UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

PRIORIZAÇÃO DA ANÁLISE PERICIAL PELA AMOSTRAGEM ESTATÍSTICA DE DADOS DE DISCOS RÍGIDOS SUSPEITOS

CRISTIANO RODRIGUES TESSMANN

ORIENTADOR: RAFAEL TIMÓTEO DE SOUSA JÚNIOR

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
ÁREA DE CONCENTRAÇÃO INFORMÁTICA FORENSE E SEGURANÇA DA INFORMAÇÃO

PUBLICAÇÃO: PPGENE.DM – 099/12

BRASÍLIA / DF: FEVEREIRO/2012

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

PRIORIZAÇÃO DA ANÁLISE PERICIAL PELA AMOSTRAGEM ESTATÍSTICA DE DADOS DE DISCOS RÍGIDOS SUSPEITOS

CRISTIANO RODRIGUES TESSMANN

DISSERTAÇÃO DE MESTRADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE PROFISSIONAL EM INFORMÁTICA FORENSE E SEGURANÇA DA INFORMAÇÃO.

APROVADA POR:

RAFAEL TIMÓTEO DE SOUSA JÚNIOR, Doutor, UnB
(ORIENTADOR)

FLAVIO ELIAS GOMES DE DEUS, Doutor, UnB
(EXAMINADOR INTERNO)

HÉLVIO PEREIRA PEIXOTO, Doutor, MJ/DPF
(EXAMINADOR EXTERNO)

DATA: BRASÍLIA/DF, 10 DE FEVEREIRO DE 2012.

FICHA CATALOGRÁFICA

TESSMANN, CRISTIANO RODRIGUES
Priorização da análise pericial pela amostragem estatística de dados de discos rígidos suspeitos [Distrito Federal] 2012. xxii, 58 pp., 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2012).

Dissertação de Mestrado – Universidade de Brasília, Faculdade de Tecnologia. Departamento de Engenharia Elétrica.

1. Amostragem Estatística 2. Informática Forense 3. Raciocínio Baseado em Casos.

I. ENE/FT/UnB. II. Título (Série)

REFERÊNCIA BIBLIOGRÁFICA

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.

CESSÃO DE DIREITOS

NOME DO AUTOR: Cristiano Rodrigues Tessmann
TÍTULO DA DISSERTAÇÃO: Priorização da análise pericial pela amostragem estatística de dados de discos rígidos suspeitos.
GRAU/ANO: Mestre/2012.

Cristiano Rodrigues Tessmann
Av. Princesa Isabel, 1056
CEP 90620-000 – Porto Alegre – RS - Brasil

A minha família,
pelo incentivo aos estudos.

AGRADECIMENTOS

Ao meu orientador Prof. Dr. Rafael Timóteo de Sousa Júnior, pela paciência, bom humor, pensamento positivo e otimismo frente aos obstáculos que foram superados.

Ao Prof. Laerte Peotta, por apontar uma janela de conhecimento onde eu só via uma parede opaca.

Ao Perito Criminal Federal, Dr. Helvio Pereira Peixoto, pelo esforço e dedicação na idealização do projeto de Mestrado junto ao PRONASCI.

A todos, os meus sinceros agradecimentos.

O presente trabalho foi realizado com o apoio do Departamento Polícia Federal – DPF e do Instituto-Geral de Perícias – IGP, da Secretaria de Segurança Pública – SSP, do Estado do Rio Grande do Sul, com recursos do Programa Nacional de Segurança Pública com Cidadania – PRONASCI, do Ministério da Justiça.

RESUMO

PRIORIZAÇÃO DA ANÁLISE PERICIAL PELA AMOSTRAGEM ESTATÍSTICA DE DADOS DE DISCOS RÍGIDOS SUSPEITOS

Autor: Cristiano Rodrigues Tessmann

Orientador: Rafael Timóteo de Sousa Júnior

Programa de Pós-graduação em Engenharia Elétrica

Brasília, Fevereiro de 2012

Os notórios avanços da informática e sua democratização foram responsáveis por facilitar no ambiente doméstico certos tipos de crime, como os relacionados à pirataria e pedofilia. O que elevou drasticamente a quantidade e a diversificação de equipamentos submetidos à análise Pericial por suspeita de envolvimento em crimes.

Frente à demanda crescente de análises, este trabalho propõe a análise de pequenas amostras de dados de forma a determinar os objetos questionados com maior potencialidade de conter os vestígios procurados em um caso envolvendo crimes relacionados à informática, priorizando estes objetos para a análise pericial, possibilitando que os primeiros resultados conclusivos sejam obtidos em um tempo reduzido.

Como forma de aumentar a precisão dos resultados, propõe-se o uso de técnicas de Raciocínio Baseado em Casos para determinar as palavras-chave a serem usadas nas buscas realizadas nos dados amostrados.

Os resultados deste trabalho apontam que, apesar do acesso a um volume de dados reduzido, o tempo necessário para a análise não é proporcionalmente menor, devido ao tempo de busca necessário a estes dados nos dispositivos analisados (HDD).

Outro importante resultado é o fato de a análise por amostragem só ter utilidade prática em busca de conteúdos muito representativos no objeto questionado.

O trabalho conclui que o método proposto é viável, e que é possível o uso dos resultados obtidos para a priorização proposta.

ABSTRACT

PRIORITIZATION IN FORENSIC ANALYSIS BASED ON STATISTICAL DATA SAMPLING OF SUSPECT HARD DRIVES

Author: Cristiano Rodrigues Tessmann

Supervisor: Rafael Timóteo de Sousa Júnior

Programa de Pós-graduação em Engenharia Elétrica

Brasília, February of 2012

The notable advances in information technology and its democratization were responsible for facilitating at the home environment certain types of crime, such as those related to piracy and pedophilia. This demonstration increased the amount and diversity of equipment subjected to forensic analysis by suspicion of involvement in certain crimes, what poses a challenge to the forensic community.

Given this growing demand for analysis, this work proposes the analysis of small samples of data to be analyzed to determine which questioned objects have the greatest potential to contain relevant evidence, prioritizing the forensic analysis of these objects in a case involving computer-related crimes, allow the first conclusive results to be obtained in a short time.

In order to increase the accuracy of the results, we propose the use of Case-Based Reasoning to determine the keywords to be used in searches to be performed on the data sampled.

The results present in this dissertation show that, despite access to a small volume of data, the time required for analysis is not proportionally lower due to the time necessary to search the data analyzed in the devices (HDD).

Another important fact about the results is that the analysis by sampling only have practical use for content very representative object in question.

The work concludes that the proposed method is feasible and that it is possible to use results obtained for the proposed prioritization.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE SÍMBOLOS, NOMENCLATURA E ABREVIAÇÕES

ASCII	AMERICAN STANDARD CODE FOR INFORMATION INTERCHANGE
CPP	CÓDIGO DE PROCESSO PENAL
DD	DATA DUPLICATION
DNA	DEOXYRIBONUCLEIC ACID
EI	ENTIDADE DE INFORMAÇÃO
FBI	FEDERAL BUREAU OF INVESTIGATION
FTK	FORENSIC TOOLKIT
HDD	HARD DISK DRIVE
MJ	MINISTÉRIO DA JUSTIÇA
MB	MEGABYTE
NCQ	NATIVE COMMAND QUEUING
NIST	NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
ONU	ORGANIZAÇÃO DAS NAÇÕES UNIDAS
PC	PERSONAL COMPUTER
PGP	PROTOCOLO-GERAL DE PERÍCIAS
RBC	RACIOCÍNIO BASEADO EM CASOS
RPM	ROTAÇÕES POR MINUTO
SSD	SOLID-STATE DRIVE
TB	TERABYTE

1. INTRODUÇÃO

Os avanços globais e notórios da informática trazem muitos benefícios para a sociedade, como a democratização da informação, a redução de custos em diversos setores, e o auxilio em praticamente todas as áreas do conhecimento (Carmo, 2010). O valor desta tecnologia e suas ferramentas não foi somente notado por empresas e a sociedade em geral, mas também pelos criminosos (Lange, 2010).

Segundo (Colli, 2010) e (Eleutério, 2011) a informática, quando usada como uma ferramenta, não se distingue da maioria das ferramentas no referente a seu objetivo final, ou seja, depende do uso que se dá a ela, e da intenção de seu utilizador. Um avião, por exemplo, pode ser um meio de transporte nas mãos de um determinado usuário, ou uma arma de guerra nas mãos de outro (Carmo, 2010).

Assim, computadores e vários outros dispositivos similares, a maioria conectada a Internet nos dias de hoje, podem ser utilizados para fins criminosos. Não é necessário que a vítima use ou mesmo saiba o que é a informática, pois sendo utilizada como uma ferramenta é um meio alternativo, e muitas vezes vantajoso, para se chegar a um fim delituoso. Tem-se como exemplo a falsificação de documentos, que não mais necessita de equipamentos gráficos especializados ou grandes conhecimentos na área de impressos, e também a Pedofilia, onde a vítima muitas vezes sequer ainda foi alfabetizada.

É classificado como crime impróprio por (Redivo, 2007) todo o crime que é facilitado, mas poderia ocorrer mesmo sem o auxílio da informática. Já (Eleutério, 2011) classifica os tipos de crimes quanto à utilização do computador, como uma ferramenta ou como meio para o crime, sendo que, neste último caso, o crime seria impossível de ocorrer sem a existência do equipamento.

Quanto ao arcabouço social para o tratamento de tais tipos de crimes, (Lange, 2010) destaca a falta de legislação específica, o que traz dificuldades para a justiça lidar com este novo tipo de delito, garantindo, muitas vezes, a impunidade por não haver tipicidade penal prévia.

Por outro lado, os produtos desta tecnologia também são utilizados por órgãos de segurança pública no Brasil e no exterior para a prevenção, combate e apuração de atos criminosos (MJ, 2011, NIST, 2011 e FBI, 2011).

O aumento da utilização da informática com fins criminosos (Colli, 2010) (Lange, 2010) (Eleutério, 2011) foi o responsável por criar áreas especializadas em seu combate.

No entanto a disseminação e diversificação dos equipamentos levaram certos crimes para o ambiente doméstico, como por exemplo, a violação de direitos autorais (comumente chamada de pirataria) e a distribuição de pornografia envolvendo crianças e adolescentes (frequentemente chamada de pedofilia) (Guedes, 2009) (Lange, 2010). Aliando-se a isto, o fato da capacidade de armazenamento dos equipamentos de informática crescer continuamente (StorageReview, 2011) (Morimoto, 2010), tem-se hoje um volume cada vez maior e diversificado de dados a serem analisados, com o objetivo de utilizá-los, ou descartá-los, como prova material de um fato criminoso.

(Hoelz, 2009) evidencia a demanda exponencial por análises periciais em informática na Polícia Federal Brasileira, citando que as unidades de discos rígidos (HDD) examinadas normalmente são do tipo daquelas comercializadas em maior quantidade nos 12 a 24 meses anteriores. Este dado é verificado também em outros órgãos, como no Instituto-Geral de Perícias do Estado do Rio Grande do Sul (Figura 1.1).

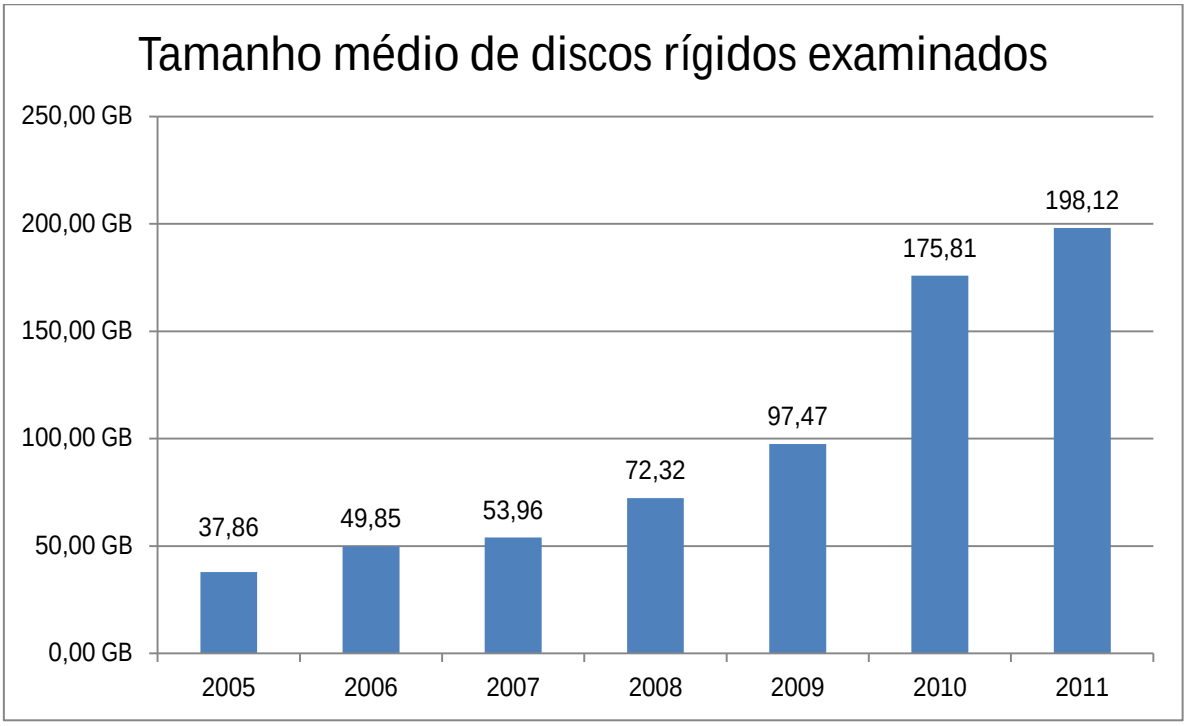

Figura 1.1 – Tamanho médio dos discos rígidos examinados no IGP/RS

Um cidadão não pode ser considerado um criminoso apenas por suspeitas, mesmo que fundamentadas (Constituição, 1988. Art. 5º), e a polícia não pode abster-se de investigar qualquer denúncia ou suspeita fundamentada (CPP, 1941. Art. 5º), o que eleva o número

global de indícios a serem analisados para materializar evidências relacionadas aos atos suspeitos investigados, e consequentemente a necessidade de um número maior de Laudos Periciais.

Tal situação motiva o desenvolvimento de abordagens para ajudar na comprovação das suspeitas, tal como a abordagem proposta na presente dissertação.

1.1. O PROBLEMA

Este trabalho aborda o problema do aumento contínuo da demanda por análises periciais em casos envolvendo crimes relacionados à informática que, aliado ao crescimento da capacidade de armazenamento de dados individuais, torna inviável financeiramente o aumento do número de Peritos como único meio para tratar todos os casos suspeitos no prazo necessário. Este problema torna evidente a necessidade do aumento da produtividade, a automação de procedimentos, e mesmo uma melhor triagem da demanda para priorização dos casos mais importantes, evitando assim a demora nos primeiros resultados, que podem ser fundamentais como apoio a decisões no processo investigatório.

