基于高阶统计的网页隐秘信息检测研究

黄华军[①] 谭骏珊[①] 孙星明[②]
[①](中南林业科技大学计算机与信息工程学院 长沙 410004)
[②](湖南大学计算机与通信学院 长沙 410082)

摘 要：隐秘信息能隐藏在网页标记字母中，虽在浏览器浏览时无法发现其存在，但却不可避免地改变了标记的内在特征 —— 标记偏移量。基于此，该文提出一种新的网页隐秘信息检测算法。根据标记偏移量在隐藏信息前和隐藏信息后的变换规律，确立高阶统计特征来检测网页标记中是否有隐秘信息。实验随机下载了 30 个不同类型网站的主页测试，实验结果验证了统计特征的正确性。检测的漏检率随嵌入信息的增大而减小，当 50%的标记字母被用来隐藏信息后，检测的漏检率为 0%。

关键词：信息隐藏；隐写术；隐写分析；网页；高阶统计；偏移
中图分类号：TP391 **文献标识码**： A **文章编号**：1009-5896(2010)05-1136-05
DOI: 10.3724/SP.J.1146.2009.00530

On Steganalysis of Information in Tags of a Webpage Based on Higher-order Statistics

Huang Hua-jun[①] Tan Jun-shan[①] Sun Xing-ming[②]
[①](*School of Computer and Information Engineering*, *Central and South University of Forestry and Technology*, *Changsha* 410004, *China*)
[②](*School of Computer and Communication*, *Hunan University*, *Changsha* 410082, *China*)

Abstract: Secret message can be embedded into letters in tags of a webpage in ways that are imperceptible to human eye viewed with a browser. These messages, however, alter the inherent characteristic of the offset of a tag. This paper presents a new higher-order statistical steganalytic algorithm for detection of secret messages embedded in a webpage. The offset is used to build the higher-order statistical models to detect whether secret messages hide in tags. 30 homepages are randomly downloaded from different websites to test, and the results show the reliability and accuracy of statistical characteristics. The probability of missing secret messages decrease as the secret message increase, and it is zero, as 50% letters of tags are used to carry secret message.

Key words: Information hiding; Steganography; Steganalysis; Webpage; Higher-order statistics; Offset

1 引言

信息隐藏(information hiding)是指在图像、视频、音频、文本、网页等载体中嵌入一些秘密信息，让第三方在主观上难以察觉秘密信息的存在，主要包括用于隐秘通信的隐写术(steganography)和用于数字媒体版权保护的数字水印技术(digital watermarking)[1,2]。目前，信息隐藏技术已成为信息安全和多媒体版权保护的一个研究热点[3–6]。

隐写术来自于希腊词根(στεγανός, γράφειν)，字面意义是“密写”(covered writing)，通常被解释为把秘密信息隐藏于其他信息当中[1]。隐写术的目的是为了避免传递的隐藏信息引起怀疑。如果隐藏的信息引起了怀疑，隐写术的目的就被破坏了。

隐写分析(steganalysis)是一门发现和破坏隐藏信息的科学，主要包括检测，提取和攻击等领域[2]。它一方面可以促进隐写算法安全性的提高，推动隐写算法实用化[5,6]；另一方面可以防止隐写术被滥用来协助犯罪活动，危害国家安全等[7]。隐写分析已经成为信息隐藏技术中一个重要的研究方向，引起了国内外众多学者的广泛关注[5–11]。隐写分析算法的研究主要集中在基于 LSB(least significant bit)图像隐写术算法上。如 Westfeld 等人根据像素值对(PoVs)的统计分布，建立卡方统计量来检测隐秘信息的存在[8]；Fridrich 等人提出的 RS(regular singular)分析方法能够检测连续和随机替换嵌入的隐秘信息[9]；针对不可见字符和标记大小写变换网页

2009-04-13 收到，2009-12-29 改回
国家 973 计划项目(2006CB303000)，国家自然科学基金重点项目(60736016)，国家自然科学基金(60973128)，湖南省教育厅资助项目(08B091)，中南林业科技大学青年基金重点项目(2008010A)和中南林业科技大学人才引进项目(104-0055)资助课题
通信作者：黄华军 hhj0906@163.com

