

一种基于内容特征的 Word 文件雕复方法

陈 默 郑 宁 徐 明 楼永坚 汪 霞
(杭州电子科技大学计算机学院 浙江 杭州 310018)

摘 要 提出一种不依赖于文件系统元信息,而凭借于文件数据内容及其内部结构特征的 Word 文件雕复方法,其基本原理是利用文件头/根存储/最大扇区、分片文件的扇区分配表和分片文件的数据流等验证方法。此雕复方法能自动雕复在原始磁盘镜像中连续和分片有序存储的 Word 文件。实验结果表明该方法可以在 Word 文件自动雕复的高准确率情况下,确保低“误报”率。

关键词 文件雕复 内容特征 Word 文件

A WORD FILE CARVING METHOD BASED ON CONTENT CHARACTER

Chen Mo Zheng Ning Xu Ming Lou Yongjian Wang Xia
(College of Computer Hangzhou Dianzi University Hangzhou 310018, Zhejiang China)

Abstract This paper presents a Word File carving method based on file's content and the character of its internal structure but not depend on the metadata of file system. Its basic theory is to make use of header/ root storage/max sector validation SAT of the fragmented file validation and data stream of the fragment file validation. This carving method can carve automatically the contiguous Word files and the fragmented Word files stored in order in original disc mirroring. Experiments show that the method can ensure lower "false positive" rate while the Word files are carved automatically with a high accuracy.

Keywords File carving Content character Word file

0 引 言

文件雕复技术是一种特殊的文件恢复技术,被广泛应用于计算机取证研究^[1-5]。它利用不同类型文件特有的文件信息从未分配的或空闲空间中雕复文件,可恢复隐藏或删除的磁盘内容、内存和交换分区中的数据而无须文件系统的支持。随着 MS Word 的普及,如何恢复无文件系统信息或文件系统元信息已损坏的磁盘镜像中的 Word 文件,特别是分片的 Word 文件已成为研究者关注的问题。目前,文件雕复的技术有文件头/嵌入长度雕复^[3-5]、分片恢复雕复^[3-5]、语义雕复^[3-5]、Smart 雕复^[4-5]和人工雕复^[6]。但是,现有的这些文件雕复技术没有充分利用 Word 文件的特殊性,在恢复 Word 文件过程中仍然存在一些缺陷:大量“误报”文件(已雕复的文件包含无效数据)、雕复过程自动性差,不能自动雕复有序分片较多(二个以上)的 Word 文件。

针对上述问题,本文在深入了解 Word 文件的内部结构和研究现有 Word 文件雕复技术的基础上,提出了基于内容特征验证的 Word 文件雕复方法。该方法主要采用 Word 文件中的根存储和最大扇区相结合来确定完整文件的尾部,通过特征匹配来确定 Word 分片文件中分散的扇区分配表,利用熵、熵差和语义验证来搜索和匹配分片的数据流。

1 Word 文件简介

Word 文件采取 Unicode 的形式,且为尽可能减少存储量,采取了压缩方式,并以 ASCII 码来表示字符。Word 文件使用微软

复合文档文件格式,以结构化形式来存储文件,基本原理与微软文件系统类似^[7]。复合文档文件的标准结构可参考 Office 文档的 Spreadsheet 方案^[8]。每个 Word 文件都由一个 512 字节的头结构和紧随的一系列扇区组成,其中包含一个主扇区分配表、一个扇区分配表、一个短扇区分配表、一个目录和一个或多个数据流^[3-8],而主扇区分配表的首部存放于头结构中(如图 1 所示)。

头结构	
0	B 数据流 0
1	A 数据流 0
2	目录 0
3	B 数据流 1
4	A 数据流 1
5	扇区分配表 0
6	A 数据流 2
7	扇区分配表 1
8	目录 1
9	短扇区分配表
...	...