O problema da filtragem de um grande volume de dados, para a realização de uma triagem, esbarra no tempo necessário para acessar e interpretar estes dados. Independente da velocidade da interface do dispositivo com o computador ou equipamento de leitura, todos os dispositivos mecânicos, comuns até hoje na maioria dos computadores, têm velocidades limites físicas de acesso (Morimoto, 2007), o que é, hoje, inferior a 100 MB/s em média[1] na maior parte dos dispositivos domésticos (StorageReview, 2011). Desta forma, o simples acesso ao conteúdo integral de um dispositivo de 3 TB de capacidade, com tecnologia atual, levaria mais de 8,3 horas por dispositivo.

1.2. OBJETIVOS

Este trabalho concentra-se na busca por meios de acelerar a obtenção dos primeiros resultados em Pericias em Informática Forense de forma confiável.

Os resultados aqui buscados podem ser aplicados na otimização do tempo e recursos em processos passíveis de automatização, como, neste caso, a redução do tempo de triagem de casos envolvendo crimes relacionados à informática. Isto possibilitará o direcionamento do

[1] Considerando a velocidade média de acesso a toda a mídia.

tempo e dos recursos na análise prioritária de dispositivos que contenham vestígios relevantes à investigação em curso.

O objetivo deste trabalho é propor um método de pesquisa automatizada de palavras-chave em dados amostrados aleatoriamente de uma unidade de armazenamento de dados questionada, usando para isto técnicas estatísticas e de Raciocino Baseado em Casos.

1.3. MÉTODO

Propõe-se a amostragem aleatória de dados armazenados digitalmente em dispositivos computacionais, como, por exemplo, unidades de discos rígidos (HDD) encaminhadas a Perícia por suspeita de crimes envolvendo informática, buscando identificar expressões (texto) possivelmente relacionadas ao fato investigado/questionado.

Conforme (Garfinkel, 2010) a amostragem aleatória de setores de um disco mostra-se eficiente para determinar o tipo de conteúdo armazenado neste dispositivo (imagens, músicas, etc.), analisando eventuais assinaturas e/ou cabeçalhos de arquivos encontrados nos setores amostrados, independentemente do sistema de arquivos utilizado.

A técnica de Garfinkel não distingue arquivos disponíveis de arquivos excluídos, mas é particularmente eficaz na estimativa de área ocupada por dados e na determinação de discos vazios, seja por se tratarem de dispositivos novos, ou pelo fato de seus dados terem sido apagados por técnicas de sobreposição de dados padronizados – wipe (Garfinkel, 2003).

Ao invés de determinar o tipo de conteúdo armazenado, propõe-se usar busca textual nos dados amostrados, buscando a identificação de expressões relevantes (por exemplo, palavras conhecidas e frequentemente encontradas em equipamentos previamente analisados por suspeitas similares), e a partir destas, determinar a relevância do equipamento analisado em um grupo de equipamentos questionados, permitindo sua ordenação e priorização.

Segundo (Eleutério, 2011), a busca por palavras chave é um meio eficiente para encontrar a maioria das evidencias digitais necessárias para a elaboração de laudos forenses.

(Beebe, 2007) afirma ainda que a evidência textual é importante para a grande maioria das investigações digitais. Isso ocorre porque uma grande quantidade de dados armazenados em formato digital é linguística por natureza (linguagens humanas, linguagens de programação, logs e registros de sistemas ou aplicações, por exemplo). Alguns exemplos de importantes

evidências baseadas em texto incluem: e-mail, histórico de navegação Internet (logs e o próprio conteúdo), mensagens instantâneas, documentos de texto, planilhas, apresentações, catálogos de endereços, agendas de compromissos, logs de atividade de rede, e os registros do sistema.

Na abordagem aqui proposta, através de uma amostragem probabilística, será utilizada inferência estatística para estender o resultado ao objeto questionado como um todo, pesando sua relação com casos similares com resultado final conclusivo e obtidos através de análise tradicional resultando em Laudo Pericial Criminal.

O tamanho da amostra inicialmente proposto é $\overline{\ }$, onde é o número total de *clusters* do dispositivo questionado. Tendo como base a recomendação da Organização das Nações Unidas – ONU em (Schiavone, 2009) para a amostragem de narcóticos. Os resultados obtidos serão confrontados com amostras de tamanho 0,48% de , julgada relevante por (Garfinkel, 2010) em sua análise de um dispositivo de 1TB de capacidade, e 1% de , proposto para fins de comparação.

A escolha das palavras-chave (*keywords*) a serem pesquisadas nos dados amostrados se dará através da recuperação da tabela de palavras-chave de casos similares arquivados, armazenada após a análise compreensiva destes e da obtenção de seus resultados (Laudo Pericial Criminal Conclusivo). Técnicas de Raciocínio Baseado em Casos serão utilizadas para determinação da similaridade entre o caso proposto e os casos arquivados.

O número de ocorrências (*hits*) de cada palavra-chave considerada relevante nos casos arquivados ou a soma das ocorrências de todas as palavras-chave analisadas nos casos suspeitos será usado para comparação entre um ou mais casos, de uma mesma investigação, submetidos à triagem com o objetivo de serem classificados por prioridade.

Desta forma, ao inserir os dados básicos do caso (tipo de perícia, data da ocorrência) e seus questionamentos (quesitos) no sistema proposto, um sistema de raciocínio baseado em casos recuperará uma lista formada pelas palavras-chaves mais relevantes do caso mais similar encontrado dentre os casos concluídos arquivados. No entanto este sistema, só armazenará novas palavras-chave para atualização de seu banco de dados (retenção de novos casos), após a análise completa do caso (de forma assíncrona).

Para os experimentos serão utilizadas cópias de dados já analisados de casos reais (imagens de HDDs), arquivados para este propósito, os quais possuem Laudos Periciais obtidos da forma tradicional (análise compreensiva de todo o conteúdo fornecido) contendo o resultado das análises solicitadas para cada caso.

1.4. ESCOLHA DA FERRAMENTA

O software EnCase, na versão "Forensic for Law Enforcement" 6.18.1 da empresa Guidance Software (Guidance, 2011), foi escolhido para executar os experimentos (buscas por palavras-chave) por ter sido a ferramenta utilizada como apoio à análise original dos dados de teste (casos arquivados). O aplicativo EnCase, juntamente com o aplicativo FTK, da empresa AccesData (AccessData, 2011), são as ferramentas periciais para Informática Forense mais usadas no Brasil por órgãos de segurança pública[2], e considerados as principais ferramentas na área de Informática Forense por (Eleutério, 2011). Soma-se a este fato, o recurso Enscript, presente no EnCase, que possibilita a execução de trechos de código personalizados (*scripts*) em uma linguagem própria, similar a Java (Garfinkel, 2009), que acrescenta uma funcionalidade importante a ferramenta, na viabilização da futura implementação da solução proposta. Embora menos intuitivo e considerado menos produtivo que o FTK (Eleutério, 2011) a funcionalidade de personalização do EnCase, não presente no aplicativo FTK, até a última versão analisada (versão 3.0), torna-se determinante na escolha da ferramenta.

Para que a comparação dos resultados finais, principalmente no tocante ao tempo de análise automatizada (tempo máquina), será utilizado para execução dos experimentos o mesmo equipamento nas mesmas circunstâncias da análise original (descrita no ANEXO A).

1.5. RESULTADOS ESPERADOS

Espera-se que os resultados obtidos com a aplicação da técnica de análise por amostragem aleatória, mostrem-se coerentes e proporcionais aos resultados obtidos com o método de análise tradicional, apresentando, neste caso, uma alta probabilidade de conter a informação questionada nos casos em que o resultado do Laudo Pericial foi positivo, em um tempo reduzido. No caso de um resultado incompatível, espera-se o aumento da precisão com uma nova amostragem ou mesmo o aumento da amostra. Casos muito específicos, onde a conclusão original não esteja baseada nas informações analisadas na amostragem (*search*

[2] Segundo levantamento realizado com Peritos Oficiais Estaduais e do Distrito Federal, bem como da Polícia Federal Brasileira.

hits), ou estes tenham sido irrelevantes, serão classificados como incompatíveis com busca textual, fazendo com que o processo de triagem, neste caso, não tenha qualquer interferência na decisão de priorização do caso em questão para a análise tradicional.

Espera-se também que os resultados dos experimentos apontem qual o tamanho de amostra mais eficaz, que traga os melhores resultados sem comprometer o desempenho esperado para um processo de triagem.

1.6. A ORGANIZAÇÃO DESTE DOCUMENTO

A organização do restante deste trabalho é descrita a seguir:

O segundo capítulo apresenta a revisão bibliográfica sobre o tema abordado, descrevendo o estado da arte dos estudos sobre análise por amostragem, raciocínio baseado em casos e informática forense, aplicáveis a este trabalho.

O terceiro capítulo apresenta a técnica de análise por amostragem utilizada, a escolha das amostras, a técnica de raciocínio baseado em casos proposta para o cálculo de similaridade entre os casos e seu método de aquisição de conhecimento.

O quarto capítulo deste trabalho apresenta os resultados obtidos com a aplicação da técnica de amostragem proposta, nos casos de teste, e o confronto deste com os resultados obtidos com a forma tradicional de análise pericial no mesmo caso, suas peculiaridades, exceções, e ajustes realizados.

Por fim, é apresentada no quinto capítulo uma análise dos resultados obtidos, suas contribuições, conclusões finais do trabalho, e uma proposta para trabalhos futuros sobre o tema.

2. TRABALHOS RELACIONADOS

Este capítulo apresentará os elementos teóricos usados para abordar o problema em estudo, discutindo como estes se integram dentro da proposta de trabalho desta dissertação. As seções a seguir apresentarão o estado da arte dos estudos sobre informática forense, análise por amostragem e raciocínio baseado em casos até o momento.

2.1. INFORMÁTICA FORENSE

Várias expressões são utilizadas para descrever a área da informática que lida com evidências digitais. Termos como Computação Forense, Forense Computacional e Informática Forense são alguns dos sinônimos mais comuns dessa área da computação. (Hoelz, 2009) comenta que o termo adequado seria "Perícia em Computadores", por ser originário do termo em inglês "*Computer Forensics*", e o profissional que atua nesta área é chamado de "Perito" ou "especialista", por ser originário da expressão em inglês "*Expert*". Neste trabalho será utilizado o termo Informática Forense por ser uma forma amplamente difundida de referenciar a perícia em equipamentos digitais.

Segundo (Eleutério, 2011), a Computação Forense tem como objetivo principal determinar à dinâmica, a materialidade e autoria de ilícitos ligados à área de informática, conferindo-lhes validade probatória em juízo.

(Hoelz, 2009) define o exame pericial de sistemas computacionais como a preservação, análise e apresentação dos vestígios relacionados à prática de algum crime ou a incidentes de segurança envolvendo computadores. Afirma ainda que esta é uma atividade complexa, que demanda grandes quantidades de recursos computacionais e humanos, além do tempo necessário para sua realização, e que estes são muitas vezes escassos ou insatisfatórios, prejudicando assim a obtenção de resultados para a elucidação de crimes.

A Informática Forense é a área da computação relacionada à coleta, preservação, identificação, análise e interpretação de evidências digitais, bem como a materialização destas como provas judicialmente.

(Beebe, 2005) afirma que investigações digitais, sejam de natureza forense ou não, necessitam de rigor científico e que são facilitadas através do uso de processos padronizados. Tais processos podem ser de natureza complexa.

Atualmente as unidades de armazenamento de dados, tanto em computadores pessoais (PCs), quanto em grandes equipamentos computacionais, são principalmente unidades de discos rígidos, que armazenam a informação de forma magnética com uso de atuadores mecânicos. Estas unidades de discos rígidos atingem hoje a capacidade de TeraBytes de armazenamento, onde milhões de arquivos e outros registros podem ser armazenados. Desta forma, a Informática Forense tem utilizado, cada vez mais, técnicas e métodos científicos para encontrar a evidência questionada em meio a este volume de dados (Beebe, 2007).

Informática forense é uma área multidisciplinar que envolve diversas áreas das Ciências da Computação, como, Criptografia, Sistemas de Arquivos, Sistemas Operacionais, Bancos de Dados, Redes de Computadores, Arquitetura de Computadores, Programação, entre outras. Além de, assim como a criminalística, estar relacionada com inúmeras áreas em outras ciências, devido ao seu uso como ferramenta em crimes convencionais (Eleutério, 2011).

Uma evidência digital trata-se de informação armazenada ou transmitida por meios digitais, sendo que esta não é necessariamente o dispositivo, mas a informação contida em algum dispositivo, mídia, ou outro portador desta. A evidência em si é volátil e sujeita a modificações (Mesquita, 2011), o que requer uma coleta adequada e uma cadeia de custódia confiável, a fim de garantir sua integridade e valor probatório (Eleutério, 2011) (Hoelz, 2009).

Segundo (Lange, 2010), é errado considerar uma evidência digital como virtual ou imaterial, uma vez que sua unidade básica é o bit, e a existência física deste pode ser comprovada nas mídias de armazenamento.

O uso de métodos científicos, comprovados e bem compreendidos nos exames periciais é importante para que o resultado destes tenha valor como prova (Hoelz, 2009).

Na área criminal, é o Perito Criminal (servidor público) o responsável por analisar objetos questionados em busca das provas (evidências). Este profissional irá realizar os exames necessários e descrever os resultados encontrados e suas conclusões em um laudo pericial remetido ao solicitante, que é normalmente uma autoridade Policial ou Judiciária (CPP, 1941).

2.1.1. Fases do exame em Informática Forense

(Eleutério, 2011) sugere quatro etapas necessárias para a realização de exames forenses em dispositivos de armazenamento computacional, como, por exemplo, um disco rígido, conforme Figura 2.1.

Figura 2.1 – Fases do exame em Informática Forense, segundo (Eleutério, 2011)

- Fase 1 – Preservação: Consiste em garantir que as informações armazenadas no equipamento questionado não sejam alteradas. Desta forma, os procedimentos seguintes podem ser replicados, verificados, ou, se necessário, auditados. A duplicação da evidência é uma das técnicas sugerida para a preservação desta;
- Fase 2 – Extração: Trata-se da recuperação, indexação e identificação das informações presentes no dispositivo questionado, preferencialmente de forma automática, para que estas possam ser mais facilmente analisadas no passo seguinte;
- Fase 3 – Análise: É a fase onde os dados são interpretados, buscando e coletando evidências que tenham relação com o crime investigado. Esta fase, apesar de poder ser apoiada por ferramentas especializadas, é a que mais necessita da experiência e capacitação do Perito Criminal;
- Fase 4 – Formalização: A última fase consiste na elaboração do Laudo Pericial pelo Perito, apresentando os resultados, as evidências e suas conclusões. Neste documento constam também as respostas aos quesitos, quando solicitados explicitamente.

(Eleutério, 2011) não prevê uma fase anterior a de preservação, assumindo que todo o objeto apresentado à análise já estará devidamente identificado e será uma possível fonte de evidências.