信息隐藏算法，黄华军等人分别提出相应的检测算法[10,11]。

早在 2000 年，就有研究者指出了在网页中通过加入 Tabs 和 Spaces 来隐藏信息。2004 年，国内研究者也提出了通过修改标记字母的大小写状态来隐藏秘密信息。隐藏在网页的秘密信息可以方便地通过 Internet 进行传递[12−15]。针对网页中隐秘信息的检测已经成为一项紧迫的任务急需得到解决。在 Fridrich 等人提出的建立高阶统计来检测 LSB 图像中隐秘信息的思想上，本文通过深入研究发现：隐藏在网页标记中的秘密信息不会影响网页在浏览器中的正常显示，但隐藏的秘密信息不可避免地改变了标记的内在特征 —— 标记偏移量，提出一种新的网页隐秘信息检测算法。根据标记偏移量在隐藏信息前和隐藏信息后的变化规律，确立高阶统计特征来检测网页标记中是否隐藏了秘密信息。实验随机选择了 30 个不同类型网站的主页测试，结果验证了统计特征的准确性。当大于 50%的标记字母被用来隐藏信息后，检测的漏检率为 0%。

2　现有的网页信息隐藏算法

网页信息隐藏是将网页作为载体的信息隐藏技术。相对于其他载体，网页使用场合更为广泛，在其中加载秘密信息可以更方便地通过因特网进行传递。与图像、视频、音频信息隐藏算法相比，网页信息隐藏所用的算法截然不同。目前，国内外研究者主要是研究基于HTML语法和标记的网页信息隐藏算法，大体上可分为3类[12−15]：第1类是已流行的Wbstego4.2，Invisible Secrets 等信息隐藏工具中利用的不可见字符方法，第2类是基于标记中字母大小写变换的方法，还有一类是基于属性对顺序的方法。

基于标记大小写变换的网页信息隐藏算法的形式化描述如下。设集合 $C=\{26$个大写英文字母$\}$，集合 $A_C=\{A(x)\mid x\in C\}$ 表示大写英文字母的 ASCII 码。再设集合 $c=\{26$个小写英文字母$\}$，$A_c=\{A(x)\mid x\in c\}$ 表示小写英文字母的 ASCII 码。其中函数 $A(\cdot)$ 是求英文字母的 ASCII 码。令 $f_{+1}(x)=x-32$，其中 $x\in A_c$，$f_{+1}(x)\in A_C$，表示将小写字母变换为大写字母的函数。令 $f_{+0}(x)=x$，其中 $x\in A_c$，$f_{+0}(x)\in A_c$，表示将小写字母变换为小写字母的函数。设 $m=\{0,1\}^n$ 表示待隐藏信息 M 的 n 位比特流，$m_i\in m$ 表示第 i 比特信息$(0\le i\le n-1)$。嵌入过程如下：顺序读入标记中的字母 x_i 和待隐藏的信息 $m_i\,(0\le i\le n-1)$。当 $m_i=0$ 时，利用 f_{+0} 对字母 x_i 进行变换。当 $m_i=1$ 时，利用 f_{+1} 对字母 x_i 进行变换。直到 $i=n-1$ 时，嵌入过程完成。

3　标记偏移量与检测算法

3.1 标记偏移量

定义 1　标记偏移量具有如下形式：

$$\sum_{i=1}^{n-1}\left|A(x_{i+1})-A(x_i)\right|,\quad x_i\in c\cup C \tag{1}$$

设 $T(x_1,x_2,\cdots,x_n)$ 表示网页中任意一个标记，其中 x_i 表示标记的第 i 个英文字母$(1\le i\le n)$，n 是标记中字母个数。经深入研究发现：一般情况下，“正常网页”标记的所有字母属于一个集合，即 $x_i\in c$，或 $x_i\in C\,(1\le i\le n)$。网页的标记隐藏信息后，一般情况下标记的字母不完全属于同一个集合。不妨假设其一般形式为：标记中有 k 对相邻字母 x_jx_{j+1}，其中 $x_j\in c$，$x_{j+1}\in C$，且 $0<k<\lceil n/2\rceil$，$1\le j\le n$。其余的 $n-k$ 个字母全部属于 c，或全部属于 C。标记偏移量在标记隐藏信息前和隐藏信息后也发生变化。

定义 2　令 $F(T(x_1,x_2,\cdots,x_n))$ 是计算标记偏移量的偏移函数。偏移函数具有如下形式：

$$F(T(x_1,x_2,\cdots,x_n))=\sum_{i=1}^{n-1}\left|A(x_{i+1})-A(x_i)\right|,\ x_i\in c\cup C \tag{2}$$

