基于加权概率模型的自适应散列值长度 Hash 算法

王杰林 [1,2]，高金定 [1,3]

（1 湖南涉外经济学院信息与机电工程学院，湖南 长沙 410205；
2 湖南遥昇通信技术有限公司，湖南 长沙 410600；
3 中南大学物理与电子学院，湖南 长沙 410083）

摘　要：针对标准 Hash 算法散列值长度不能根据攻击强度自适应调整的不足，利用分布函数区间编码方法，定义了加权概率质量函数和加权分布函数，提出了一种基于加权概论模型 Hash 算法并分析了算法碰撞的理论极限，结合分段、异或运算和加权系数的非线性轮函数给出了算法实现步骤并进行了实验测试。结果表明：本文提出的 Hash 算法散列值长度可以自定义并且可以根据攻击强度自适应调整，可达到哈希碰撞的理论极限。该算法可以广泛应用于数字签名、文件校验以及数据传输校验等领域。

关键词：单向散列函数；Hash 函数；加权概率模型；自适应散列值长度

中图分类号：TP309.7

文献标志码：A

Hash Algorithm with Adaptive Hash Value Length Based on Weighted Probability Model

WANG Jielin [1, 2], GAO Jinding [1,3]

1 School of information and mechanical and electrical engineering, Hunan International Economics University, Changsha, 410205, China

2 Hunan YESINE Communication Technology Co., Ltd., Changsha, 410600, China

3 School of physics and electronics, Central South University, Changsha, 410205, China

Abstract: In view of the deficiency that the hash length of standard hash algorithm cannot be adjusted adaptively according to the attack strength, this paper defines the weighted probability quality function and weighted distribution function using the interval coding method of distribution function, and proposes a hash algorithm based on weighted generality model. The theoretical limit of algorithm collision is analyzed, the algorithm implementation steps are given and the experimental test is carried out in combination with the nonlinear round function for weighting coefficient, exclusive OR (XOR) and encode by segment. The results show that the hash length of the proposed hash algorithm can be customized and adjusted adaptively according to the attack strength, which can reach the theoretical limit of hash collision. The algorithm can be widely used in digital signature, file verification and data transmission verification.

Keywords: One way hash function, Hash function, weighted probability model, Adaptive hash value length

1　引言

Hash 算法可以将任意长度的消息压缩到固定长度的消息摘要，在数字签名、文件校验、信息加密等领域得到了广泛的应用。目前标准的 Hash 算法主要包括 MD（Message-Digest）[1][2][3]信息摘要算法、SHA（Secure Hash Algorithm）[4][5]安全散列算法、基于格的哈希算法[6]以及 MAC（Message Authentication Code）[7][8]消息认证码等几大系列。其中 MD 信息摘要算法主要包括 MD2、MD4 和 MD5 等系列，主流的 SHA 安全散列算法主要有 SHA-1 和 SHA-2(SHA-224，SHA-256，SHA-384，SHA-512)系列，以及 SHA-3(KECCAK 算法[9])。MAC 算法主要是 HMAC 含有密钥的散列函数算法，是在原有的 MD 和 SHA 算法的基础上添加了密钥，主要有 HmacMD 系列（HmacMD2，HmacMD4，HmacMD5）以及 HmacSHA 系列（HmacSHA1，HmacSHA224，HmacSHA256，HmacSHA38，HmacSHA512），摘要长度与原有的 MD 系列和 SHA 系列一致。摘要长度不同，算法内部的运算结构不相同，使得系统需集成大量不同摘要长度的哈希算法来适应安全要求，造成系统资源和成本浪费。随着量子计算的到来，计算性能大幅度提升，安全系统需采用更长消息摘要的哈希算法或更可靠运算结构的哈希算法才能保障安全。然而消息摘要越长对网络传输、校验运算和存储也带来了巨大负担，而且复杂运算结构的哈希算法存在运算上的负担。

本文基于概率分布函数[10]定义了加权概率分布函数和加权概率模型，通过加权概率分布函数定义 Hash 算法的编码结果长度，给出了一种全新的 Hash 算法。该算法具备如下特点：

通信作者：高金定, E-mail: jdgao@qq.com

基金项目：国家重点研发计划项目（2018YFC0603202）；国家重点研发计划项目（2018YFC0807802）；湖南省自然科学基金面上项目（2019JJ40154）；湖南省教育厅优秀青年项目（20B337）。

1、线性地计算消息中每个符号对应的概率区间，构造了加权系数的非线性轮函数，使得消息的长度、符号的概率和符号的排列顺序不同则编码结果不同，消息中出现差异均会造成“雪崩效应”。

2、可自定义编码结果的长度，安全系统无需集成大量不同摘要长度的哈希算法，节约系统资源和降低成本。

3、比特为单位的线性编码过程，适用于比特流，可随时终止或启动编码运算。

相较于目前主流的 MD、SHA 和 MAC 固定散列值长度的 Hash 算法，本文算法摘要长度可以自定义，则可根据信息安全级别自适应调整，既能保障信息安全，又能有效降低消息网络传输、校验运算和存储的压力。

