第37卷 第18期
Vol.37 No.18

计 算 机 工 程
Computer Engineering

2011年9月
September 2011

·安全技术·

文章编号：1000—3428(2011)18—0124—03 文献标识码：A 中图分类号：TP309

基于MD5的迭代散列算法

张 青

(中国矿业大学计算机科学与技术学院，江苏 徐州 221008)

摘 要：分析网络中用户名、密码的存储方式及其存在的风险。在此基础上分析MD5散列算法的弱点及其破译手段。针对这些破译方法提出基于MD5的迭代散列算法。该算法可以避免第二类生日攻击，并有效提高第一类生日攻击的复杂度，对于破解效率最高的彩虹表也具有免疫性，能够加强密码的安全性能，从而提高网络中信息传递和存储的安全性。

关键词：MD5散列算法；迭代算法；生日攻击；彩虹表

Iterative Hashing Algorithm Based on MD5

ZHANG Qing

(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221008, China)

【Abstract】 The paper analyzes the user name and password storage methods and their risks on the internet. On this basis, it analyzes MD5 message digest encryption weaknesses and deciphering means. In response to these decoding methods, it proposes an iterative algorithm based on MD5. This algorithm is immune to the second birthday attack, meanwhile improve the complicates of the the first birthday attack. It is also useful for rainbow table attack. The algorithm can enhance password security features, thereby enhancing the network information transmission and storage security.

【Key words】 MD5 hashing algorithm; iterative algorithm; birthday attack; rainbow table

DOI: 10.3969/j.issn.1000-3428.2011.18.041

1 概述

随着计算机网络的飞速发展，网络安全的形势日趋严峻。其中主要有论坛、聊天室、内容管理系统等。为了有效地保护、存储、管理和使用网上的私有信息，MD5散列算法被广泛应用于网络领域。但王小云教授于2004年给出了MD5产生碰撞的一个充分条件集，并首次成功对MD5进行了碰撞攻击，而且碰撞的改进和效率在不断提高。目前，Web系统都是采用登录验证的方式进行用户验证的，而在登录时，大多数Web系统是采用MD5散列算法后的密文传输；账号与密码不但在传输过程中存在风险，在网站的数据库中也存在风险。一旦网站被攻破，黑客就可以获得数据库中的所有数据。数据库中存储的往往有用户名、密码经过MD5散列算法后的密文。通过破解可以得到用户名、密码的明文，从而得到相应的权限进而对网站构成威胁。为此，本文在MD5散列算法的基础上，使用基于MD5的迭代散列算法，对用户的账号、密码进行加密，以提高信息的安全性。

2 MD5的破译方法

MD5的破译方法主要包括以下3种破译方法：

(1)在线破译是通过网站所提供的数据库进行比对得到明文，如www.cmd5.com，用户只要输入要破解的MD5码就可以查询到相应的密文(一般小于12位)。

(2)本地破译主要有“跑字典”和运用相应的碰撞算法破解。“跑字典”类似于上面的在线查询，只是使用的是自己事先确定好的范围。由于电脑的性能有限，“跑字典”的破译率很低。碰撞算法是通过一些算法找到找一个x'使$H(x)=H(x')$，通常情况下$x\neq x'$。这种破解方法需要破解者按照一定的算法自己编写程序。由于算法的不同，破解的效率也不同。但随着好的算法的出现，这种方法已对MD5算法构成威胁[1-2]。

(3)彩虹表破译。综合上面2种破解方法，主要是运用了下面2种思想：一种是暴力破解法，把明文P中的每一个p都算一下$hash(p)$，直到结果等于q；另一种办法是查表法，把每个p和对应的q都记录下来，按q做一下索引，到时候查一下就知道了。这2种办法理论上都是可以的，但是前一种可能需要海量的时间，后一种需要海量的存储空间，以至于在使用时会受到许多限制。彩虹表的根本原理就是组合了暴力法和查表法，并在这两者之中取得一个折中，用可以承受的时间和存储空间进行破解。使用彩虹表的方法，对于任意14位以内的密码都可以平均在5 min以内破解出来。

