ÁKIO NOGUEIRA BARBOSA

**MÉTODO PARA RANQUEAMENTO E TRIAGEM DE
COMPUTADORES APLICADO À PERÍCIA DE
INFORMÁTICA**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências

São Paulo
2015

ÁKIO NOGUEIRA BARBOSA

**MÉTODO PARA RANQUEAMENTO E TRIAGEM DE
COMPUTADORES APLICADO À PERÍCIA DE
INFORMÁTICA**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências

São Paulo

2015

ÁKIO NOGUEIRA BARBOSA

**MÉTODO PARA RANQUEAMENTO E TRIAGEM DE
COMPUTADORES APLICADO À PERÍCIA DE
INFORMÁTICA**

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
título de Doutor em Ciências

Área de concentração:

Engenharia de Computação e Sistemas Digitais

Orientador:

Prof. Dr. Wilson Vicente Ruggiero

São Paulo

2015

DEDICATÓRIA

Este trabalho é dedicado aos meus pais,
minhas irmãs e a todos os professores que
muito me ensinaram ao longo de minha vida.

AGRADECIMENTOS

Primeiramente ao Grande Criador do Universo, por ter me concedido a graça e forças necessárias para alcançar mais esta importante conquista em minha vida.

Ao meu orientador Professor Wilson Vicente Ruggiero pelos preciosos ensinamentos, discussões e reflexões realizadas no decorrer deste trabalho.

Um agradecimento especial aos colegas Cleber Morio Okida, Romeo Bulla Junior e Daniel Francis Soriano pelo auxílio com as implementações das ferramentas, e as acaloradas discussões e debates, que nem sempre as opiniões convergiam, mas que foram de extrema relevância para a conclusão deste trabalho.

Ao colega Marco Antônio Torrez Rojas, que mesmo durante as correrias do dia-a-dia, sempre esteve disposto a dedicar um tempo para discussões e sugestões.

Aos Professores Dr. Edson Satoshi Gomi e Dr. Volnys Borges Bernal pelas sugestões e contribuições feitas durante o exame de qualificação deste trabalho.

Aos Professores Dr. Edson Midorikawa, Dr. Jorge Kinoshita e Dr. Stephan Kovach pelas sugestões e comentários sempre objetivos.

Aos professores do PCS que sempre me incentivaram, principalmente as professoras Dra. Regina Melo Silveira, Dra. Cintia Borges Margi e professor Dr. José Sidnei Colombo Martini.

A Superintendência da Polícia Técnico-Científica do Estado de São Paulo - SPTC, nas pessoas da Dra. Norma Sueli Bonaccorso, e os peritos criminais Antônio Vitorio Cecere, Edson do Amaral e Rafael Souza Ferraz.

As bibliotecárias da EPUSP, que sempre atenderam prontamente minhas solicitações de auxílio.

A todos que de alguma forma, me ajudaram para que este trabalho fosse concretizado, meus amigos, funcionários e colegas do PCS, da Escola Politécnica e da Universidade de São Paulo.

A minha querida noiva Núbia Cristina, meus amados sobrinhos Matheus, Mariah e Davi, pelo carinho, apoio, incentivo e compreensão pelos diversos momentos que estive ausente para realização deste trabalho.

*“O aprender pode ser considerado uma das
mais belas virtudes da espécie humana, porém,
o ensinar é a mais sublime de todas elas...”*

(autor desconhecido)

RESUMO

Considerando-se que uma das tarefas mais comuns para um perito judicial que atua na área da Informática é procurar vestígios de interesse no conteúdo de dispositivos de armazenamento de dados (DADs), que esses vestígios na maioria das vezes consistem em palavras-chave (PChs) e durante o tempo necessário para realização da duplicação do DAD o perito fica praticamente impossibilitado de interagir com os dados contidos no mesmo, decidiu-se verificar a hipótese de que seja possível na etapa de coleta, realizar simultaneamente à duplicação do DAD a varredura para procurar PChs em dados brutos (*raw data*), sem com isso impactar significativamente o tempo de duplicação. O principal objetivo desta tese é propor um método que possibilite identificar os DADs com maior chance de conter vestígios de interesse para uma determinada perícia ao término da etapa de coleta, baseado na quantidade de ocorrências de PChs encontradas por um mecanismo de varredura que atua no nível de dados brutos. A partir desses resultados é realizada uma triagem dos DADs. Com os resultados da triagem é realizado um processo de ranqueamento, indicando quais DADs deverão ser examinados prioritariamente na etapa de análise. Os resultados dos experimentos mostraram que é possível e viável a aplicação do método sem onerar o tempo de duplicação e com um bom nível de precisão. Em muitos de casos, a aplicação do método contribui para a diminuição da quantidade de DADs que devem ser analisados, auxiliando a diminuir o esforço humano necessário.

Palavras-chave: Perícia de Informática. Perícia de Computadores. Triagem de Computadores. Digital Forensics. Investigação Digital.

ABSTRACT

Considering that one of the most common tasks for a legal expert acting in the Information Technology area is to look for evidences of interest in the content data storage devices (DADs). In most cases these evidences consist of keywords. During the time necessary to perform the DAD duplication, the expert is practically unable to interact with the data contained on DAD. In this work we have decided to verify the following hypothesis: It is possible, at the collection stage, to simultaneously hold the duplication of the DAD and scan to search for keywords in raw data, without thereby significantly impact the duplication time. The main objective of this thesis is to propose a method that allows to identify DADs with a strong chance of containing evidences of interest for a particular skill at the end of the collection stage, based on the keywords occurrences found by a scanner mechanism that operates at the raw data level. Based on these results, a triage of DADs is established. With the results of the triage, a ranking process is made, providing an indication of which DADs should be examined first at the analysis stage. The results of the ours experiments showed that it is possible and feasible to apply the method without hindering the duplication time and with a certain level of accuracy. In most cases, the application of the method contributes to reduce the number of DADs that must be analyzed, helping to reduces the human effort required.

Keywords: Computer forensics, Computer triage, Digital forensics, Digital investigation, Forensics triage.

LISTA DE ILUSTRAÇÕES

Figura 1.1 - Estimativa para o aumento da capacidade dos discos	29
Figura 1.2 - Etapas para o procedimento de Investigação Digital Forense	32
Figura 3.1 – Esquema de representação do método proposto	75
Figura 3.2 – Sequência em que os processos e resultados ocorrem no tempo	77
Figura 3.3 - Inserção do método proposto no modelo de referência	79
Figura 3.4 – Fluxo de processamento dos dados em memória	82
Figura 3.5 – Processo de duplicação	83
Figura 3.6 – Processo de varredura	85
Figura 4.1 - Procedimento das operações do DD distribuído em função do tempo	105
Figura 4.2 - Fluxo completo da Rotina principal da Ferramenta CopyScan	108
Figura 4.3 – Leitura dos dados de entrada	111
Figura 4.4 – Armazenamento dos dados na fila de varredura	111
Figura 4.5 – Armazenamento dos dados na fila de duplicação	112
Figura 4.6 – Diagrama de classes da ferramenta <i>CopyScan</i>	113
Figura 4.7 – Diagrama de objetos para combinação da palavra larc	114
Figura 4.8 - Sub-rotina do módulo duplicador	116
Figura 4.9 - Sub-rotina do módulo de varredura	117
Figura 4.10 – Exemplo da máquina de estados da ferramenta <i>CopyScan</i>	119
Figura 4.11 – Exemplo de fila de varredura da ferramenta <i>CopyScan</i>	120
Figura 4.12 – Exemplo do passo do mecanismo de varredura da <i>CopyScan</i>	121
Figura 4.13 – Ilustração dos processos cobertos pelas ferramentas CS e FIP	127
Figura 4.14 – Clusters alocados e clusters não alocados	128
Figura 4.15 – Alusão às quatro partes do DAD	129

Figura 4.16 – Comparação entre resultados	131
Figura 4.17 – Primeira varredura	132
Figura 4.18 – Segunda varredura	134
Figura 4.19 – Terceira varredura.....	136
Figura 4.20 – Quarta varredura	137
Figura 5.1 – Exemplo de sobrescrita de DAD a partir da ferramenta DD.	149
Figura 5.2 - Tempo para duplicação de DADs.....	152
Figura 5.3 – Comparação do tempo para duplicação e varredura	168
Figura 5.4 – Sequências de Varreduras	171

LISTAS DE TABELAS

Tabela 2.1 - Síntese das técnicas e características dos trabalhos relacionados	69
Tabela 3.1 – Relação de caracteres imprimíveis	89
Tabela 3.2 – Exemplo de tabela agregada com ocorrências de PChs (VP) de todos os DADs.....	90
Tabela 3.3 – Exemplo de tabela agregada com os valores de VP	96
Tabela 3.4 – Resultado do ranqueamento das PChs sem prioridade.....	97
Tabela 3.5 – Resultado do ranqueamento de uma PCh com prioridade.....	98
Tabela 3.6 – Resultado do ranqueamento do grupo de PChs com prioridade.....	99
Tabela 3.7 – Resultado do ranqueamento dos grupos de PChs com prioridade.....	100
Tabela 3.8 – Comparação do resultado do ranqueamento dos grupos de PChs	101
Tabela 4.1 - Tipos de arquivos que não podem ser "interpretados" pela ferramenta <i>CopyScan</i>	124
Tabela 4.2 - Tipos de arquivos que podem ser "interpretados" pela ferramenta <i>CopyScan</i>	125
Tabela 5.1 – Informações sobre os arquivos de imagem utilizados	146
Tabela 5.2 – Relação de dias em que os DADs dos funcionários foram duplicados.	147
Tabela 5.3 – Arquivos de imagens descompactados para o formato .img	154
Tabela 5.4 – Tempo de duplicação utilizando as ferramentas DD e CS.	154
Tabela 5.5 – Lista de palavras-chave	159
Tabela 5.6 – Ocorrências de Pchs e tempo de duplicação sem combinação	160
Tabela 5.7 – Ocorrências de PChs e tempo de duplicação com combinação.....	161
Tabela 5.8 – Tempo de duplicação para a ferramenta <i>CopyScan</i>	162
Tabela 5.9 – Resumo da quantidade de ocorrências encontradas	162
Tabela 5.10 – Ocorrências de PChs (VP e FP) sem combinações	163
Tabela 5.11 – Ocorrências de PChs (VP e FP) com combinações.....	164
Tabela 5.12 – Ocorrências encontradas nos DADs pela ferramenta Autopsy.....	167

Tabela 5.13 – Exemplo da comparação do tempo entre os métodos.....	168
Tabela 5.14 – Grupos de imagens formados a partir do sorteio	170
Tabela 5.15 – Lista de palavras-chave com quatro caracteres ou mais.....	171
Tabela 5.16 – Ocorrências de PChs em cada parte do DAD #7	175
Tabela 5.17 – Ocorrências de PChs para todos DADs do subconjunto	176
Tabela 5.18 – Proporção entre os arquivos interpretáveis e o DAD todo	177
Tabela 5.19 – Quantidade de ocorrências de PChs no DAD todo (QO _{DT}).....	179
Tabela 5.20 – Tabela agregada com quantidade de ocorrências de VP _{DT}	181
Tabela 5.21 – Tabela agregada com a quantidade de ocorrências de VP _{AI}	181
Tabela 5.22 – Tabela agregada com a quantidade de ocorrências de PCh no <i>Autopsy</i>	182
Tabela 5.23 – Comparação entre o ranqueamento de QOVP _{AI} e QOVP _{DT} (sem prioridade).....	183
Tabela 5.24 – Comparação entre os ranqueamentos de QOVP _{AI} , QOVP _{DT} e <i>Autopsy</i> sem prioridade	184
Tabela 5.25 – Tabela agregada com quantidade de ocorrências de VP _{DT} com prioridade	186
Tabela 5.26 – Tabela agregada com quantidade de ocorrências de VP _{AI} com prioridade.....	186
Tabela 5.27 – Comparação entre o ranqueamento de QOVP _{AI} e QOVP _{DT} com prioridade.....	187
Tabela 5.28 – Tabela agregada com quantidade ocorrências Autopsy com prioridade	188
Tabela 5.29 – Comparação entre QOVP _{AI} , QOVP _{DT} e Autopsy com prioridade.....	189
Tabela 5.30 – Total de ocorrências disco todo (QOVP _{DT})	191
Tabela 5.31 – Total de ocorrências obtidas pela ferramenta Autopsy	192
Tabela 5.32 – Comparação entre QOVP _{DT} e Autopsy com prioridade	193

LISTA DE ABREVIATURAS E SIGLAS

AA	- Arquivos nos clusters alocados
AAFS	- American Academy of Forensic Sciences
AB	- Arquivo Binário
ACPO	- Associação de Oficiais Chefes de Polícia do Reino Unido
AI	- Arquivos Interpretáveis
AM	- Arquivos Mistos
Art.	- Artigo
ASCII	- American Standard Code for Information Interchange
CA	- Cluster Alocado
CD	- Compact Disc
CEO	- Chief Executive Officer
CF	- Computação Forense
CFFTPM	- Cyber Forensic Field Triage Process Model
CPC	- Código de Processo Civil
CPF	- Cadastro de Pessoa Física
CPP	- Código de Processo Penal
CS	- CopyScan
DAD	- Dispositivo de armazenamento de dados
DD	- Disc Duplicator
DF	- Digital Forense
DVD	- Digital Video Disc
DFRWS	- Digital Forensics Research Conference
EB	- Exabyte
EF	- <i>EraseFiles</i>

EL	- Espaço Livre
FIFO	- First in-first out
FIP	- FiltroP
FN	- Falso Negativo
FP	- Falso Positivo
FP _{DT}	- Falso Positivo para o disco todo
FTK	- Forensic Toolkit
GB	- Gigabyte
GPS	- Global Positioning System
HDD	- Hard Disk Drive
ID	- Investigação Digital
IDS	- Intrusion Detection System
IDF	- Investigação Digital Forense
IEEE	- Institute of Electrical and Electronics Engineers
IP	- Internet Protocol
ISO	- International Organization for Standardization
ISP	- Internet Service Provider
KB	- Kilobyte
NPS	- Naval Postgraduate School
PCh	- Palavra-Chave
RAM	- Memória de acesso randômico
RPM	- Rotações Por Minuto
SBC	- Sociedade Brasileira de Computação
SD	- Secure Digital
SIM	- Subscriber Identity Module
SSD	- Solid State Disc
STF	- Supremo Tribunal Federal
STJ	- Superior Tribunal de Justiça

STM	- Superior Tribunal Militar
SWGDE	- <i>Scientific Working Group on Digital Evidence</i>
TB	- Terabyte
TCE	- Tribunais de Contas dos Estados
TJ-SP	- Tribunal de Justiça do Estado de São Paulo
TRF	- Tribunal Regional Federal
TSE	- Tribunal Superior Eleitoral
TST	- Tribunal Superior do Trabalho
URL	- Uniform Resource Locator
UTF	- Unicode Transformation Format
VN	- Verdadeiro Negativo
VP	- Verdadeiro Positivo
XML	- eXtensible Markup Language

LISTA DE SÍMBOLOS

δ	- Preditividade Positiva do processo de filtragem
DAD_D	- Dispositivo de armazenamento de dados de destino
DAD_E	- Dispositivo de armazenamento de dados externo
DAD_O	- Dispositivo de armazenamento de dados de origem
k	- Índice da capacidade do dispositivo
L_H	- Lista de valores de <i>hash</i> dos dispositivos de armazenamento de dados
L_{QO}	- Listagem com a quantidade de ocorrências de palavras-chave
L_{PCh}	- Lista de palavras-chave
L_{PCh_e}	- Lista de palavras-chave expandida
m	- Quantidade de diferentes capacidades dos DADs a serem duplicados
QO	- Quantidade total de ocorrências de palavras-chaves
$QDAD_{1..m}$	- quantidade de DADs com diferentes capacidades
QO_{AA}	- Quantidade de ocorrências de PChs nos arquivos Alocados
QO_{AB}	- Quantidade de ocorrências de PChs nos arquivos Binários
QO_{AI}	- Quantidade de ocorrências de PChs nos arquivos Interpretáveis
QO_{AM}	- Quantidade de ocorrências de PChs nos arquivos Mistos
QO_{DT}	- Quantidade de ocorrências de PChs no DAD todo
QO_{EL}	- Quantidade de ocorrências de PChs no espaço Livre

QOFP _{AB}	- Quantidade de ocorrências de falso positivo nos arquivos binários
QOFP _{AI}	- Quantidade de ocorrências de falso positivo nos arquivos interpretáveis
QOFP _{AM}	- Quantidade de ocorrências de falso positivo nos arquivos mistos
QOFP _{DT}	- Quantidade de ocorrências de falso positivo no disco todo
QOFP _{EL}	- Quantidade de ocorrências de falso positivo no espaço livre
QOVP _{AB}	- Quantidade de ocorrências de verdadeiro positivo nos arquivos binários
QOVP _{AI}	- Quantidade de ocorrências de verdadeiro positivo nos arquivos interpretáveis
QOVP _{AM}	- Quantidade de ocorrências de verdadeiro positivo nos arquivos mistos
QOVP _{DT}	- Quantidade de ocorrências de verdadeiro positivo no disco todo
QOVP _{EL}	- Quantidade de ocorrências de verdadeiro positivo no espaço livre
t	- Tempo para obtenção de resultados
T	- Tempo total
T _{1..m}	- Tempo necessário para duplicar o DAD de capacidade k
T _E	- Tempo estimado para duplicar todos os DADs
VP _{AI}	- Verdadeiro Positivo dos Arquivos Interpretáveis
VP _{DT}	- Verdadeiro Positivo para o disco todo

SUMÁRIO

LISTA DE ILUSTRAÇÕES.....	XI
LISTAS DE TABELAS.....	XIII
LISTA DE ABREVIATURAS E SIGLAS	XV
LISTA DE SÍMBOLOS	XIX
1 INTRODUÇÃO	27
1.1 Motivação.....	34
1.2 Problema de Pesquisa.....	36
1.3 Objetivo.....	37
1.4 Contribuições	39
1.4.1 Produção Científica	40
1.5 Sequência para desenvolvimento do trabalho	40
1.6 Organização do Trabalho	42
2 CONCEITOS FUNDAMENTAIS E TRABALHOS CORRELATOS.....	43
2.1 Terminologias.....	44
2.1.1 Forense e Ciência Forense.....	44
2.1.2 Forense Computacional, Computação Forense, Digital Forense e Investigação Digital	45
2.1.3 Considerações Sobre as Terminologias Apresentadas	48
2.1.4 Vestígio e Indício	49
2.1.5 Prova Pericial	49
2.1.5.1 Perícia Judicial	50
2.1.6 Quesitos.....	51
2.2 Processo de Investigação Digital Forense	52
2.2.1 Etapas da Investigação Digital Forense	53

2.2.1.1	Etapa de Preparação	53
2.2.1.2	Etapa de Preservação	54
2.2.1.3	Etapa de Coleta.....	55
2.2.1.4	Etapa de Análise.....	56
2.2.1.5	Etapa de Correlação	57
2.2.1.6	Etapa de Apresentação dos Resultados	57
2.3	Investigação Envolvendo Grande Quantidade de Computadores	57
2.3.1	Duplicar Todos os DADs Para Posterior Análise	58
2.3.2	Analisar Antes de Coletar e Duplicar os DADs Parcialmente	59
2.3.3	Analisar Antes e Coletar Somente Informações Parciais.....	59
2.3.4	Escolher Aleatoriamente	60
2.3.5	Considerações Sobre IDF Envolvendo Grande Quantidade de Computadores ..	60
2.4	Trabalhos Correlatos	61
2.4.1	Técnicas de Triagem	64
2.4.2	Resumo e Análise dos Trabalhos Relacionados.....	69
3	MÉTODO PARA TRIAGEM E RANQUEAMENTO DE COMPUTADORES	71
3.1	Introdução.....	71
3.2	Considerações Iniciais Sobre o Método Proposto.....	72
3.3	Visão Geral do Método Proposto	74
3.3.1	O Método Proposto e o Modelo de Referência Utilizado	78
3.4	Características dos Processos que Compõem o Método Proposto	80
3.4.1	Processos de Duplicação e Varredura	80
3.4.1.1	Carregamento da Lista de Palavras-chave	81
3.4.1.2	Processo de Duplicação.....	83
3.4.1.3	Processo de Varredura.....	84
3.4.2	Processo de Filtragem	88
3.4.2.1	Cálculo da Preditividade Positiva do Processo de Filtragem.....	90
3.4.3	Processo de Triagem	91

3.4.3.1	Cr�terios Para o Processo de Triagem	91
3.4.4	Processo de Ranqueamento	94
3.4.4.1	Procedimentos Para o Processo de Ranqueamento	94
3.5	Exemplos Para Demonstra��o dos Processos de Triagem e Ranqueamento	95
4	VALIDA��O DO M�TODOS PARA TRIAGEM E RANQUEAMENTO DE COMPUTADORES.....	103
4.1	An�lise de Ferramentas Para Valida��o do M�todo	104
4.1.1	<i>Disk Duplicator</i> (DD).....	104
4.1.2	<i>Autopsy</i>	105
4.1.3	Resultados da An�lise das Ferramentas para Duplica��o e Varredura	106
4.2	Ferramentas Implementadas.....	107
4.2.1	Ferramenta <i>CopyScan</i>	107
4.2.1.1	Carregamento dos Dados de Entrada	110
4.2.1.2	Carregamento da Lista de PChs	113
4.2.1.3	Duplica��o e Varredura.....	115
4.2.1.4	M�dulo Duplicador	115
4.2.1.5	M�dulo de Varredura	116
4.2.1.6	Algoritmo de Compara��o	118
4.2.1.7	Defini��o do Passo de Varredura	119
4.2.1.8	Outras Funcionalidades da Ferramenta <i>CopyScan</i>	121
4.2.1.9	Restri��es da Ferramenta <i>CopyScan</i>	122
4.2.2	Ferramenta FiltroP.....	126
4.2.3	Ferramenta <i>EraseFiles</i>	126
4.3	Processos Realizados Pelas Ferramentas <i>CopyScan</i> e FiltroP	127
4.4	Procedimentos para Avalia��o dos Processos de Varredura e Filtragem	128
4.4.1	Verifica��o de Ocorr�ncias em Cada Parte do DAD	130
4.4.1.1	Verifica��o de Ocorr�ncias VP e FP em Cada Parte do DAD.....	131

4.4.1.2	Verificação de Ocorrências no DAD Todo	132
4.4.1.3	Verificação de Ocorrências em Arquivos nos <i>Clusters</i> Alocados	133
4.4.1.4	Verificação de Ocorrências em Arquivos Interpretáveis e Mistos	135
4.4.1.5	Verificação de Ocorrências em Arquivos Interpretáveis	137
4.4.1.6	Considerações Sobre os Procedimentos para Avaliação dos Processos de Varredura e Filtragem	138
4.5	Restrição Para Validação da Triagem	138
4.6	Validação da Sequência de Ranqueamento dos DADs	139
4.7	Considerações Finais Sobre o Método Proposto	139
5	EXPERIMENTOS E TESTES	141
5.1	Considerações Iniciais Sobre os Experimentos	142
5.1.1	Descrição do Cenário e do Conjunto de Dados <i>M57-Patents</i>	144
5.1.1.1	Imagens dos DADs que Compõem o Cenário <i>M57-Patents</i>	145
5.1.3	Equipamento Utilizado e DADs de Origem e Destino	147
5.1.3.1	Dispositivos de Origem (DAD _O) e Destino (DAD _D)	148
5.1.4	Esterilização dos DADs de Origem e Destino	148
5.1.4.1	Exemplo de Esterilização dos DADs de Origem e Destino	149
5.2	Experimento 1	151
5.2.1	Experimento 1 - Parte A	151
5.2.1.1	Preparação e Procedimentos	151
5.2.1.2	Duplicação dos DADs	152
5.2.1.3	Resultados do Experimento 1 - Parte A	152
5.2.2	Experimento 1 - Parte B	153
5.2.2.1	Preparação e Procedimentos	153
5.2.2.2	Escolha dos Arquivos de Imagens .E01	153
5.2.2.3	Descompactação dos arquivos .E01 e gravação no DAD _O	154
5.2.2.4	Resultados Experimento 1 - Parte B	154
5.2.2.5	Considerações Sobre os Resultados do Experimento 1	155

5.3	Estimativa de Tempo para DADs com Diferentes Capacidades.....	155
5.4	Experimento 2	157
5.4.1	Preparação e Procedimentos.....	158
5.4.1.1	Resultados do Experimento 2 - Parte A	160
5.4.1.2	Resultados do Experimento 2 - Parte B	161
5.4.1.3	Resultados do Experimento 2 - Parte C	163
5.4.2	Considerações Sobre os Resultados do Experimento 2	164
5.5	Experimento 3	166
5.5.1	Preparação e Procedimentos.....	166
5.5.2	Resultados do Experimento 3	167
5.5.2.1	Resultados do Experimento 3 - Parte A	167
5.5.2.2	Resultados do Experimento 3 - Parte B	168
5.5.3	Considerações Sobre os Resultados do Experimento 3	169
5.6	Experimento 4	170
5.6.1	Preparação e Procedimentos.....	170
5.6.1.1	Resultados do Experimento 4.....	172
5.6.1.2	Considerações Sobre os Resultados do Experimento 4	177
5.6.2	Quantidade de ocorrências de PChs no DAD todo (QO_{DT})	179
5.7	Experimento 5	180
5.7.1	Preparação e Procedimentos.....	180
5.6.2	Resultados do Experimento.....	182
5.6.2.1	Resultados do Experimento – Parte A.....	183
5.6.2.2	Resultados do Experimento – Parte B.....	185
5.7	Experimento 6	190
5.7.1	Preparação e Procedimentos.....	190
5.7.2	Resultados do Experimento 6.....	191
5.8	Considerações Finais Sobre os Experimentos.....	194

6	DISCUSSÃO DOS RESULTADOS E CONCLUSÕES.....	195
6.1	Resultados da Duplicação e Varredura	195
6.2	Resultados da Filtragem	197
6.3	Resultados da Triagem	199
6.4	Resultados do Ranqueamento	199
6.5	Considerações Finais da discussão dos Resultados.....	200
6.5	Conclusões	200
6.5.1	Contribuições	202
6.5.2	Trabalhos Futuros e Perspectivas de Continuidade.....	203
	REFERÊNCIAS	205
	APÊNDICE A: ALGORITMO DA FERRAMENTA COPYSCAN.....	225
	APÊNDICE B: UTILIZAÇÃO DA FERRAMENTA COPYSCAN.....	229

1 INTRODUÇÃO

A informatização da sociedade vem ocorrendo de forma acelerada. Nas últimas décadas houve em escala global a popularização, expansão e dependência da sociedade moderna à informação armazenada ou transmitida em formato digital, que se utiliza principalmente do acesso às tecnologias de Informação e Comunicação [NERI, 2012].

O número de computadores em uso no Brasil em 2009 era 64 milhões, o que representava um computador para cada três habitantes. Em 2012, esta quantidade aumentou para 99 milhões, o que significa um computador para cada dois habitantes. A perspectiva é de que até 2017 o número de computadores no país supere a quantidade de habitantes [GUIMARÃES, 2012].

Computadores servidores de grande porte, computadores pessoais (*desktops*), equipamentos de computação móvel (*notebooks*, *palmtops*, *tablets*) e aparelhos de telefone celular com funcionalidades avançadas (*smartphones*), entre outros equipamentos controlados por microprocessadores, são amplamente utilizados para finalidades pessoais e profissionais.

Nesse amplo universo de equipamentos é possível identificar uma variedade de dispositivos para armazenamento de dados (DADs¹) com diferentes tipos de interfaces. Dentre esses dispositivos estão unidades de discos (*Hard disk Drive* (HDD) ou *Solid State Disk* (SSD)), *pen-drives*, cartões de memória - *Multi Media Card* (MMC), *Memory Stick* (MS), *Micro Secure Digital* (MSD), *Secure Digital* (SD), *Subscriber Identity Module* (SIM), *Blu-Ray Disc*, *Compact Disc* (CD), *Digital Video Disc* (DVD) e congêneres.

Esses DADs são utilizados para armazenar o contínuo e crescente volume de dados gerado por usuários, que recursivamente necessitam cada vez mais por DAD com maior capacidade.

Em 2009, um dos grandes desafios identificados pela Sociedade Brasileira de Computação foi o tratamento de grandes volumes de dados [SBC, 2009]. Corroborando esse desafio, o relatório publicado pelo IEEE 802.3 *Working Group* (2012) ratificou a explosão do volume de dados em formato digital nas últimas décadas, que aumentou de 130 Exabytes (EB) em 2005 para 1227 EB em 2010, com previsão de crescimento exponencial que alcance 7910 EB ao final de 2015. O desafio para tratamento de grandes volumes de dados ainda permanece [KARIE, 2015].

Segundo a previsão de Mark H. Kryder², em poucos anos os consumidores americanos teriam em média 10 a 20 unidades de DADs, distribuídos entre vários computadores e outros equipamentos de uso regular [WALTER, 2005]. Análoga a Lei de Moore, a Lei de Kryder previu que a capacidade dos discos rígidos dobraria e o preço cairia pela metade a cada 18 meses a partir de 1979. O gráfico da Figura 1.1

¹Neste trabalho a terminologia "dispositivo para armazenamento - DAD" é utilizada para referir-se a disco (magnético ou estado sólido), cartão de memória ou qualquer mídia utilizada para gravar e armazenar dados.

²Mark H. Kryder (nascido em 07 de Outubro de 1943 em Portland, Oregon) é vice-presidente de pesquisas e diretor de tecnologia da empresa Seagate Technology e professor de Engenharia Elétrica e Computação na Carnegie Mellon University.
Disponível em: <http://www.aps.org/units/fiap/awards/recipient.cfm?first_nm=Mark&last_nm=Kryder&year=2007>. Acesso em: 10 ago. 2012.

ilustra a curva da estimativa para a média de crescimento da capacidade das unidades de discos entre os anos de 1979 a 2015 [ESENER, 1999; WALTER, 2005].

Figura 1.1 - Estimativa para o aumento da capacidade dos discos

Observando o gráfico nota-se o crescimento da capacidade de armazenamento dos discos rígidos ocorrida e a perspectiva de que este crescimento continue nos próximos anos, podendo alcançar 10 Terabytes (TB) em 2015 [KRYDER, 2009; WALTER, 2005]. Na prática a previsão de Kryder para 2015 foi confirmada. Analogamente, acredita-se que essa perspectiva do aumento de capacidade também pode ser estendida para os demais tipos de DADs.

Com a informatização crescente das várias atividades desenvolvidas individual ou coletivamente na sociedade houve a intensa proliferação ubíqua de computadores, o considerável aumento da quantidade e capacidade dos DADs e da mesma forma o enorme e contínuo crescimento do volume de dados gerados, armazenados e distribuídos.

A maneira como esses recursos estão sendo massivamente inseridos no cotidiano da vida das pessoas, desencadeou uma série de transformações e efeitos colaterais. Infratores perceberam nesse cenário uma ampla oportunidade de exploração para o cometimento de

atividades ilícitas. Computadores e as tecnologias associadas a estes propiciaram o surgimento de novas modalidades de crimes. Esses equipamentos passaram ser alvo de crimes ou utilizados como instrumento para sua prática [ALI, 2012; COSTA, 2012; KRUSE, 2002; LANGE, 2010].

Em outras circunstâncias, não se trata necessariamente de novas modalidades de crimes, mas de alguma maneira esses equipamentos podem ter sido utilizados indevidamente, estabelecendo alguma relação com crimes praticados, como por exemplo, homicídio, pedofilia, concorrência desleal, extorsão, furto, tráfico de entorpecente, estelionato, violação de direitos autorais, dentre outros crimes [CASTRO, 2003; COSTA, 2012; KÄLBER, 2014; LUCCA, 2005].

Devido à elevação da quantidade de atividades ilícitas envolvendo o uso de tecnologias de computadores, Internet, meios eletrônicos e outros equipamentos digitais, houve a criação de diversos órgãos e delegacias especializadas na investigação e repressão dessas atividades. No Brasil, alguns desses órgãos, como é o caso por exemplo, da Polícia Federal do Brasil e do Instituto Geral de Perícias do Estado do Rio Grande do Sul registraram o aumento significativo de crimes, nos quais algum recurso tecnológico relacionado à computadores foi utilizado [HOELZ, 2009; TESSMANN, 2012].

Da mesma maneira, também se registrou a elevação da quantidade de processos que tramitam nos diversos órgãos do Judiciário e demandam por investigações periciais envolvendo o uso de computadores ou algum tipo de tecnologia associada a estes.

Conforme levantamento realizado pelo autor deste trabalho, algumas buscas utilizando a palavra-chave "perícia de informática" junto ao Portal da Justiça Federal³, Tribunal de Justiça do Estado de São Paulo (TJ-SP⁴) e *site* JusBrasil⁵, que apresenta jurisprudências

³<<https://www2.jf.jus.br/juris/unificada/>>. Acesso em: 18 ago. 2013

⁴<<http://esaj.tjsp.jus.br/esaj/portal.do>>. Acesso em: 18 ago. 2013

para o Superior Tribunal de Justiça (STJ), Superior Tribunal Militar (STM), Supremo Tribunal Federal (STF), Tribunais de Contas dos Estados (TCE), Tribunais Regionais Federais (TRF), Tribunal Superior do Trabalho (TST) e Tribunal Superior Eleitoral (TSE) identificaram mais de 9.000 processos nos últimos cinco anos. Observando que nesse número não estão contabilizados os processos que tramitaram ou ainda tramitam em segredo de justiça, cujos números não estão disponíveis.

As informações anteriores ratificam a importância que computadores e outros equipamentos com capacidade para armazenamento de dados digitais assumem na investigação de atividades ilícitas. Por esse motivo, seus respectivos DADs são considerados peças importantes, devido à possibilidade de que no conteúdo destes sejam encontrados vestígios⁶ que servirão como pistas para constatar ou não a prática e autoria de atividades ilícitas.

Investigações envolvendo computadores podem ocorrer em diferentes âmbitos, por exemplo, corporativo, criminal, militar ou judicial. A diferença entre eles está no escopo da investigação e nos procedimentos utilizados. Neste trabalho a ênfase é voltada à esfera judicial.

Nesse contexto, para investigar e constatar se um determinado DAD contém vestígios de interesse, faz-se necessária a aplicação de um conjunto de procedimentos que sejam juridicamente válidos.

Diversos modelos têm sido propostos na literatura, conforme será visto na Seção 2.2.1. A diferença entre eles está basicamente na quantidade de etapas e nos procedimentos realizados nas respectivas

⁵<http://www.jusbrasil.com.br/jurisprudencia/busca?q=per%C3%ADcia+de+inform%C3%A1tica>

⁶Na literatura internacional é amplamente utilizado o termo "*evidence*" - evidência. Devido este termo não ser definido na legislação vigente no Brasil, neste trabalho optou-se pela utilização da palavra vestígio, como referência a qualquer dado encontrado no DAD que possa ter relação com o fato investigado. Dessa forma, vestígios são as palavras-chave de interesse para a investigação.

etapas. Não há um modelo de Investigação Digital Forense⁷ (IDF) definido como padrão [VALJAREVIC, 2015]. Neste trabalho é utilizado um modelo de seis etapas, baseado nos trabalhos de [CISAR, 2011; PALMER, 2001; PILLI, 2010; RUIBIN, 2005]. Essas etapas e a sequência para realização estão ilustradas na Figura 1.2. Este modelo engloba objetivamente as etapas necessárias para condução de uma investigação digital no âmbito da perícia judicial. Devido ao objetivo desta tese, conforme será visto na Seção 1.3, o foco está voltado para a etapa de coleta.

Figura 1.2 - Etapas para o procedimento de Investigação Digital Forense

Este conjunto de etapas têm recebido diferentes denominações: *Digital Investigation*, *Computer Forensic* e *Digital Forensics* [CASEY, 2000; HU, 2009; KRUSE, 2002; REIS, 2002; ROGERS, 2004]. Tais

⁷ No presente trabalho é utilizado o termo Investigação Digital Forense (IDF) para referir-se genericamente a toda e qualquer perícia judicial de informática, ou seja, investigações com propósitos judiciais relacionadas às áreas de Computação, Informática e Telecomunicações, que envolva a análise de redes de dados ou qualquer tipo de equipamento relacionado com estas áreas. Computação, Informática ou Telecomunicações.

denominações e a descrição de cada uma das etapas mostradas na Figura 2.1 serão aprofundadas no Capítulo 2.

Os vestígios encontrados no conteúdo dos DADs por meio da IDF podem ser utilizados para confirmar ou refutar hipóteses sobre ações que um indivíduo realizou ou não no passado [SHIELDS, 2011]. Dessa forma, auxilia o Perito Judicial⁸ em suas conclusões, para responder quesitos formulados pelo juiz da causa e pelas partes (autor/réu) envolvidas, para esclarecer questões de interesse para o litígio, como por exemplo, o que ocorreu, como ocorreu, autoria (quem praticou ou participou), quando ocorreu, onde e porque, e assim auxiliar a tomada de decisão do juiz [BODEAU, 2010; BRANTINGHAM, 2011].

Ressaltando que neste trabalho a ênfase está voltada para Investigação Digital Forense, mais especificamente em casos onde se faz necessária a análise pericial em um cenário composto por uma grande quantidade de computadores, um problema é o curto tempo que é concedido para a realização dos trabalhos, e que pode ser despendido analisando desnecessariamente DADs que não apresentam os vestígios de interesse para a investigação. Considerando o dever que o perito tem em entregar o documento de laudo pericial dentro do prazo determinado pelo juiz. Observando que normalmente este prazo já é determinado no momento da nomeação do perito, conforme o Artigo (Art.) 421 do Código de Processo Civil (CPC) [BRASIL, 1973].

Mesmo por motivos legítimos o perito tenha a possibilidade de solicitar ao juiz a prorrogação do prazo para realização dos trabalhos, o tempo concedido pode ser insuficiente para analisar todos os dispositivos, concluir os trabalhos e apresentar o documento de laudo pericial. Salientando que a extensão do prazo, não é infinita, ficando condicionado ao prudente arbítrio do juiz, mas limitada no máximo há

⁸Profissional técnico que não guarda vínculo empregatício ou obrigacional com o Poder Público estando, porém habilitado a prestar serviços junto ao Judiciário (sem prejuízo de suas atividades profissionais cotidianas) mediante a nomeação por um juiz ou contratação de uma das partes de uma demanda judicial, e recebe o encargo de prestar esclarecimentos a respeito da matéria em que é *expert*.

20 dias antes da audiência de instrução e julgamento, conforme determina a Lei.

1.1 Motivação

O aumento da quantidade e diversidade de computadores e seus respectivos DADs, juntamente com o crescente volume de dados causaram mudanças nas IDFs [BRADY, 2014; GARFINKEL, 2010; KARIE, 2015; NANCE, 2009]. Esses fatores tornaram os métodos de investigações digitais impraticáveis e as ferramentas tradicionalmente utilizadas insustentáveis, ultrapassando os limites da capacidade de investigação humana. Isso causa um enorme acúmulo de trabalho, devido à impossibilidade de analisar o conteúdo de todos os DADs encontrados no local da perícia dentro do tempo determinado pelo juiz.

Pesquisas na área de investigação digital foram realizadas tendo em vista o desenvolvimento de técnicas que possibilitem diminuir a quantidade de DADs e com isso o volume de dados a ser analisado, auxiliado para a redução do tempo e esforços gastos com a análise de DADs que não apresentam chances de conter os vestígios de interesse para uma determinada investigação [GARFINKEL, 2006; GRILLO, 2009; HALBOOB, 2014; O'CONNOR, 2004], apenas para mencionar alguns trabalhos.

Essas pesquisas têm produzido propostas de soluções baseadas nas mais variadas técnicas, dentre as quais, agrupamento (*clustering*) [MORA, 2010], amostragem estatística (*statistical sampling*) [JONES, 2012; TESSMANN, 2012], aprendizagem de máquina (*machine learning*) [NASSIF, 2011; PROVOST, 1998], mineração de dados (*data mining*) [BHAT, 2010; DAN HU, 2009; LU QIN, 2010], procura por palavras-chave (*string searching*) [BEEBE, 2007a; BEEBE, 2007b], reconhecimento de padrões (*pattern recognition*) [PUNGILA, 2012].

Embora diversos autores salientem a necessidade de desenvolvimento de novos métodos para a realização das atividades de

IDF, existe uma lacuna entre os aspectos teóricos e práticos [GARFINKEL, 2012; MARTURANA, 2012; MOHAY, 2005; ROGERS, 2006].

Uma abordagem que vem sendo amplamente discutida e explorada na literatura da área é a utilização de técnicas de triagem [CASEY, 2013; GARFINKEL, 2013; ROGERS, 2006; SHIAELES, 2013]. Por definição, a palavra triagem⁹ significa escolha ou separação em classes ou local onde se faz a seleção. No contexto de IDF, pode ser definida como o processo para identificação e seleção de DADs que contém os vestígios de interesse para determinada investigação [POLLITT, 2013].

No decorrer das diligências¹⁰ judiciais para realização de IDF, destacando medidas cautelares com mandado de busca e apreensão, nas quais o perito não conhece previamente o local, é possível que o mesmo se depare com ambientes de redes computacionais corporativas com dezenas, centenas ou até milhares de computadores conectados, cada qual possuindo um ou mais DADs (internos e externos) e não raro com diferentes tipos de interfaces.

Para cenários com essas características, por exemplo, grandes corporações ou um campus universitário, o perito deve decidir se os equipamentos que contém os DADs serão analisados no próprio local da diligência ou serão apreendidos e examinados posteriormente em laboratório. Independentemente desta decisão, é recomendado que o perito aplique as melhores práticas, aceitas pela comunidade científica e siga procedimentos juridicamente válidos. Uma prática tradicionalmente seguida, que visa preservar a integridade dos dados originais, consiste na duplicação completa (*byte-a-byte*) de todos os

⁹ Moderno Dicionário Michaelis. Editora Melhoramentos Ltda. © 2009 UOL. Disponível em: <<http://michaelis.uol.com.br/moderno/portugues/index.php?lingua=portugues-portugues&palavra=triagem>>. Acesso em: 01 mar. 2014.

¹⁰ Investigação judicial realizada fora do fórum.

DADs suspeitos de conter os vestígios de interesse para a perícia [KRUSE, 2002; CANTRELL, 2012].

Porém, nem sempre é possível identificar imediatamente quais são os dispositivos mais prováveis de conter os vestígios de interesse para a investigação. Na dúvida, o perito deve coletar os dados de todos os DADs suspeitos para posterior análise. Nessas circunstâncias, a realização de IDF pode consistir em uma tarefa árdua e morosa, demandando demasiado tempo e esforço, predominantemente na etapa de análise, o que pode implicar no prazo para a apresentação dos resultados [CLAYTON, 2012; GARFINKEL, 2010; GARFINKEL, 2012].

A redução do esforço humano e o tempo gasto na etapa de análise representa um importante desafio para a IDF [BRADY, 2014; CANTRELL, 2012; KÄLBER, 2014; LEE, 2008; YANG, 2008].

Métodos que possam ser utilizados para oferecer indicativos dos DADs com maiores chances de conter vestígios de interesse para o caso investigado e que possibilitem priorizar equipamentos que devem ser examinados na etapa de análise, podem auxiliar a IDF para reduzir o esforço e tempo gasto desnecessariamente com a análise de DADs que não apresentam chances de conter vestígios de interesse [CASEY, 2013; GRILLO, 2009; NANCE, 2009; YANG, 2008].

Os fatores anteriormente expostos motivaram a abordagem para o problema de pesquisa proposto nesta tese.

1.2 Problema de Pesquisa

Diversas abordagens foram propostas na literatura na tentativa de identificar computadores que apresentam conteúdo de interesse para investigações digitais forenses, e com isso, direcionar o foco da análise para estes computadores e, dessa forma, diminuindo o tempo e esforços necessários para alcançar os resultados.

O problema de pesquisa desta tese consiste em, a partir de um conjunto de computadores com seus respectivos dispositivos de armazenamento de dados (DADs), objetos de uma investigação digital forense – IDF, descobrir:

- I. Como identificar os DADs que apresentam maiores chances de conter os vestígios de interesse para a investigação apenas com os dados obtidos da varredura durante a duplicação?
- II. Como sugerir uma sequência que os DADs devem ser examinados na etapa de análise, baseada em prioridades de palavras-chaves¹¹ de interesse para a investigação?

1.3 Objetivo

Uma das tarefas mais comum para um perito judicial que atua na área de Investigação Digital Forense é procurar vestígios de interesse no conteúdo de DADs, esses vestígios na maioria das vezes consistem em palavras-chave - PChs [BEEBE, 2014; YANG, 2008].

Considerando que durante o tempo necessário para a realização da duplicação do DAD os dados contidos no mesmo ficam praticamente inacessíveis para o perito realizar outras ações, ou seja, o perito fica praticamente impossibilitado de interagir com os dados contidos no DAD, decidiu-se verificar a hipótese de que na etapa de coleta, seja possível realizar simultaneamente à duplicação do DAD de origem para o DAD de destino a varredura¹² dos dados para procurar por palavras-

¹¹No contexto deste trabalho o termo “palavras-chave” deve ser entendido como sendo uma estrutura de dados composta por uma sequência de caracteres [CASTRO, 1998]. Na literatura também é utilizado o termo “*string*”. Vale a pena ressaltar que embora toda palavra-chave seja uma *string*, nem toda *string* é uma palavra-chave.

¹²A palavra varredura no contexto em que foi mencionada deve ser entendida como sendo uma inspeção automática que realiza uma pré-análise para busca de *strings* no conteúdo dos dados brutos que estão sendo duplicados de um DAD_O para um DAD_D.

chave de interesse do perito, sem com isso impactar significativamente o tempo de duplicação.

Assumindo-se que a hipótese colocada anteriormente seja verdadeira, o **principal objetivo** desta tese é analisar a possibilidade de realizar simultaneamente à duplicação do DAD de origem para o DAD de destino, a varredura dos dados para procurar por palavras-chave de interesse do perito, sem com isso impactar significativamente o tempo de duplicação. Então, baseando-se na quantidade de ocorrências de palavras-chave encontradas, propor um método para triagem e ranqueamento dos DADs duplicados, de forma a apresentar uma sugestão de sequência, indicando quais desses DADs podem ser prioritariamente examinados na etapa de análise, e assim, responder:

- a) A proposta é viável? Caso afirmativo, para quais cenários?
- b) Quais tipos de dados podem ser reconhecidos?
- c) Como definir um ranqueamento baseado nos dados coletados durante a varredura?
- d) Quais os indicadores de desempenho que devem ser analisados para avaliar este tipo de varredura?
- e) A aplicação do método contribui para a diminuição do esforço humano na etapa de análise?

O intuito do método não é substituir o exame detalhado e minucioso realizado na etapa de análise, e sim auxiliar na identificação dos DADs com maior chance de conter as PChs de interesse para a investigação, indicando uma sequência em que os DADs devem ser prioritariamente examinados na etapa de análise.

Dependendo da investigação, encontrar as PChs de interesse nos primeiros DADs, pode não ser necessário analisar todos os DADs. Não está no escopo desse trabalho a realização dos procedimentos intrínsecos da etapa de análise.

Diversos trabalhos encontrados na literatura utilizam técnicas para procura por palavras-chave, duplicação e triagem, isoladamente.

Esta tese, apresenta uma abordagem que utiliza esses conceitos de maneira combinada, de forma a obter os primeiros resultados para a investigação já ao término da duplicação, visto que realiza automaticamente a procura por palavras-chave simultaneamente à duplicação. Então, utiliza estes resultados para realizar a triagem e ranqueamento dos DADs duplicados, sugerindo uma sequência em que os DADs deverão ser examinados na etapa de análise. Essa abordagem não é observada em outros trabalhos da literatura, caracterizando assim, a originalidade desta tese. Na Seção 1.4, são apresentados maiores detalhes das contribuições deste trabalho.

1.4 Contribuições

A principal contribuição desta tese está no método proposto para duplicação e varredura simultânea, triagem e ranqueamento, que fornece as primeiras informações da quantidade de ocorrências de PChs encontradas no DAD já ao término da duplicação, ainda na etapa de coleta. Para determinadas investigações, a aplicação deste método pode auxiliar para a diminuição do tempo e esforços necessários para realizar a etapa de análise. O intuito é propiciar o direcionamento do foco na etapa de análise para os dispositivos considerados mais relevantes para a investigação, o que em certas circunstâncias implica na diminuição da quantidade de DADs e dados a serem analisados.

Para validação do método foram desenvolvidas e implementadas três ferramentas, cujas funcionalidades estão descritas no Capítulo 4. Os primeiros testes realizados, indicaram resultados positivos para obtenção de ocorrências das PChs procuradas durante a duplicação.

Os resultados provenientes deste trabalho servirão como um recurso para auxiliar peritos judiciais no desempenho das atividades de IDF. Embora o método seja focado no trabalho de peritos judiciais de informática, não é restrito a estes, pois o mesmo também poderá ser

utilizado por peritos criminais e profissionais da área de auditoria e segurança da informação.

1.4.1 Produção Científica

A obtenção dos primeiros resultados deste trabalho, ainda que preliminares, permitiu a publicação do artigo:

BARBOSA, A. N.; OKIDA, C.; RUGGIERO, W. V. Método Para Triagem e Agrupamento de Computadores Baseado em Perfil Computacional Aplicado à Computação Forense. Anais da 9ª Conferencia Ibérica de Sistemas y Tecnologías de Informacion, 2014, Barcelona, España. Actas de 9ª Conferencia Ibérica de Sistemas y Tecnologías de Información. v. II. p. 445-448.

