Análise de ocorrência de falso positivos em recuperação de dados formatados.

Rubens K. P. Silva[1], **Islan A. Bezerra**[1], **Sidney M. L. de Lima**[2], **Sérgio M. M. Fernandes**[1]

[1]Departamento de Engenharia da Computação – Universidade de Pernambuco, (UPE) – Recife – PE – Brazil

[2]Departamento de Eletrônica e Sistemas – Universidade Federal de Pernambuco (UFPE) Recife – PE – Brazil.

{rkps,iab, smurilo}@ecomp.poli.br, sidney.lima@ufpe.br

Abstract. *The data carving technique makes it ineffective to attempt to hide digital evidence by deleting or formatting files. This study evaluates the ability of softwares forensic data carving in recovering various files, especially .doc and .docx, even after formatting. The following softwares were selected from data carving: Foremost, Scalpel, Recuva, PhotoRec, Autopsy and Magic Rescue, all widely used and recognized in the forensic area. The evaluation of these softwares considered the false positive and true positive rate metrics, in addition to the runtime, quantity and size of the recovered files in two distinct scenarios. A dataset was built with 16,000 of files of various extensions for the experiments. The results indicated that Recuva, Autopsy and PhotoRec presented the best performances, with true positive rates exceeding 90% in all scenarios. Regarding false positives, Recuva stood out with a rate of approximately 1%, surpassing the other softwares*

Resumo. *A técnica de data carving torna ineficaz a tentativa de ocultar evidências digitais por meio da exclusão ou formatação de arquivos. Este estudo avalia a capacidade de softwares forenses de data carving em recuperar arquivos diversos arquivo, em especial .doc e .docx, mesmo após formatação. Foram selecionados os seguintes softwares de data carving: Foremost, Scalpel, Recuva, PhotoRec, Autopsy e Magic Rescue, todos amplamente utilizados e reconhecidos na área forense. A avaliação desses softwares considerou as métricas taxa de falsos positivos e verdadeiros positivos, além do tempo de execução, quantidade e tamanho dos arquivos recuperados em dois cenários distintos. Foi construído um dataset com 16.000 exemplares de arquivos de diversas extensões para realização dos experimentos. Os resultados indicaram que o Recuva, Autopsy e PhotoRec apresentaram os melhores desempenhos, com taxas de verdadeiros positivos superiores a 90% em todos os cenários. Em relação aos falsos positivos, o Recuva se destacou com uma taxa de aproximadamente 1%, superando os demais softwares.*

1. Introdução

Ao ser armazenado em dispositivos como *smartphones* ou computadores, um arquivo é gravado de forma permanente em uma memória não volátil, por meio do sistema de armazenamento específico do Sistema Operacional (SO). Apesar da exclusão parecer definitiva, o arquivo permanece no dispositivo, aguardando ser sobrescrito. Essa persistência

de dados permite que arquivos que tenham sido eliminados acidentalmente ou intencionalmente sejam recuperados.

A recuperação de arquivos apagados é uma etapa crucial na perícia digital, fornecendo aos especialistas novas perspectivas e informações valiosas, especialmente em casos de dados formatados. Essa técnica, conhecida como *data carving*, é definida como o processo pelo qual se recuperam dados excluídos ou inacessíveis armazenados em mídias computacionais [Stanković and Khan 2022] e [Blaskovic et al. 2023]. Ela se baseia na recuperação de arquivos a partir de seu conteúdo bruto, independentemente de informações adicionais sobre o sistema de arquivos [Lin 2018]. A técnica *data carving* é um importante tópico na área da Forense Digital.

Diante da relevância dos formatos .doc e .docx em ambientes profissionais e pessoais, este trabalho busca explorar soluções eficazes para a recuperação de arquivos nesses formatos. A escolha desses formatos se justifica pela lacuna existente na literatura científica em relação à recuperação desses tipos de arquivos, em seu trabalho [Hanis et al. 2021], corrobora com essa afirmativa.

O presente estudo tem como objetivo avaliar a qualidade da recuperação de dados realizada por diferentes *softwares*, com foco na identificação de falsos positivos, verdadeiros positivos e a ocorrência de duplicações de arquivos recuperados. A análise desses indicadores permitirá avaliar a precisão e a confiabilidade dos *softwares* testados.