3 基于MD5的多重计算算法

3.1 算法描述

通过上文对网络中用户名和密码的存储方式、存在的风险和MD5信息摘要加密的破译手段的分析可见：单纯的MD5信息摘要加密的安全性已经大大降低了。针对以上这些情况提出基于MD5的新的算法，从而增加密码的安全性能，提高网络中信息传递和存储的安全性。基于MD5的多重计算算法的主要描述如下：

(1)将用户名和密码分别进行MD5信息摘要。

(2)将生成的MD5信息摘要再次进行信息摘要。

(3)重复上面的步骤N次。

(4)输出最终的结果。

如图1所示，设一明文为M，H_x为第x次运算后得到的$hash$值。基于MD5的迭代散列算法可表示为：

$H_1=hash(M)$, $H_2=hash(H_1)$,…, $H_N=hash(H_{N-1})$

基金项目：江苏省自然科学基金资助项目(BK2007035)；中国矿业大学青年科技基金资助项目(0D061035, 2007A047)

作者简介：张 青(1989－)，男，本科生，主研方向：信息安全

收稿日期：2011-02-10 **E-mail**：mailzhangqing@126.com

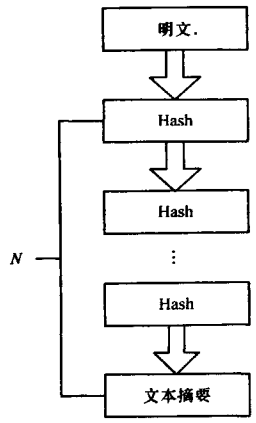

图1 基于MD5的多重计算算法

3.2 基于MD5的多重计算算法的安全性分析

随着MD5破解及其破解速度的不断提高，单纯的MD5信息摘要已经不能够保障信息的安全。采用基于MD5的多重计算的算法在某些程度上可以加强信息的安全性。

由于MD5的多重计算的算法只是单纯的将消息进行n次MD5信息摘要，所以无法抵抗第二类生日攻击。攻击者只需将最后所得的*hash*值按照加密过程倒过去进行解密(即$H_N=hash^{-1}(H_{N-1})$、$H_{N-1}=hash^{-1}(H_{N-2})$、…、$H_1=hash^{-1}(M)$)，就可以得到与明文相应的碰撞对。通过基于MD5的多重计算的算法可以将128为的MD5碰撞攻击时间复杂度由$O(2^{64})$提高到$O(n\times 2^{64})$(n为MD5摘要算法的重复的次数)。