(Reith, 2002) comenta a importância da padronização de procedimentos para exames em Informática Forense, citando suas vantagens e desvantagens, propõe uma metodologia independente de tecnologia ou tipo de crime (abstrata), com os seguintes passos:

- Identificação: Determinar o tipo de incidente, ou crime no caso;
- Preparação: Preparar as ferramentas e documentação necessárias para os procedimentos a serem adotados;
- Estratégia de Abordagem: Definir uma estratégia que possibilite a maximização da coleta de possíveis evidências com mínimo impacto;
- Preservação: Isolar e proteger as evidências;
- Coleta: Cópia dos dados a serem analisados;

- Exame: Busca nos dados coletados de evidências que possam estar relacionadas com o crime;
- Análise: Interpretação dos dados e determinação de sua relevância, tirar conclusões baseadas em evidências encontradas;
- Apresentação: Resumir e apresentar as conclusões. Esta apresentação, segundo (Reith, 2002) deve sempre ser escrita em termos leigos, utilizando uma terminologia não técnica;
- Retorno da evidência: O objeto questionado deve ser devolvido preservado.

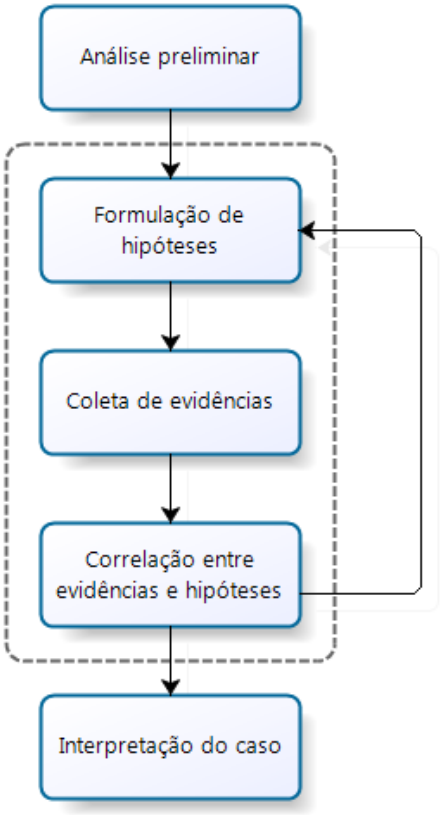

Figura 2.2 – Processo Investigativo, adaptada de Reith et al. (2002) por (Hoelz, 2009).

A Figura 2.2 mostra uma adaptação de (Hoelz, 2009) ao processo proposto por (Reith, 2002). Nesta adaptação, inicialmente é feita uma análise preliminar do caso, em seguida são formuladas hipóteses com base nesta análise, e então são coletadas possíveis evidências com base nessas hipóteses. Em seguida, as evidências coletadas são correlacionam com as hipóteses, e, quando necessário, ajustadas repetindo o processo até que a interpretação do caso esteja consistente.

(Hoelz, 2009), citando (Reith, 2002), afirma que os quatro elementos chave do processo investigativo são a identificação, preservação, análise e apresentação das evidências, e que

independente do número de passos apresentado na literatura, eles sempre podem ser enquadrados no quatro elementos da Figura 2.3.

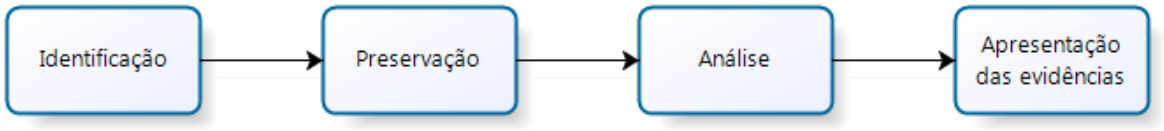

Figura 2.3 – Elementos Chave no processo de investigação, segundo (Hoelz, 2009)

Seja na identificação do que será submetido à perícia (Reith, 2002), ou análise preliminar (Hoelz, 2009), é nesta etapa que os objetos questionados podem ser classificados e priorizados, de forma a otimizar os recursos em etapas mais dispendiosas, e automatizadas, como a preservação (duplicação forense), e os exames iniciais automatizados (pesquisas, filtros, indexação), evitando assim que no momento da análise um objeto seja descartado como evidência, por algum motivo que poderia ser detectado com antecedência na triagem.

2.2. ANÁLISE POR AMOSTRAGEM

Segundo (Schiavone, 2009) e (Lucy, 2005) para tratar itens em grande quantidade, a amostragem é amplamente considerada como uma abordagem razoável, inclusive na área jurídica, muitas vezes, permitindo que um cientista chegue à conclusão final de sua análise usando inferência estatística.

(Lucy, 2005) justifica que, a exemplo dos exames de DNA que não dão certeza absoluta, mas são bem aceitos juridicamente, várias outras áreas forenses podem se valer de análises estatísticas para embasar suas conclusões.

Segundo (Neto, 1977), o sucesso de uma análise estatística envolve aspectos importantes sobre as formas de amostragem, e é preciso garantir que a amostra ou amostras que serão usadas sejam obtidas por processos adequados.

É necessário garantir que a amostra seja representativa da população. A forma mais confiável e normalmente acessível de se obter uma amostra representativa é a Amostragem Probabilística. (Neto, 1977) (Lucy, 2005) (Morettin, 2010).

2.2.1. Vícios de Seleção

Vícios, ou viés de seleção, é a tendência a escolher amostras baseado numa perspectiva parcial em detrimento de outras alternativas. Em estatística, é um termo usado para expressar

o erro sistemático ou tendenciosidade, ou seja, uma escolha que não respeite os princípios da imparcialidade.

Segundo (Morettin, 2010), uma amostragem não probabilística, onde alguns elementos da população podem ter probabilidade desconhecida ou até nula de pertencer à amostra, como amostras intencionais, a esmo ou de voluntários, pode induzir a vícios de seleção.

Uma clássica consideração sobre amostras viciadas data da Segunda Guerra Mundial quando o matemático Abraham Wald (Wald, 1943) aplicou seus conhecimentos de estatística para resolver o problema de perdas de bombardeiros ao fogo inimigo. Para estimar a vulnerabilidade dos aviões, Wald observou as aeronaves que retornavam, cheias de buracos, das missões de combate. Ao contrário de um estudo prévio onde tinha sido proposto que uma armadura (reforço) deveria ser adicionada às áreas que apresentaram o maior dano, Wald recomendou, que as partes dos aviões que não haviam recebido tiros eram as mais vulneráveis, e deviam ser reforçadas. Na visão de Wald os danos encontrados nos bombardeiros que retornaram foram ocasionados em áreas onde eles eram capazes de resistir, e as áreas não atingidas eram os pontos fracos e que devem ser reforçadas.

A resposta de Wald não estava nos aviões que ele analisou, pois estes eram os aviões que haviam retornado com sucesso, apesar dos danos. Wald levou em conta que os aviões que retornavam, ainda que cheios de buracos, haviam resistido às avarias o suficiente para fazer a viagem de volta. Os buracos indicavam os locais mais robustos, que podiam resistir aos danos.

A análise de Wald considerou o viés de seleção. O conjunto de dados analisado já tinha sido selecionado de alguma forma, neste caso, a amostra era viciada, pois só considerava os casos de sucesso, e os demais não eram passíveis de análise.

Na criminalística podemos incorrer nos mesmos vícios, principalmente por características inerentes do ser humano (viés cognitivo), como a tendência das pessoas em favor de informações que confirmem suas hipóteses, independentemente de serem ou não verdadeiras. Como resultado, as pessoas colhem evidências e trazem informações da memória de forma seletiva, interpretando-as de maneira enviesada (tendenciosa) (Baron, 2000, p.195).

Uma possível causa que pode levar a vícios na seleção dos dados a serem analisados é o simples fato de o Perito poder estar “buscando as evidências” no material questionado, ao invés de analisá-lo com isenção.

A melhor maneira de evitar vícios de seleção é o uso de sorteio aleatório considerando toda a população. Segundo (Morettin, 2010), a **amostragem probabilística é isenta de viés de seleção**.

2.2.2. Amostragem Probabilística

Amostragem Probabilística é usada quando uma amostra selecionada de tal forma que cada membro da população estudada têm uma probabilidade (não nula) conhecida de ser incluída na amostra (Neto, 1977) (Morettin, 2006). Por ser um método não determinístico aumenta a confiabilidade do resultado final, no caso da existência de um adversário que possua condições de promover alterações pontuais na população, buscando prever ou direcionar o resultado de uma análise.

2.2.2.1. Métodos de Amostragem Probabilística

(Neto, 1977) cita os tipos de Amostragem Probabilística tradicionais:

- Na Amostragem Aleatória Simples, uma amostra é escolhida de tal forma que cada item ou membro na população tem a mesma probabilidade de ser incluído. Se a população tem um tamanho , cada membro desta população tem a mesma probabilidade, igual a –, de entrar na amostra;

- Na Amostragem Aleatória Sistemática, os membros da população são ordenados de alguma forma, através de algum método conhecido ou pré-determinado. Um ponto de partida aleatório é sorteado, e então cada k-ésimo membro da população é selecionado para a amostra. Apesar de haver aleatoriedade inicial, este método fornece um padrão que pode vir a ser detectado e usado para esconder informações ou mesmo evidenciar informações pré-determinadas de forma maliciosa;

- Na Amostragem Aleatória Estratificada, a população é inicialmente dividida em subgrupos (estratos) e uma subamostra é selecionada a partir de cada estrato da população. Este método é útil quando se conhece a população, por exemplo, um sistema de arquivos conhecido poderia ser usado para diferenciar dados disponíveis de excluídos, ou mesmo definir estratos com base nos períodos de tempo que se busca analisar;

- Já na Amostragem aleatória Estratificada com Repartição Proporcional, os estratos devem ser os mais homogêneos possíveis com relação às características relevantes da

pesquisa (variáveis que se correlacionam fortemente com a variável estudada, como, por exemplo, imagens ou arquivos ASCII contendo registros de acontecimentos – *logs*). Para um mesmo tamanho amostral, a amostragem aleatória estratificada com repartição proporcional é mais precisa (menor variância do estimador) do que a amostragem aleatória simples;

- A Amostragem Aleatória Estratificada com Repartição de Neyman, é a mais adequada no caso da variância de cada extrato populacional referente a variável que se deseja estimar é conhecida, pois para um mesmo tamanho amostral a precisão é maior do que para a amostra aleatória estratificada com repartição proporcional, que por sua vez é maior do que a amostra aleatória simples;
- Na Amostragem por Conglomerados, a população é subdividida inicialmente em subgrupos (estratos) e uma amostra de estratos é selecionada com probabilidade proporcional ao tamanho de cada estrato. A seguir, amostras são selecionadas dos estratos selecionados previamente. A principal vantagem da amostra por conglomerados é a de possibilitar considerável redução de custos (em relação a uma amostragem aleatória estratificada, por exemplo) para um mesmo tamanho amostral. Este método costuma ser empregado quando não dispomos de um índice da população (como no caso da amostragem sistemática) e os custos de ser elaborado um índice para toda a população é muito elevado.

(Garfinkel, 2010) evidencia o uso da amostragem probabilística como um meio eficiente para determinar o tipo de conteúdo armazenado em um dispositivo desconhecido, reforçando o fato da técnica de amostragem não determinística diminuir a possibilidade de um adversário mal intencionado direcionar os resultados obtidos.

(Schiavone, 2009) descreve vários métodos de amostragem. Ele se concentra em amostragem nos casos envolvendo apreensões de drogas ilícitas em que um grande número de materiais relativamente homogêneo está disponível. As vantagens e desvantagens de vários métodos, incluindo seu uso na prática, são apresentadas. Ele considera que uma abordagem Bayesiana é razoável em muitos casos, mas a sua complexidade pode ser uma grande desvantagem, especialmente no meio jurídico (tribunal). Felizmente, abordagens hipergeométrica e Bayesiana parecem mostrar mais ou menos os mesmos resultados em casos onde são aplicados métodos menos complexos. O objetivo final do seu trabalho é fornecer métodos

práticos e simples de determinar o número mínimo de amostras a se obter de forma a trazer resultados confiáveis, e juridicamente aceitos, em uma análise laboratorial.

Considerando que a população a ser amostrada, no caso deste trabalho, são conjuntos de dados de uma unidade de discos rígidos, e que o conteúdo deste, em princípio, é completamente desconhecido, inclusive quanto a sua organização ou existência (o disco pode não conter dados), o método de Amostragem Aleatória Simples é o mais indicado por ser simples, seguro e, neste caso, veloz (por não envolver ou considerar a possível criação de um índice).

2.3. RACIOCÍNIO BASEADO EM CASOS

RBC é um modelo de tomada de decisão baseado na experiência adquirida com a resolução de problemas anteriores e sua adaptação e reutilização em um problema atual. Assim como no comportamento humano, RBC usa a capacidade de aprender com seus atos, e reutilizar este conhecimento na tomada de novas decisões (Kolodner, 1993).

Segundo (Wangenheim, 2003), RBC pode ser resumido como a solução de novos problemas por meio da utilização de casos anteriores já conhecidos. Um problema novo pode ser resolvido com a reutilização da solução encontrada para um problema anterior similar.

RBC é baseado em memória, o que, segundo (Wangenheim, 2003), torna um sistema simples de usar para construir sistemas computacionais inteligentes, podendo ser usado para a resolução de problemas reais em várias áreas, como diagnósticos médicos (Silva, 2005) (Deters, 2006), diagnósticos remotos em equipamentos eletrônicos (Plotegher, 2005), e Informática Forense (Mesquita, 2011).

O conceito base de RBC, vem da observação da vida humana. Todos os seres humanos ao longo de suas vidas acabam adquirindo experiência e usando a mesma para resolver mais rapidamente problemas semelhantes. Este conceito é utilizado amplamente. Um exemplo disso é justamente o trabalho de um Perito Criminal, que diante de um caso novo, irá lembrar-se de casos semelhantes, os quais ele já resolveu, e tentar adaptar as soluções usadas nestes casos para resolver o caso novo. Ao se confrontar com um caso, por exemplo, envolvendo pedofilia, mas em um contexto diferente, o Perito irá inicialmente efetuar as buscas que, baseado em sua experiência anterior, foram mais efetivas para encontrar as informações solicitadas nos casos do mesmo tipo, adaptando a solução encontrada, se necessário, às

características do novo caso. De forma análoga, soluções que possam ter gerado falsos positivos, e a consequente necessidade de reanálise, serão evitadas. Esta é uma forma naturalmente eficiente de encontrar uma solução rápida para a resolução de um problema novo. Da mesma forma, o Perito agora tem uma nova experiência, que será utilizada na resolução de casos futuros.

Há quatro elementos básicos no RBC, segundo (Wangenheim, 2003):

- Representação do Conhecimento;
- Medida de Similaridade;
- Adaptação; e
- Aprendizado.

2.3.1. Representação do Conhecimento

Uma das características principais de um Sistema de Raciocínio Baseado em Casos é a representação do conhecimento, a forma na qual irá se registrar os problemas previamente analisados (questionados) e suas soluções encontradas.

Um caso é um pedaço contextualizado de conhecimento representando uma experiência real (Kolodner, 1993).

Segundo (Wangenhein, 2003) um caso é a uma peça de conhecimento contextualizado representando uma experiência ou episódio concretos. Contém a lição passada, que é o conteúdo do caso e o contexto em que a lição pode ser usada. Um caso contém a descrição de um problema ou situação que já foi resolvida e a descrição da solução encontrada, a solução, neste caso, é a experiência adquirida com este caso. Casos podem ter muitas formas e tamanhos, inclusive ter outros casos como atributos.