2 加权概率模型 Hash 算法

令信源序列$X=(X_1,X_2,\dots,X_i,\dots,X_n)$是有限个值或可数个可能值的离散序列，$X_i\in A=\{0,1,2,\dots,k\}$。于是对于$A$中一切数值有概率空间：

$$\begin{bmatrix}X_i\\P\end{bmatrix}=\begin{bmatrix}0 & 1 & \dots & k\\ p(0) & p(1) & \dots & p(k)\end{bmatrix}$$

由于随机过程必须转移到某个符号，所以在任意时刻有：

$$\sum_{X_i=0}^{k}p(X_i)=1,\qquad 0\le p(X_i)\le 1$$

于是，任意符号X_i的分布函数为：

$$F(X_i)=\sum_{s\le X_i}p(s) \tag{2-1}$$

$p(0)\le F(x)\le 1,\ s\in A$。

定义 2.1 设离散随机变量X，$X\in A=\{0,1,\dots,k\}$，$\mathrm{P}\{X=a\}=p(a)(a\in A)$，加权概率质量函数为$\varphi(a)=r\mathrm{P}\{X=a\}=rp(a)$，$p(a)$为的概率质量函数，$0\le p(a)\le 1$，$r$为权系数，且

$$F(a)=\sum_{i\le a}p(i) \tag{2-2}$$

若$F(a,r)$满足$F(a,r)=rF(a)$，则称$F(a,r)$为加权累积分布函数，简称加权分布函数。显然，所有符号的加权概率之和为$\sum_{a=0}^{k}\varphi(a)=r$。

令离散信源序列$X=(X_1,X_2,\dots,X_n)$，$X_i\in A$，且令$F(X_i-1)=F(X_i)-p(X_i)$，序列X的加权分布函数记为$F(X,r)$。当$n=1$时：

$$F(X,r)=rF(X_1-1)+rp(X_1)$$

当$n=2$时：

$$F(X,r)=rF(X_1-1)+r^2F(X_2-1)p(X_1)+r^2p(X_1)p(X_2)$$

当$n=3$时：

$$F(X,r)=rF(X_1-1)+r^2F(X_2-1)p(X_1)+r^3F(X_3-1)p(X_1)p(X_2)+r^3p(X_1)p(X_2)p(X_3)$$

令$\prod_{j=1}^{0}p(X_j)=1$，类推得：

$$F(X,r)=\sum_{i=1}^{n}r^iF(X_i-1)\prod_{j=1}^{i-1}p(X_j)+r^n\prod_{i=1}^{n}p(X_i) \tag{2-3}$$

将满足(2-3)的加权分布函数的集合定义为加权概率模型，简称加权模型，记为$\{F(X,r)\}$。若$X_i\in A=\{0,1\}$，则称$\{F(X,r)\}$为二元加权模型。令：

$$H_n=F(X,r) \tag{2-4}$$

$$R_n=rp(X_1)rp(X_2)\dots rp(X_n)=r^n\prod_{i=1}^{n}p(X_i) \tag{2-5}$$

$$L_n=H_n-R_n \tag{2-6}$$

其中$X_i\in A,n=1,2,\dots$。当$r=1$时：

$$F(X,1)=\sum_{i=1}^{n}F(X_i-1)\prod_{j=1}^{i-1}p(X_j)+\prod_{i=1}^{n}p(X_i) \tag{2-7}$$

由(2-4)(2-5)(2-6)可得$H_n=F(X,1)$，即算术编码（区间编码）[11][12]是基于$r=1$时加权分布函数的无损编码方法。

因X_i必须取A中的值，所以$p(X_i)>0$。显然(2-4)(2-5)(2-6)为区间列，$[L_i,H_i)$是信源序列X在时刻$i(i=0,1,2,\dots,n)$变量X_i对应的区间上下标，$R_i=H_i-L_i$是区间的长度。根据(2-4)(2-5)(2-6)，设$i=0$时$R_0=H_0=1$，$L_0=0$，于是$i=1,2,\dots,n$时用运算式为：

$$\begin{aligned} R_i &= R_{i-1}\varphi(X_i) \\ L_i &= L_{i-1}+R_{i-1}F(X_i-1,r) \\ H_i &= L_i+R_i \end{aligned} \tag{2-8}$$