图 1 Word 文件结构

Word 文件的文件头“D0CF11E0A1B11AE1”是头结构的前 8 个字符,无直接的文件尾特征,而不同于其他 Office 类型文件的特有特征是“57006F007200640044006F00630075006D0065006E007400”(Word Document)。Word 文件的每个目录实体长度为 128 字节,其中根存储的特征为“52006F006F007400200045006E

收稿日期:2008-08-08 浙江省自然科学基金项目(Y106176);浙江省科技计划项目(2007C33058)。陈默,硕士生,主研领域:网络信息安全。

00740072007900" (RootEntry)。

2 基于内容和文件内部结构特征的 Word 文件雕复方法

基于内容和文件内部结构特征的 Word 文件雕复方法主要包括文件头/根存储/最大扇区、分片文件的扇区分配表和分片文件的数据流的验证等方法，其中分片文件的数据流验证方法由熵差和熵结合验证及语义验证组成。Word 雕复的流程如图 2 所示：先根据文件头/根存储/最大扇区验证方法扫描原始磁盘镜像，确认完整的 Word 文件；未确认文件都视为分片文件，由分片文件的扇区分配表验证方法寻找分散的扇区分配表并确认其是否完整，若扇区分配表不完整，则认为分片文件为破损的 Word 文件（丢失部分文件数据），否则由熵差和熵结合验证方法及语义验证方法搜索其数据流并检验是否完整，如果文件中所有分片的数据流被匹配成功，那么可确认此 Word 分片文件，如果有一处匹配失败，则视其为破损的 Word 文件；最后雕复已确认的 Word 文件。

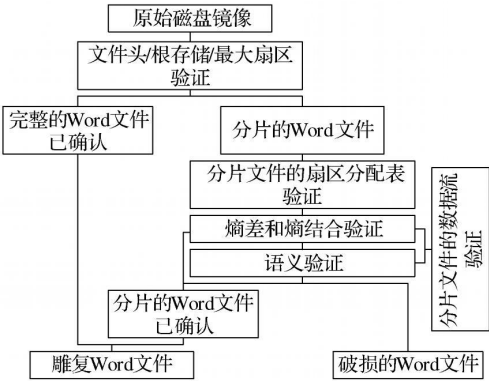


图 2 基于内容和文件内部结构特征的 Word 雕复流程

2.1 文件头/根存储/最大扇区验证方法

验证文件头/根存储/最大扇区的验证方法用于确认 Word 文件是否完整及确定其尾部位置。Word 文件没有直接的文件尾特征，但文件尾部一般为数据流或 Word 的版本信息。尾部的数据流可以通过计算扇区分配表中的最大扇区号来获得其位置，而 Word 的版本信息大多包含特征“4D693726F736F667420” (Microsoft)，可以在根存储中找到它对应的位置。算法 1 总结了验证文件头/根存储/最大扇区方法的具体操作。

- 算法 1 文件头/根存储/最大扇区验证算法
- 1) 搜索原始磁盘镜像中的 Word 文件头特征，确定 Word 文件个数 n ；Word 文件 W_i ($i=1, 2, \dots, n$) 和扇区块大小 Sec 。
 - 2) 搜索 W_i 中根存储特征，确定其位置 R_i ；搜索扇区分配表 SAT_i 。如果 R_i 不存在或 SAT_i 不完整，则视 W_i 为分片的 Word 文件，验证结束；否则比较扇区分配表中的扇区号，获得最大扇区号 Max_i 。
 - 3) 如果 $Max_i > R_i$ ，则确定 W_i 的尾部是数据流，其结束位置为 $(Max_i + 1) \times Sec + Sec$ ；算法结束；否则确定 W_i 的尾部是 Word 版本信息，在根存储中搜索 Word 的版本信息位置 R_i ，其结束位置为 $(R_i + 1) \times Sec + Sec$ ；算法结束。
- 通过算法 1 可确定 Word 完整文件的头部和尾部位置，而不可确认的 Word 文件都视为分片文件，需通过其他验证方法来鉴别和匹配分片。

2.2 分片文件的扇区分配表验证方法

分片文件的扇区分配表验证方法主要是寻找分片的 Word 文件中的扇区分配表。Word 文件的头结构含扇区分配表中的总扇区数，扇区分配表的第一个扇区的头部特征一定为“01000000020000000300000004000000”，如果总扇区数大于 1，则第二个扇区的头部特征一般为“81000000”，即扇区分配表中的两个邻近的扇区之间相差值为 128。或者其头部特征是具特殊意义的特征“FEFFFFFF”、“HFFFFFFF”和“FCFFFFFF”，分别表示数据流的终止符、作为扇区分配表的扇区和作为主扇区分配表的扇区。Word 分片文件中的扇区分配表大致可分为三种情况：1) 第一个分片含完整的扇区分配表；2) 其他分片含完整的扇区分配表；3) 扇区分配表被分散于不同的分片中。后二种情况的扇区分配表的位置发生改变，那么按上述理论从文件头往下查找头部特征符合的扇区，一旦找到，则验证成功；否则确认此 Word 文件是破损文件。其中扇区分配表因分片而发生位置改变的间隔值 Gap_i 将被记录。