2. Trabalhos Relacionados

Em seu trabalho, [Nurhayati 2017], realiza uma análise entre dois *softwares* de *data carving*: PhotoRec e Foremost. Em sua proposta de método experimental o autor propõe 8 cenários de testes considerando cinco elementos importantes: preparação do dispositivo de armazenamento; criação da imagem do dispositivo (.dd ou .raw); *softwares* de recuperação de arquivos, (PhotoRec e Foremost); os arquivos de saída dos *softwares* de *carving*, ou seja, os arquivos recuperados existentes anteriormente no dispositivo de armazenamento; validação dos arquivos recuperados. Conclui-se que o PhotoRec apresenta melhor desempenho em velocidade de processamento. Mesmo recuperando menos arquivos que seu concorrente, retornou mais verdadeiros positivos. O referido trabalho serviu como norteador metodológico e experimental para a pesquisa atual.

Com maior frequência é encontrado na literatura *softwares carving* voltados para arquivos de imagem. Por exemplo, [Uzun and Sencar. 2020] propõem o JpgScraper. Segundo os autores, o *software* é a solução inovadora para detectar dados JPEG considerando *bitstream*, quantização, largura da imagem e outras características, de modo a recuperar órfãos que são os arquivos fragmentados. No trabalho de [Hilgert et al. 2019] é proposto um *data carving* que explora ao máximo a sintaxe de formato dos arquivos PNG. A abordagem alcança resultados de recuperação de 98% de arquivos PNG considerando, inclusive, cenários complexos de arquivos fragmentados.

Os autores, [Pereira et al. 2019], em sua revisão da literatura, apontaram um panorama muito particular da ausência de trabalhos que tratem de falsos positivo em *softwares* para *Data Carving*. Evidenciando que, entre os 107 trabalhos analisados, apenas um único trabalho zelou por esse aspecto. Evidenciando uma carência de cuidados experimentais na concepção e implementação de *softwares* tendo como meta o tratamento deste

tipo de problema. Por tanto, ciente do *gap* existente na literatura apontado pelos autores e suas propostas de trabalhos futuros, que sugere uma melhor compreensão das ferramentas mais empregadas no processo de recuperação de dados, foi proposta a presente pesquisa.

3. Procedimento Metodológico

Para manter o rigor da cadeia de custódia, o procedimento metodológico utilizado foi inspirado no trabalho de [Nurhayati 2017]. A Figura 1 ilustra as etapas metodológicas realizadas pelos autores para concretização da pesquisa. Iniciando com a limpeza da mídia (*pendrive*) que foi utilizada até a análise dos arquivos que foram recuperados pelas aplicações *benchmark*. As etapas descrita na Figura 1 são:

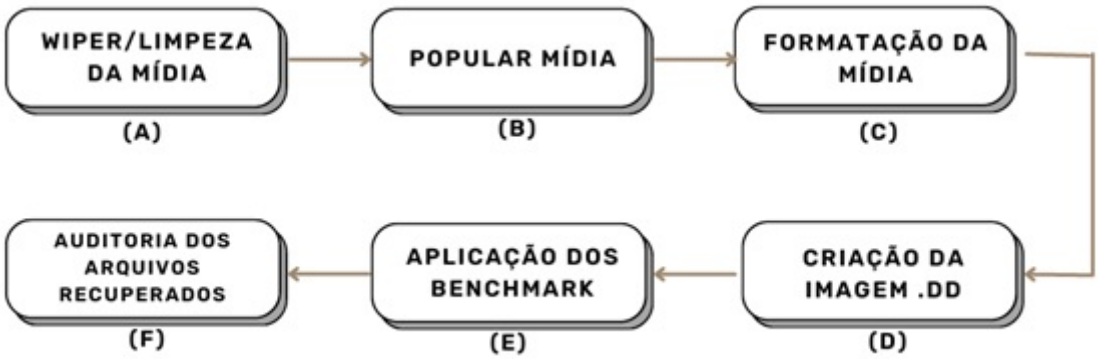

Figura 1. Metodologia experimental. Fonte: Autoria própria.