通过(2-8)对信源序列X进行加权概率模型编码运算，L_n为实数，是加权概率模型编码结果。L_n通过进制转换得到二进制序列。

3 加权概率模型信息熵与碰撞极限分析

3.1 加权概率模型信息熵

设离散无记忆信源序列$X=(X_1,X_2,\dots,X_n)(X_i\in A,A=\{0,1,2,\dots,k\})$，当$r=1$时，$\varphi(X_i)=p(X_i)$。由香农信息熵[13][14][15]定义，$X$的熵为：

$$H(X)=-\sum_{X_i=0}^{k}p(X_i)\log_{k+1}p(X_i) \tag{3-1}$$

当$r\neq 1$时，定义具有概率$\varphi(X_i)$的随机变量X_i的自信息量为：

$$I(X_i)=-\log_{k+1}p(X_i) \tag{3-2}$$

设集合$\{X_i=a\}(i=1,2,\dots,n,a\in A)$中有$c_a$个$a$。当$r$的值确定，信源序列$X$的总信息量为：

$$-\sum_{a=0}^{k}c_a\log_{k+1}p(a)$$

于是平均每个符号的信息量为：

$$-\frac{1}{n}\sum_{a=0}^{k}c_a\log_{k+1}p(a)=-\sum_{a=0}^{k}p(a)\log_{k+1}p(a)$$

定义 3.1 令$H(X,r)$为：

$$\begin{aligned} H(X,r) &= -\sum_{a=0}^{k}p(a)\log_{k+1}\varphi(a) \\ &= -\log r-\sum_{a=0}^{k}p(a)\log_{k+1}p(a) \\ &= -\log r+H(X) \end{aligned} \tag{3-3}$$

根据定义 3.1，在r的值确定时，通过加权概率模型编码后的二进制长度为$nH(X,r)(bit)$。最简单的信源序列为二进制序列，设二进制信源序列X的比特长度为n，X中符号 0 和符号 1 有概率$p(0)$和$p(1)$，且经加权概率模型编码后得到长度为$L(bit)$的序列。当$k=1$时由(3-3)可得：

$$-n\log_2 r+nH(X)=L \tag{3-4}$$

其中$H(X)$序列X的信息熵，即$H(X)=-p(0)\log_2 p(0)-p(1)\log_2 p(1)$，化简(3-4)得：

$$\begin{aligned} r &= 2^{\left(H(X)-\frac{L}{n}\right)} \\ &= 2^{\left(-p(0)\log_2 p(0)-p(1)\log_2 p(1)-\frac{L}{n}\right)} \end{aligned} \tag{3-5}$$

根据无失真编码定理，$H(X)$是离散无记忆信源序列X的无失真编码极限，所以当$H(X,r)\geq H(X)$时加权模型函数$F(X,r)$可无失真还原信源序列X。当$H(X,r)<H(X)$时，加权模型函数$F(X,r)$无法还原信源X，即当$L<nH(X)$时编码结果L_n无法还原信源X。

由(3-4)和(3-5)可得，当$H(X)>L/n$时，有$r>1$，$H(X,r)<H(X)$，于是满足(3-5)且$r>1$的加权模型函数$F(X,r)$均是单向散列函数（Hash 函数）。

3.2 碰撞极限分析

定理 3.2 任意二进制序列经加权概率模型哈希算法得出的散列值中符号 0 和符号 1 的概率均等。

证明 设二进制序列经加权概率模型哈希算法得出的散列值的比特长度为L，散列值的二进制序列记为Y，其信息熵为$H(Y)=$

$-p(0)\log_2 p(0)-p(1)\log_2 p(1)$。根据定义 3.1，$nH(X,r)=-n\log_2 r+nH(X)$（$n$为二进制序列$X$的比特长度），所以$LH(Y)=-n\log_2 r+nH(X)$。当且仅当$H(Y)=1$时，(3-4)成立，即$r$满足(3-5)。否则$r$不满足(3-5)。又当且仅当$p(0)=p(1)=0.5$时，$H(Y)=1$，所以序列$Y$中符号 0 和符号 1 的概率均等。

根据定理 3.2，可得本文哈希算法得出的散列值中符号等概率。设散列值的比特长度为L，则取值空间范围为$\{0,1,\dots,2^L-1\}$。令$d=2^L$，根据哈希碰撞（或“生日攻击”）[2][4][16][17]概率可得N次试验本算法的哈希碰撞概率为：

$$p(N,d)\approx 1-e^{\frac{-N(N-1)}{2d}} \tag{3-6}$$

3.3 非线性的分段迭代运算和加权系数的轮函数

常见 Hash 算法、AES[18][19]和 DES[20]对称加密算法均采用了非线性字节替换的轮函数来消除线性相关性，通常称为 S 盒。本节基于论函数思想，采用分段迭代和异或运算来消除线性相关性，并基于(3-5)构造了加权系数的非线性轮函数。

（1）序列X分段迭代和异或运算

序列X的比特长度为n，设每段的比特长度为m^2，于是序列X被线性地分割为$\lceil n/m^2\rceil$段。令$m=16$，$j=1,2,\dots,\lceil n/m^2\rceil$，$v$为每段对应的比特数，显然$j=\lceil n/m^2\rceil$时$v=n-(\lceil n/m^2\rceil-1)*m^2-1$，$j<\lceil n/m^2\rceil$时$v=m^2$。随机生成$m^2$个比特存于$16*16$的二维表中，如表 1，其中$x$和$y$为行列下标。

