HASH Algorithm of Weighted Probability Model

SUBMISSION DATE / POSTED DATE

19-01-2021 / 22-01-2021

CITATION

Wang, Jielin (2021): HASH Algorithm of Weighted Probability Model. TechRxiv. Preprint.
https://doi.org/10.36227/techrxiv.13606373.v1

DOI

10.36227/techrxiv.13606373.v1

HASH Algorithm of Weighted Probability Model

Jielin Wang

Hunan International Economics University, Changsha, Hunan, China

E-mail: 254908447@qq.com

***Abstract*—The weighted probability mass function of the discrete random variable X is defined as $\varphi(X) = rp(X)$, where $p(X)$ is the fixed probability of the symbol X and r is a positive real number weight coefficient. Let $F(X, r)$ be the weighted distribution function and $F(X, r) = \sum_{s \le X} \varphi(s)$. A new hash algorithm is constructed and its implementation method is provided in this paper based on the weighted distribution function $F(X, r)(r > 1)$. It is proved that the algorithm can reach the theoretical limit of hash collision. The implementation procedure of this article is published in GITHUB.**

***Keywords*—One-way hash function; HASH functions; Weighted probability**

I: INTRODUCTION

The common one-way hash functions currently include MD5, SHA, MAC, CRC, etc. MD5 Message-Digest Algorithm [1], a widely used cryptographic hash function, can generate a 128-bit (16-byte) hash value to ensure that information transmission is complete and consistent. After 1996, the algorithm proved to have weakness and could be cracked. For data fields requiring high security, other algorithms are generally used, such as SHA-2 and SM3. In 2004, experts confirmed that the MD5 algorithm cannot prevent collisions, so it is not suitable for security authentication, such as SSL public key certification or digital signature.

With the increasing computing power of computer equipment, hash algorithms that can adapt to collision strength are more suitable for security authentication and digital signatures. The hash algorithm for adaptive collision strength should satisfy:

(1) The hash value length is automatically increased with the intensity of the collision attack.

(2) Different systems and different permissions use hash values of unequal length to reduce computational waste and reduce the possibility of collision.

MD5, SHA-2 and SM3 cannot satisfy the above conditions. A brand-new hash algorithm through the weighted probability distribution function is constructed in this paper. This algorithm can encode an input message string of any length into an output string with a bit length of L, and L can be customized.

II: THE WEIGHTED PROBABILITY MODEL

In this chapter, the distribution function based on range encoding [2][3] method is discussed, the definitions of weighted probability mass function and weighted distribution function are provided, and the expression of the weighted probability model Hash function is inferred.

2.1 Range encoding

Let source sequence $X = (X_1, X_2, \dots, X_i, \dots, X_n)$ be a discrete sequence $X_i \in A = \{0,1,2, \dots, k\}$ that takes on a finite or countable number of possible values. Then for all numerical values in A we have a probability space, that is

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

Since the source sequence must be transferred into some symbol, at any time we have

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

Therefore, the distribution function of any symbol X_i is

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

$p(0) \le F(x) \le 1,\ s \in A$. According to the range encoding algorithm in reference [3], we let R_i be the length of the interval $[L_i, H_i)$ after the $i(i = 1,2, \dots, n)$-th symbol X_i. Since $L_0 = 0$ and $H_0 = R_0 = 1$, the range encoding iterative operation process of the source sequence X is as follows:

$$R_i = R_{i-1} p(x_i), \qquad L_i = L_{i-1} + R_{i-1} F(x_i - 1), \qquad H_i = L_i + R_i \tag{2-2}$$

where $F(X_i - 1)$ is distribution function.

$$F(X_i - 1) = \sum_{s < X_i} p(s) \tag{2-3}$$

Taking the binary source sequence $X = 010$ as an example, the iterative operation based on Equation (2-2) is shown in Figure 1.

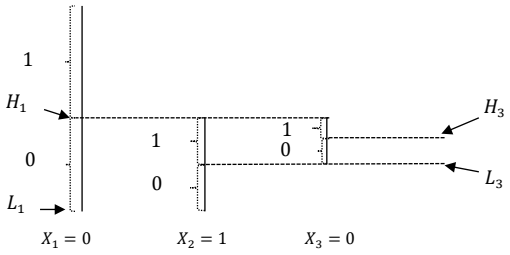

Figure 1 Schematic diagram of range encoding process

Select any real number in the interval $[L_3, H_3)$ as the encoding result. We let the encoding result be L_3. When the source sequence X is restored by L_3, $p(0)$ and $p(1)$ are known. We let $L_0 = 0$, $H_0 = R_0 = 1$ and $i = 1$, then the decoding step is

Step 1: According to $p(0)$ and $p(1)$, $F(0) = 0$ and $F(1) = p(0)$, then $H_{i=1} = p(0)$.

