ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.7, July 2006, pp.1617−1626
DOI: 10.1360/jos171617
© 2006 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

合谋安全的卷积指纹信息码*

朱 岩[1+], 杨永田[2], 冯登国[3]

[1](北京大学 计算机科学技术研究所,北京 100871)

[2](哈尔滨工程大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

[3](国家信息安全重点实验室(中国科学院 研究生院),北京 100049)

Convolutional Fingerprinting Information Codes for Collusion Security

ZHU Yan[1+], YANG Yong-Tian[2], FENG Deng-Guo[3]

[1](Institute of Computer Science and Technology, Peking University, Beijing 100871, China)

[2](Computer Science and Technology School, Harbin Engineering University, Harbin 150001, China)

[3](State Key Laboratory of Information Security (Graduate School, The Chinese Academy of Sciences), Beijing 100049, China)

+ Corresponding author: Phn: +86-10-82528564, E-mail: zhuyan@icst.pku.edu.cn, http://www.icst.pku.edu.cn

Zhu Y, Yang YT, Feng DG. Convolutional fingerprinting information codes for collusion security. *Journal of Software*, 2006,17(7):1617−1626. http://www.jos.org.cn/1000-9825/17/1617.htm

Abstract: Digital Fingerprinting is a technique for the merchant who can embed unique buyer identity marks into digital media copy, and also makes it possible to identify "traitors" who redistribute their illegal copies. At present,the fingerprinting scheme generally has many difficulties and disadvantages for large-size users, involved in code construction with shorter length and effective traitor-tracing. To resolve these problems, this paper presents the definition of Fingerprinting Information Code and a practical construction method by composing of convolutional codes and common fingerprinting codes based on Boneh-Shaw model. Its decoding algorithm is presented by improving Viterbi algorithm through introducing the idea of 'Optional Subcode Set'. The security and performance are proved and analyzed by theory and example. The results indicate that the proposed scheme has shorter information encoding length and achieves optimal traitor searching in larger-size users.

Key words: digital fingerprinting; collusion security; tracing traitor; convolutional code

摘 要: 数字指纹是一种发行商通过在数字作品拷贝中添加唯一用户身份标记,使得能够识别出制作非法拷贝"叛逆者"的技术.目前,数字指纹方案中普遍存在着大用户数下指纹码构造过长以及无法有效叛逆跟踪的问题和缺陷.为了解决这些问题,在 Boneh-Shaw 的基础上,给出了指纹信息码的定义,并将卷积码与一般指纹码相结合,提出一种实际构造方法.同时,通过引入"备选子码集"的概念对 Viterbi 译码算法予以改进,给出了指纹信息码的译码算法.在理论与实例两方面对编码安全性质和性能的证明和分析结果均表明,所提出方案具有更短的用户信息编码长度,并实现了大用户集下有效叛逆用户的搜索.

关键词: 数字指纹;合谋安全;叛逆跟踪;卷积编码

中图法分类号: TP309 文献标识码: A

Received 2004-10-22; Accepted 2005-09-05

© 中国科学院软件研究所 http://www.jos.org.cn

数字指纹是一种发行商通过在数字作品拷贝中添加每个用户唯一的身份标记,使得在发现非法拷贝后能够识别出制作非法拷贝"叛逆者"的一种知识版权保护技术.数字指纹最初只是作为一种编码技术,但随着数字媒体使用范围的扩大,对数字指纹的要求也不断提高,诸如在原有抗合谋攻击、跟踪性基础上,还要求不可感知性、不可否认性、匿名性等,逐渐形成了包含信号处理、编码理论、密码学等多学科的综合性技术,并且按照功能提出了对称指纹、非对称指纹、匿名指纹等指纹方案.传统的指纹编码是发行商独自为每位用户分配一个码字,并且在合谋攻击下依靠码字能够识别出叛逆用户,但是这种编码并不能实现对信息的编码,因此为更高级应用带来了不便,例如,为了实现用户的不可否认性,需要在拷贝中嵌入用户的秘密信息,当发现盗版拷贝时以获得秘密信息作为法庭证据.为了实现这种功能,本文将用户个人信息编码隐藏于媒体中,并通过提取这些信息达到抵抗合谋、跟踪叛逆者的目的,将这种编码称为指纹信息码,而称指纹编码为指纹码.

Boneh 和 Shaw 首先提出了 c-安全码的构造问题以及"标记假设"来模拟合谋者策略[1],并证明了在此假设下不存在完全 c-安全码,因而在跟踪错误概率为ε的情况下,用上层随机码组合下层二进制码构造出了一种二进 ε-错误 c-安全码的构造,然而此模型(BS 模型)中所含译码算法是一个 NP 困难问题.在文献[2]中,Barg 和 Blaklay 等人针对这一点构造了基于代数码的可跟踪父元码(IPP),使得对于编码长度为 n 的指纹码译码算法的复杂度为 $poly(n)$,但是这种指纹码只适用于合谋人数为 2 的情况,跟踪结果或者以概率 1 识别出其中一个,或者以概率 $1-\exp(\Omega(n))$识别出两个.此外,针对图像、语音等媒体提出了随机指纹编码方法,例如,在文献[3]中给出了一种二进制随机码,但是该方法要保证每位用户使用不同种子生成伪随机序列并保存这些种子,这对于大用户集的叛逆搜索是不现实的.

快速、准确地识别出叛逆用户由于在数字指纹中所具有的重要性,一直得到研究人员相当的重视,也提出了各种方案.其中,在指纹构造中引入结构化编码是极为可行的方法.此外,指纹码具有识别多个合谋码字的能力,如果能将这种能力应用于跟踪算法中则将增强算法的识别能力.基于这两点,本文在 BS 模型基础上提出了指纹信息码的概念,并将卷积码与指纹码相结合构成一种两层链接结构的指纹信息码,同时利用指纹码具有识别多个合谋码字的能力,引入了备选子码集并对 Viterbi 译码算法给予改进,通过对编码长度、抗合谋性、安全性、效率进行证明和分析,实现了更短的指纹码构造和多项式时间的搜索复杂度.

本文第 1 节介绍数字指纹及其相关概念.第 2 节给出 Boneh 和 Shaw 模型.第 3 节详细描述基于卷积指纹信息码的构造及编码、译码算法.第 4 节对性能进行证明和分析.第 5 节给出实例.最后总结全文.