Os casos em um sistema RBC são mantidos em um Repositório de Conhecimento (Base de casos).

O Repositório de Conhecimento é a base de dados que engloba todos os casos, o vocabulário usado para gerar estes casos, as medidas de similaridade referentes a cada caso, e os dados de como adaptar os casos similares para resolver o novo problema.

Um exemplo de vocabulário utilizado são os quesitos encaminhados ao Perito, contidos na solicitação de perícia para um determinado equipamento. Em toda solicitação de perícia há um questionamento, esteja ele implícito na solicitação ou explícito na forma de quesitos a serem respondidos. O que gera um caso a ser atendido e orienta a análise. Os quesitos, neste caso, seriam o vocabulário.

2.3.2. Cálculo de Similaridade

(Wangenhein, 2003) reforça que o objetivo do Raciocínio Baseado em Casos é a reutilização de soluções conhecidas no contexto de um problema novo, de solução ainda desconhecida. Com base nisso, definir o quão similar o caso novo é a um dos casos arquivados, torna-se crucial para o bom funcionamento do sistema.

A medida de similaridade é o ponto principal de um sistema de RBC, através dela será organizada a base de dados utilizada pelo sistema, através desta medida serão encontrados os casos com maior chance de gerarem solução para o novo caso (caso questionado). Similaridade em RBC nada mais é do que a comparação de dois casos para verificar o quanto um é similar ao outro e como os mesmos podem compartilhar soluções.

Conforme (Wangenhein, 2003), o conceito da utilidade de casos é, de um ponto de vista abstrato, central para o raciocínio baseado em casos. Durante a recuperação de casos procura-se por um problema na base de casos, que, no contexto da descrição do problema atual, é útil para determinar a sua solução.

Uma das hipóteses básicas de um sistema RBC é que problemas similares possuem soluções similares (Kolodner, 1993) (Wangenhein, 2003). Com base nesta afirmação (Wangenhein, 2003) afirma que um caso é útil se ele é similar à questão atual.

Para que seja possível a comparação entre dois casos, é importante definir quais atributos serão usados para esta comparação. (Wangenhein, 2003) nomeia estes atributos como Entidades de Informação (EIs).

A comparação utilizada para encontrar um caso parecido na tentativa da obtenção da resolução do novo problema, pode ser feita de duas formas que são: similaridade sintática ou semântica (Wangenhein, 2003).

Segundo (Wangenhein, 2003), uma medida de similaridade frequentemente utilizada é a técnica *nearest neighbour* (vizinho-mais-próximo) apresentada na Equação 2.1, que é, em sua opinião, muito mais simples que a Distância Euclidiana (distância real entre dois pontos eu um espaço de quaisquer dimenções) apresentada na Equação 2.2, mesmo considerando inúmeros parâmetros (EIs). *Nearest neighbour*, também possui uma versão ponderada, onde se pode considerar a importância de cada índice.

(EQ. 2.1)

(EQ. 2.2)

As Equações 2.1 e 2.2 apresentam os cálculos de medida de similaridade ponderada *nearest neighbour* e baseada em Euclides, onde "*A*" e "*B*" são os pontos a serem comparados e "*w*" o peso atribuído a cada parâmetro da comparação.

A Distância de Hamming é definida como o número de bits divergentes em dois vetores e de mesmo tamanho (Wangenhein, 2003).

Segundo Mesquita, (2011) apud (Liao, T. W.; Zang, Z.; Mount, C. R., 1998), um sistema RBC normalmente usa a Distância Euclidiana ou Distância Hamming como medida de similaridade.

Existem atributos binários assimétricos, como, por exemplo, "Caso possui quesito relacionado à pedofilia" é um atributo assimétrico, pois (Caso1 possui quesito relacionado à pedofilia) = 1 e (Caso2 possui quesito relacionado à pedofilia) = 1 implicam numa similaridade entre os casos (os casos são similares), enquanto (Caso1 possui quesito relacionado a pedofilia) = 0 e (Caso2 possui quesito relacionado a pedofilia) = 0 de forma alguma implica nestes casos tratarem do mesmo crime.

(Mesquita, 2011), considerando sobre as fórmulas simétricas de Distância de Hamming e Distância Euclidiana, conclui ser mais útil a fórmula de Jaccard, apresentada na Equação 2.3, para cálculo de similaridade por desconsiderar as similaridades quando uma característica não se apresenta, tratando-se de uma medição assimétrica para informações binárias.

(EQ. 2.3)

A Equação 2.3 apresenta o cálculo de medida de similaridade de Jaccard, onde "*A*" e "*B*" são *arrays* binários a serem comparados.

Quanto à influência de não haver em uma solicitação determinado quesito, a situação mostra-se muito próxima da encontrada por (Mesquita, 2011), já que o sistema não impede de encontrar evidências não solicitadas, e é comum a presença do quesito "Outros dados julgados úteis.", ou similar, nas solicitações de autoridades policiais ou judiciárias.

2.3.3. Adaptação

A reutilização de um conhecimento anterior muitas vezes necessita de adaptação da solução encontrada para o caso anterior ao caso atual, algumas técnicas são tratadas na reutilização tentam resolver os problemas gerados pela adaptação, que conforme (Wangenhein, 2003), são:

- Quais aspectos da situação devem ser adaptados;
- Quais modificações devem ser realizadas para esta adaptação;
- Que método aplicar para realizar a adaptação; e
- Como controlar este processo.

Uma adaptação é realizada para adequar uma solução encontrada anteriormente a um novo problema similar, quando o caso recuperado da base de casos não satisfazer completamente os requisitos dados pela nova situação (Wangenhein, 2003).

Todas as técnicas de adaptação automática de casos baseiam-se em dois aspectos:

- As diferenças entre o caso passado (cuja solução é conhecida), e o caso atual (cuja solução ainda é desconhecida);
- Qual parte do caso passado pode ser reaproveitada para gerar na solução para o novo caso.

Segundo (Wangenhein, 2003), na maioria das circunstâncias e domínios de aplicação de RBC, geralmente é suficiente que se copie a solução do caso similar encontrado para o caso atual

(questionado) e a aplique, ou então que se adapte o caso manualmente. Esta forma de adaptação é chamada de "Adaptação Nula", e é adequada quando, independente da complexidade da tarefa de classificação, a solução (resposta) é simples.

Considerando as características deste trabalho, sendo as soluções recuperadas, palavras-chave para buscas textuais, a abordagem de não realizar qualquer adaptação automatizada parece razoável, uma vez que, quando da necessidade de alguma adaptação manual às palavras-chave recuperadas, esta se tornará parte do sistema após a retenção do novo caso, evitando o retrabalho.

2.3.4. Revisão

Após a localização do caso mais similar na base de casos e sua solução proposta ser aplicada, adaptada ou não, pode-se concluir que esta não é satisfatória, ou ainda, que é a incorreta.

Segundo (Wangenhein, 2003), após uma avaliação criteriosa da solução gerada, pode-se seguir dois caminhos:

- Se a solução for satisfatória, ou correta, esta será Retida pelo sistema como um caso novo aprendido (será acrescentada a base de casos);
- Se a solução se mostrar insatisfatória, ou errada, deverá ser corrigida, os motivos da falha investigados, possivelmente propondo uma nova solução.

O segundo caso pode requerer uma intervenção manual (operador do sistema), pois pode-se estar diante de uma exceção que deve ser adaptada manualmente antes da aprendizagem pelo sistema RBC.

No sistema proposto não está previsto meios para revisão antecipada, no entanto revisão ocorrerá quando da análise compreensiva do objeto questionado, onde será possível detectar eventuais inconsistências nos resultados apontados na triagem com os dados resultantes da análise final. Neste ponto, o Perito pode sugerir novas palavras-chave e corrigir as já existentes (que serão acrescentadas como novas versões), reaplicando o teste de triagem. Caso, após os ajustes efetuados pelo Perito, a triagem mostre-se consistente com o resultado do Laudo Pericial Criminal, o caso será retido.

2.3.5. Retenção de Novos Casos

O modelo mais comum para representar o processo RBC é o "Ciclo RBC" proposto por Aamondt e Plaza, conhecido como 4R: *Retrieve, Reuse, Revise, Retain* (Wangenheim, 2003 apud [AP94]), como se pode ver na Figura 2.4.

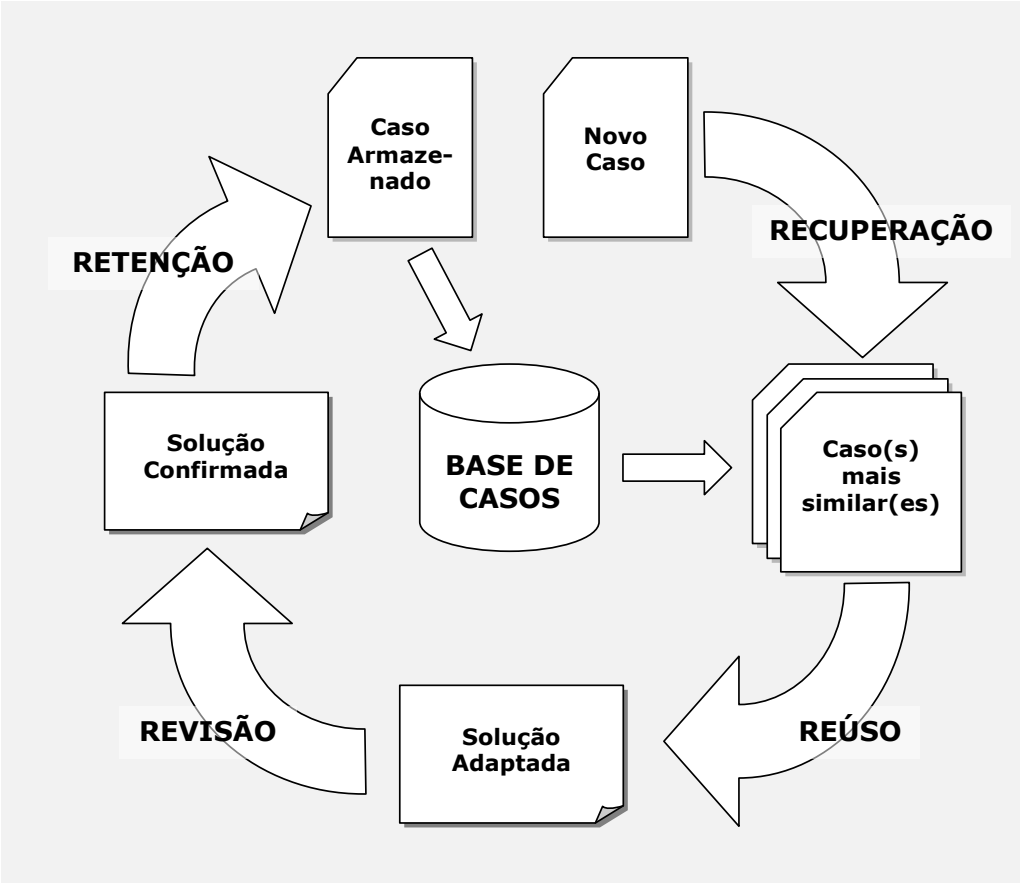

Figura 2.4 – Ciclo RBC adaptado de (Wangenheim, 2003 apud [AP94])

Neste ciclo podemos perceber que, quanto maior o número de casos que tiverem que ser solucionados, com maior facilidade estes serão resolvidos, pois a base de casos será maior e a obtenção de novas soluções será mais eficiente. Um exemplo disso é justamente as buscas por palavras-chave em informática forense, onde a frequência de uso de certas expressões está intimamente ligada à época do delito questionado, por exemplo: Quando se busca evidências da existência de arquivos compartilhados na Internet, relacionados à pedofilia, é importante saber quando o suposto fato ocorrera, pois, devido aos filtros e bloqueios usados por provedores, empresas, ou mesmo domésticos, as expressões usadas nestes arquivos são frequentemente alteradas, para que possam burlar tais filtros, mas que ainda assim sejam localizados facilmente pelos interessados. Desta forma casos contemporâneos tem uma maior

chance de serem similares entre si, logo, quanto maior a Base de Casos, maiores as chances de um caso similar ser encontrado.

(Wangenheim, 2003) destaca que a capacidade de aprendizagem é uma das funções mais importantes em um sistema de Raciocínio Baseado em Casos.

A integração com o aprendizado é uma das características mais importantes do RBC, na opinião de (Wangenheim, 2003). A retensão de conhecimento no RBC implica num paradigma de aprendizado de máquina que suporta o aprendizado sustentado pela atualização contínua da memória de casos.

Sistemas de retenção para RBC podem ser classificados em três tipos (Wangenheim, 2003):

- **Sem retenção de casos:** Neste modelo, não há retenção de conhecimento novo, a base de casos é pré-carregada durante a implementação do sistema e não sofre alterações com o tempo;
- **Retenção de soluções de problemas:** É a forma mais comum, retendo novos casos assim que são resolvidos;
- **Retenção de documentos:** Neste caso, o aprendizado se dá de forma assíncrona, ou seja, novos casos não são retidos automaticamente, o conhecimento é adquirido de forma independente, por outro processo, assim que novos conhecimentos se tornam disponíveis.

É possível reter somente os casos com soluções satisfatórias, no entanto quando um caso selecionado se revela inadequado, isso leva a necessidade de ser reindexado de forma a não ser selecionado em situações idênticas. De forma análoga, quando um caso similar resulta em uma solução ideal, seus índices devem ser alterados para que reflitam este fato.

Segundo (Kolodner, 1992), um sistema que relembre suas falhas evita sugerir a mesma solução incorreta em um caso semelhante. O sistema aprende com seus erros.

No sistema proposto o aprendizado será de forma assíncrona, apesar de propor a retenção de todos os casos questionados, com o objetivo de melhorar a precisão do sistema, este aprendizado só ocorrerá após a análise completa do caso, e se necessário, ajustes nas palavras-chave propostas. Esta escolha é devida ao fato de a proposta deste trabalho se basear em

análise por amostragem para fins de triagem, ou seja, a resposta final, com a certeza do resultado, só se dará ao final da análise pericial tradicional.

3. ANÁLISE POR AMOSTRAGEM

Este capítulo apresentará a técnica de análise por amostragem utilizada, a escolha das amostras e o uso de técnicas de Raciocínio Baseado em Casos para a recuperação da tabela de chaves de busca relevantes.

3.1. TAMANHO DA AMOSTRA

Define-se o elemento populacional como grupo de oito setores consecutivos[3] de uma unidade de discos rígidos. Embora o "setor" do disco seja a menor parte endereçável, uma amostragem que selecione apenas um setor em meio a um *cluster*, não necessariamente no seu início, tem uma maior chance de trazer dados truncados. Ainda pesa o fato de ser mais fácil encontrar expressões compostas de vários bytes (*strings*) integrais, em agrupamentos de dados maiores.

Para fins de experimentos, foram adotados os tamanhos de amostra ,

e onde é o número total de *clusters* do dispositivo questionado, considerando os dispositivos usados neste experimento, que possuem o tamanho de setor comumente encontrado em unidades de discos rígidos (HDDs), ou seja, 512 bytes, e *cluster* de oito setores (4kib , ou seja, 4096 bytes). Exemplifica-se na Tabela 3.1 as amostragens sugeridas em 16 capacidades de HDDs comerciais.