Step 2: If $L_3 \geq H_{i=1}$, symbol "1" is output; if $L_3 < H_{i=1}$, symbol "0" is output.

Step 3: $i = i + 1$, return to step 1.

According to the decoding steps, it is easy to obtain the source sequence $X = 010$. Let $\prod_{j=1}^{0} p(X_j) = 1$, then from Equation (2-2) we obtain

$$L_n = \sum_{i=1}^{n} F(X_i - 1) \prod_{j=1}^{i-1} p(X_j) \tag{2-4}$$

$$R_n = \prod_{i=1}^{n} p(X_i) \tag{2-5}$$

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

2.2 Weighted distribution function

Definition 2.1 We let discrete random variable be X, where $X \in A = \{0,1,\dots,k\}$. Furthermore, $\mathrm{P}\{X = a\} = p(a)(a \in A)$, and the weighted probability mass function is $\varphi(a) = r\mathrm{P}\{X = a\} = rp(a)$. $p(a)$ is the probability mass function, $0 \leq p(a) \leq 1$, and r is the weight coefficient, and

$$F(a) = \sum_{i \leq a} p(i) \tag{2-7}$$

If $F(a,r)$ satisfies $F(a,r) = rF(a)$, then $F(a,r)$ is called a weighted cumulative distribution function, or a weighted distribution function for short. The sum of the weighted probabilities of all the symbols is $\sum_{a=0}^{k} \varphi(a) = r$.

We let the discrete source sequence be $X = (X_1, X_2, \dots, X_n)$, where $X_i \in A$, and $F(X_i - 1) = F(X_i) - p(X_i)$. The weighted distribution function of sequence X is denoted as $F(X,r)$. When $n = 1$,

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

When $n = 2$,

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

When $n = 3$,

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

Let $\prod_{j=1}^{0} p(X_j) = 1$, by analogy we can obtain

$$F(X,r) = \sum_{i=1}^{n} r^i F(X_i - 1) \prod_{j=1}^{i-1} p(X_j) + r^n \prod_{i=1}^{n} p(X_i) \tag{2-8}$$

A set of weighted distribution functions satisfying Equation (2-2) is defined as a weighted probability model, which is henceforth referred to as a weighted model, and is denoted as $\{F(X,r)\}$. If $X_i \in A = \{0,1\}$, then $\{F(X,r)\}$ is called a binary weighted model. We let

$$H_n = F(X,r) \tag{2-9}$$

$$R_n = r^n \prod_{i=1}^{n} p(X_i) \tag{2-10}$$

$$L_n = H_n - R_n \tag{2-11}$$

where $X_i \in A, n = 1,2, \ldots$. When $r = 1$,

$$F(X,1) = \sum_{i=1}^{n} F(X_i - 1) \prod_{j=1}^{i-1} p(X_j) + \prod_{i=1}^{n} p(X_i) \tag{2-12}$$

From Equation (2-4), (2-5) and (2-6), $H_n = F(X,1)$, that is, the range coding (arithmetic coding) [2,3] is a lossless coding method based on the weighted distribution function of $r = 1$.

Since X_i must take a value in A, $p(X_i) > 0$. Equations (2-9), (2-10) and (2-11) are interval columns, and $[L_i, H_i)$ is the corresponding interval superscript and subscript of variable X_i of source sequence X at time $i(i = 0,1,2, \ldots, n)$. $R_i = H_i - L_i$ is the length of the interval. The iterative expression of Equations (2-9), (2-10) and (2-11) are

$$\begin{aligned} R_i &= R_{i-1}\varphi(X_i) \\ L_i &= L_{i-1} + R_{i-1}F(X_i - 1, r) \\ H_i &= L_i + R_i \end{aligned} \tag{2-13}$$

Through (2-13), the weighted probability model coding operation is performed on the source sequence X, and L_n is the coding result.