MD5的多重计算的算法虽然增加了破解的复杂度，但相对于单纯的MD5信息摘要并没有本质上的区别。

4 基于MD5的迭代散列算法

4.1 算法描述

基于MD5的迭代散列算法的主要描述如下：

(1)将密码以N个字符为一段进行分割(N小于密码长度)。

(2)将第一段字符进行MD5信息摘要算法，并将输出的结果迭代在后一段字符的后面。

(3)对新组合成的字符段进行MD5信息摘要算法，并将输出的结果迭代在后一段字符的后面。

(4)重复以上过程至最后一段字符。

(5)输出结果，如图2所示。

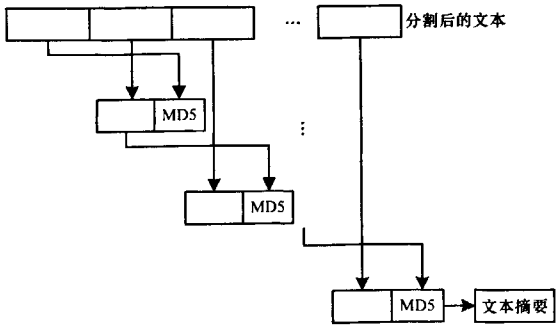

图2 基于MD5的迭代散列算法

设一明文为M，将其按照规律分成$i=M/N$段，各段为M_1、M_2、…、M_i，H_x表示第x次运算后得到的*hash*值。

基于MD5的迭代散列算法可以表示为：$H_1=hash(M_1)$，$H_2=hash(M_2+H_1)$，…，$H_i=hash(M_i+H_{i-1})$。

对于MD5信息摘要算法来说，分段过程所花费的时间是可以忽略不计的。又将明文分M成$i=M/N$段，所以生成基于MD5的迭代散列算法的时间复杂度是MD5信息摘要算法的i倍，为倍数关系。如密码12345678经过MD5信息摘要加密后为：

25d55ad283aa400af464c76d713c07ad

经过基于MD5的迭代加密后为(可选1~7中任意一个数为分割单元，此处以1为分割单元)：

d36b42db4e9c6ea995430be65b883ae4

密码12345678基于MD5的迭代加密过程如表1所示。

表1 MD5迭代计算过程

轮数	读入	输出
1	1	c4ca4238a0b923820dcc509a6f75849b
2	2c4ca4238a0b923820dcc509a6f75849b	c903e3a4ef62ea2bc5cff8847475621c
3	3c903e3a4ef62ea2bc5cff8847475621c	9f7041ef0af6cecca21243b4ce1e2078
4	49f7041ef0af6cecca21243b4ce1e2078	c32977f6781aeb76ddbebcb8cbeb64c9
5	5c32977f6781aeb76ddbebcb8cbeb64c9	bc8092e5fb4dad2a82f883f475ef749f
6	6bc8092e5fb4dad2a82f883f475ef749f	d3828bc0d3b0181583bd940862cecd08
7	7d3828bc0d3b0181583bd940862cecd08	3386a1e7be0aeef90c13ce511b9cac02
8	83386a1e7be0aeef90c13ce511b9cac02	d36b42db4e9c6ea995430be65b883ae4

4.2 基于MD5的迭代散列算法的安全性分析

采用基于MD5的迭代散列算法后，他人截获或被从网站数据库中获得的账号、密码都是密文。对于这样的密文破解可分为以下2种情况：

(1)对于第二类生日攻击来说，攻击者可以通过密文进行碰撞，得到倒数第二步的碰撞值[3]。由于碰撞值与原来的倒数第二步的*hash*值是不同的，所以无法将明文中的添加量从碰撞值中剔除，因此攻击者无法进行下一步的破解。可知基于MD5的迭代散列算法是可以抵抗第二类生日攻击的。

(2)对于第一类生日攻击来说，由于攻击者不知道明文分割的规则，无法模拟加密过程的逆过程，进行“跑字典”或暴力破解。只有结合这2个方法才有可能破译出明文。通常情况下对一个具有128位的MD5信息摘要进行碰撞攻击需要的时间复杂度是$O(2^N)$。对于一个n位的基于MD5的迭代散列算法的字符串进行碰撞攻击与暴力破解相结合的方法进行破译，需要的时间复杂度是(ceil()为向上取整函数)。

$O(2^{(\text{ceil}(n/1)+\text{ceil}(n/2)+\cdots+\text{ceil}(n/(n/2))+\cdots+\text{ceil}(n/(n-1))+\text{ceil}(n/n))\times N})$

对指数(ceil(n/1)+ceil(n/2)+…+ceil(n/(n/2))+…+ceil(n/(n-1))+ceil(n/n))×N进行观察可得：ceil(n/n)为1；ceil(n/(n/2))、ceil(n/(n/2+1))、ceil(n/(n/2+2))、…、ceil(n/(n-2))、ceil(n/(n-1))这n/2个数的值均为2，其和为n；ceil(n/3)、ceil(n/(n/3+1))、ceil(n/(n/3+2))、…、ceil(n/(n/2-2))、ceil(n/(n/2-1))这n/2-n/3个数的值均为3，其和为n/2；ceil(n/(n/4))、ceil(n/(n/4+1))、ceil(n/(n/4+2))、…、ceil(n/(n/3-2))、ceil(n/(n/3-1))这n/3-n/4个数的值均为4，其和为n/3；依次类推，Ceil(n/(n/n))、ceil(n/(n/n+1))、ceil(n/(n/n+2))、…、ceil(n/(n/(n-1)-2))、ceil(n/((n/n-1)-1))这n/(n-1)-n/n个数的值均为n，其和为n/(n-1)。

