系列单错完整性指示码及其性能分析

陈 龙[1,2] 方新蕾[1] 王国胤[1,2]
(重庆邮电大学计算机科学与技术研究所 重庆 400065)[1]
(西南交通大学信息科学与技术学院 成都 610031)[2]

摘 要 实现细粒度的取证副本完整性检验是计算机取证的新需求,但是为每个取证对象生成一个独立Hash数据的完整性检验方法会产生大量的Hash检验数据,给Hash检验数据的存储与网络传输带来不利影响。在完整性指示编码思想的指导下,引入了能提高Hash检验数据抗篡改能力的平行分组关系设计需求——将Hash检验数据分组,其中任一组Hash数据均可从某一粒度完全指示全部数据的完整性。基于方阵与超方体的空间位置关系提出了平行分组式单错指示码,可实现几十倍、几百倍的压缩。分析了该类指示码在不同参数下的性能,结论表明该类指示码具有实用价值。

关键词 计算机取证,Hash,数据完整性,完整性检验,海量数据
中图法分类号 TP309 **文献标识码** A

One Error Integrity Indication Codes and Performance Analysis

CHEN Long[1,2] FANG Xin-lei[1] WANG Guo-yin[1,2]
(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[1]
(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China)[2]

Abstract Finer-grain integrity check to forensic copies becomes a new demand of computer forensics. Unfortunately, the traditional integrity check method, which generates a hash for every forensic object will produce huge amounts of hash data. It makes a great negative impact on storing hash data and transmitting them over network. This paper introduced the idea of integrity indication codes for compressing hash data and met the need of parallel grouping design which asks for any group hashes can indicate the integrity of all data. Then, one error integrity indication codes which can compress the data tenfold or hundredfold easily were presented and the performance of these codes was analyzed. The codes are practically valuable.

Keywords Computer forensics, Hash, Data integrity, Integrity check, Mass data

信息化时代的高速发展,使得计算机取证成为研究的一个热点领域[1,2],其发展非常迅速。当今,计算机取证面临的主要难题之一是海量数据处理。由于目前的取证调查过程过于手工化,导致取证的时效性与取证成本都难以控制[1]。

1 Hash完整性检验技术

Hash检验是计算机取证分析的重要手段之一[3,4]。利用传统的SHA-1等Hash函数生成的Hash数据可以高效地检验两个数据对象(文件、数据块)是否完全相同,在潜在证据快速搜索、数据相似性判断方面有突出的应用[4,5]。

计算机取证要求在进行取证复制的过程中计算并存储取证映像的Hash数据,从而保证取证分析用副本的完整性,这种应用称为完整性指示。但是取证映像(完全复制件)的完整性不能只停留在整体是否可靠、是否未被修改的层面上,因为偶然的数据变化就能影响全部数据的可用性、可信性。为了减小这类情况带来的影响,同时也从技术的角度避免以偶然错误为借口进行人为篡改,使用细粒度的数据完整性检验成为计算机取证的必然需求,即我们需要分别判断单个文件或小数据块是否具有完整性。例如采用物理存储块大小作为划分单位[6]。而如此一来,完整性检验面临着新问题——完整性检验Hash数据也成了大规模数据。

Vassil Roussev等人在考虑衡量海量数据之间的相似性时首先意识到了Hash数据的大数据量问题[6]。Hash检验数据具有随机性,无法使用数据压缩技术进行压缩,海量检验数据会给存储及网络传输效率带来负面影响。例如,一个512GB硬盘的扇区级MD5 Hash数据将需要16GB的存储量,如果使用强度较高的SHA-256,则需要32GB。

借鉴纠错编码思想可以实现大量Hash数据的压缩,并且保持原Hash检验的强度不变[7]。纠错编码与Hash编码的编码设计、编码性质、分析方法都不相同,因而需要细粒度

到稿日期:2008-09-26 返修日期:2008-11-05 本文受国家自然科学基金(No. 60573068),重庆市自然科学基金(No. CSTC 2007BB2454)资助。
陈 龙(1970—),男,博士研究生,教授,CCF高级会员,主要研究方向为计算机取证、网络安全、智能信息处理,E-mail:chenlong@cqupt.edu.cn;方新蕾(1984—),女,硕士生,主要研究方向为信息安全;王国胤(1970—),男,博士,教授,主要研究方向为数据挖掘、粗糙集、粒计算、知识技术、智能信息安全。

