

复杂网页表格的信息隐藏算法设计

秦彩云, 张晓明, 赵国庆

(北京石油化工学院 信息工程学院, 北京 102617)

摘 要: 利用网页来隐秘传递信息, 具有重要的研究意义和实用性. 提出了双比特的隐藏方法, 使网页表格的隐藏容量比已有方法增大了一倍. 同时, 针对网页存在多个表格结构的特点, 对平行表和嵌套表的进出特性进行分析, 设计了全面的多表格隐藏算法和进出管理方法, 详细分析其变化路线. 本算法可以用于各种规则的多表结构和单表结构, 多表结构适于任意平行表和嵌套表的组合. 实验结果表明, 基于表格属性的网页隐藏方法, 具有很好的实用性、隐蔽性和透明性.

关键词: 网页; 表格属性; 信息隐藏; 嵌套结构; 算法

中图分类号: TP393.4

文献标识码: A

文章编号: 1000-7180(2011)08-0132-04

Design of Information Hiding Algorithm for Complicated Webpage Tables

QIN Cai-yun, ZHANG Xiao-ming, ZHAO Guo-qing

(College of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing 102617, China)

Abstract: It is very important and practical to transform messages securely through the web pages. A kind of two-bit hiding approach is proposed to double the hiding capacity of webpage table than the existing method. Meanwhile, considering the status of multiple table structure in a webpage, the features of parallel tables and embedding tables are analyzed. The algorithms of multiple table hiding and table in-out management are designed in detail to obtain the change route of table operating states. So, the algorithms can be available for all kinds of webpage structures of multiple tables and single table. The multiple tables can be composed with parallel table and embedding table. Experimental results show that the proposed approach is practical, secure and transparent.

Key words: Webpage; table attribute; information hiding; embedding structure; algorithm

1 引言

网站信息的传播具有实时永远在线的特点, 若将其中的网页作为载体进行信息隐藏, 虽然容量较小, 但是具有隐蔽性和快捷性. 因此, 如何利用网页安全传递重要信息已成为网络安全通信的重要研究内容.

网页上的图象、音频等多媒体都能作为载体用于嵌入水印或秘密信息^[1], 而网页文本中隐藏信息的难度更大. 近年来, 人们提出了许多网页信息隐藏方法, 多数是利用 HTML 语法特点和特殊标记来完成, 主要有基于不可见字符的方法、基于标记中字

母大小写变化的方法、基于属性对顺序的方法等^[2-8]. 常规的不可见字符方法是在行末加入空格或制表符, 其隐藏容量不受限制, 但是隐藏信息集中时容易被发现, 也增大了文件. 标记大小写的方法是利用 HTML 规范对标记大小无关的特性隐藏信息. 虽然浏览效果能达到要求, 但是通过分析源文件, 很容易发现这种大小写不规则的现象. 基于属性对顺序方法是利用顺序不同的属性对的标记之间是等价的这一性质嵌入信息, 该方法不改变文件的大小, 隐蔽性好, 但其提取时需要数据库的支持. 文献[2]提出了一种基于网页表格属性的新方法, 具有良好的隐蔽性和安全性. 但其嵌入容量小, 而且只能用于

具有单一表格的网页。

基于文献[2]的工作,本文提出了一种双比特隐藏方法,使容量增大了一倍;同时,提出了适应多个表格结构的隐藏算法,能够处理复杂嵌套等表格情况,取得了满意的效果。

2 双比特隐藏方法设计

在 HTML 规范中, <table>、<tr>、<td> 这三个标记是定义表格的最重要的标记。它们当中有一些参数设置也就是设置单元格的属性,其中 <td> 的参数设定中 width 表示宽度, height 表示高度。当不同 <table></table> 中的 width 属性值或百分比相差仅为 1 时,从视觉角度来看没有差异。文献[2]基于表格单元行列距属性值的奇偶性,提出了以下单比特的隐藏方法,总体隐藏规则设计为:

对于表格任一行的某相邻两列,若其宽度值的奇偶性为奇数对奇数(简称为奇奇,下同理),表示隐藏比特 1;若为偶偶,表示隐藏比特 0。

同理,对于表格任一列的某相邻两行,若其高度值的奇偶性为奇奇,表示隐藏比特 1;若为偶偶,表示隐藏比特 0。

2.1 双比特隐藏思路

由于奇偶性的状态表现为 4 种:奇奇、奇偶、偶奇和偶偶,而以上规则只使用了其中 2 种,因此还可以采用。进一步形成以下双比特隐藏规则,表示如下:

偶偶:表示嵌入 00

偶奇:表示嵌入 01

奇偶:表示嵌入 10

奇奇:表示嵌入 11

于是,每一种状态都能表示 2 个比特,这比原有方法提高了一倍的隐藏容量。

如果不符合以上规则,则必须调整,使奇偶性达到要求。调整算法如表 1 所示。

表 1 双比特隐藏规则的调整算法

嵌入	偶偶	偶奇	奇偶	奇奇
00	0, 0	0, +1	+1, 0	+1, +1
01	0, +1	0, 0	+1, +1	+1, 0
10	+1, 0	+1, +1	0, 0	0, +1
11	+1, +1	+1, 0	0, +1	0, 0

例如,待嵌入比特为 00 时,如果奇偶性为偶偶,则不必修改;否则,需要将其中的奇数值加 1,使其成为偶数。虽然变化量最大者是对奇奇的调整,有两

行或两列都需要加 1,但实际上,只相当于表格线往右移动了 1 个单位。

2.2 双比特隐藏算法设计

输入: 网页 H ; 待隐藏的秘密信息 W , 长度为 N 。

输出: 嵌入信息后的网页 H' 。

步骤:

(1) 取网页 H , 同一表格中每一 <tr></tr> 中的 <td> 的 height 属性值相同;

找到 <tr></tr> 中第一个 <td> 的 height 值并赋给这一 <tr></tr> 的其它 <td> 的 height;

(2) 同一表格中不同 <tr></tr> 中的 <td> 的 width 组值相同且取决于第一个 <tr></tr> 中 <td> 的 width 组; 找到第一个 <tr></tr> 中 <td> 的 width 组值, 将这一集合值赋给其他 <tr></tr> 中 <td> 的 width 组。

(3) 取表格 T_i ;

(4) 计算 T_i 的行数和列数分别为 m 和 n ,

若满足 $((m \geq 1) \&\& (n \geq 2))$ 或 $((m \geq 2) \&\& (n \geq 1))$, 则转(5), 否则 $i = i + 1$ 返回(3) 取下一个表格;

(5) 按行遍历表格 T_i 每一单元格, 遍历当前行为 R_j ;

(6) 遍历行 R_j 的当前列为 c_k ;

(7) 若 $(k+1 < n)$, 从 W 中顺序取出 2 个比特。参照式(3) 计算结果, 按照表 1 的双比特隐藏规则, 调整单元格的宽度并保存。

调整后, 令 $N = N - 2$ 且 $k = k + 2$ 。返回(6);

若 $j = m$ 且行 R_m 中的列全部遍历完成, 此时信息横向嵌入完成, 并且每个单元格 height 值保存完成, 则转(8)。否则 $j = j + 1$ 返回(5) 遍历表格 T_i 的下一行;

(8) 按列遍历表格 T_i 的每一单元格, 遍历当前列为 c_k ;

(9) 遍历列 c_k 的当前列为 R_j ; 从 W 中顺序取出 2 个比特。参照式(4) 计算结果, 按照表 1 的双比特隐藏规则, 调整单元格的高度并保存。

调整后, 令 $N = N - 2$, 且 $j = j + 2$ 。返回(9)。

若 $k = n$ 且 $i < p$, 则 $i = i + 1$ 返回(3) 取下一个表格, 重复以上步骤;

若 $i = p$, 转到(10); 否则, $k = k + 1$ 返回(8);

(10) 此时 $N = 0$, 嵌入完成。转到(11)。

(11) 输出 H' 。

提取算法与嵌入算法类似,不再赘述.

3 多表格结构的隐藏方法设计

以下研究多个表格结构的隐藏方法,包括简单和复杂的组合情况.

