基于 Shamir 门限和 html 标签 id 的网页水印方法①

杨旭光, 唐文龙

(百色学院 数学与计算机信息工程系, 百色 533000)

摘 要: html 网页水印技术是信息隐藏技术的分支, 目前虽已提出了一些方法, 但相较其它水印技术, 仍存在着水印的嵌入困难、容量有限和鲁棒性不强. 针对于目前采用网页中单个标签或某个符号来表示单个水印位而使得嵌入容量有限的问题, 提出了利用网页标签 id 来表示水印的思路, 而网页标签 id 可用来表示多位水印值. 方法中, 首先是把网页中重要内容形成消息摘要并和表示版权的二值图像异或运算后, 作为水印信息, 然后经 Shamir 门限方案后分解, 把其作为网页标签 id 值的方法来分存嵌入网页. 经实验验证, 该方案有较好的嵌入容量、鲁棒性和隐蔽性.

关键词: Shamir 门限; 网页标签标识符; 网页水印; 消息摘要; 秘密共享

A Webpage Watermarking Method Based on the Shamir Threshold and Html Label Id

YANG Xu-Guang, TANG Wen-Long

(Department of Mathmatics and Computer, Baise University, Baise 533000, China)

Abstract: Html webpage watermarking technology is a information hiding technology branch, at present several methods have been proposed, but compared to other watermark technology, there still exist problems about watermark embedding difficulties, limited capacity and weak robustness. Aiming at the limited embedding capacity question by using the single label or a symbol in webpage to indicate that a single watermark bit, a method is proposed to use the id value of label in webpage to represent watermark, and the id value of label in webpage can be used to express many watermarking bits. First in this method is to xor with the message digest of the important content in webpage and bits of binary image to represent copyright, and then by the Shamir threshold decomposition scheme, the watermark as the id value of label in webpage is deposited the webpage. Proved by experiments, the scheme has better capacity of embedding, robustness and invisibility.

Key words: Shamir threshold; webpage label identification; webpage watermark; message digest; secret sharing

信息隐藏技术是计算机安全技术领域的一个分支, 该技术目前在图像、视频、音频等多媒体领域比较成熟, 已有大量的文献进行了相关技术的阐述. 不同于数字图像或者视频文件所拥有的大量的视觉冗余, 网页的文件结构是由普通文本文件加上各种标记(Tag)所构成, 这种文件结构不存在太多的冗余信息, 从而使得将信息隐藏到网页文件中比较困难.

目前, 国内外关于网页信息隐藏技术的研究主要是基于 HTML 语法的高容错性, 以及针对网页标签的信息隐藏技术, 典型方法有: (1)基于插入不可见字符; (2)修改标签属性值的大小写[1]; (3)调整标记名称和标记属性间的空格数来隐藏信息, 因为符号之间的多个空格会被当成一个空格对待,或在符号“>”的左边插入空格来隐藏信息, 符号“>”的一个或多个空格会被浏览器忽略, 或利用标签中属性赋值号“=”左右添加空格来隐藏信息[2,3]; (4)修改标签名称字符的大小写[4]; (5)利用有些标签的结束标签可省略来隐藏信息; (6)某些标签可有两种等价格式[4]. 上述的这些方法仍有一些各自的局限性, 且大多存在着水印嵌入的容量不足、鲁棒性较差等特点. 因此, 探讨改进和提出更多的网页

① 基金项目:百色学院院级项目(2010KB13)
收稿时间:2013-01-10;收到修改稿时间:2013-03-04

水印算法, 仍是网页水印技术领域积极发展的一个方向. 例如, 在文献[5]中赵启军和卢宏涛就提出了新的基于 PCA 的网页水印技术.