Inicialmente propõe-se o tamanho amostral (Schiavone, 2009) devido seu uso juridicamente aceito no Brasil e no Exterior na análise preliminar de narcóticos, e estendido a outras áreas da criminalística.

Outra possibilidade de tamanho de amostra a ser considerada é o uso de um percentual fixo da população, como o sugerido por (Garfinkel, 2010), mesmo que, em seu exemplo, o autor tenha considerado o volume de dados possível de ser copiado em 1 minuto, ele considera como altamente relevante o volume de 4,8 GB, que, no caso, representa 0,48% do dispositivo analisado.

Para fins de comparação, será incluído nos experimentos também o tamanho amostral fixo de 1%, para que se possa relacionar os dados de desempenho e precisão dos resultados.

[3] Por ser o tamanho comum usado em *clustes* pela maior parte dos sistemas de arquivos.

Tabela 3.1 – Amostragens em tamanhos de HDDs comerciais

Capacidade Nominal	Capacidade Real[4][5]	Clusters:	Amostra: ‾	Volume de dados da Amostra[6]	Amostra: 0,48%	Volume de dados da Amostra[6]	Amostra: 1%	Volume de dados da Amostra[6]
10 GB	9,31 GiB	2.441.406	1.563	6,11 MiB	11.719	45,78 MiB	24.415	95,37 MiB
20 GB	18,63 GiB	4.882.813	2.210	8,63 MiB	23.438	91,55 MiB	48.829	190,74 MiB
40 GB	37,25 GiB	9.765.625	3.125	12,21 MiB	46.875	183,11 MiB	97.657	381,47 MiB
80 GB	74,51 GiB	19.531.250	4.420	17,27 MiB	93.750	366,21 MiB	195.313	762,94 MiB
120 GB	111,76 GiB	29.296.875	5.413	21,14 MiB	140.625	549,32 MiB	292.969	1.144,41 MiB
160 GB	149,01 GiB	39.062.500	6.250	24,41 MiB	187.500	732,42 MiB	390.625	1.525,88 MiB
200 GB	186,26 GiB	48.828.125	6.988	27,30 MiB	234.375	915,53 MiB	488.282	1.907,35 MiB
240 GB	223,52 GiB	58.593.750	7.655	29,90 MiB	281.250	1.098,63 MiB	585.938	2.288,82 MiB
320 GB	298,02 GiB	78.125.000	8.839	34,53 MiB	375.000	1.464,84 MiB	781.250	3.051,76 MiB
400 GB	372,53 GiB	97.656.250	9.883	38,61 MiB	468.750	1.831,05 MiB	976.563	3.814,70 MiB
500 GB	465,66 GiB	122.070.313	11.049	43,16 MiB	585.938	2.288,82 MiB	1.220.704	4.768,38 MiB
640 GB	596,05 GiB	156.250.000	12.500	48,83 MiB	750.000	2.929,69 MiB	1.562.500	6.103,52 MiB
750 GB	698,49 GiB	183.105.469	13.532	52,86 MiB	878.907	3.433,23 MiB	1.831.055	7.152,56 MiB
1 TB	931,32 GiB	244.140.625	15.625	61,04 MiB	1.171.875	4.577,64 MiB	2.441.407	9.536,75 MiB
2 TB	1.862,65 GiB	488.281.250	22.098	86,32 MiB	2.343.750	9.155,27 MiB	4.882.813	19.073,49 MiB
3 TB	2.793,97 GiB	732.421.875	27.064	105,72 MiB	3.515.625	13.732,91 MiB	7.324.219	28.610,23 MiB

Embora os demais tamanhos amostrais se apresentem de forma linear, a amostra inicialmente proposta por (Schiavone, 2009) – ‾ – possui a característica de ser menos suscetível ao aumento exponencial da capacidade de armazenamento das unidades atuais, conforme demonstrado na Figura 3.1.

Com relação ao desempenho esperado, o tempo decorrido para o acesso à amostra exemplificada na Tabela 3.1 seria, em tese, proporcional ao tamanho da amostra, chegando a ser desprezível (menor que 1 segundo), em qualquer um dos casos analisados para o tamanho amostral ‾, se considerarmos somente o tempo de acesso aos dados na velocidade nominal do dispositivo. Entretanto, devido à necessidade do reposicionamento do braço de leitura para o novo cilindro do disco, a cada amostra (chamado de "*seek time*"[7], com tempo médio

[4] A capacidade real possui pequenas variações de acordo com o modelo e o fabricante do dispositivo.

[5] Como forma de representar a capacidade real, foi utilizada a notação GibiByte (contração do inglês giga binary byte) que é uma unidade medida para armazenamento eletrônico de informação, estabelecida pela Comissão Eletrotécnica Internacional (IEC) para designar 2^{30} bytes de informação ou de armazenamento computacional.

[6] Capacidade expressa em MebiByte (MiB) usada para designar 2^{20} bytes de informação ou de armazenamento computacional.

[7] Chama-se de tempo de busca ou "*seek time*", o tempo de reposicionamento da cabeça de leitura, o qual varia, na maioria dos dispositivos mecânicos atuais, entre 0,2 ms e 0,8 ms para movimentos Track-to-Track (tempo que a cabeça de leitura demora para mudar de uma trilha – cilindro – para a imediatamente seguinte), podendo chegar à 20 ms o tempo de deslocamento do primeiro ao último cilindro. Em média, para mover de um cilindro aleatório para outro, considera-se um tempo de 10 ms (average).

considerado de 10 ms), e a latência[8] do disco (tipicamente 4,15 ms para um dispositivo de 7200 RPM), para cada movimento, considerando dispositivos mecânicos (Morimoto, 2007), e o fato que este acesso será aleatório e necessitará de movimentação constante – ter-se-ia um acréscimo teórico de aproximadamente 14 ms para cada amostra lida, dependendo das características do dispositivo.

O acréscimo de tempo devido ao acesso aleatório ao disco, possivelmente não foi considerado na proposta de (Garfinkel, 2010) em suas primeiras análises, pois ele apresenta o conteúdo de 4,8 GB coletado aleatoriamente de um dispositivo de 1TB de capacidade em 1 minuto (81,92 MB/s) sendo que em sua proposta são amostrados setores individuais de um disco.

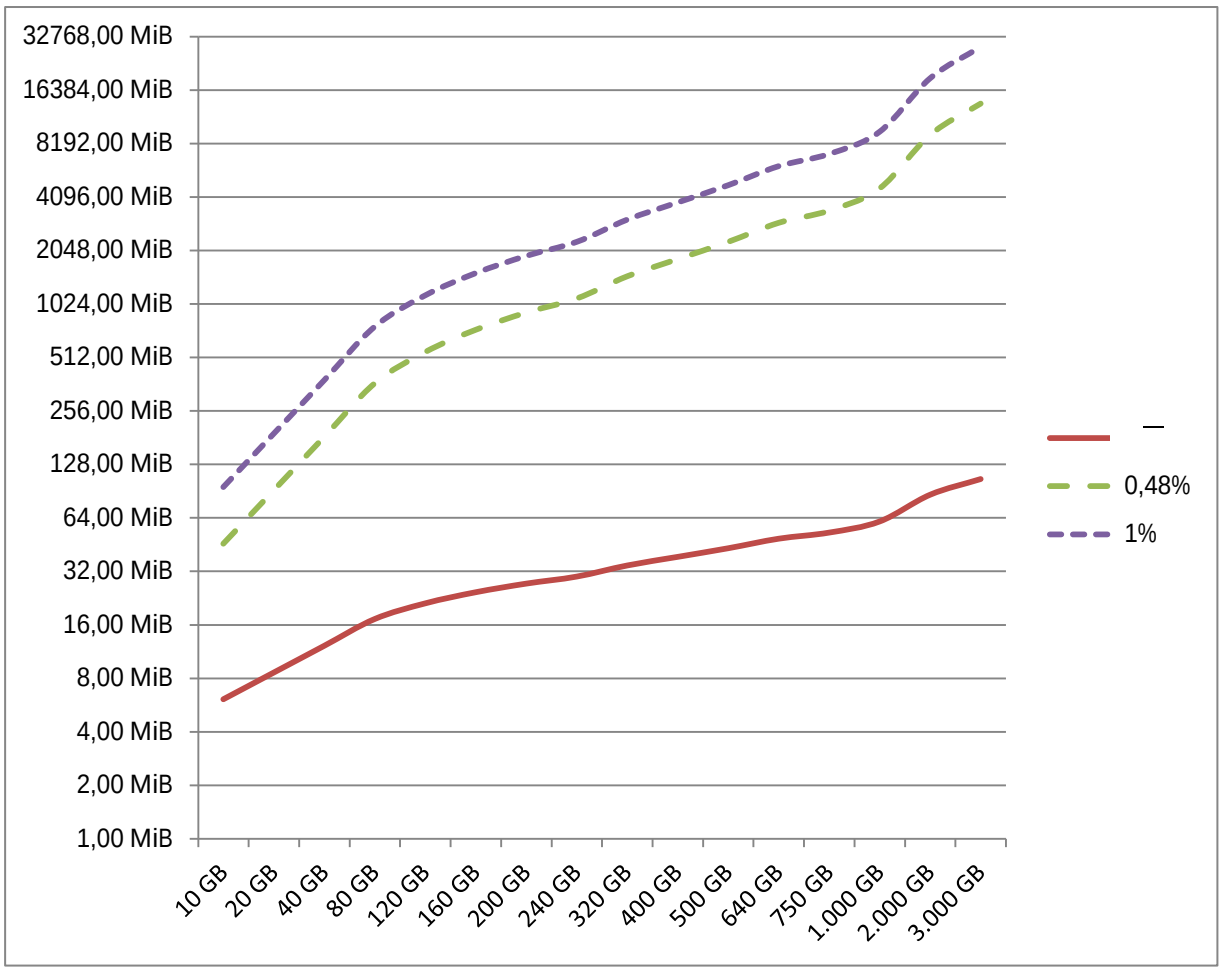

Figura 3.1 – Tamanho da amostra.

Dados os obstáculos teóricos de um acesso aleatório a um dispositivo mecânico, conclui-se que a escolha adequada da amostra é um dos pontos de maior importância para o sucesso da análise proposta.

[8] Tempo médio para um determinado setor do disco estar sob a cabeça de leitura, com esta já posicionada em um determinado cilindro do disco.

3.2. MÉTODO DE AMOSTRAGEM

Quanto ao método de amostragem, deve-se considerar que no caso da Informática Forense sempre haverá a possibilidade de um adversário com interesse na manipulação dos resultados, seja para inocentar um suspeito, ou mesmo incriminar outro.

Qualquer método determinístico, neste caso, pode ser explorado como uma vulnerabilidade do sistema, dando ao adversário a possibilidade de manipular a fila de prioridades, por meio do armazenamento de dados relevantes em áreas com baixa probabilidade de serem amostradas, ou mesmo inserindo dados manipulados em áreas com probabilidade alta.

Considerando também a possibilidade de vícios de seleção, por parte do sistema ou mesmo do Perito, amostragens intencionais devem ser evitadas.

O método escolhido para amostragem foi a escolha aleatória de *clusters* da unidade de armazenamento como um todo, independente de particionamento, sistema de arquivos, ou sistema operacional em uso. Evitando, desta forma, a possibilidade de vícios de seleção (Morrettin, 2010), e diminuindo a possibilidade de um adversário mal intencionado direcionar os resultados obtidos (Garfinkel, 2010).

Para fins de experimento, os *clusters* selecionados serão copiados para um arquivo em separado, e este será submetido à busca pelas palavras-chave escolhidas usando a ferramenta EnCase.

3.3. ESCOLHA DAS CHAVES (RBC)

Este capítulo apresentará o uso de técnicas de Raciocínio Baseado em Casos para a recuperação da tabela de chaves de busca relevantes de casos similares, já analisados compreensivamente e com resultados conclusivos.

Como atributo principal a ser considerado para determinação da similaridade entre os casos, será utilizado, para fins de experimento, a classificação dos tipos de Perícias em Informática do sistema PGP (Protocolo-Geral de Perícias) do Instituto-Geral de Perícias[9], apresentado na Figura 3.2.

[9] Instituto-Geral de Perícias, da Secretaria de Segurança Pública, do Estado do Rio Grande do Sul (IGP/SSP/RS).

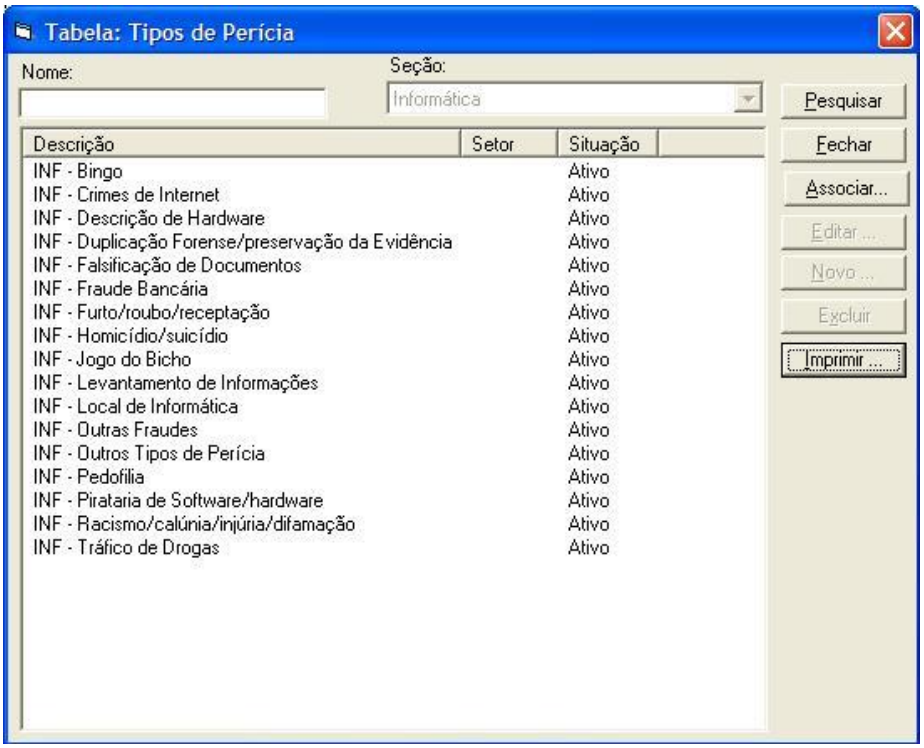

Figura 3.2 – Tipos de Perícia de Informática

Diversos tipos de perícias, relacionados a crimes ou outras contravenções, classificados na Figura 3.2, podem ser usados nos experimentos, pois possuem expressões muito particulares e não usadas comumente em outras áreas, como, por exemplo, “Pedofilia”, “Trafico de Drogas” e “Fraude Bancária”. Foi selecionado o tipo de crime “Pedofilia” para a realização dos experimentos deste trabalho.