3.1 多表格结构的隐藏思路

对具体使用的表格结构,采用双比特隐藏方法,分别利用表格行单元格方向的宽度属性和列单元格方向的高度属性进行嵌入处理.

许多网页都存在着各种表格的组合,例如图 1 示例的是平行表和嵌套表的复杂组合. 共有 6 个表格,平行表有 T1 和 T2、T4 和 T5,嵌套表结构是 T1 包含 T2、T3 包含 T4、T5 和 T6、T5 包含 T6. 图 1 的箭头线指向了各表的入口顶点.

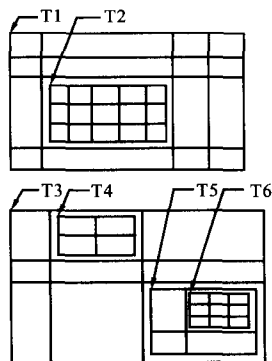


图 1 多表格结构示例

3.2 多表格进出算法设计

本算法需要解决规则表的任意组合,包括平行结构和嵌套结构,甚至是多层嵌套结构. 难点是解决表格的进出处理. 解决方案是:

(1) 设计 1 个表格状态数组 $tMark[]$, 每当进入表时, 设置为 1; 每当退出时设置为 0. 这样, 在每次循环中, 都可以检查出具体表格的使用状态.

(2) 设计 1 个表格序号 k , 表示当前所处的表格的序号.

(3) 用 (T_a, T_b) 表示当前表格的进出次数, 其中, T_a 表示多表格结构的进入次数, T_b 表示多表格结构的退出次数.

于是, 当前表格的判断规则是:

(1) 每当进入某表格时, $T_a = T_a + 1$, 且设置表序号 k 等于已进过表格的总数: $k = T_a$.

(2) 每当退出某表格时, 需要查找最近使用的表格, 其位置必定是 1. 说明当前正处于该表格的某单元格中, 令 k 等于该表格的序号.

因此, 对照图 1 的多表格结构, 可以得到其状态变化路线, 如图 2 所示.

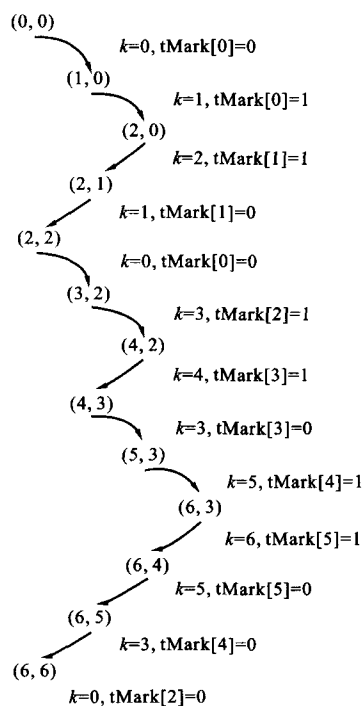


图 2 多表格结构的状态变化路线

4 实验与分析

基于规则的单表、平行表和嵌套表, 开展实验设计. 采用 C# 语言编写程序, 通过任意构造的综合性的多表格结构网页, 来进行重要信息嵌入和提取.

原始网页如图 3 所示, 其中, 包含了 2 个第 1 层次的平行表、2 个第 2 层次的平行表. 图中上方是 2 个嵌套表, 下方也是相互嵌套的表格.

Table attribute-based Web Page Watermarking			
a1	a2	a3	a4
b1	b2	b3	b4
c1	x1	x2	x3
	y1	y2	y3
	z1	z2	z3
d1	d2	d3	d4
w1	u1	u2	w3
	v1	v2	
p1	p2	p3	
s1	s2	j1	j2
		k1	k2

图 3 原始网页

信息嵌入界面分别如图 4 所示, 表明其隐藏容量为 82 个比特. 待隐藏信息为占有 13 个字符. 由于每个字符占用 7 位 ASCII 码, 则该字符串需要 $13 \times 7 = 91$ 个比特, 因此实际有 9 个比特没有嵌入. 经过提取后, 结果与实际嵌入的完全相同.