数据完整性指示编码理论与完整性检验新方法[7]。文献[7]分析了完整性指示编码的性质，并设计出了一种具有很高压缩率的组合码方案。组合码的不足在于错误放大率高。另外，考虑到 Hash 检验数据有可能受到破坏，设想存在这样几组独立的 Hash 数据：单独由某一组 Hash 数据在一中间粒度即可完全指示所有数据的完整性。那么，只需将各组 Hash 数据分离存放并由不同的人保管。在某组 Hash 数据遭破坏或被伪造时，其余 Hash 数据仍可以指示全部数据的完整性。具备这种能力的 Hash 分组称为平行分组。

根据降低错误放大率和实现平行分组这两个需求，本文设计了具有平行分组特性的超方体单错指示码。该码在单错情况下可轻易实现上百倍的压缩，在保持较高压缩率的同时，具有较低的错误放大率，并且实现完整性数据的分离式存储，单独一组 Hash 数据即可指示全部数据的完整性。

2 完整性指示编码与 Hash 压缩思想

2.1 完整性检验编码思想

设 $X_1, X_2, X_3, X_4, X_5, X_6$ 表示 6 个数据对象，参照纠错编码的一致监督方程表达式[8,9]，得到 Hash 检验监督关系如式(1)：

$$\begin{cases} X_1+X_2+X_4=h_1 \\ X_1+X_3+X_5=h_2 \\ X_2+X_3+X_6=h_3 \\ X_4+X_5+X_6=h_4 \end{cases} \tag{1}$$

式(1)中的“+”表示将数据对象连接成一个数据流，“=”表示将左端的数据流进行单向 Hash 运算(如 SHA-1)，等式右端 h_1, h_2, h_3, h_4 表示 Hash 数据。

在需要进行完整性检验时，按式(1)重新生成 Hash 数据，与事先存储的 Hash 数据进行比较，以判断数据对象是否变化。

假设只有某一个数据对象出错(数据被篡改或因偶然因素发生变化等，均简称出错)，该监督方案可以准确指示出该数据对象。例如 h_1, h_2 不相符时，表示 X_1 出错。

完整性检验要求决不能将不相同的两个数据对象误判为相同(相同对象被误判为不同则属于可接受的错误)。如果出现错误个数无法区分的情况，只能按照存在最多个错误的情况对待。

基于纠错编码思想的完整性检验的方法称为完整性指示编码，其基本原则是不能将出错块判定为正常块。

2.2 完整性指示码的基本概念

定义 1(完整性指示码) 设需要检验完整性的数据对象有 n 个，若存在一种监督关系，使得生成并存储 m 个 Hash 数据，在对 n 个对象进行检验时能准确指示任意的 t 个出错对象。而在 $n \geqslant t+1$ 时存在 $t+1$ 个错误的组合无法准确指示，其中单个数据对象参与 Hash 运算的次数最多为 $k(k \geqslant 1)$。这种监督关系称为一个完整性指示码，记为 $[n,m,t,k]$。设计这种监督关系就是设计一个编码。传统的对每个需要检验的数据对象都单独使用一个 Hash 数据进行监督的编码方式，也是一种完整性指示码，记为 $[n,n,n,1]$。

码的压缩率 η 为数据对象数 n 和使用的 Hash 数据个数 m 之比，即

$$\eta=\frac{n}{m} \tag{2}$$

为方便分析编码的性质，将式(1)的监督关系表达为一个监督矩阵 $A[m,n]$，如表 1 所列。

表 1 监督矩阵实例

A		n					
		1	2	3	4	5	6
m	1	1	1	0	1	0	0
	2	1	0	1	0	1	0
	3	0	1	1	0	0	1
	4	0	0	0	1	1	1