1.5 Sequência para desenvolvimento do trabalho

Tendo em vista realizar as constatações do problema de pesquisa, verificação da hipótese e responder as questões colocadas no objetivo, anteriormente descritas, o método de pesquisa adotado para o desenvolvimento desse trabalho pode ser sintetizado conforme descrito a seguir.

Inicialmente contemplou-se o levantamento de referências bibliográficas para estudo e identificação das principais técnicas relacionadas com os métodos utilizados para otimização, diminuição do volume de dados e do tempo para realização das etapas de IDF, mais especificamente técnicas e ferramentas utilizadas para duplicação de DADs, procura por PChs e triagem de computadores (DADs e o conteúdos de dados armazenados).

No decorrer deste levantamento, verificou-se que nenhum trabalho combinava técnicas de duplicação e varredura para procura por PChs simultaneamente, bem como a utilização desses resultados

para realização de triagem e ranqueamento de DADs, de forma a propiciar a indicação de uma sequência em que os DADs que apresentassem maiores chances de conter as PChs de interesse para a investigação poderiam ser examinados prioritariamente na etapa de análise. Dessa forma foi identificada a lacuna para o problema de pesquisa, a hipótese levantada e o vislumbre da possibilidade para a concepção do método proposto nesta tese.

Então, foram estudadas possibilidades de como PChs poderiam ser identificadas e contabilizadas durante a realização de uma varredura sobre dados brutos simultaneamente à duplicação do DAD.

Posteriormente, foi iniciada a procura por discos rígidos e outros dispositivos de armazenamento contendo dados e conjuntos de arquivos de imagens de discos rígidos que pudessem ser utilizados nos experimentos de prova de conceito e para validação do método proposto. Paralelamente às atividades anteriores, houve a definição de características necessárias e especificação de ferramentas de softwares que poderiam ser utilizadas para validação do método durante os experimentos e testes. Houve uma fase que contemplou testes e simulações com as diversas ferramentas encontradas. Posteriormente, ocorreu a implementação de ferramentas conforme as especificações necessárias. Outra etapa, consistiu no delineamento dos cenários para realização dos experimentos e testes necessários para a validação do método proposto. Então, por último foram realizadas as simulações, execução dos experimentos e testes para validação do método proposto nesta tese.

1.6 Organização do Trabalho

Este capítulo situa o leitor apresentando os fatores que motivaram a pesquisa, a proposição do problema de pesquisa, as questões que se pretende responder ao final do trabalho, o objetivo, a metodologia seguida para o desenvolvimento e obtenção dos resultados.

Os capítulos subsequentes estão organizados da seguinte forma:

- **Capítulo 2 - Conceitos Fundamentais e Trabalhos Correlatos.** Fornece uma visão geral e definição de termos e conceitos relativos à perícia judicial de informática, bem como as etapas e ciclo de vida para realização da mesma e a revisão da literatura com os principais resultados já alcançados nos trabalhos correlatos.
- **Capítulo 3 - Método para Triagem e Ranqueamento de Computadores.** Apresenta a descrição do método proposto.
- **Capítulo 4 - Validação do Método para Triagem e Ranqueamento de Computadores.** Destaca as ferramentas testadas, as ferramentas desenvolvidas e os procedimentos seguidos validação do método proposto.
- **Capítulo 5 - Experimentos e Testes.** Descreve os testes realizados para a validação do método proposto e os resultados obtidos.
- **Capítulo 6 – Discussão dos Resultados e Conclusões.** Apresenta os resultados e discussões, dificuldades encontradas e as considerações sobre os experimentos realizados. Aborda também as conclusões, as considerações sobre o objetivo do trabalho, as respostas às perguntas colocadas no Capítulo 1 e discute sucintamente as limitações encontradas. Ao final sugere indicações de possíveis caminhos de investigação que podem ser seguidos para a realização de trabalhos futuros.

2 CONCEITOS FUNDAMENTAIS E TRABALHOS CORRELATOS

Este capítulo apresenta uma visão geral de conceitos relacionados com a Investigação Digital Forense - IDF, abordando as principais terminologias utilizadas neste trabalho, bem como a descrição das etapas e procedimentos, úteis para o entendimento de assuntos aqui abordados. Em seguida, apresenta uma abordagem dos trabalhos encontrados na literatura que estão relacionados com a pesquisa desenvolvida nesta tese.

De acordo com Sammons (2012), nenhuma conversa sobre IDF pode ser completa sem incluir aspectos Legais. Porém, devido ao caráter predominantemente técnico deste trabalho, a linguagem adotada foi destituída de características jurídicas.

2.1 Terminologias

2.1.1 Forense e Ciência Forense

A palavra Forense¹³ (*Forensic*) é uma designação que diz respeito ao foro judicial ou relativo aos tribunais de justiça. Para o termo Ciência Forense¹⁴ (*Forensic science*), várias definições são encontradas na literatura. Franke (2008), a define como sendo a correta aplicação metodológica de diversas disciplinas da ciência para responder questões de interesse de um tribunal de justiça.

De acordo com a Academia Americana de Ciências Forenses (AAFS)¹⁵:

"Ciência forense é a aplicação de princípios científicos e práticas tecnológicas para propósitos judiciais no estudo e resolução de questões criminais e cíveis."

No entendimento do autor, pode ser definida como a aplicação das ciências naturais e métodos científicos para as questões judiciais. Na prática e de maneira ampla, a ciência forense baseia-se em diversas áreas do conhecimento para a identificação, individualização, avaliação e documentação de vestígios físicos, para serem apresentados em um tribunal de justiça.

¹³Em latim, a palavra *forēnsis* significa pertencente ao fórum, que é relativo aos tribunais, ao judiciário [FARIA, 1962].

¹⁴Tradução do inglês para "*Digital Forensic Science*". Disponível em: <<http://www.all-about-forensic-science.com/definition-of-forensic-science.html>>. Acesso em: 12 jun. 2012.

¹⁵*American Academy of Forensics Sciences*. Disponível em: <<http://www.thefreedictionary.com/computational>>. Acesso em: 11 jun. 2012.

2.1.2 Forense Computacional, Computação Forense, Digital Forense e Investigação Digital

O termo "computacional"¹⁶ tem sido associado com diversas disciplinas do conhecimento humano. Como por exemplo, visão computacional, linguística computacional, química computacional e etc. [FRANKE, 2009].

A área de Forense Computacional (*Computational Forensic*) funciona como ferramenta de apoio para o campo da Ciência Forense auxiliando, mediante o uso de computadores e métodos computacionais, na avaliação das bases dos métodos científicos, na investigação de hipóteses para um problema específico e na obtenção de resultados.

Dessa forma, conforme Franke (2008) e Andaló (2012) a Forense Computacional auxilia pesquisadores, peritos e profissionais da área pericial principalmente de três maneiras:

1. Provendo ferramentas de software para melhor análise e acurácia dos vestígios, superando determinadas limitações da habilidade cognitiva humana.
2. Oferecendo bases científicas para disciplinas ou procedimentos forenses que demandam por análise de grande volume de dados, tarefa que pode ser muitas vezes humanamente impossível.
3. Reproduzindo o conhecimento de especialistas pela implementação, em máquina, de suas habilidades de reconhecimento e raciocínio.

Ainda conforme Franke (2008) e Andaló (2012), de uma maneira mais abrangente, a Forense Computacional, por meio de algoritmos e

¹⁶Tradução do inglês para "*computational*". Relacionado a assuntos envolvendo computação ou computadores. Disponível em: <<http://www.thefreedictionary.com/computational>>. Acesso em: 12 jun. 2012.

métodos computacionais permite aos pesquisadores, peritos e profissionais da área:

- a) Identificar e melhorar traços de vestígios para posterior análise;
- b) Analisar vestígios de forma objetiva e reproduzível;
- c) Avaliar a qualidade de métodos de análise;
- d) Descrever e padronizar procedimentos de análise;
- e) Pesquisar grandes volumes de dados de forma eficiente;
- f) Visualizar e documentar os resultados de uma análise;
- g) Auxiliar na interpretação de resultados e em sua argumentação;
- h) Revelar padrões, ou vínculos, antes desconhecidos, para derivar novas regras e contribuir para a geração de novos conhecimentos.

A necessidade de investigações com características predominantemente técnicas de crimes nos quais de alguma maneira computadores podem estar envolvidos, levou a criação de um novo ramo na Ciência Forense denominada de Computação Forense (CF) [FERNANDEZ, 2005; GARBER, 2001]. A raiz utilizada para denominação segue outras áreas da Ciência Forense, como por exemplo, a Medicina Forense, Contabilidade Forense e a Química Forense [BERGUEL, 2003; GLADYSHEV, 2004]. A CF envolve a obtenção e análise de vestígios digitais que podem ser usados como prova em um tribunal de justiça [NELSON, 2010].

Na literatura, nota-se a utilização dos termos Computação forense e Digital Forense (DF) indistintamente, sendo empregadas como sinônimos. Observa-se também uma gama de diferentes definições. Uma das primeiras definições para CF que tem sido amplamente adotada desde o inaugural *Workshop* de Investigação em *Digital Forense* (DFRWS) é:

"A utilização de métodos cientificamente comprovados e voltados à preparação, preservação, coleta, validação,

identificação, correlação, análise, interpretação, documentação e apresentação de vestígios digitais derivados de fontes digitais com o propósito de facilitar ou possibilitar a reconstituição de eventos criminosos, ou auxiliar para identificar antecipadamente ações planejadas e não autorizadas." [PALMER, 2001].

A CF pode ser resumida como o processo de identificar, coletar, preservar, analisar e apresentar "vestígios digitais"¹⁷ relacionados com o computador de uma forma que seja legalmente aceita em um tribunal de justiça [MC KEMMISH, 1999; NOBLETT, 2000a; ROBBINS, 2000; YASINSAC, 2003; SLADE, 2004].

Entretanto, muitos pesquisadores e peritos sentem que uma definição mais precisa ainda não é possível porque vestígios digitais também podem ser recuperados de equipamentos que não são tradicionalmente considerados computadores, tais como, centrais telefônicas digitais, rádios transmissores digitais, câmeras (fotográficas, filmadoras, de monitoramento) e aparelhos de sistema de posicionamento global - GPS (*Global Positioning System*) [HARRINSON, 2002; HALL, 2005].

Conforme Cisar (2011), a Computação forense é direcionada principalmente em computadores, sendo considerada uma subárea da Digital Forense, ou seja, a Computação Forense é considerada menos abrangente do que a Digital Forense [CARRIER, 2005; PALMER, 2001; PIRES, 2004].

Assim, o termo Digital forense abrange investigações relacionadas às áreas de Computação, Informática e Telecomunicações.

O termo Investigação Digital (*Digital investigation*) é utilizado na literatura como sinônimo para CF e DF [CARRIER, 2005; 2006]. Embora possa seguir as mesmas etapas destas duas, a Investigação

¹⁷No documento original é utilizado o termo "*digital evidence*", referindo-se a qualquer informação de valor probatório que é armazenada ou transmitida em formato digital [SWGDE, 2011].

Digital (ID) não é realizada necessariamente para propósitos de âmbito criminal ou judicial, sendo voltada para finalidades de investigações particulares ou internas a uma corporação, como por exemplo, no caso de vazamento de informações confidenciais, espionagem, tentativas de invasão de sistemas de informática, etc. [NELSON, 2010].

2.1.3 Considerações Sobre as Terminologias Apresentadas

Os trabalhos realizados na área indicam uma convergência entre os procedimentos e técnicas adotadas para Computação forense, Digital forense e Investigação digital. No entendimento do autor e baseando-se nas definições encontradas na literatura, todas elas apresentam principalmente duas características em comum:

- ❖ Englobam conhecimentos multidisciplinares: Mineração de dados, Lógica, Probabilidade e Estatística, Processamento digital de Sinais e imagens, Criptografia, Redes de computadores, Direito, Telecomunicações, entre outras disciplinas.
- ❖ As três terminologias são utilizados com propósitos relacionados à condução de investigações envolvendo assuntos de informática.

Neste trabalho será adotado o termo Investigação digital forense (IDF), que será utilizado de maneira ampla para referir-se genericamente às investigações realizadas na área de informática com propósitos judiciais.

2.1.4 Vestígio e Indício

A palavra vestígio¹⁸ é definida como sinal deixado pela pisada ou passagem, tanto do homem como de qualquer outro animal; pegada, rasto; Indício ou sinal de coisa que sucedeu, resquícios [FARIA, 1962]. Para Espíndula (2009), quando os peritos examinam um local de crime, procuram todos os tipos de objetos, marcas ou sinais sensíveis que possam ter relação com o fato investigado. Todos esses elementos, individualmente, são denominados de vestígios.

A palavra Indício é definida no Artigo 239 do Código de Processo Penal (CPP):

“Considera-se indício a circunstância conhecida e provada, que, tendo relação com o fato, autorize, por indução, concluir-se a existência de outra ou outras circunstâncias.” [BRASIL, 1941].

Na literatura internacional a palavra evidência (*evidence*) é amplamente utilizada pelas áreas de Ciência forense e Jurídica, indistintamente como sinônimo para vestígio e indício. Entretanto, esta palavra não é definida na legislação jurídica vigente no Brasil. Por este motivo, no presente trabalho, optou-se pelo uso da palavra vestígio, que é utilizada como referência a qualquer dado encontrado em DAD que possa ter relação com o fato investigado.

2.1.5 Prova Pericial

A prova pericial ocorre todas as vezes que o juiz não se considera suficientemente apto para realizar a verificação dos fatos, seja pela ausência de conhecimentos técnicos ou pela impossibilidade de coletar os dados necessários, quando o trabalho será realizado por pessoas

¹⁸ Do latim, *vestígiu* [FARIA, 1962].

entendidas e com percepção técnica e científica na matéria, por meio de perícia técnica. Nessas ocasiões, é de extrema importância que os vestígios sejam analisados para determinação dos fatos, não pairando quaisquer dúvidas sobre as conclusões obtidas [NETO, 1998].

A prova pericial é dentre as provas produzidas na persecução penal, a que mais norteia a decisão dos juristas, devido seu poder de convencimento estar amparado em características como imparcialidade e embasamento científico.

2.1.5.1 Perícia Judicial

Para Alberto (1996), Perícia é um instrumento especial de constatação, prova, demonstração científica ou técnica da veracidade de situações, coisas ou fatos. Perícia pode ser entendida como sendo um meio de prova admitida no processo, com o propósito de levar ao Juiz elementos relativos a fatos que necessitam de conhecimentos técnicos específicos, podendo consistir numa declaração de ciência, na afirmação de um juízo, ou em ambas, simultaneamente.

A perícia judicial é aquela realizada dentro dos procedimentos processuais do Poder Judiciário, por determinação, requerimento ou necessidade de seus agentes ativos, e se processa segundo regras legais específicas. A perícia judicial pode ser utilizada como prova ou como arbitramento.

- Perícia judicial é utilizada com objetivo de prova: quando tem por escopo trazer a verdade, demonstrável técnica ou cientificamente, para subsidiar a formação da convicção do julgador.
- Perícia judicial é utilizada com objetivo de arbitramento: quando determinada no processo de liquidação de sentença tem por objetivo quantificar mediante critério técnico.

2.1.6 Quesitos

Para a realização da perícia, um cuidado importante está baseado à formulação de quesitos, que consistem das perguntas que as partes formulam ao perito e pelas quais delimitam o escopo da perícia. Essas perguntas são relativas aos fatos que constituem o objeto da perícia.

Em determinadas circunstâncias o próprio juiz, tendo em vista a natureza da infração, já determina a priori quais as indagações a serem respondidas pelo perito. Os casos estão contidos nos artigos 171 e 173 do Código de Processo Penal [BRASIL, 1941].

Os quesitos podem ser classificados como principais, suplementares e impertinentes.

- Quesitos principais: São formulados no primeiro momento legalmente oportuno, apresentadas pelas partes ao perito judicial, através de petição endereçada ao juiz da causa, ou apresentadas pelo próprio juiz.
- Quesitos suplementares: No decorrer da realização da perícia, ou seja, durante a diligência, as partes podem elaborar quesitos adicionais, àqueles já apresentados como principais. Os quesitos suplementares após deferidos pelo juiz deverão ser respondidos pelo perito (CPC., art. 425).
- Quesitos impertinentes: Podem ser interpretados como àqueles que extrapolam ao campo da perícia que está sendo realizada, ou não guardam relação com o objeto da perícia. Também se enquadram nesta classificação, quesitos que requerem ao perito interpretar a legislação, dizendo se ela está correta ou incorreta. Caso estes não sejam indeferidos pelo juiz, o próprio perito pode se recusar a respondê-los.

Na área de Investigação Digital Forense, alguns termos são inadequadamente utilizados para determinados contextos. Por se tratar de uma área de conhecimento relativamente recente, quando comparada com outras áreas do conhecimento, não há ontologias

(vocabulários) padronizadas e a definição dos termos utilizados ainda é escassa. Sendo assim, uma padronização de vocabulário de termos dedicados precisa ser desenvolvida [GAYED, 2014].

2.2 Processo de Investigação Digital Forense

O processo de investigação digital forense - IDF deve preservar a integridade dos dados originais, assegurando que estes não sofram alterações durante a etapa de coleta e isto se aplica igualmente aos vestígios obtidos no decorrer do processo investigativo.

Considerando-se o grau de volatilidade dos vestígios digitais, isto pode ser uma tarefa meticulosa. A simples ação de ligar ou desligar um computador pode alterar ou destruir os vestígios de interesse definitivamente. Por essa razão, é de extrema relevância que o processo investigativo seja conduzido de forma metódica e organizado [CASEY, 2004; REIS, 2002].

Estas características são fundamentais para evitar erros durante o processo investigativo, para garantir que as melhores técnicas disponíveis sejam utilizadas e para aumentar a chance de que diferentes peritos cheguem aos mesmos resultados ao examinarem os mesmos dados [REIS, 2002].

Os procedimentos seguidos e os protocolos de análise forense devem ser detalhados, documentados, revisados e aceitos pela comunidade científica da área, coerentes com os princípios legais e técnicos [NOBLETT, 2000; REIS, 2002]. Exemplo disso são os procedimentos definidos no *Standard Operating Procedures* (SOP) e sugeridos pelo SWGDE - Grupo de Trabalho Científico em Evidências Digitais [SWGE, 2000] e também pela Associação de Oficiais Chefes de Polícia do Reino Unido [ACPO, 1999].

2.2.1 Etapas da Investigação Digital Forense

Diversos modelos de estruturas para IDF têm sido propostos na literatura [ADEMU, 2011; AGARWAL, 2011; ASHCROFT, 2001; BEEBE, 2005; CASEY, 2004; CARRIER, 2004B; CIARDHUÁIN, 2004; CISAR, 2011; DITTRICH, 2000; KALBANDE, 2013; KRUSE, 2003; MANDIA, 2003; MCKEMMISH, 1999; NELSON, 2004; PALMER, 2001; PILLI, 2010; POLLITT, 1995; REITH, 2002; RUIBIN, 2005], apenas para mencionar alguns.

Cada modelo de estrutura proposto é dividido em etapas. O que muda entre eles é a quantidade de etapas, o nível de detalhamento e a sequência das etapas a serem seguidas pelo perito. Na Figura 1.2 foi ilustrado o exemplo de um modelo com as etapas consideradas tradicionais, baseado nos modelos de [CISAR, 2011; PALMER, 2001; PILLI, 2010; RUIBIN, 2005], que será utilizado como base de referência para este trabalho.

As etapas são definidas em linhas gerais, permitindo que o perito possa utilizá-las como um roteiro a ser seguido para a grande maioria das investigações periciais, possibilitando identificar as principais atividades e pontos críticos relacionados com cada fase.

2.2.1.1 Etapa de Preparação

Nesta etapa deve ser reunida a maior quantidade de informações possíveis sobre o ambiente a ser periciado, tais como, quantidade de equipamentos a serem periciados e que terão os DADs duplicados, sistemas operacionais utilizados, versão, etc. Deve-se identificar detalhes necessários que podem ser peculiares para a investigação, considerando as condições iniciais encontradas no local da perícia.

Envolve todo o estudo preliminar do ambiente em que o trabalho será realizado, os procedimentos e estratégias que serão adotadas, ferramentas de software necessárias, hardware e a respectiva capacidade necessária para o armazenamento dos dados que serão

coletados. Identificação de DADs e interfaces a serem utilizadas na etapa de coleta de dados.

No que tange ao conjunto de ferramentas de software a serem utilizadas, estas devem ser confiáveis e não intrusivas, previamente testadas, que possam ser executadas a partir de algum tipo de mídia protegida contra escrita e possibilite a leitura e coleta dos dados na máquina suspeita de conter os vestígios de interesse para a perícia.

Na etapa de preparação, normalmente já se tem definido (porém não existe uma regra ou obrigatoriedade) o que será procurado nos DADs objetos da investigação na etapa de análise. Geralmente, até mesmo antes desta etapa, se define o que será procurado, como por exemplo, um conjunto de PChs.

2.2.1.2 Etapa de Preservação

A essência desta etapa é proteger a integridade dos dados originais, garantindo que não sofram qualquer tipo de alteração. O *hash*¹⁹ do dispositivo de armazenamentos de dados de origem (DAD_o) que contém os dados que serão coletados deve ser gerado antes do início da duplicação, para posterior comparação com o *hash* da imagem gerada que contém os dados coletados, garantindo assim que originais e cópias sejam idênticos. Não devem ser utilizados programas contidos no computador sob investigação, nem escrever ou gravar nada no respectivo DAD. Em alguns casos o perito pode se deparar com o computador ainda em operação, por exemplo, o suspeito realizando o envio de um software pirateado via mensagem de email, executando *download/upload* de algum dado ou gravando simultaneamente diversas cópias sem a respectiva licença. Em uma situação como essa, o perito precisa decidir entre manter o computador em funcionamento

¹⁹Neste contexto, refere-se a uma sequência de bits gerada por um algoritmo de forma a identificar a imagem gerada unicamente.

ou desligá-lo, tendo sempre em mente a coleta da maior quantidade de vestígios possíveis [EC-COUNCIL, 2010].

2.2.1.3 Etapa de Coleta

Envolve o procedimento para a aquisição do conteúdo dos dados contidos nos DADs_O dos computadores suspeitos. Nessa etapa é realizada a duplicação *byte-a-byte* do DAD_O. Esta duplicação pode ser realizada espelhando o conteúdo para outro DAD_D (Dispositivo de trabalho) ou para um arquivo de imagem. Todos os dados coletados devem ser documentados, mantendo a cadeia de custódia²⁰. Existem dois grandes cuidados que devem ser tomados ao se coletar material para análise: perda e alteração.

Caso as devidas precauções não sejam tomadas, dados importantes podem ser sobrescritos e perdidos em sua totalidade ou parcialmente, alterando seu significado ou apagando informações críticas. Além disso, a coleta conduzida com procedimentos inadequados²¹ como exemplo, o uso de ferramentas intrusivas, pode comprometer completamente a investigação pericial, e consequentemente os resultados sejam refutados e invalidados [REIS, 2002; STEPHENSON, 2000].

Podem ser utilizados diversos mecanismos para a coleta, envio e armazenamento do material coletado. Os dados coletados podem ser enviados para uma estação forense, para um disco especialmente preparado (esterilizado) para esta finalidade, conectado diretamente ao computador suspeito ou através da rede. Também faz parte desta etapa

²⁰ A cadeia de custódia permite manter e documentar cronologicamente os eventos de manipulação e transporte da evidência [SMITH, 1990].

²¹ Que não segue as práticas adotadas pela comunidade científica e seja juridicamente aceita.

a apreensão de todo material²² que o perito julgue necessário e que possa ter alguma relação direta ou indireta com a investigação.

2.2.1.4 Etapa de Análise

Uma das tarefas habituais de um perito de informática é procurar em DAD por vestígios de interesse. Ele pode focar a procura em vestígios específicos, como por exemplo, palavras-chave, que funcionarão como indicadores de uma atividade suspeita. Esta fase consiste no objetivo principal do processo investigativo pericial, pois é o elemento caracterizador da perícia.

Primeiramente, pode ser realizado um pré-processamento automatizado dos dados com o objetivo de extrair características de interesse para identificação de dispositivos com maior relevância, sendo muitas vezes, necessária a aplicação de técnicas de recuperação de arquivos deletados, reconstituição de fragmentos e quebra de senhas em arquivos cifrados. Para isso, podem ser utilizadas ferramentas de análise de dados, técnicas de mineração de dados (*data mining*), ferramentas de “força-bruta”, entre outras.

Posteriormente, os dados coletados podem ser examinados à procura de possíveis vestígios ou indicadores que possam estar relacionados com a investigação pericial em curso. Exemplo de dados alvo para esta etapa incluem emails e arquivos anexados, identificação e análise de softwares e aplicativos instalados autorizados e não autorizados, funcionalidades de softwares, procura por PChs, de arquivos do sistema, histórico do navegador, arquivos de registros (*logs*) de computadores específicos e da rede e arquivos cifrados. Todas as

²²Neste contexto a palavra é utilizada em sentido amplo, designando tudo que pode ser coletado, seja fisicamente tangível ou não, como por exemplo, informações digitais, dvd's, cd, hardware, documentos, configurações para conexões externas, anotações e *post-its*.

tarefas realizadas nesta etapa precisam ser devidamente documentadas e descritas no documento final (Laudo pericial ou Relatório técnico).

2.2.1.5 Etapa de Correlação

Envolve a comparação ou a busca do estabelecimento de relação dos vestígios encontrados com fatos, arquivos ou outros vestígios previamente obtidos, de modo a encontrar alguma correspondência entre os mesmos, permitindo dessa forma, tirar conclusões a cerca da investigação.

Esta etapa pode ter diversas iterações (tantas quanto necessárias) com a etapa de Análise, sempre que houver a necessidade de estabelecer relacionamento entre os vestígios e o material encontrado, ou mesmo entre os vestígios.

2.2.1.6 Etapa de Apresentação dos Resultados

Os resultados das etapas anteriores e demais observações são apresentados em um documento de laudo pericial utilizando-se de linguagem clara e que não seja estritamente técnica, entendível por pessoas interessadas, mas que não sejam necessariamente da área técnica, por exemplo, juízes e advogados. O principal propósito do laudo pericial é servir de base para amparar a ação judicial.

2.3 Investigação Envolvendo Grande Quantidade de Computadores

Em determinadas circunstâncias para realização de investigação digital forense, não é possível o perito conhecer previamente o cenário em que a diligência será realizada, por exemplo, em medidas cautelares com produção antecipada de provas, sendo possível se deparar com

ambientes de redes computacionais compostas por um grande conjunto de computadores e seus respectivos dispositivos de armazenamento. Dentre as possibilidades a serem seguidas pelo perito na hipótese de não ter absoluta certeza em qual computador contém os vestígios de interesse, estão:

- I. Duplicar todos os DADs ou gerar imagens na etapa de coleta. Posteriormente em laboratório na etapa de análise, examinar aleatoriamente (sem uma ordem pré-definida) cada um dos DADs;
- II. Analisar todos os DADs e duplicar somente os que foram encontrados vestígios de interesse, na etapa de coleta. Posteriormente somente estes DADs são examinados na etapa de análise;
- III. Analisar todos os DADs e copiar somente dados parciais dos DADs que foram encontrados vestígios de interesse.
- IV. Escolher aleatoriamente ou por amostragem dentre todos os DADs encontrados uma parte que devem ser duplicados na etapa de coleta, e posteriormente examinados na etapa de análise.

2.3.1 Duplicar Todos os DADs Para Posterior Análise

Neste procedimento, na etapa de coleta, o perito utiliza ferramentas de software para realizar a duplicação *byte-a-byte* (também denominada clonagem ou imagem *raw*²³) de todos os DADs encontrados no local da diligência pericial [CANTRELL, 2012; JONES, 2009; SAMMES, 2000]. Esse procedimento tem a finalidade de evitar que os dados contidos nos DADs originais sofram qualquer tipo de

²³ Refere-se a dados (em formato binário) brutos, ainda não processados.

alteração, permanecendo preservados [BREZINSKI, 2002]. Além disso, este procedimento, segue as melhores práticas, sendo aceita e amplamente utilizada pela comunidade da área

Todo e qualquer exame é realizado posteriormente na etapa de análise, sobre o DAD duplicado, utilizando uma ou um conjunto de ferramentas de software. Atualmente podem ser encontradas ferramentas com funcionalidades para duplicação do DAD e análise, porém, separadamente, ou seja, não realizam ambas as funcionalidades simultaneamente [ABDULLAH, 2008; CANTRELL, 2012; GRILLO, 2009; JONES, 2009; KORNBLUM, 2002; KORNBLUM, 2004; REIS, 2002].

2.3.2 Analisar Antes de Coletar e Duplicar os DADs Parcialmente

Outra possibilidade a ser seguida pelo perito consiste em primeiramente analisar cada um dos DADs encontrados em modo somente de leitura (sem permissão de escrita) buscando por vestígios de interesse [HALBOOB, 2014]. Ao final da análise, e baseado nos resultados alcançados, considerar se o DAD deve ser duplicado. Observa-se que dessa forma, somente os DADs que foram duplicados serão examinados posteriormente na etapa de análise.

2.3.3 Analisar Antes e Coletar Somente Informações Parciais

Neste procedimento, semelhante ao item anterior, primeiramente é realizada a análise em cada um dos DADs encontrados em modo somente de leitura (sem permissão de escrita) buscando por vestígios de interesse. A diferença é que uma vez encontrados os vestígios de interesse, somente esses vestígios são coletados.

2.3.4 Escolher Aleatoriamente

Além dos procedimentos já descritos, outra possibilidade é escolher ao acaso dentre o total de DADs, somente uma amostra para ser duplicada na etapa de coleta e posteriormente examinados na etapa de análise.

Para os procedimentos descritos nos itens II, III e IV, o perito pode vir a se deparar com um problema. Considerando-se que eventualmente por algum motivo, como por exemplo, o esclarecimento de uma possível dúvida ou questionamento adicional, ou até mesmo alguma constatação ou análise por outro perito (assistente técnico), houver a necessidade de examinar os DADs que não foram duplicados. Neste caso não seria possível coletar dados adicionais ou realizar uma nova perícia, já que os DADs originais podem não terem sido preservados, e também não houve a duplicação dos DADs.

2.3.5 Considerações Sobre IDF Envolvendo Grande Quantidade de Computadores

Conforme foi visto neste capítulo, alguns procedimentos podem ser seguidos para investigações em cenários que envolvem uma grande quantidade de computadores. Acredita-se que cada procedimento possua seus pontos favoráveis e pontos desfavoráveis.

A escolha do procedimento geralmente fica condicionada a experiência e responsabilidade do perito, visto que a realização do procedimento de coletar tudo o que encontra no momento da perícia, como por exemplo a duplicação de todos os DADs, pode consistir em oportunidade única para que o mesmo seja conduzido. Dependendo do cenário em que a IDF é realizada, se um vestígio importante deixar de ser coletado neste momento, é possível que nunca mais possa ser coletado, considerando a possibilidade de que um DAD seja propositalmente danificado ou extraviado.

No presente trabalho, o foco está no procedimento descrito na Seção 2.4.1. O autor defende que a duplicação de todos os DADs seja a melhor maneira para evitar problemas como o que foi exposto no parágrafo anterior. Portanto, esse procedimento será adotado no decorrer desta tese.

2.4 Trabalhos Correlatos

Esta seção apresenta o levantamento bibliográfico dos trabalhos encontrados durante a revisão da literatura que tratam das técnicas de Triagem e procura por palavras-chave²⁴ aplicadas à investigação digital Forense - IDF. Estes trabalhos serviram de base para investigação e identificação do estado da arte nesta área.

Ao proceder à revisão bibliográfica foram consultadas diversas bases de dados acadêmicas e os periódicos elencados a seguir:

- ACM Digital Library;
- Digital Investigation (The International Journal of Digital Forensics & Incident Response);
- Google Scholar;
- IEEE Xplore;
- IEEE Transactions on Information Forensics and Security;
- IJOFCS (International Journal on Forensic Computer Science);
- International Journal of Digital Crime and Forensics;
- International Journal of Digital Evidence;
- Journal of Digital Forensic Practice;
- Journal of Digital Forensics, Security and Law;
- Journal of Digital Investigation;
- Science Direct;
- Scopus;
- The Journal of Digital Investigations;

²⁴ Na literatura é utilizado o termo *string-search*.

Foram buscadas publicações no período compreendido entre os anos de 1999 a Agosto de 2015, utilizando como palavras-chave os termos: *computer forensic dataset reduction*, *computer forensic volume problem*, *computer forensic string search*, *computer forensics triage*, *digital forensic challenges*, *digital forensic data volume*, *digital forensic growth*, *digital forensic string search hits*, *digital forensic triage*, *digital forensics keyword searches*, *digital investigation intelligence*, *digital investigation triage*, *digital investigation string search*, *forensic data mining*, *forensic data reduction*, *raw data string search* e *scan raw data content*.

Os trabalhos encontrados foram analisados com o intuito de identificar as principais características dos mesmos, qual o método utilizado, além de aspectos não abordados e que poderiam ser pontos de possíveis explorações.

Diversos aspectos têm aumentado os desafios para Computação Forense. Fatores como o aumento da quantidade de computadores e outros equipamentos congêneres que possuem DAD, a ubiquidade, o aumento da capacidade dos DADs, conseqüentemente o volume de dados e de investigações envolvendo a necessidade de produção de provas por meio de Computação Forense em ambientes com elevado número de computadores estão agravando o trabalho de investigação digital [GARFINKEL, 2010]. Concomitantemente, observa-se:

- a) Um aumento no número de equipamentos apreendidos por caso.
- b) O número de casos que envolvem vestígios digitais é crescente (Estima-se que mais de 85% dos crimes cometidos atualmente deixam rastros de vestígios digitais) [RAY, 2005].
- c) A quantidade de dados armazenados por cada indivíduo está aumentando.

A pressão em obter informações rapidamente a partir de dados contidos em DADs está alcançando níveis críticos, tornando as técnicas

e ferramentas utilizadas tradicionalmente insustentáveis [GOGOLIN, 2010; PARSONAGE, 2009]. Essas ferramentas e técnicas estão começando a enfrentar problemas de escalabilidade, principalmente para a análise de grandes conjuntos de dispositivos e elevado volume de dados. A quantidade de equipamentos e o volume de dados a serem analisados vêm aumentando mais rápido do que a capacidade das ferramentas podem processar e gerenciar [ROUSSEV, 2013]. Isso justifica a quantidade de chamados para trabalhos científicos tratando do problema de perícia forense em grandes conjuntos de computadores e elevado volume de dados [RICHARD, 2006; WILES et al., 2007; CASEY, 2013].

Cada vez mais os peritos e investigadores da área de investigação digital forense enfrentarão problemas com conjuntos de dispositivos envolvendo grande volume de dados, além do custo envolvido para a realização destas análises periciais. É nítida a necessidade de diminuição do tempo e esforço humano gasto desnecessariamente com a análise de DADs que não apresentam dados de interesse para uma determinada investigação. Esses fatores implicam em acúmulo de trabalho para os peritos e investigadores, causando sérias consequências, por exemplo, atrasos na obtenção dos resultados, aumento no tempo de espera para condenação de suspeitos e absolvição de inocentes [SHAW, 2013].

Neste sentido, diferentes abordagens têm sido propostas na literatura, tendo em vista o desenvolvimento e aprimoramento de técnicas e ferramentas que proporcionem aumentar a agilidade para a identificação de DADs com maior chance de conter vestígio de interesse para IDF.

Duas abordagens que têm se destacado são a utilização de técnicas para procura por palavras-chave (*string-search*) e triagem.

2.4.1 Técnicas de Triage

Diversos autores definem o termo triagem no contexto da Computação Forense.

Os trabalhos encontrados na literatura propõem soluções por meio da utilização de técnicas de triagem, conforme será visto a seguir. Alguns são baseados em determinar DADs que podem ser excluídos da análise e outros focam na priorização de dados que devem ser coletados. No contexto deste trabalho, o foco está em processos de triagem aplicados na identificação de DADs mais prováveis de conter vestígios de interesse para a IDF por meio de uma busca por palavras-chave em dados não estruturados por um sistema de arquivos.

ROGERS et al. (2006), apresentam um dos trabalhos precursores que trata da técnica de triagem aplicada a investigação digital forense. Os autores salientam que em determinadas circunstâncias não é mais apropriada a apreensão do equipamento, transportá-lo para um laboratório, para então iniciar a procura por informações de interesse, vasculhando profundamente em todo o sistema, para talvez encontrar algum vestígio que possa ser útil para a investigação. Em determinadas investigações as respostas devem ser obtidas o mais rápido possível, em questão de horas, em oposição a dias. Por exemplo, em investigações que envolvem sequestro e pedofilia, nos quais a preservação da vida pode estar em risco, obter respostas em um curto intervalo de tempo, nas primeiras horas da investigação é essencial. As informações obtidas podem ser confrontadas durante o depoimento do suspeito.

O objetivo do trabalho é realizar rapidamente a identificação dos DADs com maior potencial de conter vestígios de interesse para a investigação, dentro de um curto intervalo de tempo, no local onde a diligência de investigação é realizada. Os autores propõe *Cyber Forensic Field Triage Process Model* - CFFTPM, visando à obtenção de respostas já nos momentos iniciais da investigação, no decorrer das etapas de preparação, preservação e coleta. Na abordagem adotada, é realizada a coleta parcial de dados contidos no perfil do usuário (diretório *Home*,

propriedades dos arquivos e no *Registry*), nas atividades da Internet (histórico do navegador, emails e ferramentas de comunicação instantânea), nos softwares aplicativos instalados e arquivos acessados. Esses dados são analisados e os DADs são ranqueados em termos de importância ou prioridade, conforme a quantidade de informações relevantes encontradas.

Embora o propósito do modelo seja fornecer a identificação, análise e interpretação dos vestígios encontrados no local da investigação sem a necessidade de levar o equipamento ou o DAD diretamente para o laboratório, os autores não descartam a necessidade de que após a triagem inicial em campo seja concluída, o DAD seja levado para um laboratório e submetido à análise mais aprofundada.

O CFFTPM tem sido utilizado com sucesso em casos reais. Além disso, os vestígios e resultados obtidos nesses casos não foram contestados judicialmente.

No trabalho, embora utilizem a técnica de triagem, não é informada de que maneira as informações são procuradas, se ocorre a duplicação do DAD antes da análise e nem são apresentados os resultados dos experimentos realizados.

ROUSSEV e QUATES (2012) propõem um processo de triagem de conteúdo que é realizado subsequente ao processo de aquisição de imagens da mídia denominado de Imagem Forense, usando o resumo de similaridade e metodologias de correlação de *hash*. Eles aplicam o processo no estudo do caso M57 para demonstrar a aplicação do método proposto em um conjunto de DADs.

Os autores afirmam que a geração atual de ambientes forenses integrados como *EnCase* e *FTK* são fundamentalmente inadequados para o trabalho de triagem, pois a abordagem empregada nas ferramentas de “examinação forense” profunda possui baixo desempenho. A melhor escolha seria focar sobre o uso de resumo de similaridade para identificar as correlações entre diversas fontes e

estabelecer um quadro inicial de entendimento para os casos estudados.

Os autores identificam priorização baseados em metadados como uma opção para melhorar o desempenho ao invés de obter vantagem de desempenho por meio do acesso sequencial para o dispositivo de armazenamento físico. A técnica empregada utiliza filtros de *Bloom* para obter a similaridade baseada no conteúdo do arquivo.

Porém esta abordagem mostra-se eficiente quando o objeto de busca é conhecido, pois não há como obter o resumo da similaridade de um vestígio desconhecido, que pode resultar na grande quantidade de falsos positivos.

BERTÈ et al. (2012) chamam a atenção para o aumento do número de crimes praticados envolvendo o uso de computadores. Posteriormente salientam da crescente dificuldade de analisar a quantidade de DADs apreendidos pela polícia italiana. Então propõe uma nova abordagem para perícia computacional baseada na mineração de dados e teoria da gestão do conhecimento para realizar triagem dos dados coletados. O método baseia-se, primeiramente em espelhar o disco e depois em agrupar determinados tipos de arquivos de interesse que dependem do objetivo da investigação, *logs* (registros) de acessos encontrados, *logs* de eventos do sistema, histórico de navegadores para encontrar *sites* mais visitados, então sumarizam tudo em uma matriz, denominada de matriz completa e finalmente analisam os hábitos do usuário e a probabilidade de um computador ter sido utilizado para cometer o crime investigado.

MARTURANA et al. (2012) ratificam a necessidade de desenvolvimento de novas ideias e métodos para tratar a difícil tarefa de analisar grandes volumes de massas de dados. Destacam que o problema também é enfrentado pela polícia italiana em diversos casos. Então propõem um modelo para ser aplicado em investigações de crimes de violação de direitos autorais. No modelo apresentam uma forma de automatizar a análise de provas para ser aplicado na triagem de vestígios “*post-mortem*”. Para isso, utiliza-se de técnicas baseadas em

aprendizagem de máquina (*Machine Learning*) e mineração de dados (*Data Mining*) para classificar os DADs apreendidos baseado no conteúdo, isto é, nos tipos de arquivos encontrados e histórico de navegador e gera uma estatística desse conteúdo, antes de serem analisados em detalhe. O modelo utiliza o conceito de classe para associar os tipos de arquivos à variáveis.

Dependendo do enfoque do crime de violação de direitos autorais investigado é realizada uma associação com determinados tipos de arquivos. Dessa forma, o foco é direcionado para as unidades de armazenamento que apresentaram maior incidência dos tipos de arquivos que podem ter maior relação com a investigação.

BEEBE e LISHU (2014) apresentam um método de classificação para fazer busca textual para investigação digital forense baseado na técnica de máquina de vetores de suporte (*support vector machine* - SVM) que consiste em um modelo de aprendizado supervisionado que analisa os dados e reconhece padrões, usado para classificação e análise de regressão.

A ferramenta *Encase* foi utilizada para extração de características para buscar palavras-chave (PChs). O *SleuthKit* foi binarizado para extrair características adicionais dos arquivos alocados e *clusters* desalocados, e a ferramenta *Sceadan* foi utilizada para classificar tais *clusters*. A autora propôs dezoito características como indicadores quantitativos para medir a relevância das ocorrências.

Os autores propuseram a pesquisa baseada no comprimento da PCh buscada como uma métrica muito útil em algoritmos de classificação de ocorrências. O novo algoritmo de classificação foi chamado de algoritmo de classificação por relevância (*Relevancy Ranking Algorithm*). O conjunto de dados M57 foram utilizados como conjunto teste.

Porém, apenas três das dezoito características (número de ocorrências da PCh por dispositivo, o tamanho e a prioridade da PCh buscada) podem ser utilizados como métrica para dados brutos, sem necessariamente interpretar os arquivos com ajuda do sistema de

arquivos. As métricas apresentadas no resultado são estatísticas particulares do método SVM binário, caso houvesse a apresentação de estatísticas como distribuição, média, desvio padrão e resultados intermediários como resultados dos treinamentos da SVM, possibilitaria uma comparação justa com outros modelos que não seguem o método SVM.

Contudo, este trabalho apresenta informações importantes sobre o conjunto de dados M57, tais como a lista de PChs a ser pesquisada e as funções que os usuários do cenário desempenhavam, e uma sugestão de prioridade por tipos de arquivos.

BAGGILI et al. (2014) apresentam uma abordagem voltada para respostas a incidentes. A ferramenta foi desenvolvida para triagem de dados em investigação digital forense, baseada em uma plataforma inicializável, denominada Forensics2020. A plataforma possui cinco fases: a primeira fase coleta metadados de todos os arquivos lógicos encontrados no disco rígido de um computador. A segunda fase coleta dados de arquivos de imagens. A terceira fase analisa e categoriza cada arquivo baseado no cabeçalho do mesmo. A quarta fase analisa cada arquivo executável para prover uma auditoria completa de softwares aplicativos no sistema investigado. Uma assinatura é gerada para cada arquivo executável, que posteriormente é verificada em uma base de dados de detecção de ameaças. A quinta fase gera o resumo (*hash*) de cada arquivo e grava em um dispositivo. Todas as cinco fases são realizadas em segundo plano, enquanto a equipe de segurança corporativa interage com o sistema sob suspeita de invasão.

A ferramenta realiza alterações no disco do sistema sob investigação, similarmente a outras ferramentas vendidas comercialmente. Não executa buscas por palavras-chave. Principalmente por estas duas características, não pode ser utilizada com finalidades de investigações de caráter judicial.

2.4.2 Resumo e Análise dos Trabalhos Relacionados

Conforme os trabalhos relacionados anteriormente, cujo as principais características estão sintetizadas na Tabela 2.1, é nítida a preocupação da comunidade científica da área com o aumento desenfreado da quantidade de DADs e o enorme volume de massa de dados que devem ser coletados e analisados pelos profissionais da área de perícia em computação forense.

Tabela 2.1 - Síntese das técnicas e características dos trabalhos relacionados

Autor	Etapa	Tipo de Coleta	Técnica utilizada			Pró	Contra
			Varredura	Triagem	Ranqueamento		
ROGERS et al. (2006)	Análise	Parcial	Sim, porém não apresenta a técnica utilizada.	Sim	Sim	Prioridade baseada na triagem. Tenta reduzir a quantidade de informações a ser analisada.	Não explica como as informações são procuradas e não apresenta resultados dos experimentos.
ROUSSEV e QUATES (2012)	Análise	Total	Sim, utilizando similaridade de <i>hash</i> de arquivos.	Sim	Não	Tenta reduzir o tempo. Utiliza um arquivo do caso M57-Patents.	Tem que ter a assinatura de <i>hash</i> dos arquivos de interesse.
BERTÈ et al. (2012)	Coleta/Análise	Total	Sim, utilizando mineração de dados.	Sim	Não	Tenta reduzir o tempo. Tenta reduzir a quantidade de informações a ser analisada.	Tempo necessário para treinamento das redes neurais artificiais.
MARTURANA et al. (2012)	Análise	Total	Sim, utilizando similaridade de <i>hash</i> de arquivos.	Sim	Sim	Realiza a classificação dos dados antes de analisar, priorizando os que apresentam maior chance de conter os vestígios.	Tempo necessário para realização do método.
BEEBE e LISHU (2014)	Análise	Total	Sim, por meio de busca textual utilizando máquinas de vetores de suporte (SVM).	Sim	Sim	Propõe um algoritmo de classificação de PChs de interesse para a investigação baseando-se na relevância das PChs.	Tempo necessário para treinamento das redes neurais artificiais utilizadas. Não apresenta resultados intermediários.
BAGGILI et al. (2014)	Coleta/Análise	Total	Sim	Sim	Não	Utiliza técnicas de triagem.	Altera informações originais do DAD. Necessita do <i>hash</i> dos arquivos originais para realizar a comparação.

Também é perceptível a necessidade de desenvolver novas metodologias e ferramentas que possam ser aplicadas para auxiliar no trabalho cotidiano dos profissionais que atuam nessa área no sentido de identificar dispositivos com maiores chances de conter os vestígios de interesse para uma dada investigação.

A aplicação de técnicas baseadas em triagem indica uma tendência que pode ser uma alternativa para auxiliar em diversos casos que envolvem coleta e análise em cenários com elevado número de DADs, que por sua vez pode implicar em grande volume de massa de dados.

As publicações encontradas permitiram o conhecimento dos trabalhos que mais se aproximam com o objetivo desta tese, ratificando os problemas e dificuldades para a realização perícias de informática que envolvem a necessidade de analisar elevada quantidade de dispositivos, o aumento da capacidade de armazenamento e do crescente volume de dados e formas de como priorizar os dispositivos que devem ser analisados. Além disso, contemplam técnicas e ferramentas, considerando a diminuição de esforços e tempo necessário para busca, identificação e extração de vestígios relevantes para perícia de informática.

O que pode ser denotado dos trabalhos correlacionados é a busca por métodos para diminuir o esforço da IDF. As técnicas apresentadas focam os esforços para a etapa de análise. Nos respectivos trabalhos não foram constatadas a utilização de tentativas para busca por palavras-chave já na etapa de coleta de dados, durante a duplicação do DAD, aspectos que são abordados no presente trabalho. Além do ranqueamento para indicação dos DADs com maior chance de conter os vestígios de interesse e que devem ser prioritariamente examinados na etapa de análise.

3 MÉTODO PARA TRIAGEM E RANQUEAMENTO DE COMPUTADORES

Este capítulo apresenta e detalha o método proposto neste trabalho. A seguir, são descritos os processos que o integram, as entradas e saídas de cada processo, tipos de resultados obtidos e de que maneira esses resultados são utilizados.

3.1 Introdução

Em determinadas circunstâncias a Investigação digital forense - IDF deve ser realizada em ambientes compostos por diversos computadores e outros equipamentos cujos dados contidos em seus respectivos dispositivos de armazenamento de dados (DADs) devem ser coletados para posterior análise. Essa análise normalmente se inicia com o exame dos dados coletados, por meio da busca por palavras-chave - PChs.

Essas PChs normalmente estão relacionadas ao contexto da investigação realizada. A escolha dessas PChs e sua relevância fica a critério do perito, responsável pela condução da investigação. Entre os exemplos de PChs estão: nomes de pessoas (física ou jurídica), número do cadastro de pessoas físicas (CPF), números de telefones, números de cartões de crédito, endereços de emails, números de endereço IP - *Internet Protocol* e outros termos específicos para a investigação.

Tendo em vista a necessidade de duplicar os DADs *byte-a-byte* e que durante essa duplicação o perito fica praticamente impossibilitado de interagir com os dados que estão sendo duplicados, esta tese propõe um método que a partir de um conjunto de computadores, cujo conteúdo dos respectivos DADs serão investigados, realiza uma sequência de procedimentos na etapa de coleta, que independentemente do sistema de arquivos, simultaneamente à duplicação do DAD, procura sequencialmente por uma lista de PChs de interesse para IDF.