表 1 随机生成m^2个比特的二维表

x \ y	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	1	0	1	0	1	1	0	1	0	0	0	0	1	0	0
1	1	1	0	0	0	1	0	1	0	0	1	1	1	1	1	1
2	0	1	1	1	0	1	1	0	1	0	0	1	0	1	0	0
3	1	0	0	1	0	0	1	0	1	1	1	0	1	1	0	1
4	1	1	0	1	1	1	1	0	0	0	1	1	0	1	0	1
5	0	0	0	1	0	1	0	0	1	1	0	0	0	0	1	0
6	1	0	1	1	1	1	0	1	0	0	0	1	1	0	0	1
7	1	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1
8	0	0	0	0	1	1	1	1	1	0	1	0	1	1	1	0
9	0	1	1	1	1	1	0	0	1	0	1	0	1	1	0	0
10	0	0	0	0	1	1	0	1	0	0	1	0	0	0	1	1
11	0	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0
12	0	1	1	0	0	0	1	1	0	0	1	1	0	0	0	0
13	1	1	1	0	0	0	0	0	1	0	1	1	0	0	0	1
14	0	0	1	1	1	0	1	1	0	1	1	1	0	1	0	1
15	0	1	0	1	0	1	1	0	1	0	0	1	0	1	0	1

由于第j段中符号X_i的位置为i（$i=1,2,\dots,m^2$），表 1 的查表运算式为：

$$y=x=(i+X_i)\bmod m,\ \ y=\lfloor (i+X_i)/m\rfloor \tag{3-7}$$

位置(x,y)因X_i随机，所以比特值$f(x,y)$未知。X_i与$f(x,y)$进行异或运算，异或运算后的二进制序列记为Y：

$$Y_i=X_i\oplus f(x,y) \tag{3-8}$$

异或运算后计算序列Y中符号 0 和符号 1 的概率，并代入(3-5)计算当前段的r值。

（2）加权系数的非线性轮函数

设定L后，每段二进制序列进行加权概率模型编码时，若r不随i（$i=1,2,\dots,m^2$）变化，则称r为静态加权系数；若r随i变化，则称r为动态加权系数，记为$r(i)$。随机生成m^2个 0 到 255 的整数存于$16*16$的二维表中，如表 2，其中x和y为行列下标，x和y的计算式同(3-7)。经过查表 2 得到数值$g(x,y)$，$r(i)$的非线性轮函数为：

$$r(i)=2^{(H(X)-L/v)}-\frac{g(x,y)}{10^s} \tag{3-9}$$

(3-9)中s可取值大于 3 的整数，实际值根据计算机的运算精度而定，本文实验中$s=4$。当$r<2^{(H(X)-L/n)}$时编码结果将超过L个比特。设编码结果为$L_z=L+t$比特，则多出了t个比特，将第j段散列值的二进制序列记为Z，Z为L_n通过进制转换后的二进制序列。因$\frac{g(x,y)}{10^s}$趋于 0，所以不存在$t>L$情形。令$l=1,2,\dots,t$，则后t个比特与前L个比特进行异或运算：

$$Z_{l-1}=Z_{l-1}\oplus Z_{L+l-1} \tag{3-10}$$

(3-10)中Z_{l-1}和Z_{L+l-1}为序列Z的第$l-1$个和第$L+l-1$个二进制符号。

表 2 随机生成m^2个 0 到 255 整数的二维表

x \ y	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	143	254	212	99	131	164	146	48	66	137	58	219	171	251	193	50
1	94	43	86	32	68	48	217	36	242	56	183	150	73	48	82	213
2	143	12	156	62	127	28	182	213	23	235	49	190	164	108	121	92
3	178	59	73	121	131	85	96	153	126	112	204	252	183	237	213	249
4	220	75	88	129	35	170	182	14	162	35	179	67	94	67	132	38
5	122	96	58	232	187	17	3	218	35	105	75	82	137	114	160	2
6	140	157	84	106	186	209	79	221	213	41	37	228	47	74	141	115
7	189	203	124	9	33	28	206	90	7	206	30	230	136	89	149	128
8	221	216	214	146	133	135	208	157	79	85	45	84	1	23	95	97
9	62	148	184	231	37	249	204	152	146	227	97	35	113	170	167	197
10	107	63	233	178	99	118	249	249	155	99	152	204	23	223	214	222
11	199	86	226	230	84	90	249	64	214	173	94	242	209	202	229	147
12	146	55	115	158	202	125	19	58	180	187	127	151	24	135	11	125
13	186	2	102	215	97	187	126	146	39	91	208	115	54	174	197	18
14	41	69	133	125	6	111	70	193	192	249	161	224	78	238	178	138
15	218	250	158	88	134	71	46	8	248	233	103	224	141	74	174	5

