MD5值的电子取证应用研究

王聪 饶智韬 刘满果
（深圳海关缉私局刑事技术处，深圳518000）

【摘 要】电子数据对侦查办案起着重要作用，以MD5值为代表的哈希值能够保障电子数据的完整性和真实性。本文对常见的电子数据文件类型进行测试，分析影响MD5值的因素，总结MD5值在电子取证工作中的应用场景。
【关键词】MD5；哈希；取证；电子数据
【中图分类号】TP39【文献标识码】A【DOI】10.3969/j.issn.1672-2396.2020.01.038【文章编号】16722396[2020]58-00146-04

Application Research of MD5 Value Electronic Forensics

WANG Cong RAO Zhitao LIU Man'guo
(Criminal Technique Department of the Anti-smugging Bureau of Shenzhen Custom, Shenzhen 518000, China)

Abstract: Electronic data plays an important role in investigating and handling cases. The hash value represented by the MD5 value can ensure the integrity and authenticity of electronic data. This article tests common electronic data file types, analyzes the factors that affect MD5 values, and summarizes the application scenarios of MD5 values in electronic forensics.
Key words: MD5; Hash; forensics; digital data

0 引言

我国《刑事诉讼法》、《行政诉讼法》和《民事诉讼法》均明文规定“电子数据”作为法定证据类型之一，确定了“电子数据”的法律证据地位。作为一种新型证据，电子数据在案件侦办过程中的重要性逐渐凸显。而在实际工作中，环境影响和人为操作都会导致电子数据的改变和失真。面对电子数据改变和失真导致的完整性和真实性问题，以MD5值为代表的哈希值成为保障电子数据完整性真实性的重要手段，在案件取证以及检验鉴定的过程中具有重要意义。

1 哈希算法、MD5算法及MD5值

哈希算法又叫散列算法，是将任意长度的二进制值映射为较短的固定长度的二进制值，这个小的二进制值称为哈希值。它的原理其实很简单，就是把一段交易信息转换成一个固定长度的字符串。

MD5信息摘要算法是一种被广泛使用的哈希算法，可以产生出一个128位（16字节）的散列值，用于确保信息传输完整一致。MD5由美国密 码学家罗纳德•李维斯特设计，于1992年公开，用以取代MD4算法[1]。

MD5值是MD5算法将一个任意长度的数据经过编码得到一个128位（16字节）的哈希值。

MD5值具有以下特点：压缩性：任意长度的数据、算出的MD5值长度都是固定的；易计算：从数据计算出MD5值的速度相对其他算法较快；抗修改：对原数据进行任何改动，哪怕只修改一个字节，所得到的MD5值都有很大区别；强抗碰撞：已知原数据和其MD5值，想找到一个具有相同MD5值得数据（即伪造数据）是非常困难的；单向不可逆：给定MD5值，无法通过计算倒推出其输入原数据。

2 影响MD5值结果的因素的测试和分析

在实际案件取证和检验鉴定的过程中，MD5算法的计算对象可以是硬盘、分区或者特定文件。本文针

收稿日期：2020-04-18
作者简介：王聪（1992-），男，中国刑事警察学院，专业方向：刑事科学技术（电子物证）。

对常见的不同格式的电子数据，测试影响其MD5值结果的因素。

2.1 测试环境、测试软件和测试电子数据的文件格式

2.1.1 测试环境

操作系统：Windows 10 64 位（家庭中文版）；处理器：Intel(R)Core(TM)i7-8750H CPU @ 2.20GHz 2.21GHz；主板型号：惠普 84D8（HM370芯片组）；主硬盘：KXG50ZNV256G TOSHIBA(256GB/固态硬盘)。

2.1.2 MD5值计算软件

fHash:Flie’s Hash Calculator；软件版本号：1.8.5.0。

2.1.3 测试电子数据的文件格式

Word文档、Excel表格、PPT演示文稿、pdf文档、压缩文件、图片、视频文件、文本文档。

2.2 测试过程和结果

2.2.1 对word文档、excel表格、ppt演示文稿3种Microsoft office文档格式进行测试：

（1）新建以上3种Microsoft office文档，在文档内写入任意内容后保存，使用fHash:Flie’s Hash Calculator计算文档的MD5值。