Este resultado é utilizado para a realização de uma triagem baseada na quantidade de ocorrências de PChs e mediante critérios de importância para a investigação, indica quais DADs possuem maiores chances de conter PChs (vestígios) de interesse para a IDF. Os resultados oriundos da triagem são utilizados para realizar o ranqueamento dos dispositivos de armazenamento de dados de destino (DADs_D). Esse ranqueamento indica quais DADs_D deverão ser prioritariamente examinados na etapa de análise.

3.2 Considerações Iniciais Sobre o Método Proposto

Conforme Yang (2008), a atividade mais comum para um investigador é procurar vestígios de interesse em DADs. O investigador precisa focar em evidências específicas e indicadores-chave das atividades suspeitas, por exemplo, pesquisas por PChs específicas.

Em um primeiro momento pode parecer desnecessária a concepção de um método para procurar PChs em dados que estão sendo duplicados na etapa de coleta. Isto de fato poderia ser verdade se a quantidade de DADs que precisa ser duplicada e posteriormente analisada for pequena, composta por alguns DADs com baixa capacidade de armazenamento de dados ou o tempo para realização da análise for ilimitado. Entretanto, se a quantidade de DADs e o respectivo volume de dados for grande, por exemplo, algumas dezenas de DADs com capacidade acima de 500 GB, a realização de uma análise “às cegas” pode ser uma tarefa árdua e demorada.

Assim, identificar os DADs que possuem maiores chances de conter os vestígios de interesse, pode ser útil ao perito em casos nos quais encontrar os vestígios nos primeiros DADs pode tornar desnecessária a continuidade da análise dos demais DADs ou ainda, poupar recursos humanos (homem-hora) e de tempo, deixando de analisar DADs que não possuem chances de conter vestígios de interesse para a investigação.

Antes do início da aplicação do método proposto, algumas condições iniciais devem ser satisfeitas. Assim, será assumido que:

- ❖ Todos os DADs suspeitos encontrados no local da perícia serão duplicados;
- ❖ Que o perito possui espaço disponível e suficiente para armazenar os dados duplicados provenientes dos DAD₀;
- ❖ Que o perito leu o Processo Judicial para entendimento dos objetivos da investigação pericial e tomou conhecimento dos quesitos que devem ser respondidos, possibilitando dessa maneira a definição de uma lista de PChs que serão

utilizadas no processo de busca durante a duplicação²⁵ e varredura do DAD_O para o DAD_D (imagem do DAD_O). É importante destacar que em função dos objetivos das investigação, o perito deve definir as PChs classificando-as em níveis de prioridade, de acordo com a relevância dessas PChs para a investigação.

A lista de palavras-chave (LpCh) deve ser definida em algum momento que antecede a etapa de coleta, visto que essa lista será utilizada como entrada para a etapa de coleta, conforme será tratado na Seção 3.3.

Vale salientar que o propósito do método proposto nesta tese é fornecer um indicativo dos DADs mais prováveis de conter os vestígios de interesse para uma determinada investigação e a sequência que esses DADs devem ser examinados, sem a pretensão de substituir o exame detalhado e minucioso que deve ser realizado na etapa de análise, possibilitando que o método seja utilizado de maneira complementar aos métodos tradicionalmente utilizados.

3.3 Visão Geral do Método Proposto

Esta seção descreve o método proposto de forma geral, sem aprofundamento nos detalhes, apresentando as partes que o compõem e de que forma o mesmo se insere nas etapas do modelo utilizado como referência. O intuito é descrever o que o método faz. O detalhamento dessas partes e de que forma são realizados, será apresentado na Seção 3.4.

²⁵ Embora no decorrer do texto seja mencionado o termo “duplicação” do DAD_O para o DAD_D, em determinadas circunstâncias pode ocorrer que os dados do DAD_O sejam enviados para um arquivo de imagem, como por exemplo, para um arquivo do tipo “.img” (ferramenta dd) ou “.E01” (ferramenta Encase).

Para cada DAD_O , com base na lista de PChs que serão procuradas, são executados paralelamente os processos de duplicação e varredura. Como saída para estes processos tem-se:

- a) A duplicação (*byte-a-byte*) do DAD_O para o DAD_D . A partir do DAD_D é gerado o respectivo valor de *hash*.
- b) Uma tabela que lista a quantidade de todas as ocorrências para cada PCh encontrada nos padrões de codificação (ASCII, UTF-8, UTF-16, UTF-32, ISO-8859-1 (LATIN-1)) em que a PCh foi encontrada.
- c) Uma tabela de Registro (*Log*) de localização que lista todas as ocorrências para cada PChs encontradas, o respectivo deslocamento²⁶ (*offset*) do DAD onde cada PCh foi encontrada e ainda os 40 caracteres que a circunda, sendo 20 *bytes* anteriores e 20 *bytes* posteriores à palavra-chave.

A partir da tabela de *Log* de localização das PChs descrita no item “c” é realizado o processo de filtragem. Esse processo classifica cada ocorrência de PCh relacionada na Tabela de *Log* de localização como verdadeiro-positivo (VP) ou falso-positivo (FP). A descrição do processo de filtragem e a definição dos termos VP e FP são apresentados posteriormente na Seção 3.4.2.

Após estes procedimentos, obtém-se a tabela de ocorrências filtradas para cada DAD, ou seja, a quantidade de ocorrências de VP. Este procedimento é repetido até que sejam obtidas as tabelas para todos os DADs duplicados.

Os resultados da quantidade de ocorrências de VP de cada DAD são organizados em uma tabela agregada. Ou seja, o resultado de cada DAD representa uma linha na tabela agregada. Baseado em um conjunto de critérios, de acordo com a existência ou não de PChs e as

²⁶ Um número relativo que indica a que distância a partir do primeiro *bit* do primeiro bloco do DAD a palavra-chave (PCh) está posicionada.

prioridades de cada PChs é realizado o processo de triagem. Os resultados da triagem são utilizados com entrada para o processo de ranqueamento. Os resultados desse processo é a sugestão da sequência em que os DADs_D devem ser prioritariamente examinados na etapa de análise.

Outro aspecto a ser observado é a sequência em que os processos e a geração dos resultados ocorrem no tempo, conforme mostra a Figura 3.2. A representação ilustrada permite entender os instantes em que o resultado de cada processo é obtido. Os resultados de determinados processos só podem ser obtidos após o término de processos anteriores. Dessa forma, têm-se os tempos (t) para obtenção dos resultados (saída) para cada processo. A soma destes tempos resulta no tempo total (T) para aplicação do método.

Figura 3.2 – Sequência em que os processos e resultados ocorrem no tempo

Os resultados dos processos que ocorrem nos tempos t_1 , t_2 e t_3 são obtidos no local da perícia, durante a duplicação do DAD. Enquanto que os resultados dos processos t_4 e t_5 não precisam necessariamente ser realizados no local da perícia, podendo ser

realizados posteriormente, por exemplo, em laboratório. Assim, no decorrer do tempo tem-se:

- t_1 : Execução dos processos de duplicação e varredura simultaneamente.
- t_2 : Obtenção da tabela com a lista de ocorrências e da tabela de *Log* de localização das PChs no DAD.
- t_3 : A tabela de *Log* de localização das PChs é utilizada como entrada para o processo de filtragem. O processo de filtragem resulta na Tabela de Log de ocorrências filtradas, que contém a quantidade de ocorrências filtradas com os valores de VP (verdadeiros positivos) e FP (falsos positivos). Simultaneamente é obtido o DAD_D duplicado. Em seguida é gerado o *hash* desse DAD_D .
- t_4 : Os resultados do processo de filtragem, que é a quantidade de ocorrências (valores de VP de todos os dispositivos) são utilizados para compor a tabela agregada. As informações dessa tabela e as informações de prioridade de PChs, definidas pelo perito (se houver) são utilizados no processo de triagem.
- t_5 : Os resultados do processo de triagem são utilizados para o ranqueamento, que indica a sequência que os $DADs_D$ devem ser examinados na etapa de análise.

Esta seção apresentou uma visão geral do método proposto neste trabalho. A explicação das características de cada processo que compõe o método são apresentados na Seção 3.4.

3.3.1 O Método Proposto e o Modelo de Referência Utilizado

A seguir, no fluxograma ilustrado na Figura 3.3 é mostrado de que forma o método proposto se insere nas etapas do modelo de referência utilizado neste trabalho. Também podem ser observadas entradas e saídas de cada etapa.

Para o exemplo do fluxograma ilustrado na Figura 4.3, considere-se que a lista de PChs foi definida em alguma das etapas que antecedem a etapa de coleta, ou seja, na etapa de preparação ou preservação. Os procedimentos das etapas de preparação e preservação seguem o modelo utilizado como referência apresentado na Figura 1.2. Uma sutil diferença que deve ser observada é que a lista de PChs passa a ser utilizada como "entrada" já na etapa de coleta. Enquanto no modelo de referência, geralmente as PChs são utilizadas somente na etapa de análise.

Figura 3.3 - Inserção do método proposto no modelo de referência

O foco principal do método proposto está direcionado para a etapa de coleta, porque é nesta etapa que o mesmo é aplicado, conforme destacado na Figura 3.3. Os resultados de saída do método fornecem as duplicações, ou seja, os DADs_D e o ranqueamento, que serve como “entrada” para a etapa de análise, informado a sequência que esses DADs_D devem ser analisados.

3.4 Características dos Processos que Compõem o Método Proposto

Esta seção descreve os processos que integram o método com explicações adicionais, apresentando o funcionamento dos processos que compõem o método, explicando as entradas e saídas de cada processo, como ocorre o fluxo de informações e como os resultados são utilizados. Observando a Figura 3.1, pode-se identificar os quatro processos principais:

- ❖ Duplicação e Varredura
 - ▶ Geração de *hash* do DAD_D
- ❖ Filtragem
- ❖ Triagem
- ❖ Ranqueamento

3.4.1 Processos de Duplicação e Varredura

Esta seção descreve os procedimentos para a realização dos processos de duplicação e varredura, que consistem nos processos iniciais do método proposto, conforme já introduzido na Seção 3.3. Apresenta detalhes complementares que visam ampliar as explicações para entendimento do funcionamento dos processos de duplicação e varredura.

Considerando-se que a velocidade para processamento (leitura e escrita) de dados em memória é maior do que a velocidade para leitura e escrita (gravação) em disco [TANENBAUM, 2010], buscou-se explorar o tempo gasto na etapa de coleta para realizar dois processos simultaneamente, duplicação do DAD e varredura para procurar por PChs.

Os dados utilizados como entrada para os processos de duplicação e varredura são os DADs de origem ($DADs_O$) e a tabela

contendo a lista de PChs que serão procuradas (L_{PCh}). Além disso, é necessário que o perito tenha providenciado:

- ❖ A lista (L_H) - é a relação com os valores de *hashes* dos $DADs_O$ a serem duplicados. Considerando que esta lista já tenha sido gerada na etapa de apresentação ou preservação. Esta lista pode ser utilizada para comparação dos valores do DAD original e do DAD duplicado.
- ❖ Um DAD_E (dispositivo de armazenamento de dados externo) a ser utilizado para armazenamento das listas (L_{PCh} , L_H , L_{QO}). L_{QO} é a listagem com a relação de quantidade de ocorrências de PChs que deve ser obtida para cada DAD.
- ❖ $DADs_D$ que receberão os dados a serem duplicados. Dependendo dos critérios adotados pelo perito, ele poderá realizar a duplicação (*byte-a-byte*) do DAD_O para um DAD_D ou criar um arquivo da imagem do DAD_O e armazenar em um DAD_D local ou remoto (na rede local ou em uma nuvem);
- ❖ Uma ou mais estações para executar os processos de duplicação e varredura. A utilização de maior quantidade de estações pode diminuir proporcionalmente o tempo necessário para duplicação e varredura de todos os $DADs_O$.

Os processos de duplicação e varredura são executados para cada DAD_O . Relembrando que estes processos ocorrem simultaneamente, estes processos iniciam e terminam ao mesmo tempo.

3.4.1.1 Carregamento da Lista de Palavras-chave

Em algum instante que antecede o início dos processos de duplicação e varredura, a lista de PChs deve ser carregada na memória

de acesso randômico – RAM (*Random Access Memory*) da estação utilizada **(A)**, conforme é destacado na área pontilhada na Figura 3.4.

Figura 3.4 – Fluxo de processamento dos dados em memória

- (A)** Alocação de memória para expansão e carregamento da lista de PCh na memória da estação utilizada.
- (B)** Alocação de memória para expansão e carregamento dos blocos do DAD_O na memória da estação utilizada.
- (C)** Varredura – Busca pelas palavras-chave.
- (D)** Threads são apagadas e a memória liberada (desalocada).
- (E)** Gravação dos blocos do DAD_D.

Após lida em algum dispositivo conectado à estação, a lista de PChs é carregada em memória e expandida com as possíveis combinações maiúsculo/minúsculo. Então, a partir da combinação de combinações de cada PCh, são geradas as respectivas codificações nos padrões ASCII, UTF-8, UTF-16, UTF-32 e ISO-8859-1 (LATIN-1), que é a lista de palavras-chave expandida (L_{PChe}), identificada na Figura 3.4 por **(C)**.

3.4.1.2 Processo de Duplicação

O processo de duplicação é realizado sobre dados “brutos” (*raw data*), ou seja, dados binários que não possuem uma estrutura de arquivo. Conforme é ilustrado na área pontilhada da Figura 3.5, no processo de duplicação, a partir do início (*offset 0*) do DAD_O, cada bloco é lido sequencialmente **(B)** e copiado na mesma sequência para uma área da memória RAM da estação utilizada.

Figura 3.5 – Processo de duplicação

Estes blocos são enfileirados até no máximo da metade da quantidade de memória RAM disponível na estação utilizada. Simultaneamente ao enfileiramento desses blocos, os primeiros que entraram na fila da memória RAM são gravados no DAD de destino (DAD_D), na mesma sequência em que estavam armazenados na fila **(E)**.

Este ciclo se repete até que todos os blocos do DAD_O sejam duplicados para o DAD_D .

Após o término da duplicação do DAD_O é gerado o *hash* do conteúdo do DAD_D .

Assim, para o processo de duplicação tem-se como resultados intermediários as duplicações, os respectivos *hashes* de cada DAD_O e o resultado global, que é o conjunto da duplicação e os *hashes* de todos os DAD_O . Nesse trabalho, é admitido que os DADs (DAD_O e DAD_D) são ideais, ou seja, que estão em perfeitos estados de funcionamento e sem setores ou blocos defeituosos (*bad sectors* e *bad blocks*).

Geração de *hash* do DAD Duplicado

Após o término do processo de duplicação de cada DAD_O é gerado o *hash* do DAD_D . A geração dos respectivos *hashes* tem o intuito de prover uma forma para verificação de integridade, para garantir que os dados contidos em ambos dispositivos não foram modificados durante o procedimento de duplicação e varredura. Esse valor é comparado com o respectivo valor do *hash* do DAD_O , conforme a lista de *hashes* (L_H). Se os valores forem iguais o processo de duplicação foi realizado com sucesso, ou seja, não houve modificação dos dados durante a duplicação e a integridade foi preservada. Caso contrário, se os valores obtidos forem diferentes, é necessário repetir o processo de duplicação até que os valores sejam iguais.

3.4.1.3 Processo de Varredura

Após o carregamento das PChs em memória RAM, geração da L_{PCh_e} e início do carregamento dos blocos em memória da estação utilizada, o processo de varredura é iniciado, conforme ilustrado na área pontilhada da Figura 3.6.

Figura 3.6 – Processo de varredura

Na medida em que os blocos que estão sendo copiados do DAD_O para o DAD_D , eles entram na FILA, o processo de varredura cria *threads* que duplicam cada bloco da FILA para outra área da memória RAM e procura as $LPCh_{se}$ sequencialmente do início para o final do bloco (C). A quantidade máxima de *threads* que podem ser criadas simultaneamente deve ser igual ao número de núcleos do processador.

A varredura é realizada por meio da comparação entre os caracteres que formam as palavras da $LPCh_{se}$ e o conteúdo do bloco que está sendo varrido. A sequência de *bytes* que formam as palavras da $LPCh_{se}$ vai se deslocando de caractere em caractere (*byte-a-byte*) em cada bloco (em cada *thread*). Na medida em que a varredura do bloco de uma determinada *thread* é concluída, esta *thread* é descartada (D). Este

ciclo se repete até que todos os blocos que estavam na fila sejam completamente varridos.

Tabela de Lista de quantidade de Ocorrências de PCh

Considerando-se que o tamanho do caractere pode variar dependendo do padrão de codificação utilizado na varredura, a procura é realizada em diferentes padrões de codificação (ASCII, UTF-8, UTF-16, UTF-32, ISO-8859-1 (LATIN-1)). Para cada ocorrência de uma determinada PCh encontrada é realizado o incremento de uma ocorrência na tabela de lista de ocorrências de PChs, no respectivo padrão de codificação em que a PCh foi encontrada. Embora esta tabela contenha menos detalhes quando comparada com a tabela de *Log* de localização, ela é muito útil em consultas rápidas para obter informações mais simples, já que apresenta a quantidade de ocorrências para cada palavra-chave em cada padrão de codificação encontrado e a quantidade total de ocorrências para uma PCh específica.

Tabela de *Log* de localização das PCh

Da mesma forma que ocorre para tabela de lista de ocorrências de PChs, cada ocorrência também é registrada na tabela de *Log* de localização. Nessa tabela, além das ocorrências para cada PCh encontrada, são adicionadas as informações do respectivo *offset* do DAD onde a mesma foi encontrada, os 20 *bytes* anteriores à palavra-chave e 20 *bytes* posteriores. O motivo de apresentar os 20 *bytes* antes e após a PCh é justificado pela possibilidade de mostrar até cinco caracteres de padrões de codificação de até 32 bits, como por exemplo, o UTF-32, que usa quatro *bytes* para representar cada caractere. Sendo assim, é possível visualizar até cinco caracteres nesse padrão de codificação. Para outros padrões de codificação que utilizam menor

quantidade de *bits* para representar seus respectivos caracteres, como por exemplo, o UTF-8, é possível visualizar 20 caracteres, antes e após a PCh. Isto torna possível para um humano verificar visualmente trechos onde a PCh foi localizada.

Os resultados destas tabelas são as saídas do processo de varredura. Assim, têm-se como resultados intermediários: as tabelas de lista de ocorrências e as tabela de *Log* de localização das PChs de cada DAD_O. E como resultado global, o conjunto de tabelas para todos os DADs_O. A quantidade total de ocorrências encontradas em cada DAD pode ser obtida a partir da somatória das ocorrências para cada PCh e adicionalmente, um conjunto global formado pelas respectivas listas com a quantidade total de ocorrências de PChs de cada um dos DADs duplicados.

As listas L_{PChs}, L_H e L_{QO}, não devem ser armazenadas ou gravadas nos DADs_O nem nos DADs_D, e sim armazenadas em um DAD externo, como por exemplo, um *pendrive* ou no próprio DAD da estação na qual a ferramenta de duplicação e varredura é executada, caso não seja um CD ou DVD. Este cuidado deve ser seguido para evitar que os dados dos DADs_O e DADs_D sejam modificados com o conteúdo das respectivas listas, comprometendo a integridade dos dados.

Processo de Varredura em Modo Somente de Inspeção

Se por algum motivo for necessário procurar por outra lista de PChs de interesse do perito, não é necessário realizar novamente o processo de duplicação. Nessas circunstâncias é realizar uma nova varredura sobre o DAD_D em modo de inspeção, prosseguindo com a aplicação dos processos subsequentes normalmente. A necessidade de uma nova busca pode surgir, por exemplo em decorrência de uma nova lista de PChs.

3.4.2 Processo de Filtragem

O objetivo do processo de filtragem é a partir dos resultados obtidos da tabela de *Log* de localização de PChs de cada DAD_O, verificar individualmente cada uma das ocorrências encontradas e classificá-las como verdadeiro-positivo (VP) ou falso-positivo (FP).

No processo de filtragem, para cada ocorrência de PCh encontrada, são analisados os 40 bytes que circundam a PCh, ou seja, 20 bytes anteriores e os 20 bytes posteriores a PCh. Se nessas sequências todos os caracteres forem identificados como imprimíveis²⁷, conforme mostra a Tabela 3.1, a ocorrência é classificada como VP. Caso contrário, se dentre essa sequência de caracteres for identificado pelo menos um caractere classificado como não imprimível, a ocorrência é diagnosticada como FP. Assim, tem-se:

- ❖ Verdadeiro-positivo (VP): Ocorre todas as vezes que os caracteres nas vizinhanças (20 *bytes* anteriores e 20 *bytes* posteriores) de uma ocorrência de PCh são identificados como imprimíveis, conforme mostra a Tabela 3.1.
- ❖ Falso-positivo (FP): Ocorre quando pelo menos um caractere nas vizinhanças de uma ocorrência de PCh não é identificado como imprimível, ou seja, não fazem parte da Tabela 3.1.

²⁷ Neste trabalho considera-se caracteres não-imprimíveis (valores 0 a 31, 127, 129, 141, 143, 144 e 157 do conjunto de caracteres Unicode). <https://support.office.com/pt-br/article/Remover-espacos-e-caracteres-nao-imprimiveis-do-texto-023f3a08-3d56-49e4-bf0c-fe5303222c9d?ui=pt-BR&rs=pt-BR&ad=BR>

Tabela 3.1 – Relação de caracteres imprimíveis

a - z	]
A - Z	{
0 - 9	}
Caracteres alfanuméricos	\$
Caracteres acentuados e Unicode	,
<	:
>	;
@	\
.	/
\w	!
\n - caractere para inserir uma nova linha	~
\r - caractere de "retorno de carro"	^
&	%
?	
"	-
'	+
-	=
	*
_	

Após o processo de filtragem ser realizado para cada um dos DADs, obtém-se a lista de ocorrências filtradas para cada um destes. Essa lista apresenta a quantidade total de VP e por complemento a quantidade de FP de cada DAD. Os resultados da quantidade de ocorrências de VP para cada palavra, obtidos para cada DAD são organizados em uma tabela agregada, na qual o resultado de VP de cada DAD torna-se uma linha dessa tabela, conforme exemplo mostrado na Tabela 3.2.

Tabela 3.2 – Exemplo de tabela agregada com ocorrências de PChs (VP) de todos os DADs

n°do DAD	Palavra 1	Palavra 2	Palavra 3	Palavra 4	Palavra 5	⋮	Palavra n-ésima	Total de Ocorrências
1	96	45	83	17	65	35	79	420
2	15	57	24	39	27	39	49	250
3	53	89	89	23	23	53	31	361
⋮	77	34	68	57	69	35	66	361
⋮	19	32	39	82	70	45	79	366
n	21	77	57	87	38	08	89	377

Na tabela, o números que aparece em cada linha da coluna "n° do DAD" refere-se ao número de um DAD. Assim, para o DAD1 foram encontradas 96 ocorrências da Palavra1, 45 ocorrências da Palavra2, ..., 79 ocorrências para a Palavra n-ésima. No final da linha é mostrado o total de ocorrências para o respectivo DAD, no caso do DAD1, o total de ocorrências classificadas como VP foi 420.

Os dados da tabela agregada são utilizados como entrada para o processo de triagem, conforme será descrito na Seção 3.4.3.

3.4.2.1 Cálculo da Preditividade Positiva do Processo de Filtragem

Denomina-se preditividade positiva do processo de filtragem (δ), a proporção de verdadeiro-positivos (VP) em relação ao total de ocorrências encontradas (VP+FP). Assim o valor da preditividade positiva do processo de filtragem é calculado por:

$$\delta = \frac{VP}{VP+FP} \quad (1)$$

3.4.3 Processo de Triagem

Conforme definida na Seção 1.1, a palavra triagem significa escolha, separação em classes ou local onde se faz a seleção. No contexto desta tese, é definida como o processo para selecionar os DADs que apresentam ocorrências de PChs de interesse para uma determinada investigação dos que não apresentam ocorrências de PChs.

As entradas para o processo de triagem são: a tabela agregada com as ocorrências de PChs classificadas como VP para cada um dos DADs, obtida do processo de filtragem e as PChs com prioridades, que fazem parte da lista de PChs. No processo de triagem, diferentes prioridades podem ser atribuídas às PChs. O estabelecimento ou não de prioridades para a PChs fica a cargo do perito.

Desse modo, é no processo de triagem que ocorre a definição dos critérios de como os resultados da tabela agregada serão utilizados no processo de ranqueamento.

3.4.3.1 Critérios Para o Processo de Triagem

Esta seção descreve as possibilidades em que o processo de triagem pode ocorrer. Primeiramente, serão abordadas as circunstâncias em que a ocorrência de PChs nos DADs são diferentes de zero, ou seja, não há ocorrências de PChs e em seguida são abordadas as circunstâncias em que as ocorrências de PChs nos DADs são iguais a zero.

A triagem pode ocorrer baseada na prioridade de: uma única PCh, de um grupo de PChs individualmente (cada PCh neste grupo pode ter a mesma prioridade ou prioridades distintas) ou para grupos de PChs com prioridades distintas. Reiterando que a definição das PChs que irão integrar cada grupo de prioridade, bem como a definição das prioridades das PChs são de escolha do perito. Este pode ser um

processo iterativo no qual diferentes possibilidades podem ser experimentadas.

Os critérios para o processo de triagem são:

- ❖ Todos os DADs apresentam ocorrências de palavras-chave;
- ❖ Nenhum dos DAD apresentam ocorrências de palavras-chave;
- ❖ Alguns DADs apresentam ocorrências de palavras-chave.

I) Todos os DADs Apresentam Ocorrências de PChs

Para circunstâncias nas quais houveram ocorrências para as PChs procuradas nos DADs é possível:

- a) Utilizar a quantidade de ocorrências de todas as PChs, se o perito não definiu prioridade para as PChs. Neste caso todas as PChs terão a mesma prioridade.
- b) Atribuir prioridade para uma única PCh frente às demais PChs.
- c) Atribuir prioridade para um grupo de PChs frente às demais PChs. Isso significa que todas as PChs pertencentes ao grupo terão a mesma prioridade, porém estas PChs têm maior prioridade frente às outras que não pertencem ao grupo de prioridade. É como se a soma das ocorrências dessas PChs correspondesse a uma única palavra em todos os DADs.
- d) Priorizar mais de um grupo de PChs frente às demais PChs. Significa que o grupo com maior prioridade sempre terá maior importância sobre outros de menor prioridade e assim sucessivamente.

II) Nenhum dos DADs Apresentam Ocorrência de PChs

Para circunstâncias nas quais não houveram ocorrências das PChs procuradas em parte dos DADs, isso pode ser em decorrência de:

- a) As palavras-chave utilizadas realmente não estão no conteúdo dos DADs.
- b) O conteúdo dos dados só apresenta arquivos compactados.
- c) O conteúdo dos dados está cifrado.

A ação a ser tomada é que os DADs que não apresentaram ocorrências sejam separados dos demais. Posteriormente, os DADs compactados devem ser descompactados e varridos novamente para procurar as PChs de interesse.

Para DADs cujo conteúdo dos dados estejam cifrados, não é possível inferir nada sobre este conteúdo. Alternativamente a critério do perito, técnicas de força-bruta podem ser utilizadas.

III) Somente Parte dos DADs Apresentam Ocorrências de PChs

Nas circunstâncias nas quais apenas uma parte dos DADs contém ocorrências de PChs e outra parte não contém. Este conjunto será dividido em dois subconjuntos de DADs.

Para os DADs que apresentaram ocorrências de PChs, devem ser seguidos os procedimentos descritos no item I desta Seção. E serem aplicados os critérios de ranqueamento baseados em prioridades.

Para os DADs que não contém ocorrências de PChs devem ser seguidos os procedimentos descritos no item II da Seção 3.4.3.1.

Dos critérios apresentados acima, para o processo de triagem, somente os apresentados nos itens I e III serão utilizados para a descrição do ranqueamento na Seção 3.4.4.1. Visto que DADs que não apresentam ocorrências de PChs, não seguem para o processo de ranqueamento.

3.4.4 Processo de Ranqueamento

O objetivo do processo de ranqueamento é realizar a ordenação e sugerir a sequência em que os DADs_D devem ser examinados na etapa de análise. No ranqueamento é estabelecida uma classificação mediante a comparação da pontuação entre os DADs_D. A pontuação refere-se à quantidade de ocorrências de PChs, considerando as prioridades, quando houverem.

3.4.4.1 Procedimentos Para o Processo de Ranqueamento

Somente os DADs que apresentaram ocorrências de PChs serão considerados no processo de ranqueamento. A pontuação para cada DAD é obtida a partir das informações do processo de triagem, que são: a tabela agregada com a quantidade de ocorrências de PChs dos DADs e as prioridades das PChs, quando houverem.

Assim, os DADs provenientes da triagem necessariamente se enquadram em algum dos procedimentos descrito a seguir:

I) Procedimentos Para DADs que Apresentaram PChs

- a) PChs sem prioridade: é realizada a soma de todas as ocorrências de PChs do DAD, obtendo-se a quantidade total de ocorrências de PChs do DAD. O ranqueamento é realizado classificando-se a quantidade total de ocorrências de cada DAD em ordem decrescente.
- b) Prioridade para uma única PCh: o ranqueamento é realizado em função da PCh com prioridade. Assim, o DAD que apresentar maior quantidade de ocorrências para esta PCh terá maior importância sobre os que apresentarem menor quantidade de ocorrências desta PCh. Se ocorrer empate entre os DADs, deve ser utilizado o total de ocorrências de PChs.

- c) Prioridade para um único grupo de PChs: O ranqueamento é realizado em função da soma das ocorrências das PChs do grupo de PChs com prioridades. Se ocorrer empate, devem ser somadas as demais PChs e os resultados ordenados. Assim, primeiramente o ranqueamento é realizado pelas ocorrências do grupo com prioridade e o desempate é baseado na ordenação decrescente de ocorrências de todas as PChs.
- d) Prioridade para mais de um grupo de PChs: Neste caso, o ranqueamento é realizado pela soma das ocorrências de cada grupo. Sendo que o grau de importância é do grupo com prioridade mais alta para o grupo com prioridade mais baixa. O resultado do grupo de menor prioridade é utilizado somente em circunstâncias nas quais ocorra empate entre DADs do grupo de maior prioridade. Caso persista o empate para o grupo de menor prioridade, deverá ser utilizada a ordenação decrescente de ocorrências de todas as PChs.

3.5 Exemplos Para Demonstração dos Processos de Triagem e Ranqueamento

A seguir, são apresentados exemplos ilustrativos com o intuito de mostrar de que forma os critérios de triagem e procedimentos para ranqueamento são utilizados. Para isso, será utilizada a tabela agregada resultante de um processo de filtragem de 25 DADs, cujas PChs são formadas pelos nomes de 29 países. Na Tabela 3.3, são apresentados os resultados da quantidade de ocorrências de PChs para cada DAD.

Tabela 3.3 – Exemplo de tabela agregada com os valores de VP

	Palavras-chave (PChs)																														
nº do DAD	Alemanha	Argentina	Brasil	Canadá	Estados Unidos	Espanha	Guatemala	Holanda	Hungria	Irã	Jamaica	Kuwait	Marrocos	Noruega	Nepal	Portugal	Suriname	Uruguai	Venezuela	Groelândia	Paraguai	Luxemburgo	Suécia	Japão	França	Inglaterra	Itália	China	Austrália	Total Ocorrênc.	
1	1	0	26	620	1	0	24	9	246	39	0	0	576	4	9	0	0	40	0	0	3	0	33	39	0	0	24	155	0	1849	
2	1	0	22	774	1	0	26	10	142	45	0	0	744	13	8	0	0	28	0	0	3	0	36	65	0	0	99	138	246	2401	
3	1	0	21	906	1	0	39	10	184	18	0	0	620	3	14	0	0	35	180	0	3	0	37	77	0	0	83	145	142	2519	
4	1	0	24	715	1	3	55	9	0	42	0	0	774	7	22	0	0	32	143	0	3	363	35	61	0	0	25	227	184	2726	
5	1	0	26	576	1	22	59	11	0	0	0	0	906	6	17	0	0	32	1378	0	3	475	34	23	0	0	17	255	182	4024	
6	2	0	31	696	2	27	90	11	0	0	0	0	715	3	17	0	0	35	1273	0	3	426	37	35	0	1	45	288	150	3887	
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	64	0	22	49	12	11	1	3	5	43	142	0	696	6	0	14	19	6	0	0	3	444	16	90	0	0	58	152	3	1859	
9	402	0	25	64	89	11	28	3	52	180	184	0	731	12	1	106	123	5	0	0	3	521	33	5	26	67	61	159	22	2913	
10	347	0	24	96	87	10	38	4	67	155	182	0	723	6	1	95	104	37	0	0	3	1	27	0	26	51	2	246	27	2359	
11	550	0	36	75	49	0	2	4	10	33	150	7	949	6	7	156	321	11	4	0	4	0	19	19	15	164	33	142	27	2793	
12	639	0	42	0	64	0	3	4	23	58	253	5	1154	12	3	222	363	77	2	0	5	0	22	5	48	186	65	184	11	3450	
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	761	0	38	0	75	246	2	4	24	105	0	3	1280	10	3	298	426	132	3	0	4	0	17	8	68	208	23	150	10	3898	
15	873	0	35	0	106	142	2	3	40	126	0	6	1431	9	3	341	477	141	2	0	4	0	20	9	71	214	29	253	0	4337	
16	752	0	31	0	79	184	2	3	29	102	0	7	1378	3	3	304	444	131	2	0	3	0	18	14	66	197	63	221	0	4036	
17	941	0	21	0	105	0	2	1	39	127	87	6	1273	3	3	369	521	136	2	0	3	0	15	27	91	239	39	124	246	4420	
18	93	0	56	0	4	0	1	4	0	13	49	0	998	22	0	2	31	0	0	0	2	0	16	32	1	32	45	247	142	1790	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	14	0	475	0	1	0	0	0	0	1	96	0	823	27	0	0	6	4	1	0	0	0	8	98	3	4	42	204	22	1829	
21	93	0	426	0	16	0	0	0	1	21	75	0	433	11	0	1	31	3	1	0	0	0	2	55	6	31	97	103	25	1431	
22	139	0	477	0	21	0	0	0	1	22	0	0	341	11	0	3	50	1	1	0	0	0	3	80	8	52	55	110	24	1399	
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	9	0	521	0	1	0	0	0	0	0	0	0	227	0	0	0	5	3	1	0	0	0	3	321	6	3	87	71	42	1300	
25	45	0	77	96	3	2	5	75	173	96	82	37	91	37	89	98	65	65	32	0	54	87	56	23	25	58	74	71	23	1639	

Seguindo os critérios para triagem, conforme a Seção 3.4.3.1, observa-se na Tabela 3.3 que os DADs 7, 13, 19 e 23 não apresentam ocorrências de PChs. Baseando-se nos critérios de triagem explicados anteriormente, esses DADs são separados dos demais DADs. Para os DADs que apresentam ocorrências de PChs serão exemplificados nos casos a seguir.

Caso 1: Deseja-se identificar quais DADs possuem maior quantidade de ocorrência dos países procurados. Neste caso, os nomes dos países são as PChs e não há PChs com prioridade.

- Triagem: sem prioridade

Por não ter PChs com prioridade, serão utilizados o valor total de todas as ocorrências de VP em cada DAD.

- Ranqueamento:

A partir da tabela de todos DADs que passaram pela triagem, é realizada a ordenação decrescente do total de ocorrências de

PChs de todos DADs. Então, a classificação é obtida, conforme o resultado mostrado na Tabela 3.4.

Tabela 3.4 – Resultado do ranqueamento das PChs sem prioridade

	Palavras-chave (PChs)																																
nº do DAD	Alemanha	Argentina	Brasil	Canadá	Estados Unidos	Espanha	Guatemala	Holanda	Hungria	Irã	Jamaica	Kuwait	Marrocos	Noruega	Nepal	Portugal	Suriname	Uruguai	Venezuela	Groelândia	Paraguai	Luxemburgo	Suécia	Japão	França	Inglaterra	Itália	China	Austrália	Total Ocorrenc.			
17	941	0	21	0	105	0	2	1	39	127	87	6	1273	3	3	369	521	136	2	0	3	0	15	27	91	239	39	124	246	4420			
15	873	0	35	0	106	142	2	3	40	126	0	6	1431	9	3	341	477	141	2	0	4	0	20	9	71	214	29	253	0	4337			
16	752	0	31	0	79	184	2	3	29	102	0	7	1378	3	3	304	444	131	2	0	3	0	18	14	66	197	63	221	0	4036			
5	1	0	26	576	1	22	59	11	0	0	0	0	906	6	17	0	0	32	1378	0	3	475	34	23	0	0	17	255	182	4024			
14	761	0	38	0	75	246	2	4	24	105	0	3	1280	10	3	298	426	132	3	0	4	0	17	8	68	208	23	150	10	3898			
6	2	0	31	696	2	27	90	11	0	0	0	0	715	3	17	0	0	35	1273	0	3	426	37	35	0	1	45	288	150	3887			
12	639	0	42	0	64	0	3	4	23	58	253	5	1154	12	3	222	363	77	2	0	5	0	22	5	48	186	65	184	11	3450			
9	402	0	25	64	89	11	28	3	52	180	184	0	731	12	1	106	123	5	0	0	3	521	33	5	26	67	61	159	22	2913			
11	550	0	36	75	49	0	2	4	10	33	150	7	949	6	7	156	321	11	4	0	4	0	19	19	15	164	33	142	27	2793			
4	1	0	24	715	1	3	55	9	0	42	0	0	774	7	22	0	0	32	143	0	3	363	35	61	0	0	25	227	184	2726			
3	1	0	21	906	1	0	39	10	184	18	0	0	620	3	14	0	0	35	180	0	3	0	37	77	0	0	83	145	142	2519			
2	1	0	22	774	1	0	26	10	142	45	0	0	744	13	8	0	0	28	0	0	3	0	36	65	0	0	99	138	246	2401			
10	347	0	24	96	87	10	38	4	67	155	182	0	723	6	1	95	104	37	0	0	3	1	27	0	26	51	2	246	27	2359			
8	64	0	22	49	12	11	1	3	5	43	142	0	696	6	0	14	19	6	0	0	3	444	16	90	0	0	58	152	3	1859			
1	1	0	26	620	1	0	24	9	246	39	0	0	576	4	9	0	0	40	0	0	3	0	33	39	0	0	24	155	0	1849			
20	14	0	475	0	1	0	0	0	0	1	96	0	823	27	0	0	6	4	1	0	0	0	8	98	3	4	42	204	22	1829			
18	93	0	56	0	4	0	1	4	0	13	49	0	998	22	0	2	31	0	0	0	2	0	16	32	1	32	45	247	142	1790			
25	45	0	77	96	3	2	5	75	173	96	82	37	91	37	89	98	65	65	32	0	54	87	56	23	25	58	74	71	23	1639			
21	93	0	426	0	16	0	0	0	1	21	75	0	433	11	0	1	31	3	1	0	0	0	2	55	6	31	97	103	25	1431			
22	139	0	477	0	21	0	0	0	1	22	0	0	341	11	0	3	50	1	1	0	0	0	3	80	8	52	55	110	24	1399			
24	9	0	521	0	1	0	0	0	0	0	0	0	227	0	0	0	5	3	1	0	0	0	3	321	6	3	87	71	42	1300			

Mediante os resultados, o DAD que contém a maior quantidade dos nomes dos países procurados é o 17. A sugestão da sequência que os DADs devem ser analisados é: 17, 15, 16, ..., 24.

Caso 2: Deseja-se identificar quais DADs possuem maior quantidade de ocorrência da PCh “Brasil”.

- Triagem: com prioridade para uma única PCh.

Neste caso será utilizada a PCh “Brasil”. A PCh “Brasil” possui maior prioridade frente às demais

- Ranqueamento:

O ranqueamento deve ser realizado em função dos resultados da triagem. Deve ser organizado em ordem decrescente da quantidade de ocorrência para esta PCh, conforme mostra a Tabela 3.5.

Tabela 3.5 – Resultado do ranqueamento de uma PCh com prioridade

	Palavras-chave (PChs)																																
nº do DAD	Alemanha	Argentina	Brasil	Canadá	Estados Unidos	Espanha	Guatemala	Holanda	Hungria	Irã	Jamaica	Kuwait	Marrocos	Noruega	Nepal	Portugal	Suriname	Uruguai	Venezuela	Groelândia	Paraguai	Luxemburgo	Suécia	Japão	França	Inglaterra	Itália	China	Austrália	Total Ocorrênc.			
24	9	0	521	0	1	0	0	0	0	0	0	0	227	0	0	0	5	3	1	0	0	0	3	321	6	3	87	71	42	1300			
22	139	0	477	0	21	0	0	0	1	22	0	0	341	11	0	3	50	1	1	0	0	0	3	80	8	52	55	110	24	1399			
20	14	0	475	0	1	0	0	0	0	1	96	0	823	27	0	0	6	4	1	0	0	0	8	98	3	4	42	204	22	1829			
21	93	0	426	0	16	0	0	0	1	21	75	0	433	11	0	1	31	3	1	0	0	0	2	55	6	31	97	103	25	1431			
25	45	0	77	96	3	2	5	75	173	96	82	37	91	37	89	98	65	65	32	0	54	87	56	23	25	58	74	71	23	1639			
18	93	0	56	0	4	0	1	4	0	13	49	0	998	22	0	2	31	0	0	0	2	0	16	32	1	32	45	247	142	1790			
12	639	0	42	0	64	0	3	4	23	58	253	5	1154	12	3	222	363	77	2	0	5	0	22	5	48	186	65	184	11	3450			
14	761	0	38	0	75	246	2	4	24	105	0	3	1280	10	3	298	426	132	3	0	4	0	17	8	68	208	23	150	10	3898			
11	550	0	36	75	49	0	2	4	10	33	150	7	949	6	7	156	321	11	4	0	4	0	19	19	15	164	33	142	27	2793			
15	873	0	35	0	106	142	2	3	40	126	0	6	1431	9	3	341	477	141	2	0	4	0	20	9	71	214	29	253	0	4337			
16	752	0	31	0	79	184	2	3	29	102	0	7	1378	3	3	304	444	131	2	0	3	0	18	14	66	197	63	221	0	4036			
6	2	0	31	696	2	27	90	11	0	0	0	0	715	3	17	0	0	35	1273	0	3	426	37	35	0	1	45	288	150	3887			
5	1	0	26	576	1	22	59	11	0	0	0	0	906	6	17	0	0	32	1378	0	3	475	34	23	0	0	17	255	182	4024			
1	1	0	26	620	1	0	24	9	246	39	0	0	576	4	9	0	0	40	0	0	3	0	33	39	0	0	24	155	0	1849			
9	402	0	25	64	89	11	28	3	52	180	184	0	731	12	1	106	123	5	0	0	3	521	33	5	26	67	61	159	22	2913			
4	1	0	24	715	1	3	55	9	0	42	0	0	774	7	22	0	0	32	143	0	3	363	35	61	0	0	25	227	184	2726			
10	347	0	24	96	87	10	38	4	67	155	182	0	723	6	1	95	104	37	0	0	3	1	27	0	26	51	2	246	27	2359			
2	1	0	22	774	1	0	26	10	142	45	0	0	744	13	8	0	0	28	0	0	3	0	36	65	0	0	99	138	246	2401			
8	64	0	22	49	12	11	1	3	5	43	142	0	696	6	0	14	19	6	0	0	3	444	16	90	0	0	58	152	3	1859			
17	941	0	21	0	105	0	2	1	39	127	87	6	1273	3	3	369	521	136	2	0	3	0	15	27	91	239	39	124	246	4420			
3	1	0	21	906	1	0	39	10	184	18	0	0	620	3	14	0	0	35	180	0	3	0	37	77	0	0	83	145	142	2519			

Conforme o resultado obtido, o DAD contém a maior quantidade da PCh “Brasil” é o 24. A sugestão da sequência para início da análise é: 24, 22, 20, ..., 3.

Observe que os DADs 16 e 6, nesta ordem, aparecem empatados. Aplicando o critério de desempate, deve ser considerado o total de ocorrências desses DADs. Ou seja, o que apresentar maior quantidade no total de ocorrências será considerado de maior importância. Neste caso, o DAD 16 terá maior importância, pois apresenta 4036 ocorrências de PCh enquanto o DAD 6 apresenta 3887 ocorrências. O mesmo procedimento pode ser seguido para os DADs 5 e 1, 4 e 10, 2 e 8, 17 e 3.

Caso 3: Deseja-se identificar quais DADs possuem maior quantidade de ocorrência de um grupo de PChs: {Brasil, Paraguai, Uruguai e Venezuela}.

- Triagem: com prioridade para um grupo de PCh. A soma das ocorrências dessas PChs correspondem a uma única palavra em todos os DADs. Estas PChs terão prioridade frente às demais.

- **Ranqueamento:**

O ranqueamento é realizado em função da soma das ocorrências das PChs do grupo de PChs com prioridades. A sequência de ranqueamento é o resultado da classificação da soma da quantidade de ocorrências de PChs do grupo de PCh em ordem decrescente, conforme mostra a Tabela 3.6

Tabela 3.6 – Resultado do ranqueamento do grupo de PChs com prioridade

nº do DAD	Palavras-chave																												GRUPO PRIOR.
	Alemanha	Brasil	Canadá	Estados Unidos	Espanha	Guatemala	Holanda	Hungria	Irã	Jamaica	Kuwait	Marrocos	Noruega	Nepal	Portugal	Suriname	Uruguai	Venezuela	Paraguai	Luxemburgo	Suécia	Japão	França	Inglaterra	Itália	China	Austrália	total	
5	1	26	576	1	22	59	11	0	0	0	0	906	6	17	0	0	32	1378	3	475	34	23	0	0	17	255	182	4024	1439
6	2	31	696	2	27	90	11	0	0	0	0	715	3	17	0	0	35	1273	3	426	37	35	0	1	45	288	150	8887	1342
24	9	521	0	1	0	0	0	0	0	0	0	227	0	0	0	5	3	1	0	0	3	321	6	3	87	71	42	1300	525
20	14	475	0	1	0	0	0	0	1	96	0	823	27	0	0	6	4	1	0	0	8	98	3	4	42	204	22	1829	480
22	139	477	0	21	0	0	0	0	1	22	0	341	11	0	3	50	1	1	0	0	3	80	8	52	55	110	24	1399	479
21	93	426	0	16	0	0	0	1	21	75	0	433	11	0	1	31	3	1	0	0	2	55	6	31	97	103	25	1431	430
3	1	21	906	1	0	39	10	184	18	0	0	620	3	14	0	0	35	180	3	0	37	77	0	0	83	145	142	2519	239
25	45	77	96	3	2	5	75	173	96	82	37	91	37	89	98	65	65	32	54	87	56	23	25	58	74	71	23	1639	228
4	1	24	715	1	3	55	9	0	42	0	0	774	7	22	0	0	32	143	3	363	35	61	0	0	25	227	184	2726	202
15	873	35	0	106	142	2	3	40	126	0	6	1431	9	3	341	477	141	2	4	0	20	9	71	214	29	253	0	4337	182
14	761	38	0	75	246	2	4	24	105	0	3	1280	10	3	298	426	132	3	4	0	17	8	68	208	23	150	10	8898	177
16	752	31	0	79	184	2	3	29	102	0	7	1378	3	3	304	444	131	2	3	0	18	14	66	197	63	221	0	4036	167
17	941	21	0	105	0	2	1	39	127	87	6	1273	3	3	369	521	136	2	3	0	15	27	91	239	39	124	246	4420	162
12	639	42	0	64	0	3	4	23	58	253	5	1154	12	3	222	363	77	2	5	0	22	5	48	186	65	184	11	3450	126
1	1	26	620	1	0	24	9	246	39	0	0	576	4	9	0	0	40	0	3	0	33	39	0	0	24	155	0	1849	69
10	347	24	96	87	10	38	4	67	155	182	0	723	6	1	95	104	37	0	3	1	27	0	26	51	2	246	27	2359	64
18	93	56	0	4	0	1	4	0	13	49	0	998	22	0	2	31	0	0	2	0	16	32	1	32	45	247	142	1790	58
11	550	36	75	49	0	2	4	10	33	150	7	949	6	7	156	321	11	4	4	0	19	19	15	164	33	142	27	2793	55
2	1	22	774	1	0	26	10	142	45	0	0	744	13	8	0	0	28	0	3	0	36	65	0	0	99	138	246	2401	53
9	402	25	64	89	11	28	3	52	180	184	0	731	12	1	106	123	5	0	3	521	33	5	26	67	61	159	22	2913	33
8	64	22	49	12	11	1	3	5	43	142	0	696	6	0	14	19	6	0	3	444	16	90	0	0	58	152	3	1859	31

Conforme o resultado obtido, o DAD contém a maior quantidade de PChs é DAD 5. A sugestão da sequência que os DADs devem ser analisados apresentada na coluna da extrema esquerda: 5, 6, 24, ..., 8.

Caso 4: Deseja-se identificar quais DADs possuem maior quantidade de ocorrências entre dois grupos de PChs. O grupo 1 (G1) é formado pelos países {Noruega e Suécia} e grupo 2 (G2) é formado pelos países {Brasil, Paraguai, Uruguai e Venezuela}.