1 数字指纹信息码

数字指纹流程大致可分为版权信息编码、标记嵌入和检测、叛逆用户跟踪 3 部分,其中:标记嵌入与检测涉及到数字水印、广播通信等多媒体技术;而版权信息编码和叛逆用户跟踪是与指纹编码直接相关的.一般地,在 s 个字符组成的字符表Σ中,指纹编码是由指纹序列集合$\Gamma\subseteq\Sigma^l$和算法对(E,T)构成,E 是根据用户 u 和相关信息 r 生成指纹码字的指纹生成算法,即 $I=E(u,r)\in\Gamma$;T 是根据码字 $I'\in\Sigma^l$和销售记录 m 识别出叛逆用户的跟踪算法,即 $U=T(I',m)$,这里,U 可以是叛逆用户集合.

对作品 O 嵌入用户 u_i 的指纹信息的过程如下:首先,指纹生成算法 E 根据信息 r_i 为用户 u_i 生成指纹码字 $I_{u_i}=(i_1,i_2,...,i_L)$,即 $I_{u_i}=E(u_i,r_i)$ $(i_j\in\Sigma,1\le j\le L)$;其次,将作品分割成段,选择 L 个段构成序列 $O=(o_1,o_2,\ldots,o_L)$,再使用水印嵌入算法将 i_j 嵌入 o_j 中生成含指纹的作品拷贝 $O^{(u_i)}$ 并分发给用户和保留销售信息 m_i.当分发商发现盗版拷贝后,水印检测算法在作品的段序列中提取嵌入标记并输出盗版指纹 I';最后,指纹跟踪算法 T_j 根据销售记录 $M=(m_1,m_2,\ldots,m_n)$追查出叛逆者集合 $C=\{u_1,u_2,\ldots,u_b\}$,即 $C=T_j(I',M)$.

数字水印算法根据指纹码字的每个字符 i_j 对作品所作的修改称为一个标记.令 $C=\{u_1,u_2,\ldots,u_c\}$是 c 个叛逆者组成的合谋集合,集合中每人拥有作品 O 的一个带指纹拷贝.在没有原作品的情况下,构造盗版拷贝的基本策略是:通过逐段对比彼此的拷贝确定标记位置,对已确定的标记,使用合谋集合中任意标记或者混杂所有标记来构造一个新的标记;对未检测标记不予改变,从而伪造使指纹失效的盗版拷贝.当指纹编码(E,T)算法只与确定用户相关,即当 $I=E(u)$时,我们称其为指纹码[4],我们对指纹码定义如下:

© 中国科学院软件研究所 http://www.jos.org.cn

定义 **1**(指纹码). (l,n)指纹码是由函数 $E(u)$将用户编号 $u(1\le u\le n)$映射到字符表Σ上长为 l 的字符串集合Σ^l中 n 个序列构成的码字集合.当至多 c 个成员的合谋集合 $C=\{u_1,u_2,\ldots,u_c\}$利用码字 $E(u_1),E(u_2),\ldots,E(u_c)$伪造出一个码字 $z\in\Sigma^l$时,如果存在一个概率跟踪算法 A 以至少 $1-\varepsilon$的概率识别出 C 中至少一名成员,即 $\Pr\{A(z)\in C\}\ge 1-\varepsilon$,我们称这种编码为$\varepsilon$错误的 c-安全码.其中,错误概率由算法 A 和合谋集合 C 的随机选择决定.

与上面的指纹码不同,在指纹码抵抗合谋攻击的基础上,指纹信息码能够对信息进行编码和译码,一般可由销售记录、用户个人信息等来构造指纹信息码.对指纹信息码的定义如下:

定义 **2**(指纹信息码). (L,N)指纹信息码是由函数 $E(u,r)$将用户信息 $m^u(1\le u\le N)$和随机比特串 r 映射到字符表Σ上长为 L 的字符串集合Σ^L中 N 个码字构成的码字集合.当至多 c 个用户的合谋集合 $C=\{u_1,u_2,\ldots,u_c\}$利用码字集合 $C'=(E(m^{u_1},r_1),E(m^{u_2},r_2),\ldots,E(m^{u_c},r_c))$ 伪造出一个码字 $z\in\Sigma^L$ 时,如果存在一个概率跟踪算法 A,以至少 $1-\varepsilon$ 概率识别出 C'中至少一个码字 $E(m^{u_i},r_i)$,即 $\Pr\{A(z)\in C'\}\ge 1-\varepsilon$,进而获得其中随机比特信息 r_i,则称这种编码为ε错误的 c-安全码.其中,错误概率由算法 A、合谋集合 C 和比特串 $r_i(1\le i\le c)$的随机选择决定.

2 Boneh-Shaw 指纹模型

在文献[1]中,Boneh-Shaw 在合谋攻击假设基础上,利用(l,p)-防诬陷码Γ和$(L,N,D)_p$-ECC 码的两层组合给出了 c-防诬陷码模型.首先,为了确定合谋用户所能检测出的标记数量以及伪造码字的策略,提出了标记假设,并要求模型能够经受此假设下的任何攻击.令Σ表示大小为 s 的字符表,Σ中每个字母表示一个不同标记状态并被映射到$[1,s]$中的整数[5].设集合$\Gamma=\{w^{(1)},w^{(2)},\ldots,w^{(n)}\}\in\Sigma^l$被称为一个$(l,n)$-码,$\Gamma$中的每个码字 $w^{(i)}=(w_1^{(i)},w_2^{(i)},\ldots,w_l^{(i)})$ 被分配到用户 u_i.我们称Γ中所有码字构成的集合为Γ的码书,则标记假设可定义为:

定义 **3** (标记假设 **marking assumption**). 令$\Gamma=\{w^{(1)},w^{(2)},\ldots,w^{(n)}\}$是一个$(l,n)$-码,并且 $C=\{u_1,u_2,\ldots,u_c\}$是 c 个叛逆用户的合谋集合,如果 C 中所有码字在位置 i 处字母相同,即 $w_i^{(u_1)}=\ldots=w_i^{(u_c)}$,则称位置 i 为隐藏位且不能被修改,而只有非隐藏位能被修改,定义合谋集 C 所能产生的可行码字集合$\Gamma(C)$为

$$\Gamma(C)=(x_1,\ldots,x_l)\in\Sigma^l\,|\,x_j\in W_j,1\le j\le l \tag{1}$$

其中,

$$W_j=\begin{cases}\{w_j^{(1)}\}, & w_j^{(1)}=\ldots=w_j^{(l)}\\ \{w_j^{(1)}\,|\,1\le i\le c\}\cup\{\perp\}, & \text{otherwise}\end{cases} \tag{2}$$

其中,$\perp$称为擦除标记.

