MD5 散列函数的结果是均匀分布吗？

发表于 2015 年 08 月 12 日 | 分类于 IT | 本文共被围观 179 次 | 暂无评论

今天，大学同学昊轩在微信群里问到「MD5 散列函数的结果是不是均匀分布的」。询问之后才知道，昊轩在工作中需要一个快速的、均匀分布的 8 – 16 bytes 的散列函数。

回忆本科学过的内容，似乎并没有提及散列函数的结果服从何种统计分布。不过，一个合格的散列函数应当包含三个特征：

- 单向性：容易计算输入的散列结果，但是从散列结果无法推出输入内容；
- 抗碰撞性：很难找到两个不同的输入内容，得到相同的输出结果；
- 映射均匀性和差分均匀性：散列结果中 bit 位上的 0 的数量和 1 的数量应当大致相等；改变输入内容的 1 个 bit 信息会导致散列结果一半以上的 bit 位变化（雪崩效应）。

雪崩效应的本质就是散列结果的均匀性，因此，基本上可以说 MD5 散列函数的结果应当服从均匀分布。

受篇幅所限，这里无法给出详细的证明。不过我们可以以大量测试来说明这一结果。

如果我们将 MD5 的散列结果以十六进制的形式表达出来，那么表达的结果中可能出现 `1234567890abcdef` 这十六个可能的结果。如果十六个可能的结果等可能出现（均匀），那么对于某个十六进制位来说，它的信息熵等于 $\sum_0^{15} -\frac{1}{16}\log_2\frac{1}{16} = 4$。

这就是说，如果我们将 MD5 的散列结果以十六进制的形式表达出来，计算每一个数位的信息熵，如果（约）等于 4，就说明 MD5 的散列结果是（接近）均匀的。为此，我用 Python 写了一份代码作为测试。这份代码用到了 Python 的 `hashlib` 库，可能需要额外安装。

```
from hashlib import md5
from math import log

def entropy(wkList):
    wkSet   = set(wkList)
    rate    = {}
    lenList = len(wkList)
    for i in wkSet:
        rate[i] = float(wkList.count(i)) / lenList
    return sum([-p * log(p, 2) for p in rate.values()])

if __name__ == '__main__':
    wkDict = {} # key: index number; value: list of appeared chars
    for i in xrange(1000000):
        s = md5(str(i)).hexdigest()
```

```
        for j in xrange(32):
            if not j in wkDict:
                wkDict[j] = [s[j]]
            else:
                wkDict[j].append(s[j])

    for j in xrange(32):
        print j, '\t', entropy(wkDict[j])
```

运行之后结果如下：

```
0   3.99997252009
1   3.99999302786
2   3.99999519608
3   3.99998883009
4   3.99999198139
5   3.99999277151
6   3.99998601745
7   3.99998913662
8   3.99998403144
9   3.99998997451
10  3.9999888372
11  3.99999167561
12  3.99998973775
13  3.99998788689
14  3.99998465031
15  3.99999026671
16  3.99998951346
17  3.9999925505
18  3.99999145863
19  3.99999118615
20  3.99999429662
21  3.99998661919
22  3.99999172588
23  3.99998375623
24  3.99999562943
25  3.99998534411
26  3.9999892247
27  3.99998627499
28  3.99999072251
29  3.99999088822
30  3.99999107997
31  3.99998905455
```

经过一百万次的计算，我们发现，MD5 散列值的每一个十六进制位的信息熵都大致相等，且它们的值都约等于 4。根据之前的讨论，我们可以认为「MD5 散列函数的结果服从均匀分布」。