- **Triagem:** com prioridades diferentes para dois grupos de PChs. Deve ser definido pelo perito qual dos dois grupos de PCh terá maior importância.
- **Ranqueamento:** Realiza a soma das ocorrências das PChs do grupo de maior prioridade, neste caso G1 e a soma do grupo de menor

prioridade, neste caso G2. O resultado do ranqueamento ocorre em função da classificação da quantidade de ocorrências em ordem decrescente de G1 e depois da classificação da quantidade de ocorrências em ordem decrescente de G2, conforme mostra a Tabela 3.7

Tabela 3.7 – Resultado do ranqueamento dos grupos de PChs com prioridade

	Palavras-chave (PChs)																													
nº do DAD	Alemanha	Brasil	Canadá	Estados Unidos	Espanha	Guatemala	Holanda	Hungria	Irã	Jamaica	Kuwait	Marrocos	Noruega	Nepal	Portugal	Suriname	Uruguai	Venezuela	Paraguai	Luxemburgo	Suécia	Japão	França	Inglaterra	Itália	China	Austrália	Total Ocorrênc.	G1	G2
25	45	77	96	3	2	5	75	173	96	82	37	91	37	89	98	65	65	32	54	87	56	23	25	58	74	71	23	1639	93	228
2	1	22	774	1	0	26	10	142	45	0	0	744	13	8	0	0	28	0	3	0	36	65	0	0	99	138	246	2401	49	53
9	402	25	64	89	11	28	3	52	180	184	0	731	12	1	106	123	5	0	3	521	33	5	26	67	61	159	22	2913	45	33
4	1	24	715	1	3	55	9	0	42	0	0	774	7	22	0	0	32	143	3	363	35	61	0	0	25	227	184	2726	42	202
5	1	26	576	1	22	59	11	0	0	0	0	906	6	17	0	0	32	1378	3	475	34	23	0	0	17	255	182	4024	40	1439
6	2	31	696	2	27	90	11	0	0	0	0	715	3	17	0	0	35	1273	3	426	37	35	0	1	45	288	150	3887	40	1342
3	1	21	906	1	0	39	10	184	18	0	0	620	3	14	0	0	35	180	3	0	37	77	0	0	83	145	142	2519	40	239
18	93	56	0	4	0	1	4	0	13	49	0	998	22	0	2	31	0	0	2	0	16	32	1	32	45	247	142	1790	38	58
1	1	26	620	1	0	24	9	246	39	0	0	576	4	9	0	0	40	0	3	0	33	39	0	0	24	155	0	1849	37	69
20	14	475	0	1	0	0	0	0	1	96	0	823	27	0	0	6	4	1	0	0	8	98	3	4	42	204	22	1829	35	480
12	639	42	0	64	0	3	4	23	58	253	5	1154	12	3	222	363	77	2	5	0	22	5	48	186	65	184	11	3450	34	126
10	347	24	96	87	10	38	4	67	155	182	0	723	6	1	95	104	37	0	3	1	27	0	26	51	2	246	27	2359	33	64
15	873	35	0	106	142	2	3	40	126	0	6	1431	9	3	341	477	141	2	4	0	20	9	71	214	29	253	0	4337	29	182
14	761	38	0	75	246	2	4	24	105	0	3	1280	10	3	298	426	132	3	4	0	17	8	68	208	23	150	10	3898	27	177
11	550	36	75	49	0	2	4	10	33	150	7	949	6	7	156	321	11	4	4	0	19	19	15	164	33	142	27	2793	25	55
8	64	22	49	12	11	1	3	5	43	142	0	696	6	0	14	19	6	0	3	444	16	90	0	0	58	152	3	1859	22	31
16	752	31	0	79	184	2	3	29	102	0	7	1378	3	3	304	444	131	2	3	0	18	14	66	197	63	221	0	4036	21	167
17	941	21	0	105	0	2	1	39	127	87	6	1273	3	3	369	521	136	2	3	0	15	27	91	239	39	124	246	4420	18	162
22	139	477	0	21	0	0	0	1	22	0	0	341	11	0	3	50	1	1	0	0	3	80	8	52	55	110	24	1399	14	479
21	93	426	0	16	0	0	0	1	21	75	0	433	11	0	1	31	3	1	0	0	2	55	6	31	97	103	25	1431	13	430
24	9	521	0	1	0	0	0	0	0	0	0	227	0	0	0	5	3	1	0	0	3	321	6	3	87	71	42	1300	3	525

Os resultados obtidos indicam que o DAD com maior quantidade de ocorrências de PChs de G1 é o DAD 25 e o último é o 24. Porém, em caso de empate na classificação de G1, deve ser utilizado o resultado da classificação de G2, como por exemplo, o que ocorre com os DADs 3, 5 e 6, todos apresentam 40 de ocorrências de PChs. Sendo assim, a sequência de classificação decrescente é 5, 6 e 3. Dessa forma, a sugestão da sequência para análise é: 25, 2, 9, 4, 5, 6, 3, ... e 24.

Um aspecto a ser observado no processo de triagem é a possibilidade do perito comparar o resultado obtido anteriormente com o resultado da classificação pela soma das ocorrências dos dois grupos de PChs, conforme mostra a Tabela 3.8. Neste último caso, é atribuído o mesmo grau de importância para os dois grupos de PChs, o que resulta em um único grupo com a quantidade de ocorrências somadas.

A seguir na Tabela 3.8 é mostrada a classificação obtida a partir da soma de ocorrências de G1 e G2.

Tabela 3.8 – Comparação do resultado do ranqueamento dos grupos de PChs

	Palavras-chave																														
nº do DAD	Alemanha	Brasil	Canadá	Estados Unidos	Espanha	Guatemala	Holanda	Hungria	Irã	Jamaica	Kuwait	Marrocos	Noruega	Nepal	Portugal	Suriname	Uruguai	Venezuela	Paraguai	Luxemburgo	Suécia	Japão	França	Inglaterra	Itália	China	Austrália	Total Ocorrênc.	G1	G2	G1+G2
5	1	26	576	1	22	59	11	0	0	0	0	906	6	17	0	0	32	1378	3	475	34	23	0	0	17	255	182	4024	40	1439	1479
6	2	31	696	2	27	90	11	0	0	0	0	715	3	17	0	0	35	1273	3	426	37	35	0	1	45	288	150	3887	40	1342	1382
24	9	521	0	1	0	0	0	0	0	0	0	227	0	0	0	5	3	1	0	0	3	321	6	3	87	71	42	1300	3	525	528
20	14	475	0	1	0	0	0	0	1	96	0	823	27	0	0	6	4	1	0	0	8	98	3	4	42	204	22	1829	35	480	515
22	139	477	0	21	0	0	0	1	22	0	0	341	11	0	3	50	1	1	0	0	3	80	8	52	55	110	24	1399	14	479	493
21	93	426	0	16	0	0	0	1	21	75	0	433	11	0	1	31	3	1	0	0	2	55	6	31	97	103	25	1431	13	430	443
25	45	77	96	3	2	5	75	173	96	82	37	91	37	89	98	65	65	32	54	87	56	23	25	58	74	71	23	1639	93	228	321
3	1	21	906	1	0	39	10	184	18	0	0	620	3	14	0	0	35	180	3	0	37	77	0	0	83	145	142	2519	40	239	279
4	1	24	715	1	3	55	9	0	42	0	0	774	7	22	0	0	32	143	3	363	35	61	0	0	25	227	184	2726	42	202	244
15	873	35	0	106	142	2	3	40	126	0	6	1431	9	3	341	477	141	2	4	0	20	9	71	214	29	253	0	4337	29	182	211
14	761	38	0	75	246	2	4	24	105	0	3	1280	10	3	298	426	132	3	4	0	17	8	68	208	23	150	10	3898	27	177	204
16	752	31	0	79	184	2	3	29	102	0	7	1378	3	3	304	444	131	2	3	0	18	14	66	197	63	221	0	4036	21	167	188
17	941	21	0	105	0	2	1	39	127	87	6	1273	3	3	369	521	136	2	3	0	15	27	91	239	39	124	246	4420	18	162	180
12	639	42	0	64	0	3	4	23	58	253	5	1154	12	3	222	363	77	2	5	0	22	5	48	186	65	184	11	3450	34	126	160
1	1	26	620	1	0	24	9	246	39	0	0	576	4	9	0	0	40	0	3	0	33	39	0	0	24	155	0	1849	37	69	106
2	1	22	774	1	0	26	10	142	45	0	0	744	13	8	0	0	28	0	3	0	36	65	0	0	99	138	246	2401	49	53	102
10	347	24	96	87	10	38	4	67	155	182	0	723	6	1	95	104	37	0	3	1	27	0	26	51	2	246	27	2359	33	64	97
18	93	56	0	4	0	1	4	0	13	49	0	998	22	0	2	31	0	0	2	0	16	32	1	32	45	247	142	1790	38	58	96
11	550	36	75	49	0	2	4	10	33	150	7	949	6	7	156	321	11	4	4	0	19	19	15	164	33	142	27	2793	25	55	80
9	402	25	64	89	11	28	3	52	180	184	0	731	12	1	106	123	5	0	3	521	33	5	26	67	61	159	22	2913	45	33	78
8	64	22	49	12	11	1	3	5	43	142	0	696	6	0	14	19	6	0	3	444	16	90	0	0	58	152	3	1859	22	31	53

Ao somar as ocorrências dos grupos G1 e G2 observa-se a mudança na classificação dos DADs, conforme pode ser observado na primeira coluna. O DAD 5 assume a primeira colocação e a última é ocupada pelo DAD 8.

Comparando-se ambas as classificações para os DADs, a partir das Tabelas 3.7 e 3.8, o DAD 24 que ocupa a última colocação na Tabela 3.7 passa para a terceira colocação na tabela 3.8.

Na prática, o que se observa é que a classificação realizada a partir de grupos com diferentes prioridades implica na qualidade no ranqueamento. Esta característica evidencia a importância estratégica ao analisar corretamente os critérios para realização do processo de triagem.

Nos exemplos apresentados, foi possível mostrar de que forma os critérios de triagem e procedimentos para ranqueamento são utilizados. Além disso, apresentam como é obtida a sugestão da sequência em que os DADs devem ser analisados.

4 VALIDAÇÃO DO MÉTODO PARA TRIAGEM E RANQUEAMENTO DE COMPUTADORES

Após a descrição do método e seus respectivos processos, conforme descrito no Capítulo 3, o próximo passo foi o planejamento da sua validação. Foi verificada a necessidade de duas importantes comparações.

A primeira é a comparação entre o tempo necessário para duplicar um DAD sem realizar a varredura para busca por PChs e o tempo necessário para duplicar um DAD com a varredura para busca por PChs simultaneamente. Essa avaliação permite conhecer qual o impacto da varredura sobre o tempo de duplicação. O desejável é que a varredura onere o mínimo possível o tempo de duplicação.

A segunda importante comparação é a quantidade de ocorrências de PChs encontradas por uma ferramenta que realiza a procura em dados interpretados por um sistema de arquivos e a quantidade de ocorrências encontradas por uma ferramenta que realiza a procura em dados não interpretados por um sistema de arquivos.

4.1 Análise de Ferramentas Para Validação do Método

Foram procuradas e analisadas ferramentas de *software* comerciais e de código livre e que possibilitassem realizar a duplicação e outras para realizar a varredura nos dados a serem duplicados para procurar por uma lista de PChs.

Foram encontradas ferramentas, como por exemplo, Encase Forensic^{®28} e Forensic Toolkit - FTK^{®29}, de códigos proprietários. Embora apresentassem as funcionalidades para duplicação e varredura, são limitadas a realizar apenas um processo de cada vez. Ressaltando ainda que as versões disponibilizadas gratuitamente nos respectivos sites, só apresentam a funcionalidade para duplicação, ou seja, sem a possibilidade para que após a duplicação as mesmas pudessem ser utilizadas para procura por PChs, por esse motivo não foram incluídas neste trabalho.

As ferramentas que mais se aproximaram das funcionalidades necessárias para realização da validação do método e que também são de código livre foi a ferramenta “DD” (*Disk Duplicator*) para duplicação e *Autopsy* para varredura do DAD.

4.1.1 *Disk Duplicator (DD)*

A ferramenta DD é um utilitário nativo do sistema operacional Unix e seus derivantes. É executado via linha de comando e suas principais funcionalidades são duplicação (*byte-a-byte*) e conversão de formatos de arquivos de imagem. A duplicação pode ser realizada entre DADs, por exemplo, de um DAD_O para um DAD_D ou de um DAD_O para um arquivo de imagem de disco.

²⁸Disponível em: <<https://www.guidancesoftware.com/products/Pages/Product-Forms/Forensic-Imager-download.aspx>>. Acesso em: 01 de mai. de 2014.

²⁹Disponível em: <<http://accessdata.com/product-download>>. Acesso em: 01 de mai. de 2014.

A ferramenta DD executa as operações de leitura e escrita alternadamente de acordo com as fatias de tempo do escalonador (*scheduler*) de processos, como ilustrado na Figura 4.1.

Figura 4.1 - Procedimento das operações do DD distribuído em função do tempo.

O tempo de execução do DD é diretamente proporcional aos tempos de leitura e de escrita do dispositivo (DAD). Estes tempos dependem de diversos fatores, tais como, o melhor tamanho de bloco a ser lido/escrito, que pode variar de DAD para DAD (DADs mais antigos utilizam 512 *bytes* e os mais novos utilizam 4096 *bytes*), da velocidade do barramento do equipamento utilizado para duplicação, da quantidade de memória RAM disponível e do desempenho do processador.

4.1.2 *Autopsy*

A ferramenta *Autopsy*³⁰ é uma plataforma utilizada para análise Digital Forense com interface gráfica para o *software The Sleuth Kit*³¹ e outras ferramentas de análise Digital Forense. É possível utilizá-la para análise de dados em DADs a partir de um arquivo de imagem, entre outras funcionalidades. O *The Sleuth Kit* é um conjunto de ferramentas que são invocadas via linha de comando que permite analisar dados contidos no DAD, igualmente ao *Autopsy*, permite a análise de arquivos de imagens de um DAD.

³⁰Disponível em: <<http://www.sleuthkit.org/autopsy/>>. Acesso em: 01 de mai. de 2014.

³¹Disponível em: <<http://www.sleuthkit.org/sleuthkit/>>. Acesso em: 01 de mai. de 2014.

4.1.3 Resultados da Análise das Ferramentas para Duplicação e Varredura

Após a análise das ferramentas encontradas, decidiu-se que ferramenta DD seria utilizada para os testes de medição do tempo de duplicação. Essa ferramenta possibilita realizar buscas simples de apenas um PCh por vez, invocando o comando ‘dd’ juntamente com os comandos ‘strings’ e ‘grep’ via linha de comando. Porém, não é possível carregar uma lista de PChs a serem buscadas durante a duplicação, característica essa considerada importante para os testes do método proposto. Nos testes iniciais, foi observado que mesmo a procura para uma única PCh simultaneamente a duplicação, o tempo de duplicação é significativamente onerado.

A ferramenta encontrada para medição da quantidade de ocorrências de PChs foi a *Autopsy*. A ferramenta é executada em alto nível, mas realiza buscas por PCh em baixo nível a partir de um DAD. O mecanismo de busca de PCh da *Autopsy* é uma interface para os comandos ‘strings’ e ‘grep’ que se encontram em todas as versões atuais do sistema operacional Unix e seus derivantes. A ferramenta *Autopsy* executa o comando ‘strings’ primeiro, e depois o comando ‘grep’ durante a saída do comando ‘strings’. O motivo da execução de ambos os comandos conjuntamente é para suprir a deficiência do comando ‘grep’, já que ele é mais lento e examina apenas caracteres em ASCII, portanto não encontra PCh com codificação Unicode³², função que é realizada pelo comando ‘strings’. Este é mais rápido do que o comando ‘grep’ ao inspecionar um sistema de arquivo de imagem. A execução desses comandos conjuntamente compromete o desempenho da ferramenta. Mesmo utilizando esses comandos conjuntamente, a ferramenta *Autopsy* pode deixar de encontrar PCh localizadas em

³² Unicode é um padrão que permite aos computadores representar e manipular, de forma consistente, texto de qualquer sistema de escrita existente.

blocos não consecutivos³³ do DAD. Porém, possibilita o carregamento de uma lista de PChs.

4.2 Ferramentas Implementadas

Para realização dos testes de prova de conceito do método proposto nesta tese, foi necessário a implementação de três ferramentas, que foram denominadas de *CopyScan* (CS), *EraseFiles* (EF) e FiltroP (FIP), conforme serão descritas respectivamente nas Seções 4.2.1, 4.2.2 e 4.2.3.

4.2.1 Ferramenta *CopyScan*

A ferramenta CS foi desenvolvida para suprir a necessidade de uma ferramenta automática que possibilite a realização das funcionalidades de duplicação de blocos *byte-a-byte* de um DAD_O para um DAD_D ou para um arquivo de imagem e a varredura simultaneamente. Para utilização dessa ferramenta é necessário o conhecimento de três informações básicas: a lista de PChs, que contém as palavras-chave que serão procuradas, os caminhos do DAD_O e do DAD_D. A rotina principal é apresentada no fluxograma da ferramenta *CopyScan* na Figura 4.2.

³³ Disponível em: <<http://www.sleuthkit.org/informer/sleuthkit-informer-16.txt>>. Acesso em: 01 de mai. de 2014.

Figura 4.2 - Fluxo completo da Rotina principal da Ferramenta CopyScan

A ferramenta *CopyScan* lê e escreve os dados em baixo nível³⁴, isto é, abstrai os conceitos de partições e sistemas de arquivo. Esta abstração garante que o procedimento de duplicação ocorra de modo

³⁴Dados em baixo nível são dados em formato binário não estruturados, para os quais não há conhecimento sobre qual o tipo de codificação, tão pouco qual sistema de arquivos utilizado, sendo que a única referencia da localização de um *bit* é a posição em relação ao primeiro *bit*.

independente de sistemas operacionais e outras diferenças de configuração contidos no DAD a ser duplicado.

Visando permitir que a ferramenta seja executada em uma diversidade maior de plataformas e sistemas operacionais, ela foi desenvolvida em linguagem *Java*. Foram utilizados os recursos do modelo de orientação a objetos para construir uma arquitetura modular a fim de permitir a inserção de novas funcionalidades, futuramente, caso necessário.

A ferramenta implementa a duplicação de forma síncrona com as tarefas de busca por PChs. As funções de busca atuam em *threads* separadas aproveitando o fato da velocidade da memória principal ser da ordem de grandeza 10^5 vezes mais rápida que a velocidade de acesso ao DAD (memória secundária), e assim não acrescentará tempo significativo durante o procedimento de duplicação dos dados [TANENBAUM, 2010].

Os procedimentos de leitura e escrita agrupam os *bits* em blocos de tamanho que por padrão são 4096 *bytes*, mas que pode ser modificado via linha de comando, visto que o tamanho padrão dos blocos de discos magnéticos com capacidade menor que 2 TB podem ter blocos de 512 *bytes*. Desta forma, o desempenho da duplicação é otimizado considerando-se que os DADs fazem a leitura e escrita por blocos especificados na fabricação do dispositivo.

A duplicação é executada copiando-se o disco de origem *byte-a-byte* para o disco de destino, assim, o MBR (*Master Boot Record*) com a tabela de partições, os registros de *boot*, a tabela de alocação de arquivos, o início da área de dados e os *clusters* são duplicados fielmente para o disco de destino.

O fato evidencia a importância de essa característica estar presente nas ferramentas forenses, pois algumas ferramentas disponíveis fazem a cópia do DAD e desconsideram setores que

poderiam conter fragmentos de arquivos apagados, exemplo disso é a ferramenta *Norton Ghost*³⁵.

No início do procedimento de operação, a ferramenta *CopyScan* obtém os dados de entrada (palavras-chave e DADs de origem e destino) e adicionalmente é configurada por meio de um parâmetro sobre qual modo de operação executará, entre os dois possíveis: modo de Duplicação/Varredura ou somente no modo de Varredura.

4.2.1.1 Carregamento dos Dados de Entrada

Conforme mostra a Figura 4.3, no modo de Duplicação é realizada a leitura do DAD_O *byte-a-byte* e a duplicação para o DAD_D (ou para um arquivo de imagem).

O carregamento dos dados de entrada para processamento em memória, isto é, o preenchimento dos *buffers*, é realizado abrindo-se o DAD_O para leitura e copiando o conteúdo de cada bloco para memória. Esse fluxo de carregamento é ilustrado pela Figura 4.3.

³⁵ Disponível em:
<http://symantec-norton.com/Backup_e_Restaura%C3%A7%C3%A3o_c3.aspx?lang=pt-BR>.
Acesso em: 28 de abr. de 2014.

Figura 4.3 – Leitura dos dados de entrada

Os dados carregados nos *buffers* são inseridos na fila de varredura, cujos dados são denominados de elementos da fila de varredura. Esses elementos serão retirados da fila de varredura pela sub-rotina de varredura, conforme mostra a Figura 4.4.

Figura 4.4 – Armazenamento dos dados na fila de varredura

O primeiro elemento inserido na fila é a cópia idêntica do primeiro bloco do DAD. Mas a partir do segundo bloco, haverá a repetição dos últimos *q-bytes* do bloco anterior (onde *q* é o tamanho (em *bytes*) da maior PCh menos um *byte*) para permitir a procura por ocorrências de PChs entre os blocos. Por isso os primeiros *q-bytes* dos elementos seguintes serão os últimos *q-bytes* do bloco anterior.

Os dados carregados nos *buffers* também serão inseridos na fila de duplicação, onde serão chamados de elementos da fila de duplicação. Porém, este processo de inserção não pode permitir alterações nos blocos do disco, pois será o responsável pela duplicação do DAD, conforme apresentado na Figura 4.5.

Figura 4.5 – Armazenamento dos dados na fila de duplicação

Observa-se que os elementos da fila de duplicação possuem o mesmo tamanho do bloco do DAD₀. Assim, mantém a integridade do conteúdo do DAD a ser duplicado.

A partir da geração dessas filas pelo mecanismo de leitura dos dados de entrada, os mecanismos de duplicação e varredura poderão iniciar suas operações.

4.2.1.2 Carregamento da Lista de PChs

Primeiramente, a lista de palavras-chave a serem procuradas é carregada a partir de um arquivo texto intitulado `palavras.txt`³⁶ localizado no DAD (que não seja o dispositivo de origem ou destino) ou pasta onde a linha de comando para a execução do programa é invocada.

Uma maneira para ampliar as chances de encontrar as PChs procuradas foi convertê-las nos padrões de codificação mais comumente utilizados, que são: ASCII, UTF-8, UTF-16, UTF-32 e ISO-8859-1 (LATIN1).

A seguir, a Figura 4.6 ilustra o diagrama de classes utilizado na implementação da ferramenta *CopyScan*.

Figura 4.6 – Diagrama de classes da ferramenta *CopyScan*

Descrição das classes:

- *MainKeyword*: Classe que representa uma única palavra configurada no arquivo "palavras.txt" para a realização da busca. Cada palavra configurada no referido arquivo deve ser uma instância de *MainKeyword*.
- *Keyword*: Classe que representa uma das combinações possíveis da palavra representada por *MainKeyword* em todas possíveis

³⁶ Este nome pode ser modificado pelo usuário; e o conteúdo, que são as *strings*, deve estar na codificação UTF-8.

combinações de letras maiúsculas e minúsculas. Exemplo: se a palavra "larc" for uma *MainKeyword*, teremos as combinações larc, Larc, lArc, LArc, laRc, LaRc, lARc, LARc, larC, LarC, lArC, LArC, laRC, LaRC, lARC, LARC como instâncias de *Keyword*.

- *EncodedKeyword*: Classe que representa uma *Keyword* como vetor de *bytes* em uma determinada codificação (*encoding*) como ASCII e UTF-8 entre outras. Por exemplo: a combinação Larc (*Keyword*) da palavra "larc" (*MainKeyword*) pode ser representada pelo vetor de *bytes* [4C, 61, 72, 63] em hexadecimal na codificação ASCII.

Na Figura 4.7 a seguir, é mostrado o diagrama de objetos que exemplifica as classes do diagrama da Figura 4.6.

Figura 4.7 – Diagrama de objetos para combinação da palavra larc

No Diagrama da Figura 4.7, está ilustrado um exemplo da PCh "larc" (*MainKeyword*), respectivamente com suas possíveis combinações em caracteres maiúsculos e minúsculos, representadas por instâncias de *Keyword* as quais, por sua vez, são codificados em diversos vetores

de *bytes* (nas codificações ASCII, UTF-8, UTF-16, UTF-32, ISO-8859-1 (LATIN-1) representadas por instâncias de *EncodedKeyword*.

4.2.1.3 Duplicação e Varredura

Após o carregamento em memória bem sucedido das palavras a serem procuradas, inicia-se a varredura e duplicação, que ocorrem em paralelo. A ferramenta realiza a leitura de dados a partir do disco de origem ou da imagem de origem especificada como parâmetro na linha de comando. A leitura de dados é realizada em blocos do DAD e armazenada em *buffers* de memória, sendo sucessivamente realizada até que os dados de origem sejam completamente “consumidos”, como apresentado na Figura 4.3.

Cada um dos elementos obtidos é adicionado em duas filas (do tipo FIFO - *First-in - First-out*) distintas:

- ❖ Fila de duplicação;
- ❖ Fila de varredura.

Dessa forma, parte da concorrência entre o processo de varredura das PChs sobre os elementos e a gravação dos dados no DAD_D foi eliminada. O funcionamento de cada uma dessas filas é abordado em cada um dos subitens a seguir.

4.2.1.4 Módulo Duplicador

O módulo duplicador é executado em um *thread* separado (que funciona em paralelo com a varredura) e tem como objetivo consumir os elementos da fila de duplicação colocando-os em um arquivo de destino. Ele é o responsável pela realização da cópia (*byte-a-byte*) do DAD_O. O fluxograma a seguir, mostrado na Figura 4.8 ilustra seu funcionamento.

Figura 4.8 - Sub-rotina do módulo duplicador

O módulo duplicador é “*single thread*”, que assegura que os dados serão gravados na mesma ordem do DAD original.

4.2.1.5 Módulo de Varredura

O módulo de varredura é responsável por executar a varredura (inspeção) nos dados que estão sendo duplicados. Este módulo é executado em um *thread* principal separado (funcionamento paralelo) e tem como objetivo consumir os elementos da fila de varredura. O fluxograma representado na Figura 4.9 ilustra o funcionamento do módulo de varredura.

Figura 4.9 - Sub-rotina do módulo de varredura

Contudo, ele não processa diretamente os elementos obtidos da fila. Ao invés disso, ele apenas os distribui para um *pool* de *threads*, estes por sua vez realizarão a varredura. A quantidade de *threads* que compõe o referido *pool* de *threads* é inicialmente obtida como um parâmetro na linha de comando para a execução do programa. Caso tal parâmetro não seja encontrado, a quantidade de *cores* ou *threads* (o que for maior) do processador da máquina no qual a ferramenta *CopyScan* é executada determinará a quantidade de *threads* do *pool*.

Cada um dos *threads* recebe um dos elementos obtidos e realiza a busca pelas palavras configuradas usando um algoritmo de máquina de

estados. Cada palavra encontrada é relacionada na estrutura de armazenamento de resultados, a qual será processada após o término da fase de comparação e cópia.

4.2.1.6 Algoritmo de Comparação

Um diferencial da ferramenta *CopyScan* é o algoritmo de comparação. Ele permite que as palavras procuradas em diversas combinações (variações da palavra em letras maiúsculas e minúsculas) em diferentes codificações sejam buscadas simultaneamente.

Outros algoritmos varrem um elemento da fila procurando uma palavra por vez, isto é, o elemento é varrido diversas vezes com palavras diferentes.

Seguindo o modelo de máquina de estados, todas as combinações e codificações possíveis são procuradas simultaneamente, gerando uma quantidade menor de comparações, o que se traduz em menor tempo de processamento. A diminuição do tempo de processamento é uma característica imprescindível para a ferramenta, visto que a comparação deve ser concluída antes do término da cópia.

Para a construção da máquina de estados, primeiramente é necessário processar cada uma das representações binárias das palavras que são representadas por instâncias de *EncodedKeyword*, conforme foi mostrado nas Figuras 4.6 e 4.7.

No processamento dessas representações binárias, é construída uma árvore de possibilidades com diversos níveis. Nessa árvore, cada nó significa a possibilidade de ocorrência de um determinado *byte*, exceto o nó raiz, o qual representa o estado inicial (do conceito de máquina de estados). Na Figura 4.10 é mostrada a máquina de estados exemplificando a palavra “larc” com as combinações Larc e LarC nas codificações ASCII e UTF-16.

Figura 4.10 – Exemplo da máquina de estados da ferramenta *CopyScan*

Sendo assim, quando uma busca precisa ser realizada em um elemento, um *byte* por vez é inserido na máquina de estados representado pela árvore para consultar o estado final. Esse passo está ilustrado na Figura 4.9 pelas atividades: “Realizar varredura no elemento da fila a partir do índice inicial” e “Realizar varredura no elemento da fila a partir do próximo índice”.

Caso a sequência procurada atinja um nó que significa a possibilidade de uma palavra buscada, uma ocorrência é gerada com sucesso. Caso contrário, quando uma ramificação não mapeada for procurada, significa que a sequência de *bytes* consultada não representa nenhuma das palavras procuradas.

4.2.1.7 Definição do Passo de Varredura

A sub-rotina de varredura retira o elemento da fila e avalia a ocorrência das PChs (nas codificações mencionadas na Seção 3.3) *byte* por *byte* até o último *byte* do elemento. O índice do *byte* no elemento é a posição que o *byte* analisado ocupa no elemento. Esta sequência está representada na Figura 4.4.

Figura 4.11 – Exemplo de fila de varredura da ferramenta CopyScan

O algoritmo percorre o elemento da fila buscando as PChs com auxílio da máquina de estados, conforme exemplo que foi mostrado na Figura 4.10.

Conforme é ilustrado na Figura 4.12, o algoritmo compara o *byte* do elemento ❶ retirado da fila de varredura com o primeiro *byte* a ser reconhecido pela máquina de estados ❷ sequencialmente ❸ até encontrá-lo ❹, este *byte* do elemento é denominado de pivô. Depois disso, o algoritmo muda para o próximo estado ❺ e para o próximo *byte* do elemento ❻, e assim sucessivamente ❼, ❽, ❾ e ❿ até que toda a PCh seja reconhecida ⓫. Ao reconhecer a PCh, o algoritmo retorna ao estado inicial da máquina de estados ⓬ e no elemento o

algoritmo retorna para o próximo *byte* após o pivô 13. Então a varredura continua 14, 15, 16 até o final do elemento. Dessa forma, é possível demonstrar que o passo do mecanismo de varredura é de um *byte*.

Figura 4.12 – Exemplo do passo do mecanismo de varredura da *CopyScan*

4.2.1.8 Outras Funcionalidades da Ferramenta *CopyScan*

A ferramenta desenvolvida é composta por várias operações, que permitem diferentes modos de operação. Uma funcionalidade operacional importante é a possibilidade de realizar apenas a varredura, caso não seja informado o caminho do DAD ou arquivo de saída. Esta funcionalidade torna-se interessante quando, por exemplo, se deseja repetir a varredura para uma verificação de resultados. Portanto, neste caso não há interesse em realizar a cópia do dispositivo.

Outra funcionalidade de igual importância, como uma forma de verificar os resultados encontrados pela *CopyScan*, é a apresentação do trecho do disco onde a PCh foi localizada.

Esta funcionalidade indica em um arquivo de registro de eventos (*log*) a PCh encontrada e 40 *bytes* adjacentes (20 *bytes* antes e 20 *bytes*

após)³⁷ que permite a análise da localização da PCh. Por isso, esta funcionalidade foi denominada de “dedo-duro”. Dessa forma, é possível ao perito visualizar o local em que a PCh foi encontrada, ou seja, se é um trecho de dados binários ou texto plano.

A forma de utilização da ferramenta *CopyScan* é apresentado no Apêndice B.

4.2.1.9 Restrições da Ferramenta *CopyScan*

O procedimento de cópia sequencial possui restrições em comparação ao acesso aleatório, como em situações onde o dispositivo todo já está disponível para consulta. A leitura sequencial impossibilita a interpretação dos sistemas de arquivos durante a cópia, pois os sistemas de arquivos dos DADs exigem a leitura de vários blocos (conjunto de *clusters*) diferentes e não contíguos do disco para acessar um arquivo.

Durante o acesso a um arquivo, o sistema operacional consulta um índice, geralmente localizado no início ou fim de uma partição, para identificar quais blocos armazenam o conteúdo do arquivo. Nos sistemas de arquivos atuais há a possibilidade de ocorrer fragmentação, e neste caso o conteúdo do arquivo pode estar espalhado por diversos blocos disjuntos na superfície do disco.

Para interpretar um sistema de arquivos de disco em tempo real a partir de uma leitura sequencial seria necessário armazenar em memória RAM todos os blocos já lidos, para que os blocos lidos estivessem disponíveis durante o processo de leitura completa de

³⁷ Dependendo da codificação (*encoding*) dos dados em que a *string* foi encontrada, cada caractere dessa *string* pode ser representado por um determinado número de *bytes*. Por exemplo, o encoding UTF-32 utiliza 4 *bytes* (32 bits) para representar 1 caractere, UTF-16 utiliza 2 *bytes* (16 bits) para representar 1 caractere, ASCII e UTF-8 utilizam 01 *byte* (8bits) para representar um caractere. A codificação que utilizada maior quantidade de *bytes* para representar um único caractere, necessita 4 *bytes*, portanto é necessário que os trechos antes e após as PCh tenham uma quantidade múltiplo de 4. Assim com 20 *bytes* é possível representar PCh de no mínimo 5 e no máximo 20 caracteres.

arquivos. O problema neste caso é que a quantidade de memória RAM deveria ser pelo menos do mesmo tamanho que o DAD duplicado.

Do mesmo modo, a leitura sequencial não garante que os arquivos são lidos sempre do começo ao fim, e assim não é possível efetuar durante a cópia o mesmo tipo de investigação feita pelas ferramentas de perícia tradicionais que realizam a leitura e busca por PCh sobre o sistema de arquivos encontrado no dispositivo.

Os tipos de arquivos que a ferramenta *CopyScan* não reconhece são apresentados na Tabela 4.1 e os tipos de arquivos texto que podem ser reconhecidos são apresentados na Tabela 4.2.

Os arquivos com extensão .pdf evoluíram com o passar do tempo, o seu desempenho com recursos de armazenamento de dados, algoritmos de compressão para que os arquivos ficassem menores e, portanto melhores para serem distribuídos e enviados por correio eletrônico. Porém, isso gera uma grande quantidade de versões e variações do mesmo tipo de arquivo, na versão 1.2, por exemplo, iniciou-se a utilização de filtros e armazenamento dos dados em *stream* para que o programa leitor conseguisse tratar melhor as imagens contidas no arquivo.

Para armazenar esses objetos foram implementados algoritmos de compressão que além de compactar as imagens começaram a codificar os textos. Isto perdurou até a versão 1.5, quando foi introduzida a funcionalidade de fluxos de referência cruzada aumentando a quantidade de algoritmos de compressão que podem ser usados neste tipo de documento [WHITINGTON, 2011].

Tabela 4.1 - Tipos de arquivos que não podem ser "interpretados" pela ferramenta *CopyScan*.

Tipo do Arquivo	Extensão	Identificação
Chart	odc	application/vnd.oasis.opendocument.chart
Chart template	otc	application/vnd.oasis.opendocument.chart-template
Database	odb	application/vnd.sun.xml.base
Database	odb	application/vnd.oasis.opendocument.base
Database	odb	application/vnd.oasis.opendocument.database
Drawing	odg	application/vnd.oasis.opendocument.graphics
Drawing Template	otg	application/vnd.oasis.opendocument.graphics-template
Formula	odf	application/vnd.oasis.opendocument.formula
Formula template	otf	application/vnd.oasis.opendocument.formula-template
Image	odi	application/vnd.oasis.opendocument.image
Image template	oti	application/vnd.oasis.opendocument.image-template
Master Document	odm	application/vnd.oasis.opendocument.text-master
Microsoft Office Open XML Format Document	docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document
Microsoft Office Open XML Format Presentation	pptx	application/vnd.openxmlformats-officedocument.presentationml.presentation
Microsoft Office Open XML Format Presentation Slide Show	ppsx	application/vnd.openxmlformats-officedocument.presentationml.slideshow
Microsoft Office Open XML Format Presentation Template	potx	application/vnd.openxmlformats-officedocument.presentationml.template
Microsoft Office Open XML Format Spreadsheet	xlsx	application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
Microsoft Office Open XML Format Template	dotx	application/vnd.openxmlformats-officedocument.wordprocessingml.template
Microsoft Office Open XML Format Template	xltx	application/vnd.openxmlformats-officedocument.spreadsheetml.template
Portable Document Format	pdf	application/pdf
Presentation	odp	application/vnd.oasis.opendocument.presentation
Presentation Template	otp	application/vnd.oasis.opendocument.presentation-template
Spreadsheet	ods	application/vnd.oasis.opendocument.spreadsheet
Spreadsheet Template	ots	application/vnd.oasis.opendocument.spreadsheet-template
Text	odt	application/vnd.oasis.opendocument.text
OpenDocument Text Template	ott	application/vnd.oasis.opendocument.text-template
Web page template	oth	application/vnd.oasis.opendocument.text-web

Tabela 4.2 - Tipos de arquivos que podem ser "interpretados" pela ferramenta CopyScan.

Tipo do Arquivo	Extensão	Identificação
ASCII Text File Format	asc	text/plain
Binary Workbook	xlsb	application/vnd.ms-excel.sheet.binary.macroEnabled.12
C Header file	h	text/x-chdr
C Source Code	c	text/x-c
C++ Source Code	cpp	text/x-c++
Cerius2 by Accelrys Software Inc.	msi	application/x-ole-storage
Comma-Separated Variables	csv	text/csv
Compiled Help File	chm	application/vnd.ms-htmlhelp
Document Template	dot	application/msword-template
E-mail Folder	dbx	application/database
Excel Addin	xla	application/vnd.ms-excel
Excel Worksheet	xls	application/vnd.ms-excel
Extensible HyperText Markup Language File	xhtml	application/xhtml+xml
Extensible Markup Language	xml	application/xml
Google document	gdoc	application/x-gdoc
HTML File Containing Server Side Directives	shtm	wwwserver/html-ssi
HTML File Containing Server Side Directives	shtml	wwwserver/html-ssi
Hypertext Cascading Style Sheet	css	text/css
Hypertext Markup Language	html	text/html
Hypertext Markup Language	htm	text/html
Java Source Code	java	text/x-java
LaTeX Source	tex	text/x-tex
Log text file	log	text/x-log
Macro-Enabled Template	xltm	application/vnd.ms-excel.template.macroEnabled.12
Macro-Enabled Workbook	xlsm	application/vnd.ms-excel.sheet.macroEnabled.12
Markup language	mhtml	application/x-mimearchive
MHTML Document	mht	message/rfc822
Microsoft Office Open XML Format Add-in With Macros	ppam	application/vnd.ms-powerpoint.addin.macroEnabled.12
Microsoft Office Open XML Format Document with Macros Enabled	docm	application/vnd.ms-word.document.macroEnabled.12
Microsoft Office Open XML Format Presentation Slide Show with Macros Enabled	ppsm	application/vnd.ms-powerpoint.slideshow.macroEnabled.12
Microsoft Office Open XML Format Presentation Template with Macros Enabled	potm	application/vnd.ms-powerpoint.template.macroEnabled.12
Microsoft Office Open XML Format Presentation with Macros Enabled	pptm	application/vnd.ms-powerpoint.presentation.macroEnabled.12
Microsoft Office Open XML Format Template with Macros Enabled	dotm	application/vnd.ms-word.template.macroEnabled.12
Microsoft Office Open XML Formats Guide	xlam	application/vnd.ms-excel.addin.macroEnabled.12
Offline Cache	ost	text/x-ost
Outlook Personal Folder File	pst	text/x-pst
PostScript	ps	application/postscript
Power Point Presentation	ppt	application/vnd.ms-powerpoint
PowerPoint Add-in	ppa	application/vnd.ms-powerpoint
Powerpoint Template	pot	application/vnd.ms-powerpoint
Python Script or Library	py	text/x-python
README files	README	text/x-readme
Registry Data File	reg	text/x-ms-regedit
Rich Text Format File	rtf	application/rtf
Settings File (Resource Description Framework)	rdf	application/rdf+xml
S-HTML Document	sht	text/s-html
Slideshow	pps	application/vnd.ms-powerpoint
Source Language	asm	text/x-asm
Structured Query Language Data	sql	text/x-sql
Template	xlt	application/vnd.ms-excel
TeX Device Independent Document	dvi	application/x-dvi
Text File	txt	text/plain
Thumbnail Database	db	application/database
Tool Command Language	tcl	text/x-tcl
TrueType Compressed Font	ttc	text/ttcf
TrueType Font	ttf	text/ttf
Visual Basic Extension	vbx	text/plain
Word Document	doc	application/msword
XML Paper Specification	xps	application/oxps
Outros tipos de arquivos texto	***	text/*

Vale ressaltar que as restrições apresentadas, não invalidam a ferramenta CS e nem o método proposto, pois apesar desses arquivos que apresentam algum tipo de compressão, outros tipos de arquivos podem ser lidos e interpretados pela ferramenta.

4.2.2 Ferramenta FiltroP

A ferramenta denominada de FiltroP (FIP) é uma aplicação escrita em *Python* que tem o objetivo de verificar se nas vizinhanças da PCh no *Log* de dedo-duro, gerado pela ferramenta CS, há apenas caracteres de natureza textual (letras, números e determinados símbolos especiais utilizados em documentos texto, e-mails) ou não. Se as vizinhanças forem constituídas apenas por caracteres de natureza textual, então a ocorrência para a PCh será classificada como verdadeiro-positivo (VP), caso contrário, será classificada como falso-positivo (FP).

4.2.3 Ferramenta *EraseFiles*

A ferramenta *EraseFiles* (EF) é uma aplicação desenvolvida em *Java* (versão 7) que tem duas funções principais:

- ❖ Criar um arquivo do tamanho do espaço livre no disco contendo apenas caracteres zero.
- ❖ Identificar e abrir os tipos de arquivos: binário, interpretáveis ou mistos e preencher seu respectivo conteúdo com o caracteres zero (*wipe*).

A identificação dos tipo de arquivos ocorre por meio do cabeçalho e rodapé do arquivo (por meio da execução da função *Java probeContentType*) podendo ser dos tipos:

- ❖ Arquivo binário: engloba os arquivos do tipo “*application/octet-stream*”, imagem, vídeo e áudio;

- ❖ Arquivo Misto ou texto compactado: são os arquivos com extensão PDF e da família XML-Based do Microsoft Office ou Libre Office que seriam dos tipos ODT, DOCX, XLSX;
- ❖ Arquivo Interpretáveis: são os arquivos listados conforme apresentado na Tabela 4.2.

Dessa forma, é possível isolar as ocorrências de cada PCh por “parte” do DAD, conforme mostra a Figura 4.15.

4.3 Processos Realizados Pelas Ferramentas *CopyScan* e *FiltroP*

Os processos dos métodos cobertos pela ferramenta *CopyScan* são duplicação e varredura. Os resultados dessa ferramenta são as duplicações dos DADs. A tabela com a lista de ocorrências e a tabela de Log de localização das PChs no DAD, que denominamos de Log Dedo-duro. A ferramenta *FiltroP* a partir do Log Dedo-duro identifica quais ocorrências são VP e FP, conforme os critérios já descritos na Seção 3.4.2. Os processos cobertos por essas ferramentas são ilustrados na Figura 4.13.

Figura 4.13 – Ilustração dos processos cobertos pelas ferramentas CS e FIP

4.4 Procedimentos para Avaliação dos Processos de Varredura e Filtragem

Inicialmente, considera-se que um DAD esteja dividido em duas partes principais: Espaço ocupado por dados (*clusters* alocados (CA)) e Espaço livre (EL) (*clusters* não alocados), que está livre para gravação de dados, mas que pode conter fragmentos de dados de arquivos, conforme mostra a Figura 4.14.

Figura 4.14 – Clusters alocados e clusters não alocados

Considerando uma abstração, suponha que os *clusters* alocados sejam ocupados por arquivos que possam ser classificados como: texto simples (*plain text*), documentos baseados no formato de linguagem de marcação extensível (*eXtensible Markup Language* (XML)), documento de formato portátil (*portable documents files* (pdf)), binários, compactados e cifrados.

Considere ainda que o espaço ocupado possa ser dividido em três partes e os arquivos que ocupam os *clusters* alocados possam ser classificados e distribuídos em cada uma das três partes como: Arquivos Binários (AB), Arquivos Interpretáveis (AI) e Arquivos Mistos (AM).

Dessa forma, o DAD fica formado por quatro partes: AB, AI, AM e EL, conforme representado na Figura 4.15.

Figura 4.15 – Alusão às quatro partes do DAD

Cada uma dessas classificações são descritas a seguir, conforme as quatro partes consideradas e como os diferentes tipos de arquivos mencionados anteriormente podem ser distribuídos em cada uma dessas partes.

- ❖ AB: Não contém em sua maior parte arquivos em linguagem natural (que não podem ser "legível por humanos"). Exemplos: arquivos cifrados (.tc), .dat, .bin, .mov, .mp3, .jpeg, .tif, .exe, .dll, .tc, .arj., .zip., .gz, dentre outros tipos de arquivos.
- ❖ AI: texto simples - Contém em sua maior parte arquivos em linguagem natural (que podem ser "legível por humanos"). Exemplos: .txt, .doc., .tex., .xls, .ps, .html, entre outros tipos de arquivos.
- ❖ AM: arquivos de documentos baseados no formato XML e pdf.
- ❖ EL: Não contém dados (está disponível para gravação de dados), mas pode conter fragmentos de arquivos apagados (partes de arquivos que foram apagados, mas não removidos completamente do DAD). Os fragmentos de arquivos remanescentes nos clusters não alocados podem conter dados de interesse para a perícia.

4.4.1 Verificação de Ocorrências em Cada Parte do DAD

Durante a varredura do DAD, quando há a identificação de uma ocorrência de PCh por uma ferramenta, que neste caso é a ferramenta CS, faz-se necessário identificar se esta ocorrência é VP ou FP, conforme critérios explicados nas Seção 3.4.2.

A identificação de PChs no conteúdo de arquivos de áudio, vídeo, imagens, som, executáveis, compactados dos tipos (*.zip, .arj, *.rar, .tar) e cifrados, tem grande chance de ser uma simples coincidência na correspondência da sequência de caracteres das PChs procuradas, e dessa forma, ser um FP. Por esse motivo os resultados da tabela de *Log* de localização das PChs no DAD (que foi apelidada de "*Log* dedo-duro") são submetidos ao processo de filtragem. Dessa forma, definimos que:

- ❖ QO_{DT} : é o total da quantidade de ocorrências de PChs encontradas no DAD todo, obtido por meio do *Log* dedo-duro da ferramenta CS.
- ❖ Os resultados da ferramenta FiltroP são a quantidade de ocorrências de PChs encontradas classificadas como VP ($QOVP$) e a quantidade de ocorrências de PChs encontradas classificadas como FP ($QOFP$).

Mediante o exposto, deseja-se avaliar a convergência dos resultados do *Log* dedo-duro da ferramenta CS, bem como os resultados da ferramenta FiltroP, tendo como métrica a quantidade de ocorrências de PChs encontradas (QO).

Para isso, serão comparados os resultados da quantidade de ocorrências de PChs encontradas: no DAD todo (QO_{DT}), VP no DAD todo ($QOVP_{DT}$), Arquivos Interpretáveis (QO_{AI}) e VP nos Arquivos Interpretáveis ($QOVP_{AI}$).

4.4.1.1 Verificação de Ocorrências VP e FP em Cada Parte do DAD

Esta seção descreve os procedimentos que foram adotados para verificação dos resultados do *Log* dedo-duro da ferramenta CS, bem como os resultados da ferramenta FiltroP.

O intuito é realizar os processos de duplicação e varredura do DAD utilizando a ferramenta CS e obter os resultados do *Log* dedo-duro, identificado na Figura 4.16 por R_1 . O resultado R_1 é submetido ao processo de filtragem, obtendo-se assim o resultado R_2 , que é a quantidade de ocorrências classificadas como VP.

Figura 4.16 – Comparação entre resultados

Posteriormente, cada uma das quatro partes do DAD são "isoladas", ou seja, em cada parte é realizada a varredura individualmente para obter a quantidade de ocorrências de PChs, identificados na Figura 4.16 por R_3 . Observe que para cada parte do DAD, também são obtidos dois resultados, um que é proveniente do *Log* dedo-duro e outro da filtragem, como foi feito para o DAD todo. Em

cada parte do DAD são identificadas as respectivas quantidades de ocorrências de VP.

Ao final da varredura de todas as partes do DAD, são somadas as quantidades de ocorrências de VP obtidas. Então, essa quantidade é comparada com a quantidade de ocorrências VP obtidas em R2.

4.4.1.2 Verificação de Ocorrências no DAD Todo

Considerando que o DAD original não deve ser modificado, primeiramente é realizada a duplicação deste DAD. Então, sobre essa duplicação por meio da ferramenta CS é realizada a primeira varredura no DAD. Conforme ilustra a Figura 4.17.

