
N. A. Kostilova, A. V. Sorokin
On Elaboration of the Method of Identifying Clusters Belonging to JPEG Files

Keywords: data recovery, file carving, file fragmentation

The results of the experimental attempts of elaboration of the previously proposed method for identifying clusters belonging to JPEG files are described. Type II errors rates for different file formats are estimated. The possible ways of their improvement for some file formats are offered.

Н. А. Костылова, А. В. Сорокин

**ОБ УТОЧНЕНИИ МЕТОДА ОПРЕДЕЛЕНИЯ ПРИНАДЛЕЖНОСТИ
КЛАСТЕРОВ ФАЙЛАМ ФОРМАТА JPEG**

Введение

Основными методами восстановления удаленных файлов с цифровых носителей можно считать восстановление с использованием доступной информации файловой системы, чему посвящена, например, книга Брайана Кэрриэ «Криминалистический анализ файловых систем» [1], и попытки извлечения файлов из последовательности двоичных данных, зафиксированной на носителе. Главным методом такого извлечения является описанный в той же книге метод «вырезания» (carving), основанный на поиске пар известных сигнатур начала и завершения файла в зависимости от его формата. Далее последовательность байт, включающая эти сигнатуры и все данные между ними, представляется в качестве восстановленного файла. При этом нетрудно убедиться, что такой метод корректно работает при условии, что восстановленный файл расположен на носителе последовательно, то есть не фрагментирован. В посвященной различным аспектам восстановления данных книге Г. Сенкевича [2] явно отмечается, что восстановление фрагментированного файла является совершенно отдельной задачей с совершенно иным прогнозом успешного результата по сравнению с последовательно расположенным файлом. В статье [3], авторы которой предлагают вариант развития метода извлечения файлов из последовательности данных носителя, отмечается важность возможности определить, принадлежит ли отдельный кластер файлам того или иного типа. Особенную трудность представляют форматы файлов, не допускающие программное распознавание отдельных элементов и автоматическую проверку корректности. Одним из таких форматов является JPEG. Методу определения принадлежности произвольного кластера файлам такого типа с кодовыми таблицами, описанными в спецификации формата [4], посвящена серия работ А. В. Сорокина. Суть метода подробно изложена в статье [5], где приводятся значения вероятностей ошибок первого и второго рода. Снижение этих вероятностей позволило бы расширить применимость метода и сделать его использование более эффективным.

Распределение ошибок по форматам файлов

В указанной статье [5] приведены значения ошибок первого и второго рода на уровне примерно 0,03 и 0,3 соответственно. При этом не указано, кластерами каких именно типов файлов был заполнен тестовый образец или образцы для исследования ошибки второго рода. Поэтому авторами данной статьи были отдельно составлены экспериментальные подборки файлов различных типов, объемом не менее 100 Мб для каждого типа, на которых были получены экспериментальные оценки значений ошибок обоих родов. Описанный критерий рассматривает две альтернативные гипотезы: исследуемый кластер принадлежит файлу формата JPEG и кластер не принадлежит файлу данного формата. Поэтому ошибкой первого рода следует считать случай, когда кластер, в действительности



принадлежащий файлу формата JPEG, признается не принадлежащим. Ошибкой же второго рода — случай, когда кластер файла иного формата признается кластером формата JPEG.

Для оценки ошибки первого рода изучалась подборка файлов формата JPEG, суммарно занимающая 25 623 кластера. Описанные в [5] сигнатуры ни разу не встретились в 258 кластерах, а более 19 раз они были обнаружены в 450 кластерах. Ошибка первого рода, таким образом, наступила в 708 случаях, что составило 0,027 и соответствует ранее приведенному значению.

Для оценки ошибок второго рода были использованы подборки файлов 27 различных типов. На рис. 1 приведены процентные доли кластеров, в которых описанные сигнатуры не встретились ни разу. Можно видеть, что результаты от 70 % и более (то есть обеспечивающие уровень ошибки второго рода гарантированно не выше 0,3) получены для таких типов файлов, как RTF и TXT (текстовые файлы), HTML (веб-страницы), CPP (файлы исходных текстов языка программирования C++), а также MUI, TLB (системные файлы) и XML (текстовый файл в формате XML — в некотором смысле сходен с HTML).

