International Conference on Network Computing & Information Security.IEEE Computer Society,2011:417-420.

[5] 陈韦旭,陈建平,文万志,等.基于空样式的网页水印方法[J].计算机科学,2018,45(S2):338-341.

[6] CHOU Y C,LIAO H C.A Webpage Data Hiding Method by Using Tag and CSS Attribute Setting[C]//2014 Tenth

•••••• 下转第149页

3.2 蚁群算法的改进

基于蚁群算法存在的问题，可以从信息素初始化、信息素更新规则、参数优化设置、路径选择算法、路径搜索策略等途径进行改进。文献 [15] 提出一种改进信息素二次更新局部优化蚁群算法（IPDULACO），实验证明，通过二次更新信息素、跳出局部最优解等措施，该方法能够在较少的迭代次数内获得更精确的解，从而具备更强的全局搜索能力和更快的收敛速度。

4 结语

路径规划广泛应用于日常生活生产中。本文以 TSP 问题为研究对象，讨论了蚁群算法的原理，并分别使用蚁周模型、蚁量模型、蚁密模型的定义给出了蚁群算法的实现。通过对蚁群算法的参数分析，找出特定场景下的最优参数组合。最后研究讨论了蚁群算法存在的问题及改进途径。实验结果表明：相对于蚁量模型和蚁密模型，蚁周模型具有更强的全局搜索能力；组合参数优化有利于蚁群算法的收敛；可以从信息素初始化、信息素更新规则、参数优化设置、路径搜索策略等途径改进蚁群算法。

参考文献

[1] 刘砚菊,杨青川,辜吟吟.蚁群算法在机器人路径规划中的应用研究[J].计算机科学,2008,35(5):263-265.

[2] 仪孝展.基于改进遗传算法的物流车辆路径规划方法研究与应用[D].西安:西安理工大学,2018.

[3] 伟伟,王伟,陈能成,等.一种利用改进A*算法的无人机航迹规划[J].武汉大学学报(信息科学版),2015,40(3):315-320.

[4] 尚兴宏.无线传感器网络若干关键技术的研究[D].南京:南京理工大学,2013.

[5] 黄辰,费继友,刘洋,等.基于动态反馈A*蚁群算法的平滑路径规划方法[J].农业机械学报,2017,48(4):34-40.

[6] 宋彬.结合粒子群算法和改进蚁群算法的机器人混合路径规划[D].徐州:中国矿业大学,2018.

[7] 荀燕琴,田竹梅,任国凤,等.基于遗传算法的智能扫地机器人路径规划研究[J].高师理科学刊,2020,40(3):56-60.

[8] 许亚.基于强化学习的移动机器人路径规划研究[D].济南:山东大学,2013.

[9] 王沛栋.改进蚁群算法及在路径规划问题的应用研究[D].青岛:中国海洋大学,2012.

[10] 崔志华,孙佑强,任叶青.改进蚁群算法在机器人路径规划中的应用[J].南昌工程学院学报,2020,39(1):15-19+24.

[11] 刘军.基于改进蚁群算法的移动机器人路径规划研究[D].郑州:郑州大学,2010.

[12] 王胜训.蚁群算法的改进及TSP仿真研究[D].西安:西安电子科技大学,2014.

[13] 张松灿,普杰信,司彦娜,等.蚁群算法在移动机器人路径规划中的应用综述[J].计算机工程与应用,2020,56(8):10-19.

[14] 吴华锋,陈信强,毛奇凰,等.基于自然选择策略的蚁群算法求解TSP问题[J].通信学报,2013,34(4):165-170.

[15] 许凯波,鲁海燕,程毕芸,等.求解TSP的改进信息素二次更新与局部优化蚁群算法[J].计算机应用,2017,37(6):1686-1691.

•••••• 上接第141页

Intertional Conference on Intelligent Information Hiding and Multi-media Signal Processing.Kitakyushu,2014:122-125.

[7] RAFAT K F,SHER M.Innocuous Communication Via HTML Hiding Data in Plain Sight[J].Arabian Journal for Science & Engineering,2014,39(2):783-798.

[8] 杜耀刚,薛飞.一种基于类名的大容量网页信息隐藏算法[J].密码学报,2017,4(1):29-37.

[9] 陈丽.基于XML文档的文本数字水印技术研究[D].北京:北京印刷学院,2015.

[10] 张玉梅,和红杰,陈帆.浏览器端定位篡改的网页脆弱水印算法[J].计算机研究与发展,2014,51(12):2604-2613.

[11] 曾凡涛.一种基于特殊字符串的网页防篡改方法[J].计算机安全,2012(2):40-41+44.