МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»
(СГУ)

Кафедра теоретических основ компьютерной безопасности и криптографии

Восстановление повреждённых файлов

АВТОРЕФЕРАТ

дипломной работы

студента 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Кобзева Михаила Алексеевича

Научный руководитель
доцент _____ И. Ю. Юрин
18.01.2019 г.

Заведующий кафедрой
д. ф.-м. н., доцент _____ М. Б. Абросимов
18.01.2019 г.

Саратов 2019

ВВЕДЕНИЕ

На текущий момент сложилась ситуация, когда существует очень мало способов восстановить поврежденный файл, несмотря на то, что повреждаются они довольно часто. Например, сейчас существует множество программ, которые восстанавливают удаленные файлы или файлы с поврежденных носителей, однако эти программы не учитывают внутреннюю структуру файла. В результате, после восстановления может оказаться нечитаемым.

Целью данной дипломной работы было изучить виды повреждений файлов и форматы файлов, употребляемые наиболее часто. Было необходимо выделить один из форматов, изучить его более подробно и составить программу для восстановления поврежденных файлов этого формата.

Дипломная работа состоит из введения, четырех разделов, заключения, списка использованных источников и 1 приложения. Общий объем работы – 64 страницы, из них 44 страницы – основное содержание, включая 38 рисунков и 2 таблицы, список использованных источников из 20 наименований.

КРАТКОЕ СОДЕРЖАНИЕ

1. Причины возникновения повреждений в файлах

Сначала рассмотрим неповрежденный — или целостный файл — и его расположение на носителе данных. Целостный файл состоит из блоков, записанных на какой-либо носитель данных. Для простоты далее в работе любой носитель будет называться «диском».

Целостный файл лежит на диске в виде блоков. При чтении файла блоки, из которых состоит файл, составляются в определенной последовательности, и в результате получается образ файла в памяти, готовый к обработке программой.

Поврежденный файл — это файл, претерпевший ошибочные изменения внутренней структуры и данных. В результате таких изменений файл приходит в непригодное или неудовлетворительное для использования состояние. [1]

Повреждение может носить как критический характер, когда пользователь полностью теряет доступ к данным, так и частичный, когда теряется только часть данных, но остается доступ к остальной, нетронутой повреждениями, части информации.

Повреждение файла может возникнуть в следующих случаях:

- Восстановление файлов после удаления с диска. Чаще всего в этом случае получается несколько нечитаемых секторов в конце. [1]
- Сбой в работе диска. В этом случае в файле находятся несколько поврежденных секторов, данные из которых некорректны или их нельзя считать. [3]
- Воздействие на файл вредоносной программы или антивируса. В этом случае поврежденные участки могут иметь самый разный размер и не совпадать с размерами сектора.
- Ошибки работы приложений. Чаще всего файлы могут быть повреждены при сохранении их на диск. При этом так же, как и в предыдущем пункте, часть файла может быть перезаписана мусорными данными. [4]

2. Структуры форматов, употребляемых наиболее часто.

2.1 Структура файлов формата JPEG

Файл формата JPEG состоит из заголовка и кодовой секции. Заголовок является последовательностью маркеров, после каждого из которых идет размер отмечаемой им секции, а затем сама информация, хранящаяся внутри маркера. [5] Наиболее часто встречаются следующие маркеры:

[FFD8] — маркер начала. Он всегда находится в начале всех JPEG-файлов.

[FFFE] — маркер, означающий начало секции с комментарием.

[FFE0] — маркер секции приложений (APP0) — обозначает, что файл использует спецификацию JFIF[6].

[FFC0] — маркер секции кодировки (базовая или progressive).

[FFC4] — маркер, указывающий на секцию с таблицами Хаффмана.

[FFDA] — маркер секции начала сканирования.

[FFD9] — маркера конца файла

Сразу после секции начала сканирования и до маркера конца файла — идут закодированные данные. [8].

2.2 Описание формата docx

Одним из наиболее используемых текстовых форматов является формат docx. Сам файл формата docx представляет собой zip-архив, который содержит два типа файлов:

1. Xml-файлы с расширениями xml и rels;
2. Медиа-файлы (изображения, аудиофайлы и т.д.).

Логически эти файлы представляют собой три вида элементов:

1. Типы (Content Types) — список типов медиа-файлов (например, png) встречающихся в документе и типов частей документов;
2. Части (Parts) — отдельные части документа.;
3. Связи (Relationships) — идентифицируют части документа для ссылок, а также тут определены внешние части.

Теперь рассмотрим составные части более подробно.

[Content_Types].xml

Находится в корне документа и перечисляет MIME-типы[1] содержимого документа.

_rels/.rels

Главный список связей документа.

word/document.xml

В этом файле содержится основное содержимое документа.

word/_rels/document.xml.rels

Здесь содержится список связей части word/document.xml. Название файла связей создаётся из названия части документа, к которой он относится, и добавления к нему расширения rels. Даже если в этой части документа связей нет, этот файл должен существовать. [13]

2.3 Описание формата MP3

Файл формата MP3 состоит из заголовка, использующего формата метаданных ID3v2(от англ. Identity a MP3), и последовательности MP3-фреймов. Структура заголовка ID3v2:

- маркер всегда равен ‘ID3’;
- В данный момент имеются три версии ID3v2.2, ID3v2.3 и ID3v2.4;
- Флаги. В настоящее время используются только три (5,6,7) бита;
- Длина.