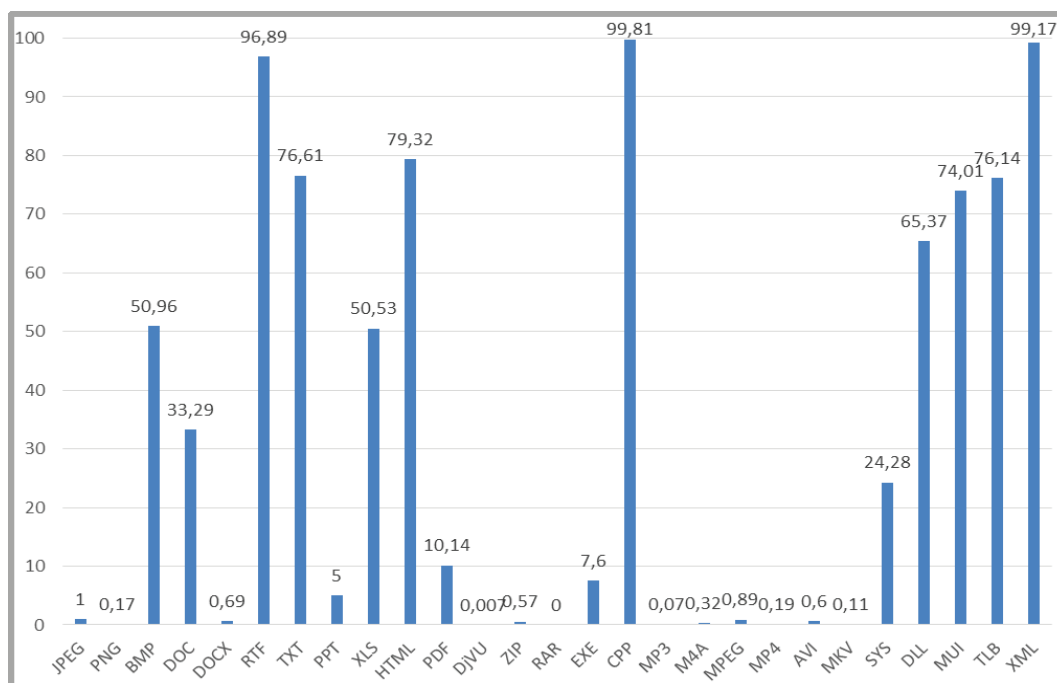


Рис. 1. Процентные доли кластеров файлов различных типов, в которых искомые сигнатуры не найдены

Как можно видеть, первая часть описанного в [5] критерия, то есть отсутствие искомых сигнатур, практически не работает для графических файлов формата PNG, файлов DOCX (текстовые документы версий MS Office после 2003), сжатых и архивированных файлов, а также для мультимедиаформатов. На рис. 2 приведено процентное распределение доли кластеров файлов различных типов, отсеиваемых второй частью критерия — обнаружением искомых сигнатур в кластере более 19 раз.

