ALONSO DECARLI

RECUPERAÇÃO DE INFORMAÇÕES EM DOCUMENTOS TEXTUAIS: APLICAÇÃO EM LAUDOS PERICIAIS DE DISPOSITIVOS MÓVEIS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

CURITIBA

2015

ALONSO DECARLI

RECUPERAÇÃO DE INFORMAÇÕES EM DOCUMENTOS TEXTUAIS: APLICAÇÃO EM LAUDOS PERICIAIS DE DISPOSITIVOS MÓVEIS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de Concentração: *Ciência da Computação*

Linha de Pesquisa: *Processamento de Imagens, Visão Computacional e Computação Forense*

Orientador: Dra. Cinthia Obladen de Almendra Freitas
Co-orientador: Dr. Emerson Cabrera Paraíso

CURITIBA
2015

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

D291r 2015

Decarli, Alonso
Recuperação de informações em documentos textuais: aplicação em laudos periciais de dispositivos móveis / Alonso Decarli; orientador, Cinthia Obladen de Almendra Freitas; co-orientador, Emerson Cabrera Paraíso. -- 2015
85 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2015
Bibliografia: f. 77-83

1. Informática. 2.Telefone celular. 3. Recuperação da informação. 4. Evidencia. 5. Crime por computador – Investigação. I. Freitas, Cinthia Obladen de Almeida. II. Paraiso, Emerson Cabrera. III. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. III. Título.

CDD 20. ed. – 004.068

ATA

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

PUCPR
GRUPO MARISTA

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

DEFESA DE DISSERTAÇÃO DE MESTRADO Nº 05/2015

Aos 25 dias do mês de Setembro de 2015 realizou-se a sessão pública de Defesa da Dissertação **" Recuperação de Informações em Documentos Textuais: Aplicação em Laudos Periciais de Dispositivos Móveis"** apresentado pelo aluno **Alonso Decarli,** como requisito parcial para a obtenção do título de Mestre em Informática, perante uma Banca Examinadora composta pelos seguintes membros:

Prof.ª Dr.ª Cinthia Obladen de Almendra Freitas
PUCPR (Orientadora) _____ APROVADO
(assinatura) (Aprov/Reprov)

Prof. Dr. Emerson Cabrera Paraiso
PUCPR _____ Aprov.
(assinatura) (Aprov/Reprov)

Prof.ª Dr.ª Andreia Malucelli
PUCPR _____ APROVADO
(assinatura) (Aprov/Reprov)

Prof. Dr. Celso Antonio Alves Kaestner
UTFPR _____ APROVADO
(assinatura) (Aprov/Reprov)

Conforme as normas regimentais do PPGIa e da PUCPR, o trabalho apresentado foi considerado APROVADO (aprovado/reprovado), segundo avaliação da maioria dos membros desta Banca Examinadora. Este resultado está condicionado ao cumprimento integral das solicitações da Banca Examinadora registradas no Livro de Defesas do programa.

Prof.ª Dr.ª Andreia Malucelli.
Coordenadora do Programa de Pós-Graduação em Informática.

Programa de Pós-Graduação em Informática PPGIa PUCPR

Rua Imaculada Conceição, 1155 Prado Velho CEP 80215 901 Curitiba Paraná Brasil
Tel.: (41) 3271 1669 Fax.: (41) 3271 2121 www.ppgia.pucpr.br

Dedicatória

Dedico essa dissertação a todos os meus familiares,
meus pais **Alberto Decarli** e **Maria Carmen Decarli**,
meus irmãos **Tatiana Decarli** e **Pablo Roberto Decarli** e seus respectivos cônjuges.

Agradecimentos

Primeiramente, à minha orientadora, aquela que me aceitou e apresentou à pesquisa. Com seu gesto, a Profª. Cinthia proporcionou uma das melhores oportunidades de aprimoramento e desenvolvimento de conhecimentos técnico-científicos que tive em minha vida até o momento. Seu exemplo de vida, seriedade, objetividade, presteza e disponibilidade foram essenciais no desenvolvimento deste trabalho.

Ao Diretor Geral do IC-PR, Hemerson Bertassoni Alves, assim como todos os peritos da Sessão de Computação Forense do Instituo de Criminalística do Paraná em especial ao Luiz Rodrigo Grochocki e Alexandre Vrubel. A participação destes foi fundamental em todas as etapas do desenvolvimento do presente trabalho, desde sua concepção até a realização dos experimentos em uma base de dados real.