（2）分别对新建的以上3种Microsoft office文档进行以下操作：复制、剪切、修改文件名、修改文件内容、修改文件后缀、转版本（使用不同Microsoft office文档软件打开同一Microsoft office文档并保存）、转wps（设置wps软件为文档默认打开方式后打开后保存文档）删除后还原、使用Microsoft office文档自带加密功能加密、使用Microsoft office文档自带加密功能加密后取消加密、修改文件属性中的只读和隐藏属性、不做任何操作只点击保存后关闭文档。

（3）使用fHash:Flie’s Hash Calculator计算进行以上操作后的Microsoft office文档的MD5值与其原MD5值进行对比。其中使用Microsoft office文档自带加密功能加密后取消加密的文档MD5值要分别与加密文档的MD5值以及原未加密文档的MD5值进行对比。

（4）以上测试结果如表1-表3所示。

表1 影响word文档类文件MD5值因素的测试

word	操作是否改变MD5值												
版本	复制	剪切	修改文件名	修改文件内容	修改文件后缀	转版本	转wps	删除后还原	加密	加密后取消加密与加密文件	加密后取消加密与原未加密文件	属性中修改只读和隐藏	不做任何操作只点击保存后关闭
2003	×	×	×	√	×	×	×	×	√	√	√	×	×
2007	×	×	×	√	×	×	×	×	√	√	√	×	×
2010	×	×	×	√	×	×	×	×	√	√	√	×	×
2013	×	×	×	√	×	×	×	×	√	√	√	×	×
2016	×	×	×	√	×	×	×	×	√	√	√	×	×

表2 影响excel表格类文件MD5值因素的测试

excel	操作是否改变MD5值												
版本	复制	剪切	修改文件名	修改文件内容	修改文件后缀	转版本	转wps	删除后还原	加密	加密后取消加密与加密文件	加密后取消加密与原未加密文件	属性中修改只读和隐藏	不做任何操作只点击保存后关闭
2003	×	×	×	√	×	×	×	×	√	√	√	×	√
2007	×	×	×	√	×	×	×	×	√	√	√	×	√
2010	×	×	×	√	×	×	×	×	√	√	√	×	√
2013	×	×	×	√	×	×	×	×	√	√	√	×	√
2016	×	×	×	√	×	×	×	×	√	√	√	×	√

表3 影响ppt演示文稿类文件MD5值因素的测试

ppt	操作是否改变MD5值												
版本	复制	剪切	修改文件名	修改文件内容	修改文件后缀	转版本	转wps	删除后还原	加密	加密后取消加密与加密文件	加密后取消加密与原未加密文件	属性中修改只读和隐藏	不做任何操作只点击保存后关闭
2003	×	×	×	√	×	×	×	×	√	√	√	×	×
2007	×	×	×	√	×	×	×	×	√	√	√	×	×
2010	×	×	×	√	×	×	×	×	√	√	√	×	×
2013	×	×	×	√	×	×	×	×	√	√	√	×	×
2016	×	×	×	√	×	×	×	×	√	√	√	×	×

2.2.2 对pdf文档格式进行测试

（1）新建pdf文档，在文档内写入任意内容后保存，分别设置Adobe Reader、福昕阅读器、得力阅读器、闪电阅读器、wps文档等不同pdf阅读器为pdf文档打开方式后打开并保存pdf文档，使用fHash:Flie’s Hash Calculator分别计算以上不同打开方式的pdf文档的MD5值。

（2）分别对以上不用打开方式的pdf文档进行以下操作：复制、剪切、修改文件名、修改文件内容、删除后还原、使用pdf文档自带加密功能加密。

（3）使用fHash:Flie’s Hash Calculator计算进行以上操作后的pdf文档的MD5值与其原MD5值进行对比。其中使用pdf文档自带加密功能加密的文档MD5值要与原未加密文档的MD5值进行对比。