综上可得：

$$\text{ceil}(n/1)+\text{ceil}(n/2)+\cdots+\text{ceil}(n/(n/2))+\cdots+\text{ceil}(n/(n-1))+\text{ceil}(n/n)=1+n+\frac{n}{2}+\frac{n}{3}\ldots+\frac{n}{n-1}=1+n\sum_{k=2}^{n}\frac{k}{k-1}$$

所以时间复杂度为：

$$O(2^{(\text{ceil}(n/1)+\text{ceil}(n/2)+\cdots+\text{ceil}(n/(n/2))+\cdots+\text{ceil}(n/(n-1))+\text{ceil}(n/n))\times N})=O(2^{(1+n\sum_{k=2}^{n}\frac{k}{k-1})\times N})$$

由于$\sum_{k=2}^{n}\frac{k}{k-1}$为调和级数，所因此函数随$n$的变化无限增大，且增速极快。假设密码的长度为8位，即当n=8时，

$1+n\sum_{k=2}^{n}\frac{k}{k-1}=24$ 即在长度为 8 的条件下，基于 MD5 的迭代散列算法破解的算法复杂度是普通 MD5 破解的算法复杂度的 24 次方。随着字符长度的变化，指数的大小将快速增长。对于现有的技术及其发展速度而言，在很长一段时间内基于 MD5 的迭代散列算法是安全的。

4.3 Rainbow 对基于 MD5 的迭代散列算法的攻击分析

针对基于 MD5 的迭代散列算法的特点使用彩虹表对其进行破解：

(1)通过彩虹表对每一步的 hash 值进行破解；

(2)对分组状况进行穷举并将其使用到每一步的破解中；

(3)重复上面的步骤，直至求得明文。

具体步骤如下：

(1)对于单步基于 MD5 的迭代散列算法的破解：给定一个明文 P_0 和与之对应的密文 C_0，试图找到用密码算法 S 加密 P_0 时所需要的密钥 $k\in N$ (N 为密钥空间)，使得：$C_0=S_k(P_0)$。

使用所有可能的 $k'\in N$ 的密钥去加密 P_0，这样就预计算出所有可能的密文。而所有的密文是按链得方式组织的，在内存中只保存链首和链尾，这样就表现出了时间和空间的折中策略。这个链是使用映射函数 R 来生成的，映射函数将一个密文映射到一个密钥。链的组织如下：

$$k_i \xrightarrow{S_{k_i}(P_0)} C_i \xrightarrow{R(C_i)} k_{i+1}$$

用 $f(k)$ 表示 $R(S_k(P_0))$，这样就得到了一个密钥的链：

$$k_i \xrightarrow{f} k_{i+1} \xrightarrow{f} k_{i+2} \longrightarrow \cdots$$

生成一个表，表中包含 m 个链，每个链的长度为 t，但是为了节约空间，只保存这 m 个链的链首和链尾元素。给定一个密文 C，在这个表中找出生成这个密文的密钥。首先以 $R(C)$ 为开始元素，生成一个长度为 t 的链。如果生成密文 C 的密钥确实在表中出现，肯定能在表中找到一个与链尾元素匹配的链。由于只有链首和链尾元素被保存了，所以需要从链首元素开始重构这个链，在 $R(C)$ 之前的那个密钥就是生成密文 C 的密钥[4]。