2.3 Information entropy of weighted model

We let the discrete memoryless source sequence X be $X = (X_1, X_2, \ldots, X_n)(X_i \in A, A = \{0,1,2, \ldots, k\})$. When $r = 1$, $\varphi(X_i) = p(X_i)$. According to the definition of Shannon information entropy, the entropy of X is

$$H(X) = -\sum_{X_i=0}^{k} p(X_i) \log p(X_i) \tag{2-14}$$

When $r \neq 1$, we define the self-information of the random variable X_i with probability $\varphi(X_i)$ as

$$I(X_i) = -\log \varphi(X_i) \tag{2-15}$$

We let the number of a in set $\{X_i = a_j\}(j = 0,1, \ldots, k; i = 1,2, \ldots, n)$ be c_a. When the value of r is determined, the total amount of information of source sequence X is

$$-\sum_{a=0}^{k} c_a \log \varphi(a)$$

Thus, the average amount of information per symbol is

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

Definition 2.2 Let $H(X,r)$ be

$$\begin{aligned} H(X,r) &= -\sum_{a=0}^{k} p(a)\log\varphi(a) \\ &= -\log r - \sum_{a=0}^{k} p(a)\log p(a) \\ &= -\log r + H(X) \end{aligned} \tag{2-16}$$

According to definition 2.2, when the value of r is determined, the binary length encoded by the weighted probability model is $nH(X,r)(bit)$. Then, the source sequence X encoded by a weighted probability model obtains a sequence of length $L(bit)$, and the symbols "0" and "1" in the message have fixed probabilities $p(0)$ and $p(1)$. Using 2 as logs base, from Equation (2-16) we obtain

$$-n\log_2 r + nH(X) = L \tag{2-17}$$

where $H(X)$ is the information entropy of sequence X. $H(X) = -p(0)\log_2 p(0) - p(1)\log_2 p(1)$, so simplify (2-17) and obtain

$$\begin{aligned} r &= 2^{H(X)-\frac{L}{n}} \\ &= 2^{-p(0)\log_2 p(0) - p(1)\log_2 p(1) - L/n} \end{aligned} \tag{2-18}$$

According to the lossless coding theorem of information theory, $H(X)$ is the lossless coding limit of the discrete memoryless source sequence X, so when $H(X,r) \geq H(X)$, the weighted model function $F(X,r)$ can restore source X without loss. When $H(X,r) < H(X)$, the weighted model function $F(X,r)$ cannot restore the source X, that is, when $L < nH(X)$, the encoding result L_n cannot restore the source X.

From (2-17) and (2-18), when $H(X) > L/n$, $r > 1$ and then $H(X,r) < H(X)$, so the weighted model functions $F(X,r)$ that satisfy Equation (2-18) and $r > 1$ are all one-way hash function (Hash function).

2.4 Collision limit

Theorem 2.1 The probabilities of symbol "0" and symbol "1" in the hash value obtained by the weighted probability model hash algorithm for any binary sequence are equal.

Proof Let the bit length of the hash value obtained by the weighted probability model hash algorithm of the binary sequence is L. The binary sequence of the hash value is recorded as Y. Its information entropy is $H(Y) = -p(0)\log_2 p(0) - p(1)\log_2 p(1)$. According to Definition 2.2, $nH(X,r) = -n\log_2 r + nH(X)$ (n is the bit length of the binary sequence X), so $LH(Y) = -n\log_2 r + nH(X)$. If and only if $H(Y) = 1$, Equation (2-17) is tenable, that is, r satisfies Equation (2-18). Otherwise, r does not satisfy Equation (2-18). And if and only if $p(0) = p(1) = 0.5$, $H(Y) = 1$, so the probabilities of symbol "0" and symbol "1" in sequence Y are equal.

According to Theorem 2.1, the probabilities of symbols in the hash value obtained by the hash algorithm in this paper are equal. Let the bit length of the hash value be L, then the value range of the value space is $\{0,1,\dots,2^L - 1\}$. Let $d = 2^L$, according to the probability of hash collision (or "birthday paradox"), the probability of hash collisions through N test is:

$$p(N,d) \approx 1 - e^{\frac{-N(N-1)}{2d}} \tag{2-19}$$

Obviously, this algorithm can reach the theoretical limit of hash collision.

III: HASH ALGORITHM OF BINARY WEIGHTED PROBABILITY MODEL

3.1 Hash algorithm steps

L is the bit length of the customized hash value, and the interval encoding steps of the binary source sequence X using the weighted probability model are as follows.