定义 2 利用完整性指示码 $C=[n,m,t,k]$ 进行完整性检验时，若实际出现的错误数 e 大于编码设计时可准确指示的错误数 t，则可能出现将正常对象判定为出错对象的情况，即指示出错的对象大于 e，这种现象称错误放大。由于错误对象的分布不同，实际指示错误数也可能不同，考察 e 个错误对象的所有分布，可得其平均值。实际指示错误数和实际出错数的比值称错误放大率，记为 $\beta(e)$。由于码 C 能准确指示 t 个错误，出现 $t+1$ 个错误时最能体现码的基本特性，将 $e=t+1$ 时的 $\beta(e)$ 简记为 β，称为码 C 的基准错误放大率。

除了完整性指示码本身已表明的性能指标 t,k 外，完整性指示码的其他主要性能指标包括压缩率、错误放大率、编码(Hash 生成)计算复杂性、完整性检验复杂性。

3 超方体单错完整性指示码

3.1 超方体单错完整性指示码概念

定义 3 如果需要检验完整性的数据对象有 n 个，设 $n=r^k$(实际应用中取 $r^k \geqslant n$ 的参数即可)，那么 n 个数据对象可排成 k 维 r 阶的超方体结构。从任一维(第 d 维，$d=1,2,\cdots,k$)把 n 个数据对象分为 r 份分别进行监督，参与每个 Hash 计算的数据对象有 r^{k-1} 个。此监督方案共需要 rk 个 Hash 数据，形成超方体单错完整性指示码 $[r^k, rk, 1, k]$。

定理 1 超方体单错完整性指示码 $[r^k, rk, 1, k]$ 可准确指示单个错误。

证明：先证明单错情况下的指示能力。设出错数据对象对应向量 $X=(x_1, x_2, \cdots, x_k)$，任取另外一个数据对象对应向量 $Y=(y_1, y_2, \cdots, y_k)$，其中 $x_i, y_i (i=1,2,\cdots,k)$ 代表数据对象的坐标。

已知每个数据对象受到 k 个 Hash 监督，若 $x_j=y_j$ 则说明两个数据对象在第 j 维被同一个 Hash 监督。由于两个不同对象的空间坐标不会完全相同，因此它们至多同时被相同的 $k-1$ 个 Hash 监督，即至少存在一个 Hash 可以区分这两个数据对象。因此，当仅有一个错误时，该码能正确指示出错数据对象。

再证明多于一个错误的情况。设有两个数据对象出错，$X=(x_1, x_2, \cdots, x_k)$，$Y=(y_1, y_2, \cdots, y_k)$，取另一个数据对象 $Z=(x_1, x_2, \cdots, x_{k-1}, y_k)$，则 $k-1$ 个 Hash 无法区分对象 X 和 Z，1 个 Hash 无法区分对象 Y 和 Z，在数据对象 X 和 Y 出错的情况下，不能正确指示 Z 的出错情况，说明该码不能正确指示两个错误。

因此，超方体单错完整性指示码 $[r^k, rk, 1, k]$ 可准确指示单个错误。

3.2 Hash 生成与检验

Hash 的生成算法：将 n 个数据块从 0 到 $n-1$ 进行编号。将每个编号 i $(i=0,1,\cdots,n-1)$ 表示为具有 k 位的 r 进制数，

形如 $r_k,\cdots,r_2r_1$，对应一个 k 维向量。每个数据块 i（$i=0,1,\cdots,n-1$），分别参与 $h_{r_j,j}$（$j=1,2,\cdots,k$）的计算，得到 r 行 k 列的 Hash 矩阵，如表 2 所列。表 2 中每个 $h_{i,j}$ 都是由数据块序列矩阵第 i 行 j 列的 r^{k-1} 个数据块计算 Hash 数据得到的。

表 2 $[r^k,rk,1,k]$码 Hash 数据生成

hash	k	…	2	1
1	$h_{1,k}$	…	$h_{1,2}$	$h_{1,1}$
2	$h_{2,k}$	…	$h_{2,2}$	$h_{2,1}$
…	…	…	…	…
r	$h_{r,k}$	…	$h_{r,2}$	$h_{r,1}$