集合 A_c，A_C 中两元素最大差值是 25，如：字母“a”和“z”的 ASCII 码相差 25，“A”和“Z”相差 25。而集合 A_c 的元素与集合 A_C 中元素最小差值是 6，如：字母：“Z”和“a”相差 6。为了使集合 A_c 的元素与集合 A_C 中元素最小差值大于 25，令集合 $A_c'=\{x\mid \mathrm{ASCII}(x)+20,x\in c\}$ 是集合 A_c 的偏移集合。此时，集合 A_c' 的元素与集合 A_c 的元素的最小差值为 26。在计算标记偏移量时，采用 A_c' 代替 A_c。

定理 1　$\exists T(x_1,x_2,\cdots,x_n)$，当标记中任意一个字母 $x_i\in c$ 或 $x_i\in C(1\le i\le n)$ 时，标记偏移量最小。

推论 1　$\exists T(x_1,x_2,\cdots,x_n)$，当标记的任意相邻的两个字母 x_jx_{j+1}，$x_j\in c$，$x_{j+1}\in C$，或者 $x_j\in C$，$x_{j+1}\in c\,(1\le j\le n)$ 时，标记偏移量最大。

为节省篇幅，省略了证明过程。

3.2 高阶统计特征

基于标记大小写变换的网页信息隐藏算法中已经定义了函数 f_{+1} 和 f_{+0}。为完备性考虑，令 $f_{-1}(x)=x+32$，其中 $x\in A_C$，$f_{-1}(x)\in A_c$，表示将大写字母变换为小写字母的函数。令 $f_{-0}(x)=x$，其中 $x\in A_C$，$f_{-0}(x)\in A_C$，表示将大写字母变换为大写字母的函数。

定义 3　函数 f_{+1}，f_{+0}，f_{-1} 和 f_{-0} 统称为变换函数。

对标记 $T(x_1,x_2,\cdots,x_n)$ 中的第 i 个字母应用变换

函数，记为

$$f_{M(i)}(x_i),\ M_{(i)} \in \{+1,+0,-1,-0\},(1 \le i \le n) \quad (3)$$

设 $T'(x_1,x_2,\cdots,x_n)=(f_{M(1)}(x_1),f_{M(2)}(x_2),\cdots,f_{M(n)}(x_n))$ 是经过变换函数变换后的标记。若标记的偏移量 $F(T'(x_1,x_2,\cdots,x_n))>F(T(x_1,x_2,\cdots,x_n))$，称 $T(x_1,x_2,\cdots,x_n)$ 是正常的(regular)；若 $F(T'(x_1,x_2,\cdots,x_n))<F(T(x_1,x_2,\cdots,x_n))$，称 $T(x_1,x_2,\cdots,x_n)$ 是异常的(singular)。标记 $T'(x_1,x_2,\cdots,x_n)$ 的偏移量计算如下：

$$F\left(T'(x_1,x_2,\cdots,x_n)\right)=\sum_{i=1}^{n-1}\left|A\left(f_{M(i+1)}(x_{i+1})\right)-A\left(f_{M(i)}(x_i)\right)\right|,\quad x_i \in c\cup C,\ M_{(i)} \in \{+1,+0,-1,-0\},1 \le i \le n \quad (4)$$

采用变换函数对标记 $T(x_1,x_2,\cdots,x_n)$ 进行变换，比较变换前和变换后标记的偏移量，可以得到 3 个集合：R，S，U：

$$R=\{T(x_1,x_2,\cdots,x_n)\,F(T'(x_1,x_2,\cdots,x_n))>F(T(x_1,x_2,\cdots,x_n))\}$$

$$S=\{T(x_1,x_2,\cdots,x_n)\,F(T'(x_1,x_2,\cdots,x_n))<F(T(x_1,x_2,\cdots,x_n))\}$$

$$U=\{T(x_1,x_2,\cdots,x_n)\,|\,F(T'(x_1,x_2,\cdots,x_n))=F(T(x_1,x_2,\cdots,x_n))\}$$