1): Initialize parameters, $p = 0$, $L_0 = 0$, $H_0 = R_0 = 1$, $i = 1$;

2): Count the number of symbols 0 in the source sequence X as c, and the bit length of the source sequence X is n;

3): Calculate the probability of symbol 0, $p = \frac{c}{n}$;

4): Calculate the weight coefficient, $r = 2^{-p\log_2 p-(1-p)\log_2(1-p)-L/n}$;

5): Calculate the weighted probability, $\varphi(0) = rp$, $\varphi(1) = r(1-p)$;

5): Obtain the i-th symbol X_i of the sequence X;

6): If $X_i = 0$, $L_i = L_{i-1}$ and $R_i = R_{i-1}rp$, otherwise $L_i = L_{i-1} + R_{i-1}rp$ and $R_i = R_{i-1}r(1-p)$;

7): $i = i + 1$. If $i < n$, repeat steps 5 to 7, and then obtain L_n;

8): End encoding and output L_n (L_n is the hash value).

The above steps can be illustrated in Figure 2.

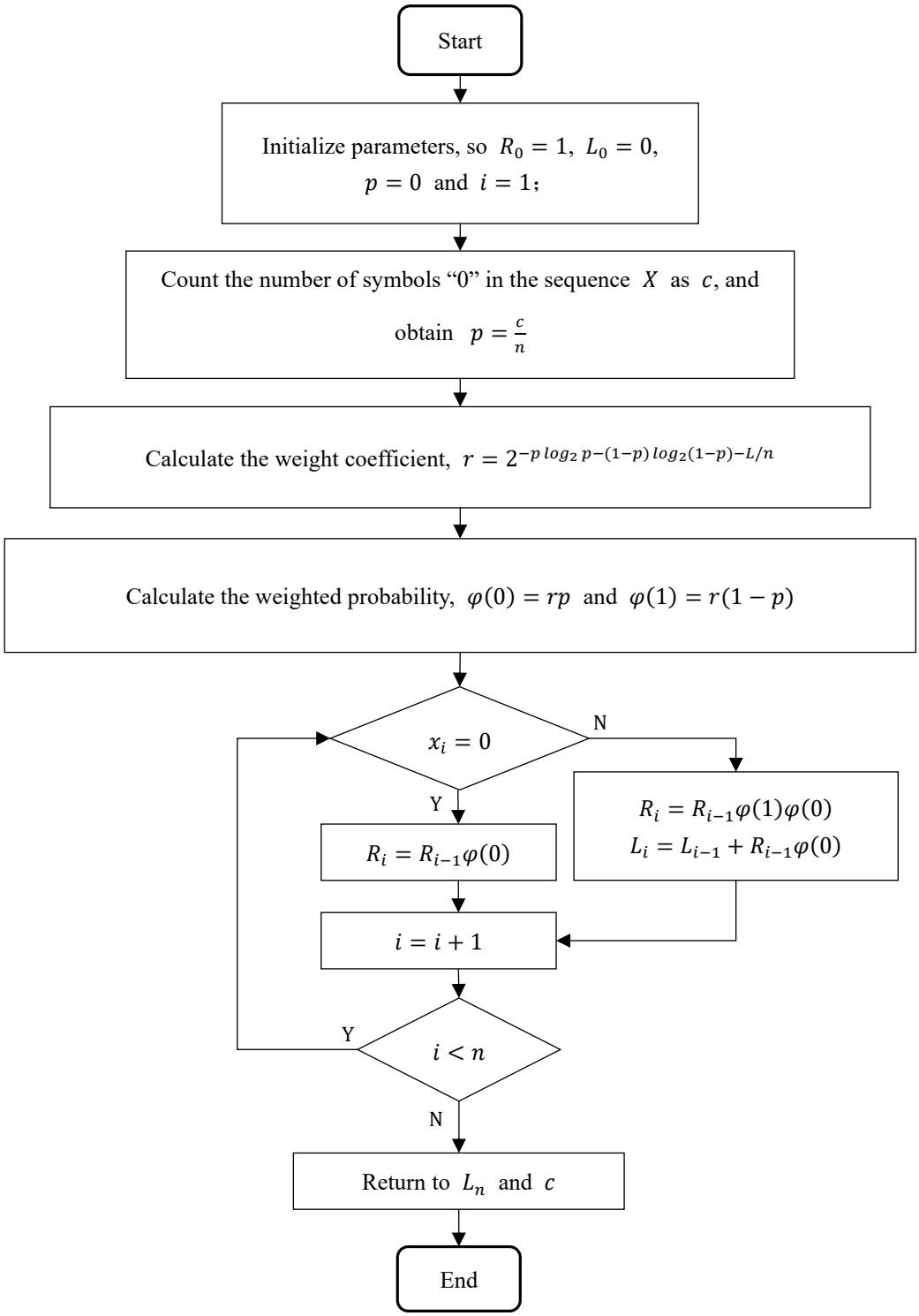

Figure 2 Schematic diagram of binary weighted model hash algorithm

3.2 Experiment and comparison

(1) customize hash value length $bytelength$

Input: abcdefghijklmnopqrstuvwxyz123456789

The output $bytelength$ is 4, 8, 15, 20, 32 (unit: byte).

$bytelength$	Hash value
4byte	b21527d1
8byte	7a63428361901e1c

15byte	d99067a5ebd9dd1f5eefc63bddcfa
20byte	cd5c2c83aa1cf1468590303811895af11bf06943
32byte	233652b6ff6365c634bd1f7dbe58967ebde64b849b5beed3a145f4b83efe365

Table 1 The customized hash values of different lengths

(2) The probability of symbol 0 in the hash value

Input(byte)	$p(0)$
105 250 74 59 245 144 19	0.500000
0 1 2 3 4 5 6 7	0.505859
212 168 167 102 99 110 58	0.507813
185 226 64 200 184 255 211	0.498047

Table 2 The value of $p(0)$ when the hash value length is 512 bits

(3) Comparison of the efficiency of this algorithm with MD5, SHA256 and SHA512

Input A: abcdefghijklmnopqrstuvwxyz123456789

Input B: abcdefghijklmnopqrstuvwxyz123456789~!@#$%^&*()_+

Output hash values of equal length, and compare algorithm efficiency under the same device.

Algorithm	MD5	$bytelength = 16$
Bit length	128bit	128bit
A	3ms	5ms
B	4ms	6ms

Table 3 Comparison of the efficiency of this method and MD5

Algorithm	SHA256	$bytelength = 32$