1 Shamir门限和html标签id的网页水印

1.1 网页标签 id 属性

网页标签是由尖括号包围的关键词, 如<html>. 网页标签的特点是: (a)通常是成对出现的, 如 和 ; (b)标签名称使用英文字母, 一般标签可含有属性定义. 如某网页中段落标记的标签: <input type="submit" value="Go" class="button" title="搜索！"/> , 该标签定义了段落内含有一个提交按钮, 同时拥有一些属性定义. (c)html 规范中, 标签和属性名称大小写, 及属性值使用单双引号没有区分. 而在标签外则是网页的主要信息表达成分, 因此, 这一部分也是网页安全重点关注的内容.

在本文的网页水印设计中, 对标签外的内容保护是使用了消息摘要的方法, 同时, 为了能够避免搜索和处理整个网页及满足可嵌入性, 水印的嵌入位置是考虑利用了大部分网页中常见的、可用于分组定义 html 块级元素的 标签位置. 网页标签一般可设置 id 属性, 该属性可被 CSS 样式表或客户端 JS 脚本引用. 而服务器端 ASP 等脚本一般是通过 name 属性引用. 也就是说, 如有用户想盗版网页且不被指证而篡改了网页标签 id, 那么, 他需知道该标签 id 在何处曾被引用过, 并也要做出相应的修改, 否则, 网页中关于该标签的显示和处理可能会不正常. 从某种意义上讲, 网页标签 id 一旦在网页中被定义, 由于修改有风险性而其自身则相对存在一定的安全性. 网页水印的嵌入一般是在已经设计好的网页中进行, 那么相反, 嵌入时, 可以利用那些已设计好的网页且存在还没有被指定 id 属性的标签. 由于没有指定 id , 对这些标签的识别和引用, 相关的样式表文件或脚本可能是通过其它方法来获取的, 如引用了标签的 name 名称属性. 因此, 如给这些标签添加 id, 并不会影响样式表或脚本对该标签的显示和处理等操作. 在本文的水印方法中, 是把水印值添加作为标签的新 id 属性值.

1.2 Shamir 门限

秘密共享是一种将秘密分割存储的密码技术, 目的是风险分散和容忍入侵、阻止过于集中秘密, 它是信息安全和数据保密中的一个重要手段. Shamir (t, n) 秘密共享[6]是由 SHAMIR 于 1979 年提出. 该算法思想是: 将秘密 S 分为 n 个子秘密, 任意 k 个子秘密都可以恢复出 S, 而任意 k-1 个子秘密无法恢复出 S. 秘密被分割的过程如下:

假设有秘密 S, 任取随机数 a_1 a_2, …, a_{k-1}. 令 a_0=S, 构造如下多项式:

$f(x)= a_0+a_1x+a_2x^2+\cdots+a_{k-1}x^{k-1}$, 其中所有的运算都在有限域 F 中进行.

任取 n 个数 x_1, ...,x_n 分别带入多项式得到 $f(x_1)$, ..., $f(x_n)$. 将$(x_1, f(x_1))$, ..., $(x_n, f(x_n))$分别存储.

秘密获取:

任取 k 个数据, 假设取$\{x_1, y_1\}$, ...,$\{x_k, y_k\}$, 代入并求解多项式系数. 用矩阵乘法可求得 a_0, a_1,...,a_{k-1}, 构造多项式 $f(x)= a_0 +a_1x+a_2x^2+...+a_{k-1}x^{k-1}$, 如将 x=0 代入多项式, 则可求得原秘密 S=a_0.

在下面小节中, 将讲述利用 Shamir 门限来完成网页水印嵌入与提取的方法.

1.3 水印嵌入和提取过程

水印嵌入首先进行预处理, 过程如下图 1 所示:

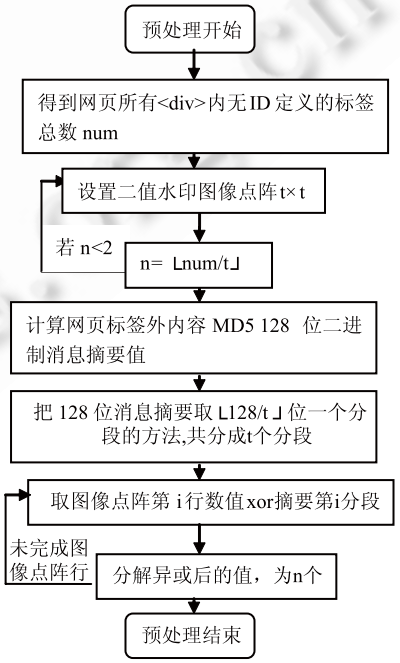

图 1 水印嵌入预处理过程

图 1 中要求选取的最低门限 n 值不小于 2, 同时假定水印图像点阵不超过 128×128. 其中, $\llcorner x \lrcorner$ 表示 x 数值的整数部分, 最后一个分段长度为: 128-(t-1)* $\llcorner x \lrcorner$. 而 t、n、及用户选取的门限 k 值和随机产生的子密钥 sk、 n

个数 $x_1, ...,x_n$ 将共同构成水印提取的密钥 Key. 为了能检测网页的重要内容是否被篡改, 水印值的定义是: 使用网页中标签外的内容进行 MD5 消息摘要算法[7]而得到 128 位二进制的摘要, 并把其与用来表示版权的二值水印图像信息异或运算后作为水印值 w; 采用异或运算只是做了简单的加密处理, 实际应用中可采用更复杂的密码算法.

水印具体嵌入步骤:

(1) 首先, 扫描网页, 把网页中的所有单引号改写为双引号, 同时提取出网页中所有标签的 id 属性值到集合 C 中. 然后, 统计网页中 标签, 计算出 标签内的其它标签总个数 num, 这些标签要求满足: id 属性无定义(html 有些标签本没有 id 属性, 应事先排除掉这些标签). 统计结束后, 使用该 num 值作为 Shamir 门限的分解水印总个数;

(2) 根据实验采取的二值图像信息, 计算 Shamir 门限 n 值. 如嵌入的水印二值图像是 32×32 点阵大小, 那么, 可取 n 等于 num 除以 32 来得到.

(3) 把 128 位消息摘要值, 依据二值图像点阵大小分解成 t 个小的数, 和二值图像对应位进行异或加密运算得到新的水印值. 比如, 如采用上述的 32×32 点阵二值图像, 那么可把 128 位消息摘要值每 4 位取一个小的数, 共分解成 32 个数, 把该 32 个数分别与二值图像的 32 行位图信息进行异或运算;

(4) 选择好定义的门限值 $k(k<n)$, 并构造多项式 $f(x)$, 取 x=1, 2, ... , n, 对每一个经异或运算后的位图信息, 计算得到 t 组 n 个分解的水印值 $w_i(i$=1, ... , $n)$;

(5) 每一组 n 个分解后水印值 $wi(i$=1, ... , $n)$, 反复执行(a)、(b):

(a) w_i 在集合 C 中, 可修改 w_i=w_i &"_", 以避免和网页中已有标签的 id 属性相同; (虽然网页中某些标签的 id 值可能会经网页脚本来动态设定, 并可能造成作为水印值的 id 值与这些动态修改的标签 id 值冲突, 但只要网站平台系统的构建者简单地调整脚本代码, 把代码中动态设置的标签 id 值改为非数字开头就可以避免冲突, 这是因为, 水印值设计都是数字构成的. 而标签 id 值的静态设定, 当水印嵌入时, 上述中已介绍了对冲突做出的调整.)

(b) 在该 标签内, 取出存在没有定义 id 的某个标签, 增加该标签属性, 令 id='wi'该属性值添加采用了单引号, 以方便水印提取的识别;

修改结束后, 水印即分存嵌入到网页中.