Figura 4.17 – Primeira varredura

O resultado obtido dessa varredura é o *Log* do dedo-duro com a quantidade de ocorrências no DAD todo (QO_{DT}). Submetendo-se QO_{DT} ao processo de filtragem, são obtidas as quantidades de ocorrências classificadas como verdadeiro-positivos para o DAD todo ($QOVP_{DT}$) e as quantidades de ocorrências classificadas como falso-positivo para o DAD todo ($QOFP_{DT}$). QO_{DT} é calculada conforme a equação 2.

$$QO_{DT} = QOVP_{DT} + QOFP_{DT} \quad (2)$$

Observe que QO_{DT} também pode ser obtida pela somatória das ocorrências em cada uma das partes principais do DAD, conforme a equação 3.

$$QO_{DT} = QO_{EL} + QO_{CA} \quad (3)$$

Mas,

$$QO_{CA} = QO_{AB} + QO_{AI} + QO_{AM} \quad (3.1)$$

Então,

$$QO_{DT} = QO_{EL} + QO_{AB} + QO_{AI} + QO_{AM} \quad (4)$$

Onde:

- QO_{AA} - quantidade de ocorrências de PChs nos arquivos Alocado
- QO_{AB} - quantidade de ocorrências de PChs nos arquivos Binários
- QO_{AI} - quantidade de ocorrências de PChs nos arquivos Interpretáveis
- QO_{AM} - quantidade de ocorrências de PChs nos arquivos Mistos
- QO_{DT} - quantidade de ocorrências de PChs no DAD todo
- QO_{EL} - quantidade de ocorrências de PChs no espaço Livre

4.4.1.3 Verificação de Ocorrências em Arquivos nos *Clusters* Alocados

Primeiramente, é verificado o tamanho do espaço livre (EL) nos *clusters* não alocados do DAD por meio da função do Java (*getFreeSpace*) da classe *File*. Então, por meio da ferramenta *EraseFiles*, o EL é sobrescrito com zeros. Esta ação é realizada para que não permaneçam quaisquer fragmentos de arquivos no EL, conforme ilustrado na Figura 4.18.

Figura 4.18 – Segunda varredura

Por meio da ferramenta CS, é realizada a segunda varredura, obtendo-se o resultado das ocorrências de PChs em arquivos nos *clusters* alocados (QO_{AA}), ou seja, nas partes do DAD identificadas na figura por Arquivos Binários (AB), Arquivos Interpretáveis (AI) e Arquivos Mistos (AM).

Dessa forma, a quantidade de ocorrências no espaço livre (QO_{EL}) pode ser obtida subtraindo-se o resultado da primeira varredura (QO_{DT}) do resultado da segunda varredura (QO_{AA}).

$$QO_{EL} = QO_{DT} - QO_{AA} \quad (5)$$

O resultado do *Log* do dedo-duro dessa varredura é submetido ao processo de filtragem e com isso, são obtidas as ocorrências classificadas como verdadeiro-positivos ($QOVP_{AA}$) e as ocorrências classificadas como falso-positivo ($QOFP_{AA}$), apenas sobre o conteúdo dos arquivos nos *clusters* alocados.

$$QO_{AA} = QOVP_{AA} + QOFP_{AA} \quad (6)$$

Entretanto, não se tem o resultado do *Log* do dedo-duro para o EL, por isso, não é possível obter os valores de $QOVP_{EL}$ e $QOFP_{EL}$ por meio do FiltroP. Dessa forma, os valores de $QOVP_{EL}$ e $QOFP_{EL}$ serão calculados da seguinte forma:

Subtraindo-se (4) de (6):

$$QO_{DT} = QOVP_{DT} + QOFP_{DT}$$

$$QO_{AA} = QOVP_{AA} + QOFP_{AA}$$

Fica:

$$QO_{DT} - QO_{AA} = QOVP_{DT} + QOFP_{DT} - QOVP_{AA} - FP_{AA}$$

$$QO_{DT} - QO_{AA} = QOVP_{DT} - QOVP_{AA} + QOFP_{DT} - QOFP_{AA}$$

De (5), sabemos que $QO_{EL} = QO_{DT} - QO_{AA}$, então:

$$QO_{EL} = QOVP_{EL} + QOFP_{EL}$$

Logo:

$$QOVP_{EL} = QOVP_{DT} - QOVP_{AA} \quad (7)$$

$$QOFP_{EL} = QOFP_{DT} - QOFP_{AA} \quad (8)$$

Nas partes onde não se tem o Log dedo-duro individualmente, não é possível se obter os valores de QOVP e QFP diretamente. Assim, para as partes AB e AM os respectivos valores de QOVP e QOFP serão obtidos de maneira análoga.

4.4.1.4 Verificação de Ocorrências em Arquivos Interpretáveis e Mistos

Por meio da ferramenta *EraseFiles* os Arquivos Binários (AB) são sobrescritos com zeros, conforme representado na Figura 4.19.

Figura 4.19 – Terceira varredura

Em seguida, é realizada a terceira varredura, obtendo-se o resultado da quantidade das ocorrências de PChs em AI e AM, ou seja, QO_{IM} (Quantidade de Ocorrências de PChs nos Arquivos Interpretáveis e Mistos).

Com isso, é possível obter a quantidade de ocorrências no espaço dos arquivos binários (QO_{AB}), subtraindo-se o resultado da segunda varredura (QO_{AA}) do resultado da terceira varredura (QO_{IM}).

$$QO_{AB} = QO_{AA} - QO_{IM} \quad (9)$$

Esse resultado é submetido ao processo de filtragem e com isso são obtidas as ocorrências classificadas como verdadeiro-positivos (VP_{IM}) e as ocorrências classificadas como falso-positivo (FP_{IM}), apenas sobre o conteúdo dos arquivos Interpretáveis e Mistos.

$$QO_{IM} = QOVP_{IM} + QOFP_{IM} \quad (10)$$

Os valores de $QOVP_{AB}$ e $QOFP_{AB}$ podem ser obtidos subtraindo-se (6) de (10):

$$QOVP_{AB} = QOVP_{AA} - QOVP_{IM} \quad (11)$$

$$QOFP_{AB} = QOFP_{AA} - QOFP_{IM} \quad (12)$$

4.4.1.5 Verificação de Ocorrências em Arquivos Interpretáveis

Primeiramente, por meio da ferramenta *EraseFiles* os arquivos Mistos (AM) são sobrescritos com zeros, conforme representado na Figura 4.20.

Figura 4.20 – Quarta varredura

Em seguida, é realizada a quarta varredura, obtendo-se o resultado da quantidade das ocorrências de PChs em arquivos Interpretáveis (QO_{AI}).

Com isso, é possível obter a quantidade de ocorrências de PChs no espaço dos arquivos Mistos (QO_{AM}), subtraindo-se o resultado da terceira varredura (QO_{IM}) do resultado da quarta varredura (QO_{AI}):

$$QO_{AM} = QO_{IM} - QO_{AI} \quad (13)$$

O resultado do *Log* do dedo-duro da varredura sobre os arquivos Interpretáveis é submetido ao processo de filtragem e com isso, são obtidas as ocorrências classificadas como verdadeiro-positivos ($QOVP_{AI}$) e as ocorrências classificadas como falso-positivo ($QOFP_{AI}$), apenas sobre o conteúdo dos arquivos interpretáveis (QO_{AI}).

$$QO_{AI} = QOVP_{AI} + QOFP_{AI} \quad (14)$$

As ocorrências classificadas como verdadeiro-positivos ($QOVP_{AM}$) e as ocorrências classificadas como falso-positivo ($QOFP_{AM}$), podem ser obtidos subtraindo-se (10) de (14):

$$QOVP_{AM} = QOVP_{IM} - QOVP_{AI} \quad (15)$$

$$QOFP_{AM} = QOFP_{IM} - QOFP_{AI} \quad (16)$$

Logo,

$$QO_{AM} = QVP_{AM} + QFP_{AM} \quad (17)$$

4.4.1.6 Considerações Sobre os Procedimentos para Avaliação dos Processos de Varredura e Filtragem

A partir dos resultados que devem ser obtidos por meio dos procedimentos para avaliação descritos anteriormente, será possível realizar comparações da quantidade de ocorrências de PChs entre: o DAD todo, QOVP do DAD todo, total de Arquivos Interpretáveis (QO_{AI}) e VP de Arquivos Interpretáveis ($QOVP_{AI}$).

A partir desses resultados, será possível verificar a convergência entre as quantidades de ocorrências de PChs obtidas em Arquivos Interpretáveis (QO_{AI} e $QOVP_{AI}$) e no DAD todo (QO_{DT} e $QOVP_{DT}$), possibilitando assim, determinar se para o propósito do método proposto nesta tese os resultados provenientes de QO_{DT} e $QOVP_{DT}$ são suficientes para executar os processos de triagem e ranqueamento.

No Capítulo 5, o **Experimento 4** será utilizado para avaliação dos procedimentos de validação dos processos de varredura e filtragem.

4.5 Restrição Para Validação da Triagem

Em razão das características da Triagem, que depende dos critérios adotados pelo perito para ser realizada, não cabe estabelecer uma regra para validação de resultados da triagem, visto que isso

poderia influenciar na decisão do perito e dessa forma interferir diretamente nos resultados do ranqueamento.

4.6 Validação da Sequência de Ranqueamento dos DADs

É importante para constatar se o ranqueamento dos DADs indicados como mais relevantes para serem priorizados na etapa de análise correspondem de fato aos DADs com maiores chances de conter os vestígios de interesse para a perícia.

Esta comparação será realizada por meio da ferramenta *Autopsy*, tradicionalmente utilizada na área de perícia de informática.

4.7 Considerações Finais Sobre o Método Proposto

Este capítulo apresentou o método para triagem e ranqueamento de computadores, mostrando que de maneira o mesmo se encaixa no modelo utilizado como referência. Abordou também os critérios para o processo de triagem, para o processo de ranqueamento, ferramentas testadas, requisitos necessários, as especificações e implementação de um conjunto de ferramentas para que o método proposto pudesse ser validado. Para consolidar o método, também foi especificado como os resultados dos processo de varredura e ranqueamento são validados.

Pela característica deste capítulo ser em sua maior parte teórico, já que descreve o método proposto, a parte experimental foi organizada no Capítulo 5. Os resultados obtidos no Capítulo 5, servirão para validar o método proposto.

5 EXPERIMENTOS E TESTES

Este capítulo descreve os experimentos que foram realizados com o intuito de verificar a hipótese (de que na etapa de coleta, seja possível realizar simultaneamente à duplicação do DAD de origem para o DAD de destino a varredura dos dados para procurar por palavras-chave de interesse do perito, sem com isso impactar significativamente o tempo de duplicação) e validar o método proposto, que consiste em, a partir das PChs obtidas, realizar a triagem e ranqueamento dos DADs duplicados, de forma a apresentar uma sugestão de sequência, indicando quais desses DADs deverão ser prioritariamente examinados na etapa de análise.

Serão realizados 6 experimentos com o objetivo de avaliar o método proposto. Esses experimentos serão apresentados e descritos detalhadamente a seguir.

5.1 Considerações Iniciais Sobre os Experimentos

Uma das maiores dificuldades enfrentadas no decorrer do trabalho foi a obtenção de base de dados, ou seja, DADs contendo informações que pudessem ser utilizadas nos experimentos para estudo de caso, validação e prova de conceito do método proposto.

A dificuldade pode ser justificada pelo fato de que a perícia judicial de informática normalmente envolve dados de pessoas físicas ou jurídicas, muitas vezes considerados confidenciais, que normalmente por restrições de privacidade e sigilo, se recusam a disponibilizar os respectivos dados. Outro fator que também deve ser considerado é que conjuntos de dados mais acessíveis ao público são muito limitados no que tange à quantidade e volume de dados.

Adicionalmente, um aspecto não menos importante está relacionado com a possibilidade da reprodutibilidade dos ensaios por outros pesquisadores.

Para realização dos testes iniciais para análise de desempenho foram utilizados discos contendo dados reais dos laboratórios didáticos de redes de computadores do Departamento de Engenharia de Computação e Sistemas Digitais PCS/EPUSP. Porém, devido à capacidade de armazenamento dos DADs, que variam de 500 GB a 1 TB, o tempo gasto para realizar os primeiros testes de duplicação necessários foram relativamente altos. Considerando que as duplicações dos DADs teriam que ser repetidas por diversas vezes, o tempo necessário para duplicação, tornou-se uma operação exaustiva, o que impactou na realização dos testes iniciais. Observando que os primeiros discos utilizados possuíam 512 bytes por setor, em diversos testes realizados a duplicação de um disco com capacidade de 1 TB demorava aproximadamente 8 horas.

Observou-se que para realizar os experimentos necessários para prova de conceito seria mais viável trabalhar com DADs de menor capacidade de armazenamento, e conseqüentemente com menor quantidade de dados, devido ao tempo necessário para a duplicação

dos DADs. Porém, sem desconsiderar duas características importantes, a quantidade de DADs e a diversidade no conteúdo dos dados desses DADs.

Assim, para contornar as dificuldades anteriormente descritas, procuramos por bases públicas de dados que disponibilizassem imagens de DADs e que pudessem ser utilizadas na realização dos experimentos necessários. O conjunto mais completo de imagens e que constitui uma das únicas bases de dados públicas que pode ser utilizada sem restrições ou permissões de uso, denominado “Corpo Real de Dados Forense”³⁸, foi criado pela Escola Naval de Pós-graduação (*Naval Postgraduate School* - NPS) da Marinha dos Estados Unidos. Essas imagens foram criadas para finalidades de pesquisas (desenvolvimento de novas tecnologias) e educacionais (treinamentos na área de investigação digital).

Os dados para a criação do “Corpo Real de Dados Forense” foram provenientes de mais de 2000 discos rígidos usados que foram comprados no mercado secundário (*sites* como o *Ebay* e lojas de revendas de computadores usados) [GARFINKEL et al., 2009; WOODS, 2011].

Dentre as imagens disponíveis, decidiu-se utilizar nos experimentos e testes iniciais o *corpus* de dados M57-Patents³⁹ [WOODS, 2011]. Trata-se de um conjunto com 79 arquivos de imagens de discos, cujo tamanho dos arquivos de imagens varia entre 10 e 40 GB. Este conjunto de arquivos de imagens apresentam algumas características importantes para o presente trabalho, tais como, a quantidade de imagens e o tamanho relativamente menor destas, quando comparadas com a capacidade dos discos que estavam sendo

³⁸DigitalCorpora.org é um *website* que contém diversos arquivos de imagens digitais para uso em pesquisa educacional forense. Todos os arquivos das imagens de DADs, *dumps* de memória e capturas de pacotes de rede estão disponibilizados livremente e disponíveis, podem ser utilizados sem autorização prévia ou aprovação IRB (*Institutional Review Board*). Disponível em: <<http://digitalcorpora.org/corp/nps/>>. Acesso em: 01 Ago. 2014.

³⁹Disponível em: <<http://digitalcorpora.org/corpora/scenarios/m57-patents-scenario>>. Acesso em: 01 Ago. 2014.

utilizados inicialmente, propiciando que os experimentos não fossem demasiadamente demorados, e assim, pudessem ser realizados em um espaço de tempo mais viável. Além disso, propicia que os experimentos realizados possam ser reproduzidos por pessoas interessadas. Vale destacar que embora o tamanho das imagens sejam menores, quando comparado com os testes preliminares, isso não prejudica em nada a generalização dos resultados obtidos.

5.1.1 Descrição do Cenário e do Conjunto de Dados M57-*Patents*

O cenário simula dados do mundo real, porém não contém informações que podem ser consideradas particulares e sensíveis. Entretanto, apresentam complexidade e desafios genuínos. No cenário é representada a companhia M57-*Patents*, especializada em realizar pesquisas sobre informações de patentes para seus clientes. A empresa iniciou suas atividades em uma sexta-feira, dia 13 de Novembro de 2009 e foi encerrada no sábado, dia 12 de Dezembro de 2009. Foram coletados dados durante 18 dias, no período compreendido entre 12 de Novembro e 11 de Dezembro de 2009.

O quadro de funcionários da companhia é formado por quatro pessoas: Pat (*Chief Executive Officer* - CEO), Terry Administrador de Tecnologia da Informação (TI), Jo e Charlie pesquisadores de patentes, que estão envolvidos em uma variedade de atividades.

No cenário, cujo principal propósito é simular as interações com o mundo real, os funcionários interagem com várias outras “pessoas” que estão fora do ambiente da companhia M57. Essas pessoas representam clientes, amigos e outros indivíduos que mantêm algum contato com os funcionários da M57. As informações sobre os dados que compõem o cenário são:

- ❖ Imagens dos discos
- ❖ Imagem instantânea da memória RAM
- ❖ Tráfego da Rede
- ❖ Imagem de *pendrives*
- ❖ Conjunto de fotos e vídeos

Para os experimentos deste trabalho serão utilizadas somente as imagens dos discos, devido às características dos testes, que envolve a duplicação de DADs.

5.1.1.1 Imagens dos DADs que Compõem o Cenário M57-*Patents*

A seguir na Tabela 5.1, é apresentada a relação com as principais informações para cada arquivo de imagem de DAD que compõem o cenário M57-*Patents*. Tem-se o número da imagem do DAD (Nº), o Nome e o tamanho (T) em GigaByte (GB) de cada um deles. Também são mostrados os respectivos tamanhos de cada arquivo no formato .E01 (compactado) e .img (descompactado).

Tabela 5.1 – Informações sobre os arquivos de imagem utilizados

Nº	Nome	.E01	.img	Nº	Nome	.E01	.img
		(GB)	(GB)			(GB)	(GB)
#01	charlie-2009-11-12	3,0	9,6	#41	41_pat-2009-11-12	3,0	13
#02	02_charlie-2009-11-12start	3,0	9,6	#42	42_pat-2009-11-12start	3,0	13
#03	03_charlie-2009-11-16	3,4	9,6	#43	43_pat-2009-11-16	3,5	13
#04	04_charlie-2009-11-17	3,4	9,6	#44	44_pat-2009-11-17	3,5	13
#05	05_charlie-2009-11-18	3,8	9,6	#45	45_pat-2009-11-18	3,8	13
#06	06_charlie-2009-11-19	3,8	9,6	#46	46_pat-2009-11-19	3,8	13
#07	07_charlie-2009-11-20	3,9	9,6	#47	47_pat-2009-11-20	3,8	13
#08	08_charlie-2009-11-23	4,0	9,6	#48	48_pat-2009-11-23	4,1	13
#09	09_charlie-2009-11-24	4,0	9,6	#49	49_pat-2009-11-24	4,4	13
#10	10_charlie-2009-11-30	4,0	9,6	#50	50_pat-2009-11-30	5,0	13
#11	11_charlie-2009-12-01	4,0	9,6	#51	51_pat-2009-12-01	5,0	13
#12	12_charlie-2009-12-02	4,0	9,6	#52	52_pat-2009-12-02	5,0	13
#13	13_charlie-2009-12-03	4,2	9,6	#53	53_pat-2009-12-03	5,0	13
#14	14_charlie-2009-12-04	4,2	9,6	#54	54_pat-2009-12-04	5,0	13
#15	15_charlie-2009-12-07	4,0	9,6	#55	55_pat-2009-12-07	5,4	13
#16	16_charlie-2009-12-08	4,2	9,6	#56	56_pat-2009-12-08	5,6	13
#17	17_charlie-2009-12-09	4,2	9,6	#57	57_pat-2009-12-09	5,6	13
#18	18_charlie-2009-12-10	3,6	9,6	#58	58_pat-2009-12-10	5,7	13
#19	19_charlie-2009-12-11	3,6	9,6	#59	59_pat-2009-12-11	5,7	13
#20	20_jo-2009-11-12	3,0	13	#60	60_terry-2009-11-12	7,0	13
#21	21_jo-2009-11-12start	3,0	13	#61	61_terry-2009-11-12start	6,5	20
#22	22_jo-2009-11-16	3,3	13	#62	62_terry-2009-11-16	7,3	20
#23	23_jo-2009-11-17	3,4	13	#63	63_terry-2009-11-17	7,4	20
#24	24_jo-2009-11-18	3,6	13	#64	64_terry-2009-11-18	7,4	20
#25	25_jo-2009-11-19	3,6	13	#65	65_terry-2009-11-19	7,2	20
#26	26_jo-2009-11-20-newComputer	2,2	15	#66	66_terry-2009-11-20	7,9	39
#27	27_jo-2009-11-20-oldComputer	3,7	13	#67	67_terry-2009-11-23	9,9	39
#28	28_jo-2009-11-23	2,8	15	#68	68_terry-2009-11-24	9,7	39
#29	29_jo-2009-11-24	3,0	15	#69	69_terry-2009-11-30	9,6	39
#30	30_jo-2009-11-30	3,8	15	#70	70_terry-2009-12-01	9,9	39
#31	31_jo-2009-12-01	3,9	15	#71	71_terry-2009-12-02	9,7	39
#32	32_jo-2009-12-02	3,9	15	#72	72_terry-2009-12-03	9,9	39
#33	33_jo-2009-12-03	4,5	15	#73	73_terry-2009-12-04	9,9	39
#34	34jo-2009-12-04	4,5	15	#74	74_terry-2009-12-07	9,9	39
#35	35_jo-2009-12-07	4,8	15	#75	75_terry-2009-12-08	9,9	39
#36	36_jo-2009-12-08	4,8	15	#76	76_terry-2009-12-09	10,0	39
#37	37jo-2009-12-09	5,1	15	#77	77_terry-2009-12-10	10,0	39
#38	38_jo-2009-12-10	5,1	15	#78	78_terry-2009-12-11-001	10,0	39
#39	39_jo-2009-12-11-001	5,5	15	#79	79_terry-2009-12-11-002	10,0	39
#40	40_jo-2009-12-11-002	5,5	15				

Por questões de praticidade, algumas vezes ao invés do nome completo do arquivo de imagem é utilizado o respectivo número para fazer menção ao DAD no corpo do texto deste trabalho.

O DAD da estação de trabalho de cada um dos funcionários da companhia M57-Patents foi duplicado diariamente nos dias mostrados na Tabela 5.2. Estas duplicações foram compactadas para o formato .E01.

Tabela 5.2 – Relação de dias em que os DADs dos funcionários foram duplicados.

Mês	Dias
Novembro	12, 16, 17, 18, 19, 20, 23, 24, 30
Dezembro	01, 02, 03, 04, 07, 08, 09, 10, 11

Todas as imagens⁴⁰ dos discos rígidos utilizam o sistema de arquivos NTFS, variando no tamanho entre 9,6 e 39 GB (sem compactação). Ao todo o conjunto é formado por 79 imagens, totalizando 1,5 TB de dados (460 GB compactados).

5.1.3 Equipamento Utilizado e DADs de Origem e Destino

O equipamento principal utilizado para realização dos testes possui a configuração apresentada abaixo. Dessa forma, todos os tempos medidos será em relação à este equipamento.

- ❖ Placa mãe: Intel modelo DX58SO-AAE29331-703.
- ❖ Processador Intel Core i7 - 950 (Octa-Core) - 3.07 GHz.
- ❖ 16 GB de memória RAM - DDR3 - 665 MHz - *Dual Channel*.
- ❖ Placa controladora de disco SATA 3 Gbits/s.
- ❖ 02 DADs SATA 1 TB - marca: Seagate - modelo ST31000528AS - SATA, 7200 rpm, bs: 4096 bytes.
- ❖ 02 DADs SATA 2 TB - marca: Seagate - modelo ST2000DM001 - SATA, 7200 rpm, bs: 4096 bytes.

⁴⁰ Disponível em: <<http://digitalcorpora.org/corpora/disk-images/nps-2009-m57-patents-redacted/>>. Acesso em: 18 de Março de 2015.

5.1.3.1 Dispositivos de Origem (DAD_O) e Destino (DAD_D)

Os principais discos utilizados foram:

- ❖ 02 DADs de 500 GB
 - Marca: Seagate, SATA, modelo: ST500DM002, número de série: S2A6GC3S, SATA, número do produto: 1BD142-302, Date code: 13035, 7200 rpm, bs: 4096 bytes.
 - Marca: Seagate, SATA, modelo: ST500DM002, número de série: S2A6A4QX, número do produto: 1BD142-302, Date code: 13023, 7200 rpm, bs: 4096 bytes.
- ❖ 01 DAD de 160 GB
 - Marca: Samsung, SATA, modelo: HD161HJ, 7200 rpm
- ❖ 01 DAD de 80 GB
 - Marca: Samsung, SATA, modelo: HD080HJ, 7200 rpm
- ❖ 01 DAD de 80 GB
 - Marca: Samsung, SATA, modelo: SP0842N, 7200 rpm

5.1.4 Esterilização dos DADs de Origem e Destino

Antes da utilização dos DAD_O e DAD_D em cada teste de duplicação ambos são sobrescritos com zeros ("*wipe*"). O propósito desta ação é apagar todo e qualquer dado ou resquício de fragmento de dados de arquivos contidos nesses dispositivos, garantindo que os dados do arquivo de imagem que serão abertos nesses DADs não sejam “contaminados” com quaisquer dados que estavam armazenados anteriormente. Exemplo da utilização deste comando será mostrado no item 5.1.4.1.

5.1.4.1 Exemplo de Esterilização dos DADs de Origem e Destino

Conforme descrito na Seção 5.1.4 esta seção ilustra a linha de comando utilizada para invocar a ferramenta DD para realizar o procedimento de esterilização dos discos. Este procedimento é ilustrado na Figura 5.1.

Figura 5.1 – Exemplo de sobrescrita de DAD a partir da ferramenta DD.

```

Terminal
root@mba2:~# time dd if=/dev/zero bs=4096 count=2499966 of=/dev/sdc
2499966+0 registros de entrada
2499966+0 registros de saída
10239860736 bytes (10 GB) copiados, 72,8383 s, 141 MB/s

real    1m12.840s
user    0m0.164s
sys     0m7.972s
root@mba2:~#

```

Respectivamente os parâmetros utilizados na linha de comando para sobrescrever o DAD, são explicados:

- ❖ `if`: é o arquivo de entrada (*input file*).
- ❖ `/dev/zero`: é o mapeamento que retorna valor zero.
- ❖ `bs`: é o tamanho do bloco (*block size*), que para este DAD é igual a 4096 bytes.
- ❖ `count`: é o número de blocos que serão escritos no DAD. O valor é obtido tomando-se o tamanho do DAD em bytes e dividindo-o pelo tamanho do bloco (`bs`):

$$\text{count} = \frac{\text{Tamanho do DAD (em bytes)}}{\text{bs}} \quad (18)$$

Para este exemplo temos:

$$\text{count} = \frac{10239860736}{4096} = 2499966$$

- ❖ `of`: é o arquivo de saída (*output file*), neste caso refere-se ao DAD que será sobrescrito, e está mapeado no diretório `/dev/sdc`.

- ❖ Os valores de registro de entrada e de saída referem-se à quantidade de blocos de 4096 bytes sobrescritos.

No DD, a conversão de bytes para gigabytes é efetuada considerando que 01 byte equivale a um bilhão (1000^3) de bytes e não 1024^3 bytes.

5.2 Experimento 1

O objetivo deste experimento é medir o tempo de duplicação (sem realizar a varredura para procurar por PChs) das ferramentas *CopyScan* e DD para DADs com diferentes capacidades. Então, os tempos de duplicação de ambas as ferramentas são comparados. É esperado que o tempo de duplicação entre ambas as ferramentas sejam semelhantes. O experimento é dividido em duas partes: A e B, conforme descritas a seguir.

- ❖ Na parte A, são utilizados DADs com capacidades de: 80 GB, 160 GB, 500 GB e 1 TB. Estes discos utilizam tamanho de blocos de 4096 bytes.
- ❖ Na parte B, são utilizados os arquivos de imagem do cenário *M57-Patents*, conforme descrito na Seção 5.1.

5.2.1 Experimento 1 - Parte A

5.2.1.1 Preparação e Procedimentos

A preparação para esta parte do experimento pode ser resumida nos seguintes procedimentos:

- a) Definir o Equipamento a ser utilizado (configuração do equipamento a ser utilizado como estação forense).
- b) Definir os discos de origem e destino que serão utilizados.
- c) Esterilizar os discos de origem e destino. Nesses experimentos o processo de esterilização utilizado foi a sobrescrita do DAD com zero e um do início do primeiro bloco até o final do último bloco do DAD.

5.2.1.2 Duplicação dos DADs

Foram realizadas as duplicações utilizando as ferramentas DD e CS, para os DADs com capacidade de 80 GB, 160 GB, 500 GB e 1 TB.

5.2.1.3 Resultados do Experimento 1 - Parte A

O gráfico da Figura 5.2 mostra os resultados do tempo necessário para duplicação dos DADs utilizando as ferramentas DD e CS. Observando as curvas obtidas, é possível notar que não houve diferença nos tempos de duplicação entre as duas ferramentas, quando comparados DADs com a mesma capacidade. Assim, o desempenho de ambas as ferramentas foram praticamente idênticos.

Figura 5.2 - Tempo para duplicação de DADs

Os resultados obtidos são compatíveis com os apresentados em Roussev, 2013.

Vale observar que os DADs utilizados nos testes não continham blocos com setores defeituosos (*bad blocks*). Salientando que a incidência de blocos defeituosos nos DADs aumenta significativamente o tempo para duplicação do DAD, podendo duplicar ou até triplicar esse tempo.

5.2.2 Experimento 1 - Parte B

5.2.2.1 Preparação e Procedimentos

A preparação para esta parte do experimento é apresentada a seguir. Porém, nesta parte são utilizados os arquivos de imagens do cenário M57-Patents. Por esse motivo foi necessário procedimentos não adotados na Parte - A. Foi necessária a descompactação de cada arquivo de imagem, para então “abri-la” no DAD_O. Em síntese, os procedimentos foram:

- a) Escolher os arquivos de imagem .E01 que serão utilizadas para duplicação.
- b) Descompactar e converter o formato dos arquivos (.E01 para .img).
- c) Abertura do arquivo no formato .img para o DAD_O.
 - ▶ Geração dos *hash* do DAD_O. em cada passo descrito anteriormente.
- d) Duplicar o DAD_O para o DAD_D.
 - ▶ Geração do *hash* do DAD_D.

5.2.2.2 Escolha dos Arquivos de Imagens .E01

A escolha dos quatro arquivos de imagem .E01, foi baseada no trabalho de BEEBE, 2014. No trabalho, a autora utilizou esse conjunto de arquivos de imagens para testar sua técnica de procura palavra-chave (*string-search*). Estes arquivos são identificados respectivamente por #19, #40, #59 e #79, conforme a Tabela 5.1.

5.2.2.3 Descompactação dos arquivos .E01 e gravação no DAD₀

Após a descompactação dos referidos arquivos, é possível obter as informações de espaço alocado e espaço não alocado, conforme pode ser observado na Tabela 5.3.

Tabela 5.3 – Arquivos de imagens descompactados para o formato .img

Arquivo	Tamanho do Arquivo de imagem			
	.E01 (GB)	.img (GB)		
		alocados	não alocados	Total
#19.E01	3,6	8,43	1,09	9,52
#40.E01	5,5	6,82	7,49	14,31
#59.E01	5,7	6,30	5,80	12,1
#79.E01	9,8	21,9	16,2	38,1

5.2.2.4 Resultados Experimento 1 - Parte B

Os procedimentos de duplicação com as ferramentas DD e CS foram aplicados para cada um dos arquivos relacionados na Tabela 5.3. Com isso, foram obtidos os respectivos tempos de duplicação para as ferramentas DD e CS. Os resultados estão dispostos na Tabela 5.4.

Tabela 5.4 – Tempo de duplicação utilizando as ferramentas DD e CS.

Arquivo	Ferramenta	
	DD	CS
#19.img	01min19s	01min19s
#40. img	01min56s	01min55s
#59. img	01min39s	01min38s
#79. img	05min02s	04min59s

Observando a tabela, é possível notar que os tempos decorridos para as duplicações dos DADs por ambas as ferramentas foram praticamente os mesmos, sendo que a ferramenta CS apresentou um desempenho ligeiramente superior, mas não significativo, quando comparado com a ferramenta DD. Assim, para a funcionalidade de duplicação a ferramenta CS apresenta um desempenho equiparável com a DD, tradicionalmente utilizada para duplicação de DADs.

5.2.2.5 Considerações Sobre os Resultados do Experimento 1

Os resultados obtidos no Experimento 1 mostraram que o tempo necessário para duplicação utilizando a ferramenta CS é praticamente igual ao da ferramenta DD. O que significa que o desempenho da ferramenta CS para realizar a duplicação é igual ao de uma ferramenta já conceituada e amplamente utilizada entre a comunidade da área.

5.3 Estimativa de Tempo para DADs com Diferentes Capacidades

Considerando que no local de perícias é possível encontrar DADs com diferentes capacidades e quantidades, a partir das informações do gráfico da Figura 5.2 torna-se possível estimar o tempo que será necessário (T_E) para duplicar dispositivos com capacidades intermediárias, diferentes das apresentadas no gráfico, por meio da interpolação dos valores para as grandezas de tempo e capacidade.

O cálculo para o tempo estimado (T_E) necessário para duplicar diversos DAD pode ser realizado a partir da Equação 19.

$$T_E = \sum_{k=1}^m T_k * QDAD_k \quad (19)$$

onde:

T_E : tempo estimado para duplicar todos os DADs;

k : índice da capacidade do dispositivo;

m : quantidade de diferentes capacidades dos DADs a serem duplicados;

$T_{1..m}$: tempo necessário para duplicar o DAD de capacidade k ;

$QDAD_{1..m}$: quantidade de DADs com diferentes capacidades;

A Equação 19 é válida no caso de ser utilizada uma única estação para a duplicação dos DADs, ou seja, a duplicação de um DAD de cada vez. Caso o perito tenha a possibilidade de utilizar mais de uma estação para realizar as duplicações em paralelo, o tempo T_E diminuirá

proporcionalmente à quantidade de estações utilizadas, conforme o desempenho do *hardware* (processador, memória e placa controladora de discos) das estações utilizadas para duplicação.

5.4 Experimento 2

O objetivo deste experimento é verificar a funcionalidade da ferramenta *CopyScan* (CS) para executar a varredura e duplicação simultaneamente no conteúdo dos dados que estão sendo duplicados de um DAD_O para um DAD_D . A varredura procura por PChs a partir de uma lista fornecida. O experimento é dividido em três partes: A, B e C, conforme descritas a seguir.

❖ Na parte A, deseja-se verificar:

- I. Qual o efeito que a procura por PChs simultaneamente à duplicação causa no **tempo de duplicação**. Para este teste, serão utilizados como base o tempo os resultados das duplicações já obtidos no Experimento 1. Não é esperado impacto no tempo de duplicação. Caso ocorra, este fenômeno não deve onerar significativamente o tempo de duplicação.
- II. A **quantidade de ocorrências** de PChs encontradas no decorrer da duplicação dos $DADs_O$. Para este teste a possibilidade de combinação (maiúsculo/minúsculo) entre caracteres não é considerada.

❖ Na parte B, deseja-se testar as funcionalidades da ferramenta *CopyScan* para verificar:

- I. Qual o efeito que a procura por PChs com combinações (maiúsculo/minúsculo) entre caracteres, simultaneamente à duplicação causa no **tempo de duplicação**. Espera-se que este fenômeno não onere significativamente o tempo de duplicação.
- II. A **quantidade de ocorrências** de PChs encontradas no decorrer da duplicação dos $DADs_O$. Para este teste serão consideradas as combinações (maiúsculo/minúsculo) entre caracteres. Caso sejam encontradas ocorrências com combinações maiúsculo/minúsculo entre caracteres para uma determinada PCh, espera-se que o resultado da quantidade de ocorrências

deverá ser maior do que a quantidade de ocorrências sem combinações, observado no item “b” da Parte A.

❖ Na parte C, deseja-se verificar a **quantidade de ocorrências de VP e FP** de cada PCh, encontradas no DAD. Para isso será utilizada a ferramenta FiltroP, descrita na Seção 4.7.2. Dessa forma, deseja-se obter:

- I. A partir dos resultados obtidos no item “II” da Parte A (quantidade de ocorrências sem combinação maiúscula/minúscula), será aplicado o processo de filtragem. Os resultados desse processo será a quantidade de ocorrências verdadeiro-positivo (VP) e falso-positivo (FP) para cada PCh, encontrada no DAD.
- II. A partir dos resultados obtidos no item “II” da Parte B (quantidade de ocorrências com combinação maiúscula/minúscula), será aplicado o processo de filtragem. Os resultados desse processo será a quantidade de ocorrências VP e FP para cada PCh, encontrada no DAD.

5.4.1 Preparação e Procedimentos

Neste experimento será utilizado o conjunto de imagens #19, #40, #59 e #79, conforme já explicado na Seção 5.2.2.2, o motivo da utilização desse conjunto de arquivos.

Adicionalmente, será utilizada uma lista de PChs, que também é proveniente do trabalho de BEEBE, 2014. Conforme a autora, essa lista foi formada pela recomendação de investigadores digitais forenses, após terem recebido informações básicas sobre o caso e os objetivos da investigação, porém, esses investigadores não tiveram nenhum contato com os dados contidos nos DADs. O conteúdo da lista de PChs é apresentado na Tabela 5.5.

Tabela 5.5 – Lista de palavras-chave

nº	palavra-chave	nº	palavra-chave	nº	palavra-chave	nº	palavra-chave
01	mcgoo01	10	VNC	19	Terry Johnson	28	@m57.biz
02	johnson01	11	nc	20	Jo Smith	29	immortality
03	smith01	12	prior art	21	Pat McGoo	30	prostate
04	brown01	13	Nitroba	22	Charlie Brown	31	Epithelial
05	VNC.exe	14	Patent	23	terry@m57.biz	32	US000
06	netcat	15	ftp	24	jo@m57.biz	33	M57
07	Project2400	16	virus	25	pat@m57.biz	34	CEO
08	swexpert.com	17	jaime	26	charlie@m57.biz	35	IT
09	steg	18	money	27	cancer	36	PR

Para este experimento, as PChs que compõem a lista foram armazenadas no arquivo WORDLIST-03_09MAR2015, que será lido e carregado em memória pela ferramenta CS no início dos processos de duplicação e varredura e procuradas no decorrer desses processos.

Antes da gravação de dados é realizada a esterilização do DAD_O e DAD_D. Considerando que já foi realizada a conversão do arquivo no formato “.E01” para “.img”.

Na parte A, foi realizado o processo de duplicação e varredura para cada um dos quatro DADs, sem combinação (maiúsculo/minúsculo) entre caracteres. Na parte B, foi realizado o processo de duplicação e varredura para cada um dos quatro DADs, considerando-se a combinação (maiúsculo/minúsculo) entre caracteres.

Na parte C, diferentemente dos procedimentos anteriores, será realizado o processo de filtragem, utilizando-se a ferramenta FiltroP (FIP), sobre os resultados da Tabela de Log de localização das PChs no DAD (Log Dedo-duro), dos resultados da varredura com combinação e sem combinação. Os resultados do Experimento 2 são mostrados a seguir.

5.4.1.1 Resultados do Experimento 2 - Parte A

Tabela 5.6 – Ocorrências de Pchs e tempo de duplicação sem combinação

	Lista de PCh: WORDLIST-03_09MAR2015.txt			
	Arquivos utilizados			
	#19	#40	#59	#79
Palavras-Chave	ocorrências	ocorrências	ocorrências	ocorrências
mcgoo01	0	0	11	0
johnson01	0	0	0	2
smith01	0	0	0	0
brown01	3	0	0	0
VNC.exe	0	0	0	0
netcat	28	24	105	1
Project2400	0	0	0	0
swexpert.com	10	0	0	0
steg	575	175	175	289
VNC	589	1266	1955	2436
nc	1196135	1117358	1531141	5141668
prior art	492	210	198	9
Nitroba	214	39	86	16
Patent	7436	2646	3362	1952
ftp	5890	6033	6717	12036
virus	4043	2442	1433	3529
jaime	60	36	17	5
money	6741	7872	20008	9542
Terry Johnson	47	199	412	296
Jo Smith	50	406	175	126
Pat McGoo	168	565	1590	372
Charlie Brown	4	4	0	4
terry@m57.biz	139	579	1267	2769
jo@m57.biz	159	1557	765	706
pat@m57.biz	195	771	2656	1811
charlie@m57.biz	538	673	668	666
cancer	3540	1664	935	1879
@m57.biz	1265	3721	5528	6183
immortality	3067	830	12	13
prostate	547	223	47	47
Epithelial	128	20	23	0
US000	103	0	0	24
M57	9175	11323	15730	4207
CEO	1433	4978	6603	2646
IT	273731	291253	343211	647221
PR	232302	300249	365619	836878
Total de ocorrências	1.748.807	1.757.116	2.310.449	6.677.333
Tempo	01min26s	02min11s	01min51s	05min53s

Conforme pode ser observado o acréscimo de tempo para a duplicação com varredura dos DADs #19, #40, #59 e #79 foi 7s, 16s, 13s e 54s respectivamente. Dessa forma, a varredura não onerou o tempo de duplicação, conforme era previsto. A quantidade de ocorrências de PChs em cada DAD, é destacada na Tabela 5.6.

5.4.1.2 Resultados do Experimento 2 - Parte B

O resultado do Experimento 2 – Parte B é mostrado a seguir na Tabela 5.7.

Tabela 5.7 – Ocorrências de PChs e tempo de duplicação com combinação

	Lista de PCh: WORDLIST-03_09MAR2015.txt			
	Arquivos			
	#19	#40	#59	#79
palavras-Chave	ocorrências	ocorrências	ocorrências	ocorrências
mcgoo01	0	0	11	0
johnson01	0	0	0	2
smith01	0	0	0	2
brown01	3	0	0	0
VNC.exe	1	0	1	0
netcat	32	28	111	13
Project2400	0	0	0	0
swexpert.com	10	0	3	0
steg	910	249	275	446
VNC	589	1266	1955	2436
nc	1490268	1539659	1989987	6401639
prior art	591	269	404	10
Nitroba	219	87	226	16
Patent	7798	2850	3691	2201
ftp	20110	22517	28956	45251
virus	5083	3854	3521	5422
jaime	74	44	41	30
money	7023	8111	21263	10518
Terry Johnson	47	199	412	296
Jo Smith	50	406	175	126
Pat McGoo	168	565	1590	372
Charlie Brown	4	4	0	4
terry@m57.biz	139	579	1269	2769
jo@m57.biz	159	1557	765	706
pat@m57.biz	195	771	2658	1811
charlie@m57.biz	538	673	669	666
cancer	4730	1952	1640	1976
@m57.biz	1265	3721	5533	6259
immortality	3091	837	15	23
prostate	662	305	71	47
Epithelial	128	20	24	0
US000	103	0	0	24
M57	9175	11323	15730	4207
CEO	1433	4978	6603	2646
IT	273731	291253	343211	647221
PR	232302	300249	365619	836878
Total de ocorrências	2.060.631	2.198.326	2.796.429	7.974.017
Tempo	1m27s	2m11s	1m51s	5min55s

A varredura procurando por diferentes padrões de codificação simultaneamente à duplicação, não onera significativamente o tempo de duplicação do DAD. Pode ser observado que a quantidade de ocorrências de cada PCh em cada DAD, aumentou significativamente.

Na Tabela 5.8, é apresentado um resumo que permite comparar os tempos para duplicação sem varredura e com varredura procurando as PChs com combinações de caracteres maiúsculo/minúsculo.

Tabela 5.8 – Tempo de duplicação para a ferramenta *CopyScan*.

Arquivo	sem varredura	com varredura	com varredura (combinações maiús./minús.)
#19	01min19s	01min26s	01min27s
#40	01min55s	02min11s	02min11s
#59	01min38s	01min51s	01min51s
#79	04min59s	05min53s	05min55s

Comparando os resultados é possível notar que o tempo necessário para realizar a duplicação sem varredura, com varredura e varredura com combinações maiúscula/minúscula são relativamente próximos. Assim, não houve um impacto significativo no tempo de duplicação de forma a inviabilizar a varredura para procurar pela PChs simultaneamente.

Na Tabela 5.9 é apresentada a sumarização dos resultados obtidos nas Tabelas 5.6 e 5.7, mostrando os resultados obtidos para cada DAD e total de quantidade de ocorrências para cada DAD.

Tabela 5.9 – Resumo da quantidade de ocorrências encontradas

Arquivo	Quantidade de ocorrência de PChs por DAD		Diferença
	varredura (sem combinação)	varredura (com combinação)	
#19	1.748.807	2.060.631	311.824
#40	1.757.116	2.198.326	441.210
#59	2.310.449	2.796.429	485.980
#70	6.677.333	7.974.017	1.296.684
Total de ocorrências nos DAD	12.493.705	15.029.403	2.535.698

Conforme pode ser visto, a varredura com combinações de caracteres maiúsculo/minúsculo, causou um aumento de 2.535.698 ocorrências de PChs.

5.4.1.3 Resultados do Experimento 2 - Parte C

A seguir, conforme os critérios vistos nas Seções 4.4.2 e 4.7.2, que tratam do Processo de Filtragem da Ferramenta FiltroP, serão utilizados os resultados obtidos nas Tabelas 5.6 e 5.7 para classificar as ocorrências como VP e FP.

Observe na Tabela 5.10 que para cada DAD é mostrado a quantidade de ocorrências de cada PCh com as respectivas classificações de VP e FP e o total de ocorrências de PChs em cada DAD. Como referência foi utilizada a Tabela 5.6.

Tabela 5.10 – Ocorrências de PChs (VP e FP) sem combinações

PCh	Imagens											
	#19			#40			#59			#79		
	VP	FP	Total	VP	FP	Total	VP	FP	Total	VP	FP	Total
mcgoo01	0	0	0	0	0	0	3	8	11	0	0	0
johnson01	0	0	0	0	0	0	0	0	0	2	0	2
smith01	0	0	0	0	0	0	0	0	0	0	0	0
brown01	2	1	3	0	0	0	0	0	0	0	0	0
VNC.exe	0	0	0	0	0	0	0	0	0	0	0	0
netcat	21	7	28	8	16	24	90	15	105	1	0	1
Project2400	0	0	0	0	0	0	0	0	0	0	0	0
swexpert.com	9	1	10	0	0	0	0	0	0	0	0	0
steg	520	55	575	142	33	175	150	25	175	269	20	289
VNC	79	510	589	192	1074	1266	298	1657	1955	876	1560	2436
nc	863954	332181	1196135	727431	389927	1117358	1065526	465615	1531141	4662031	479637	5141668
prior art	491	1	492	210	0	210	189	9	198	6	3	9
Nitroba	124	90	214	36	3	39	78	8	86	16	0	16
Patent	7132	304	7436	2574	72	2646	3033	329	3362	1913	39	1952
ftp	3937	1953	5890	3634	2399	6033	4611	2106	6717	8977	3059	12036
virus	3676	367	4043	1995	447	2442	1076	357	1433	3221	308	3529
jaime	45	15	60	27	9	36	13	4	17	5	0	5
money	5833	908	6741	6754	1118	7872	17458	2550	20008	8830	712	9542
Terry Johnson	45	2	47	94	105	199	168	244	412	262	34	296
Jo Smith	47	3	50	193	213	406	76	99	175	126	0	126
Pat McGoo	166	2	168	238	327	565	740	850	1590	372	0	372
Charlie Brown	4	0	4	4	0	4	0	0	0	4	0	4
terry@m57.biz	132	7	139	385	194	579	733	534	1267	2748	21	2769
jo@m57.biz	149	10	159	1075	482	1557	485	280	765	702	4	706
pat@m57.biz	193	2	195	420	351	771	1517	1139	2656	1795	16	1811
charlie@m57.biz	514	24	538	416	257	673	443	225	668	662	4	666
cancer	3445	95	3540	1574	90	1664	857	78	935	1805	74	1879
@m57.biz	1213	52	1265	2572	1149	3721	3455	2073	5528	6145	38	6183
immortality	2739	328	3067	711	119	830	10	2	12	13	0	13
prostate	538	9	547	218	5	223	41	6	47	47	0	47
Epithelial	128	0	128	20	0	20	23	0	23	0	0	0
US000	103	0	103	0	0	0	0	0	0	21	3	24
M57	2227	6948	9175	1796	9527	11323	4648	11082	15730	1568	2639	4207
CEO	280	1153	1433	2105	2873	4978	2443	4160	6603	993	1653	2646
IT	159031	114700	273731	136366	154887	291253	161238	181973	343211	427761	219460	647221
PR	67335	164967	232302	64502	235747	300249	84834	280785	365619	467873	369005	836878
Total	1124112	624695	1748807	955692	801424	1757116	1354236	956213	2310449	5599044	1078289	6677333

Na Tabela 5.11, tem-se a quantidade de ocorrências de PChs de VP e FP, e o total de ocorrências de PChs em cada DAD. Como referência foi utilizada a Tabela 5.7.

Tabela 5.11 – Ocorrências de PChs (VP e FP) com combinações

PCh	Imagens											
	#19			#40			#59			#79		
	VP	FP	Total	VP	FP	Total	VP	FP	Total	VP	FP	Total
mcgoo01	0	0	0	0	0	0	3	8	11	0	0	0
johnson01	0	0	0	0	0	0	0	0	0	2	0	2
smith01	0	0	0	0	0	0	0	0	0	2	0	2
brown01	2	1	3	0	0	0	0	0	0	0	0	0
VNC.exe	0	1	1	0	0	0	0	1	1	0	0	0
netcat	25	7	32	8	20	28	96	15	111	13	0	13
Project2400	0	0	0	0	0	0	0	0	0	0	0	0
swexpert.com	9	1	10	0	0	0	2	1	3	0	0	0
steg	826	84	910	196	53	249	211	64	275	361	85	446
VNC	80	509	589	192	1074	1266	299	1656	1955	876	1560	2436
nc	975793	514475	1490268	853966	685693	1539659	1203978	786009	1989987	5503922	897717	6401639
prior art	562	29	591	257	12	269	301	103	404	7	3	10
Nitroba	128	91	219	72	15	87	173	53	226	16	0	16
Patent	7344	454	7798	2754	96	2850	3256	435	3691	2138	63	2201
ftp	8930	11180	20110	8641	13876	22517	11210	17746	28956	22978	22273	45251
virus	4339	744	5083	2980	874	3854	2502	1019	3521	4657	765	5422
jaime	57	17	74	35	9	44	35	6	41	16	14	30
money	6054	969	7023	6926	1185	8111	18503	2760	21263	9710	808	10518
Terry Johnson	45	2	47	94	105	199	168	244	412	262	34	296
Jo Smith	47	3	50	193	213	406	76	99	175	126	0	126
Pat McGoo	166	2	168	238	327	565	740	850	1590	372	0	372
Charlie Brown	4	0	4	4	0	4	0	0	0	4	0	4
terry@m57.biz	132	7	139	385	194	579	734	535	1269	2748	21	2769
jo@m57.biz	149	10	159	1075	482	1557	485	280	765	702	4	706
pat@m57.biz	193	2	195	420	351	771	1518	1140	2658	1795	16	1811
charlie@m57.biz	514	24	538	416	257	673	443	226	669	662	4	666
cancer	4613	117	4730	1847	105	1952	1507	133	1640	1894	82	1976
@m57.biz	1213	52	1265	2572	1149	3721	3460	2073	5533	6145	114	6259
immortality	2750	341	3091	716	121	837	13	2	15	23	0	23
prostate	651	11	662	300	5	305	65	6	71	47	0	47
Epithelial	128	0	128	20	0	20	24	0	24	0	0	0
US000	103	0	103	0	0	0	0	0	0	21	3	24
M57	2227	6948	9175	1796	9527	11323	4648	11082	15730	1568	2639	4207
CEO	280	1153	1433	2105	2873	4978	2443	4160	6603	993	1653	2646
IT	159031	114700	273731	136366	154887	291253	161238	181973	343211	427761	219460	647221
PR	67328	164974	232302	64495	235754	300249	84827	280792	365619	467873	369005	836878
TOTAL	1243723	816908	2060631	1089069	1109257	2198326	1502958	1293471	2736429	6457694	1516323	7974000

5.4.2 Considerações Sobre os Resultados do Experimento 2

Foram verificadas as funcionalidades da ferramenta CS para executar a varredura e duplicação simultaneamente no conteúdo dos dados que foram duplicados de um DAD_O para um DAD_D.