Outro atributo considerado importante é a contemporaneidade do delito investigado, uma vez que, como já comentado anteriormente, as expressões mais relevantes em alguns tipos de crimes tendem a sofrer alterações com o tempo, de forma a burlar certos tipos de bloqueios, fiscalizações ou investigações automatizadas, como o caso de filtros de navegação para conteúdo adulto na Internet.

Como forma de identificar casos mais específicos, e melhorar a precisão da ferramenta, um ponto importante da solicitação a ser inserido no sistema são os quesitos ou objetivo da análise solicitada (que traz implícitos os quesitos na descrição da solicitação).

Os quesitos, seja explícitos ou implícitos na solicitação, são solicitados de forma textual, muitas vezes com particularidades na forma de escrita, mas com o mesmo objetivo final. A

seguir são exemplificados[10] alguns quesitos comumente encontrados em solicitações relacionadas ao tipo de crime Pedofilia (a relação completa encontra-se no ANEXO B).

- "Os arquivos contém elementos que possam indicar envolvimento ou sejam compatíveis com crime de "PEDOFILIA", como fotos de adolescentes ou crianças nuas?";
- "É possível revelar o período (dia e hora) em que estes possíveis arquivos foram baixados para o computador referido?";
- "É possível revelar se houve troca de fotos e mensagens entre o usuário do computador e terceiros? Se estas fotos e mensagens trocadas são de menores ou adolescentes nuas ou assuntos referentes ao crime de "PEDOFILIA"?";
- "É possível identificar quem recebeu ou enviou material de cunho pornográfico usando a máquina periciada? Se possível, quem o fez? Quem recebeu?";
- Existem, neste material, qualquer arquivo, bem como jogos, endereço eletrônico ou site com indícios ou vestígios de imagens com pornografia ou cenas de sexo ?";
- "Se houve acesso a sites de pornografia envolvendo crianças ou adolescentes?";
- "Existem, neste material cenas de sexo explícito ou pornográfica utilizando criança ou adolescente?";
- "O material encaminhado contém fotografias ou vídeos com imagens de pornografia infanto-juvenil? Caso positivo, quais as datas de criação dos arquivos?".

Nota-se que vários quesitos acima solicitam a mesma informação, e poderiam ser classificados como um mesmo tipo de quesito abstrato, como, por exemplo, "[Imagens pedofilia]". Essa abstração é fundamental para o funcionamento do sistema RBC, pois possibilitará a verificação de similaridade entre casos, solicitados por diferentes instituições ou autoridades.

Foi construída uma lista de quesitos abstratos, obtida através da análise de diferentes solicitações relacionadas á pornografia envolvendo crianças e/ou adolescentes atendidas no ano de 2011 no Instituto-Geral de Perícias/RS. Esta lista é mostrada na Tabela 3.2.

[10] Quesitos copiados integralmente de solicitações reais, mantendo a grafia original. Foram suprimidos dados identificadores que eventualmente estejam protegidos por segredo de justiça.

Tabela 3.2 – Abstração de Quesitos.

Quesito Abstrato	Interpretação do quesito original
[Imagens Pedofilia]	Busca por imagens e/ou vídeos, armazenados no equipamento, com conteúdo relacionado à pornografia infanto-juvenil.
[P2P Pedofilia]	Busca por registros de compartilhamento de arquivos envolvendo pornografia infanto-juvenil.
[Web Pedofilia]	Busca por registros de acessos, criação ou mantenimento de conteúdo on-line relacionado à pornografia infanto-juvenil.
[Troca de mensagens]	Busca por registros de conversas, onde o assunto esteja relacionado à pornografia infanto-juvenil, ou aliciamento de menores.
[Equipamento Vítima]	Quando a solicitação deixa claro que o equipamento analisado era usado pela vítima[11].
[Equipamento Suspeito]	Quando a solicitação deixa claro que o equipamento analisado era usado pelo suspeito[9].

A lista de palavras-chave armazenadas na base de casos para cada caso analisado podem ser qualquer chave de busca usada em pesquisas textuais, desde palavras individuais como "lolita", frases complexas como "underage first time sex", ou mesmo expressões regulares de busca como "[[^1-9][1-9] | [[1][0-7]]] [^A-Z#]?years?.?old".

3.4. REPRESENTAÇÃO DO CASO

A representação dos casos já analisados e armazenados na base de casos se dará pela notação da Equação 3.1

(EQ. 3.1)

Tipo: Enquadrado em um dos itens da lista de tipos previamente definida pelo órgão (Figura 3.2);

Data: Data da ocorrência (fato), apreensão do equipamento, ou último uso registrado;

Quesitos: Lista binária de quesitos padronizados (abstratos) relacionado com o tipo de perícia, onde 0 (zero) representa a ausência do quesito e 1 (um) a presença do quesito, sendo o número de quesitos disponíveis na base de dados para este tipo de perícia.

[11] O equipamento analisado pode ter sido usado pela vítima, suspeito, ou mesmo compartilhado por ambos. Neste último caso, ambos os Quesitos Abstratos devem ser marcados.

3.5. CÁLCULO DE SIMILARIDADE

O cálculo de similaridade entre os casos presentes na base de casos e o caso proposto para análise se dará pela comparação direta da EI *Tipo* do caso questionado com os casos arquivados. Casos que não possuam o mesmo *Tipo* serão considerados **não similares**, sem a necessidade de avaliação de outras EIs.

Após a definição da lista de casos de mesmo *Tipo*, serão considerados para análise de quesitos os casos contemporâneos, ou seja, datados (EI *Data*) em uma faixa de tempo próxima ao caso questionado.

Com base na análise de Laudos Periciais, relacionados à Pedofilia, dos anos de 2010 e 2011, estabeleceu-se que, inicialmente, o período de 12 meses é adequado para considerar um caso contemporâneo a outro.

(Le Grand, 2009) descreve a dinâmica da popularidade de palavras-chave relacionadas à pedofilia em duas amostras realizadas em um período de 2 anos. Seus dados mostram que, mesmo mantendo a mesma representatividade com relação ao total de buscas, a popularidade (total de ocorrências) das expressões buscadas varia em média[12] 46,18%

Propõe-se que, a cada 12 meses de distância, seja decrementado um percentual fixo de 23%, refletindo desta forma na similaridade dos casos a tendência de mutação das expressões relacionadas aos delitos investigados.

Por fim o *array* binário resultante da lista de Quesitos Abstratos (EI *Quesitos*) será confrontado com os *Quesitos* dos casos armazenados, através da fórmula de Jaccard, obtendo assim o índice de similaridade entre os casos, possibilitando a ordenação da Base de Casos e o uso do caso mais similar, conforme exemplo a seguir.

[12] Considerando a variação do número de ocorrências de cada palavra-chave encontrada com pelo menos 1% de representatividade no total de palavras-chave.

Passo 1: Tratam-se do mesmo *Tipo* (Pedofilia).

Passo 2: Ocorreram dentro de um período de 7 meses. São contemporâneos, logo não haverá decréscimo em sua similaridade.

Passo 3: Segundo Jaccard:

(EQ. 3.2)

Nota-se claramente que Jaccard desconsiderou os três últimos *Quesitos*, o que é importante resaltar, pois o fato de não ter sido fornecida a informação sobre quem era o usuário do equipamento (vítima, suspeito ou ambos) em ambos os casos não nos permite inferir qualquer similaridade útil entre eles.

Considerando outro caso no Banco de Casos:

Teríamos:

(EQ. 3.3)

Considerando o decréscimo devido à distância de 13 meses da *Data* dos casos:

(EQ. 3.4)

Ainda assim, apesar de não tão contemporâneo, ter-se-ia um caso mais similar que o caso2, o que se justifica pelo fato de no caso3 ter sido realizada a análise de registros de compartilhamento de arquivos, tendo este maior possibilidade de possuir palavras-chave relevantes para o caso1, onde se questiona o mesmo fato.

3.6. RETENÇÃO DE NOVOS CASOS

Neste trabalho optou-se por realizar a realimentação da Base de Casos de forma assíncrona, usando o sistema classificado por (Wangenheim, 2003) como “Retenção de documentos”.

A retenção de novos casos se dará de forma independente e manualmente pelo Perito após a análise final e entrega do Laudo Pericial respondendo as solicitações. No caso de uma análise conclusiva, onde a busca textual tenha sido relevante, um novo caso será inserido no sistema e sua lista de palavras-chave carregada. Quando da pré-existência de um caso idêntico ao caso sendo retido, a lista de palavras-chave será incrementada.

Durante a retenção de um novo caso, o *array* de quesitos pode ser ampliado para refletir novas solicitações. No entanto novas classificações para tipos de perícia devem sempre ser discutidas, pois podem provocar a subdivisão de um *Tipo* pré-existente, o que pode acarretar inconsistências não previstas no sistema.

A retenção assíncrona garante a revisão e validação dos dados obtidos após a análise compreensiva do caso, já que o resultado da aplicação direta do sistema proposto consideraria apenas uma pequena amostra dos dados questionados.

A proposta de solução apresentada neste capítulo é composta de duas etapas: Definição da amostra dos dados e Escolha das palavras-chaves de busca com base em RBC. Os experimentos realizados com base na solução proposta e seus resultados são descritos no Capítulo 4.

4. EXPERIMENTOS E RESULTADOS

Este capítulo apresentará os resultados obtidos com a aplicação da técnica proposta, nos casos de teste, e o confronto destes com os resultados obtidos com a forma tradicional de análise pericial nos mesmos casos, suas peculiaridades, exceções, e ajustes realizados.

Para fins de experimento, foram selecionados dois casos considerados similares, ou seja, classificados como *Tipo* Pedofilia, contemporâneos, e ambos quesitados sobre o armazenamento de imagens com conteúdo pornográfico envolvendo crianças e/ou adolescentes, seu possível compartilhamento e/ou distribuição, e sobre o possível acesso a sítios na Internet que possam conter ou estar relacionado de alguma forma a este tipo de conteúdo.

Ambos os casos foram analisados compreensivamente por dois Peritos Criminais e tiveram respostas conclusivas. O caso1 teve respostas positivas aos quesitos e concluiu que o equipamento questionado foi utilizado em ações ligadas a crimes relacionados á Pedofilia. Já no caso2 não foi encontrado qualquer indício de sua vinculação ao crime questionado.

Os dados originais encontravam-se armazenados em unidades de discos rígidos – HDDs pertencentes a computadores pessoais tipo desktop, com 160GB de capacidade nominal em cada um dos dispositivos. Os dados analisados nos experimentos originaram-se das copias previamente realizadas no momento da análise original, e armazenados em discos individuais para os experimentos.

Para ambos os casos (cujos discos originais possuíam a mesma capacidade) foram calculadas as amostras para os três tamanhos amostrais propostos. A seguir foram geradas duas listas aleatórias para cada tamanho amostral, sendo uma delas posteriormente ordenada. Por fim, cada disco questionado teve seus dados amostrados, sendo esta amostra armazenada em um segundo disco para análise.

4.1. CÁLCULO DE DESEMPENHO

Para o cálculo de desempenho das amostragens, foi utilizado o equipamento descrito no Anexo A, executando o Sistema Operacional Linux (Kernel versão 2.6.32) a partir da distribuição Ubuntu 10.04 LTS 32 bits. Para o procedimento foi utilizado o terminal sobre

interface gráfica com as configurações padrão, usando os comandos “dd” para acesso a mídia e “time” para o registro preciso do tempo de execução.

O resultado de desempenho da amostragem (valor médio entre os casos analisados) para cada tamanho amostral, considerando ainda o acesso aleatório e ordenado à lista de amostras, é apresentado na Tabela 4.1.

Tabela 4.1 – Desempenho da amostragem

	amostra √N	amostra 0,48%	amostra 1%	Total da Mídia
Clusters Acessados	6251	187550	390728	39072726
Volume de dados acessado	24,42 MiB	732,62 MiB	1,49 GiB	149,05 GiB
Tempo acesso aleatório	0min 20.05s	9min 10.44s	15min 53.22s	não se aplica
Velocidade média ac. aleatório	1,22 MiB/s	1,33 MiB/s	1,60 MiB/s	não se aplica
Tempo acesso ordenado	0min 19.31s	7min 53.93s	13min 7.83s	65min 32.16
Velocidade média ac. Ordenado	1,26 MiB/s	1,55 MiB/s	1,94 MiB/s	38,82 MiB/s
Ganho do desempenho	3,83%	16,15%	20,99%	não se aplica

A Tabela 4.1 mostra um desempenho superior quando o acesso se dá de forma ordenada aos grupos de dados amostrados. Neste caso, não havendo a necessidade do retorno do braço de leitura, o que reduz a distância percorrida e as perdas devido ao “*seek-time*”.

Notou-se na análise destes resultados um ganho menor que o esperado em termos de desempenho para as amostras menores, com relação ao tipo de acesso (aleatório/ordenado), e o aumento gradual deste nas amostras maiores. Isto se explica devido à tecnologia empregada nos HDDs utilizados para os experimentos (descritos no Anexo A). Os dispositivos testados possuem uma tecnologia chamada NCQ (*Native Command Queuing* – Comando Nativo de Enfileiramento) que possui a capacidade de rearranjar a fila de requisições de acesso, de forma a otimizá-la para um melhor desempenho através da ordenação dos acessos, visando reduzir a movimentação desnecessária da cabeça de leitura. Este recurso aliado à capacidade de *buffer* de 32 MB do dispositivo resultou na completa reordenação da primeira amostragem, reduzindo o impacto do acesso a uma lista desordenada. Esta característica reduziu as perdas de desempenho em todas as amostragens com acesso aleatório.

Comparando o tempo necessário para acesso a todos os dados do disco com o tempo necessário para acesso as amostras de seu conteúdo, nota-se o evidente ganho de performance com o aumento da amostra. Isto se explica pelo fato do acesso a um volume maior de dados

resultar em amostras fisicamente mais próximas, sendo necessária uma movimentação linear menor da cabeça de leitura, o que resulta em uma velocidade média maior para o acesso, conforme Figura 4.1.