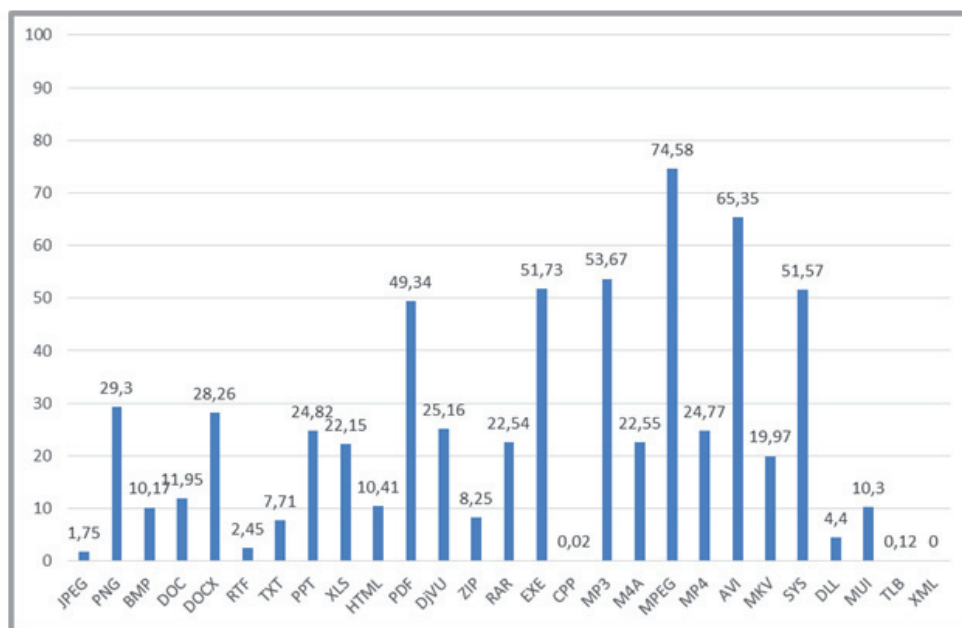


Рис. 2. Процентные доли кластеров файлов различных типов, в которых сигнатуры найдены более 19 раз

Можно видеть, что отсеивается более 70 % кластеров файлов формата MPEG (мультимедиаформат), а также более половины кластеров форматов SYS (системные файлы), MP3, AVI (аудио- и мультимедиафайлы), файлов EXE (исполняемые файлы), чуть менее половины кластеров файлов формата PDF (электронные документы). Для некоторых типов файлов хорошие значения дает сочетание обоих критериев, что и предлагается в [5].

Дальнейшее уточнение метода

Сопоставляя рис. 1 и 2, можно отметить, что метод явно слабо отсеивает кластеры файлов архивов (RAR и ZIP), а также ряда форматов документов и изображений. Поэтому представляется полезным изучить возможность дальнейшего совершенствования данного метода. Одним из очевидных улучшений представляется снижение верхней границы числа обнаруженных вхождений сигнатур в кластере с 19 до того количества, который позволит сохранить вероятность ошибки первого рода на приемлемом уровне. В ходе экспериментов было установлено, что при снижении этой границы до 17 указанная вероятность не превышает 5 %. При этом почти для всех типов файлов удалось достичь повышения доли отсеиваемых кластеров. Значения более 50 % удалось достичь для 16 типов файлов из 26. Одновременно для всех форматов кроме ZIP это значение не уступало 33 %. Такие значения, при вероятности ошибки первого рода в пределах 5 %, позволяют применять данный критерий.

Отдельного внимания заслуживает формат ZIP, поскольку на его примере можно предложить дальнейшие усиления метода. Для файлов формата JPEG число вхождений сигнатур от 0 до 20 распределяется так, как показано на рис. 3.

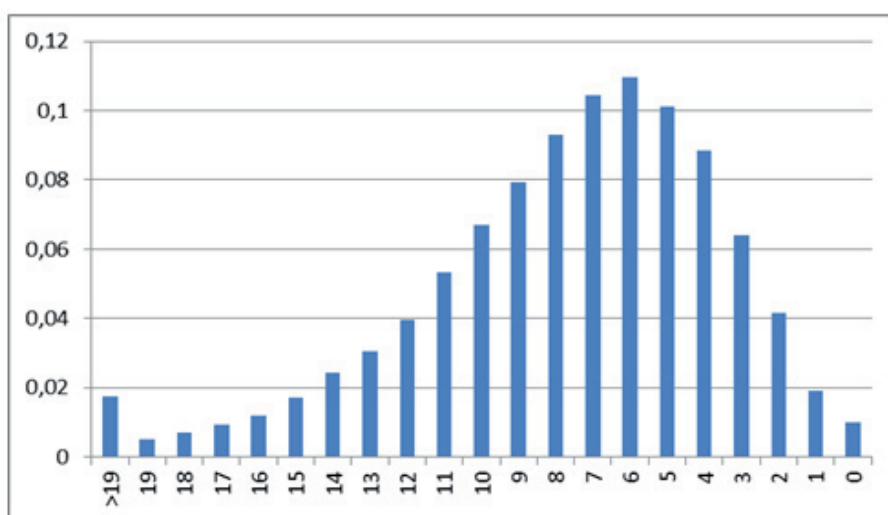


Рис. 3. Долевое распределение числа вхождений сигнатур в кластерах файлов формата JPEG

Здесь можно отметить, что число вхождений сигнатур имеет распределение с явным максимумом на уровне 6 вхождений. При этом случаи 15 и более вхождений имеют довольно невысокую частоту, что и позволило снизить верхнюю границу критерия без значительного увеличения вероятности ошибки первого рода. На рис. 4 и 5 приведены аналогичные распределения для форматов ZIP и RAR.