A partir dos resultados, nota-se a possibilidade para executar a varredura e duplicação simultaneamente no conteúdo dos dados que estão sendo duplicados.

Com relação ao tempo decorrido para a varredura por PChs, independentemente de a varredura ser realizada com combinação e sem combinação (maiúsculo/minúsculo) dos caracteres, não foi observado o aumento significativo no tempo de duplicação, o que pode ser explicado pelo fato do tempo de busca pela ferramenta CS ser menor do que o tempo necessário para escrita (gravação) no DAD.

A realização de varredura com combinação maiúsculo/minúsculo dos caracteres, possibilita encontrar uma determinada PCh independentemente da forma como a mesma estiver escrita. Porém, observa-se um aumento significativo na quantidade de ocorrências de PChs.

Mediante os resultados, é possível notar que as PChs compostas com menos de 4 caracteres produziram uma elevada quantidade de ocorrências. Decorrente disso, observa-se uma elevada quantidade de ocorrências de Falso-positivo para estas PChs. Isso pode ser um indicativo de que as PChs escolhidas pelo perito devem ter quatro caracteres ou mais.

5.5 Experimento 3

Na prática, para obter-se os resultados que são produzidos pela ferramenta *CopyScan* - CS, seriam necessárias a utilização de duas ferramentas distintas: uma para duplicação e outra para varredura.

Este experimento visa comparar os resultados da quantidade de ocorrências de palavras-chave - PChs obtidos pela ferramenta CS com outra ferramenta amplamente utilizada por peritos e pesquisadores, a ferramenta *Autopsy*. Além disso, será comparado o tempo necessário para obtenção dos resultados.

Este experimento será dividido em duas partes: A e B, conforme descrito a seguir.

- ❖ Na parte A, será utilizada a ferramenta *Autopsy* para verificar a quantidade de ocorrências de PChs encontradas no decorrer da varredura dos DADs₀, bem como o tempo gasto pela ferramenta para realizar a varredura.
- ❖ Na parte B, deseja-se comparar os tempos para obtenção dos resultados. Ou seja, qual seria o tempo necessário para obter os resultados semelhantes aos da ferramenta CS utilizando-se duas ferramentas (uma para duplicação e outra para procurar por PChs).

5.5.1 Preparação e Procedimentos

Neste experimento será utilizado o conjunto de imagens #19, #40, #59 e #79, conforme já descrito na Seção 5.2.2.2, a lista de PChs da Tabela 5.6., as ferramentas CP, *Autopsy* e DD.

5.5.2 Resultados do Experimento 3

5.5.2.1 Resultados do Experimento 3 - Parte A

A seguir, na Tabela 5.12 são apresentados principalmente dois resultados obtidos pela ferramenta *Autopsy*, a quantidade de ocorrências que foram encontradas a partir da lista de PChs da Tabela 5.5 e o tempo gasto para realizar a varredura para procurar pelas PChs.

Tabela 5.12 – Ocorrências encontradas nos DADs pela ferramenta Autopsy

PChs	DADs			
	#19	#40	#59	#79
mcgoo01	0	0	3	0
johnson01	0	0	0	0
smith01	0	0	0	0
brown01	2	0	0	0
VNC.exe	0	0	0	0
netcat	4	4	6	2
Project2400	7	2	12	0
swexpert.com	4	0	1	0
steg	24	23	22	44
VNC	40	47	109	246
nc	1161	1608	1222	2696
prior art	175	56	46	4
Nitroba	55	2	16	0
Patent	538	310	410	271
ftp	841	899	813	1098
virus	346	330	359	377
jaime	19	12	20	6
money	114	104	242	186
Terry Johnson	7	10	42	121
Jo Smith	8	61	17	38
Pat McGoo	7	45	197	140
Charlie Brown	4	4	0	1
terry@m57.biz	7	21	88	168
jo@m57.biz	7	68	53	67
pat@m57.biz	7	36	168	156
charlie@m57.biz	41	10	39	60
cancer	120	72	65	39
@m57.biz	99	173	253	212
immortality	93	31	10	10
prostate	55	29	23	11
Epithelial	57	29	19	8
US000	0	0	0	0
M57	521	549	779	511
CEO	70	83	151	164
IT	0	0	0	0
PR	2140	2274	2339	5957
Total de ocorrências	6573	6892	7524	12593
Tempo decorrido	24 min	60 min	37 min	107min

5.5.2.2 Resultados do Experimento 3 - Parte B

Observe que o tempo total para realizar a duplicação e varredura utilizando as duas ferramentas, será a soma do tempo necessário para duplicação (t_1) e do tempo necessário para a varredura para procurar pelas PChs (t_2). A Figura 5.3 ilustra a comparação de tempo necessário para duplicação e varredura utilizando as ferramentas CS e DD+Autopsy.

Figura 5.3 – Comparação do tempo para duplicação e varredura

A Tabela 5.14 mostra a comparação do tempo entre os métodos e as ferramentas utilizadas para obter a quantidade de ocorrências das PChs.

Tabela 5.13 – Exemplo da comparação do tempo entre os métodos

	Método			
	PROPOSTO	TRADICIONAL		
Arquivo	CS	DD (t ₁)	Autopsy (t ₂)	T = t ₁ + t ₂
#19	01min19s	01min19s	24min	25min19s
#40	01min55s	01min56s	60min	61min56s
#59	01min38s	01min39s	37min	38min39s
#79	04min59s	05min02s	107min	112min02s

5.5.3 Considerações Sobre os Resultados do Experimento 3

Baseado nos resultados, é possível notar que com a utilização da ferramenta CS os resultados podem ser obtidos em um tempo significativamente menor quando comparado com a utilização de duas ferramentas distintas.

Supondo que o tempo de duplicação fosse desconsiderado, por serem iguais, observamos que o maior “custo” de tempo é decorrente da varredura para busca pelas PChs.

A diferença de tempo para obtenção da quantidade de ocorrências de PChs é que a ferramenta *CopyScan* realiza a busca sequencial, enquanto que a ferramenta *Autopsy*, primeiro realiza a indexação dos arquivos para depois iniciar a busca pelas PChs.

No decorrer deste experimento foi observado que PChs formadas com menos de quatro caracteres apresentavam uma quantidade significativamente maior de ocorrências quando comparadas com palavras com quatro ou mais caracteres. Isso foi observado em ambas as ferramentas. Inspeccionando visualmente essas ocorrências foi possível notar que muitas ocorrências de PChs com menos de quatro caracteres eram provenientes de falsos positivos – FP. Assim, para os testes subsequentes, essas palavras foram excluídas da lista de palavras-chave utilizada.

5.6 Experimento 4

O objetivo de Experimento 4 é avaliar a convergência dos resultados do *Log* dedo-duro da ferramenta CS (QO_{DT}), dos resultados da ferramenta FiltroP (VP no DAD todo ($QOVP_{DT}$)) e Arquivos Interpretáveis ((QO_{AI}) e VP nos Arquivos Interpretáveis ($QOVP_{AI}$)), tendo como métrica a quantidade de ocorrências de PChs encontradas. Caso a convergência seja constatada, os resultados de QO_{DT} e $QOVP_{DT}$ poderão ser utilizados para a realização dos processos de triagem e ranqueamento.

Para alcançar este objetivo, será necessário determinar a quantidade de ocorrências de PChs em cada “parte” do DAD (QO_{EL} , QO_{AB} , QO_{AI} e QO_{AM}), aplicando os procedimentos para avaliação dos processos de varredura e filtragem que foram descritos na Seção 4.9.1. Com isso, será possível avaliar se este método de validação é adequado.

5.6.1 Preparação e Procedimentos

Serão utilizadas imagens do caso M57-*Patents*. Relembrando, conforme foi visto na Seção 5.1., o conjunto completo é composto por 79 imagens. Para os propósitos desse experimento, o conjunto foi dividido em grupos. Estes grupos foram formados a partir de um sorteio ao acaso e sem reposição, cujo resultado é mostrado a Tabela 5.14.

Tabela 5.14 – Grupos de imagens formados a partir do sorteio

1º GRUPO	2º GRUPO	3º GRUPO	4º GRUPO	5º GRUPO	6º GRUPO	7º GRUPO
20	26	1	5	3	8	7
24	31	2	12	19	10	9
30	32	4	15	23	11	16
34	36	6	18	25	22	21
35	40	13	38	27	29	33
41	42	14	49	28	46	44
45	43	17	50	37	47	48
56	59	51	58	39	63	52
57	60	64	61	55	71	53
66	65	67	62	76	72	54
69	73	68	74	79	77	70
					78	75

Inicialmente, os procedimentos para avaliação dos processos de varredura e filtragem seriam aplicados para todos os grupos, porém,

por ser tratar de um experimento para prova de conceito, optou-se por juntar os grupos 6 e 7, por serem os maiores quando comparado com os demais grupos e aplicar os procedimentos nesse subconjunto de 24 imagens.

A lista de PChs que será utilizada é mostrada a seguir na Tabela 5.15. Em relação a lista anterior foram excluídas as PChs: VNC, nc, ftp, M57, CEO, IT e PR.

Tabela 5.15 – Lista de palavras-chave com quatro caracteres ou mais

nº	palavra-chave	nº	palavra-chave	nº	palavra-chave	nº	palavra-chave	nº	palavra-chave
01	mcgoo01	07	Project2400	13	virus	19	Charlie Brown	25	@m57.biz
02	johnson01	08	swexpert.com	14	jaime	20	terry@m57.biz	26	immortality
03	smith01	09	steg	15	money	21	jo@m57.biz	27	prostate
04	brown01	10	prior art	16	Terry Johnson	22	pat@m57.biz	28	Epithelial
05	VNC.exe	11	Nitroba	17	Jo Smith	23	charlie@m57.biz	29	US000
06	netcat	12	Patent	18	Pat McGoo	24	cancer		

Os procedimentos descritos nas Seções 4.9.1.1 a 4.9.1.4 serão aplicados em cada “parte” do DAD. A Figura 5.4 ilustra a sequência em que a varredura em cada parte é realizada para contagem da quantidade de ocorrências de PChs.

Figura 5.4 – Sequências de Varreduras

5.6.1.1 Resultados do Experimento 4

Para este experimento, foi escolhida o arquivo de imagem #7, o primeiro do subconjunto, para exemplificar os procedimentos descritos na Seção 4.9.1, a serem aplicados em todas as 24 imagens do subconjunto.

A partir da descrição da Seção 4.9.1.1, é realizada a verificação da quantidade de ocorrências no DAD todo (QO_{DT}). Por meio da ferramenta CS é realizada a **primeira varredura** no DAD. Mas:

$$QO_{DT} = QOVP_{DT} + QOFP_{DT}$$

Logo:

$QO_{DT} = 14090$ (Resultado proveniente da Tabela de *Log* Dedo-duro da ferramenta CS)

Aplicando-se o processo de filtragem sobre a quantidade total de ocorrências, obtém-se:

$QOVP_{DT} = 12914$ e $QOFP_{DT} = 1176$ (Valores provenientes da ferramenta FiltroP)

Cálculo de ocorrências em arquivos nos *Clusters* Alocados:

O espaço livre EL é sobrescrito com zeros utilizando a ferramenta EraseFile. Então é realizada a **segunda varredura** no DAD para obtenção de QO_{AA} :

$QO_{AA} = 12080$

A quantidade de ocorrências no espaço livre (O_{EL}) pode ser calculada por:

$$QO_{EL} = QO_{DT} - QO_{AA}$$

$QO_{EL} = 14090 - 12080$

$QO_{EL} = 2010$

O *Log* do Dedo-duro da segunda varredura é submetido ao processo de filtragem e obtém-se: $QOVP_{AA}$ e $QOFP_{AA}$.

$QOVP_{AA} = 10996$

$QOFP_{AA} = 1084$

Somando os valores de $QOVP_{AA}$ e $QOFP_{AA}$, confirma-se o total de ocorrências obtido da segunda varredura:

$$QO_{AA} = QOVP_{AA} + QOFP_{AA}$$

$$QO_{AA} = 10996 + 1084 = 12080$$

Como já foi obtido os valores de ocorrências de $QOVP_{AA}$ e $QOFP_{AA}$, os valores de $QOVP_{EL}$ e $QOFP_{EL}$ podem ser calculados por meio das equações (7) e (8) da Seção 4.9.1.2:

$$QOVP_{EL} = QOVP_{DT} - QOVP_{AA}$$

$$QOVP_{EL} = 12914 - 10996 = 1918$$

$$QOFP_{EL} = QOFP_{DT} - QOFP_{AA}$$

$$QOFP_{EL} = 1176 - 1084 = 92$$

Conforme descrito na Seção 4.9.1.3, realiza-se a operação de sobrescrever os arquivos binários AB, por meio da ferramenta EraseFiles. Em seguida realiza-se a **terceira varredura** para obter a quantidade de ocorrências de PChs em arquivos interpretáveis AI e mistos AM, ou seja, QO_{IM} (Quantidade de ocorrências de PCH nos arquivos Interpretáveis e Mistos). Assim é possível obter QO_{IM} .

$$QO_{IM} = 5691$$

Com isso é possível obter QO_{AB} :

$$QO_{AB} = QO_{AA} - QO_{IM}$$

$$QO_{AB} = 12080 - 5691 = 6389$$

Aplicando o processo de filtragem no Log do Dedo-Duro da terceira varredura, realizado sobre os arquivos Interpretáveis e Mistos (QO_{IM}), tem-se:

$$QO_{IM} = QOVP_{IM} + QOFP_{IM}, \text{ da equação (10)}$$

$$QVP_{IM} = 5488$$

$$QOFP_{IM} = 203$$

$$QO_{IM} = 5691$$

Da equação (11), tem-se que:

$$QOVP_{AB} = QOVP_{AA} - QOVP_{IM}$$

$$QOVP_{AB} = 10996 - 5488 = 5506$$

Da equação (12), tem-se que:

$$QOFP_{AB} = QOFP_{AA} - QOFP_{IM}$$

$$QOFP_{AB} = 1084 - 203 = 881$$

Para a calcular a quantidade de ocorrências em arquivos interpretáveis, sobrescreve-se os arquivos Mistos AM com zeros, obtendo-se a quantidade de ocorrências de PChs nos arquivos interpretáveis AI. Então, é realizada a **quarta varredura**, obtendo-se o resultado da quantidade das ocorrências de PChs em arquivos Interpretáveis (QO_{AI}).

$$QO_{AI} = 5674$$

Da equação (13), tem-se que:

$$QO_{AM} = QO_{IM} - QO_{AI}$$

$$QO_{AM} = 5691 - 5674 = 17$$

A partir do *Log* do dedo-duro da varredura sobre os arquivos Interpretáveis é submetido ao processo de filtragem, obtendo-se:

$QOVP_{AI}$ e $QOFP_{AI}$.

$$QOVP_{AI} = 5471$$

$$QOFP_{AI} = 203$$

Assim, verifica-se que:

$$QO_{AI} = QOVP_{AI} + QOFP_{AI} = 5674$$

A partir dos valores já obtidos de $QOVP_{AI}$ e $QOFP_{AI}$, é possível calcular $QOVP_{AM}$ e $QOFP_{AM}$ pelas equações (15) e (16):

$$QOVP_{AM} = QOVP_{IM} - QOVP_{AI}$$

$$QOVP_{AM} = 5488 - 5471 = 17$$

$$QOFP_{AM} = QOFP_{IM} - QOFP_{AI}$$

$$QOFP_{AM} = 203 - 203 = 0$$

Os resultados obtidos para a quantidade de ocorrências de cada PCh no DAD #07 é apresentado na Tabela 5.16.

Nesta tabela, é possível verificar que os resultados apresentados no exemplo anterior, é igual ao total da quantidade de ocorrências de PChs em cada parte do DAD. Para popular a tabela, o mesmo procedimento do exemplo foi aplicado para todas as PChs. Isso foi feito para obtenção da QO para as 29 PChs em cada parte do DAD (QO_{EL} , QO_{AB} , QO_{AM} e QO_{AI}).

Tabela 5.16 – Ocorrências de PChs em cada parte do DAD #7

PChs	QO _{DT}	QOVP _{DT}	QOFP _{DT}	QO _{AA}	QOVP _{AA}	QOFP _{AA}	QO _{EL}	QOVP _{EL}	QOFP _{EL}	QO _{IM}	QOVP _{IM}	QOFP _{IM}	QO _{AB}	QOVP _{AB}	QOFP _{AB}	QO _{AI}	QOVP _{AI}	QOFP _{AI}	QOVP _{AM}	QOFP _{AM}	QO _{AM}
@m57.biz	418	413	5	412	409	3	6	4	2	5	3	2	407	406	1	5	3	2	0	0	0
brown01	2	2	0	2	2	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0
cancer	1184	1107	77	1178	1103	75	6	4	2	1135	1086	49	43	17	26	1135	1086	49	0	0	0
Charlie Brown	2	2	0	2	2	0	0	0	0	2	2	0	0	0	0	2	2	0	0	0	0
charlie@m57.biz	179	176	3	173	172	1	6	4	2	1	1	0	172	171	1	1	1	0	0	0	0
Epithelial	7	7	0	7	7	0	0	0	0	7	7	0	0	0	0	7	7	0	0	0	0
immortality	456	422	34	382	356	26	74	66	8	251	249	2	131	107	24	246	244	2	5	0	5
jaime	42	32	10	31	22	9	11	10	1	5	5	0	26	17	9	5	5	0	0	0	0
Jo Smith	12	12	0	12	12	0	0	0	0	2	2	0	10	10	0	2	2	0	0	0	0
jo@m57.biz	74	74	0	74	74	0	0	0	0	0	0	0	74	74	0	0	0	0	0	0	0
johnson01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mcgoo01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
money	5917	5229	688	4644	4001	643	1273	1228	45	594	527	67	4050	3474	576	594	527	67	0	0	0
netcat	27	23	4	25	21	4	2	2	0	0	0	0	25	21	4	0	0	0	0	0	0
Nitroba	48	28	20	38	23	15	10	5	5	20	14	6	18	9	9	20	14	6	0	0	0
Pat McGoo	57	57	0	57	57	0	0	0	0	0	0	0	57	57	0	0	0	0	0	0	0
pat@m57.biz	84	84	0	84	84	0	0	0	0	0	0	0	84	84	0	0	0	0	0	0	0
Patent	2447	2395	52	2072	2029	43	375	366	9	1384	1381	3	688	648	40	1372	1369	3	12	0	12
prior art	109	109	0	106	106	0	3	3	0	96	96	0	10	10	0	96	96	0	0	0	0
Project2400	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
prostate	62	58	4	61	58	3	1	0	1	57	57	0	4	1	3	57	57	0	0	0	0
smith01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
steg	517	473	44	416	375	41	101	98	3	255	247	8	161	128	33	255	247	8	0	0	0
swexpert.com	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Terry Johnson	6	6	0	6	6	0	0	0	0	0	0	0	6	6	0	0	0	0	0	0	0
terry@m57.biz	36	36	0	36	36	0	0	0	0	0	0	0	36	36	0	0	0	0	0	0	0
US000	81	81	0	66	66	0	15	15	0	3	3	0	63	63	0	3	3	0	0	0	0
virus	2323	2088	235	2196	1975	221	127	113	14	1874	1808	66	322	167	155	1874	1808	66	0	0	0
VNC.exe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total:	14090	12914	1176	12080	10996	1084	2010	1918	92	5691	5488	203	6389	5508	881	5674	5471	203	17	0	17

(*)⁴¹

A partir do total da quantidade de ocorrências de PChs para o Espaço Livre (QOVP_{EL} e QOFP_{EL}), Binários (QOVP_{AB} e QOFP_{AB}), Mistos (QOVP_{AM} e QOFP_{AM}), Interpretáveis (QOVP_{AI} e QOFP_{AI}) e DAD todo (QOVP_{DT} e QOFP_{DT}), é obtida uma tabela mostrando este valor para cada DAD do subconjunto, cuja finalidade é comparar a quantidade de ocorrências entre os DADs, mostrado a seguir na Tabela 5.17.

⁴¹ O resultado da quantidade de ocorrências de PCh para cada um dos 24 DADs é apresentado na Seção 5.6.2, na Tabela 5.19. Os resultados apresentados nesta tabela são importantes quando se deseja comparar a quantidade de ocorrências de PCh em cada DAD.

Tabela 5.17 – Ocorrências de PChs para todos DADs do subconjunto

Nº DAD	QOVP					QOFP				
	QOVP _{EL}	QOVP _{AB}	QOVP _{AM}	QOVP _{AI}	QOVP _{DT}	QOFP _{EL}	QOFP _{AB}	QOFP _{AM}	QOFP _{AI}	QOFP _{DT}
07	1918	5508	17	5471	12914	92	881	0	203	1176
08	2165	4914	17	5884	12980	154	1030	0	209	1393
09	2617	6343	17	8019	16996	176	929	0	227	1332
10	2377	5469	17	8679	16542	264	1190	0	242	1696
11	2237	5586	17	9453	17293	224	1351	0	255	1830
16	3180	6395	11	11235	20821	332	1235	0	272	1839
21	475	3017	0	1318	4810	50	723	0	156	929
22	1156	5239	2	1826	8223	71	1065	0	206	1342
29	480	5201	0	4835	10516	14	2045	0	171	2230
33	3434	6771	0	6258	16463	1001	2042	0	213	3256
44	658	6932	0	4242	11832	50	2489	0	210	2749
46	969	11521	175	5927	18592	507	3116	0	271	3894
47	1926	13484	177	6369	21956	466	3875	0	291	4632
48	4040	12527	179	7691	24437	1426	3611	0	303	5340
52	5883	13567	184	8470	28104	2082	4157	0	331	6570
53	5966	12383	184	8514	27047	2038	3775	0	333	6146
54	6008	12586	184	8729	27507	2045	4021	0	338	6404
63	231	3374	0	7040	10645	53	1522	0	470	2045
70	8096	2613	0	12016	22725	423	1104	0	672	2199
71	4270	2699	0	18165	25134	234	1100	0	895	2229
72	3958	3237	0	17139	24334	196	755	0	804	1755
75	3091	3366	0	20118	26575	157	767	0	887	1811
77	5245	3642	0	20436	29323	221	616	0	828	1665
78	4653	3638	0	20726	29017	202	330	0	799	1331

Dentre as partes que foram consideradas anteriormente, a de maior importância é a que foi denominada de Arquivos Interpretáveis (AI). Visto tratar-se da parte que contém os arquivos provenientes de texto simples, que faz deles sempre possíveis de serem lidos pela ferramenta *CopyScan* enquanto realiza a varredura em dados brutos para busca por PChs.

Para avaliar se a quantidade de ocorrências do DAD todo (QO_{DT}) e a quantidade de ocorrências de verdadeiro-positivos do DAD todo (QOVP_{DT}) são suficientes para realizar a triagem e ranqueamento, serão comparados os resultado de:

$$\textcircled{1} \quad \frac{\text{QOVP}_{\text{AI}}}{\text{QOVP}_{\text{DT}}}$$

$$\textcircled{2} \quad \frac{QO_{AI}}{QO_{DT}}$$

$$\textcircled{3} \quad \frac{QO_{AI}}{QOVP_{DT}}$$

conforme mostra a Tabela 5.18.

Tabela 5.18 – Proporção entre os arquivos interpretáveis e o DAD todo

Nº DAD	QOVP			Qtd. Total de Ocorr.				Predição positiva	
	QOVP _{AI}	QOVP _{DT}	QOVP _{AI}	QO _{AI}	QO _{DT}	QO _{AI}	QO _{AI}	QOVP _{DT}	QOVP _{AI}
			QOVP _{DT}			QO _{DT}	QOVP _{DT}		
7	5471	12914	0,4	5674	14090	0,4	0,4	0,92	0,96
8	5884	12980	0,5	6093	14373	0,4	0,5	0,90	0,97
9	8019	16996	0,5	8246	18328	0,5	0,5	0,93	0,97
10	8679	16542	0,5	8921	18238	0,5	0,5	0,91	0,97
11	9453	17293	0,6	9708	19123	0,5	0,6	0,90	0,97
16	11235	20821	0,5	11507	22660	0,5	0,6	0,92	0,98
21	1318	4810	0,3	1474	5739	0,3	0,6	0,84	0,89
22	1826	8223	0,2	2032	9565	0,2	0,2	0,86	0,90
29	4835	10516	0,5	5006	12746	0,4	0,5	0,83	0,97
33	6258	16463	0,4	6471	19719	0,3	0,4	0,83	0,97
44	4242	11832	0,4	4452	14581	0,3	0,4	0,81	0,95
46	5927	18592	0,3	6198	22486	0,3	0,3	0,83	0,96
47	6369	21956	0,3	6660	26588	0,3	0,3	0,83	0,96
48	7691	24437	0,3	7994	29777	0,3	0,3	0,82	0,96
52	8470	28104	0,3	8801	34674	0,3	0,3	0,81	0,96
53	8514	27047	0,3	8847	33193	0,3	0,3	0,81	0,96
54	8729	27507	0,3	9067	33911	0,3	0,3	0,81	0,96
63	7040	10645	0,7	7510	12690	0,6	0,7	0,84	0,94
70	12016	22725	0,5	12688	24924	0,5	0,6	0,91	0,95
71	18165	25134	0,7	19060	27363	0,7	0,8	0,92	0,95
72	17139	24334	0,7	17943	26089	0,7	0,7	0,93	0,96
75	20118	26575	0,8	21005	28386	0,7	0,8	0,94	0,96
77	20436	29323	0,7	21264	30988	0,7	0,7	0,95	0,96
78	20726	29017	0,7	21525	30348	0,7	0,7	0,96	0,96
Média (M)			0,5			0,4	0,5	0,87	0,96
Desvio Padrão (DP)			0,2			0,2	0,2	0,05	0,02

Observe que os valores demarcados na tabela, encontram-se fora do intervalo da média e desvio-padrão.

5.6.1.2 Considerações Sobre os Resultados do Experimento 4

Analisando a proporção entre a QOVP_{AI} e QOVP_{DT}, tem-se a média de 0,5 e desvio padrão de 0,2. Mediante os resultados observa-se que mais de 90% das proporções estão dentro da média.

Analisando a proporção entre a QO_{AI} e $QOVP_{DT}$, tem-se a média de 0,5 e desvio padrão de 0,2. Mediante os resultados observa-se que mais de 80% das proporções estão dentro da média.

Analisando a proporção entre QO_{AI} e QO_{DT} , tem-se a média de 0,4 e desvio padrão de 0,2. Mediante os resultados observa-se que mais de 80% das proporções estão dentro da média.

A predição positiva para todos os DADs ficou acima de 81%, que pode ser considerada um boa representatividade da quantidade de valores verdadeiros-positivos.

Para a quantidade de ocorrências de VP dos arquivos interpretáveis ($QOVP_{AI}$), a predição positiva foi superior a 89%, que também pode ser considerada uma boa representatividade.

Ao analisar os resultados das proporções de ❶, ❷ e ❸, verifica-se que as quantidades de ocorrências submetidas ao processo de filtragem (❶ e ❸) apresentaram maior quantidade de proporções dentro da média e desvio padrão.

Conforme pode ser observado, houve convergência entre os resultados do *Log* dedo-duro da ferramenta CS (QO_{DT}), dos resultados da ferramenta FiltroP (VP no DAD todo ($QOVP_{DT}$)) e Arquivos Interpretáveis ((QO_{AI}) e VP nos Arquivos Interpretáveis ($QOVP_{AI}$)), tendo sido utilizada como métrica, a quantidade de ocorrências de PChs encontradas.

Na prática, o tempo necessário para obter a $QOVP_{AI}$ demanda um tempo e esforço maior, quando comparado com o tempo e esforço para obter QO_{DT} e $QOVP_{DT}$.

Por meio dessas análises, é possível admitir que para os propósitos do método proposto nesta tese, os resultados provenientes da quantidade de ocorrências de PChs de QO_{DT} e $QOVP_{DT}$ são satisfatórios para executar os processos de triagem e ranqueamento.

5.6.2 Quantidade de ocorrências de PChs no DAD todo (QO_{D_T})

A seguir, na Tabela 5.19 é apresentada a quantidade de ocorrências de PChs por DAD. Isso possibilita uma visualização global de cada PCh em cada um dos 24 DADs.

Tabela 5.19 – Quantidade de ocorrências de PChs no DAD todo (QO_{D_T})

	Palavras-chave	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	immortality	jaimé	Jo Smith	jo@m57.biz	johnson01	mcgoo01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	Total
7	418	21184	2179	7456	4212	740	0	0	5917	2748	5784	2447	1090	62	0	5170	6	3681	2323	0	14090										
8	452	41292	2199	8664	3812	770	0	0	4989	2746	6194	3012	1250	68	0	6530	6	39130	2375	0	14373										
9	571	41907	2250	531332	3814	900	0	0	5765	2582	6799	3802	1650	2960	6460	1863	1062933	0	18328												
10	564	42043	2247	421615	3614	850	0	0	5137	30107	68103	3930	1750	3420	6050	1556	982920	0	18238												
11	631	42148	2277	491679	3616	930	0	0	5364	18107	76111	4063	2200	3360	5540	1868	923161	0	19123												
16	909	42571	2381	882272	5430	1170	0	0	5533	14121	127158	5379	3660	3740	5904	31101	923342	0	22660												
21	0	0	2490	0	0	714	0	0	4322	128	0	0	600	320	1470	0	0	8880	5739												
22	367	0	2580	490	816	14130	1	0	6508	158	43	68862	240	327	1540	0	490	9520	9565												
29	1282	01259	2261	16546	4192	5210	0	0	3761	4318	188241	2072	2040	560	1970	49219	0	16780	12746												
33	2181	01549	4437	23775	28204	9270	0	0	7137	1329	331419	2442	2100	701	2220	93319	0	23050	19719												
44	1313	0	8190	1450	913	12125	0	17	8708	1870	315678	746102	0	380	1560	31332	0	9340	14581												
46	1942	0	8130	2400	1316	60244	0	8	13705	2365	516976	2007	950	400	1720	107427	0	10170	22486												
47	2876	0	7890	3040	1215	60367	0	10	15336	1884	8461377	2201	1530	400	1670	222746	0	9650	26588												
48	3133	0	9920	3520	1218	72428	0	6	16845	1883	9551579	3004	2050	440	1730	220698	0	9400	29777												
52	3724	0	9200	4880	2612	19114	5550	0	920162	7276	9901780	3127	1480	500	1780	253797	0	11740	34674												
53	3550	0	9140	43118	1219	95503	0	10	19216	6376	9481738	3061	1440	480	1780	240790	0	11390	33193												
54	3892	0	9310	47027	1216	107543	0	9	19071	6576	10341897	3151	1440	480	1730	268871	0	11060	33911												
63	1051	0	3360	330	94	0	8510	0	7404	258	66302	1490	0	350	2240	5596	192329	0	12690												
70	2602	01438	0	3130	160	67333	2	0	13045	3216	149658	1378	10	0	560	3260	1721190	273094	24924												
71	2986	01806	0	3850	170	81400	4	0	13913	3122	157736	1863	10	0	640	3300	1911344	312992	27363												
72	3464	01830	0	4430	161	81468	5	0	11624	1416	172899	1867	12	0	520	3010	1911522	313080	26089												
75	5284	01434	0	6450	145	102644	4	0	10750	1516	2631491	1482	12	0	480	2990	2302308	403300	28386												
77	6188	01957	4	7000	145	127713	3	0	10120	1116	3421785	1964	9	0	480	2920	2862756	303618	30988												
78	6244	01876	4	6800	135	122709	2	0	9536	116	3661837	1936	9	0	470	2930	2882784	253555	30348												

5.7 Experimento 5

O objetivo deste experimento é a partir dos resultados provenientes do Experimento 4, realizar a triagem e ranqueamento dos 24 DADs, baseando na quantidade de ocorrências de PChs dos arquivos interpretáveis (QOVPAI) e DAD todo (QOV PDT).

Serão realizadas duas comparações: para o ranqueamento da triagem sem prioridade de PChs e com prioridade. O experimento é dividido em duas partes: A e B, conforme descritas a seguir.

- ❖ Na parte A, é realizado o ranqueamento sem prioridades, por este motivo não há escola de PChs.
- ❖ Na parte B, é realizado o ranqueamento com prioridades. As palavras chaves utilizadas são: @m57.biz, Jo Smith, pat@m57.biz

Então, são analisados os ranqueamentos obtidos para verificar se há defasagem⁴² entre QOVPAI e QOV PDT. Além disso, deseja-se verificar a aderência da sequência desses ranqueamentos baseados nos resultados da ferramenta *CopyScan* com uma ferramenta amplamente utilizada pela comunidade da área e de código livre, que neste caso é a ferramenta *Autopsy*.

5.7.1 Preparação e Procedimentos

Neste experimento serão utilizados os resultados já obtidos no Experimento 4, ou seja, VP_{PDT} apresentados na Tabela 5.20 e VP_{AI} apresentados na Tabela 5.21.

A lista de PChs, também será a mesma já utilizada no Experimento 4, conforme a Tabela 5.15, que contém a lista de PChs com 4 caracteres ou mais.

⁴² Quanto a posição da classificação de um DAD está deslocada em relação à posição em outra classificação (utilizada como referência) para o mesmo DAD.

Tabela 5.20 – Tabela agregada com quantidade de ocorrências de VP_{DT}

nº do DAD	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	immortality	jaime	Jo Smith	jo@m57.biz	johnson01	mcgoo01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	TOTAL
7	413	2	1107	2	176	7	422	32	12	74	0	0	5229	23	28	57	84	2395	109	0	58	0	473	0	6	36	81	2088	0	12914
8	447	4	1219	2	196	8	618	27	12	77	0	0	4130	14	27	61	94	2923	125	0	64	0	599	0	6	39	130	2158	0	12980
9	566	4	1833	2	247	53	1271	27	14	90	0	0	5028	22	49	67	99	3688	165	0	292	0	587	0	18	63	106	2705	0	16996
10	555	4	1967	2	244	42	1449	26	14	85	0	0	4255	23	54	68	103	3808	175	0	338	0	548	0	15	56	98	2613	0	16542
11	619	4	2070	2	272	49	1538	24	16	93	0	0	4349	12	55	75	111	3944	220	0	332	0	504	0	18	68	92	2826	0	17293
16	901	4	2487	2	377	88	2059	42	30	117	0	0	4673	11	52	127	158	5217	365	0	369	0	538	4	31	100	92	2977	0	20821
21	0	0	186	0	0	0	6	11	0	0	0	0	3661	6	8	0	0	60	0	0	28	0	125	0	0	0	0	719	0	4810
22	301	0	193	0	37	0	7	13	9	85	1	0	5690	9	8	29	49	822	22	0	29	4	132	0	0	49	0	734	0	8223
29	876	0	1183	2	172	16	515	38	39	341	0	0	2980	31	17	82	118	2066	204	0	53	0	156	0	23	152	0	1452	0	10516
33	1508	0	1464	4	271	23	681	23	92	634	0	0	6289	7	27	138	216	2387	210	0	66	0	183	0	42	215	0	1983	0	16463
44	760	0	738	0	96	0	7	8	2	92	0	10	7630	12	63	158	354	731	98	0	33	0	133	0	16	168	0	723	0	11832
46	1142	0	724	0	171	0	10	10	29	176	0	3	12333	11	60	244	512	1895	91	0	33	0	145	0	47	199	0	757	0	18592
47	1791	0	710	0	203	0	10	11	22	236	0	5	13953	7	79	449	803	2068	147	0	35	0	143	0	127	444	0	713	0	21956
48	1921	0	905	0	241	0	10	12	28	275	0	3	15197	8	76	494	892	2813	198	0	38	0	150	0	104	361	0	711	0	24437
52	2314	0	835	0	339	26	10	13	46	359	0	3	17982	63	69	444	996	2862	142	0	43	0	151	0	127	445	0	835	0	28104
53	2261	0	831	0	309	18	10	12	38	331	0	3	17139	60	69	439	987	2803	138	0	41	0	153	0	119	455	0	831	0	27047
54	2414	0	854	0	322	27	10	11	40	346	0	3	17094	62	69	459	1069	2885	138	0	41	0	151	0	122	494	0	896	0	27507
63	913	0	228	0	29	0	8	0	0	57	10	0	6196	3	8	64	248	147	0	0	33	0	201	0	4	558	17	1921	0	10645
70	2569	0	1315	0	310	0	16	0	67	328	2	0	11610	5	16	149	643	1315	6	0	56	0	299	0	149	1177	22	2671	0	22725
71	2946	0	1696	0	381	0	17	0	81	395	4	0	12444	4	22	157	722	1794	6	0	64	0	302	0	163	1326	25	2585	0	25134
72	3346	0	1751	0	423	0	16	1	80	443	5	0	10620	3	16	171	863	1815	8	0	52	0	279	0	161	1475	25	2781	0	24334
75	5114	0	1358	0	620	0	14	5	101	617	4	0	9833	4	16	260	1432	1434	8	0	48	0	276	0	194	2239	32	2966	0	26575
77	6149	0	1881	4	696	0	14	5	127	709	3	0	9112	1	16	342	1769	1924	6	0	48	0	271	0	244	2733	26	3243	0	29323
78	6206	0	1802	4	676	0	13	5	122	705	2	0	8769	1	16	366	1821	1896	6	0	47	0	273	0	253	2764	22	3248	0	29017

Tabela 5.21 – Tabela agregada com a quantidade de ocorrências de VP_{AI}

	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	immortality	jaime	Jo Smith	jo@m57.biz	johnson01	mcgoo01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	TOTAL
7	3	0	1086	2	1	7	244	5	2	0	0	0	527	0	14	0	0	1369	96	0	57	0	247	0	0	0	3	1808	0	5471
8	0	0	1148	2	0	8	322	5	2	0	0	0	528	0	14	0	0	1549	111	0	62	0	247	0	0	0	3	1883	0	5884
9	12	0	1639	2	9	40	787	5	3	1	0	0	657	0	27	1	1	1937	136	0	283	0	248	0	1	1	3	2226	0	8019
10	1	0	1845	2	1	40	880	5	2	0	0	0	661	0	26	0	0	2173	152	0	323	0	248	0	0	0	3	2317	0	8679
11	13	0	2031	2	10	47	971	5	2	0	0	0	664	0	26	1	1	2401	197	0	329	0	249	0	0	0	3	2501	0	9453
16	17	0	2346	2	13	86	1292	6	2	0	0	0	687	0	26	1	1	3176	322	0	363	0	274	0	0	0	4	2617	0	11235
21	0	0	175	0	0	0	6	5	0	0	0	0	490	0	8	0	0	47	0	0	27	0	115	0	0	0	0	445	0	1318
22	0	0	181	0	0	0	7	5	1	0	0	0	560	0	8	0	0	458	11	0	28	0	119	0	0	0	0	448	0	1826
29	47	0	1116	1	2	16	340	5	4	17	0	0	486	0	8	16	18	1278	155	0	51	0	122	0	3	10	0	1140	0	4835
33	71	0	1362	2	2	23	457	5	14	37	0	0	525	0	9	21	18	1656	190	0	63	0	122	0	3	14	0	1664	0	6258
44	98	0	674	0	18	0	7	5	0	18	0	2	2036	0	24	28	36	635	17	0	29	0	122	0	0	25	0	468	0	4242
46	212	0	678	0	38	0	7	5	6	50	0	2	2706	0	32	69	76	1330	40	0	29	0	125	0	6	47	0	469	0	5927
47	419	0	405	0	44	0	7	5	10	67	0	2	3060	0	38	108	171	1198	49	0	29	0	125	0	35	135	0	462	0	6369
48	441	0	410	0	43	0	7	5	10	73	0	2	3980	0	38	119	185	1527	60	0	29	0	125	0	37	139	0	461	0	7691
52	511	0	590	0	57	18	7	5	11	89	0	2	4060	0	38	132	207	1793	60	0	37	0	129	0	39	157	0	528	0	8470
53	535	0	590	0	57	18	7	5	11	89	0	2	4061	0	38	137	217	1795	60	0	37	0	129	0	40	171	0	515	0	8514
54	556	0	611	0	57	23	7	5	11	89	0	2	4078	0	38	142	229	1854	60	0	37	0	129	0	40	180	0	581	0	8729
63	287	0	189	0	11	0	8	0	0	19	0	0	4123	0	8	8	91	123	0	0	29	0	163	0	0	162	10	1809	0	7040
70	675	0	735	0	116	0	12	0	9	85	1	0	7024	0	8	19	153	679	0	0	31	0	172	0	28	269	15	1985	0	12016
71	1343	0	1200	0	206	0	15	0	25	176	2	0	10313	0	22	49	321	1208	2	0	55	0	256	0	60	565	20	2327	0	18165
72	1565	0	1260	0	228	0	14	0	30	204	3	0	8769	0	16	64	389	1228	4	0	45	0	235	0	62	666	22	2335	0	17139
75	2850	0	1310	0	371	0	14	4	46	328	4	0	8693	0	16	136	806	1312	4	0	45	0	229	0	89	1237	23	2601	0	20118
77	3411	0	1357	1	400	0	13	4	68	374	2	0	7596	0	16	164	986	1350	0	0	43	0	225	0	111	1497	19	2799	0	20436
78	3766	0	1272	1	433	0	13	2	59	406	2	0	7356	0	16	177	1108	1335	0	0	43	0	225	0	120	1641	16	2735	0	20726

A seguir, na Tabela 5.22 são apresentados os resultados da quantidade de ocorrências de PCh para cada DAD, obtidos por meio da ferramenta *Autopsy*.

Tabela 5.22 – Tabela agregada com a quantidade de ocorrências de PCh no *Autopsy*

	Nº DAD		mcgoo01	johnson01	smith01	brown01	VNC.exe	netcat	Project2400	swexpert.com	steg	prior art	Nitroba	Patent	virus	jaimé	money	Terry Johnson	Jo Smith	Pat McGoo	Charlie Brown	terry@m57.biz	jo@m57.biz	pat@m57.biz	charlie@m57.biz	cancer	@m57.biz	immortality	prostate	Epithelial	US000	total
	7	0	0	0	1	0	4	1	0	22	47	33	225	333	13	92	2	7	6	2	2	6	6	24	47	77	29	25	16	0	1020	
	8	0	0	0	2	0	4	1	0	26	58	36	253	324	13	93	2	7	7	2	3	7	7	27	57	84	32	29	21	0	1095	
	9	0	0	0	2	0	4	1	0	25	75	45	297	356	13	96	9	8	8	2	10	8	8	35	76	97	52	39	35	0	1301	
	10	0	0	0	2	0	4	1	0	23	85	47	330	323	13	95	5	7	6	2	5	6	6	38	79	95	63	41	32	0	1308	
	11	0	0	0	2	0	4	1	0	23	97	48	347	326	13	94	5	6	6	2	5	5	6	40	80	98	63	40	36	0	1347	
	16	0	0	0	2	0	4	6	3	24	143	53	453	331	19	109	7	8	8	2	7	7	8	39	101	93	84	49	45	0	1605	
	21	0	0	0	0	0	1	0	0	16	0	0	65	282	9	73	0	0	0	0	0	0	0	25	31	8	14	10	0	534		
	22	0	0	3	0	0	3	0	0	15	8	0	94	292	10	89	0	9	9	0	6	12	9	6	31	56	9	14	11	0	686	
	29	0	0	0	0	0	2	1	0	21	45	2	210	300	10	79	2	17	17	2	5	23	17	7	53	64	28	21	23	0	949	
	33	0	0	1	0	0	3	2	0	25	62	2	342	333	11	92	3	34	26	6	6	40	17	8	89	112	40	29	36	0	1319	
	44	3	0	0	0	0	2	9	1	15	13	6	128	300	12	130	4	2	34	0	17	14	34	14	36	77	10	17	12	0	890	
	46	3	0	0	0	0	2	9	1	15	20	11	232	309	13	177	8	8	63	0	28	24	49	24	41	103	11	18	13	0	1182	
	47	3	0	0	0	0	2	13	1	16	28	14	281	294	12	188	27	11	105	0	48	32	85	31	41	148	10	17	12	0	1419	
	48	3	0	0	0	0	3	13	1	16	36	14	320	289	13	199	25	10	111	0	46	35	87	30	46	156	10	19	12	0	1494	
	52	3	0	0	0	0	5	12	1	19	36	14	388	308	20	224	28	12	138	0	53	42	115	35	59	210	10	24	18	0	1774	
	53	3	0	0	0	0	5	12	1	19	36	13	391	296	21	225	29	12	146	0	58	43	122	36	59	215	10	24	17	0	1793	
	54	3	0	0	0	0	5	12	1	19	36	14	400	287	20	235	29	12	153	0	61	43	127	36	62	215	10	26	20	0	1826	
	63	0	2	0	0	0	2	0	0	40	0	0	207	365	8	158	20	0	45	0	69	12	56	12	28	88	6	12	9	0	1139	
	70	0	0	0	0	0	3	0	0	45	4	0	245	348	7	181	81	32	90	0	114	54	102	47	41	153	9	15	12	0	1583	
	71	0	0	0	0	0	2	0	0	43	2	0	246	350	7	179	81	34	90	0	115	52	102	47	38	152	8	13	8	0	1569	
	72	0	0	0	0	0	2	0	0	44	2	0	249	352	7	197	93	37	105	0	131	57	119	49	41	168	10	13	8	0	1684	
	75	0	0	0	0	0	2	0	0	43	2	0	253	364	6	191	94	35	107	0	141	57	119	57	44	179	9	12	8	0	1723	
	77	0	0	0	0	0	2	0	0	46	3	0	269	370	7	192	123	36	140	1	167	65	156	56	41	212	10	12	9	0	1917	
	78	0	0	0	0	0	2	0	0	42	3	0	269	385	6	192	121	37	141	1	168	69	158	60	42	214	11	11	8	0	1940	

A partir dessas tabelas, serão realizadas as triagens e ranqueamentos necessários para obter os resultados deste experimento.

5.6.2 Resultados do Experimento

Analisando a Tabela 5.20, verifica-se que não ocorre casos em que a quantidade de todas as PChs para o mesmo DAD são iguais a zero, portanto, todos passam pela triagem e são válidos para o ranqueamento.

5.6.2.1 Resultados do Experimento – Parte A

Pelo fato de não ter prioridade para as PChs, então, é analisada quantidade total de ocorrências de PChs em cada DAD. E a partir desse valor, os DADs são ranqueados em ordem decrescente, obtendo-se a sequência mostrada na coluna de Ranqueamento de $QOVP_{DT}$ da Tabela 5.23.

Analogamente, os procedimentos realizados para a Tabela 5.20, são executados para a Tabela 5.21, obtendo-se a sequência mostrada na coluna de Ranqueamento de $QOVP_{AI}$.

A defasagem entre os ranqueamentos de $QOVP_{AI}$ e $QOVP_{DT}$ é obtida a partir da diferença entre as posições relativas para os mesmos DADs.

Tabela 5.23 – Comparação entre o ranqueamento de $QOVP_{AI}$ e $QOVP_{DT}$ (sem prioridade)

Sequência de Classificação	Ranqueamento de $QOVP_{AI}$	Ranqueamento de $QOVP_{DT}$	Defasagem entre $QOVP_{AI}$ e $QOVP_{DT}$
1	78	77	1
2	77	78	1
3	75	52	3
4	71	54	3
5	72	53	4
6	70	75	4
7	16	71	5
8	11	48	6
9	54	72	5
10	10	70	6
11	53	47	6
12	52	16	9
13	9	46	2
14	48	11	6
15	63	9	6
16	47	10	5
17	33	33	0
18	46	8	5
19	8	7	1
20	7	44	1
21	29	63	1
22	44	29	2
23	22	22	0
24	21	21	0
MEDIA			3,4
D. PADRÃO			2,4

A coluna defasagem cujos valores são iguais a zero, mostra as posições das sequências em que houve coincidência na classificação dos DADs. Também, é possível verificar as posições das sequências em

que não houveram coincidências e quanto os DADs estão defasados entre si nas posições das sequências de classificação.

Conforme pode ser observado, os valores da média e do desvio padrão são 3,4 e 2,4 respectivamente, o que significa que em média, os mesmos DADs estão deslocados de 3,4 posições entre os ranqueamentos. Os sete primeiros colocados estão entre 1 e 5,8. De acordo com os resultados da tabela, estão dentro da média.