2.3 分片文件的数据流验证方法

Word 文件的数据流包括文本数据、压缩数据和格式信息等。若一个 Word 分片文件含大量的文字、图片等数据信息时，那么文件中的数据流很可能被分片。分片文件的数据流验证方法针对不同的数据流采取多种验证方法结合的方式，以提高验证结果的准确性。

- 熵差和熵结合验证方法
- 熵又叫香农熵，是一个随机事件的不确定性的量度^[9]。Shannon^[10]定义熵的公式为：

$$H(x) = - \sum_{i=1}^n P(i) \log_2 P(i) \tag{1}$$

其中 $H(x)$ 表示熵， $P(i)$ 表示 i 事件的概率。熵值越大，表明事件的不确定性越大，信息量也越大。所以很多研究者就利用熵来鉴别不同的文件类型^[4,5,11]，且 Luan 等指出压缩文件和加密文件的熵值较高^[9]，一般大于 7。Word 文件包含不同数据类型，所以不同类型的数据流有不同的熵值，而一般数据流的熵值为 3—6 之间。本文为了能够鉴别同一类型文件中的相似数据，在研究熵的基础上提出了熵差的概念，其算式为：

$$HD(x) = \left| \left(- \sum_{j=1}^m P(j) \log_2 P(j) \right) - \left(- \sum_{i=1}^n P(i) \log_2 P(i) \right) \right| \tag{2}$$

($j=i+1, m=n=256$)

其中 $HD(x)$ 表示熵差， n 表示某一个扇区的末尾 256 个字节数据， m 表示紧随的扇区的头 256 个字节数据。

熵差和熵结合验证方法先用熵差验证数据流中的邻近两个扇区 (S_x 和 S_y) 是否存在分片，如果存在分片，则根据间隔值 Gap_i 计算出 S_y 对应扇区 S'_y ，采用熵差验证 S_x 和 S'_y 是否为相连的扇区。如果匹配正确，则确认两个分片；否则表示 S_x 并非为分片尾部扇区或者 S_x 和 S'_y 之间还存在其他分片，采用熵值验证扇区。算法 2 总结了熵差和熵结合验证方法的具体操作。

- 算法 2 熵差和熵结合验证算法
- 1) 根据分片文件 W_i 的扇区分配表初步确定可能含分片的数据块 D 。
 - 2) 根据熵差公式 (2) 计算出 D 中熵差值大于 1 的扇区 S 并获得最小的扇区号 Mip 。
 - 3) 取 Mip 对应的扇区 S_x 及下一个扇区 S_y 对应的扇区 S'_y ，计算两者熵差值，若此值 < 1 ，则确认 S_x 和 S'_y 分别是分片的最

后一个扇区和下一个分片的第一个扇区, 算法结束, 否则继续;

4) 根据熵公式, 计算 S_x 和 S_y 的熵值, 如果 S_x 的熵值 < 6 且 S_y 的熵值不属于 $1-7$ 之间, 则确定 S_x 和 S_y 之间还有分片, 确认 S_x 是分片的最后一个扇区, 继续, 否则确认 S_x 不是最后的扇区, 算法结束;

5) 根据熵公式, 在 S_x 后寻找熵值为 $3-6$ 之间的扇区, 一旦找到, 则确认为下一个分片的第一个扇区 S'_y , 计算 S_x 和 S'_y 的间隔值 Gap_i , 根据熵差和熵继续寻找其他分片的扇区, 直到所有间隔值的和等于 Gap_i , 则确认此 Word 分片文件, 算法结束; 否则, 算法结束。