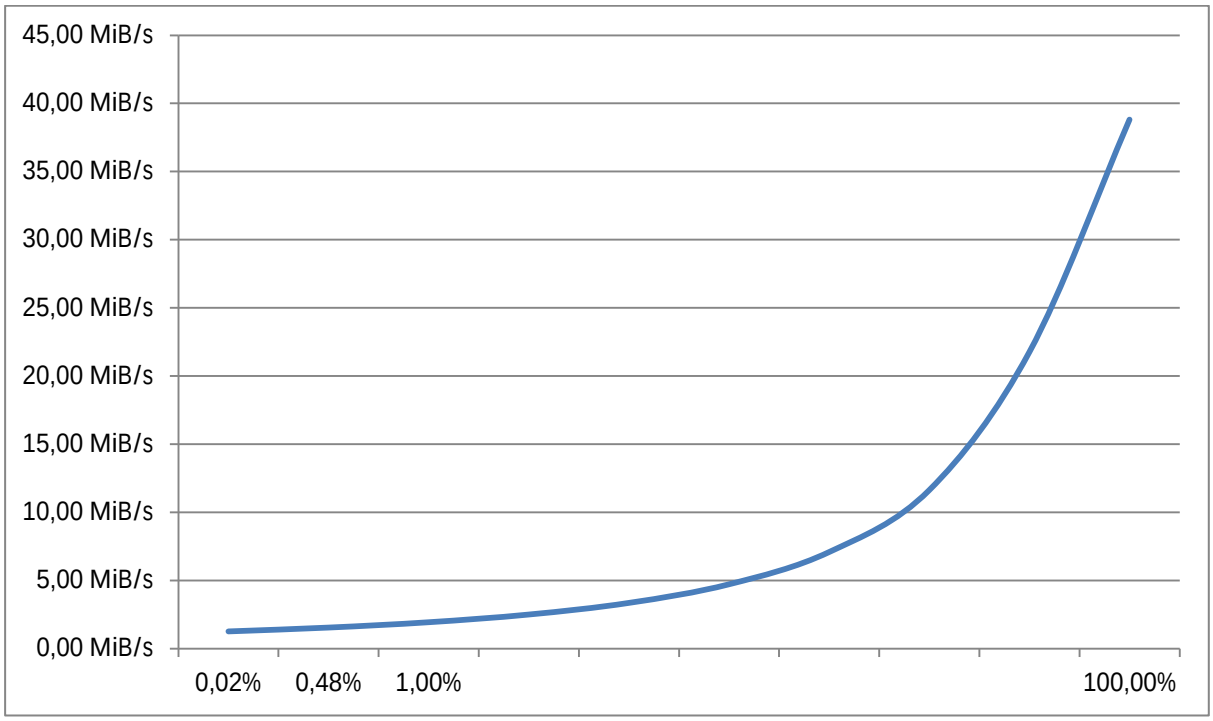

Figura 4.1 – Desempenho da Amostragem com relação à velocidade de acesso.

Ainda considerando o tempo necessário para o acesso aos dados, nota-se que este impõe uma limitação ao tamanho da amostra, que, em termos de desempenho, não traria vantagens significativas em amostragens superiores a 1%, como mostrado na Figura 4.2.

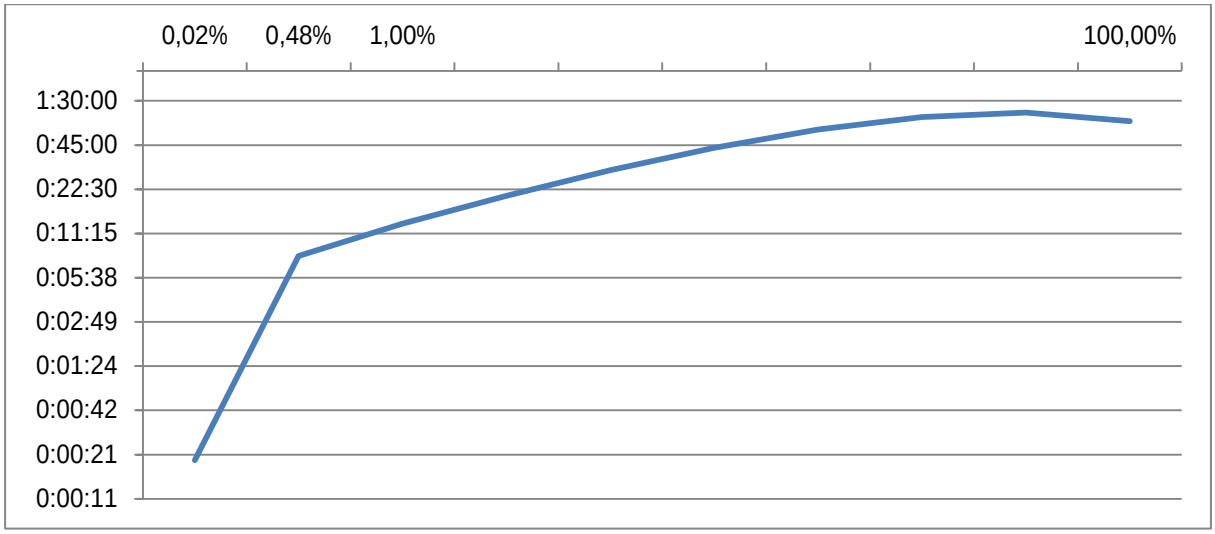

Figura 4.2 – Desempenho da Amostragem com relação ao tempo.

4.2. RESULTADOS OBTIDOS

Para obtenção dos resultados das buscas por palavras-chave, foi considerada, inicialmente, a lista de expressões usadas para orientar as buscas nos Laudos Periciais referentes a cada caso. Trata-se de expressões relacionadas à pornografia infanto-juvenil, frequentemente encontradas em casos envolvendo Pedofilia.

As palavras-chave (*keywords*) foram inseridas no Software Encase (Figura 4.3), e a seguir foi realizada a busca em cada um dos casos do experimento.

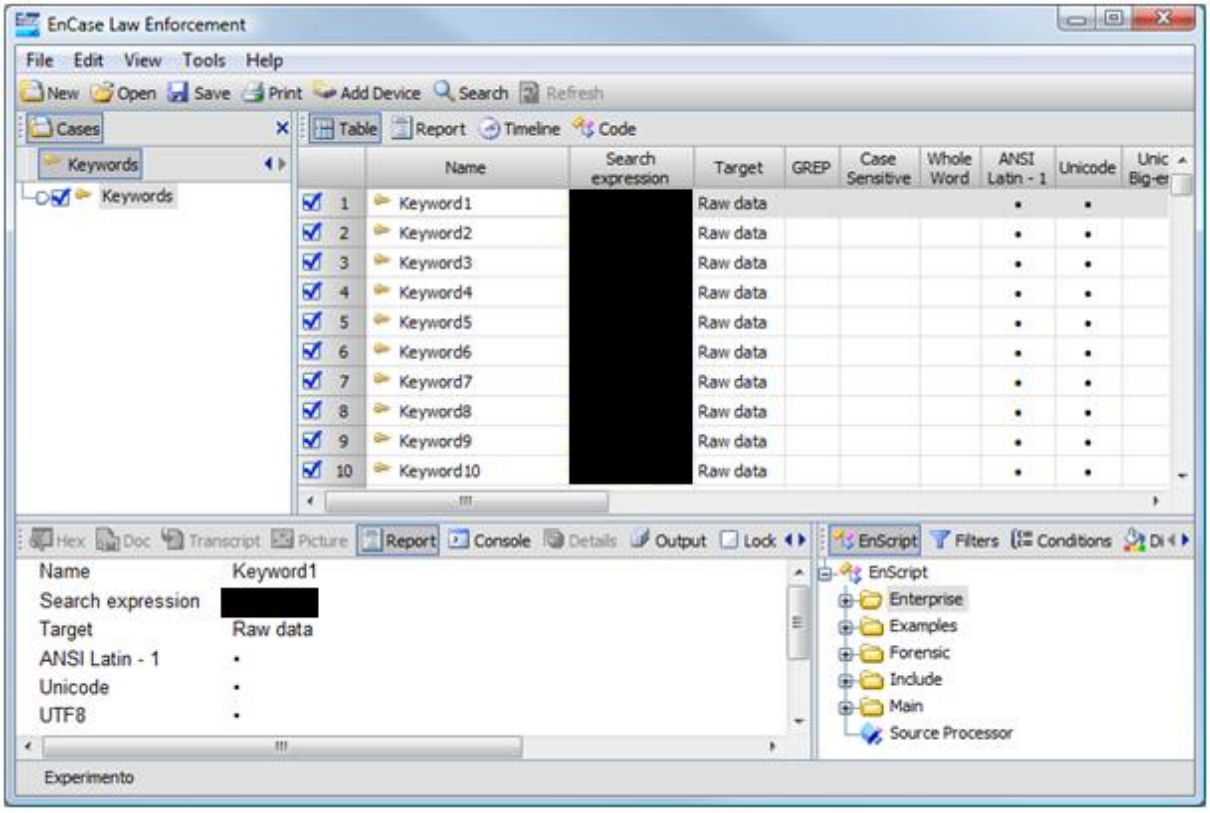

Figura 4.3 – Aplicativo Pericial Encase

Os resultados das buscas foram compilados, sendo os mais significativos, ou seja, com maior número de ocorrências, para o caso1, listados conforme Tabela 4.2.

Tabela 4.2 – Resultado preliminar das buscas nos dados no caso1

PALAVRA-CHAVE	**TOTAL DE HITS**
Keyword1	12.773
Keyword2	7.866
Keyword3	2.816
Keyword4	2.168
Keyword5	1.410

PALAVRA-CHAVE	**TOTAL DE HITS**
Keyword6	88
Keyword7	1.083
Keyword8	348
Keyword9	318
Keyword11	213
Keyword12	96
Keyword13	72
Keyword14	64
Keyword17	55
Keyword25	6.122
Keyword26	488
Keyword27	122
TOTAL	**36.102**

Como forma de reduzir a incidência de falsos-positivos, a mesma lista de expressões foi submetida ao caso2, conforme mostrado na Tabela 4.3.

Tabela 4.3 – Resultado preliminar das buscas nos dados no caso2

PALAVRA-CHAVE	**TOTAL DE HITS**
Keyword1	16.876
Keyword2	2.674
Keyword3	227
Keyword4	77
Keyword5	268
Keyword6	216
Keyword7	42
Keyword8	4
Keyword9	2
Keyword11	10
Keyword12	40
Keyword13	18
Keyword14	916
Keyword17	0
Keyword25	1.662
Keyword26	11
Keyword27	464
TOTAL	**23.507**

Após a análise das duas tabelas, notou-se que as *keywords* de números 1, 14, 25 e 27 faziam referência a temas pornográficos, mas não necessariamente com relação à Pedofilia, e os *keywords* 2 e 6 integravam outras expressões não relacionadas à pornografia. Removidas as palavras-chave com alta incidência de falsos-positivos, foi definida a Tabela 4.4para o caso1 da base de casos:

Tabela 4.4 – Tabela de palavras-chave para o caso1 comparada com os resultados do caso2

PALAVRA-CHAVE	**CASO1**	**CASO2**
Keyword3	2.816	227
Keyword4	2.168	77
Keyword5	1.410	268
Keyword7	1.083	42
Keyword8	348	4
Keyword9	318	2
Keyword11	213	10
Keyword12	96	40
Keyword13	72	18
Keyword17	55	0
Keyword26	488	11
TOTAL	**9.067**	**699**

Então o possuiria como solução proposta o conjunto sugerido de chaves-de-busca relevantes {Keyword3, Keyword4, Keyword5, Keyword7, Keyword8, Keyword9, Keyword11, Keyword12, Keyword13, Keyword17, Keyword26}.

As palavras-chave obtidas pela recuperação do caso1, após o cálculo de similaridade, serão então buscadas nos dados resultantes da amostragem de todos os equipamentos enviados para exame. Supondo, por exemplo, que os casos 1 e 2, pertencessem a mesma investigação e fossem submetidos à triagem, o primeiro caso seria priorizado para análise final.

Foram realizadas duas amostragens para cada tamanho amostral sugerido em ambos os casos. Analisando os resultados das buscas nos dados amostrados para os casos 1 e 2 tem-se:

Tabela 4.5 – Resultados das amostragens para o caso1

	√N	**√N'**	0,48%	0,48%'	1%	1%'
Keyword3	0	0	3	8	20	50
Keyword4	0	0	2	4	7	21
Keyword5	0	0	9	14	15	10
Keyword7	0	0	1	3	13	11
Keyword8	0	0	0	1	0	4
Keyword9	0	0	2	3	0	3
Keyword11	0	0	0	0	0	0
Keyword12	0	0	0	3	0	1
Keyword13	0	0	0	0	0	0
Keyword17	0	0	0	0	0	0
Keyword26	0	0	0	12	7	6
Total	0	0	17	48	62	106

Tabela 4.6 – Resultados das amostragens para o caso2

	√N	**√N'**	0,48%	0,48%'	1%	1%'
Keyword3	0	0	0	3	0	3
Keyword4	0	0	0	0	0	3
Keyword5	0	0	0	0	2	0
Keyword7	0	0	0	0	0	0
Keyword8	0	0	0	0	0	0
Keyword9	0	0	0	0	0	0
Keyword11	0	0	0	0	0	0
Keyword12	0	0	0	1	2	0
Keyword13	0	0	0	0	0	0
Keyword17	0	0	0	0	0	0
Keyword26	0	0	0	0	0	0
Total	0	0	0	4	4	6

As tabelas Tabela 4.5 e Tabela 4.6 demonstram a eficácia do procedimento proposto, onde **√N**, 0,48% e 1% mostram o número de ocorrências (*hits*) de cada *keyword* na primeira amostragem para cada caso, e **√N'**, 0,48%**'** e 1%**'** mostram o número de ocorrências na segunda amostragem.

Os dados avaliados para a tomada de decisão quanto à priorização de um caso, será o valor total de ocorrências das expressões buscadas, ou seja, o somatório das ocorrências de cada palavra-chave buscada nos dados amostrados.

Considerando os valores esperados para os resultados das amostras como proporcionais ao volume de dados amostrado, pode-se inferir a quantidade total de ocorrências na mídia questionada, conforme demonstrado nas Tabelas 5.7 e 5.8.

Tabela 4.7 – Comparação com os valores esperados para o caso1

	Total da Mídia	**E√N**	**√N**	E0,48%	0,48%	E1%	1%
Keyword3	2.816	1	0	14	6	28	35
Keyword4	2.168	0	0	10	3	22	14
Keyword5	1.410	0	0	7	12	14	13
Keyword7	1.083	0	0	5	2	11	12
Keyword8	348	0	0	2	1	3	2
Keyword9	318	0	0	2	3	3	2
Keyword11	213	0	0	1	0	2	0
Keyword12	96	0	0	0	2	1	1
Keyword13	72	0	0	0	0	1	0
Keyword17	55	0	0	0	0	1	0
Keyword26	488	0	0	2	6	5	7
Total	9.067	1	0	43	33	91	84

Tabela 4.8 – Comparação com os valores esperados para o caso2

	Total da Mídia	**E√N**	**√N**	E0,48%	0,48%	E1%	1%
Keyword3	227	0	0	1	2	2	2
Keyword4	77	0	0	0	0	1	2
Keyword5	268	0	0	1	0	3	1
Keyword7	42	0	0	0	0	0	0
Keyword8	4	0	0	0	0	0	0
Keyword9	2	0	0	0	0	0	0
Keyword11	10	0	0	0	0	0	0
Keyword12	40	0	0	0	1	0	1
Keyword13	18	0	0	0	0	0	0
Keyword17	0	0	0	0	0	0	0
Keyword26	11	0	0	0	0	0	0
Total	699	0	0	2	2	6	5

As tabelas Tabela 4.7 e Tabela 4.8 mostram a comparação dos resultados esperados das buscas para cada tamanho amostral, com o resultado obtido a partir da análise das amostras do disco, onde **E.√N**, **E.0,48%** e **E.1%** são os valores esperados para cada tamanho de amostra e

¯, **0,48%** e **1%** a média das amostragens realizadas respectivamente para cada tamanho de amostra.

A coerência apresentada entre os valores esperados e os valores obtidos confirma a possibilidade de conclusões baseadas nos dados amostrados, podendo também inferir sobre o resultado esperado durante a análise compreensiva.

Por outro lado, o número total de ocorrências encontradas não permite, para todos os casos, o uso de uma amostra de tamanho ¯ para a obtenção de dados confiáveis.