Bit length	256bit	256bit
A	4ms	7ms
B	4ms	7ms

Table 4 Comparison of the efficiency of this method and SHA256

Algorithm	SHA512	$bytelength = 64$
Bit length	512bit	512bit
A	7ms	9ms
B	7ms	9ms

Table 5 Comparison of the efficiency of this method and SHA512

It can be concluded from the experimental results that the hash length of the method in this paper can be customized and conforms to Theorem 2.1. As it involves floating-point operations, the efficiency is slower than MD5, SHA256, and SHA512. However, efficiency problems can be solved through code optimization or parallelization, and this article does not make a specific analysis.

IV: CONCLUSION

In this paper, a brand-new hash algorithm is provided. The length of the hash value can be customized. It is theoretically proved that the hash collision limit can be reached. The algorithm can be applied flexibly, and it can be applied to digital signatures, document verification and other fields in the future.

REFERENCES

[1] Rivest, R. "The MD5 Message Digest Algorithm." RFC 1321 (1992).

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

[3] 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.

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

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

[6] 张鸣然.Hash 算法[J].中国科技投资,2019,(2).doi: 10.3969/j.issn.1673-5811.2019.02.172

[7] 张裔智,赵毅,汤小斌. MD5 算法研究[J]. 计算机科学,2008,(7):295-297.

[8] 赵玉鑫,刘光杰,戴跃伟,等. 一种新的视觉 Hash 算法[J]. 光学精密工程,2008,(3):551-557.doi:10.3321/j.issn:1004-924X.2008.03.029.

[9] 易红军,佘名高. MD5 算法与数字签名[J]. 计算机与数字工程,2006,(5):44-46.

[10] 曹安丽,谢长生. 利用 Hash 算法优化网络备份系统[J]. 计算机工程与科学,2005,(3):4-6,12.doi:10.3969/j.issn.1007-130X.2005.03.002.

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

[12] Wang X , Yin Y L , Yu H . Finding Collisions in the Full SHA-1[C]// Advances in Cryptology - CRYPTO 2005: 25th Annual International Cryptology Conference, Santa Barbara, California, USA, August 14-18, 2005, Proceedings. Springer, Berlin, Heidelberg, 2005.

[13] Gilbert H , Handschuh H . Security Analysis of SHA-256 and Sisters[J]. 2003.