在(l,n)-码上的指纹算法是指将用户 u 的标记通过随机置换位串 $r\in\{0,1\}^*$映射到Σ^l中码字的函数$\Gamma(u,r)$,其中 r 是发行商随机选择但对用户保密的参数.指纹算法用Γ_r表示,其安全性可由下面定义表示:

定义 **4**($\boldsymbol{\varepsilon}$-错误 $\boldsymbol{c}$-安全指纹码). 一个指纹编码Γ_r被称为是ε-错误 c-安全码,如果存在一个跟踪算法 A,使得对任何至多 c 个叛逆者构成的合谋集 C 所伪造的码字 x,都有

$$\Pr\{A(x)\in C\}>1-\varepsilon.$$

其中,概率分布由 C 中成员的随机选择和 r 的随机掷币决定.

在(l,n)-码设计中,码长 l 与用户数目 n、合谋人数 c 的关系是 $l=16c^2\log n$.当用户数目较大时,这是不现实的.为了在有限长度媒体中嵌入指纹和存储用户信息,BS 模型给出了一种两层编码的混合方案:上层为 ECC 纠错码层,下层为(l,n)-指纹码层.这种$(L,N,D)_p$-ECC 码定义如下:

定义 **5** ($\boldsymbol{(L,N,D)_p}$-**ECC** 码). 在 p 个字符组成的字符表Σ中,码长 L 的 N 个码字构成的集合$\Im$.如果在$\Im$中,每两个码字的汉明距离至少为 D,那么,称集合$\Im$为$(L,N,D)_p$-ECC 码.

由(l,n)-码Γ和$\Im$组合构造出$(L,N,D)_p$-ECC 指纹码Γ'中包含Γ和$\Im$的任意组合,码字是独立随机均匀分布的,并且没有规定具体的组合形式.对指纹码Γ'的安全性有如下定理:

定理 **1**. 给定整数 $N,c,\varepsilon>0$,令 $n=2c,L=2c\log(2N/\varepsilon),d=2n^2\log(4nL/\varepsilon)$,那么,$\Gamma'(L,N,n,d)$是一个$\varepsilon$-错误 c-安全码,编码中包含 N 个码字,并且编码长度为 $O(Ldn)=O(c^4\log(N/\varepsilon)\log(1/\varepsilon))$.

在定理 1 中,n 为指纹码Γ的码字数;d 为每块指纹码Γ中重复的比特数.虽然 BS 模型较好地解决了短指纹码

© 中国科学院软件研究所 http://www.jos.org.cn

相互连接构造长指纹码的问题,但当发现盗版拷贝时,模型利用跟踪算法从每块指纹码中任意选择一个叛逆者符号并相互连接作为 ECC 层输入,再通过全局搜索找到一个篡改前的码字.由此便产生了几个问题[6]:一方面,从*I*指纹码的特点分析,跟踪算法有能力找出多个乃至所有叛逆者的能力,而 BS 模型仅仅利用了其中一个叛逆者信息,因此增大了解码难度;另一方面,模型中使用了全局搜索的策略(见文献[1]中的算法 2),现实中指纹系统用户数目很大,不仅要存储所有这些编码,而且全局逐一搜索也是不现实的.

3 基于卷积码的数字指纹

本文在 BS 模型基础上,针对模型中存在的问题加以改进,提出一种基于卷积码的数字指纹方案,BS 模型与本文所提方案的主要区别在于:

(1) 将结构化编码引入指纹编码中,用卷积码代替 BS 模型中 ECC 编码,从而采用最大似然译码提高译码效率和减少编码长度,并且实现指纹信息长度与用户数和指纹码无关;

(2) 根据卷积码的特点,通过引入备选子码集的概念,对卷积码的 Viterbi 译码算法进行改进,从而最大限度地利用指纹跟踪算法结果来减少跟踪错误概率.

本文的目标是将 N 个用户的身份信息嵌入到媒体中,在标记假设下构造一个ε-错误 c-安全编码方案,以达到改善运行时间和存储要求的目的,其核心是指纹信息码构造.下面,我们分别对编码和译码两方面进行阐述.

3.1 指纹信息码编码

卷积码[7]是 1955 年由 Elias 提出的,与分组码不同,在(n_0,k_0,m_0)-码编码中,本组的 n_0-k_0 个校验元不仅与本组的 k_0 个信息有关,而且与以前各时刻输入到编码器的信息组有关.同样,译码过程也要利用以前或以后各时刻收到的码组中的相关信息.由卷积码构造出来的自正交码、删余码等都有较好的性能,且实现最佳和准最佳译码也比分组码容易.因此,本文采用卷积码作为指纹信息码构造的基础.