提取待检测网页的标记，并对每个字母应用非负变换，即 $M_{(i)} \in \{+1,+0\}$，$1 \le i \le n$。利用式(4)计算标记的偏移量，并得到集合 R、S 和 U。设 R_M 表示集合 R 中标记个数与网页中标记个数的比值，S_M 表示集合 S 中标记个数与网页中标记个数的比值，总有 $R_M+S_M \le 1$。类似地，对每个字母应用非正变换 $M_{(i)} \in \{-1,-0\}$，$1 \le i \le n$，R_{-M} 表示集合 R 中标记个数与网页中标记个数的比值，S_{-M} 表示集合 S 中标记个数与网页中标记个数的比值，总有 $R_{-M}+S_{-M} \le 1$。图 1 是非负 (非正)变换函数变换标记字母的图。从图 1 可以看出，集合 c 中的字母在 f_{+0} 下变换到集合 c，在 f_{+1} 下变换到集合 C；集合 C 中的字母在 f_{-0} 下变换到集合 C，在 f_{-1} 下变换到集合 c。

如果待检测的网页标记没有隐藏信息，那么无论应用非负变换还是非正变换，从统计上来说，会增加标记偏移量。$\forall T(x_1,x_2,\cdots,x_n)$，当 $\forall x_i \in c(1 \le i \le n)$ 时，一般会有 $R_M \ne 0$，$R_M >> S_M$，且 $S_M \approx 0$，$R_{-M} \approx S_{-M} \approx 0$；当 $\forall x_i \in C(1 \le i \le n)$ 时，一般会有 $R_{-M} >> S_{-M}$，且 $S_{-M} \approx 0$，$R_M \approx S_M \approx 0$。

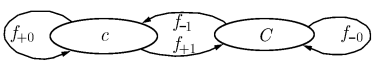

图 1 非负(非正)变换函数变换标记字母

如果待检测的网页标记隐藏有秘密信息，应用非负变换和应用非正变换，从统计上来说，会减小标记偏移量。一般会有 $S_M > R_M$，$S_{-M} > R_{-M}$，且 $S_M > 0$，$S_{-M} > 0$。当网页标记中所有字母隐藏信息，会有，$S_M > R_M$，$S_{-M} > R_{-M}$，$S_M > 0$，$S_{-M} > 0$ 且 $S_M \approx S_{-M}$，$R_M \approx R_{-M}$。

3.3 检测算法

由高阶统计的统计特征知，根据 R_M，S_M，R_{-M} 和 S_{-M} 的相互关系能判断出网页是否隐藏有秘密信息。为简便起见，下面只给出检测算法关键步骤的伪代码：

输入：待检测的网页 H;

输出：“T”表示隐藏了信息，“F”表示没有隐藏信息。

步骤：

$S1$. 提取 H 中的所有标记，并使用数组记录；

$S2$. 利用式(2)计算标记的偏移量，并记录标记的偏移量；

$S3$. 利用 f_{+1}，f_{+0} 对标记中所有字母进行变换，由式(4)计算标记变换后的偏移量，并记录；

$S4$. 比较记录的两次偏移量，获得集合 R，S 和 U。计算出 R_M 和 S_M；

$S5$. 利用 f_{-1}，f_{-0} 对标记中所有字母进行变换，由式(4)计算标记变换后的偏移量，并记录；

$S6$. 比较记录的两次偏移量，获得集合 R，S 和 U。计算出 R_{-M} 和 S_{-M}；

$S7$. 根据 R_M，S_M，R_{-M} 和 S_{-M} 的相互关系满足高阶统计特征，输出“T”，或输出“F”。

4 实验结果与分析

4.1 实验结果

随机下载了 Internet 上 30 个不同类型的网站主页，包括门户网站、新闻网站、会议网站、政府网站、个人网站，以及国外网站等。图 2 给出了没有隐藏信息时，检测出的 R_M 和 R_{-M}（$S_M=S_{-M}\approx 0$，没有在图中标出）从图 2 中得知，网页 4，5，6 和 8 的 $R_M \ne 0$，$R_M >> S_M$ 且 $S_M \approx 0$，$R_{-M} \approx S_{-M} \approx 0$；其余的网页 $R_{-M} \ne 0$，$R_{-M} >> S_{-M}$，且 $S_{-M} \approx 0$，$R_M \approx S_M \approx 0$。此时，$R_M$，$S_M$，$R_{-M}$ 和 S_{-M} 满足高阶统计模型中没有隐藏信息的规律，检测的误报率为 0%。图 3 给出了 100%字母用来隐藏信息后，检测出的 R_M，S_M，R_{-M} 和 S_{-M}。从图 3 中得知，当 100%的标记字母隐藏了信息，有 $S_M > R_M$，$S_{-M} > R_{-M}$，且 $S_M \approx S_{-M}$，在 50%处上下波动；$R_M \approx R_{-M}$，在 15%处上下波动。