После ID3v2-заголовка следует MP3-заголовок. Заголовок состоит из 4 байт, где:

- Биты 0-10 всегда единицы — это маркер фрейма;
- Далее 2 бита являются индексом версии MPEG;
- Ещё два бита указывают на версию Layer; [15]
- Далее следует бит, определяющий наличие защиты;

[1] (Multipurpose Internet Mail Extension или Многоцелевые расширения почты Интернета — спецификация для передачи по сети файлов различного типа: изображений, музыки, текстов, видео, архивов и др.) [11]

- Следующие 4 бита определяют индекс битрейта;
- Ещё 2 бита определяют индекс частоты;
- Далее идет бит смещения. Если он есть, то данные смещаются на 1 байт;
- Следующий бит — бит private. Не используется;
- Затем два бита определяют режим канала;
- Ещё два бита используются только вместе со смешанным стерео-режимом. Определяют расширение режима канала;
- Следующие два бита определят наличие копирайта и факт копирования файла соответственно;
- Последние два бита в настоящий момент не используются.

Длина самого фрейма вычисляется по следующей формуле:

$$\text{Размер фрейма} = \frac{\left(\frac{\text{количество сэмплов в фрейме}}{8} * \text{битрейт}\right)}{\text{частота сэмпла}}$$

Вычислив размер фрейма остается только считать идущие дальше данные, представленные в виде набора частот и амплитуд. [16]

3. Обзор существующих решений для восстановления поврежденных файлов

Ниже представлены существующие решения, которые работают с наиболее часто встречающиеся типами файлов:

3.1 PixRecovery

PixRecovery — программа для восстановления поврежденных фотографий в формате JPEG, снятых на Kodak, Nikon, Sony, Fuji и другие цифровые камеры.

3.2 RS File Repair

С помощью RS File Repair можно восстановить поврежденные изображения, фотографии в формате JPEG.

3.3 JPEGfix

JPEGfix — программа для ремонта и анализа поврежденных JPEG-файлов.

3.4 JPEG Recovery Pro

JPEG Recovery Pro — программа для восстановления файлов с расширением JPEG.

3.5 All Media Fixer

All Media Fixer — простая программа для восстановления разнообразных файлов.

3.6 MP3val

MP3val — программа для восстановления файлов формата MP3.

3.7 Recovery Toolbox for Word

Recovery Toolbox for Word — программа для восстановления поврежденных файлов формата doc.

4. Практическая часть

Для реализации программы для восстановления файлов был выбран формат JPEG. В ходе работы была создана программа на языке Python[18], которая анализирует поданный на вход файл и производит восстановление его внутренней структуры.

4.1 Описание принципа работы программы

Рассмотрим возможные варианты восстановления фотографии для каждого из случаев.

Первый случай — поврежденный фрагмент целиком находится в заголовке файла. В этом случае возможность восстановления зависит от местоположения поврежденного участка. [20]

Второй случай — повреждения затрагивают стык заголовка и секции данных. В этом случае возможность восстановления зависит от местоположения поврежденного участка.

Третий случай — повреждения затрагивают только секцию данных. В таком случае можно попытаться восстановить большую часть информации из поврежденного файла.

Повреждение делит секцию данных на три части: участок, идущий до поврежденного фрагмента, сам поврежденный фрагмент, участок после него.

Очевидно, что участок, идущий до поврежденного фрагмента, можно восстановить без потерь качества изображения. Затем добавляется заглушка — пустые MCU-блоки, которые вставляются вместо тех, что содержались в удаленном фрагменте. Далее надо попытаться восстановить часть файла после поврежденного фрагмента.

При восстановлении части файла, которая идет следом за поврежденным участком, могут возникнуть проблемы с цветопередачей. Однако, как показало проведенное мною исследование, существует возможность с помощью автоматизированного перебора значений в первом MCU-блоке приблизить результат работы программы к изначальному виду файла.

ЗАКЛЮЧЕНИЕ

В ходе работы были рассмотрены разнообразные причины повреждений и их возможные размер и местоположение в файле. Были рассмотрены наиболее часто встречающиеся форматы файлов и сделаны выводы о возможности их восстановления. Особое внимание было уделено формату JPEG. Этот формат был наиболее детально разобран и была составлена программа для восстановления поврежденных JPEG-файлов.

Полученный в ходе работы опыт по восстановлению поврежденных файлов может быть применен и к другим типам файлов. Например, с некоторыми доработками, алгоритм может быть применен и к файлам формата MP3.