水印提取主要步骤:

(1) 扫描网页 标签, 找到该标签内含有的带有 id 属性的其它标签且 id 属性值使用单引号定义, 提取出该 id 值(若该 id 含有"_"字符, 应去掉该字符), 即得到了一个水印值;

(2) 继续选择网页后续内容, 反复执行(1), 直到找到满足最低门限值的 k 个水印值;

(3) 利用该 k 个水印值和矩阵计算, 构造出多项式的 k 个系数, 其中第一个系数即是分解前的水印值;

(4) 反复执行(1)、(2)、(3), 直到得到 t 个分解前的水印值;

(5) 再次得到标签外的网页主要内容, 通过 MD5 函数计算 128 位消息摘要值, 按 t 个分段该 128 位二进制, 并和得到的分解前的 t 个水印值进行异或运算, 来还原出嵌入的二值水印图像信息.

为了验证上述方案, 经实践编程设计了一个水印系统原型, 原型主界面如图 2 所示.

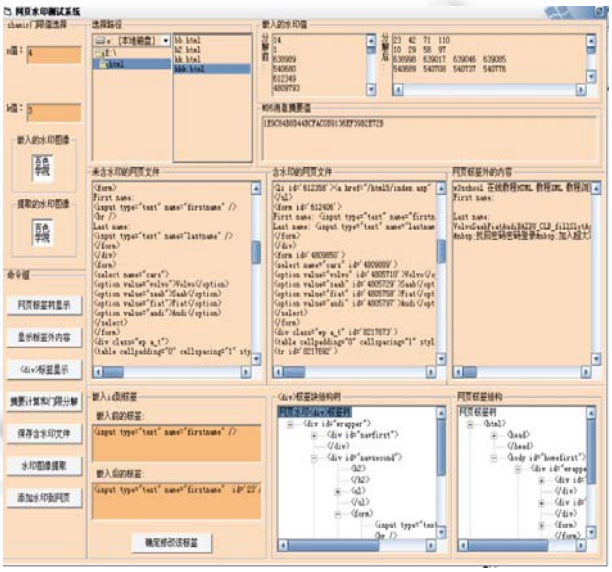

图 2　网页水印系统原型

该系统在接收网页文件后, 首先分析网页的标签结构, 标签结构的分析没有使用网页解析 htmlparser 工具, 是利用了递归匹配扫描方法完成了网页 标签的提取, 并采用树型控件显示出标签结构. 水印 Shamir 门限分解时需要提供 n、k、随机数 a_2, …, a_{k-1} 及 $x_1, ...,x_n$, 其中, 随机数 a_2, …,a_{k-1} 及 $x_1, ...,x_n$ 是定义了一个伪随机数序列发生器来产生(门限计算中的 a_0、

a_1 值实际定义为: a_0=异或运算后的值 mod 子密钥 *sk*, a_1=异或运算后的值\子密钥 *sk*; 由 *k* 值确定产生 a_i 随机数的个数, *n* 值确定产生 x_j 随机数的个数), *k* 值由用户选择, 但要小于等于 *n*, 而 *t* 值、*n* 值是按照上述图 1 中的水印预处理后得到.

系统中, 消息摘要的计算是单独定义了一个类模块来实现. 水印嵌入时, 简化了嵌入位置的选择, 并没有置乱嵌入, 而是依序嵌入在满足嵌入条件的位置处, 实际操作时, 可由用户手工逐个嵌入 id 值或由系统自动逐个添加 id 值到网页中. 提取水印时, 定义提取密钥 Key={*t*,*n*,*k*,*sk*,x_1, ...,x_n}, 密钥输入正确后, 通过门限矩阵计算求得水印 w i 值, 该过程需要反复执行 *t* 次. 图 2 中提取的二值图像是在网页没有遭受到篡改的情况下得到的.