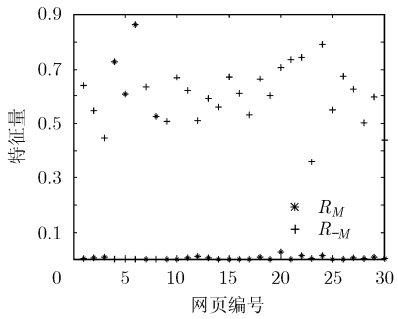

图 2 没有隐藏信息时 R_M 和 R_{-M}

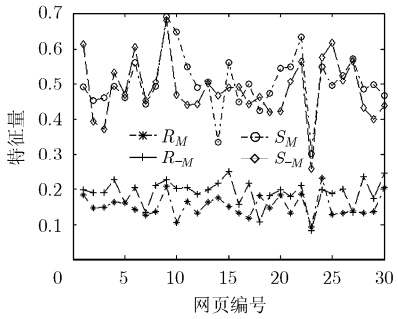

图 3 当 100%字母隐藏信息时 R_M , S_M , R_{-M} , S_{-M}

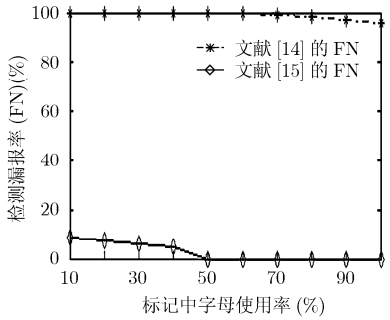

图 4 检测的漏检率

随机从网上下载了 175 份网页测试检测算法的漏检率。从 10%-100%，每次增加 10%信息量的方式，分别利用文献[14]和文献[15]的方法隐藏信息。图 4 中，符号“*”表示是针对文献[14]的检测漏检率 (False Negative， FN)，符号“◊”表示的是针对文献[15]的检测漏检率。从图 4 看出，对文献[14]的检测漏检率在 95%以上；对文献[15]的检测漏检率在标记中字母使用率10%时的漏检率最大，为8.5%。随着隐藏信息的比例增大，漏检率再慢慢减小。当字母比例使用达到 50%以上时，检测的漏检率为 0%。

4.2 分析

检测算法无法准确检测出利用文献[14]的隐藏方法隐藏信息的网页。由推论 1 知，当标记中任意相邻的两个字母属于不同集合时，标记偏移量最大。非负函数 f_{+1} 和 f_{+0}，或非正函数 f_{-1} 和 f_{-0} 对隐藏信息的标记进行变换后，一般情况下，标记的偏移量会变大。因此，，一般会有 $R_M \neq 0$， $R_M >> S_M$，且 $S_M \approx 0$， $R_{-M} \approx S_{-M} \approx 0$。或者 $R_{-M} >> S_{-M}$，且 $S_{-M} \approx 0$， $R_M \approx S_M \approx 0$。从而不能检测出网页中是否隐藏了信息。

为此，我们改进了检测算法。在检测算法加入预处理过程：给定一个待检测的网页，顺序获取标记的首字母，形成向量 $\boldsymbol{G}=(x_1,x_2,\cdots,x_n)$。设置长度为 3(经过统计发现，标记中字母的平均数是 2.3)的滑动窗口，通过移动窗口顺序得到标记组 $G_1(x_1, x_2, x_3)$, $G_2(x_2,x_3,x_4)$, $\cdots$, $G_{n-1}(x_{n-2},x_{n-1},\cdots,x_n)$。图 5 给出了滑动窗口划分标记组的过程。因此，标记组中所有字母都隐藏了信息，可以用检测算法来检测网页中是否隐藏了秘密信息。

5 结束语

随着计算机技术和网络技术的迅猛发展，网络已成为传递信息的主要方式之一。其带来的直接后果之一是给不法分子借助信息隐藏工具，传递秘密信息提供了便利。本文通过对修改标记大小写状态的信息隐藏算法的深入分析，发现隐藏秘密信息后会破坏标记偏移量的内在统计规律。基于此，提出了一种新的网页检测算法。算法能够准确地检测出网页中是否隐藏了信息，而且检测的漏检率较低。