基于(3-9)经加权概率模型编码得到第j段对应的实数L_v，然后通过(3-10)可得当前段的散列值Z。

4 二进制加权概率模型 Hash 算法实现

4.1 自定义散列值长度的算法流程

L为自定义散列值的比特长度，采用加权概率模型对长度为n的二进制信源序列X编码步骤如下。

（1）初始化参数，$p=c=L_0=0$，$H_0=R_0=1$，$i=t=L_z=0$，$m=16$，$j=l=1$，$s=4$，$v=m^2-1$，$T=x=y=\varphi(0)=\varphi(1)=0$；

（2）将序列X线性地分割成$\lceil n/m^2\rceil$段；

（3）当$j=\lceil n/m^2\rceil$时$v=n-(\lceil n/m^2\rceil-1)*m^2-1$；

（4）获取序列X的第j段二进制序列，共v个比特；

（5）获取第j段第i个符号X_i；

（6）计算x和y，$x=(i+X_i)\bmod m$，$y=(i+X_i)/m$；

（7）查表 1 获取$f(x,y)$；

（8）$Y_i=X_i\oplus f(x,y)$；

（9）$i=i+1$，若$i\le v$，重复（5）到（9）；

（10）$i=0$，统计序列Y中符号 0 的个数c，得出$p=\frac{c}{v}$和$H(Y)=-p\log_2 p-(1-p)\log_2(1-p)$；

（11）计算x和y，$x=(i+X_i)\bmod m$，$y=(i+X_i)/m$；

（12）查表 2 获取$g(x,y)$；

（13）计算$r(i)$、$\varphi(0)$和$\varphi(1)$，$r(i)=2^{(H(Y)-L/v)}-\frac{g(x,y)}{10^s}$，$\varphi(0)=r(i)p$和$\varphi(1)=r(i)(1-p)$；

（14）加权模型编码运算，若$Y_i=0$，$R_i=R_{i-1}\varphi(0)$，否则$L_i=L_{i-1}+R_{i-1}\varphi(0)$且$R_i=R_{i-1}\varphi(1)$；

（15）$i=i+1$，若$i\le v$，重复（11）到（15）；

（16）$L_v=L_v+T$，$T=L_v$；

（17）L_v经进制转换为二进制序列Z；

（18）统计Z的比特长度L_z并计算t，$t=L_z-L$；

（19）当$l\le t$时，$Z_{l-1}=Z_{l-1}\oplus Z_{L+l-1}$

（20）$l=l+1$，重复（19）到（20）；

（21）$i=0$，$j=j+1$；

（22）若$j\le\lceil n/m\rceil$，重复（3）到（22）；

（23）结束编码，输出Z，Z为散列值（hash value）。

4.2 自适应攻击强度的系统流程

将 4.1 节自定义散列值长度的算法记为 WPMHA（Weighted Probability Model Hash Algorithm）。下面以信息传输校验为例基于 WPMHA 给出自适应攻击强度的系统流程。

发送端的系统流程如图 1 所示，步骤如下：

（1）初始化参数，定义增幅d（d可以是一定范围内的随机数，比如属于$\{1,2\ldots,m\}$的随机数，也可以是自定义大于 0 的整数），校验次数$i=3$，L的值（比如$L=512$，L为 WPMHA 算法中散列值的比特长度），$sign=0$；