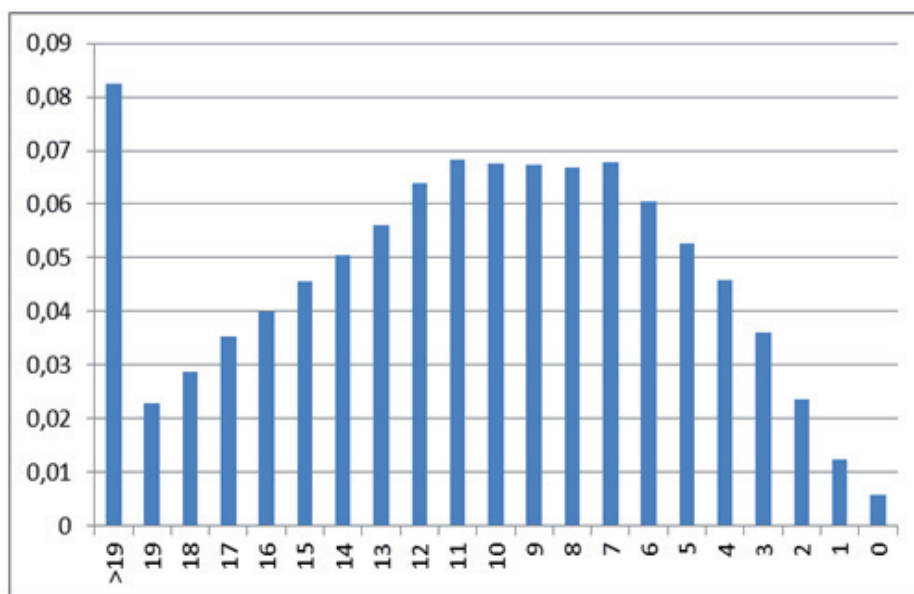


Рис. 4. Долевое распределение числа вхождений сигнатур в кластерах файлов формата ZIP

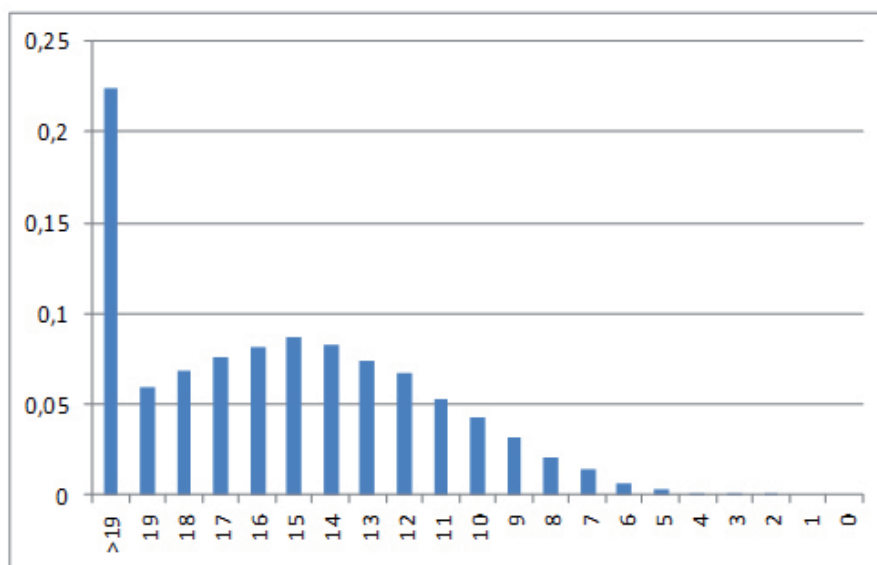


Рис. 5. Долевое распределение числа вхождений сигнатур в кластерах файлов формата RAR