算法 2 中不能被确认的 Word 分片文件存在二种情况: 1) 此分片文件实际上是破损的 Word 文件; 2) 用熵差和熵结合的验证方法不能匹配某些分片。这二种情况也说明分片的数据流与冗余数据之间的相似度很高, 需要采用其他验证方法加以识别, 比如语义验证。

。语义验证方法

语义验证方法是用于确定 Word 分片文件中的文本数据流与邻近的数据是否一致, 如果不一致, 则会自动寻找相一致的数据流, 其原理与语义修复类似。

如果扇区不能被熵差和熵结合的验证方法所鉴别, 则取其尾部最后 16 个字符, 通过非字母字符的匹配和过滤, 获得最后一个单词的部分字母或者空值 (扇区最后以标点符号结尾) C_1 ; 然后取邻近扇区的头 16 个字符, 通过非字母字符的匹配和过滤, 获得第一个单词的部分字母或空值 (扇区第一个字符为空格或标点符号) C_2 , 连接 C_1 和 C_2 , 得到单词 C 匹配单词库, 看单词 C 是否有效, 如果有效, 则确认两个扇区邻近; 否则说明邻近扇区为冗余数据, 需向下继续匹配扇区数据, 直到单词 C 为有效或者测试的数据块全部匹配失败; 如果数据块全部匹配失败, 则确认此 Word 分片文件为破损文件。

3 实验比较与分析

实验环境: Fedora 5 操作系统 (2.6.20 内核), Intel P4 (2.66GHz), 512MB 内存, 双硬盘 (40GB+120GB)。实验数据集为数字取证研究组 2006 年文件修复难题挑战数据集^[13], 大小为 50MB, 包含 5 个 Word 文件, 分别处于五种不同的情况: 1) 完整文件; 2) 含 3 个分片, 随机数据穿插其中; 3) 嵌入一个 JPEG 文件中; 4) 2 个分片, 分片中间是一个 JPEG 文件; 5) 2 个分片, 分片中间是一个文本文件。对我们的方法同 Foremost (一种常用的文件修复工具^[14,21]) 和 Revit (Mez 等人设计的文件修复工具^[14,9]) 进行比较实验, 结果如表 1 所示。

表 1 实验结果

实验算法	执行时间	修复结果	准确度
Foremost	1 149 秒	2 个完整文件	40%
Revit	1 分 19.633 秒	3 个完整文件, 2 个部分数据文件	60%
本文方法	14 236 秒	5 个完整文件, 7 个临时的分片	100%

1) 执行时间 由于实验数据集本身比较小, 三者执行时间都比较短。Foremost 采用文件头/文件尾的修复技术^[12], 所以执行时间非常短, 只有 1.149 秒; Revit 和本文方法都需对文件分片进行匹配处理, 而 Revit 同时还要恢复数据集中的其他类型文件, 所以 Revit 的执行时间最长, 为 1 分 19.633 秒, 本文方法次之, 为 14.236 秒。

2) 修复结果 本文方法将修复的文件分为完全修复和破损修复两种, 由于实验数据集中不存在破损的 Word 文件, 所以在实验结果中只有完全修复的结果, 但会区分实验数据集中的完整文件和分片文件。Foremost 只修复实验数据集中的完整文件, 所以基本没有“误报”文件。Revit 将文件分为三种类型——完整、嵌入和部分数据, 它能修复实验数据集中的 2 个完整的 Word 文件和 1 个含 2 个有序分片的 Word 文件, 但同时修复出 2 个“误报”文件。本文方法修复分片文件时, 先将文件中的各分片分别修复出来, 以文件名+“_”+编号的形式命名, 最后将各分片连接并修复出完整的分片文件, 故这些临时的分片不能算“误报”文件。从本文方法的修复结果可知, 本文方法能够自动修复实验数据集中的 2 个完整的 Word 文件和 3 个含有序分片的 Word 文件。

3) 准确度 通过求准确度 $E(x) = Cn(x)/Tn$ 来衡量修复后的文件的完整性, 其中 x 表示修复方法, $Cn(x)$ 表示采用 x 修复出的 Word 完整文件数量, Tn 表示实验数据集中的 Word 文件总数。Foremost 只修复出实验数据集中的 2 个完整的 Word 文件, 故准确度为 40%。Revit 可以修复完整的和含有序分片的 Word 文件, 但修复分片文件尚存在不足, 只能修复 1 个含 2 个有序分片的 Word 文件且修复结果存在“误报”文件, 不过它的准确度比 Foremost 要高, 达到 60%。本文方法能够自动匹配同一文件的分片, 准确地识别 Word 完整文件和含有序分片的 Word 文件, 所以能修复实验数据集中的所有 Word 文件, 准确度达到 100%。