（2）发送端利用 WPMHA 得出信息（信息标记为 Data）的散列值 A，并发送散列值 A；

（3）发送端发送信息 Data，等待接收端返回$sign$；

（4）超时或$sign=1$时$i=i+1$，$L=L+d$；

（5）$i>0$时重复（2）到（5），否则结束。

接收端的系统流程如图 2 所示，步骤如下：

（1）初始化参数，校验次数$i=3$，$sign=0$；

（2）接收散列值 A；

（3）接收信息 Data；

（4）根据散列值 A 的长度，基于 WPMHA 计算出 Data 的散列值 B；

（5）当 A 等于 B，回复成功标识$sign=0$并结束，否则丢弃 Data，$j=j-1$；

（6）$j>0$时回复失败标识$sign=1$，否则结束。

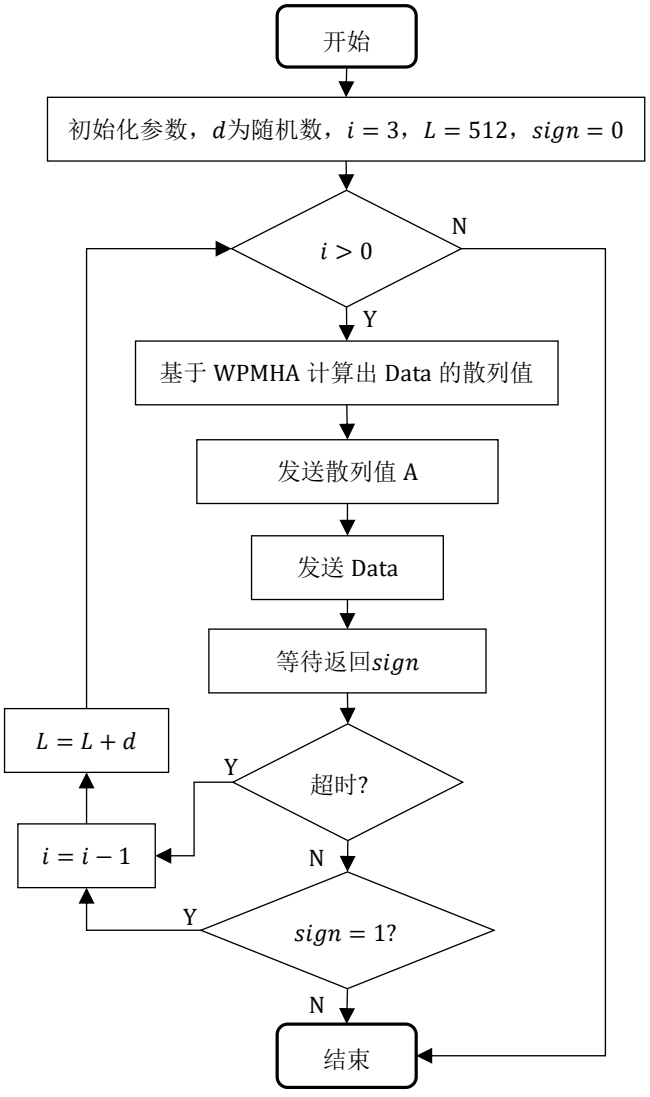

图 1 发送端系统示意图

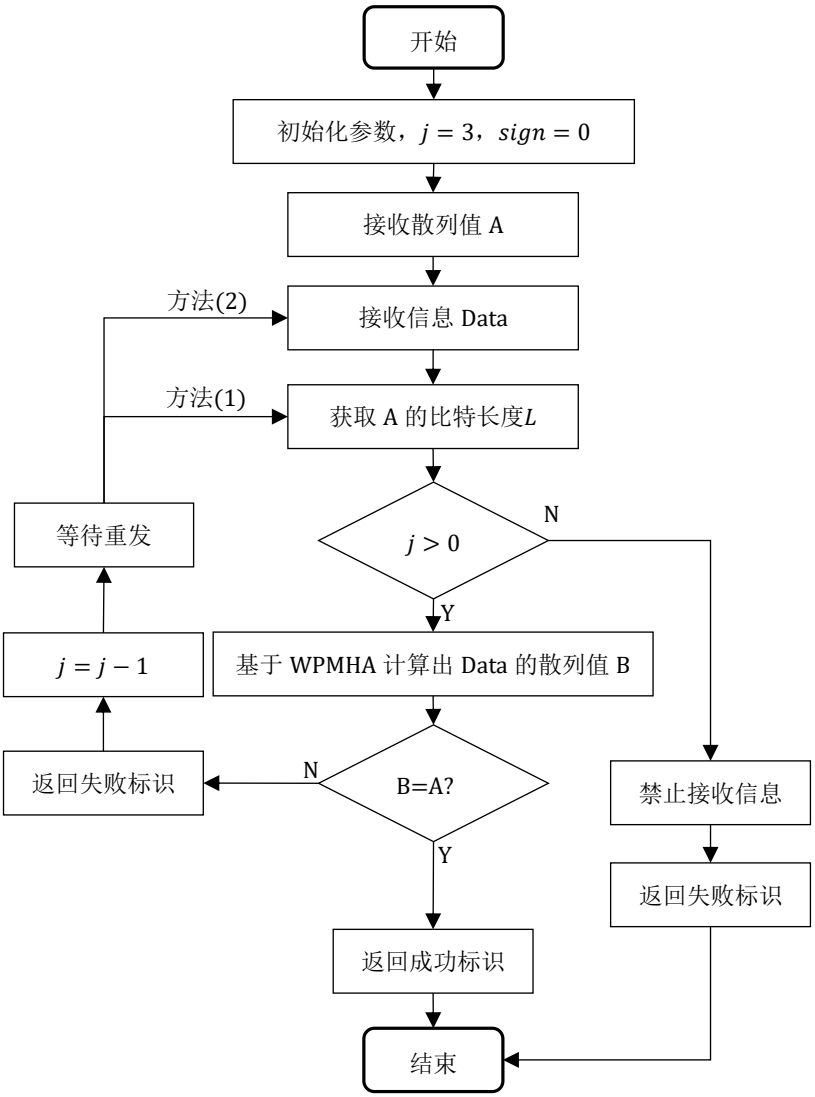

图 2 接收端系统示意图

图 2 中提出了方法(1)和方法(2)，是根据系统安全要求而定。比如，当$j=2$时 Data 无需重新传输校验，当$j=1$时必须重新传输 Data，目的是减少重传负担。