以 $k=2,r=3$ 的方阵结构为例加以说明。此时，9 个数据对象可排成 3 阶方阵，形成方阵单错完整性指示码[9,6,1,2]。将 9 个数据块从 0 到 8 进行编号，把编号表列为 2 位的 3 进制数，对每个数据块判断 $r_v=u$（$v=1,2,\cdots,k$）的成立情况，生成数据块序列矩阵。数据块序列矩阵生成后，对每个位置的 3 个数据块计算一个 Hash 数据，即得到 6 个监督 Hash 数据。

平行分组单错完整指示码[9,6,1,2]的方阵监督关系可表示为表 3 所列的监督矩阵 $A[6,9]$。

表 3 [9,6,1,2]码的监督矩阵

A		n								
		0	1	2	3	4	5	6	7	8
m	1	1	0	0	1	0	0	1	0	0
	2	0	1	0	0	1	0	0	1	0
	3	0	0	1	0	0	1	0	0	1
	4	1	1	1	0	0	0	0	0	0
	5	0	0	0	1	1	1	0	0	0
	6	0	0	0	0	0	0	1	1	1

从表 3 可以看出，每个数据对象参与 Hash 计算的次数为 2，参与每个 Hash 计算的数据对象有 3 个。任意两列中 1 的分布情况都不相同，前(后)3 行中 1 的分布涵盖到所有列。从该监督矩阵中可以看出该码可以准确指示单个错误，该监督关系是均匀遍历的。

Hash 检验算法：当一个数据块发生变化时，Hash 矩阵中每列有唯一的一个 Hash 数据不相等。Hash 表共有 k 列，将不相等的 Hash 数据所在的行号从左往右依次记为 $r_k,\cdots,r_2,r_1$。这些数字按此顺序排列表示了一个 k 位的 r 进制数，按照式(3)求得出错数据块编号 i 为

$$i=\sum_{j=1}^{k} r_j \cdot r^{j-1} \tag{3}$$

若多于 1 个数据块发生变化，设 Hash 表第 j 列有 x_j（$j=k,\cdots,2,1$）个 Hash 出错。$r_k,\cdots,r_2,r_1$ 分别从对应列的 $x_k,\cdots,x_2,x_1$ 个 Hash 中选取，对每种组合按式(3)计算，可以得到所有的出错数据块编号。

3.3 完整性指示码的性能

超方体单错完整性指示码的压缩率为

$$\eta=\frac{1}{k}r^{k-1} \tag{4}$$

取 $k=2,3,4,5$，在各种 r 值下得到这些组合的压缩率如图 1 所示。

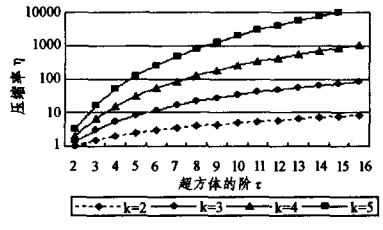

图 1 超方体单错完整性指示码的压缩率

从图 1 可以看出，数据对象个数越多，超方体单错完整性指示码的压缩效果越好，该码具有较高的数据压缩率。

下面考察基准错误放大率。

第 1 个出错点任意给定，那么 2 个出错点同在超方体的某个一维子空间上时，第 2 个出错点有 $k(r-1)$ 种选择，2 个错误均可准确指示出；若 2 个出错点同在某个二维子空间上，但不同在一维子空间上，第 2 个出错点有 $C_k^2(r-1)^2$ 种选择，2 个错将被放大，判断为 4 个错；依次类推，同在某个三维子空间但不同在低维子空间上的第 2 点有 $C_k^3(r-1)^3$ 种选择，将被判断为 2^3 个出错；依此类推，同在某个 k 维子空间但不同在任意的小于 k 维的子空间上的点有 $C_k^k(r-1)^k$，将被判断为 2^k 个出错。

所以超方体单错完整性指示码$[r^k,rk,1,k]$的基准错误放大率为：

$$\beta=\frac{1}{2\cdot(r^k-1)}\sum_{i=1}^{k}(2^i\cdot C_k^i\cdot(r-1)^i) \tag{5}$$

k 维超方体的基准错误放大率如图 2 所示。

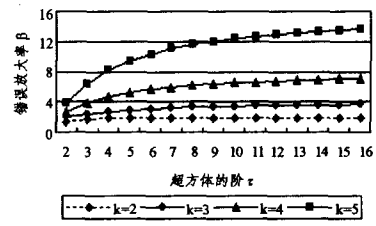

图 2 超方体单错完整性指示码的基准错误放大率

超方体单错完整性指示码的阶 r 可以根据实际需要选择任意的整数，综合图 1、图 2 的结果，$k=3,4$ 是实际应用中较为理想的选择，既有很高的压缩率，又有较低的错误放大率。