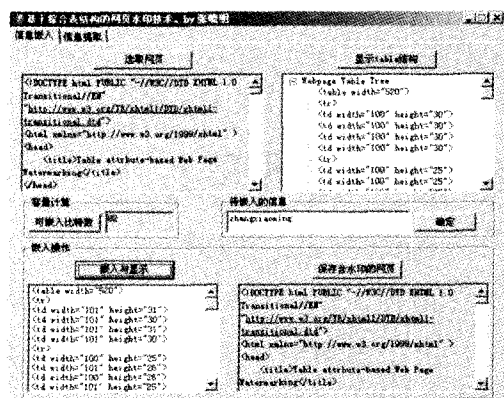


图4 多表格结构的信息嵌入界面

5 结束语

基于网页表格属性,提出了双比特的隐藏规则和算法实现,使嵌入容量增大了一倍。提出的多表格进出管理方法,解决了多个表格结构的网页信息隐藏困难,能够适应各种表格的组合和嵌套情况,使算法具有很好的实用性,适应面广。

参考文献:

- [1] 李文治, 张晓明, 殷雄. 一种基于能量关系的鲁棒音频扩频水印算法[J]. 微电子学与计算机, 2009, 26(8): 144—147.
- [2] 张晓彦, 张晓明. 一种基于表格属性的网页信息隐藏算

法[J]. 北京石油化工学院学报, 2009, 17(1): 43—47.

- [3] 沈勇. 一种基于 HTML 文档的信息隐藏方案[J]. 武汉大学学报: 自然版, 2004, 50(s1): 217—220.
- [4] Mohamed Lahcen BenSaad, Sun Xingming. Techniques with statistics for web page watermarking [J]. Proceedings of World Academy of Science, Engineering and Technology, 2005, 6(7): 300—303.
- [5] 孙星明, 黄华军, 王保卫, 等. 一种基于等价标记的网页信息隐藏算法[J]. 计算机研究与发展, 2007, 44(5): 756—760.
- [6] 龙银香. 基于 HTML 标签的信息隐藏模型[J]. 计算机应用研究, 2007, 24(5): 137—140.
- [7] Qijun Zhao, Hongtao Lu. PCA-based web page watermarking [J]. Pattern Recognition, 2007, 40: 1334—1341.
- [8] 李建国, 马小虎, 沈晓峰. 一种基于重复标记属性的多网页信息隐藏方法[J]. 计算机应用与软件, 2009, 26(8): 62—63.

作者简介:

秦彩云 女, (1963—), 研究生, 实验师. 研究方向为网络信息安全方向.

张晓明 男, (1968—), 博士, 教授. 研究方向为网络系统与网络安全方向.

赵国庆 男, (1965—), 硕士, 副教授. 研究方向为信息安全.

(上接第 131 页)

式(6)通过计算所有通路的相似度值,并选出最大值作为两个概念的相似度值,来解决问题 2. 因为所有通路都考虑了,因此避免了问题 2 中所描述的情况.

4 结束语

本文首先分析常用的几种基于路径的相似度计算算法,然后提出了一种改进的基于路径的概念相似度计算算法,通过对 Wu 和 Palmer 算法进行改进,弥补了文中提出的两个问题. 这种改进较为直观,容易实现,而且算法复杂度和 Wu 和 Palmer 算法类似,没有增加多余的时间开销.

参考文献:

- [1] George A Miller, Richard Beckwith, Christiane Fellbaum, et al. Introduction to wordNet: an on-line lexical database[J]. International Journal of Lexicography,

1990, 3(4): 235—244.

- [2] Leacock Claudia, Martin Chodorow. Combining local context and WordNet similarity for word sense identification [C]// Christiane Fellbaum. WordNet, An electronic lexical database. Cambridge: MIT Press, 1998: 265—283.
- [3] Wu Zhibiao, Martha Palmer. Verb semantics and lexical selection [C]// ACL '94 Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1994: 133—138.
- [4] 徐德智, 郑春卉, Pass K. 基于 SUMO 的概念语义相似度研究[J]. 计算机应用, 2006, 26(1): 180—183.

作者简介:

曾 诚 男, 副教授. 研究方向为网络化软件工程.

韩光辉 男, 副教授. 研究方向为软件工程形式化.

李 兵 男, 教授. 研究方向为网络化软件工程.