2 实验结果分析

为了检验篡改网页后, 二值图像水印的提取情况, 同时进行了仿真实验攻击(*t*=3, *n*=4), 并整理得到了如图 3 所示的实验结果:

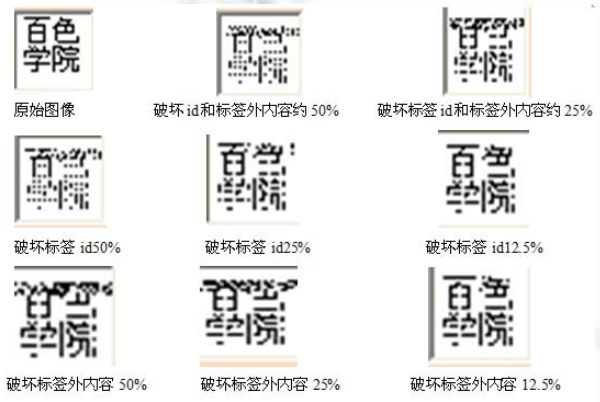

图3 篡改网页标签和标签外信息水印图像的提取效果

图 3 中显示了不同攻击程度下的, 水印图像的提取结果, 当网页内标签 id 和标签外信息被破坏 50%, 图像仍然能够依稀辨别, 而单独篡改标签 id 或标签外网页信息, 图像分辨效果较好.

对网页系统的攻击类型一般有以下几种[5]: 网页内容被篡改, (1) 但没有生成新的水印. (2) 水印被破坏. (3) 增加伪水印. 对于第一、二种情况, 篡改后提取的水印值将可能会不一致, 可以用来检测是否篡改. 对于第三种情况, 由嵌入方法生成的水印可进一步进行置乱处理再嵌入, 置乱后的水印可能性取值有 $2^{h\times h}$, 其中 *h* 为图像一行像素个数, 如对于 *h* 等于 16, 则可能性取值将为 2^{256} 种. 又因提取需密钥 Key, 因此, 篡改者为了增加伪水印而剔除真实水印, 想由此先断定出真实水印的可能性非常小. 另外, 采用门限秘密共享, 也提高了真实水印提取的成功率. 如果主要是考虑提高水印鲁棒性, 可以不与标签外消息摘要进行异或运算, 只是对二值水印图像进行置乱处理后即分解嵌入.

网页在嵌入水印后, 网页大小可能也会发生变化, 前述中某些利用标签属性值的单双引号或大小写、及等价标记排列等网页水印方法, 可以不改变网页大小, 但嵌入量也相对较少. 而利用空格表示水印, 虽易被用户剔除掉, 但嵌入量可较多, 相应文件大小也会增加较多. 在本文方法中, 为了评估因嵌入水印而对文件大小的影响, 基于不同的 Shamir 门限 n 值, 实验统计了水印嵌入前后文件大小的变化如表 1 所示:

表 1 不同 Shamir 门限 *n* 值, 水印嵌入前后文件大小对比(水印: 32×32 像素图像)

n 值	嵌入前网页大小	嵌入后网页大小	增加比率
n=3	13.7kb	14.825 kb	8.21%
n=4	13.7kb	15kb	9.5%
n=6	21.1kb	22.97kb	8.9%
n=8	24.5kb	27.4kb	11.84%

表 1 中, 取不同 Shamir 门限 *n* 值时, 水印分解数目也会变化, 实验采用的网页大小也会变化, 文件大小的增加比率相应变化并不大, *n* 值取较大时, 则适宜在较大网页文件中进行.

文献[8]中, 采用了不可见字符来编码水印值, 一共定义了 16 种组合, 平均 5.18 个字符编码为 4 位二进制, 如在 32×32 像素的二值图像同等条件下, 需要的字符数目约为 1326 个, 采用 ASCII 码字符编码, 空间大小约为 1.295kb. 从表 2 中比较可知, 当门限 *n* 值取 3 时, 空间所需相对较小, 与表 1 数据相比, 当门限 *n* 值取 4 时, 空间所需大小相当.

表 2 水印嵌入所需空间大小对比

像素位数值 (像素)	不可见字符编码水印方法	本文方法 (门限 *n* 值=3)
16×16	0.3237kb	0.4208 kb
32×32	1.295kb	1.125kb
64×64	5.18kb	4.5kb
96×96	11.655kb	9.563kb

在信息隐藏技术的水印方法分类中, 按可见性, 可分为不可见和可见水印两类, 其中不可见水印强调的是隐藏特性. 在本文方法中, 采用了网页标签常见的 id, 并且和其它标签正常的 id 相混合, 识别是采用了不易引起视觉敏感的单双引号. 其实, 在方法中, 没有把单双引号用于水印表示, 只是识别标记. 因此, 像在属性值前后添加空格或取值等于符号前后添加空格, 也可用于识别的标记, 这也在一定程度上增加了水印方法的透明性.