Os experimentos realizados buscam avaliar a viabilidade da solução proposta através de sua aplicação em casos oriundos de exames periciais reais do Instituto-Geral de Perícias do Estado do Rio Grande do Sul. Os resultados obtidos apresentam tempos de execução viáveis para o processo proposto com grande precisão nas amostras de 0,48% e 1%. As conclusões a respeito do trabalho realizado e as propostas de melhorias e trabalhos futuros serão apresentadas Capítulo 5.

5. CONCLUSÕES

Este capítulo apresenta uma análise dos resultados obtidos, suas contribuições, conclusões finais do trabalho, e uma proposta para trabalhos futuros sobre o tema.

Considerando a proposta da triagem de casos, com o objetivo de determinar o equipamento com maior possibilidade de conter os dados questionados, dentre um grupo de equipamentos apreendidos em uma mesma investigação, a análise proposta mostrou-se adequada e viável para tamanhos amostrais entre 0,48% e 1% do total de dados disponíveis.

É possível o uso dos resultados obtidos a partir da análise dos dados amostrados, como subsídio para determinação de qual equipamento possui maior possibilidade de conter os dados questionados.

Os resultados deste trabalho ainda apontam que, apesar do acesso a um volume de dados reduzido, o tempo necessário para a análise por amostragem não é proporcional ao volume de dados acessados, devido ao tempo de busca necessário aos dados selecionados nos dispositivos mecânicos analisados (HDDs).

Outro dado importante quanto aos resultados, é o fato da análise por amostragem só ter utilidade prática em busca de conteúdos muito representativos no objeto questionado, ou seja, que possam conter um número significante de ocorrências de uma mesma expressão textual relevante, e não encontrada facilmente em outros dispositivos não relacionados ao fato buscado. Crimes envolvendo Pedofilia enquadram-se neste quesito, conforme observado nos experimentos.

Este trabalho avança o estado da arte na área relacionada a buscas textuais, acrescentando a possibilidade de decisão baseada em análises em volumes menores de dados, o que permite uma análise preliminar mais rápida e eficiente, uma vez que pode priorizar ou descartar grandes volumes de dados sem sua análise compreensiva.

5.1. TRABALHOS FUTUROS

Entre as diversas possibilidades de trabalhos futuros e oportunidades para aperfeiçoar a solução proposta, destacam-se:

- A realização dos experimentos em unidades SSD (*solid-state drive* – unidade de estado sólido), e outros dispositivos não mecânicos, já que estes possuem tempos de

busca muito reduzidos, quando comparados a discos mecânicos;

- A realização dos experimentos em outros tipos de perícia que também contenham expressões muito particulares e não usadas comumente em outras áreas, como Fraudes Bancárias e Tráfico de Drogas;
- A implementação de uma Base de Casos inicial do sistema RBC, junto a um órgão Pericial, para coleta progressiva de dados para implementação futura da solução proposta ou outras que possam vir a usar o mesmo conhecimento armazenado.

REFERÊNCIAS BIBLIOGRÁFICAS

AccessData Group. Products. Forensics. FTK. Disponível em: <*http://accessdata.com/products/computer-forensics/ftk*> . Acesso em: 01 julho 2011.

Baron, Jonathan (2000), Thinking and deciding (3rd ed.), New York: Cambridge University Press, 2000, pag. 195.

Bowman, A. W., Gibson, I., Scott, E. M., & Crawford, E. (2010). Interactive Teaching Tools for Spatial Sampling. Journal Of Statistical Software, 36(13), 1-17. Retrieved from *http://www.stats.gla.ac.uk/~adrian/rpanel/spatial-sampling-paper.pdf*.

Beebe, N. L. & Clark, J. G. (2005). A hierarchical, objectives-based framework for the digital investigations process. Digital Investigation, Junho 2005, Volume 2, Issue 2, 2005.

Beebe, N., & Clark, J. (2007). Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. Digital Investigation, 4 (Supplement 1), 2007.

Le Grand, Bénédicte et al. (2009). Dynamics of Paedophile Keywords in eDonkey Queries. Technical report. Retrieved from *http://antipaedo.lip6.fr/T24/TR/kw-dynamics.pdf*.

BRASIL. **Constituição** (1988). Constituição da República Federativa do Brasil:

BRASIL. **CPP** (1941). Decreto-Lei n. 3.689, de 3 de outubro de 1941. Código Processual Penal.

BRASIL. **ECA** (1990). Lei n. 8.069, de 13 de julho de 1990. Dispõe sobre o Estatuto da Criança e do Adolescente e dá outras providências.

Bunting, Steve & Wei, Willian (2006). The Official EnCE: EnCase Certified Examiner. Indianapolis, IN. Wiley Publishing, Inc, 2006. 527 p.

Carmo, Antonio Rosemir do (2010). Refletindo as Mudanças Ocorridas na Sociedade nos Últimos Cinco Anos. Artigonal, 08 agosto 2010. Disponível em: <*http://www.artigonal.com/educacao-artigos/refletindo-as-mudancas-ocorridas-na-sociedade-nos-ultimos-cinco-anos-2991298.html*>. Acesso em: 29 junho 2011.

Carrier, Brian (2005). File System Forensic Analysis. Crawfordsville, In. Pearson, 2005, 569 pp.

Colli, Maciel (2010) Cibercrimes: Limites e Perspectivas à Investigação Policial de Crimes Cibernéticos. Curitiba: Juruá, 2010, 208 pp.

Deters, J. I. et al. (2006). Desenvolvimento de um Sistema de Raciocínio Baseado em Casos na Identificação de Transtornos Mentais. In *II Congresso Sul Catarinense de Computação - SulComp 2006, 2006*, Criciúma, SC. Anais do II Congresso Sul Catarinense de Computação, 2006.

Eleutério, P. & Machado, M. (2011). Desvendando a Computação Forense. Editora Novatec, 2011, 200 pp.

Federal Bureau of Investigation (**FBI**). About Us. Information Technology. Disponível em: <*http://www.fbi.gov/about-us/itb*>. Acesso em: 30 junho 2011.

Garfinkel, S. & Shelat, A. (2003). Remembrance of Data Passed: A Study of Disk Sanitization Practices, In *IEEE Security and Privacy*, Janeiro/Fevereiro 2003.

Garfinkel, S. L. (2009). Automating Disk Forensic Processing with SleuthKit, XML and Python. 2009 *Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, 73-84. Ieee. doi: 10.1109/SADFE.2009.12.

Garfinkel, Simson L. (2010). Automated Digital Forensics. In *Talk, Harvard School of Engineering and Applied Sciences*, Cambridge, MA, USA, em 18 outubro 2010.

Guedes, Igor Rafael de Matos Teixeira (2009). A pedofilia no âmbito da Internet. Monografia (Graduação em Direito) - Faculdades Integradas Pitágoras, Montes Claros, MG, 2009.

Guidance Software, Inc. EnCase Forensic Version 6 User Manual. Passadena, CA, 2006.

Guidance Software, Inc. (2011) Encase Forensic. Disponível em: <*http://www.guidancesoftware.com/forensic.htm*>. Acesso em: 01 julho 2011.

Hoelz, Bruno 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.

Kolodner, J. L. (1992). An introduction to case-based reasoning. Artificial Intelligence Review, 6(1), 3-34. Springer Netherlands, Volume 6, Issue 1, 1993.

Kolodner, J. L. (1993). Case-Based Reasoning. Morgan Kaufmann Pub., Inc. 1993.

Lange, Rodrigo (2010). Crimes cibernéticos: Novos desafios do Direito. Monografia (Graduação em Direito) - Faculdades Integradas do Brasil - UniBrasil, Curitiba, PR, 2010.

Lucy, David (2005). Introduction to Statistics for Forensic Scientists. University of Edinburgh, UK. John Wiley & Sons, Ltd, 2005. 251p.

Menegazzo, Cinara T. (2001). Raciocínio baseado em casos aplicado a diversos domínios de problema. Dissertação (Mestrado em Computação) - Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, 2001.

Mesquita, F., Hoelz, B. & Ralha, C. (2011). Raciocínio Baseado em Casos Aplicado em Análise Live, In *Proceeding of the Sixth International Conference on Forensic Computer Science – ICoFCS 2011*, Florianópolis, SC. ABEAT (ed.), 2011, 210 pp.

Ministério da Justiça do Brasil (**MJ**). Segurança Pública. Disponível em: <*http://www.portal.mj.gov.br*>. Acesso em: 30 junho 2011.

Morettin, Luiz Gonzaga (2006). Estatística Básica. São Paulo, SP. Pearson. 7.ed., 2006.

Morettin, Luiz Gonzaga (2010). Estatística Básica: Probabilidade e Inferência. São Paulo, SP. Pearson, 2010, 375 pp.

Morimoto, Carlos E. (2007). Hardware II, o Guia Definitivo. GDH Press e Sul Editores, 2007, 1088 pp.

National Institute of Standards and Technology (**NIST**). Computer Forensics Portal. Disponível em: <*http://www.nist.gov/computer-forensics-portal.cfm*>. Acesso em: 30 junho 2011.

Neto, Pedro L. C. (1977). Estatística. Ed. Blücher Ltda, 1977, 264 pp.

Plotegher, S. L. & Fernandes, M. M. (2005).Raciocínio Baseado em Casos Aplicado a um Sistema de Diagnóstico Remoto. In: *XXXII Seminário Integrado de Software e Hardware (SEMISH), XXV Congresso da SBC*, 2005, São Leopoldo, RS. Anais do SBC - XXXII SEMISH, 2005.

Reith, M., Carr, C., & Gunsch, G. (2002). An Examination of Digital Forensic Models. In *International Journal of Digital Evidence*, Fall 2002, Volume 1, Issue 3, 2002.

Redivo, Rafaella & Monteiro, Gabriela Loosli (2007). O direito frente à era da informática, Presidente Prudente: Faculdades Integradas Antônio Eufrásio de Toledo, 2007.

Roussev, V., & Garfinkel, S. L. (2009). File Fragment Classification-The Case for Specialized Approaches. 2009 *Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, 3-14. Ieee. doi: 10.1109/SADFE.2009.21.

Schiavone, S. et al. (2009). Guidelines on Representative Drug Sampling (ONU, 2009, ST/NAR/38). New York, NY. United Nations Publication, 2009. 54 p.

Silva, R. P. et al. (2005). Proposta de um Modelo RBC para Construção de um Sistema de Apoio ao Diagnóstico Médico, In *IV Simpósio Brasileiro de Qualidade de Software - V Workshop de Informática Médica, 2005, Porto Alegre, RS.* Anais do IV SBQS. Porto Alegre, RS: PUCRS, 2005.

StorageReview.com's Drive Performance Resource Center (2011). Disponível em: <*http://www.storagereview.com/comparison.html*>. Acesso em 01 julho 2011.

Wald, Abraham (1947). Sequential analysis. Oxford, England. John Wiley, 1947, 212pp.

Wald, Abraham. (1943). A Method of Estimating Plane Vulnerability Based on Damage of Survivors. Statistical Research Group, Columbia University. Center for Naval Analyses, 1943.

Wangenheim, C. & Wangenheim, A. (2003). Raciocínio Baseado em Casos. Editora Manole, 2003, 300 pp.

ANEXOS

A – EQUIPAMENTO UTILIZADO NOS EXPERIMENTOS

O equipamento utilizado para realização dos experimentos deste trabalho foi uma das estações periciais da Seção de Informática Forense do Departamento de Criminalística do Instituto-Geral de Perícias / SSP / RS, descrito a seguir:

- Tipo: Computador Pessoal.
- Placa de Sistema: ASUS P5Q-E.
- CPU: Intel Core 2 Quad Q9550 à 2,83GHz.
- Memória RAM: Tipo DDR2, capacidade de 8GB, configurada em duplo canal à 400MHz com latência 4-4-4-12.
- Disco de Sistema: Seagate ST3250318AS (Windows).
- Disco Auxiliar: Samsung HM160HC (Linux).
- Sistema Operacional: Microsoft Windows Vista Ultimate 64 Bits SP2.
- Sistema Operacional Auxiliar: Ubuntu 10.04 LTS 32 bits.
- Aplicativo usado para os experimentos: Guidance Software Encase, versão 6.18.1.3 Law Enforcement 64 bits.
- Demais componentes de interconexão, alimentação, interface, auxiliares ou não utilizados para os experimentos.

Ao equipamento citado, foram acrescentadas duas unidades de discos rígidos – **HDDs marca Samsung, modelo HD753LJ**, novas e devidamente testadas para a condução dos experimentos, de forma a não haver interferências com os casos armazenados decorrentes do uso normal da estação (as unidades de armazenamento de casos originais em análise foram temporariamente removidas durante os experimentos).

B – EXEMPLOS DE QUESITOS

Exemplos de quesitos encontrados em solicitações relacionadas ao tipo de crime Pedofilia, copiados integralmente de solicitações reais, mantendo a grafia original.

Foram suprimidos dados identificadores que eventualmente estejam protegidos por segredo de justiça.

- "Houve acesso a sites de pornografia infanto-juvenil? Caso positivo, qual o endereço eletrônico e datas de acesso?";
- "Pelos padrões dos arquivos é possível identificar o equipamento para sua geração (marca ou modelo da máquina fotográfica ou filmadora)?";
- "Caso houver conteúdo pornográfico envolvendo crianças e adolescentes no material, é possível observar-se alguma característica antropométrica (sexo, idade aparente, sinal) que permita uma posterior identificação das vítimas?";
- "Através do material apreendido foram enviadas, recebidas ou compartilhadas mensagens eletrônicas com conteúdo de pornografia infanto-juvenil? Caso positivo, qual(is) o(s) e-mails(s) e datas de envio?";
- "No material apreendido foram utilizadas fotografias ou vídeos de conteúdo pornográfico infanto-juvenil para publicação em redes de relacionamento? Caso positivo, qual o endereço eletrônico, perfil do usuário e demais dados utilizados no(s) site(s)?";
- "Teve algum arquivo de fotos e/ou vídeos contendo cenas relacionadas à pedofilia enviado por mensagens ou, de algum outro modo, repassado a terceiros através do equipamento em análise?";
- "Há alguma fotografia ou vídeo no equipamento contendo cenas de pornografia e/ou pedofilia?";
- "Os arquivos contém elementos que possam indicar envolvimento ou sejam compatíveis com crime de "PEDOFILIA", como fotos de adolescentes ou crianças nuas?";
- "É possível revelar o período (dia e hora) em que estes possíveis arquivos foram baixados para o computador referido?";

- “É possível revelar se houve troca de fotos e mensagens entre o usuário do computador e terceiros? Se estas fotos e mensagens trocadas são de menores ou adolescentes nuas ou assuntos referentes ao crime de “PEDOFILIA”?”;
- “É possível identificar quem recebeu ou enviou material de cunho pornográfico usando a máquina periciada? Se possível, quem o fez? Quem recebeu?”;
- “Se houve acesso à internet ou outros meios com conteúdo pornográfico relacionado a crianças e adolescentes ou adultos.”;
- “Se foram encontradas imagens ou vídeos pornográficos envolvendo crianças e adolescentes ou adultos.”;
- “Outros dados julgados úteis.”.