3.4 设计实例

设实际需要检验完整性的数据对象个数 $n=4096$，可采取不同的 r 和 k 的组合（$n=r^k$），计算其压缩率 η 和错误放大率 β。系列编码及性能如表 4 所列。可综合各方面性能和现实的需求，选择最合适的编码方案。

在实际应用中，显然 $r=4,k=6$ 只用 24 个 Hash 存储，从压缩的角度看，超方体码[4096,24,1,6]是最优的。综合考虑较高的压缩率和较低的错误放大率，则超方体码 [4096,32,1,4]，[4096,24,1,6]，[4096,48,1,3] 较好。

表 4 $n=4096$ 时的系列编码及性能

r	k	m	η	β
64	2	128	32	1.97
16	3	48	85.33	3.64
8	4	32	128	6.18
4	6	24	170.67	14.36
2	12	24	170.67	64.89

结束语 本文提出了系列单错完整性指示码，分析了其性能，设计出具有实用价值的高压缩率、较低错误放大率的完整性指示方案。该系列单错码可选取连续的整数作为超方体的阶，在实际应用中具有较好的灵活性。该系列码可轻易实现上百倍的 Hash 数据压缩，码的平行分组关系为实现 Hash 数据的分离存储提供了条件。下一步的研究主要考虑准确指示多个错误的编码方案。

参 考 文 献

[1] Richard III G G, Roussev V. Next-generation Digital Forensics

[J]. Communications of the ACM, 2006, 49(2): 76-80
[2] 王玲,钱华林. 计算机取证技术及其发展趋势[J]. 软件学报, 2003, 14 (9): 1635-1644
[3] [美]Steel C. Windows 取证:企业计算机调查指南[M]. 吴渝,唐红,陈龙,译. 北京:科学出版社, 2007
[4] Kornblum J. Identifying Almost Identical Files Using Context Triggered Piecewise Hashing[J]. Digital Investigation, 2006, 3 (s1): 91-97
[5] Chen Long, Wang Guoyin. An Efficient Piecewise Hashing Method for Computer Forensics[C]// International Workshop on Knowledge Discovery and Data Mining. Adelaide, Australia, January 2008: 635-638
[6] Roussev V, Chen Yixin, Bourg T, et al. md5bloom: Forensic File System Hashing Revisited[J]. Digital Investigation, 2006, 3 (s1): 82-90
[7] 陈龙,王国胤. 一种细粒度数据完整性检验方法[J]. 软件学报, 2009, 20(4): 902-909
[8] 靳蕃,陈志. 组合编码原理及应用[M]. 上海:上海科学技术出版社, 1995
[9] Bose R. Information Theory Coding and Cryptography[M]. China Machine Press, 2003

(上接第 94 页)

②目录服务器首先根据证书序列号查询证书是否存在。如果证书存在,则查询该证书的策略,确定是否允许 Bob 访问其隐蔽字段;如果证书不存在,则返回查询失败(每个证书的策略由证书主体自行定义,目录服务器只进行策略检查)。

③如果策略允许 Bob 访问证书的隐蔽字段,则用 Bob 的公钥加密随机数 r,并将证书、预映射值 $p_1, p_2, \cdots, p_n$($p_i = v_i \oplus r$)和加密后的结果 $Ep_B(r)$传给 B。

④Bob 首先计算预映射值对应的散列值 $a_1, a_2, \cdots, a_n$($a_i = H(p_i)$),再将计算出的散列值与证书中存储的散列值进行对比。如果两者一致,则说明传递过来的预映射值是可信的。然后 Bob 用私钥脱密 $Ep_B(r)$得到随机数 r,最后将预映射值 $v_i \oplus r$ 和随机数 r 进行异或得到该隐秘字段所对应的原文 v_i。