下一步工作是针对现有网页信息隐藏算法，找出各种算法的共性和特性，以及在隐藏信息后引起的网页异常分析和统计，研究覆盖面宽、效率高和准确率好的网页隐秘信息检测算法。

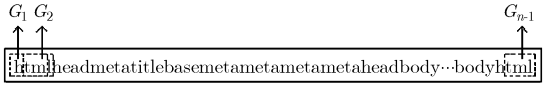

图 5 滑动窗口

参 考 文 献

[1] Petitcolas F A P, Anderson R J, and Kuhn M G. Information hiding-a survey. *Proceedings of the IEEE*, 1999, 87(7): 1062-1078.

[2] Johnson N F and Jajodia S. Steganalysis: the investigation of hidden information. Proceedings of the IEEE Information Technology Conference, Syracuse New York, Sep. 1-3, 1998: 113-116.

[3] 张新鹏, 王朔中. JPEG 图像中的安全密写方案. 电子与信息学报, 2005, 27(11):1813-1818.

Zhang X P, Wang S Z. Secure steganogrphic algorithm in JPEG images. *Journal of Electronics & Information Technology*, 2005, 27(11): 1813-1818.

[4] 胡云, 钮心忻, 杨义先. 静止图像的信息隐藏容量研究. 电子与信息学报, 2004, 26(11): 1681-1685.

Hu Y, Niu X X, and Yang Y X. The study on the capacity of information hiding system on still images. *Journal of Electronics & Information Technology*, 2004, 26(11): 1681-1685.

[5] Fridrich J and Goljan M. Practical steganalysis: State of the art. Proceeding of the SPIE, Security and Watermarking of Multimedia Contents IV, 2002, Vol.4675: 1-13.

[6] Provos N. Defending against statistical steganalysis. Proceedings 10th Advanced Computing System Association, Security Symposium, Washington DC, Aug 2001: 323-335.

[7] Jack K. Terror groups hide behind Web encryption. http://www.usatoday.com/tech/news/2001-02-05-binladen.htm, 2005–10–20.

[8] Westfield A and Pfitzmann A. Attacks on steganographic systems. Lecture Notes in Computer Science, 1999, Vol.1768: 61–76.

[9] Fridrich J, Goljan M, and Du R. Reliable detection of LSB steganography in grayscale and color images. Proceedings of the ACM Workshop on Multimedia and Security, Ottawa, Sep. 19–22, 2001: 27–30.

[10] Huang H J, Sun X M, and Li Z S, *et al*.. Detection of hidden information in webpage. Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery, Haikou, Aug. 25–28, 2007: 317–320.

[11] Huang HJ, Sun XM, and Sun G. Detection of hidden information in tags of webpage based on tag-mismatch. Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Taiwan, Nov. 11–13, 2007: 257–260.

[12] 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. 计算机研究与发展, 2007, 44(5): 756–760.
Sun X M, Huang H J, and Wang B W, *et al*.. An algorithm of webpage information hiding based on equal tag. *Journal of Computer Research and Development*, 2007, 44(5): 756–760.

[13] Zhao Q J and Lu H T. A PCA-based watermarking scheme for tamper-proof of web pages. *Journal of the Pattern Recognition*, 2005, 38(8): 1321–1323.

[14] 沈勇. 一种基于 HTML 文档的信息隐藏方案. 武汉大学学报(理学版), 2004, 50(s1): 217–220.
Shen Y. A scheme of information hiding based on HTML document. *Journal of Wuhan University*, 2004, 50(s1): 217–220.

[15] Sui X G and Luo H. A new steganography method based on hypertext. Proceedings of Asia-Pacific Radio Science Conference, Qingdao, Aug. 24–28, 2004: 181–184.

黄华军： 男，1978年生，博士，副教授，硕士生导师，研究方向为网络与信息安全、网页信息隐藏、网页隐秘信息检测、反网络钓鱼.

谭骏珊： 男，1963年生，教授，博士生导师，研究方向为数据库信息与管理、数据挖掘.

孙星明： 男，1963年生，博士，教授，博士生导师，研究方向为网络信息安全、数字水印、自然语言处理.