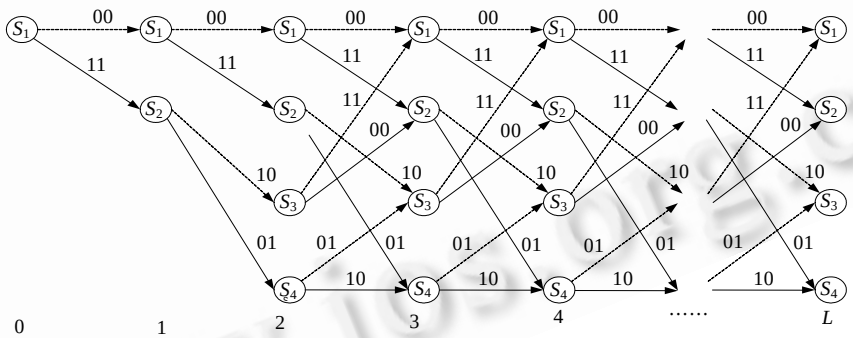


Fig.1 Trellis diagram for (2,1,2) convolutional code

图 1 (2,1,2)卷积码的译码网格图

我们构造的ε-错误 c-安全指纹信息码$\Phi(L,N,n,d)$采用两层结构:上层为卷积纠错码层,下层为指纹码层.令 c 表示最大合谋人数,N 为用户人数,L 为指纹码个数, $m^{(u_i)}$ 为用户 $u_i(1\le i\le N)$的身份信息,它由用户随机选择并满足独立均匀分布,并且保证每位用户具有唯一性.通过提取盗版拷贝的身份信息,不仅可以知道叛逆者,而且在某些情况下可作为指认嫌疑人的证据.卷积纠错码层完成信息 m_i 的分组卷积编码,再对每组码字进行指纹编码来实现编码的抗合谋攻击,最终通过指纹码的联接构成指纹信息码.合谋用户的攻击策略采用定义 3 中的标记假设,但是本文不考虑擦除标记的情况,否则,可选用具有可增/删性质的卷积码.

3.1.1 指纹码层编码

指纹码层是以在有限码书内抵抗合谋攻击并识别叛逆者所用码字为目标.目前研究较多的跟踪码、父元码、防诬陷码都适合作为指纹码层,但必须保证码书大小 $n\ge c$(c 为整个指纹信息码的合谋人数).这里采用 c-防诬陷(l,n)-码Γ,并以$\Gamma_0(n,d)$码构造的(l,n)-码进行说明[1].$\Gamma_0(n,d)$码是将字符串$(c_1,c_2,\ldots,c_{n-1})$中的每个字符重复 d

© 中国科学院软件研究所 http://www.jos.org.cn

次得到的编码,其中:n 为码字数目;d 决定了译码错误概率ε.c_i 的 d 次重复称为一块,用 $B_i(1\le i\le n-1)$表示.令 $\{w^{(1)},w^{(2)},\ldots,w^{(n)}\}$表示$\Gamma_0(n,d)$中的码字,那么,码字 $w^{(i)}$定义如下:

$$w^{(i)}=\underbrace{\underbrace{00\ldots0}_{d},\ldots,\underbrace{00\ldots0}_{d}}_{(i-1)\times d},\underbrace{\underbrace{11\ldots1}_{d},\ldots,\underbrace{11\ldots1}_{d}}_{(n-i)\times d},\ 1\le i\le n \tag{3}$$

其中:当 i=0 时,$w^{(0)}=\{1\}^{d(n-1)}$.指纹码的码字长度为 $d(n-1)$.

3.1.2 卷积纠错码层编码

令卷积码$\Im$为(n_0,k_0,m_0)-码,将用户身份信息 $m^{(u_i)}$ 中每 k_0 比特分为一组,共 L 组(包括 m_0k_0 比特构成结尾码字).在卷积编码后得到码字 $v=(v_1,v_2,\ldots,v_L)$,长度为 n_0L 比特,卷积码易于选择码间自由距离较大的自正交码或纠突发错误码.如果用状态图来描述卷积码,共有 $2^{k_0m_0}$ 种状态,每个字符 $v_i(1\le i\le L)$将有 $2^{k_0m_0}$ 种可能的输出,那么,在得到的 L 组编码中,令 $n\ge 2^{k_0m_0}$ 且每组的输出 v_i 对应指纹码的一个码字 $w^{(v_i)}$,因此可得到 L 个码字:$w^{(v_1)},w^{(v_2)},\ldots,w^{(v_L)}$.将这些码字相互链接再根据$\pi$对其进行置换,就得到最终的$\varepsilon$-错误 c-安全指纹信息码.置换π由发行商密钥通过伪随机序列发生器(pseudorandom sequence generator)随机生成,并对所有用户保持不变,指纹信息码的安全性完全依赖于秘密置换π,而于两层编码无关.因此,两层编码算法对用户是公开的.

由(l,n)-码Γ和(n_0,k_0,m_0)卷积码$\Im$构造出的指纹信息码Φ的过程可以简单描述如下:对用户 u_i 身份信息 m 和编号 i 进行卷积编码,得到码字 $v=(v_1,v_2,\ldots,v_L)$,再分别对每一分量 v_i 在指纹码书中查找到相应的码字 $w^{(v_i)}$,将所有码字进行链接,即 $W_v=(w^{(v_1)}\|w^{(v_2)}\|\ldots\|w^{(v_L)})$,并完成秘密置换 $W^{(u_i)}=\pi(W_v)$,得到用户 u_i 在Φ中的一个码字 W_v.由此可见,Φ的码字长度为 $Ld(n-1)$,使用伪码编写的编码算法如下:

算法 **1**. 指纹信息码编码算法.

Encoding (message $m^{(u_i)}$, user u_i, permutation π)

Let $v=$Convolutional-Encoding $(m^{(u_i)},i)$ //卷积编码

For each $1\le k\le L$

$w^{(v_k)}=$Fingerprinting-Encoding(v_k) //指纹编码

Return $W^{(u_i)}=\pi(w^{(v_1)}\|w^{(v_2)}\|\ldots\|w^{(v_L)})$

End

3.2 指纹信息码译码

指纹信息解码采用自底向上的方法进行:首先从盗版拷贝中提取嵌入的码字序列,用发行商密钥获得秘密置换π的逆π^{-1} 来恢复指纹信息码并进行分组;再对每组序列进行指纹码解码得到备选子码集;最后,综合所有备选子码集进行卷积译码,从而识别出叛逆者.下面,分别就指纹码解码和卷积解码两方面进行阐述.

3.2.1 指纹码译码

(l,n)-指纹码译码采用类似文献[1]中给出的算法,但是我们注意到:当合谋用户对多个作品标记进行任意混合时,(l,n)码有能力找出多个叛逆码字.由于$(L,N,D)_p$-ECC 码的限制,BS 模型没有考虑这种能力,所提出的解码算法输出仅为合谋集合中任意一个码字.与原有译码算法不同,在我们提出的算法中,算法输出为合谋者的码字集合,我们称此集合为备选子码集.令 $R_i=B_{i-1}\cup B_i$,指纹码译码算法表示如下:

算法 **2**. 指纹码译码算法.