- Figura 1 (a): processo de *wiper* (limpeza) com o *software* de *antiforense* Shred[1]. A ação consiste em consecutivas varreduras na mídia, sobrescrevendo o sistema de arquivos, apagando permanentemente qualquer dado que estiver armazenado.
- Figura 1 (b): consiste em transferir os arquivos do *dataset* para a mídia, a fim de replicar uma mídia que será investigada.
- Figura 1 (c): formatação da mídia através do *software* de formatação rápida do SO Windows. Temos o intuito de replicar o comportamento intencional de ocultar informações.
- Figura 1 (d): criação de imagem *forense bit-a-bit* através do *softwares* dd (*Disk Dump*), próprio de distribuições Linux. A perícia é realizada na imagem *forense*, enquanto a mídia original é preservada.
- Figura 1 (e): os *softwares* utilizados são referências de soluções comerciais tradicionalmente utilizados em *forense* digital para recuperar dados formatados. Os *softwares* utilizados na pesquisa são: (i) Recurva, (ii) Foremost, (iii) Scalpel, (iv) MagicRescue, (v) PhotoRec e (vi) Autopsy.
- Figura 1 (f): avaliação das métricas de interesse da análise inclui o número de falsos positivos gerados, verdadeiros positivos e o tempo de execução dos *softwares* durante o processo.

3.1. *Dataset* Utilizado

Para atingir o objetivo da pesquisa, previamente, foi necessário a construção de um *dataset* que serviu como massa de dados para realização dos experimentos. O referido *dataset*

[1]Documentação: www.gnu.org/software/coreutils/manual/html_node/shred-invocation.html

consiste em um diretório contendo 16.000 arquivos de diferentes extensões. As extensões coletadas são variadas e incluem textos, planilhas, imagens, slides, entre outros, além de serem provenientes de *softwares* proprietários e *open source*.

Todos os arquivos presentes no *dataset* foram obtidos da *internet* ou criados pelos próprios autores, além disso, o conjunto contém apenas arquivos exclusivos.O tamanho total do *dataset* é de 5,1 GB. O *dataset*, cópia *forense* das partições utilizadas e resultados obtidos, estão disponíveis para consulta no *link*[2].

4. Estudo de Caso e Resultados

Com o propósito de analisar a versatilidade das ferramentas de *data carving*, o estudo de caso foi estruturado em dois cenários experimentais, cada um com características específicas e desafios distintos para a recuperação de dados.

4.1. Estudo de Caso: Cenário 1

O primeiro cenário, aqui denominado como Cenário 1, foi projetado como ponto de partida para a pesquisa, considerando apenas um número limitado de arquivos com extensão .doc e .docx a serem submetidos às ferramentas de recuperação. Inicialmente, a estratégia de pesquisa foi adotada para fornecer *insights* sobre o desempenho das ferramentas em um contexto mais simplificado. Para o Cenário 1, foi utilizado um conjunto de 1.000 arquivos do tipo .doc e.docx, totalizavam 66 MB, que foram gravados na mídia e posteriormente formatados. A imagem resultante foi apresentada a cada uma das ferramentas de recuperação e os resultados dos cenários são apresentados na Tabela 1. Esses resultados ajudarão os pesquisadores a determinar o desempenho das ferramentas em um ambiente controlado.

Tabela 1. Comparação de *softwares carving* recuperando .doc/.docx

Ferramenta	Qtd. de arquivos gerados	Tempo de Execução	Tamanho do diretório recuperado	Falso positivos (%)	Verdadeiro positivos (%)	Verdadeiro positivos repetidos (%)
Recuva	1000	8min 22s	64 MB	$< 1\%$	$> 99\%$	0%
Photorec	768	20min 53s	47,7 MB	$< 1\%$	$> 99\%$	0%
Autopsy	1743	40min 51s	112 MB	$< 1\%$	$> 99\%$	42,4%
Foremost	1563	4min 16s	191,2 MB	38,7%	61,3%	0%
Scalpel	25077	1h 37min 24s	14,5 GB	100%	0%	0%
MagicRescue	162	8min 10s	56,8 MB	94%	6%	0%

O Recuva obteve uma alta precisão de recuperação de arquivos, recuperando o número exato de arquivos e o mesmo tamanho dos originais, resultando em uma acurácia da ferramenta de 99%. O PhotoRec recuperou 768 dos 1000 arquivos (76,8%), com uma taxa de verdadeiros positivos de 99%. Por sua vez, o Autopsy recuperou 743 arquivos a mais do que o número de arquivos originalmente armazenados. Além disso, 42,4% dos arquivos recuperados eram duplicados, o que significa que a ferramenta gerou cópias de arquivos existentes.