4 结 论

本文利用 Word 文件中数据内容及其内部结构特征, 提出了一种能自动修复在磁盘镜像中连续和分片有序存储的 Word 文件方法。实验结果表明, 基于文件内容和内部结构特征的 Word 文件修复方法, 可以在高准确率情况下, 确保低“误报”率, 并实现了修复过程的自动化。今后的工作是用该方法测试更多的实际的磁盘镜像数据, 并研究破损的 Word 文件的修复技术。

参 考 文 献

[1] Golden G Richard III, Vassil Roussey. Scalpel: A frugal high performance file carving [J] // Proceedings of the 2005 Digital Forensic Research Workshop. New Orleans, LA, 2005

[2] Nicholas Mikus. An analysis of disc carving techniques [D]. Monterey Naval Postgraduate School, 2005

[3] Simson L Garfinkel. Carving contiguous and fragmented files with fast object validation [J]. Digital Investigation, 2007, 4 (supplement 1), 2-12

[4] Joachin Metz, Robert Jan Moray. Analysis of 2006 DFRWS forensic carving challenge [EB/OL]. (2007-3-21). http://sandbox.dfrws.org/2006/moray_dfrws2006.Pdf

[5] Joachin Metz, Bas Kooet, Robert Jan Moray. Analysis of 2007 DFRWS forensic carving challenge [EB/OL]. (2007-8-28). http://sandbox.dfrws.org/2007/metz_dfrws2007_carving_challenge.Pdf

[6] Glenn Henderson, David Horvath, Jeff Jones. Submission for the 2006 DFRWS Forensics Challenge [EB/OL]. (2007-3-21). http://sandbox.dfrws.org/2006/buchholz_inuwriteup.Pdf

(下转第 126 页)

Step6 提取图像边缘。采用“max”和“min”相结合的运算对增强后的图像作边缘检测。图像的边缘定义如下:

$$Edges=\bigcup_m\bigcup_n x'_{mn} \tag{11}$$

式中:

$$x'_{mn}=\left|\max_Q(x)-\min_Q(x)\right|,\left(i,j\right)\in Q \tag{12}$$

Q可取以坐标(m,n)为中心的3×3的窗口。

2.2 仿真分析

针对本文提出的算法,我们选用医学院提供的一个脑部CT图像在MATLAB环境下进行了仿真实验。为了对比实验结果,我们还用各种边缘检测方法在MATLAB环境下分别进行了仿真工作。我们先对图1利用传统的LOG算子和Sobel算子在MATLAB环境下进行边缘检测,检测结果如图2—图3所示。接着我们利用传统的Pai King算法对图1进行仿真,仿真结果如图4所示。然后我们在Pai King算法的基础上将隶属函数改成简单的P=1/MAX形式^[10],又进行了简单的仿真实验,仿真结果如图5所示。最后利用我们提出的改进算法在MATLAB环境下进行仿真实验,其过程为先用Ostu算法来确定最佳阈值参数,并在此基础上提出新的隶属函数,再对图像进行模糊增强,并且对增强后图像进行平滑处理、边缘检测等操作,最终仿真结果如图6所示。