Fingerprinting-Decoding (fingerprinting-code x)

Let $U=\{\}$

IF $weight(x|_{B_1})>0$ Then $U=U\cup\{1\}$

For each $1\le k\le n-1$

Let $s=weight(x|_{B_k})/2$

IF $weight(x|_{B_{k-1}})<s-\sqrt{s\log(2n/\varepsilon)}$

© 中国科学院软件研究所 http://www.jos.org.cn

Then $U=U\cup\{k\}$

IF $weight(x|_{B_{n-1}})<d$

Then $U=U\cup\{n\}$

Return U

End

3.2.2 卷积纠错码层译码

在(n_0,k_0,m_0)卷积码译码中,由于备选子码集的使用扩展了译码码字空间,因此采用改进的 Viterbi 算法进行最大似然译码(MLD).Viterbi 译码是一种概率译码,在网格图(trellis)中,设译码器的接收序列为 $R=(r_1,r_2,\ldots,r_L)$,其中每个备选子码集 r_i 由可选的多个码元组成,即 $r_i=\{r_{i,1},r_{i,2},\ldots,r_{i,l}\}$且 $l\in\mathbb{N}$.最大似然译码准则力图找出一条编码器在网格图上所走过的路径,即寻找长度为 L 的 2^{k_0L} 条路经中最大似然路径,即 $\max_j(\log\Pr(R\mid H_i))$ $(1\le i\le 2^{k_0L})$,其中 $\Pr(R|H_i)$为路径 R 与 H_i 的似然函数.令 R_i 表示接收序列 R 的前 i 步的所有路径集合,即 $R_i=(r_1,r_2,\ldots,r_i)$;C_i 表示所有 i 步能到达的路径;$C_{i,j}$ 表示所有第 i 步到达状态 S_j 的路径;$e_{i,j}$ 表示网格中由状态 i 到状态 S_j 的边;函数 $D(X|Y)$表示路径 X 与路径 Y 的汉明距离.在第 i 时刻到达第 S_j 状态的路径 $C_{i,j}$ 有很多条,其中一条最大似然路径被称为留选路径 $sp_{i,j}=(e_{1,i_1},e_{2,i_2},\ldots,e_{i,i_i})$,它与 R_i 的路径度量称为部分路径度量.对于二进对称信道(BSC),最大似然译码与最小距离译码等价.因此,部分路径度量具有最小汉明距离,即 $d_{i,j}=D(sp_{i,j})=\mathrm{mind}(R_i|C_{i,j})$.同时,本节只使用最小距离对算法进行说明.

改进后的 Viterbi 译码算法中需要保留搜索路径中每步的留选路径和部分路径度量,算法步骤是:(1) 初始化:设置初始状态 S_1 的部分路径度量为 0(其他状态为∞)以及每个状态的留选路径为空;(2) 逐步计算:从接收码字中的第 1 块开始,假定第 k 步中已经得到了每一状态的留选路径和部分路径度量.在第 k+1 步,把进入每一状态 $S_j(1\le j\le n)$的所有分支 $e_{i,j}$ 和与此分支相连的前一时刻留选路径 $sp_{k,i}$ 相加得到本状态的候选路径 $sq_{i,j}=(sp_{k,i},e_{i,j})$.同时,为了计算候选路径 $sq_{k+1,j}$ 与 R 的前 k+1 步路径 R_{k+1} 的最小路径度量作为留选路径 $d_{k+1,j}$,我们对每条进入分支 $e_{i,j}$ 计算 $e_{i,j}$ 与第 k+1 步备选子码集 r_{k+1} 的最小度量 $\min D(r_{k+1}|e_{i,j})$和状态 S_i 的部分路径度量 $d_{k,i}$ 相加和的最小值作为部分路径度量 $d_{k+1,j}$:

$$d_{k+1,j}=\min D(R_{k+1}|sq_{i,j})=\min(d_{k,i}+\min D(r_{k+1}|e_{i,j})) \tag{4}$$

其中,备选子码集 $r_{k+1}=\{r_{k+1,1},r_{k+1,2},\ldots,r_{k+1,m}\}$与分支 $e_{i,j}$ 的最小度量为 $\min D(r_{k+1}|e_{i,j})=\min_{1\le l\le m}D(r_{k+1,l}|e_{i,j})$,相应的路径为留选路径 $sp_{k+1,j}$.最后,对留选路径 $sp_{k+1,j}$ 及其部分路径度量 $d_{k+1,j}$ 加以存储并去掉其他候选路径;(3) 终止:若 $k\le L$,则重复第(2)步;否则算法终止,译码器得到最小部分路径度量 $\min_{1\le l\le m}(d_{L,j})$对应的留选路径的路径编码序列作为最大似然解码序列输出.详细的指纹信息码译码算法表示如下:

算法 **3**. 指纹信息码译码算法.

Convolutional-Decoding (code x, permutation π)

Let $d_{0,i}=\infty, sp_{0,i}=\{\}$ for $(1\le i\le n)$ except $d_{0,1}=0$

Let $W=\pi^{-1}(x)$

For each $1\le k\le L$

Let r_k=Fingerprinting-Decoding(W_k) //指纹码译码,见算法 2

For each state s_j

For each the incoming branch $e_{i,j}$

For each the element $r_{k,l}\in r_k$ $(1\le l\le m)$

$c_{k,l}=D(r_{k,l}|e_{i,j})$

Let $sq_{i,j}=(sp_{k,i},e_{i,j}), t_{i,j}=d_{k-1,i}+\min_{1\le l\le m}c_{k,l}$

Let $d_{k,j}=\min_i(t_{i,j})$ for exist $e_{i,j}$

Let $sp_{k,j}=sq_{l,j}$ for all $d_{k,j}==t_{l,j}$ //如果有多条最小度量路径,将其全部存储

Let $d_l=\min_{1\le j\le n}(d_{L,j})$

© 中国科学院软件研究所 http://www.jos.org.cn

Return $M(sp_{L,l})$ //函数 M 是计算最小留选路径的路径编码序列

End

4 指纹信息码性能分析

定理 **2**. 在改进的 Viterbi 译码算法 3 中,留选路径是最大似然路径,即存在留选路径 sp,对所有候选路径 $sq\neq sp$,$D(R|sp)\geq D(R|sq)$.