Aos funcionários dos setores administrativos e professores que compõem o corpo docente da PUC-PR e da UTFPR, um agradecimento especial pelo apoio e considerações feitas sobre o trabalho apresentado. Registro a valorosa participação dos professores Emerson Cabrera Paraíso, Andreia Malucelli e Celso Antonio Alves Kaestner.

Agradeço ainda os colegas de curso que além da companhia diária, a amizade e a troca de experiências, foi possível vivenciar gratos momentos de fraternidade e solidariedade.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro cedido a um considerável período deste projeto.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ANATEL	*Agencia Nacional de Telecomunicações*
CODIS	*Combined DNA Index System*
GPS	*Global Positioning System*
HTML	*Hypertext Markup Language*
ICCID	*International Circuit Card ID*
IMEI	*International Mobile Equipment Identity*
IMSI	*International Mobile Subscriber Identity*
MD5	*Message-Digest Algorithm 5*
NIST	*National Institute of Standards and Technology*
ODT	*Open Document Format for Text*
PDAs	*Personal Digital Assistants*
RAM	*Random Access Memory*
RI	*Recuperação de Informação*
ROM	*Read Only Memory*
SGIP	*Sistema de Gerenciamento de Informações Periciais*
SHA	*Secure Hash Algorithm*
SI	*Sistema de Informação*
SiCReT	*Sistema de Cruzamento de Registros Telefônicos*
SisBala	*Sistema de Indexação Balística*
TI	*Tecnologia da Informação*
UFED	*Universal Forensic Extraction Device*
W3C	*World Wide Web Consortium*
XML	*eXtensible Markup Language*

Resumo

Os Sistemas de Recuperação de Informações possuem o objetivo de fazer com que o usuário encontre a informação que está precisando rapidamente, de modo que este usuário não necessite analisar todas as informações existentes na base de informações. Mesmo com a existência de uma considerável quantidade de sistemas de recuperação de informações, um sistema de recuperação de informações nunca irá atender a todas as necessidades de todos os usuários. O presente trabalho apresenta uma ferramenta computacional com o propósito de atender uma lacuna que está em aberto nessa área nas Sessões de Computação Forense dos Institutos de Criminalística. Com o intuito de fornecer meios de processamento e otimização das atividades realizadas por peritos, a Ciência da Informação possibilita unir em uma só estrutura os conceitos de Recuperação de Informações e Descoberta do Conhecimento e aplicar esses conceitos nas informações contidas em laudos periciais de dispositivos móveis. O trabalho propõem um método para Cruzamento de Registros Telefônicos a partir de dados extraídos de laudos periciais de dispositivos móveis, o SiCReT (Sistema de Cruzamento de Registros Telefônicos), o qual foi testado e validado com experimentos realizados em 200 arquivos de laudos periciais de dispositivos móveis da base de dados real do Instituto de Criminalística do Paraná. A visualização dos resultados, especialmente em formato de grafos, permite que os usuários possam analisar rapidamente grandes quantidades de informações detectando e visualizando os cruzamentos de informações de interesse entre laudos distintos. Finalmente, o trabalho fornece uma ferramenta de apoio aos Serviços de Inteligência e Policiamento Preditivo, evitando a subjetividade no exercício da atividade pericial e proporcionando a produção de provas e constatação de evidências forenses que até então estavam ocultas em documentos dispersos nas Sessões de Informática Forense.

Palavras-Chave: Telefones Celulares, Evidência Digital, Análise Forense em Telefones Celulares, Recuperação de Informações.

Abstract

The Information Retrieval Systems have the objective of helping the user find the desired information quickly, in a way that this same user doesn't need to review every information in a given database. Even with the existence of a sizable amount of Information Retrieval Systems, such systems cannot respond to the needs of every users. This dissertation presents a computacional tool, with the objective of filling this gap, which is wide open in this area, the Computer Forensics Session in the Institutes of Criminology nationwide. With the intent of providing ways of processing and optimization of the activities done by forensic experts, the Information Science allows the merging, in one greater structure, the concepts of Information Retrieval and Knowlegde Discovery and the application of these concepts in the information located in the expert reports from mobile devices. The dissertation proposes a method for the Mobile Phones Registry Crossing out of data from forensic reports from mobile phones, SiCReT (Telephonic Registry Crossing System), which has been tested and validated with experiments done in two hundred expert reports of mobile phones from the Institute of Criminology the state of Paraná. The visualization of the results, especially in graphs, allow the users to quickly review great quantities of information, detecting and visualizating the crossing of the desired information, from different reports. Finally, this dissertation offers a tool with the intent of aiding the Intelligence and Predictive Policing Services, avoing the subjectivity in the application of forensics activity and allowing the production of proof and the findings of forensic evidences, till then hidden in dispersed documents in the Computer Forensics Sessions.