Так как большая часть существующих решений распространяется на коммерческой основе, выводы относительно их функционала делались на основании данных, имеющихся в открытом доступе. Созданная мною программа очень проста в применении и не требует от пользователя каких-либо специальных знаний о формате JPEG, в отличие от свободно распространяемой JPEGFix. При этом качество восстановления не сильно уступает коммерческим решениям.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Хабрахабр. Что такое поврежденный файл? OfficeRecovery классифицирует и иллюстрирует характерные примеры испорченных файлов [электронный ресурс] — URL: https://habr.com/company/officerecovery/blog/149640/ (дата обращения 28.09.2018) Загл. с экрана. Яз. рус.
2. Хабрахабр. Двадцать способов потерять свои данные [электронный ресурс] — URL: https://habr.com/post/150499/ (дата обращения 30.12.2018) Загл. с экрана. Яз. рус.
3. Service812. Диагностика жесткого диска. Типичные неисправности HDD [электронный ресурс] — URL: http://www.service812.ru/news/neispravnosti-zhestkih-diskov-i-ih-diagnostika.php (дата обращения 28.09.2018) Загл. с экрана. Яз. рус.
4. Hetman Software. Причины повреждения файлов и способы их исправления [электронный ресурс] — URL: https://hetmanrecovery.com/ru/recovery_news/causes-of-file-corruption-and-how-to-fix-it.htm (дата обращения 29.09.2018) Загл. с экрана. Яз. рус.
5. Wikipedia. JPEG [электронный ресурс] — URL: https://en.wikipedia.org/wiki/JPEG (дата обращения 8.10.2018) Загл. с экрана. Яз. англ.
6. Hamilton E. JPEG File Interchange Format [электронный ресурс] — URL: http://www.martinreddy.net/gfx/2d/JPEG.txt (дата обращения 8.10.2018) Загл. с экрана. Яз. англ.
7. CodeNet. JPG [электронный ресурс] — URL: http://www.codenet.ru/progr/formt/jpeg_13.php#A14 (дата обращения 8.10.2018) Загл. с экрана. Яз. рус.
8. THE INTERNATIONAL TELEGRAPH AND TELEPHONE CONSULTATIVE COMMITTEE. TERMINAL EQUIPMENT AND PROTOCOLS FOR TELEMATIC SERVICES. [электронный ресурс] —

URL: https://www.w3.org/Graphics/JPEG/itu-t81.pdf (дата обращения 9.10.2018) Загл. с экрана. Яз. англ.

9. Хабрахабр. Декодирование JPEG для чайников [электронный ресурс] — URL: https://habr.com/post/102521 (дата обращения 16.10.2018) Загл. с экрана. Яз. англ.
10. Impulseadventure. JPEG Huffman Coding Tutorial [электронный ресурс] — URL: https://www.impulseadventure.com/photo/jpeg-huffman-coding.html (дата обращения 17.10.2018) Загл. с экрана. Яз. англ.
11. Хабрахабр. Как я разбирал docx с помощью XSLT [электронный ресурс] — URL: https://habr.com/company/intersystems/blog/321044/ (дата обращения 6.12.2019) Загл. с экрана. Яз. рус.
12. Microsoft. Introducing the Office (2007) Open XML File Formats [электронный ресурс] — URL: https://docs.microsoft.com/en-us/previous-versions/office/developer/office-2007/aa338205(v=office.12) (дата обращения 7.12.2019) Загл. с экрана. Яз. англ.
13. Htmlbook. MIME-типы. [электронный ресурс] — URL: http://htmlbook.ru/html/value/mime (дата обращения 7.12.2019) Загл. с экрана. Яз. рус.
14. Хабрахабр. Внутри MP3. А как оно всё устроено? [электронный ресурс] — URL: https://habr.com/post/103635/ (дата обращения 7.12.2019) Загл. с экрана. Яз. рус.
15. CodeProject. MPEG Audio Frame Header [электронный ресурс] — URL: https://www.codeproject.com/Articles/8295/MPEG-Audio-Frame-Header (дата обращения 7.12.2019) Загл. с экрана. Яз. англ.
16. Wikipedia. MP3 [электронный ресурс] — URL: https://en.wikipedia.org/wiki/MP3 (дата обращения 8.12.2019) Загл. с экрана. Яз. англ.

17. W3Techs Web Technology Surveys. Usage of JPEG for websites [электронный ресурс] — URL: https://w3techs.com/technologies/details/im-jpeg/all/all (дата обращения 20.09.2018) Загл. с экрана. Яз. англ.
18. Прохоренок, Н. А. Python 3 Самое необходимое / Н. А. Прохоренок, В. А. Дронов. М.: «БХВ-Петербург», 2016. 461 с.
19. Sencar H. Identification and Recovery of JPEG Files with Missing Fragments [электронный ресурс] — URL: https://www.dfrws.org/sites/default/files/session-files/paper-identification_and_recovery_of_jpeg_files_with_missing_fragments.pdf (дата обращения 1.11.2018) Загл. с экрана. Яз. англ.
20. Uzun E. Carving Orphaned JPEG File Fragments [электронный ресурс] — URL: http://www.prism.gatech.edu/~euzun3/projects/jpgcarving/euzunTIFS2015.pdf (дата обращения 9.12.2019) Загл. с экрана. Яз. англ.