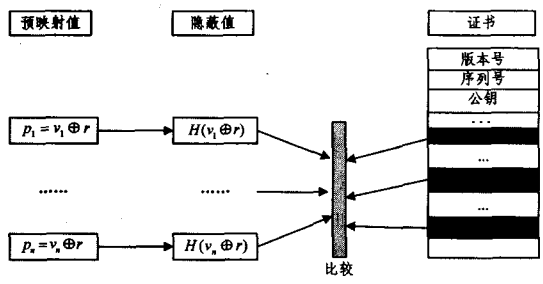

图 2 查询验证过程

4 安全性与效率分析

本文提出的位承诺协议的安全性取决于单向函数和公钥密码算法的强度。我们假定单向函数和使用的公钥密码算法是足够安全的,下面我们对第 3 节提出的位承诺协议的安全性和效率进行分析。

4.1 机密性

对证书中原始特征信息存储的机密性保证是基于单向散列函数和预映射值构造的强度。预映射值是将一个随机位串与特征属性值异或后产生的(如图 1 所示),那么攻击者在确定一个预映射值对应的原始信息时将需比较 2^{128} 次,这足以对抗穷尽攻击。

对证书中原始特征信息传输的机密性保证是基于公钥密码算法。采用公开密钥长度大于 2304 位能够保护随机数 r 足够安全[7]。由于 r 在查询过程中被加密,攻击者即使得到预映射值 $v_i \oplus r$ 也无法还原原始特征信息 v_i。

4.2 不可否认性

Alice 对隐蔽字段承诺的结果是对预映射值 $p_i = v_i \oplus r$ 进行单向函数变换得到的,只要单向函数足够安全,Alice 想伪造预映射值 p' 使得 $H(p_i') = H(p_i)$,在计算上不可行。Bob 通过计算预映射值所对应的散列值并与证书中的散列值进行比较,来验证隐蔽字段的值是否真实可信。

4.3 访问控制

X. 509V3 格式的数字证书提供了丰富的扩展字段,证书主体可以灵活定义自己的访问控制策略,来实现用户的选择性揭示和内容的选择性揭示。由于每个证书主体身份、职务级别的不同,其访问控制策略可能千差万别,因此由证书主体自行定义策略,而不是由签发中心集中定义,可以提高策略定义的灵活性。由目录服务器代替用户进行策略检查,可以实现对证书的统一访问控制。

4.4 性能

以证书内仅存储一个隐秘字段为例,忽略异或运算的计算量,证书申请过程比普通的证书申请过程仅多出一次哈希运算;忽略策略查询所花费的时间,证书查询验证过程比普通的证书查询验证过程多出一次加密运算和一次哈希运算。

结束语 本文针对证书隐私泄露问题,提出了基于位承诺的数字证书敏感信息保护方案。该方案能够阻止非法用户恶意收集他人证书,实现证书敏感信息的隐藏和敏感信息的选择性揭示。针对合法用户恶意收集他人证书及其隐私的问题,将是我们下一步研究的重点。

参 考 文 献

[1] Park J S, Sandhu R. Smart Certificates: Extending X. 509 for Secure Attribute Services on the Web[C]// 22th National Information Systems Security Conference. Crystal City, Virginia, October 1999
[2] Renfro S G. VeriSign CZAG: Privacy Leak in X. 509 Certificates [C]// 11th USENIX Security Symposium. San Francisco, California, August 2002
[3] Persiano P, Visconti I. User Privacy Issues Regarding Certificates and the TLS Protocol[C]// 7th ACM Conference of Computer and Communications Security. Athens, Greece, November 2000
[4] Naor M. Bit Commitment Using Pseudorandomness[C]// Advances in Crypology-Crypto 89, Lecture Notes in Computer Science. Vol. 435, New York: Spring-Verlage, 1990: 128-137
[5] Fischlin M, Fischlin R. Efficient Non - malleable Commitment Schemes. CRYPTO, 2000: 413-431
[6] 王永钊. 信任协商过程中证书上敏感信息的保护[D]. 天津:天津大学, 2006
[7] Schneier B. 应用密码学[M]. 吴世忠,等. 北京:机械工业出版社, 2000