（4）以上测试结果如表4所示。

表4 影响pdf文档MD5值因素的测试

pdf	操作是否改变MD5值					
阅读器	复制	剪切	修改文件名	修改文件内容	删除后还原	加密
adobe reader	×	×	×	√	×	√
福昕阅读器	×	×	×	√	×	√
得力阅读器	×	×	×	√	×	√
闪电阅读器	×	×	×	√	×	√
wps	×	×	×	√	×	√

2.2.3 对压缩文件格式进行测试

（1）使用rar、zip、7z等不同压缩软件对任一word文档进行压缩，使用fHash:Flie’s Hash Calculator分别计算以上使用不同压缩软件得到的压缩文件的MD5值。

（2）分别对以上使用不同压缩软件得到的压缩文件进行以下操作：复制、剪切、修改文件名、修改文件内容、修改文件后缀、删除后还原、使用压缩软件自带加密功能加密。

（3）使用fHash:Flie’s Hash Calculator计算进行以上操作后的压缩文件的MD5值与其原MD5值进行对比。其中使用压缩软件自带加密功能加密的文档MD5值要与原未加密压缩软件的MD5值进行对比。

（4）以上测试结果如表5所示。

表5 影响压缩文件MD5值因素的测试

压缩文件	操作对MD5的影响						
格式	复制	剪切	修改文件名	修改文件内容	修改文件后缀	删除后还原	压缩同一文件加密
rar	×	×	×	√	×	×	√
zip	×	×	×	√	×	×	√
7z	×	×	×	√	×	×	√

2.2.4 对图片格式进行测试

（1）使用fHash:Flie’s Hash Calculator分别计算jpg、jpeg、png等不同格式的图片的MD5值。

（2）分别对以上不同格式的图片进行以下操作：复制、剪切、修改文件名、修改文件后缀、删除后还原、使用PS软件PS图片、使用图片隐写工具对图片进行隐写。

（3）使用fHash:Flie’s Hash Calculator计算进行以上操作后的图片文件的MD5值与其原MD5值进行对比。

（4）以上测试结果如表6所示。

表6 影响图片MD5值因素的测试

图片	操作是否改变MD5值						
格式	复制	剪切	修改文件名	修改文件后缀	删除后还原	PS	图片隐写
jpg	×	×	×	×	×	√	√
jpeg	×	×	×	×	×	√	√
png	×	×	×	×	×	√	√

2.2.5 对视频文件格式进行测试

（1）使用fHash:Flie’s Hash Calculator分别计算wmv、rmvb、mp4、m4v、avi、mkv等不同格式的视频文件的MD5值。

（2）分别对以上不同格式的视频文件进行以下操作：复制、剪切、修改文件名、修改文件后缀、删除后还原、剪辑修改。

（3）使用fHash:Flie’s Hash Calculator计算进行以上操作后的视频文件的MD5值与其原MD5值进行对比。

（4）以上测试结果如表7所示。

表7 影响视频文件MD5值因素的测试

视频文件	操作是否改变MD5值					
格式	复制	剪切	修改文件名	修改文件后缀	删除还原	剪辑修改
wmv	×	×	×	×	×	√
rmvb	×	×	×	×	×	√
mp4	×	×	×	×	×	√
m4v	×	×	×	×	×	√
avi	×	×	×	×	×	√
mkv	×	×	×	×	×	√