在一个有 m 个链，每个链长度为 t 的表中，查找一个密钥成功的概率为：

$$P_{table}\geqslant\frac{1}{N}\sum_{i=1}^{m}\sum_{j=0}^{t-1}\left(1-\frac{it}{N}\right)^{j+1}$$

(2)对分组状况进行穷举求解。从上面得到的单步的明文 M_x，穷举分组长度(1~N)，设为 y。从单步明文的首部截取前 y 个字符作为最终明文的后 $x\times y$ 位。对截取明文剩下的密文 H_{x-1} 继续用彩虹表进行破译。即 M_x=rainbow(H_x)=m_x+H_{x-1}。

(3)以 y 为分组长度重复上面的步骤，得到最终明文 M。

虽然对于明文小于 14 位的 MD5 散列，彩虹表可以在 5 min 内找出明文。然而对于明文小于 14 位的基于 MD5 的迭代散列算法需要 $300^{1+14\sum_{k=2}^{24}\frac{k}{k-1}}=300^{228}$ s $\approx 0.856\times 300^{225}$ y。所以基于 MD5 的迭代散列算法对彩虹表是免疫的。

5 结束语

MD5 是一种非常易用和安全的散列算法，但随着技术的发展，MD5 信息摘要算法不断受到威胁。本文采用基于 MD5 的迭代散列算法可以抵抗第二类生日攻击，增加第一类生日攻击的攻击复杂度，避免目前出现的对 MD5 的破译方法，如彩虹表攻击。通过基于 MD5 的迭代散列算法可提高网络中信息传递和存储的安全性。

参考文献

[1] Rivest R. The MD5 Message-Digest Algorithm[EB/OL]. (2005-09-05). http://www.ietf.org/rfc/rfc1321.txt.

[2] Wang Xiaoyun, Feng Dengguo, Lai Xuejia, et al. Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD[EB/OL]. (2004-08-17). http://eprint.iacr.org/2004/199.pdf.

[3] 孔 政，姜秀柱. DNS 欺骗原理及其防御方案[J]. 计算机工程，2010, 36(3): 125-127.

[4] Philippe O. Rainbow Table[EB/OL]. (2005-05-16). http://en.wikipedia.org/wiki/Rainbow_tables.

编辑 陈 文

(上接第 123 页)

4 结束语

本文提出了一个针对二值图像的信息隐藏算法。该算法中可修改得分的计算考虑到像素点的区域连通性，避免了像素点修改对区域连通性的破坏，更合乎人眼的视觉特性，同时利用 STC 编码方法提高了嵌入效率。实验结果表明，本文方法不仅嵌入容量较大，而且载密图像整体失真很小，较好地保护了隐藏信息的不可见性。

参考文献

[1] Tzeng C, Tsai W. A New Approach to Authentication of Binary Images for Multimedia Communication with Distortion Reduction and Security Enhancement[J]. IEEE Communications Letters, 2003, 7(9): 443-445.

[2] Wu Min, Liu Bede. Data Hiding in Binary Images for Authentication and Annotation[J]. IEEE Transactions on Multimedia, 2004, 6(4): 528-538.

[3] Yang Huijuan, Kot A C. Pattern-based Data Hiding for Binary Image Authentication by Connectivity-preserving[J]. IEEE Transactions on Multimedia, 2007, 9(3): 475-486.

[4] Lee Younho, Kim Heeyoul, Park Yongsu. A New Data Hiding Scheme for Binary Image Authentication with Small Image Distortion[J]. Information Sciences, 2009, 179(22): 3866-3884.

[5] Filler T, Judas J, Fridrich J. Minimizing Embedding Impact in Steganography Using Trellis-coded Quantization[C]//Proc. of SPIE'10. San Jose, USA: [s. n], 2010.

[6] 蒋 斌，平西建，张 涛. 基于模式分析的二值文本图像隐写分析算法[J]. 计算机工程，2009, 35(8): 176-178.

编辑 张正兴