3 结论

随着互联网的普及应用, 作为互联网常见的信息传递媒体之一-网页, 关于它的安全保护技术也变得越来越重要. 对于网页水印技术, 因网页本身的格式特性, 相较其它水印技术, 有其自身的弱点, 尝试采用不同方法是网页水印技术将来积极发展的方向. 在本文中, 使用 Shamir 门限秘密共享, 提高了嵌入水印的安全和可靠性, 并利用了网页文件标签的 id 值来嵌入水印, 嵌入容量可较大, 但可能存在因门限 n 值过大, 而使得在较小的单个网页的适合点处水印完全嵌入较困难. 未来设想是改进扫描算法, 找出与网页存在关联链接的其它网页并用来水印嵌入. 另外, 也将进一步考虑如何利用结合网页标签的 name 名称属性来表示水印的方法.

参考文献

1 丁伟.基于 Web 网页的文本水印技术的研究[学位论文].武汉:武汉理工大学,2012.

2 傅瑜,王保保.文本水印附加空格编码方法的实现及其性能.长安大学学报,2002,22(5):85–87.

3 Low SH, Maxemchuk NF, Lapone AM. Document Identification for Copyright Protection using Centroid Detection. IEEE Trans Communications,1998,46(3):372–383.

4 万唯一.基于数字水印的网页防篡改技术研究[学位论文].成都:西南交通大学,2010.

5 Zhao QJ, Lu HT. A PCA-based watermarking scheme for tamper-proof of web pages.Pattern recognition,2005,38(8):1321–1323.

6 雷红艳,邹汉斌.基于 Shamir 秘密共享的隐私保护分类算法.计算机工程与设计,2010,31(6):1271–1273.

7 Rivest R. The Md5 Message-Digest Algorithm.RFC. RFC Editor, 1992.

8 孙鹏.网页水印技术研究[学位论文].上海:上海交通大学,2010.

(上接第 97 页)

of 2010 IEEE International Symposium on Circuits and Systems (ISCAS),.2010.Paris: IEEE.

5 Hsu CC, Hung TY, Lin CW, et al. Video forgery detection using correlation of noise residue.2008:IEEE.

6 王俊文,刘光杰,张湛,等.基于模式噪声的数字视频篡改取证.东南大学学报(自然科学版),2008,38(A02):13–17.

7 黄添强,吴铁浩,袁秀娟,等.利用模式噪声聚类分析的视频非同源篡改检测.计算机科学与探索,2011,5(10):914–920.

8 韩家炜,坎伯.数据挖掘:概念与技术.北京:机械工业出版社,2001

9 周志华,王珏.机器学习及其应用.北京:清华大学出版社,2007.

10 赵倩,尚学群,王淼.基于 seeds 集和频繁项集挖掘的半监督聚类算法.计算机工程与应用,2010,46(8):123–126.

11 赵凤,焦李成,刘汉强,等.半监督谱聚类特征向量选择算法.模式识别与人工智能,2011,24(1):48–56.

12 王玲,薄列峰,焦李成.密度敏感的半监督谱聚类.软件学报,2007,18(10):2412–2422.

13 肖宇,于剑.基于近邻传播算法的半监督聚类.软件学报,2008,19(11):2803–2813.

14 宋凌,李枚毅,李孝源.一种新的半监督入侵检测算法.计算机应用,2008,28(7):1781–1783.

15 武伟,詹玲超,基于数码相机固有特性的篡改检测.微计算机信息,2007,23(28).

16 黄添强,秦小麟,叶飞跃.基于方形邻域的离群点查找新方法.控制与决策,2006,21(5):123–126.