证明:在 Viterbi 译码器的网格图中,令第 k 时刻以前所有状态的留选路径就是最大似然路径.当译码器从时刻 k 转换到时刻 k+1 时,根据我们所提出算法求得的状态 S_j 的留选路径为 $sp_{k+1,j}=(sp_{k,i},e_{i,j})$,相应的部分路径度量 $d_{k+1,j}$ 具有如下关系:

$$d_{k+1,j}=\min D(R_{k+1}|sp_{k+1,j})=d_{k,i}+\min D(r_{k+1}|e_{i,j}) \tag{5}$$

其中,对所有 $r_{k+1,l}\in r_{k+1}$ 有,$\min D(r_{k+1}|e_{i,j})=\min_l D(r_{k+1,l}|e_{i,j})$.使用反证法进行证明:设存在另一条与 $sp_{k+1,j}$ 不同的最大似然路径,并且其度量小于 $sp_{i+1,j}$ 的度量.不妨令这条路经为 $sp'_{k+1,j}=(p_{k,i'},e_{i',j})$,使得 $i\neq i'$,其中 $p_{k,i'}$ 表示这条路径的前 k 条边,那么,这条路经 $sp'_{k+1,j}$ 的部分路径度量为

$$d'_{k+1,j}=\min D(R\mid sp'_{k+1,j})=\min D(R_k\mid p_{k,i'})+\min D(r_{k+1}\mid e_{i',j})=d'_{k,i'}+\min D(r_{k+1}\mid e_{i',j}) \tag{6}$$

其中,令 $d'_{k,i'}=\min D(R_k\mid p_{k,i'})$.此外,由于在算法中 $sp_{k+1,j}$ 是留选路径,因而在第 k+1 时刻到达状态 S_j 的路径中 $sp_{k+1,j}$ 的度量值最小,即对任意 $1\leq l\leq n$ 且 $i\neq l$,有

$$\min D(R_{k+1}|sp_{k+1,j})=d_{k,i}+\min D(r_{k+1}|e_{i,j})\leq\min D(R_{k+1}|(C_{k,l},e_{l,j}))=d_{k,l}+\min D(r_{k+1}|e_{l,j}) \tag{7}$$

对 i'而言,$i\neq i'$,根据式(7)有

$$d_{k,i}+\min D(r_{k+1}\mid e_{i,j})\leq d_{k,i'}+\min D(r_{k+1}\mid e_{i',j}) \tag{8}$$

因为 $sp'_{k+1,j}$ 的度量小于 $sp_{i+1,j}$ 的度量,即 $d'_{k+1,j}<d_{k+1,j}$,根据式(5)、式(6)和式(8)有

$$d'_{k+1,j}=d'_{k,i'}+\min D(r_{k+1}\mid e_{i',j})<d_{k+1,j}=d_{k,i}+\min D(r_{k+1}\mid e_{i,j})\leq d_{k,i'}+\min D(r_{k+1}\mid e_{i',j}) \tag{9}$$

可知,$d'_{k,i'}<d_{k,i'}$.这与状态 $S_{i'}$ 在时刻 k 是最大似然路径矛盾,命题得证. □

根据对算法的分析以及卷积码、(l,n)码的性质,可以证明本文所提出的指纹方案具有至少 $1\sim\varepsilon$的概率识别出叛逆用户的能力.

定理 **3**. 给定整数 $N,c,\varepsilon>0$,令 $n=2c,d=2n^2(\log(8n)+r)$, $r=(2/d_f)\log(A_{d_f}/\varepsilon)$,$d_f$为码的自由距离, A_{d_f} 为卷积码中重量为 d_f的码序列数,那么,$\Phi(L,N,n,d)$是一个ε-错误 c-安全码,编码中包含 N 个码字.对于任何至多 c 个用户的合谋集合 C 生成的码字 x,算法 2 和算法 3 能以至少 $1-\varepsilon$的概率识别出 C 中至少一个码字.

证明:由卷积码概率译码的性质,在 BSC 信道下,Viterbi 译码器产生错误译码概率 P_e 为

$$P_e\approx A_{d_f}2^{d_f}p^{d_f/2} \tag{10}$$

其中:d_f为码的自由距离; A_{d_f} 为卷积码中重量为 d_f的码序列数;p 是信道转移概率.令在指纹信息码 Φ中,(l,n)指纹码 Γ的译码错误概率为ε',那么,信道转移概率就是指纹码 Γ的错误概率,即 $p=\varepsilon'=\frac{1}{4}(P_e/A_{d_f})^{2/d_f}$.根据指纹码 Γ 的性质可知,码字长度为 $d=2n^2\log(2n/\varepsilon')$,则在 Φ中,码字长度为

$$d=2n^2(\log(8n)+(2/d_f)\log(A_{d_f}/P_e)) \tag{11}$$

注意,卷积码性质决定了编码长度 L 与错误概率无关.此外,在指纹信息码 Φ中有 $P_e=\varepsilon$,这表明,当至多 c 个用户的合谋集合 C 生成码字 x 时,指纹信息码将以至多ε的概率产生错误译码.再由定理 2 可知,改进后的卷积算法能有效地实现最大似然搜索,从而证明算法能以至少 $1-\varepsilon$的概率识别出 C 中至少一个码字,命题得证. □

本文所提出的编码方案与 BS 模型及其他指纹方案相比,在编码长度、译码性能等方面都有所改进,具体如下:

1) 在编码长度方面:本文所提出的编码方案与 BS 模型比较,两者编码长度都为 $Ld(n-1)$,但不同在于:BS 模型中,L 由 N 和ε决定,且$\Gamma_0(l,n)$-码承受的错误概率是$\varepsilon/2L$,使得块长度 d 随 L 的增长而增大;但是根据卷积码性质,卷积码中的本组码字仅与相邻 m_0 组相关,如定理 3 所示,块长度 d 只增加固定的 $2rn^2=8rc^2$ 比特.因而,在我们所

© 中国科学院软件研究所 http://www.jos.org.cn

提出的算法中,L 仅由用户信息决定,而与 d 无关,因而具有更短的编码长度.同时,可事先确定(l,n)码的参数;

2) 在译码性能方面:L 组指纹码译码复杂度与编码长度 $O(Ldn)$成正比,而(n_0,k_0,m_0)卷积码的译码器需要执行 L 步处理,每步搜索 n 个状态,因此复杂度为 $O(Ln)$;