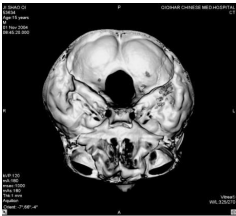


图1 原始图像



图2 LOG算法结果

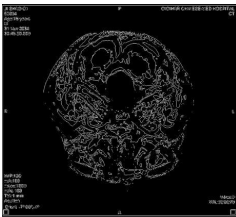


图3 Sobe算法结果

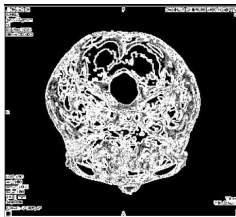


图4 Pai king算法结果



图5 P=1/MAX改进算法

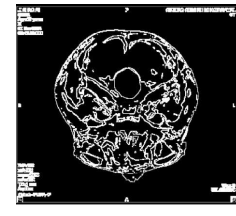


图6 本文改进算法结果

由于人脑CT图像的复杂性,边缘提取具有相当的难度,如图1所示,图像中间区域的灰度变化情况比较复杂,常用的边缘检测算子无法将其较好地检测出来。从实验结果可看出,图3 Sobe算法的检测效果最差,甚至出现边缘的损失。图2 LOG算法和图4 Pai King算法虽然能将边缘检测出来,但是出现了很多无用的边缘,这是由于人脑图像表面光线强度不均造成的。如图1中人脑顶部因光线原因产生的明亮区域边缘也被检测出来了,这不利于以后对图像作进一步处理与重建工作。图5与图6中的两种算法检测的效果较好,但是通过进一步的观察不难看出,本文改进算法的边缘检测效果从整体和细节上都优于图5检测效果。

经过反复实验证明,通过设置合适的迭代次数和P,可以

得到符合要求的检测图像,这里的取2或3,P取0.7以上均有较好的检测效果。

3 结 论

本文针对Pai King算法存在的问题,提出新的算法。通过选取最佳阈值定义一个新的隶属函数,再在模糊特征平面上分区应用模糊增强算子多次提高不同区域对比度,最后提取边缘。仿真结果表明,经过这样模糊增强后的图像区域之间的对比度大幅增强,层次更加清楚,最后检测出的边缘更加清晰。因此,本算法是一个适用于人脑CT图像的新型高效实用的边缘检测算法。今后的研究重点是能否将此算法应用到更多类型图像的图像分割预处理中。

参 考 文 献

[1] 罗玉玲,唐贤英.基于阈值优化的图像模糊边缘检测算法[J].微计算机信息,2007,23(23):286-288.
[2] Lee J,Haralick R M,Shapiro LG.Morphology Edge Detection[J].IEEE Trans RA,1987,3(4):142-156.
[3] Mallat S,Hwangw J.Singularity detection and processing with wavelets[J].EEE Trans on Information Theory,1992,38(2):679-681.
[4] 万寿红,龚育昌,等.基于多分辨率分析的肺部边缘检测[J].计算机应用与软件,2007,24(10):109-111.
[5] Pal S,K King R A.On edge detection of X-ray images using fuzzy sets[J].EEE Trans Patt Ana and Machine Intell,1983,5(1):69-77.
[6] 成培,李峰.图像模糊边缘检测算法的改进[J].电子技术应用,2006(12):31-32.
[7] 马志峰,杨水超,赵保军,等.改进的快速图像模糊边缘检测算法[J].激光与红外,2005,35(4):300-302.
[8] 蒋家伏,刁洪祥,唐贤瑛,等.一种基于粗糙记得图像增强方法[J].计算机工程与应用,2002,39(19):109-110.
[9] 商泽利.图像随机值脉冲噪声去除[D].西安:西安电子科技大学,2007.
[10] 郑春红,焦李成,等.一种快速模糊图像边缘检测算法[J].计算机工程与应用,2004,32:48-50.

(上接第102页)

[7] Hyukdon Kwop,YeogK in,Sangjin Lee.A Tool for the detection of hidden data in Microsoft compound document file format[C]//CISS 2008,Proceedings of the 2008 International Conference on Information Science and Security.Washington DC,USA,2008.
[8] Daniel Renz.OpenOffice.org's documentation of the Microsoft compound document[EB/OL].(2007-8-7).http://sc.openoffice.org/comfileformat/Pdf.
[9] Haiying Luap,Simon Mackey.Entropy analysis[EB/OL].(2006-4-21).http://polya.computing.dcu.ie/wiki/index.php/Entropy_Analysis.
[10] Shannon C.E.A mathematical theory of communication[J].Bell Sys.tem Technical Journal,1948,27:379-423,623-656.
[11] Cor J,Veerman.Statistical disk cluster classification for file carving[C]//AS2007,Proceedings of the Third International Symposium on Information Assurance and Security.Manchester,UK,2007.
[12] DFRWS.DFRWS 2006 forensics challenge[OL].(2006-8-14).http://www.dfrws.org/2006/index.shtml.