4.3 实验

（1）自定义散列值长度$bytelength$

输入：abcdefghijklmnopqrstuvwxyz123456789

输出$bytelength$分别为 4,8,15,20,32（单位为字节）的散列值。

表 3 自定义不同长度的散列值

$bytelength$	散列值
4byte	b21527d1
8byte	7a63428361901e1c
15byte	d99067a5ebd9dd1f5eefc63bddcfa
20byte	cd5c2c83aa1cf1468590303811895af11b f06943
32byte	233652b6ff6365c634bd1f7dbe58967ebd e64b849b5beed3a145f4b83efe365

（2）输入数据不同实验得出的散列值不同，散列值中符号 0 的概率$p(0)$趋近于 0.5

表 4 散列值长度为 128 位（16 字节）时$p(0)$的值

输入	散列值	$p(0)$
0123456789	9083a6bf012fe055381bb0d1cee1b307	0.53125
abcdefghijklmnopqrstuvwxyz123456789	8aa3404995ca1e7c33a3785df04c9059	0.54687
0	14e5c8666b999536301be0347d0e625b	0.53125
AAAAAAA	4a96d54d2f5dea8f0cb955643eaaf1fe	0.44531
加权概率模型 Hash 算法	b45dd174a5dcc8bf04827f868366515f	0.49219

表 5 散列值长度为 512 位（64 字节）时$p(0)$的值

输入	散列值	$p(0)$
0123456789	bfb98538a61b26ecc102452b00b9cb63 b25be606c2707dcd0b7dbf6f411593cd a1025c280f78cc704f345d7264d4fa16 c7a5a3b9f518db58e6d56fddbcf2e107	0.49218
abcdefghijklmnopqrstuvwxyz123456789	a84a4fb59a0a1d7c7169a40c8e28c7f4 b601634e5e073eee47e20ad270fcf7be 4bfde766b4d33450d69cafc26c4fad7b ecbc3442de5a9a776dd50890a6a3db06	0.48046
0	b8ce085f673a3de4f67196de9b8edf78 6d07606234baa19b5c2899040a11175e 56c8f52e9287a87224a8dc111cddc6f1 fae3597c7bb20795235de3a014c223ab	0.50585
AAAAAAA	f1fbff90d76d3750840ae113c1c6e648 035e5d123ae9c5854b64b98cab90c3c0 b364c2dcaf26bd12bd6913fb38d8ec2b 9227576c0f3c9ed5a1897febbeaaf8f7	0.47851
加权概率模型 Hash 算法	0964b505ce0da7e49dc15a0c0163fa47 c979fbe51b60c0efb96111a299155e22 c3b0e1e0e064dd12f5db9ab034c190bd f1d6fe0acce1612f6707484b41c7227e	0.51953

从实验结果可得出，本文方法的哈希长度可自定义，且散列值中符号 0 和符号 1 的概率趋近于均等。WPMHA 作为自适应攻击强度安全系统的核心，仿真实验基于 TCP/IP 协议 Socket 实现，数据传输时将 Data 和列值 A 进行随机比特错误处理。实验设定d为 1-128 随机数，$i=j=3$，每次仿真L的初始值为 512 位，当仿真次数不小于 10^5时得出实验结论为：基于图 1 和图 2 的安全系统可实现自适应攻击强度。

4.4 安全分析

分段迭代、异或运算和轮函数是抵御线性攻击和差分攻击常用手段。根据 4.1 节的编码流程，首先，因$X_i=0$和$X_i=1$查表 1 或表 2 时二维坐标(x,y)是随机的，所以$f(x,y)$和$g(x,y)$的值是随机的。其次，序列X中符号 0 的概率是固定的，但每段二进制序列中符号 0 的概率p存在差异，p的精度决定了取值空间，于是$H(Y)$不可知。因$2^{(H(Y)-L/v)}$和$\frac{g(x,y)}{10^s}$均未知，所以L_v的二进制长度不可知。又因$L_v=L_v+T$，即编码结果以加法的方式分段迭代运算，使得L_v在运算时位变化存在随机性。于是，可分析出充分条件为：编码序列E（E为任意非X的二进制序列）第i个符号时，$f(x,y)$、$g(x,y)$、$r(i)$、R_i和L_i与编码X_i时保持一致，即可产生相同的散列值。充分条件存在不确定性，可分析每一段编码时满足某一充分条件的概率。

（1）$X_i\in\{0,1\}$，$f(x,y)$仅作用于符号X_i，每段有v个比特且每个符号均正确选择$f(x,y)$的概率为$p(x,y,f)=\frac{1}{2^v}$。

（2）$g(x,y)$仅作用于权系数$r(i)$，根据（1）每个符号均正确选择$g(x,y)$的概率为$p(x,y,g)=\frac{1}{2^v}$。

（3）设p的精度为u位二进制，可得每段二进制序列中符号 0 的概率p存在2^u种可能值，因 $r=2^{(H(Y)-L/v)}$，所以$p(r)=\frac{1}{2^u}$。