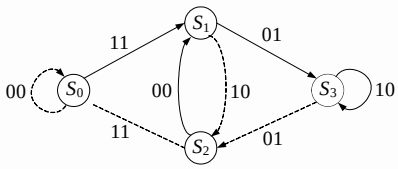
3) 在存储性能方面:卷积码译码器需要保留 $n=2^{k_0m_0}$ 个状态,并且对每一状态必须有一个路径寄存器以及一个部分路径度量值存储器,所以译码器的存储复杂性为 $O(Ln)$,其中, $n=2^{k_0m_0}$ 且 $n=2c$,则 $k_0=(\log\ c+1)/m_0$,一般要求 $m_0\leq 10$,用户信息 m 长度满足$|m|=k_0L$.由于受媒体长度限制,希望尽量缩短 L,因而可以适当调整 k_0 和 m_0 的取值,即使对较大合谋人数 c,也不难选择较好的卷积编码.当 L 较大时,也可以采用 Viterbi 译码的结尾译码法.

目前,大多数编码方案都是在 BS 模型基础上构造的,如可验证父原码(IPP)[8]、c-跟踪码(c-TA)、防诬陷码(FP)[9]等.它们的编码和译码形式虽各不相同,但基本保留了 BS 模型的性质,而且普遍存在上面编码长度和译码效率上的问题[10].总之,本文提出的编码方案已经接近于安全指纹码构造的下限[11],并可实现多项式内的有效搜索.

5 应用实例

为了更好地说明本文所提出的指纹信息码,下面以实例加以阐述.令卷积码采用常见的(2,1,2)-码,其生成多项式为 $g^{(1)}(D)=1+D^2$,$g^{(2)}(D)=1+D+D^2$.码的自由距离 d_f 为 5,卷积码中重量为 5 的码序列数的 A_{d_f} 为 1.编码器延迟编码数 $m=2$,状态数为 4,其编码器状态图如图 2(a)所示.其中,实线表示 1,虚线表示 0.

指纹码采用$\Gamma=\Gamma_0(4,d)=\{111,011,001,000\}$,并且对$\Gamma$中每个用户码字逐位扩展 d 次.例如,当 $d=3$ 时,$\Gamma_0(4,3)$码如图 2(b)所示.根据定理 3,这种指纹信息码抵抗合谋攻击的人数为 $c=2$ 人,$n=4$,当跟踪错误概率$\varepsilon=0.001$,扩展 $d\geq 300$ 时,能够保证指纹信息码是 2-安全的.不失一般性,我们这里令 $L=7$ 来阐述指纹信息码的编码和译码过程.



(a) State diagram for (2,1,2) convolutional code and

(a) (2,1,2)卷积码的编码器状态图

$w^{(1)}$	111	111	111
$w^{(2)}$	000	111	111
$w^{(3)}$	000	000	111
$w^{(4)}$	000	000	000

(b) Coding table for $\Gamma_0(4,3)$ code

(b) $\Gamma_0(4,3)$码编码表

Fig.2 State diagram for (2,1,2) convolutional code and coding table for $\Gamma_0(4,3)$ code

图 2 (2,1,2)卷积码的编码器状态图和$\Gamma_0(4,3)$码编码表

假设两名用户 u_1 和 u_2 分别随机选择长为 $L=7$ 的身份信息 M_1=(1011100)和 M_2=(0110100),其中最后两位是添入的终止字符.他们的指纹信息码编码过程如下:首先,身份信息经卷积码编码器得到输出码字序列 $R^{(1)}$=(11,10,00,01,10,01,11)和 $R^{(2)}$=(00,11,01,01,00,10,11);再令每块两位信息表示指纹码字符表中的一个字符,假设字符表中的字符用{1,2,3,4}表示,则 R^1=(4,3,1,2,3,2,4)和 R^2=(1,4,2,2,1,3,4);根据指纹信息码Φ的定义,获得两个用户的抗合谋编码 $W^{(1)}=(w^4,w^3,w^1,w^2,w^3,w^2,w^4)$和 $W^{(2)}=(w^1,w^4,w^2,w^2,w^1,w^3,w^4)$,其中每块 $w^i(1\leq i\leq 4)$的定义可查找指纹码表Γ(如图 2(b)所示);最后,作品发行商随机选择一个置换π,求取$\pi(W^{(1)})$和$\pi(W^{(2)})$作为用户的指纹信息码并嵌入到作品中分发给用户.

当发行商发现一份原作品的非法拷贝时:首先,他提取拷贝中的指纹信息,并对信息进行反向置换π^{-1} 得到指纹信息码 W';再根据算法 2,分别对每一块求取合谋集合 $r_i(1\leq i\leq L)$(注意,r_i 可能含有多个字符);最终,连接每个集合得到序列 R(这里假设用户 u_1 和 u_2 进行合谋).根据标记假设,不妨令合谋后提取的码字是 $R=(\{w^4\},\{w^3,w^4\},\{w^1,w^2\},\{w^2\},\{w^1\},\{w^2,w^3\},\{w^4\})$,其中,第 4 位和第 7 位是隐藏位,根据标记假设不能被改变.根据算法 3,首先将 R 转化为二进制序列:$R'=(\{11\},\{10,11\},\{00,01\},\{01\},\{00\},\{01,10\},\{11\})$,再将 R'作为卷积译码算法的输入.详细译码过程如图 3 所示.在图中,译码过程由 7 步组成.注意,本文提出的算法与原算法的差异是:在原算法中,备选子码集 $r_i(1\leq i\leq L)$中只有一个元素;而这里的算法中,r_i 可由多个元素组成.例如第 3 步中,r_3={00,01},状态 S_1 有两条

© 中国科学院软件研究所 http://www.jos.org.cn

进入分支 $e_{1,1}$ 和 $e_{3,1}$,$e_{1,1}$ 与 r_3 的最小距离是 $M(r_3|e_{1,1})=\min(D(00|00),D(01|00))=0$,其候选路径距离为 $d'_{2,1}+M(r_3|e_{1,1})=3$;$e_{3,1}$ 与 r_3 的最小距离是 $M(r_3|e_{3,1})=\min(D(00|11),D(01|11))=1$,候选路径距离为 $d'_{2,3}+M(r_3|e_{3,1})=1$.因此,留选路径经过 $e_{3,1}$,最小距离是 $d_{3,1}=1$.