Na prática, os DADs #77 e #78 estão nas duas primeiras posições em ambos ranqueamentos. Assim, baseando-se nesses resultados, a sequência sugerida para análise é: #77, #78, #52, #54, #53, #75 e #71.

Os ranqueamentos obtidos anteriormente, serão comparados com o resultado da ferramenta *Autopsy*, mostrado na Tabela 5.22. Com isso, deseja-se verificar qual dos dois ranqueamentos mais se aproximam com o resultado da ferramenta *Autopsy*.

Tabela 5.24 – Comparação entre os ranqueamentos de QOVP_{AI}, QOVP_{DT} e *Autopsy* sem prioridade

Sequência de Classificação	Ranqueamento de QOVP _{AI}	Ranqueamento de QOVP _{DT}	Ranqueamento da AUTOPSY	Defasagem entre QOVP _{AI} x AUTOPSY	Defasagem entre QOVP _{DT} x AUTOPSY
1	78	77	78	0	1
2	77	78	77	0	1
3	75	52	54	3	2
4	71	54	53	6	1
5	72	53	52	2	1
6	70	75	75	3	0
7	16	71	72	1	3
8	11	48	16	5	3
9	54	72	70	6	2
10	10	70	71	5	1
11	53	47	48	7	1
12	52	16	47	7	4
13	9	46	11	3	4
14	48	11	33	3	1
15	63	9	10	3	1
16	47	10	9	4	1
17	33	33	46	3	3
18	46	8	63	1	1
19	8	7	8	0	1
20	7	44	7	0	2
21	29	63	29	0	3
22	44	29	44	0	1
23	22	22	22	0	0
24	21	21	21	0	0
MÉDIA				2,6	1,6
D. PADRÃO				2,4	1,2

A média da defasagem de entre QOVP_{AI} e *Autopsy* é 2,6 e o desvio padrão 2,4. Enquanto que a média da defasagem de entre QOVP_{DT} e

Autopsy é 1,6 e o desvio padrão 1,2. QOVP_{AI} apresenta duas correspondências com *Autopsy* nas duas primeiras colocações. Enquanto QOVP_{DT}, a primeira correspondência ocorre somente na sexta colocação. Analisando estas informações, tanto QOVP_{AI} quanto QOVP_{DT}, apresentam vantagem e desvantagens.

Na prática, se forem considerados os resultados de QOVP_{AI} os DADs #78, #77 e #75, nesta ordem devem ser analisados primeiramente, observando que até esta colocação estes DADs se encontram dentro da média da defasagem entre QOVP_{AI} e *Autopsy*.

Por outro lado, se forem considerados os resultados de QOVP_{DT}, a sequência sugerida é #77, #78#, #52, #54, #53 e #75.

5.6.2.2 Resultados do Experimento – Parte B

Nesta parte do experimento, serão consideradas as prioridades para um grupo de PChs, neste caso, essas palavras são: @m57.biz, Jo Smith, pat@m57.biz, as três possuem a mesma prioridade. Então, é efetuada a soma da ocorrência entre elas, conforme o resultado mostrado na Tabela 5.25.

Verifica-se a quantidade total de ocorrências do grupo de PChs para cada DAD, e classifica estas ocorrências em decrescente. Os resultados dessa ordenação são apresentados na Tabela 5.25, na coluna do n° de DADs.

Analogamente, os procedimentos são executados para a Tabela 5.26, obtendo-se a sequência mostrada na coluna do n° de DADs.

Tabela 5.25 – Tabela agregada com quantidade de ocorrências de VP_{DT} com prioridade

nº do DAD	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	immortality	jaime	Jo Smith	jo@m57.biz	johnson01	mcgoo01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	TOTAL	Σ PChs com Prior.
78	6206	0	1802	4	676	0	13	5	122	705	2	0	8769	1	16	366	1821	1896	6	0	47	0	273	0	253	2764	22	3248	0	29017	8149
77	6149	0	1881	4	696	0	14	5	127	709	3	0	9112	1	16	342	1769	1924	6	0	48	0	271	0	244	2733	26	3243	0	29323	8045
75	5114	0	1358	0	620	0	14	5	101	617	4	0	9833	4	16	260	1432	1434	8	0	48	0	276	0	194	2239	32	2966	0	26575	6647
72	3346	0	1751	0	423	0	16	1	80	443	5	0	10620	3	16	171	863	1815	8	0	52	0	279	0	161	1475	25	2781	0	24334	4289
71	2946	0	1696	0	381	0	17	0	81	395	4	0	12444	4	22	157	722	1794	6	0	64	0	302	0	163	1326	25	2585	0	25134	3749
54	2414	0	854	0	322	27	10	11	40	346	0	3	17094	62	69	459	1069	2885	138	0	41	0	151	0	122	494	0	896	0	27507	3523
52	2314	0	835	0	339	26	10	13	46	359	0	3	17982	63	69	444	996	2862	142	0	43	0	151	0	127	445	0	835	0	28104	3356
53	2261	0	831	0	309	18	10	12	38	331	0	3	17139	60	69	439	987	2803	138	0	41	0	153	0	119	455	0	831	0	27047	3286
70	2569	0	1315	0	310	0	16	0	67	328	2	0	11610	5	16	149	643	1315	6	0	56	0	299	0	149	1177	22	2671	0	22725	3279
48	1921	0	905	0	241	0	10	12	28	275	0	3	15197	8	76	494	892	2813	198	0	38	0	150	0	104	361	0	711	0	24437	2841
47	1791	0	710	0	203	0	10	11	22	236	0	5	13953	7	79	449	803	2068	147	0	35	0	143	0	127	444	0	713	0	21956	2616
33	1508	0	1464	4	271	23	681	23	92	634	0	0	6289	7	27	138	216	2387	210	0	66	0	183	0	42	215	0	1983	0	16463	1816
46	1142	0	724	0	171	0	10	10	29	176	0	3	12333	11	60	244	512	1895	91	0	33	0	145	0	47	199	0	757	0	18592	1683
63	913	0	228	0	29	0	8	0	0	57	10	0	6196	3	8	64	248	147	0	0	33	0	201	0	4	558	17	1921	0	10645	1161
44	760	0	738	0	96	0	7	8	2	92	0	10	7630	12	63	158	354	731	98	0	33	0	133	0	16	168	0	723	0	11832	1116
16	901	4	2487	2	377	88	2059	42	30	117	0	0	4673	11	52	127	158	5217	365	0	369	0	538	4	31	100	92	2977	0	20821	1089
29	876	0	1183	2	172	16	515	38	39	341	0	0	2980	31	17	82	118	2066	204	0	53	0	156	0	23	152	0	1452	0	10516	1033
11	619	4	2070	2	272	49	1538	24	16	93	0	0	4349	12	55	75	111	3944	220	0	332	0	504	0	18	68	92	2826	0	17293	746
9	566	4	1833	2	247	53	1271	27	14	90	0	0	5028	22	49	67	99	3688	165	0	292	0	587	0	18	63	106	2705	0	16996	679
10	555	4	1967	2	244	42	1449	26	14	85	0	0	4255	23	54	68	103	3808	175	0	338	0	548	0	15	56	98	2613	0	16542	672
8	447	4	1219	2	196	8	618	27	12	77	0	0	4130	14	27	61	94	2923	125	0	64	0	599	0	6	39	130	2158	0	12980	553
7	413	2	1107	2	176	7	422	32	12	74	0	0	5229	23	28	57	84	2395	109	0	58	0	473	0	6	36	81	2088	0	12914	509
22	301	0	193	0	37	0	7	13	9	85	1	0	5690	9	8	29	49	822	22	0	29	4	132	0	0	49	0	734	0	8223	359
21	0	0	186	0	0	0	6	11	0	0	0	0	3661	6	8	0	0	60	0	0	28	0	125	0	0	0	0	719	0	4810	0

Tabela 5.26 – Tabela agregada com quantidade de ocorrências de VP_{AI} com prioridade

Palavras-chave	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	Immortality	jaime	Jo Smith	jo@m57.biz	johnson01	mcgoo01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	TOTAL	Σ PChs com Pr..
78	3766	0	1272	1	433	0	13	2	59	406	2	0	7356	0	16	177	1108	1335	0	0	43	0	225	0	120	1641	16	2735	0	20726	4933
77	3411	0	1357	1	400	0	13	4	68	374	2	0	7596	0	16	164	986	1350	0	0	43	0	225	0	111	1497	19	2799	0	20436	4465
75	2850	0	1310	0	371	0	14	4	46	328	4	0	8693	0	16	136	806	1312	4	0	45	0	229	0	89	1237	23	2601	0	20118	3702
72	1565	0	1260	0	228	0	14	0	30	204	3	0	8769	0	16	64	389	1228	4	0	45	0	235	0	62	666	22	2335	0	17139	1984
71	1343	0	1200	0	206	0	15	0	25	176	2	0	10313	0	22	49	321	1208	2	0	55	0	256	0	60	565	20	2327	0	18165	1689
70	675	0	735	0	116	0	12	0	9	85	1	0	7024	0	8	19	153	679	0	0	31	0	172	0	28	269	15	1985	0	12016	837
54	556	0	611	0	57	23	7	5	11	89	0	2	4078	0	38	142	229	1854	60	0	37	0	129	0	40	180	0	581	0	8729	796
53	535	0	590	0	57	18	7	5	11	89	0	2	4061	0	38	137	217	1795	60	0	37	0	129	0	40	171	0	515	0	8514	763
52	511	0	590	0	57	18	7	5	11	89	0	2	4060	0	38	132	207	1793	60	0	37	0	129	0	39	157	0	528	0	8470	729
48	441	0	410	0	43	0	7	5	10	73	0	2	3980	0	38	119	185	1527	60	0	29	0	125	0	37	139	0	461	0	7691	636
47	419	0	405	0	44	0	7	5	10	67	0	2	3060	0	38	108	171	1198	49	0	29	0	125	0	35	135	0	462	0	6369	600
63	287	0	189	0	11	0	8	0	0	19	0	0	4123	0	8	8	91	123	0	0	29	0	163	0	0	162	10	1809	0	7040	378
46	212	0	678	0	38	0	7	5	6	50	0	2	2706	0	32	69	76	1330	40	0	29	0	125	0	6	47	0	469	0	5927	294
44	98	0	674	0	18	0	7	5	0	18	0	2	2036	0	24	28	36	635	17	0	29	0	122	0	0	25	0	468	0	4242	134
33	71	0	1362	2	2	23	457	5	14	37	0	0	525	0	9	21	18	1656	190	0	63	0	122	0	3	14	0	1664	0	6258	103
29	47	0	1116	1	2	16	340	5	4	17	0	0	486	0	8	16	18	1278	155	0	51	0	122	0	3	10	0	1140	0	4835	69
16	17	0	2346	2	13	86	1292	6	2	0	0	0	687	0	26	1	1	3176	322	0	363	0	274	0	0	0	4	2617	0	11235	20
9	12	0	1639	2	9	40	787	5	3	1	0	0	657	0	27	1	1	1937	136	0	283	0	248	0	1	1	3	2226	0	8019	16
11	13	0	2031	2	10	47	971	5	2	0	0	0	664	0	26	1	1	2401	197	0	329	0	249	0	0	0	3	2501	0	9453	16
7	3	0	1086	2	1	7	244	5	2	0	0	0	527	0	14	0	0	1369	96	0	57	0	247	0	0	0	3	1808	0	5471	5
10	1	0	1845	2	1	40	880	5	2	0	0	0	661	0	26	0	0	2173	152	0	323	0	248	0	0	0	3	2317	0	8679	3
8	0	0	1148	2	0	8	322	5	2	0	0	0	528	0	14	0	0	1549	111	0	62	0	247	0	0	0	3	1883	0	5884	2
22	0	0	181	0	0	0	7	5	1	0	0	0	560	0	8	0	0	458	11	0	28	0	119	0	0	0	448	0	1826	1	
21	0	0	175	0	0	0	6	5	0	0	0	0	490	0	8	0	0	47	0	0	27	0	115	0	0	0	445	0	1318	0	

A defasagem entre os ranqueamentos de $QOVP_{AI}$ e $QOVP_{DT}$ é obtida a partir da diferença entre as posições relativas para os mesmos DADs.

Tabela 5.27 – Comparação entre o ranqueamento de $QOVP_{AI}$ e $QOVP_{DT}$ com prioridade

Sequência de Classificação	Ranqueamento de $QOVP_{AI}$	Ranqueamento de $QOVP_{DT}$	Defasagem entre $QOVP_{AI}$ e $QOVP_{DT}$
1	78	78	0
2	77	77	0
3	75	75	0
4	72	72	0
5	71	71	0
6	70	54	3
7	54	52	1
8	53	53	0
9	52	70	2
10	48	48	0
11	47	47	0
12	63	33	2
13	46	46	0
14	44	63	1
15	33	44	3
16	29	16	1
17	16	29	1
18	11	11	0
19	9	9	0
20	7	10	2
21	10	8	1
22	8	7	1
23	22	22	0
24	21	21	0
MED.			0,8
D. Padrão.			1,0

Igualmente ao apresentado na parte A, a defasagem informa quais são as posições em que houveram coincidências (valor igual a zero), e nas posições das sequências em que não houveram coincidência, o quanto os DADs estão defasados entre si nas posições da sequência de classificação.

O valor médio da defasagem entre o ranqueamento de $QOVP_{AI}$ e o ranqueamento $QOVP_{DT}$ é 0,8 com desvio padrão de 1,0.

Baseando-se nos resultados apresentados na tabela, tem-se duas possíveis sequências para análise. A análise dos resultados mostra que

as cinco primeiras posições são coincidentes quando se compara a quantidade de ocorrências de PChs de QOVP_{AI} com QOVP_{DT}.

A Tabela 5.28 apresenta o ranqueamento dos DAD obtido pelo resultado da quantidade de ocorrências da ferramenta *Autopsy*, considerando que as PChs: @m57.biz, Jo Smith e pat@m57.biz, possuem a mesma prioridade.

Tabela 5.28 – Tabela agregada com quantidade ocorrências Autopsy com prioridade

	Nº DAD	mcgoo01	johnson01	smith01	brown01	VNC.exe	netcat	Project2400	swexpert.com	steg	prior art	Nitroba	Patent	virus	jaimé	money	Terry Johnson	Jo Smith	Pat McGoo	Charlie Brown	terry@m57.biz	jo@m57.biz	pat@m57.biz	charlie@m57.biz	cancer	@m57.biz	immortality	prostate	Epithelial	US000	total	SOMA C/ PRIOR.
	78	0	0	0	0	0	2	0	0	42	3	0	269	385	6	192	121	37	141	1	168	69	158	60	42	214	11	11	8	0	1940	409
	77	0	0	0	0	0	2	0	0	46	3	0	269	370	7	192	123	36	140	1	167	65	156	56	41	212	10	12	9	0	1917	404
	54	3	0	0	0	0	5	12	1	19	36	14	400	287	20	235	29	12	153	0	61	43	127	36	62	215	10	26	20	0	1826	354
	53	3	0	0	0	0	5	12	1	19	36	13	391	296	21	225	29	12	146	0	58	43	122	36	59	215	10	24	17	0	1793	349
	52	3	0	0	0	0	5	12	1	19	36	14	388	308	20	224	28	12	138	0	53	42	115	35	59	210	10	24	18	0	1774	337
	75	0	0	0	0	0	2	0	0	43	2	0	253	364	6	191	94	35	107	0	141	57	119	57	44	179	9	12	8	0	1723	333
	72	0	0	0	0	0	2	0	0	44	2	0	249	352	7	197	93	37	105	0	131	57	119	49	41	168	10	13	8	0	1684	324
	71	0	0	0	0	0	2	0	0	43	2	0	246	350	7	179	81	34	90	0	115	52	102	47	38	152	8	13	8	0	1569	288
	70	0	0	0	0	0	3	0	0	45	4	0	245	348	7	181	81	32	90	0	114	54	102	47	41	153	9	15	12	0	1583	287
	48	3	0	0	0	0	3	13	1	16	36	14	320	289	13	199	25	10	111	0	46	35	87	30	46	156	10	19	12	0	1494	253
	47	3	0	0	0	0	2	13	1	16	28	14	281	294	12	188	27	11	105	0	48	32	85	31	41	148	10	17	12	0	1419	244
	33	0	0	1	0	0	3	2	0	25	62	2	342	333	11	92	3	34	26	6	6	40	17	8	89	112	40	29	36	0	1319	163
	46	3	0	0	0	0	2	9	1	15	20	11	232	309	13	177	8	8	63	0	28	24	49	24	41	103	11	18	13	0	1182	160
	63	0	2	0	0	0	2	0	0	40	0	0	207	365	8	158	20	0	45	0	69	12	56	12	28	88	6	12	9	0	1139	144
	9	0	0	0	2	0	4	1	0	25	75	45	297	356	13	96	9	8	8	2	10	8	8	35	76	97	52	39	35	0	1301	113
	44	3	0	0	0	0	2	9	1	15	13	6	128	300	12	130	4	2	34	0	17	14	34	14	36	77	10	17	12	0	890	113
	11	0	0	0	2	0	4	1	0	23	97	48	347	326	13	94	5	6	6	2	5	5	6	40	80	98	63	40	36	0	1347	110
	16	0	0	0	2	0	4	6	3	24	143	53	453	331	19	109	7	8	8	2	7	7	8	39	101	93	84	49	45	0	1605	109
	10	0	0	0	2	0	4	1	0	23	85	47	330	323	13	95	5	7	6	2	5	6	6	38	79	95	63	41	32	0	1308	108
	8	0	0	0	2	0	4	1	0	26	58	36	253	324	13	93	2	7	7	2	3	7	7	27	57	84	32	29	21	0	1095	98
	29	0	0	0	0	0	2	1	0	21	45	2	210	300	10	79	2	17	17	2	5	23	17	7	53	64	28	21	23	0	949	98
	7	0	0	0	1	0	4	1	0	22	47	33	225	333	13	92	2	7	6	2	2	6	6	24	47	77	29	25	16	0	1020	90
	22	0	0	3	0	0	3	0	0	15	8	0	94	292	10	89	0	9	9	0	6	12	9	6	31	56	9	14	11	0	686	74
	21	0	0	0	0	0	1	0	0	16	0	0	65	282	9	73	0	0	0	0	0	0	0	25	31	8	14	10	0	534	31	

Os resultados das Tabelas 5.25 e 5.26 serão comparados com os resultado da ferramenta *Autopsy* para verificar qual dos dois ranqueamentos (QOVP_{AI} e QOVP_{DT}) converge com o ranqueamento da ferramenta *Autopsy*. Essa comparação é apresentada a seguir na Tabela 5.29.

Tabela 5.29 – Comparação entre QOVP_{AI}, QOVP_{DT} e Autopsy com prioridade

Sequência de Classificação	Rankeamento de QOVP _{AI}	Rankeamento de QOVP _{DT}	Rankeamento da AUTOPSY	Defasagem entre QOVP _{AI} x AUTOPSY	Defasagem entre QOVP _{DT} x AUTOPSY
1	78	78	78	0	0
2	77	77	77	0	0
3	75	75	54	3	3
4	72	72	53	3	3
5	71	71	52	3	3
6	70	54	75	3	3
7	54	52	72	4	2
8	53	53	71	4	4
9	52	70	70	4	0
10	48	48	48	0	0
11	47	47	47	0	0
12	63	33	33	2	0
13	46	46	46	0	0
14	44	63	63	2	0
15	33	44	9	3	1
16	29	16	44	5	2
17	16	29	11	1	4
18	11	11	16	1	1
19	9	9	10	4	4
20	7	10	8	2	1
21	10	8	29	2	1
22	8	7	7	2	0
23	22	22	22	0	0
24	21	21	21	0	0
MEDIA				2,0	1,3
D. PADRÃO				1,6	1,5

A média da defasagem e o desvio padrão entre QOVP_{AI} e a ferramenta *Autopsy*, considerando a prioridade das três PChs é 2,0 e 1,6, respectivamente. Enquanto a média da defasagem entre QOVP_{DT} e a ferramenta *Autopsy*, é 1,3 e o desvio padrão 1,5, considerando as mesmas PChs.

Em ambas as comparações, os DADs #78 e #77 se apresentaram no início da sequência sugerida para análise.

Neste caso, a defasagem entre o ranqueamento de QOVP_{DT} e a ferramenta *Autopsy* é menor ou igual a defasagem do ranqueamento de QOVP_{AI} e a ferramenta *Autopsy* nas dezesseis primeiras posições da sequência.

5.7 Experimento 6

O objetivo do Experimento 6 é obter a sequência em que os DADs devem ser analisados na etapa de análise para o conjunto das 79 imagens do caso M57-*Patents*.

Para isso serão aplicados os procedimentos de (Duplicação, Varredura e Filtragem) descritos na Seção 4.4, para obter a tabela agregada. Então sobre a tabela deve ser aplicado os processos de triagem e ranqueamento. Com isso, apresenta-se o método em sua integridade, com todos os processos.

5.7.1 Preparação e Procedimentos

Para cada imagem foram realizados os processos de duplicação, varredura e filtragem, obtendo-se a tabela agregada, conforme mostra a Tabela 5.30.

Para todos os 79 DADs é necessário executar CS e FIP, obtendo assim a lista de $QOVP_{DT}$ de cada um dos DADs. Esta tabela é organizada agregando-se todos os resultados de $QOVP_{DT}$.

Baseado no que foi abordado na Seção 4.4.3.1 (Critérios Para o Processo de Triagem), em síntese, a partir da tabela agregada, verifica-se dentre os resultados se há DADs no qual a quantidade de ocorrências de PChs são todas iguais a zero. Caso haja, este DAD, deve ser separado dos demais para verificação posterior. Para os DADs que não estejam no caso anterior, devem ser escolhidas a PCh ou PChs que serão usadas na Triagem. Os resultados da Triagem são utilizados para o ranqueamento.

Neste Experimento a título de exemplo, serão utilizadas as PChs: @m57.biz, Jo Smith e pat@m57.biz. Estas terão prioridade frente as demais PChs. A seguir, são apresentados os resultados de ocorrências de $QOVP_{DT}$, Total de ocorrências obtidas pela ferramenta Autopsy e Comparação entre $QOVP_{DT}$ e Autopsy com prioridade, nas Tabelas: 5.30, 5.31 e 5.32 respectivamente.

5.7.2 Resultados do Experimento 6

Tabela 5.30 – Total de ocorrências disco todo (QOVP_{DT})

DAD nº	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	Immortality	jaimé	Jo Smith	jo@m57.biz	johnson01	mcgoc01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	Total Ocorre.
1	13	0	263	0	13	0	9	18	0	0	0	0	4676	14	8	0	0	83	0	0	33	0	152	0	0	0	0	934	0	6216
2	0	0	252	0	0	0	7	20	0	0	0	0	4897	11	8	0	0	61	0	0	31	0	147	0	0	0	0	913	0	6347
3	174	10	264	0	94	0	9	20	0	7	0	0	5448	17	8	30	44	326	9	0	33	0	154	0	0	7	0	937	0	7591
4	152	2	890	2	80	1	201	18	0	14	0	0	4394	14	15	23	38	1098	48	0	38	0	155	0	0	5	0	1262	0	8450
5	346	2	1597	3	167	7	444	21	4	54	0	0	5217	36	16	45	66	2107	99	0	61	0	166	0	6	31	0	1494	0	11989
6	388	2	1111	2	169	7	275	42	5	64	0	0	4775	24	57	58	82	3026	92	0	59	0	1043	0	6	34	128	1515	0	12964
7	418	2	1184	2	179	7	456	42	12	74	0	0	5917	27	48	57	84	2447	109	0	62	0	517	0	6	36	81	2323	0	14090
8	452	4	1292	2	199	8	664	38	12	77	0	0	4989	27	46	61	94	3012	125	0	68	0	653	0	6	39	130	2375	0	14373
9	571	4	1907	2	250	53	1332	38	14	90	0	0	5765	25	82	67	99	3802	165	0	296	0	646	0	18	63	106	2933	0	18328
10	564	4	2043	2	247	42	1615	36	14	85	0	0	5137	30	107	68	103	3930	175	0	342	0	605	0	15	56	98	2920	0	18238
11	631	4	2148	2	277	49	1679	36	16	93	0	0	5364	18	107	76	111	4063	220	0	336	0	554	0	18	68	92	3161	0	19123
12	634	4	2222	2	279	49	1825	49	16	93	0	0	6084	20	106	74	111	4448	241	0	340	0	533	0	17	67	84	3201	0	20499
13	713	4	2271	2	318	49	1733	51	18	104	0	0	5567	19	120	83	118	4145	235	0	340	0	549	0	18	68	92	3148	0	19765
14	770	4	2281	2	337	49	2032	53	23	108	0	0	5969	37	112	88	133	4570	300	0	340	0	579	3	22	87	89	3221	0	21209
15	788	4	2468	2	335	80	2107	53	27	111	0	0	5748	18	118	98	138	4940	316	0	372	0	574	4	26	91	89	3311	0	21818
16	909	4	2571	2	381	88	2272	54	30	117	0	0	5533	14	121	127	158	5379	366	0	374	0	590	4	31	101	92	3342	0	22660
17	912	6	3288	3	387	124	2848	54	30	116	0	0	5695	14	138	127	157	6593	431	0	414	0	674	4	31	100	157	3803	0	26106
18	1265	3	3540	4	543	128	3118	60	50	158	0	0	6740	28	214	167	192	7620	492	0	547	0	575	10	47	138	103	4043	0	29785
19	1265	3	3539	4	538	128	3067	60	50	159	0	0	6741	28	214	168	195	7426	492	0	545	0	575	10	47	139	103	4041	0	29537
20	1	0	255	0	0	0	8	14	0	1	0	0	4026	19	8	0	0	72	0	0	32	0	151	0	0	0	0	929	0	5516
21	0	0	249	0	0	0	7	14	0	0	0	0	4322	12	8	0	0	60	0	0	32	0	147	0	0	0	0	888	0	5739
22	367	0	258	0	49	0	8	16	14	130	1	0	6508	15	8	43	68	862	24	0	32	7	154	0	0	49	0	952	0	9565
23	351	0	834	0	77	0	147	16	14	152	1	0	5870	33	22	63	81	1516	44	0	38	3	157	0	0	33	0	1593	0	11045
24	1936	0	1253	0	509	14	281	17	63	736	1	0	6900	45	50	265	353	2032	95	0	42	7	158	0	69	303	0	1752	0	16881
25	1512	0	1372	0	355	9	333	15	79	668	1	0	6803	36	40	223	272	2255	80	0	44	4	159	0	49	184	0	1692	0	16185
26	1055	0	248	0	172	0	8	9	107	392	0	0	3581	11	18	142	255	79	10	0	32	9	146	0	95	209	0	959	0	7537
27	1161	0	1034	0	271	9	257	30	79	478	1	0	11789	23	26	202	215	2102	68	0	43	7	177	0	47	152	0	1727	0	19898
28	1020	0	905	0	195	7	402	14	103	446	0	0	4759	23	18	146	186	1189	74	0	34	3	203	0	48	157	0	1134	0	11066
29	1282	0	1259	2	261	16	546	41	92	521	0	0	3761	43	18	188	241	2072	204	0	56	0	197	0	49	219	0	1678	0	12746
30	1402	0	1229	1	295	17	530	26	106	552	0	0	7206	21	18	227	256	2074	173	0	58	3	195	0	61	236	0	1966	0	16652
31	1091	0	1278	2	214	17	524	26	98	453	0	0	6841	25	18	172	214	1915	186	0	59	1	195	0	49	169	0	1797	0	15344
32	1109	0	1572	3	219	23	755	26	98	461	0	0	7902	24	19	175	219	2376	206	0	70	3	196	0	49	169	0	2327	0	18001
33	2181	0	1549	4	437	23	775	28	204	927	0	0	7137	13	29	331	419	2442	210	0	70	1	222	0	93	319	0	2305	0	19719
34	2320	0	1589	5	439	23	784	29	207	939	0	0	7672	50	29	357	490	2408	210	0	73	3	195	0	115	368	0	2390	0	20695
35	2467	0	1600	5	457	23	939	30	239	994	0	0	7304	29	29	386	529	2724	210	0	71	1	175	0	138	392	0	2393	0	21135
36	2467	0	1595	5	457	23	843	33	239	994	0	0	7364	13	29	386	529	2588	210	0	71	0	173	0	138	392	0	2414	0	20963
37	3598	0	1869	6	676	24	966	36	339	1458	0	0	7973	35	39	560	761	2946	227	0	269	1	178	0	203	562	0	2605	0	25331
38	3627	0	1699	4	669	20	842	38	358	1480	0	0	8104	35	39	561	764	2660	210	0	223	4	181	0	201	573	0	2771	0	25063
39	3639	0	1662	4	670	20	830	36	358	1483	0	0	7633	16	39	564	769	2645	210	0	223	0	174	0	199	576	0	2409	0	24159
40	3721	0	1664	4	673	20	830	36	406	1557	0	0	7872	24	39	565	771	2646	210	0	223	0	175	0	199	579	0	2442	0	24656
41	1	0	259	0	0	0	9	12	0	0	0	0	4050	28	8	0	0	72	0	0	32	0	151	0	0	0	0	978	0	5601
42	0	0	249	0	0	0	7	12	0	0	0	0	4168	14	8	0	0	59	0	0	31	0	146	0	0	0	0	853	0	5547
43	705	0	307	0	53	0	9	13	8	46	0	15	10529	23	8	139	423	303	23	0	37	0	153	0	0	163	0	948	0	13905
44	1313	0	819	0	145	0	9	13	12	125	0	17	8708	18	70	315	678	746	102	0	38	0	156	0	31	332	0	934	0	14581
45	2038	0	830	0	273	0	9	16	58	299	0	7	9157	22	102	476	967	2134	105	0	39	0	158	0	95	460	0	1052	0	18297
46	1942	0	813	0	240	0	13	16	60	244	0	8	13705	23	65	516	976	2007	95	0	40	0	172	0	107	427	0	1017	0	22486
47	2876	0	789	0	304	0	12	15	60	367	0	10	15336	18	84	846	1377	2201	153	0	40	0	167	0	222	746	0	965	0	26588
48	3133	0	992	0	352	0	12	18	72	428	0	6	16845	18	83	955	1579	3004	205	0	44	0	173	0	220	698	0	940	0	29777
49	3920	0	920	0	469	0	12	21	93	537	0	7	18281	46	89	1044	1945	2767	170	0	49	0	178	0	277	872	0	1023	0	32720
50	4190	0	881	0	508	0	12	21	108	574	0	7	19195	40	88	1100	2084	2848	163	0	46	0	180	0	280	896	0	1138	0	34359
51	3733	0	966	0	489	39	12	19	107	554	0	11	19720	55	76	993	1776	2833	148	0	50	0	178	0	246	817	0	1194	0	34016
52	3724	0	920	0	488	26	12	19	114	555	0	9	20162	72	76	990	1780	3127	148	0	50	0	178	0	253	797	0	1174	0	34674
53																														

Tabela 5.31 – Total de ocorrências obtidas pela ferramenta Autopsy

Nº do DAD	@m57.biz	brown01	cancer	Charlie Brown	charlie@m57.biz	Epithelial	immortality	jaimé	Jo Smith	jo@m57.biz	johnson01	mcgoo01	money	netcat	Nitroba	Pat McGoo	pat@m57.biz	Patent	prior art	Project2400	prostate	smith01	steg	swexpert.com	Terry Johnson	terry@m57.biz	US000	virus	VNC.exe	total
1	35	0	33	0	6	12	10	11	0	0	0	0	79	2	0	0	0	74	0	0	15	0	15	0	0	0	295	0	587	
2	27	0	27	0	1	10	8	9	0	0	0	0	73	1	0	0	65	0	0	13	0	17	0	0	0	276	0	527		
3	51	2	31	0	16	12	10	11	0	3	0	0	81	3	0	7	7	79	3	0	15	0	16	0	0	3	298	0	648	
4	53	1	39	2	16	15	20	11	0	5	0	0	85	3	2	5	5	130	22	1	19	0	17	0	0	2	307	0	760	
5	57	1	45	4	17	16	24	12	2	6	0	0	90	4	3	6	6	161	37	1	25	0	18	0	2	3	314	0	854	
6	66	1	43	2	19	16	22	12	2	6	0	0	90	4	22	6	6	185	40	1	23	0	22	0	2	2	322	0	914	
7	77	1	47	2	24	16	29	13	7	6	0	0	92	4	33	6	6	225	47	1	25	0	22	0	2	2	333	0	1020	
8	84	2	57	2	27	21	32	13	7	7	0	0	93	4	36	7	7	253	58	1	29	0	26	0	2	3	324	0	1095	
9	97	2	76	2	35	35	52	13	8	8	0	0	96	4	45	8	8	297	75	1	39	0	25	0	9	10	356	0	1301	
10	95	2	79	2	38	32	63	13	7	6	0	0	95	4	47	6	6	330	85	1	41	0	23	0	5	5	323	0	1308	
11	98	2	80	2	40	36	63	13	6	5	0	0	94	4	48	6	6	347	97	1	40	0	23	0	5	5	326	0	1347	
12	92	2	84	2	39	36	68	18	6	5	0	0	95	3	51	5	6	377	110	5	43	0	21	0	5	5	327	0	1405	
13	94	2	85	2	41	36	68	19	7	6	0	0	96	4	53	6	6	376	105	6	42	0	21	0	6	6	326	0	1413	
14	94	2	88	2	39	37	71	18	7	6	0	0	103	4	53	6	7	402	118	6	42	0	24	3	6	6	331	0	1475	
15	103	2	92	2	42	40	80	18	8	7	0	0	107	4	58	7	7	430	133	5	45	0	24	3	7	7	338	0	1569	
16	93	2	101	2	39	45	84	19	8	7	0	0	109	4	53	8	8	453	143	6	49	0	24	3	7	7	331	0	1605	
17	94	3	117	3	40	56	90	19	8	7	0	0	112	4	53	7	7	502	167	6	56	0	27	3	7	7	342	0	1737	
18	96	2	120	4	40	57	92	19	8	7	0	0	114	4	54	7	7	538	175	7	55	0	24	4	7	7	345	0	1793	
19	99	2	120	4	41	57	93	19	8	7	0	0	114	4	55	7	7	538	175	7	55	0	24	4	7	7	346	0	1800	
20	34	0	30	0	0	11	9	10	0	2	0	0	76	2	0	0	0	70	0	0	14	0	15	0	0	0	287	0	560	
21	31	0	25	0	0	10	8	9	0	0	0	0	73	1	0	0	0	65	0	0	14	0	16	0	0	0	282	0	534	
22	56	0	31	0	6	11	9	10	9	12	0	0	89	3	0	9	9	94	8	0	14	3	15	0	0	6	0	292	0	686
23	56	0	40	0	4	16	17	10	6	11	0	0	90	3	3	6	6	139	17	1	16	2	15	0	0	4	0	299	0	761
24	73	0	44	0	20	18	24	11	14	26	0	0	93	4	11	22	22	177	32	5	18	3	15	0	10	17	0	304	0	963
25	65	0	48	0	10	21	24	12	15	19	0	0	97	3	6	16	10	192	30	2	21	3	16	0	9	10	0	307	0	936
26	40	0	26	0	6	11	9	10	12	17	0	0	72	1	2	16	16	71	1	1	14	1	13	0	5	5	0	257	0	606
27	75	0	49	0	10	20	20	12	18	22	0	0	95	2	4	16	10	209	27	1	21	3	18	0	9	9	0	283	0	933
28	58	0	38	0	5	15	21	10	16	22	0	0	76	2	2	16	15	128	15	1	15	1	14	0	2	3	0	283	0	758
29	64	0	53	2	7	23	28	10	17	23	0	0	79	2	2	17	17	210	45	1	21	0	21	0	2	5	0	300	0	949
30	88	0	57	1	8	25	29	11	25	31	0	0	88	2	2	23	17	243	47	1	22	1	19	0	3	5	0	316	0	1064
31	94	0	62	3	7	25	29	11	25	31	0	0	90	2	1	21	16	250	48	1	24	1	21	0	2	6	0	315	0	1085
32	99	0	93	5	7	37	47	11	26	33	0	0	95	2	1	21	16	341	65	1	30	1	20	0	2	6	0	339	0	1298
33	112	0	89	6	8	36	40	11	34	40	0	0	92	3	2	26	17	342	62	2	29	1	25	0	3	6	0	333	0	1319
34	135	0	89	7	8	33	39	13	36	42	0	0	117	3	2	44	35	329	60	2	30	1	24	0	12	24	0	326	0	1411
35	147	0	87	6	9	30	38	12	42	48	0	0	113	4	2	46	37	327	59	2	29	1	22	0	12	25	0	319	0	1417
36	147	0	78	5	9	30	36	12	43	49	0	0	113	4	2	47	38	316	58	2	28	0	22	0	12	26	0	308	0	1385
37	152	0	81	6	10	32	37	12	49	55	0	0	111	5	3	46	37	323	61	3	32	1	25	0	11	23	0	315	0	1430
38	164	0	73	4	8	29	32	12	56	62	0	0	104	4	2	43	34	308	56	2	29	1	23	0	10	20	0	332	0	1408
39	164	0	72	4	8	29	31	12	56	62	0	0	104	4	2	43	34	310	56	2	29	0	23	0	10	20	0	330	0	1405
40	173	0	72	4	10	29	31	12	61	68	0	0	104	4	2	45	36	310	56	2	29	0	23	0	10	21	0	330	0	1432
41	34	0	31	0	0	12	10	10	0	0	0	0	75	2	0	0	1	70	0	0	14	0	15	0	0	0	0	285	0	559
42	31	0	25	0	0	10	8	9	0	0	0	0	71	1	0	0	0	65	0	0	13	0	15	0	0	0	0	278	0	526
43	65	0	34	0	9	12	10	12	2	9	0	3	129	2	0	19	26	110	9	0	17	0	15	2	0	9	0	293	0	787
44	77	0	36	0	14	12	10	12	2	14	0	3	130	2	6	34	34	128	13	9	17	0	15	1	4	17	0	300	0	890
45	102	0	40	0	26	12	10	13	7	26	0	3	150	2	15	60	53	206	23	9	18	0	15	2	11	35	0	308	0	1146
46	103	0	41	0	24	13	11	13	8	24	0	3	177	2	11	63	49	232	20	9	18	0	15	1	8	28	0	309	0	1182
47	148	0	41	0	31	12	10	12	11	32	0	3	188	2	14	105	85	281	28	13	17	0	16	1	27	48	0	394	0	1419
48	156	0	46	0	30	12	10	13	10	35	0	3	199	3	14	111	87	320	36	13	19	0	16	1	25	46	0	289	0	1494
49	168	0	54	0	36	16	10	18	12	42	0	3	212	4	14	121	97	345	39	16	23	0	17	1	27	52	0	298	0	1625
50	196	0	55	0	33	16	10	17	11	40	0	3	220	5	15	132	109	354	37	14	22	0	19	1	28	50	0	306	0	1693
51	206	0	58	0	33	19	10	18	12	40	0	3	216	5	15	133	112	360	37	12	25	0	19	1	28	53	0	308	0	1723
52	210	0	59	0	35	18	10	20	12	42	0	3	224	5	14	138	115	388	36	12	24	0	19	1	28	53	0	308	0	1774
53	215	0	59	0	36	17	10	21	12	43	0	3	225	5	13	146	122	391	36	12	24	0	19	1	29	58	0	296	0	1793
54	215	0	62	0	36	20	10	20	12	43	0	3	235	5	14	153	127	400	36	12	26	0	19	1	29	61	0	287	0	1826
55	225	0	61	0	36	19	11	21	15	45	0	3	234	6	14	169	144	406	36	12	25	0	19	1	34	69	0	355	0	1960
56	233	0	63	0	36	19	10	19	15	45	0	3	235	7	14	171	147	408	38	12	23	0	21	1	34	69	0	363	0	1986
57	227	0	68	0	36	19	10	19	15	45	0	3	238	6	15	175	150	412	41	12	23	0	22	1	34	76	0	359	0	2006
58	247	0	66																											

Tabela 5.32 – Comparação entre QOVP_{DT} e Autopsy com prioridade

Sequência de Classificação	Ranqueamento da AUTOPSY	Ranqueamento de QOVP _{DT}	Defasagem entre QOVP _{DT} xAUTOPSY
1	59	59	0
2	58	58	0
3	78	78	0
4	79	79	0
5	77	77	0
6	56	57	2
7	57	76	1
8	55	56	5
9	76	75	2
10	54	50	4
11	53	74	7
12	74	49	1
13	73	55	5
14	52	54	1
15	75	52	7
16	51	51	0
17	72	53	7
18	50	73	8
19	71	40	7
20	70	48	7
21	49	39	9
22	40	38	3
23	38	37	1
24	39	72	2
25	48	47	5
26	69	71	10
27	47	70	2
28	68	68	0
29	37	67	6
30	36	35	1
31	35	36	1
32	67	45	3
33	34	34	0
34	64	46	6
35	66	33	8
36	33	69	1
37	45	24	5
38	46	44	4
39	65	25	22
40	63	64	10
41	32	30	6
42	31	29	7
43	30	66	2
44	62	19	11
45	15	18	11
46	19	27	2
47	9	32	15
48	44	26	10
49	18	31	4
50	11	63	10
51	17	28	2
52	16	43	2
53	24	17	16
54	14	16	3
55	10	62	8
56	13	15	2
57	12	14	2
58	27	13	12
59	8	12	5
60	29	11	18
61	43	65	9
62	25	9	23
63	7	10	2
64	28	8	13
65	6	7	1
66	22	6	1
67	23	22	1
68	26	23	20
69	5	5	0
70	4	3	1
71	3	4	1
72	1	60	1
73	41	1	1
74	20	41	1
75	21	20	2
76	42	2	2
77	2	21	1
78	60	42	6
79	61	61	0
Média			5,0
Desvio Padrão			5,3

Analisando-se os resultados, tem-se que a média da defasagem é 5,0 e o desvio padrão 5,3.

Para estas PChs, observa-se correspondência entre as cinco primeiras posições. Na prática, a sequência sugerida para análise é: #59, #58, #78, #79 e #77.

5.8 Considerações Finais Sobre os Experimentos

Este capítulo descreveu os 6 experimentos que foram realizados com o intuito de verificar a hipótese colocada e avaliar o método proposto.

Os resultados mostraram que é possível realizar na etapa de coleta, simultaneamente à duplicação de DADs de origem para DADs de destino a varredura dos dados para procurar por PChs de interesse do perito, sem com isso impactar significativamente o tempo de duplicação. A partir das PChs obtidas durante a duplicação, foi possível realizar a triagem e ranqueamento dos DADs duplicados, de forma a apresentar uma sugestão de sequência, indicando quais desses DADs devem ser prioritariamente examinados na etapa de análise. Dessa forma, estes resultados validam o método proposto.

6 DISCUSSÃO DOS RESULTADOS E CONCLUSÕES

Este capítulo apresenta objetivamente as considerações dos resultados obtidos no Capítulo 5. Seu principal intuito é discutir os resultados, estabelecer as generalizações pertinentes, indicar a aplicabilidade dos resultados obtidos e suas limitações. A complexidade dos experimentos foi incrementada gradativamente a partir dos resultados obtidos em cada experimento, até que o método proposto fosse testado em sua integridade. Ao final são apresentadas as conclusões e trabalhos futuros.

6.1 Resultados da Duplicação e Varredura

Primeiramente são tratados os resultados com relação ao fator tempo e posteriormente os resultados em relação à quantidade de ocorrências de palavras-chave.

Um dos requisitos iniciais para a viabilidade do método proposto é que o tempo necessário para a realização da duplicação e varredura pela ferramenta a ser utilizada nos experimentos de prova de conceito não fosse

maior do que o tempo de duplicação gasto por uma ferramenta utilizada para a finalidade de duplicação.

Os resultados obtidos nos experimentos mostraram que desempenho da ferramenta CS para realizar a duplicação é igual ao de uma ferramenta já conceituada e amplamente utilizada entre a comunidade da área. Esse tempo de duplicação obtido, foi utilizado como referência para os próximos experimentos.

Prosseguindo com os requisitos necessários, foi inserido à duplicação o processo de varredura simultaneamente para procurar por uma lista com 36 palavras-chave (PCh), variando entre dois e quinze caracteres. Os resultados mostraram que é possível realizar o processo de duplicação e varredura simultaneamente sem onerar o tempo de duplicação.

Um diferencial do mecanismo para duplicação e varredura implementado é a possibilidade de procurar por cinco padrões de codificação e combinações dos caracteres maiúsculo/minúsculo. Essa funcionalidade também não onerou o tempo de duplicação utilizado como referência.

Analizando os resultados com relação à quantidade de ocorrências de PChs, foi observado um aumento na quantidade de ocorrências quando a varredura é realizada procurando por PCh combinando os caracteres (maiúsculo/minúsculo). Esse resultado mostra que esse aumento de ocorrências é maior para PChs com até três caracteres. Mediante esses resultados, foi possível notar que as PChs compostas com menos de quatro caracteres produziram uma elevada quantidade de ocorrências. Decorrente disso, observou-se uma elevada quantidade de ocorrências de falso-positivo para estas PChs. Isso pode ser um indicativo de que as PChs escolhidas pelo perito devem ter no mínimo quatro caracteres.

Além da quantidade de caracteres que compõem as PChs, é importante não deixar espaço em branco antes e após a palavra-chave, porque isso pode influenciar nos resultados da quantidade de ocorrências encontradas.

A comparação dos resultados entre a ferramenta CS e a ferramenta *Autopsy*, pode ser realizada sob duas métricas: o tempo de execução e a quantidade de ocorrências de PChs encontradas. O tempo para execução da ferramenta *Autopsy* foi significativamente maior quando comparado com a

ferramenta CS. Embora tenham demorado mais para procurar pelas PChs, encontrou uma quantidade menor de ocorrências de PChs.

A diferença de desempenho com relação ao tempo é justificada pela forma como a ferramenta *Autopsy* realiza a busca. Primeiro ela realiza a indexação do conteúdo dos dados para depois iniciar a busca aleatória. Já a em relação à quantidade de ocorrências de PChs encontradas, só procura por três tipos de codificação e não realiza combinações maiúsculo/minúsculo. Além disso, não identifica ocorrências de PChs entre blocos.

A ferramenta CS, realiza busca sequencial, não é necessário fazer indexação do conteúdo dos dados. Com relação à quantidade de ocorrências de PChs, procura por cinco padrões de codificação e combinações de maiúsculo/minúsculo.

Durante a concepção da ferramenta CS tinha-se em mente que a capacidade máxima de "carga" para a ferramenta seria de até 10 palavras com até 10 caracteres cada, ou seja, 10 palavras de 80 bytes, totalizando 800 bytes. Salientando que para cada PCh da lista a ser procurada na varredura, a ferramenta CS realiza todas as combinações possíveis entre os caracteres (maiúsculo/minúsculo) nas codificações ASCII, UTF-8, UTF-16, UTF-32 e ISO 8859-1.

Após os resultados dos experimentos, notou-se que somente a partir da varredura para procura de 40 PChs de 10 caracteres cada, o tempo para duplicação apresentou um acréscimo de 1,2 segundos.

Comparando-se o método tradicional e o método proposto, partindo do pressuposto que um DAD tivesse que ser duplicado para posteriormente ser pré-analisado, para obter os primeiros resultados, utilizando o método proposto, os primeiros resultados já são obtidos ao término da duplicação.

6.2 Resultados da Filtragem

Durante os primeiros testes de avaliação da ferramenta CS, foi identificada uma elevada quantidade de falso-positivo de ocorrências de

PChs, o que gerou a necessidade de filtrar as ocorrências e classificá-las como verdadeiro-positivo e falso-positivo. Para realizar esse processo, foi implementada a ferramenta FiltroP. O processo de filtragem possibilitou uma melhora dos resultados da ferramenta CS.

Além disso, a filtragem possibilitou identificar que grande parte das ocorrências de PChs com menos que quatro caracteres eram falsos-positivo.

Ao término da etapa de filtragem, são obtidos os valores de VP e FP. Comparando os resultados obtidos com a utilização dos valores de VP ($QOVP_{DT}$), foi notado uma melhora nos resultados do ranqueamento quando comparado com os resultados utilizando a quantidade de ocorrências total (QO_{DT}).

Dentre as “partes” do disco, a que apresenta maior quantidade de arquivos que certamente pode ser lido e interpretado pela ferramenta CS é a ($QOVP_{AI}$), porém é muito difícil de ser obtido. Então, foram comparadas as quantidade de ocorrências de PChs entre (QO_{AI}), (QO_{DT}), ($QOVP_{AI}$) e ($QOVP_{DT}$). O valor que apresentou melhor convergência com ($QOVP_{AI}$), foi ($QOVP_{DT}$). Isso reforça a importância de utilizar os valores da quantidade de ocorrência de PChs filtradas ($QOVP_{DT}$).

Durante a validação dos processos de varredura e filtragem, no qual foi utilizado o procedimento de “dividir” o DAD, as quantidades de ocorrências de PChs verdadeiro-positivo encontradas nos arquivos mistos ($QOVP_{AM}$), são procedentes de metadados de arquivos, pois é o único tipo de dado nesta “parte” que a ferramenta CS reconhece. Considerando que nesta parte do DAD, todos os arquivos apresentam algum tipo de codificação, que impede a interpretação de seu conteúdo por uma ferramenta que lê dados brutos.

6.3 Resultados da Triagem

Os resultados mostraram que a ausência de prioridade de PChs para realização da triagem, coloca PChs que apresentam grande quantidade de ocorrência em evidência. Em outras palavras, isso pode inserir erros para realização do ranqueamento.

Na eventualidade de não haver prioridade, pode ser utilizada a tabela agregada, resultante do Processo de Filtragem, determinando as palavras que possuem maior variabilidade como prioritárias e examinar o ranqueamento.