2.2.6 对txt文本文档格式进行测试

（1）新建一个txt文本文档，写入任意内容后并保存，使用fHash:Flie’s Hash Calculator计算txt文本文档的MD5值。

（2）分别对以上txt文本文档进行以下操作：复制、剪切、修改文件名、修改文件内容、修改文件后缀、删除后还原。

（3）使用fHash:Flie’s Hash Calculator计算进行以上操作后的txt文本文档的MD5值与其原MD5值进行对比。

（4）以上测试结果如表8所示。

表8 影响txt文本文档MD5值因素的测试

txt	操作对MD5的影响					
	复制	剪切	修改文件名	修改文件内容	修改文件后缀	删除还原
txt	×	×	×	√	×	×

2.3 测试结果分析

根据以上实验结果，计算机文件的MD5值的改变与文件的实际内容有关，而不受文件名、文件后缀等文件属性的影响，验证了计算文件的MD5值，实际上是对文件的数据区（文件的实际内容）做校验，而文件的元数据（文件属性）的改变并不影响MD5值。即计算机文件MD5值的不变，其实际内容就没有改变，从而验证了计算机文件的完整性和真实性。其中Excel表格存在不进行任何操作点击保存后文件MD5值仍会发生变化的情况；而经过图片隐写术修改过的图片虽然表面上看没有发生变化，但其图片内容二进制数据已经被修改，所以MD5值也会发生变化。针对硬盘、光盘等存储介质计算的MD5值，主要因存储数据的改变而改变，在运输、保存以及使用的过程中，产生的磁盘坏道、物理损耗等因素也会影响其MD5值。

3 MD5值在电子取证中的应用场景

3.1 案件电子数据取证

不同于载体相对固定、易于保存且修改难度较大的传统证据，电子数据固有的易改变性给电子数据的真实性保障提出了新的要求。目前办案人员提取案件

电子数据的场景主要有电子数据现场勘查和网络在线提取电子数据，无论哪个场景，办案人员都需要确保电子数据的真实性。

以MD5值为代表的哈希值检验便是一个能有效解决当前电子数据真实性保障问题的手段。哈希值就像电子数据的“指纹”一样，不同的电子数据所对应的哈希值也有所不同。故可通过对比同一电子数据前后两组哈希值的方式来判断该数据在取得哈希值的两个时点间是否存在更改。

两高一部下发的《关于办理刑事案件收集提取和审查判断电子数据若干问题的规定》明文规定对作为证据使用的电子数据，可以采取以计算电子数据完整性校验值的方式保护电子数据的完整性[2]。这里的完整性校验值就是指以MD5值为代表的哈希值。具体来说，侦查人员在进行电子数据现场勘查以及网络在线提取电子数据时，可以计算提取到的案件电子数据的哈希值并做好记录。在电子数据审查判断环节，只需要对比电子数据的哈希值即可确认其真实性和完整性。

3.2 检验鉴定

为了保证流程的严谨、证据链的完整，鉴定人员进行电子数据检验鉴定时可以选择制作原始电子数据存储介质的镜像文件。制作镜像文件时应计算源数据及目标数据（克隆盘或镜像文件）的哈希值以确保电子数据的原始性和完整性，而计算镜像文件最常用的哈希值就是MD5值[3]。

电子数据检验鉴定提取的电子数据以及存放提取电子数据的存储介质同样需要进行完整性校验并记录在鉴定报告中以确保鉴定人员提取电子数据的完整性和真实性。CNAS-CL27文件《司法鉴定—法庭科学机构能力认可准则在电子物证鉴定领域的应用说明》明确了进行电子数据检验鉴定应记录检出数据的完整性校验值。

CNAS认可的司法鉴定/法庭科学机构认可领域分类（详见CNAS—AL13文件）中电子数据鉴定一共有三类，分别是电子数据的提取、固定与恢复和电子数据真实性（完整性）鉴定以及电子数据同一性、相似性鉴定。

3.3 哈希库比对

取证人员可以将已知文件（如常见软件、操作系统文件、恶意程序或图像文件）的形成哈希值的集合，构建哈希库，使用哈希库可以在电子数据取证的过程中快速发现或者排除此类文件。哈希库可以用于识别那些没有取证意义的文件，如操作系统文件和常用的应用程序文件。哈希库也同样可以用于检测对取证有重要意义的特定文件。例如取证人员建立了一些重点文件哈希库（如儿童色情图片、涉恐音视频文件和用于非法目的的特定应用程序），一旦在电子数据取证的过程中发现存储介质中（硬盘或手机等）存储了此类文件，即可进行快速比对，检查出与哈希库匹配的文件，从而提升取证的排查效率。

参考文献

[1]李皓.聊聊常见的三大加密算法[J].计算机与网络,2017,43(11):52-54.
[2]孙君梁.关于刑事诉讼中电子取证问题的相关研究[J].法制博览,2019(25):105+107.
[3]马锐.刑事案件电子数据取证规范化研究[D].兰州大学,2018.