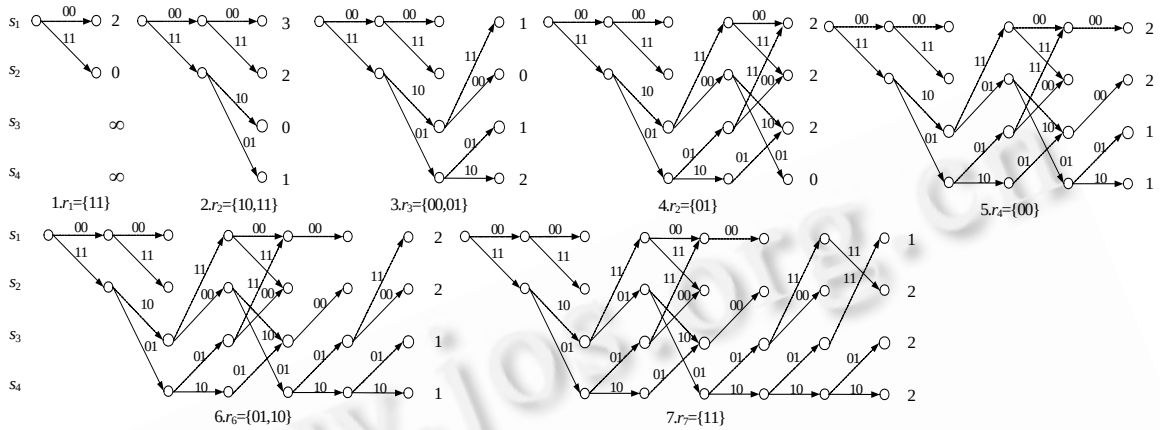


Fig.3 Trellis diagrams of the decoding process for the improved Viterbi algorithm

图 3 用改进的 Viterbi 译码算法在网格图上的译码过程

由最后的译码结果可知,最小度量为 1 的留选路径为(11,10,00,01,10,01,11),它与 R'的最小汉明距离为 1.相应的信息为 M_1=(1011100),即用户 u_1 参与了盗版行为.由此可见,本文所提出的指纹信息码能够快速而有效地完成叛逆跟踪.对比 BS 模型及其他方法,它们所采用的方法通常是从敌手合谋攻击后码字 R'=({11},{10,11},{00,01},{01},{00},{01,10},{11})中的每个备选子码集中随机抽取一个码字构成译码器的输入,如选择 R'=(11,10,01,01,00,10,11),因此减少了可用于跟踪的信息量,容易产生译码错误.同时,为了得到较为正确的结果,需要多次随机选择并译码,最终根据多次结果确定叛逆者.但这样做有时是不可取的.

6 总 结

本文将卷积码与指纹码相结合构成一种两层链接结构的指纹信息码,同时引入备选子码集,并且提出了改进的 Viterbi 译码算法,实现了更短的指纹码构造和多项式时间的搜索复杂度.本文为数字指纹编码提供了一条新思路,对知识产权保护的其他领域,在理论和实践上都具有指导意义.

References:

[1] Boneh D, Shaw J. Collusion-Secure fingerprinting for digital data. IEEE Trans. on Information Theory, 1998,44(5):1897−1905.

[2] Barg A, Blakly GR, Kabatiansky G. Digital fingerprinting codes: Problem statements, constructions, identification of traitors. Technical Report, DIMACS 2001-52, Piscataway: Center for Discrete Mathematics and Theoretical Computer Science Founded as NSFSTC, 2001. 1−26.

[3] Wang Y, Lu SW, Xu HL. A digital fingerprinting algorithm based on binary codes. Journal of Software, 2003,14(6):1172−1177 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1172.htm

[4] Biehl I, Meyer B. Protocols for collusion-secure asymmetric fingerprinting (extended abstract). In: Rüdiger R, Michel M, eds. Proc. of the 14th Annual Symp. on Theoretical Aspects of Computer Science. London: Springer-Verlag, 1997. 399−412.

[5] Naini RS, Wang Y. Collusion secure q-ary fingerprinting for perceptual content. In: Sander T, ed. Security and Privacy in Digital Rights. Berlin: Springer-Verlag, 2002. 57−75.

[6] Lindkvist T. Characteristics of some binary codes for fingerprinting. In: Pieprzyk J, Okamoto E, Seberry J, eds. Proc. of the ISW 2000. Berlin: Springer-Verlag, 2000. 97−107.

© 中国科学院软件研究所 http://www.jos.org.cn

[7] Chan F, Haccoun D. Adaptive viterbi decoding of convolutional codes over memoryless channels. IEEE Trans. on Communications, 1997,45(11):1389−1400.

[8] Cohen G, Encheva S, Litsyn S. Intersecting codes and partially identifying codes. In: Daniel A, ed. Int'l Workshop on Coding and Cryptography. Paris: Elsevier Press, 2001. 139−147.

[9] Staddon JN, Stinson DR, Wei R. Combinatorial properties and constructions of traceability schemes and flameproof codes. SIAM Journal on Discrete Math, 1998,11(1):41−53.

[10] Guruswami V, Sudan M. Improved decoding of reed-solomon and algebraic-geometry codes. IEEE Trans. on Information Theory, 1999,45(6):1757−1767.

[11] Peikert C, Shelat A, Smith A. Lower bounds for collusion-secure fingerprinting. In: Littleton R, ed. Proc. of the 14th Annual ACM-SIAM Symp. on Discrete Algorithms. Edmonton: ACM Press, 2003. 472−479.

附中文参考文献:

[3] 王彦,吕述望,徐汉良.一种二进制数字指纹编码算法.软件学报,2003,14(6):1172−1177. http://www.jos.org.cn/1000-9825/14/1172.htm

朱岩(1974－),男,黑龙江大庆人,博士生,主要研究领域为应用密码学,信息完全理论与技术.

杨永田(1939－),男,教授,博士生导师,主要研究领域为网络及信息安全理论与技术.

冯登国(1965－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为密码学,信息安全理论与技术.

© 中国科学院软件研究所 http://www.jos.org.cn