A escolha de PChs influencia diretamente o ranqueamento, portanto, é importante que a preferência seja dada a PChs que apresentam maior variabilidade. Em razão da sequência de classificação ser prioritariamente ranqueada pela quantidade de ocorrência das PChs selecionadas.

Na prática, durante a realização dos experimentos, não houve situação para a qual todas as PChs procuradas apresentaram zero ocorrências no mesmo DAD. Porém, caso houvesse algum DAD, deveria ser separado dos demais DADs, e outras técnicas (para verificação do conteúdo) devem ser utilizadas, como por exemplo, verificação se está com algum tipo de compactação ou se realmente não há nenhuma ocorrência de PCh no DAD.

6.4 Resultados do Ranqueamento

Para comparar quanto a posição da classificação de um DAD está deslocada em relação à posição em outra classificação (utilizada como referência) para o mesmo DAD, utilizou-se a defasagem.

Com a finalidade de avaliar a convergência dos ranqueamentos entre (QOVP_{AI}) e (QOVP_{DT}) foi realizada a comparação da defasagem do ranqueamento entre eles e a ferramenta *Autopsy*. Coerentemente o ranqueamento que apresentou a melhor convergência nos critérios com prioridade e sem prioridade foi (QOVP_{DT}).

O ranqueamento é diretamente influenciado pelos resultados da triagem. A partir do ranqueamento obtido, é possível que o perito interativamente e estabeleça um novo critério na triagem, obtendo um novo ranqueamento.

6.5 Considerações Finais da discussão dos Resultados

Os resultados dos experimentos permitiram avaliar as funcionalidades mais importantes para uma ferramenta que deve executar a varredura de dados em dados brutos e duplicação simultaneamente. Também indicaram ajustes necessários para melhorar o desempenho das ferramentas utilizadas para prova de conceito do método proposto nesta tese.

Para investigações nas quais é suficiente encontrar os vestígios nos primeiros DADs, este método pode auxiliar. Como por exemplo para responder se em algum dos DADs examinados contém determinados vestígios.

Os experimentos também serviram para indicar algumas limitações, como por exemplo, a procura de PChs em dados não estruturados não poder ler o conteúdo de arquivos compactados de uma forma geral. Além disso, a maneira como a busca é realizada sequencialmente não permite ler o conteúdo de um bloco do DAD e retornar para outro, ou ler blocos distintos simultaneamente. Embora estas limitações sejam intrínsecas do nível de dados em que é realizada, o método não é inviabilizado.

6.5 Conclusões

Retomando o problema de pesquisa desta tese e o objetivo, apresentados no Capítulo 1:

O problema de pesquisa desta tese consiste em, a partir de um conjunto de computadores com seus respectivos dispositivos de armazenamento de dados (DADs), objetos de uma investigação digital forense – IDF, descobrir:

- I. Como identificar os DADs que apresentam maiores chances de conter os vestígios de interesse para a investigação apenas com os dados obtidos da varredura durante a duplicação?
- II. Como sugerir uma sequência que os DADs devem ser examinados na etapa de análise, baseada em prioridades de palavras-chaves de interesse para a investigação?

Foi colocada a hipótese:

- ...de que na etapa de coleta, seja possível realizar simultaneamente à duplicação do DAD de origem para o DAD de destino a varredura dos dados para procurar por palavras-chave de interesse do perito, sem com isso impactar significativamente o tempo de duplicação.
- ...O principal objetivo desta tese é demonstrar a possibilidade de realizar simultaneamente à duplicação do DAD de origem para o DAD de destino, a varredura dos dados para procurar por palavras-chave de interesse do perito, sem com isso impactar significativamente o tempo de duplicação. Então, baseando-se na quantidade de ocorrências de palavras-chave encontradas, propor um método para triagem e ranqueamento dos DADs duplicados, de forma a apresentar uma sugestão de sequência, indicando quais desses DADs podem ser prioritariamente examinados na etapa de análise. E assim, responder:
 - a) A proposta é viável? Caso afirmativo, para quais cenários?
 - b) Quais tipos de dados podem ser reconhecidos?
 - c) Como definir um ranqueamento baseado nos dados coletados durante a varredura?

d) Quais os indicadores de desempenho que devem ser analisados para avaliar este tipo de varredura?

e) A aplicação do método contribui para a diminuição do esforço humano na etapa de análise?

Baseado nos experimentos e testes realizados, sim, o método proposto é viável. O método é voltado para cenários nos quais o perito deve duplicar todos os DADs encontrados no local da perícia e não possui nenhum indicativo de qual dos DADs podem conter vestígios, que neste caso são PChs de interesse para uma determinada investigação.

Considerando que a varredura é realizada sobre dados brutos, neste nível de dados não há um sistema de arquivos, o reconhecimento dos *bytes* pode ocorrer somente na forma sequencial. Podem ser lidos todos os tipos de dados, porém somente os dados em texto plano podem ser reconhecidos.

Conforme foi visto nos experimentos apresentados no Capítulo 5, as palavras-chaves coletadas, passam pelos processos de filtragem e triagem. A partir das informações da triagem e conforme critérios estabelecidos pelo perito, tais como a prioridade de PChs, pode então ser definida uma sequência de análise para os DADs.

Considerando que a varredura para procurar por palavras-chave ocorre em dados brutos e concorrente com a duplicação, os indicadores de desempenho que devem ser analisados são o tempo necessário para varredura e duplicação e a quantidade de palavras-chave encontradas.

Baseando-se nos resultados da triagem, ranqueamento e sugestão da sequência para início da análise, é possível encontrar os vestígios de interesse nos primeiros DADs indicados no ranqueamento. Dependendo dos objetivos da investigação, pode propiciar ao perito a desnecessidade de analisar os demais dispositivos.

6.5.1 Contribuições

A principal contribuição está no método proposto, a ser aplicado na etapa de Coleta de dados, que realiza a varredura para procura por palavras-

chave - PChs de interesse para a investigação simultaneamente a duplicação do DAD.

A partir dos resultados obtidos ao término da duplicação dos DADs é realizada a triagem e ranqueamento dos DADs, indicando quais destes devem ser prioritariamente examinados na etapa de Análise. Dependendo do contexto da perícia, localizar as PChs de interesse e os DADs que apresentam maiores chances de conter os vestígios de interesse nos primeiros DADs pode tornar desnecessária a continuação da análise de todos os DADs encontrados no local da perícia.

Como contribuições indiretas pode-se considerar as ferramentas *CopyScan* (CS), *FiltroP* (FiP) e *EraseFiles* que foram desenvolvidas para validação e prova de conceito do método proposto.

6.5.2 Trabalhos Futuros e Perspectivas de Continuidade

Uma vez que este trabalho amplia o horizonte com respeito à exploração de dados brutos simultaneamente a duplicação de DADs na etapa de coleta, pode-se empregar os conhecimentos e experiências aqui adquiridas e considerar-se a possibilidade de que um sistema de arquivos, o conteúdo dos arquivos que apresentam algum tipo de compactação e o respectivo aplicativo necessário para interpretá-lo, como por exemplo, .pdf e arquivos de textos que apresentam compactação, possa ser carregado em memória RAM e dessa forma varridos para a procura por PChs de interesse. Outra possibilidade é a criação de uma ontologia que possa ser utilizada nessa área.

REFERÊNCIAS

ABDULLAH, M. T.; et. al. Advances in Computer Forensics. IJCSNS International Journal of Computer Science and Network Security, Vol.8, nº2, Feb. 2008. Disponível em: <http://paper.ijcsns.org/07_book/200802/20080228.pdf>. Acesso em: 14 ago. 2012.

ACPO CRIME COMMITTEE. Good Practices Guide for Computer Based Evidence. Association of Chief Police Officers of England, Wales and Northern Ireland, Jun. 1999. Disponível em: <<http://www.acpo.police.uk/documents/crime/2011/201103CRIECI14.pdf>>. Acesso em: 14 ago. 2010.

ADEMU, I. O.; IMAFIDON, C. O.; PRESTON, D. S.; A New Approach of Digital Forensic Model for Digital Forensic Investigation. International Journal of Advanced Computer Science and Applications – IJACSA. vol. 2, n.12, 2011.

AGARWAL, A.; GUPTA, S.; M.; CHAND, S.; GUPTA, M.; Systematic Digital Forensic Investigation Model. In: International Journal of Computer Science and Security (IJCSS) 2011. Volume 5, Issue 1. p.1-167. CSC Journals, Kuala Lumpur, Malaysia.

ALI, K. M., "Digital Forensics Best Practices and Managerial Implications," Computational Intelligence, Communication Systems and Networks (CICSyN), 2012. Fourth International Conference on, vol., no., pp.196-199, 24-26 July 2012. doi: 10.1109/CICSyN.2012.44

ANDALÓ, F. A., Métodos de visão computacional aplicáveis à ciência forense. 2012. 89 p. Tese. Instituto de Computação da Universidade Estadual de Campinas - IC-UNICAMP. Campinas, SP. 09 nov. 2012.

BAGGILI, I.; Marrington, A., Jafar, Y. Performance of a Logical, Five- Phase, Multithreaded, Bootable Triage Tool. *Advances in Digital Forensics X*. Volume 433 of the series IFIP Advances in Information and Communication Technology pp 279-295

BEEBE, N. L.; CLARK, J.; Dealing with Terabyte Data Sets in Digital Investigations. *Advances in Digital Forensics IFIP — The International Federation for Information Processing*. Vol. 194, 2005, pp 3-16.

BEEBE, N.; DIETRICH, G.; A New Process Model for Text String Searching. *Advances in Digital Forensics III*. IFIP — The International Federation for Information Processing Vol. 242, 2007a, pp. 179-191. Disponível em: <http://link.springer.com/chapter/10.1007%2F978-0-387-73742-3_12>. Acesso em: 12 mar. 2013.

BEEBE, N. L.; CLARK, J. G.; Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. *Digital Investigation* 4S, 2007b S49–S54. Disponível em: <<http://www.dfrws.org/2007/proceedings/p49-beebe.pdf>>. Acesso em: 12 mar. 2013.

BEEBE, N. L.; CLARK, J. G.; DIETRICH, G. B., KO, M. S.; KO , D.; Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies. *Decision Support Systems*, vol. 51, Issue 4, nov. 2011, pag. 732-744, ISSN 0167-9236. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167923611000388>>. Acesso em: 05 jan. 2012.

BEEBE, N. L.; LISHU Liu, Ranking algorithms for digital forensic string search hits. *Digital Investigation*, vol. 11, Supplement 2, Aug. 2014, pag. S124-S132, ISSN 1742-2876.

BERGHEL, H. Digital village: The discipline of Internet forensics. *Communications of the ACM*, 46 (8). 15-20. 2003.

BERTÈ, R.; MARTURANA, F.; ME GIANLUIGI; TACCONI, S. **Data mining based crime-dependent triage in digital forensics analysis**. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ICACII)* and *IERI Lecture Notes in Information Technology*, 2012.

BHAT H. V. et al.; A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application. *IACSIT - International Journal of Engineering and Technology*. Vol. 2, No. 3. June 2010. ISSN: 1793-8236.

BODEAU, D. J.; GRAUBART, R.; FABIUS-GREENE, J.; Improving Cyber Security and Mission Assurance Via Cyber Preparedness (Cyber Prep) Levels. *Social Computing (SocialCom)*, 2010 IEEE Second International Conference on IEEE International Conference on Social Computing/IEEE International Conference on Privacy, Security, Risk and Trust. Digital Object Identifier: 10.1109/SocialCom.2010.170 Publication Year: 2010 , Page(s): 1147 – 1152.

BRADY, O.; OVERILL, R.; KEPPENS, J., "Addressing the Increasing Volume and Variety of Digital Evidence Using an Ontology," *Intelligence and Security Informatics Conference (JISIC)*, 2014 IEEE Joint , vol., no., pp.176,183, 24-26 Sept. 2014. doi: 10.1109/JISIC.2014.34

BRANTINGHAM, P.L. Computacional Criminology. *Intelligence and Security Informatics Conference (EISIC)*, 2011 European Digital Object Identifier: 10.1109/EISIC.2011.79. Publication Year: 2011, Page(s): 3. IEEE CONFERENCE PUBLICATIONS.

BRASIL. Código de Processo Penal. Decreto-Lei nº 3.699, de 03 de outubro de 1941. Diário Oficial da União, Poder Executivo, Brasília, DF, 24 out. 1941. P.1. Disponível em: <http://www.planalto.gov.br/ccivil_03/decreto-lei/del3689.htm>. Acesso em: 19 set. 2012.

BRASIL. Código de Processo Civil. Lei nº 5.869, de 11 de janeiro de 1973. Diário Oficial da União, Poder Executivo, Brasília, DF, 17 jan. 1973. P.1. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L5869compilada.htm>. Acesso em: 28 jan. 2012.

BREZINSKI, D.; KILLALEA, T.; Request for Comments: 3227 - Guidelines for Evidence Collection and Archiving. 2002. Disponível em: <<http://tools.ietf.org/pdf/rfc3227.pdf>>. Acesso em: 11 abr. 2010.

CANTRELL, G. et al. Research toward a Partially-Automated, and Crime Specific Digital Process Model. Computer and Information Science. Vol. 5, nº 2; Mar. 2012. Disponível em: <<http://www.ccsenet.org/journal/index.php/cis/article/view/13527/0>>. Acesso em: 14 abr. 2012.

CARRIER, B. *File System Forensic Analysis*. Upper Saddle River – NJ: 2005.

CARRIER BRIAN D. A hypothesis-based approach to digital forensic investigations. PhD thesis, Purdue University; 2006. Disponível em: <https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2006-06.pdf>. Acesso em: 11 nov. 2012.

CASEY, E. Triage in Digital Forensics. Digital Investigation. Volume 10, Issue 2. p. 85-204, sep. 2013. Disponível em: <<http://dx.doi.org/10.1016/j.diin.2013.08.001>>. Acesso em: 30 out. 2013.

CASEY, E. Digital Evidence and Computer Crime. Academic Press, San Diego, California, 2004. ISBN:0121631044.

CASEY, E. Digital Evidence and Computer Crime: Forensic Science, Computer and the Internet. Boston: Academic Press. 2000.

CASTRO, C. R. A. Crimes de informática e seus aspectos processuais. Rio de Janeiro : Lumen Juris. 2003. 2. ed., rev., ampl. e atual. 230p.

CASTRO, G.; CHAMON, V.; Dicionário de informática. Microsoft PRES - trad. 3. ed. americana. Rio de Janeiro: Campus, 1998. ISBN 85-352-0255-2.

CISAR, P.; CISAR, S. M. *Methodological frameworks of digital forensics*. Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium on.08-10 Sept. 2011. Telekom Srbija, Subotica, Page(s): 343 - 347.

CLAYTON, J. Investigation Into A Digital Forensics Triage Tool Using Sampling, Hashes And Bloom Filters – SHAFT. 2012. 101p. Dissertação (Mestrado) – Edinburgh Napier University, Edinburgh, UK, 2012. Disponível em: <<http://www.iidi.napier.ac.uk/c/publications/publicationid/13373237>>. Acesso em: 02set. 2012.

COSTA, M. A. R. da; Crimes de Informática. Jus Navigandi, Teresina, ano 2, n. 12, 05 maio1997. Disponível em:<<http://jus.com.br/revista/texto/1826>>. Acesso em: 18 ago. 2012.

DAN HU, Exploratory Study on Computer Forensics Technology. *Computer and Electrical Engineering*, 2009. ICCEE '09. *Second International Conference on*. vol. 2, p. 608-611. 28-30 Dec. 2009. doi: 10.1109/ICCEE.2009.258

EC-COUNCIL. Computer forensics : investigation procedures and response/EC-Council Press. 1 ed. Volume 1 of 5 mapping to C-HFI (Computer Hacking Forensic Investigator) certification. Clifton Park, NY : Course Technology Cengage Learning, 2010. ISBN-13: 978-1435483491 ISBN-10: 1435483499.

ESENER, S. C.; et al. *WTEC Panel Report on: The Future of Data Storage Technologies*. International Technology Research Institute. Jun1999. Disponível em: <<http://www.wtec.org/pdf/hdmem.pdf>>. Acesso em: 19 Mar. 2012.

ESPÍNDULA, A. Curso Preservação de local de crime – Módulo 1. SENASP/MJ - Última atualização em 09/07/2009. Disponível em: <<http://www.espindula.com.br/>>. Acesso em: 10 nov. 2012.

FARIA, E., *Dicionário Escolar Latino-Português*. 3ª. ed. Ministério da Educação e Cultura - Departamento Nacional de Educação - Campanha Nacional de Material de Ensino. Rio de Janeiro. 1962.

FARMER, D.; VENEMA, W. *Forensic discovery*. Upper Saddle River, NJ: Pearson Education, Inc., 2005. 217p.

FERNANDEZ, J. D.; SMITH, S.; GARCIA, M.; KAR, D. Computer forensics: a critical need in computer science programs. *Journal of Computing Sciences in Colleges*, ACM, 20 (4). 315-322. 2005.

FRANKE, K.; SRIHARI S. Computational forensics: An overview. *Computational Forensics*, pages 1-10. 2008.

GARBER, L. Computer Forensics: High-Tech Law Enforcement. *IEEE Computer Society's Computer Magazine*, 34 (1). 22-27. 2001.

GARFINKEL, S. L. Forensic feature extraction and cross-drive analysis, *Digital Investigation*, Volume 3, Supplement, September 2006, Pages 71-81, ISSN 1742-2876, <http://dx.doi.org/10.1016/j.diin.2006.06.007>.

GARFINKEL et al., Bringing science to digital forensics with standardized forensic corpora, *Digital Investigation*, Volume 6, Supplement, Sep. 2009, Pages S2-S11, ISSN 1742-2876.

GARFINKEL, S. L. Digital Forensics research: The next 10 years. *Journal Digital Investigation: The International Journal of Digital Forensics & Incident Response*. v 7.S64-S73.Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands. Ago. 2010. Disponível em: <<http://dl.acm.org/citation.cfm?id=2296192>>. Acesso em: 10 mar. 2012.

GARFINKEL, S. L. Lessons learned writing digital forensics tools and managing a 30 TB digital evidence corpus. *The Proceedings of the Twelfth Annual DFRWS Conference. 12th Annual Digital Forensics Research Conference. Digital Investigation, Volume 9, Supplement, August 2012, Pages S80-S89*. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287612000278#>>. Acesso em: 12 set. 2012.

GARFINKEL, S. L. Digital media triage with bulk data analysis and bulk_extractor. *Computers and Security*. v. 32, p. 56-72, Feb. 2013.

GAYED, T. F., LOUNIS, H., BARI, M.; Creating proprietary terms using lightweight ontology: A case study on acquisition phase in a Cyber Forensic process. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE* , Jan. 2014. pp. 76 - 81.

GLADYSHEV, P. Formalising Event Reconstruction in Digital Investigation, Phd. Dissertation, Department of Computer Science, University College Dublin. 2004.

GOGOLIN, G., The Digital Crime Tsunami, Digital Investigation, Volume 7, Issues 1-2, October 2010, Pages 3-8, ISSN 1742-2876. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287610000526>>. Acesso em: 21 jan. 2011

GRILLO, A.; LENTINI, A.; ME, G.; OTTONI, M. Fast User Classifying to Establish Forensic Analysis Priorities. IT Security Incident Management and IT Forensics, 2009. IMF '09. Fifth International Conference on Digital Object Identifier: 10.1109/IMF.2009.16. Publication Year: 2009, Page(s): 69 - 77.

GUIMARÃES, R.; SANTOS, A. L. dos. Mais computadores entre os brasileiros. GV-executivo, v. 11, n. 2, julho-dezembro, 2012. ISSN 1806-8979. Disponível em <<http://rae.fgv.br/gv-executivo/vol11-num2-2012/mais-computadores-entre-brasileiros>>. Acesso em: 12 nov. de 2012.

HALBOOB, W.; et al. An Efficient Computer Forensics Selective Imaging Model. Future Information Technology. Lecture Notes in Electrical Engineering. Volume 276, 2014, pp 277-284. DOI: 10.1007/978-3-642-40861-8_41. Springer-Verlag Berlin Heidelberg. 2014.

HALL, G. A. and Davis, W. P. Toward Defining the Intersection of Forensics and Information Technology. International Journal of Digital Evidence, 4 (1). p. 1-20. 2005.

HARRINSON, W.; et al. A Lessons Learned Repository for Computer Forensics. International Journal of Digital Evidence. Fall 2002. Vol. 1, Issue 3. Disponível em <http://www.dfrws.org/2002/papers/Papers/Warren_Harrison.pdf>. Acesso em: 12 nov. de 2012.

HOELZ, B. W. P. (2009). MADIK: Uma abordagem Multiagente para o Exame Pericial de Sistemas Computacionais. Dissertação (Mestrado em Informática) - Universidade de Brasília - UNB, Brasília, DF, 2009.

HONG I. et al., A new triage model conforming to the needs of selective search and seizure of electronic evidence, Digital Investigation (2013). Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287613000042>>. Acesso em: 21 jan. 2013.

HOPFIELD, J.; TANK, D. Neural Computation of Decision in Optimization Problems. Biological Cybernetics, p.52: 141-152, 1985.

HU, D. Exploratory Study on Computer Forensics Technology. Second International Conference on Computer and Electrical Engineering. 2009. DOI 10.1109/ICCEE.2009.258.

IEEE 802.3 ETHERNET WORKING GROUP. **IEEE 802.3™ Industry Connections Ethernet Bandwidth Assessment.** Disponível em: <http://www.ieee802.org/3/ad_hoc/bwa/BWA_Report.pdf>. Acesso em 08 ago. 2012.

JONES, B.; PLENO, S.; WILKINSON, M. The use of random sampling in investigations involving child abuse material. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287612000369>>. Acesso em: 07 set. 2012.

KALBANDE, D.; JAIN, N.; Comparative Digital Forensics Model. International Journal of Innovative Research in Science. Engineering and Technology. Vol. 2, Issue 8, August 2013. ISSN: 2319-8753. Disponível em: <http://www.ijirset.com/upload/august/6_COMPARATIVE.pdf>. Acesso em: 10 set. 2013.

KÄLBER, S. et al. Forensic Zero-Knowledge Event Reconstruction on Filesystem Metadata. Lecture Notes in Informatics. Sicherheit 2014. Sicherheit, Schutz und Zuverlässigkeit. P-228, 331-343 (2014). Gesellschaft für Informatik, Bonn. 2014. ISBN 978-3-88579-622-0. Disponível em: <<http://subs.emis.de/LNI/Proceedings/Proceedings228/article3.html>>. Acesso em: 29 dez. 2014.

KARIE, N. M.; VENTER, H. S.; Taxonomy of Challenges for Digital Forensics. Journal of Forensics Sciences. Jul. 2015, Vol. 60, n. 4. doi: 10.2222/1556-4029.12809. Disponível em: <<http://onlinelibrary.wiley.com/doi/10.1111/1556-4029.12809/pdf>>. Acesso em: 15 ago. 2015.

KNUTH, D.; MORRIS JR., J.; PRATT, V. Fast pattern matching in strings. SIAM Journal on Computing, v. 6, n. 2, p. 323-350, 1977. Disponível em: <<http://dx.doi.org/10.1137/0206024>>. Acesso em: 02 mai. 2012.

KOHONEN, T. Associative Memory: A System Theoretic Approach. [S.l.]: Springer Verlag, Berlin, 1978.

KORNBLUM, J. Preservation of Fragile Digital Evidence by First Responders. In: Proceedings of Digital Forensic Research WorkShop 2002. Syracuse, NY. International Journal of Digital Evidence. p. 1-11. 2002.

KORNBLUM, J.D. The Linux Kernel and the Forensics Acquisition of Hard Discs with an Odd Number of Sectors. International Journal of Digital Evidence, 3. 2004.

KRUSE, W.; HEISER, J.; Computer Forensics: Incident Response Essentials. New York: Addison Wesley. 2002.

KRYDER, M. H.; KIM, C. S. After Hard Drives—What Comes Next?. IEEE Transactions on Magnetics, VOL. 45, NO. 10, OCTOBER 2009.

LANGE, R. Crimes cibernéticos: Novos desafios do Direito. Monografia. (Graduação em Direito) - Faculdades Integradas do Brasil - UniBrasil, Curitiba, PR, 2010.

LEE, H. C. Cyber Crime and Challenges for Crime Investigation in the Information Era. IEEE ISI 2008 Keynote Talk (II). Disponível em: <http://isi2008.cpu.edu.tw/isi2008/docs/ISI2008_06_02_Keynote_Talk_Dr_Henry_C_Lee.pdf>. Acesso em: 12 jun. 2012.

LU QIN, Data mining method based on computer forensics-based ID3 algorithm. *Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on*, vol., no., p. 340-343, 16-18 April 2010. doi: 10.1109/ICIME.2010.5477817

LUCCA, N. DE; SIMÃO FILHO, A. (coordenadores) e outros. **Direito e Internet – aspectos jurídicos relevantes**. São Paulo. Ed. Quartier Latin, 2ª ed. 2005.

LYNCH, P.; LUAN, X.; PRETTYMAN, M.; MERICLE, L.; BORKMANN, E.; SCHLAIFER, J., "An evaluation of new and old similarity ranking algorithms." *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, vol.2, no., pp.148,149 Vol.2, 5-7 April 2004. doi: 10.1109/ITCC.2004.1286615.

MARTURANA, F.; BERTE, R.; TACCONI, S.; ME, G. Triage-based automated analysis of evidence in court cases of copyright infringement. *Communications (ICC), 2012 IEEE International Conference on Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software); Signal Processing & Analysis. 2012*, page(s): 6668 – 6672.

MC KEMMISH, R. What is forensic computing? Trends and Issues in Crime and Criminal Justices (118). 1-6. jun. 1999.

MEIKLE, N.; Case Study: Big Data Forensics. Cyber Security Malasia Awards, Conference e Exhibition CSM-ACE 2012.06 Nov. 2012. Disponível em: <http://www.csm-ace.my/presentation/Day1_Track2_2012/Neil%20Meikle_%20Case%20Study%20Big%20Data%20Forensics.pdf>. Acesso em: 12 mai. 2013.

MOHAY, G. Technical challenges and directions for digital forensics. Topic(s): Communication, Networking & Broadcasting; Components, Circuits, Devices & Systems; Computing & Processing (Hardware/Software). Digital Object Identifier: 10.1109/SADFE.2005.24. Publication Year: 2005, Page(s): 155 - 161.

MORA, R.; KLOET, B. Digital forensic sampling: The application of statistical sampling in digital forensics. Disponível em: <<https://blogs.sans.org/computer-forensics/files/2010/03/statisticalforensictriage.pdf>>. Acesso em: 10 abr. 2013.

NANCE, K.; HAY, B.; BISHOP, M.; Digital Forensics: Defining a Research Agenda. System Sciences, HICSS '09. 42nd Hawaii International Conference on System Sciences (2009), PP. 1-6. DOI: 10.1109/HICSS.2009.160.

NASSIF, L. F. DA C., HRUSCHKA, E. R., Computação Forense via Agrupamento Hierárquico de Documentos. IcoFCS 2011. The Sixth International Conference on Computer Forensics Science. Florianópolis, Brasil. Oct. 5th to 7th. 2011a. Disponível em: <<http://www.icofcs.org/2011/IcoFCS2011-PP19.pdf>>. Acesso em: 12 dez. 2011.

NASSIF, L. F. DA C., HRUSCHKA, E. R., Document Clustering for Forensic Computing: An Approach for Improving Computer Inspection. Machine Learning and Applications and Workshops (ICMLA), 2011.10th International Conference on. Volume: 1. Digital Object Identifier: 10.1109/ICMLA.2011.59. 2011b. Page(s): 265 - 268.

NELSON, B.; PHILLIPS, A., STEUART, C. Guide to Computer Forensics and Investigations. Fourth Edition. Ed. Cengage Learning. 2010. ISBN-13: 978-1-435-49883-9. ISBN-10: 1-435-49883-6. Boston, MA.

NERI, M., Mapa da Inclusão Digital. Rio de Janeiro: FGV, CPS. 2012. 190p.

NETO, F. M. Prova pericial. Belo Horizonte: Del Rey, 1998. 88p. ISBN 85-7308-208-9.

NOBLETT, M. G.; POLLIT, M. M.; PRESLEY, L. A. Recovering and Examining Computer Forensic Evidence. Forensic Science Communications, 2. Out. 2000. US. Department of Justice, FBI.

NOBLETT, M. G., Pollitt, M. and Presley, L. A. Recovering and Examining Computer Forensic Evidence Forensic Science Communications. Disponível em: <<http://www.fbi.gov/hq/lab/fsc/backissu/oct2000/computer.htm>>. Acesso em: 15 jun. 2013.

NOGUEIRA, J. H. M.; JÚNIOR, J. C. Autonomic Forensics a New Frontier to Computer Crime Investigation Management. The International Journal of Forensic Computer Science, 2009, p. 29-41.

O'CONNOR, O. Deploying forensic tools via PXE. Digital Investigation: The International Journal of Digital Forensics & Incident Response archive. vol. 1, Issue 3, Sep. 2004. pp. 173-176.

PALMER, G. A Road Map for Digital Forensic Research, Report for the First Digital Forensic Research Workshop (DFRWS), DTR-T001-01FINAL, November 2001. Disponível em: <<http://www.dfrws.org/dfrws-rm-final.pdf>>. Acesso em: 14 jun. 2010.

PARSONAGE, H. Computer forensics case assessment and triage - some ideas for discussion. Disponível em: <<http://computerforensics.parsonage.co.uk/triage/ComputerForensicsCaseAssessmentAndTriageDiscussionPaper.pdf>>. Acesso em: 14 set. 2012.

PEARSON, S.; WATSON, R.; Digital Triage Forensics: Processing the Digital Crime Scene. Syngres, 1st ed. 280 p. 2010. ISBN-10: 1597495964 ISBN-13: 978-1597495967.

PILLI, E. S.; JOSHI, R. C.; NIYOGI, R. **Network forensics frameworks: Survey and research challenges**. Digital Investigation: The International Journal of Digital Forensics & Incident Response. Vol. 7. Issues 1-2, Out. 2010. p. 14-27. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287610000113>>. Acesso em: 10 nov. 2010.

PIRES, P. S. M. Forense Computacional: Uma Proposta de Ensino. Anais do 6º SSI – Simpósio de Segurança em Informática – São José dos Campos - SP, 2004.

POLLITT, M. M. Computer Forensics: An Approach to Evidence in Cyberspace. In: Proceeding of the National Information Systems Security Conference. 1995. Baltimore, MD, Vol. II, pp.487-491.

POLLITT, M. M. Triage: A practical solution or admission of failure. Digital Investigation. Volume 10, Issue 2, September 2013, Pages 87-88. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287613000030>>. Acesso em: 10 out. 2013.

PRESSMAN, R. S.; Engenharia de Software. Tradução Rosângela Delloso Penteado, revisão técnica Fernão Stella R. Germano, José Carlos Maldonato, Paulo Cesar Masiero. 6. Ed. São Paulo: McGraw-Hill, 2006.

PROVOST, F.; FAWCETT, T.; Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. *In: Proc. Third Internet. Conf. On Knowledge Discovery and Data Mining (KDD-97)*. AAAI Press, Menlo Park, CA, pp. 43-48, 1997.

PROVOST, F.; FAWCETT, T.; KOHAVI, R.; The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning*, 445-453, 1998.

PUNGILA, C. Improved File-Carving Through Data-Parallel Pattern Matching for Data Forensics. In: Applied Computational Intelligence and Informatics (SACI) 2012. 7th IEEE International Symposium on. pp.197-202. 24-26 Mai. 2012. doi: 10.1109/SACI.2012.6250001.

QUICK, D.; CHOO, K. R.; Impacts of increasing volume of digital forensic data: A survey and future research challenges, *Digital Investigation*, Volume 11, Issue 4, Dec. 2014, Pages 273-294, ISSN 1742-2876. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1742287614001066>>. Acesso em: 11/Jan./2015.

RAY, I.; Presentation on “Digital forensics: Cyber sleuthing solves the case,” 1st International Conference on Information Systems Security (ICISS 2005). Disponível em: < <http://www.cdc-ju.org.in/iciss/tutorials.htm>>. Acesso em: 01 jan. 2011.

REIS, M. A.; GEUS, P. L. Análise Forense de Intrusões em Sistemas Computacionais: Técnicas, Procedimentos e Ferramentas. Em Anais do I Seminário Nacional de Perícia em Crimes de Informática e IV Seminário Nacional de Fonética Forense, Maceió, AL. Nov. 2002. Associação Brasileira de Criminalística. Disponível em: <<http://www.las.ic.unicamp.br/paulo/papers/2002-Pericia-marcelo.reis-forense.tecnicas.procedimentos.pdf>>. Acesso em: 01 ago. 2010.

RICHARD III, G.G., ROUSSEV, V. Digital forensics tools: The next generation. Digital Crime and Forensic Science in Cyberspace. 2006. pages 75-90. ISBN: 978-159140872-7.

ROBBINS, J. Global Information Assurance Certification Paper. SANS Institute 2000 - 2002. Disponível em: <<http://www.giac.org/paper/gsec/584/computer-forensics/101347>>. Acesso em: 13/Jun./2013.

ROGERS, K.; SEIGFRIED, K.; The future of computer forensics: a needs analysis survey. Computers & Security.Vol. 23, Issue 1, Feb. 2004, pag. 12-16. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167404804000100>>. Acesso em: 19Dez. 2011.

ROGERS, M. K., et al. Computer Forensics Field Triage Process Model. JDFSL - The Journal of Digital Forensics, Security and Law. vol. 1., n. 2, p. 19-38. Conference on Digital Forensics, Security and Law, 2006.

ROUSSEV, V.; QUATES, C.; Content triage with similarity digests: the M57 case study. Digital Investigation, Digital Investigation, Vol. 9, Supplement, Aug. 2012, Pages S60–S68. doi:10.1016/j.diin.2012.05.012.

ROUSSEV, V.; In RAM We Trust: A Modern Approach to Forensics Processing. 3rd. Annual The Sleuth Kit and Open Source Digital Forensic Conference. Oct. 2012. Disponível em: <<http://www.basistech.com/wp-content/uploads/2014/04/OSDF-2012-In-RAM-We-Trust-A-Modern-Approach-to-Forensic-Processing-Vassil-Roussev.pdf>>. Acesso em: 20 dez. 2014.

ROUSSEV, V., QUATES, C., MARTELL, R.; Real-time digital forensics and triage. Digital Investigation, v. 10, Issue 2, Sep. 2013, Pages 158–167.

RUI XU, WUNSCH, D., Survey Of Clustering Algorithms. Neural Networks, IEEE Transactions on, Vol.: 16, Issue: 3. Digital Object Identifier:10.1109/TNN.2005.845141. Publication Year: 2005, Page(s): 645 - 678.

RUIBIN, G.; YUN, C. K. Abstract Case-Relevance Information Investigation: Binding Computer Intelligence to the Current Computer Forensic Framework. International Journal of Digital Evidence. Spring 2005, Vol. 4, Issue 1. Disponível em: <<http://www.utica.edu/academic/institutes/ecii/publications/articles/B4A6A102-A93D-85B1-96C575D5E35F3764.pdf>>. Acesso em: 11 out. 2010.

SAMMES, T; JENKINSON, B. Forensic Computing: A Practitioner's Guide. Springer, London, 2000.

SAMMONS, J. The basics of digital forensics : the primer for getting started in digital forensics. ed. Syngress. 1 ed. 2012.

SBC, Sociedade Brasileira de Computação. II Seminário sobre Grandes Desafios da Computação no Brasil – Integração com a Indústria e uma Perspectiva para 2020. Manaus, 2009. Disponível em: <http://www.sbc.org.br/index.php?option=com_jdownloads&Itemid=195&task=view.download&catid=50&cid=237>. Acesso em: 10 nov. 2011.

SHAW, A.; BROWNE, A.; A practical and robust approach to coping with large volumes of data submitted for digital forensic examination. *Digital Investigation. Triage in Digital Forensics*. Vol. 10, Issue 2, Sep. 2013, Pages 116–128. doi:10.1016/j.diin.2013.04.003.

SHIAELES, S.; Chryssanthou, A.; Katos, V. On-scene triage open source forensic tool chests: Are they effective?. *Digital investigation*. v. 10, Issue 2., p. 99-115. Sep. 2013. doi:10.1016/j.diin.2013.04.002.

SLADE, R. M. *Software Forensics: Collecting Evidence from the Scene of a Digital Crime*. McGraw Hill, New York. 2004.

SHIELDS, C.; FRIEDER, O.; MALOOF, M.; A system for the proactive, continuous, and efficient collection of digital forensic evidence. *Digital Investigation*, v. 8, Supplement, Aug. 2011, p. S3–S13. The Proceedings of the Eleventh Annual DFRWS Conference.

STEPHENSON, P., *Investigating Computer Related Crime*. CRC Press, Boca Raton, Flórida, 2000.

SWGDE, SCIENTIFIC WORKING GROUP ON DIGITAL EVIDENCE AND INTERNATIONAL ORGANIZATION ON DIGITAL EVIDENCE (IOCE). *Digital evidence: Standards and principles*. Forensic Science Communications, Number 2, Vol. 2.Abr. 2000. U.S. Departament of Justice, FBI. Disponível em: <<http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/april2000/swgde.htm/>>. Acesso em: 12 out. 2011.

SWGDE. SCIENTIFIC WORKING GROUPS ON DIGITAL EVIDENCE AND IMAGING TECHNOLOGY. SWGDE AND SWGIT Digital & Multimedia Evidence Glossary. Version:2.4. Jan. 2011. Disponível em: <http://www.crime-scene-investigator.net/swgde_swgit_glossary_v2-4.pdf>. Acesso em: 21 abr. 2011.

TANENBAUM, A. S. Sistemas operacionais modernos. Prentice-Hall do Brasil, 2010. 3ª. Ed. ISBN 9788576052371.

TESSMANN, C. R. (2012). Priorização da análise pericial pela amostragem estatística de dados de discos rígidos suspeitos. Dissertação de Mestrado, Publicação PPGENE.DM – 099/12, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 58 pp.

VALJAREVIC, A.; VENTER, H. S.; A Comprehensive and Harmonized Digital Forensic Investigation Process Model. Journal of Forensics Sciences. Article first published online: 10 Aug. 2015. DOI: 10.1111/1556-4029.12823. Disponível em: <<http://onlinelibrary.wiley.com/doi/10.1111/1556-4029.12823>>. Acesso em: 18 ago. 2015.

WALTER, C. *Kryder's Law*. Scientific American, 25 Jul 2005. Disponível em: <<http://www.scientificamerican.com/article.cfm?id=kryders-law>>. Acesso em: 02 mai. 2013.

WHITINGTON, J. PDF Explained. Publisher: O'Reilly Media. Print: December 2011. Pages: 142. Ebook ISBN: 978-1-4493-1001-1. ISBN 10: 1-4493-1001-X.

WILES, J., et al. Forensic examination in a terabyte world. Techno security's guide to e-discovery and digital forensics. Elsevier; 2007. p. 129e46. doi:10.1016/B978-159749223-2.50009-1.

WOODS, K.; LEE, C. A.; GARFINKEL, S.; Extending Digital Repository Architectures to Support Disk Image Preservation and Access. Joint Conference on Digital Libraries - JCDL'11 Proceedings of the 2011 ACM/IEEE. Jun. 13–17, 2011, Ottawa, Ontario, Canada. Disponível em: <<http://www.ils.unc.edu/callee/p57-woods.pdf>>. Acesso em: 05 mai. 2012.

YANG, H.; XU, M.; ZHENG, N.; An Improved Method for Ranking of Search Results Based on User Interest. Network and Parallel Computing, 2008. NPC 2008. IFIP International Conference on , vol., no., pp.132,138, 18-21 Oct. 2008. doi: 10.1109/NPC.2008.6

YASINSAC, A. E., R.F.; Marks, D.G.; Pollitt, M.M.; Sommer,P.M. Computer forensics education. Security & Privacy Magazine, 1 (4). 15 - 23. 2003.

APÊNDICE A: ALGORITMO DA FERRAMENTA COPYSCAN

```
//Programa Principal do CopyScan
CopyScan()
    se existir DAD de origem
        então abra o DAD para leitura
    senão
        informe que não foi possível abri-lo
        saia do programa
    se houver lista de varredura
        então carregue e processe todas palavras com CriaListaPalavrasChave()
            obter o tamanho máximo possível dentre todas EncodedKeyword (MAX_ENCODED_SIZE) com
                IniciaPesquisador(FilaDeVarredura, ListaPalavraChavePrincipal)
            inicia as threads de varredura com RodaThreadVarredura()
    se houver DAD de destino
        então abre o destino para gravação
            inicia as threads de duplicação com RodaThreadDuplicação()

    obtém o buffer de um bloco
    se o buffer do bloco é nulo
        então cria um buffer, consultando o tamanho do bloco e o tamanho da maior EncodedKeyword.

    lê dados do DAD de origem
    se a leitura for realizada com sucesso
        então realiza leitura dos dados do disco pelo elemento da fila a partir do índice inicial

    se houver arquivo de destino
        então adiciona a fila do duplicador
    se houver lista de palavras-chave
        então adiciona a fila da varredura

    bloco anterior = bloco atual
    enquanto a quantidade de dados lidos for > 0 faça
        obtém o buffer do próximo bloco
        se o buffer do bloco é nulo
            então cria um buffer, consultando o tamanho do bloco e o tamanho da maior
EncodedKeyword.

        adiciona os últimos q-bytes da leitura anterior no começo do bloco atual
(MAX_ENCODED_SIZE -1)
        se foram lidos os últimos bytes do DAD a ser periciado
            então leia quantidade de bytes com base no limite dos últimos bytes do DAD
        senão
            realiza leitura dos dados do disco no elemento da fila a partir do índice inicial
                de modo que haja espaço no início do elemento da fila para armazenar o
                    trecho de dados do bloco anterior.

    se houver dados disponíveis para leitura
        então obtém o elemento da fila do próximo bloco
```

```

        se houver arquivo de destino
            então adiciona o bloco sem o trecho de dados do bloco anterior na fila de
duplicação
        senão só inspeciona

        se houver lista de palavras-chave
            então adiciona o bloco com o trecho de dados do bloco anterior na fila de
varredura
        senão só copia

        bloco anterior = bloco atual

    adiciona à fila de duplicação um bloco de tamanho 0 e início de tamanho 0.
    adiciona à fila da varredura um bloco de tamanho 0 e início de tamanho 0.
    aguarde o término do duplicador
    aguarde o término da varredura

    feche o DAD de origem
    imprima o resultado final
    se houver ocorrências
        então grave o arquivo do "dedo-duro"

// Inicia a Thread de Duplicação
Procedimento RodaThreadDuplicacao()
    enquanto a thread estiver rodando faça
        obtém um bloco do poll
        se existir um bloco diferente de nulo faça
            escreve no DAD de destino
            se chegou um bloco indicando o fim dos dados
                então pare a thread
                prepare para desligar as atividades após esperar terminar
            senão
                realiza escrita do bloco no DAD de destino

        se não veio um bloco
            então aguarda um pouco (15 milissegundos)
            se não foi possível aguardar
                então mostre a exceção

    desligue esta atividade
    calcule o tempo de execução da thread

// Cria a lista de palavras-chave usando as palavras-chave contidas no arquivo .txt de entrada
Função CriaListaPalavrasChave() {
    calcula todas as combinações de uma palavra-chave em letras maiúsculas e minúsculas
    para todas as combinações da palavra-chave faça
        inclua a combinação na lista de palavras-chave codificada com PalavrasChavesCodificadas()
    se a palavra codificada possui tamanho maior que a atual
        então atualize o tamanho da maior

```

```

devolve lista de palavras-chave

// Cria a lista de palavras-chave codificadas usando as combinações das palavras-chave
Função PalavrasChavesCodificadas() {
    para cada codificação na lista de codificações faça
        obtém os bytes da palavra-chave a ser codificada
        se os bytes da palavra-chave não estão na lista de bytes de palavras-chave
            então insira-os na lista de palavras-chave codificadas
                se o tamanho da palavra-chave for maior que as anteriores
                    então atualize o tamanho da maior palavra-chave
    devolva tamanho da maior palavra-chave

// Inicia o "Módulo de varredura"
Procedimento IniciaPesquisador(FilaDeVarredura, ListaPalavraChavePrincipal)
    para cada palavra-chave principal (mainKeyword) faça
        para cada palavra permutada (keyword) faça
            para cada codificação da palavra faça
                adiciona buffer da palavra-chave codificada no compilador com AdicionaBuffer(buffer)
                insere a palavra-chave codificada na lista de palavras-chave codificada
    monta árvore de busca com CriaMaquinaDeEstados()

// Adiciona um buffer ao compilador
// Entrada: buffer
// Saida: ID do buffer adicionado (retorna -1 se não pode ser adicionado).
Função AdicionaBuffer(ByteBuffer buffer)
    se não existir um buffer no compilador
        então insira o buffer no compilador
            atualize o Id do compilador
    devolva Id

// Cria a árvore abstrata que representa a árvore de possibilidades.
// Saida: Nó raiz da árvore
Função CriaMaquinaDeEstados() {
    para cada caractere da palavra-chave a ser buscada até último faça
        apontar nó pai para o nó raiz.
        para índice = 0 até tamanho palavra-chave faça
            obter o valor corrente
            obter o filho que representa o byte

            se o nó corrente não possuir um filho representando o valor
                então cria um nó
                    adiciona ao nó pai
            se é o último byte da palavra-chave
                então adiciona o caractere da palavra-chave como nó filho
                    nó pai recebe nó filho
    devolva nó raiz da árvore

// Inicia a Thread de Varredura
Procedimento RodaThreadVarredura()
    enquanto a thread estiver rodando faça

```

```

    obtém um bloco do poll
    se existir um bloco diferente de nulo faça
        faz a varredura com Varredura(elemento, 0, Resultado)
    se chegou um bloco indicando o fim dos dados
        então pare a thread
            prepare para desligar as atividades após esperar terminar
    senão
        realiza a busca no bloco com dados

    se não veio um bloco
        então aguarda um pouco (50 milissegundos)
            se não foi possível aguardar
                então mostre a exceção

desligue esta atividade
calcule o tempo de execução da thread

// Efetua a varredura propriamente dita em um vetor de bytes e anota o resultado
// Entrada: um vetor de bytes e Resultado é uma lista de IDs das palavras encontradas
Procedimento Varredura(Vetor, índice, Resultado){
    se existe um vetor de string neste estado (nó)
        então assinala o ID do vetor encontrado no resultado
    obtém o próximo byte do vetor
    se existe um próximo estado (nó)
        então prosseguir a Varredura(Vetor, índice+1, Resultado)

```

APÊNDICE B: UTILIZAÇÃO DA FERRAMENTA COPYSCAN

Depois de compreender o princípio de funcionamento da ferramenta *CopyScan*, é apresentado o modo de operação pelo usuário da ferramenta.

Lembrando que o usuário está encarregado de identificar e extrair as PChs a partir da leitura do processo judicial, entendimento do contexto do processo e nos quesitos a serem respondidos. E inserir essas PChs em um arquivo texto (lista de PChs) salvo em um DAD externo, por exemplo.

O usuário deve inserir esta lista no equipamento onde a duplicação será realizada e disparar a execução do duplicador do primeiro DAD a partir do seguinte comando:

```
java -jar CopyScan.jar if=<caminho DAD origem> of=<caminho DAD
destino> keywords=<Palavras.txt> [bs=<tamanho dos blocos>]
[numcore=<quantidade threads ativas simultaneamente>]
[maxsize=<tamanho DAD origem>] cs |tee Contador_<identificador
do DAD origem>.txt
```

onde:

- <caminho DAD origem>: é o caminho do DAD a ser duplicado
- <caminho DAD destino>: é o caminho do DAD para onde os dados serão copiados
- <Palavras.txt>: é a lista das PChs elaboradas pelo perito
- <tamanho dos blocos>: é o tamanho do buffer que a ferramenta utilizará para fazer a leitura e a gravação dos DADs
- dontcopy: indica se a ferramenta irá realizar apenas a varredura
- hash: indica se o hash será no buffer intermediário
- <quantidade threads ativas simultaneamente>: é a quantidade de *threads* que serão ativadas simultaneamente para o processo de varredura. Sugere-se que este número seja igual a quantidade de cores existentes na máquina onde a busca será executada. Este parâmetro é igual a quatro por padrão.

- <tamanho DAD origem>: é o tamanho do espaço máximo do DAD de origem, ou seja, é o tamanho de uma parte do DAD que será duplicada/inspecionada, por exemplo, caso um DAD de 16GB seja duplicado para um DAD de capacidade superior, desde que a duplicação se inicie no setor zero do DAD de destino, então este parâmetro irá ser igual ao tamanho do DAD de 16GB.
- <identificador do DAD origem>: é o índice do DAD_O, para identificar o arquivo de resposta da busca que será necessário para fazer o *ranking* do conjunto de dispositivos que foram duplicados.
- cs: é um marcador para desativar a opção de combinação de PCh (maiúsculo/minúsculo). Assim, as PChs serão procuradas da mesma forma como aparecem na lista de PChs.

A função do comando `tee`, neste caso, é reproduzir o conteúdo fornecido pela entrada padrão, o resultado da ferramenta *CopyScan*, tanto na saída padrão (vídeo) quanto em um ou mais arquivos para registrar a saída do processamento.

O usuário terá que executar este comando para cada um dos DADs que serão duplicados, alterando o identificador do DAD_O e os respectivos caminhos caso necessário.

A seguir, será apresentado o ranqueamento e triagem que descreve os procedimentos a serem seguidos pelo perito, cujo resultado será utilizado para definir quais serão os DADs a serem analisado na etapa de triagem.