Как можно видеть, максимальные значения для кластеров файлов формата ZIP здесь находятся в диапазоне от 7 до 11, где наблюдается плато. Для формата RAR такого плато не наблюдается, при этом явно видно, что наибольшее число кластеров имеет более 19 вхождений искомым сигнатур. На основании рассмотрения этих двух распределений и их сравнения с аналогичным распределением для формата JPEG можно предложить еще одно направление уточнения описанного метода. Наиболее перспективным представляется дальнейшее снижение верхней границы критерия. Для снижения же нарастающей при этом вероятности ошибки первого рода следует стремиться снизить число вхождений сигнатур в кластеры файлов формата JPEG. Это, в свою очередь, приведет к снижению среднего числа вхождений сигнатур. Для этого следует произвести поиск таких сигнатур, которые вносили бы существенный вклад в распределение для формата JPEG, но при этом практически не встречались бы в кластерах других форматов. Исключение таких сигнатур из поиска, во-первых, не ухудшит ситуацию с теми кластерами, в которых искомые сигнатуры не встретились вообще, а во-вторых, снизит число вхождений сигнатур в кластеры формата JPEG, без снижения такого показателя для кластеров другого типа. Следствием этого может стать возможность снижения верхней границы критерия без существенного возрастания вероятности ошибки первого рода.

Заключение

Определение принадлежности произвольного кластера конкретному формату файлов может существенно ускорить основные существующие методики восстановления данных. Решение этой задачи наиболее затруднительно для тех форматов, которые не позволяют легко получать доступ к содержимому отдельных кластеров на смысловом уровне, а также для тех, для которых является затруднительной проверка корректности восстановленного файла или его части. Поскольку формат JPEG обладает этими свойствами, вызывающими трудности, совершенствование методов идентификации именно этого формата представляется востребованным. Указанные авторами данной статьи пути усиления предложенного ранее метода направлены на еще большее, по сравнению с приведенными в статье [6] оценками, увеличение быстродействия существующих алгоритмов. Вместе с описанным зарубежными авторами в работе [7] методом поиска точки начала фрагментации он может быть использован для разработки эффективного средства восстановления файлов указанного формата в особо сложных случаях.



СПИСОК ЛИТЕРАТУРЫ:

1. Кэрриэ Б. Криминалистический анализ файловых систем. СПб.: Питер, 2007.
2. Сенкевич Г. Е. Искусство восстановления данных. СПб.: БХВ-Петербург, 2011.
3. Pal A., Memon N. The evolution of file carving // IEEE Signal Processing Magazine. Vol. 26. Issue 2. March 2009. P. 59–71.
4. CCITT T.81 (09 /92) Digital Compression and Coding of Continuous-tone Still images. The international telegraph and telephone consultative committee, ITU. 1993.
5. Сорокин А. В. Об определении принадлежности кластеров диска файлам формата JPEG // Проблемы информационной безопасности. Компьютерные системы. 2012. № 4. С. 61–67.
6. Сорокин А. В. О возможном развитии метода извлечения данных для восстановления файлов формата JPEG // Безопасность информационных технологий. 2013. № 4. С. 97–100.
7. Pal A., Sencar T. and Memon N. Detecting File Fragmentation Point Using Sequential Hypothesis Testing // Digital Investigations. 2008. № 5. P. S2–S13.

REFERENCES:

1. Carrier B. File System Forensic Analysis (in Russian). St. Petersburg: Piter, 2007.
2. Senkevich G. E. Iskusstvo vosstanovleniya dannih [The art of data recovery]. St. Petersburg: BHV-Peterburg, 2011.
3. Pal A., Memon N. The evolution of file carving // IEEE Signal Processing Magazine. Vol. 26. Issue 2. March 2009. P. 59–71.
4. CCITT T.81 (09 /92) Digital Compression and Coding of Continuous-tone Still images. The international telegraph and telephone consultative committee, ITU. 1993.
5. Sorokin A. V. On differentiation of clusters belonging to JPEG files from clusters belonging to other format files // Problems of information security. Computer systems. 2012. № 4. P. 61–67.
6. Sorokin A. V. On possible expansion of data carving for JPEG files recovery // Information systems security. 2013. № 4. P. 97–100.
7. Pal A., Sencar T. and Memon N. Detecting File Fragmentation Point Using Sequential Hypothesis Testing // Digital Investigations. 2008. № 5. P. S2–S13.