[2]https://drive.google.com/drive/folders/16rmys5T8TBXEQKVjKZ7OZZivXGVRJZBa

Em comparação com os outros *softwares* testados, o Scalpel, Foremost e MagicRescue obtiveram os piores resultados. Isso ocorreu tanto ao recuperar um número muito maior ou muito menor de arquivos que o esperado, apresetando, também, uma alta taxa de falsos positivos. O Scalpel obteve o pior desempenho, com um volume de dados recuperados 25 vezes maior que o volume original, com todos os arquivos recuperados identificados como falsos positivos.

Em relação ao tempo de execução, os *softwares* Foremost, MagicRescue e Recuva foram as mais rápidas a recuperar os arquivos formatados. De todos, o Recuva se mostrou rápido e o melhor resultado em quantidade e qualidade de arquivos recuperados. O Autopsy, por sua vez, obteve um tempo de recuperação maior, pois, por padrão, o *software* tenta recuperar os metadados atrelados aos arquivos. Caso seja retirada essa etapa, o tempo de recuperação seria de aproximadamente 13 minutos e 45 segundos.

4.2. Estudo de Caso: Cenário 2

O Cenário 2, projetado para simular um ambiente de recuperação de dados mais realista, apresentando conjunto de dados significativamente mais diverso. Nesse caso, um número muito maior de dados, 16.000 exemplares distribuídos em 16 extensões. Isso ajudou a avaliar o quão bem as ferramentas respondem a níveis de trabalho significativamente maiores. Os resultados do cenário são apresentados na Tabela 2 e discutidos a seguir.

Tabela 2. Comparação de *softwares carving* recuperando múltiplas extensões

Ferramenta	Qtd. de arquivos gerados	Tempo de Execução	Tamanho do diretório recuperado	Falso positivos (%)	Verdadeiro positivos (%)	Verdadeiro positivos repetidos (%)
Recuva	13952	13h 14min 54s	117 GB	1,1%	98,8%	0,05%
Photorec	14830	06min 50s	5 GB	8,2%	>91,7%	0%
Autopsy	32149	14h 20min 3s	193 GB	10%	90%	44,5%
Foremost	20150	4min 6s	2,2 GB	65,4%	34,5%	0%
Scalpel	508482	13h 54min 15s	100,9 GB	> 99%	< 1%	0%
MagicRescue	15950	2h 4min 47s	2,2 GB	80,1 %	19,9%	0%

Com base nessa métrica de verdadeiros positivos, a análise apresenta que o Recuva, PhotoRec e Autopsy obtiveram as melhores taxas de verdadeiros positivos. O PhotoRec superou todas as outras ferramentas, obtendo o melhor resultado em todas as outras métricas também. Isso inclui a velocidade de execução, o número de arquivos recuperados mais próximo de 16.000 e o tamanho total dos arquivos recuperados. Embora o PhotoRec não tenha recuperado cerca de 1.170 arquivos originais, seu desempenho superou o Recuva, que deixou de recuperar 2.048 arquivos.

Uma observação intrigante sobre os resultados do Autopsy é o número de arquivos recuperados, que são mais 32.000, dos quais quase 29.000 arquivos são verdadeiros positivos. No entanto, o *dataset* original continha em apenas 16.000 arquivos. Portanto, pode-se concluir que o *software* Autopsy gerou múltiplas cópias de cada arquivo durante o processo de recuperação (44,5% de verdadeiros positivos repetidos), semelhante ao observado no Cenário 1. Embora a quantidade de verdadeiros positivos dos resultados sejam

altos, a duplicação dos arquivos leva ao aumento substancial do tamanho total dos dados recuperados, o que, neste caso, gera efeito negativo semelhante aos falsos positivos.

Os resultados obtidos dos *softwares* Foremost, MagicRescue e Scalpel foram insatisfatórios, com destaque para taxa de falsos positivo maior que 65%. Embora o Foremost e MagicRescue tenham apresentado uma quantidade razoável de arquivos recuperados, além de tempo relativamente satisfatório, a baixa precisão dos resultados os tornam inviáveis. Já o Scalpel apresentou-se como ineficiente para a recuperação de arquivos. Com uma quantidade elevada de arquivos falsos positivos, a ponto do experimento ter sido interrompido após atingir apenas 11,6% do processo e ter gerado mais de 100GB de falsos positivos.