（4）T和L_v的二进制位数为L，迭代运算为$L_v=T+L_v$，则L_v和T都正确的概率为$p(T,L_v)=\frac{1}{2^L}\frac{1}{2^L}=\frac{1}{4^L}$。

上述充分条件均全部满足时可得出相同的编码结果。$f(x,y)$和$g(x,y)$为 S 盒，基于 S 盒分段迭代和异或运算，一定程度上消除

线性相关性。因每段序列符号 0 和符号 1 的概率不尽相同，且每个符号编码所使用的加权系数因轮函数变化，使得加权概率模型编码后的长度具有随机性。WPMHA 中散列值的比特长度随机，即L为随机值，相比于L固定时碰撞概率更小。又因散列值越长，碰撞概率也越小。于是基于图 1 和图 2 的安全系统随校验次数增加，L随机性增加，使得碰撞的概率越接近于 0。

5 结束语

本文在分布函数区间编码技术的基础上，提出了一种加权概论模型 Hash 算法，该算法具有散列值长度可自定义并且可以根据安全等级及攻击强度可自适应调整的优势，既能保障信息安全，又能够有效降低消息网络传输、校验运算和存储的压力，可广泛应用于数字签名、文件校验以及数据传输校验等领域。为方便检验和应用，本文算法已经公开在 GITHUB 中，获取地址为 https://github.com/Jielin-Code/WjlHashAlgorithm。

参考文献：

[1] Rivest R L . The MD5 Message-digest Algorithm. 1992.

[2] Wang X， D Feng， Lai X , et al. Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD. cryptology eprint archive report, 2004.

[3] Yu H， Wang G， Zhang G , et al. The Second-Preimage Attack on MD4[C] International Conference on Cryptology & Network Security. Springer Berlin Heidelberg, 2005.

[4] Wang X， Yin Y L， Yu H . Collision search attacks on SHA1. Springer Berlin Heidelberg, 2005.

[5] H Choi， Seo S C . Optimization of PBKDF2 using HMAC-SHA2 and HMAC-LSH Families in CPU Environment[J]. IEEE Access, 2021, PP(99):1-1.

[6]来齐齐, 杨波, 禹勇,等. 基于格的哈希证明系统的构造综述[J]. 密码学报, 2017(05):474-484.

[7] Nist F . The Keyed-Hash Message Authentication Code[J]. 2008.

[8] Michail H E， Kakarountas A P， Milidonis A , et al. Efficient implementation of the keyed-hash message authentication code (HMAC) using the SHA-1 hash function[C] IEEE International Conference on Electronics. IEEE, 2004.

[9] Vamsi T S， Kumar T S， Krishna M V . Impact Analysis of Black Hole, Flooding Attacks and Enhancements in MANET Using SHA-3 KeccakAlgorithm[M]. 2021.

[10]刘轩黄. 关于随机过程一阶概率分布函数的遍历性[J]. 大学数学, 1989.

[11] G. N. N. Martin, Range encoding: an algorithm for removing redundancy from a digitised message. Video & Data Recording Conference, held in Southampton July 24-27 1979.

[12] Ian H.Witten, Radford M.Neal,John G.Cleary. Arithmetic Coding for Data Compression.Communications of the ACM. 1987,30(6):520~539.

[13] C. E. Shannon. A mathematical theory of communication. Bell Syst. Tech. J., 27:379–423,623–656, 1948.

[14] T.M.Cover and J.A.Thomas,Elements of Information Theory.New YorK,Wiley 1991.

[15] Wang W， Feng A . Self-Information Loss Compensation Learning for Machine-Generated Text Detection[J]. Mathematical Problems in Engineering, 2021, 2021(1):1-7.

[16] Manral V . Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD[J]. Cryptology Eprint Archive Report, 2004, 2004.

[17] Weber B， Zhang X . Parallel hash collision search by Rho method with distinguished points[C] 2018 IEEE Long Island Systems, Applications and Technology Conference (LISAT). IEEE, 2018.

[18] Keller, Nathan, Shamir, et al. Improved Single-Key Attacks on 8-Round AES-192 and AES-256[J]. Journal of cryptology: the journal of the International Association for Cryptologic Research, 2015.

[19] Tromer E， Osvik D A， Shamir A . Efficient Cache Attacks on AES, and Countermeasures[J]. Journal of Cryptology, 2010, 23(1):37-71.

[20] Biham E， Biryukov A . An Improvement of Davies' Attack on DES[J]. Journal of Cryptology, 1997, 10(3):195-205.

作者简介：王杰林，男，1985 年生，湖南平江人，湖南涉外经济学院特聘教授，主要研究方向为编码技术。

高金定，男，1981 年生，湖南桃江人，博士（后），电子学教授，主要研究方向为数字信号处理 FPGA 实时实现技术、信息编码。