5. Conclusão e Trabalhos Futuros

A análise comparativa de diversos *softwares* de *data carving*, amplamente utilizados na literatura, visa fornecer ao perito uma base sólida para a escolha da ferramenta mais adequada a cada caso. Os resultados obtidos neste estudo permitem ao profissional avaliar as vantagens e desvantagens de cada *software*, considerando métricas como taxa de falsos positivos, tempo de execução e quantidade de arquivos recuperados. Dessa forma, o perito poderá tomar decisões mais acertadas e selecionar o *software* que melhor se adapta às características específicas de cada investigação.

A taxa de falsos positivos foi o principal critério de avaliação neste estudo. A identificação incorreta de arquivos pode levar a conclusões equivocadas e comprometer a integridade da investigação. Os resultados apresentados demonstram que a taxa de falsos positivos varia significativamente entre os diferentes *softwares*, evidenciando a importância de selecionar a ferramenta mais adequada para cada caso.

Ao comparar os *softwares*, observou-se que o PhotoRec e o Recuva apresentaram resultados mais consistentes em relação à quantidade de arquivos recuperados. A capacidade de recuperar a quantidade correta de arquivos é um fator determinante na escolha de um *software* de *data carving*, pois evita a geração de falsos negativos, garantindo a qualidade da análise *forense*.

A eficiência na coleta de evidências digitais não deve comprometer a qualidade da análise. A utilização de ferramentas de *data carving* adequadas permite conciliar a necessidade de celeridade com a garantia da integridade dos dados. Ao otimizar o processo de extração e recuperação de informações, o perito pode dedicar mais tempo à análise e interpretação das evidências, contribuindo para a elucidação dos fatos.

Como proposta de trabalhos futuros, os autores propõem: (i) Ampliar escopo: ampliando a gama de *softwares* de *data carving* e abrangendo outras métricas de análise. (ii) Enriquecer o *dataset*: incorporar outras extensões de arquivos, a fim de tornar o resultado mais representativo. (iii) Propor cenários individualizados: incluir cenários de teste especializados na recuperação de uma única extensão. (iv) Explorar técnicas de Inteligência Artificial: Explorar o potencial de técnicas de aprendizado de máquina [Lima et al. 2022], como redes neurais profundas [Ali and Mohamad 2021], para aprimorar a precisão e a eficiência dos processos de *data carving*.

Referências

Ali, R. R. and Mohamad, K. M. (2021). Rx_mykarve carving framework for reassembling complex fragmentations of jpeg images. *Journal of King Saud University - Computer and Information Sciences*, 33(1):21–32.

Blaskovic, A. K., Rusk, J. D., Parker, V. C. J., and Payne, B. R. (2023). Cybercrime and intellectual property theft: An analysis of modern digital forensics. *Proceedings of the Future Technologies Conference (FTC), Springer International Publishing*, 2.

Hanis, F. M., Khoshvaghti, H., Teimouri, M., and Veisi, H. (2021). A language-independent approach to classification of textual file fragments: Case study of persian, english, and chinese languages. In *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 254–259.

Hilgert, J., Lambertz, M., and M. Rybalka, R. S. (2019). Syntactical carving of pngs and automated generation of reproducible datasets. *Digital Investigation*, 29.

Lima, S. M. L., S. H. M. T. Silva, R. P. Pinheiro, D. M. S., Lopes, P. G., Lima, R. D. T., Monteiro, J. R. O. T. A., Fernandes, S. M. M., Albuquerque, E. Q., and Santos., W. W. A. S. W. P. D. (2022). Next-generation antivirus endowed with web-server sandbox applied to audit fileless attack. *Springer Nature*.

Lin, X. (2018). Introductory computer forensics - a hands-on practical approach. *Springer Nature*.

Nurhayati, N. F. (2017). The analysis of file carving process using photorec and foremost. *International Conference on Computer Applications and Information Processing Technology*.

Pereira, E., Silva, W., Bezerra, S., and Araújo, J. (2019). Análise de métodos para o tratamento de arquivos falso-positivos a partir de ferramentas de recuperação de dados digitais: Uma revisão sistemática da literatura. pages 1–10.

Stanković, M. and Khan, T. M. (2022). Digital forensics tool evaluation on deleted files. digital forensics and cyber crime. *ICDF2C Springer*, 508.

Uzun, E. and Sencar., H. T. (2020). Jpg scraper : An advanced carver for jpeg files. *IEEE Transactions on Information Forensics and Security*, 15.