Keywords: Mobile Phones, Digital Evidence, Mobile Phones Forensics, Information Retrieval.

Capítulo 1

Introdução

O volume de dados em formato digital gerados e armazenados cresceu exponencialmente com a evolução das tecnologias de informação aplicadas nas mais diversas áreas e organizações.

Em 2012, Gantz e Reinsel fizeram uma publicação a respeito do crescimento do volume de informações no planeta, apontando que no período de 2003 a 2010 o volume de dados digitais armazenados passou de 5 hexabytes para 988 hexabytes. Fazem ainda uma previsão de que até 2020 a quantidade de informações digitais criadas e replicadas chegará a aproximadamente 35 bilhões de hexabytes (GANTZ; REINSEL, 2012).

Aproximadamente 80% dos dados encontram-se armazenados em arquivos não estruturados, parte significativa destes dados encontra-se no formato de texto. Os arquivos textuais quando estruturados tornam-se fator de grande interesse para as organizações, possibilitando a agilidade nos processos de busca e de recuperação de informações (KUECHLER, 2007).

A transformação desse grande volume de dados textuais não estruturados em informação útil fornece elementos para a reorganização, avaliação, utilização, compartilhamento e armazenamento do conhecimento gerado a partir de dados extraídos destes (REZENDE et al., 2011).

Este mesmo cenário é também observado na Computação Forense, onde laudos de perícia criminal e evidências forenses são armazenadas em documentos não estruturados, geralmente contendo anexos de diversas formas e formatos de arquivo.

O presente Capítulo também apresenta a motivação, os objetivos, a hipótese do trabalho, bem como a sua contribuição e a organização deste documento.

1.1 Motivação

O Brasil demonstra um crescente aumento no uso de dispositivos móveis. Na publicação do mês de julho de 2015 a ANATEL (*Agencia Nacional de Telecomunicações*) relatou que o país alcançou a marca de 281,45 milhões de linhas ativas na telefonia móvel, apresentando dessa forma, uma teledensidade de 137,65 acessos por 100 habitantes[1] (ANATEL, 2014).

De acordo com a IDC Brasil, as vendas de celulares no ano de 2014 ultrapassaram a marca de 15 milhões de unidades, com crescimento de 49% na comparação com o mesmo período do ano de 2013[2].

O tratamento e uso da informação pela sociedade têm se modificado nas últimas décadas como consequência do surgimento de novos modelos sociais, econômicos ou tecnológicos. Estes modelos promoveram uma mudança de paradigma tão importante quanto à invenção da imprensa, ou ainda, quanto à própria revolução industrial. A crescente utilização de meios de comunicação com alto grau de mobilidade e o uso cada vez maior da Internet, definem outros espaços e demarcam novas fronteiras para a sociedade contemporânea (RIBEIRO, 2008).

A Computação Forense permeia as mais diversas áreas de perícia, seja trabalhando em conjunto, seja preparando a evidência para exames em outras áreas, o reflexo deste crescimento foi sentido com um crescente aumento na demanda por exames envolvendo dispositivos móveis.

O trabalho de Silva (SILVA, 2011) mostrou o volume de exames relacionados a Computação Forense no Estado do Paraná por meio do SGIP (*Sistema de Gerenciamento de Informações Periciais*), sendo possível dimensionar precisamente a crescente demanda por perícias na área de computação forense.

A explosão do crescimento do volume de informações gerados no planeta apresenta um forte reflexo na computação forense, onde laudos de perícia criminal e evidências forenses são armazenadas em documentos não estruturados, geralmente contendo anexos de diversas formas e formatos de arquivo.

---

[1] Disponível em <http://www.anatel.gov.br/institucional/index.php?option=com_content&view=article&id=621:julho-de-2015-fecha-com-281-45-milhoes-de-acessos-moveis&catid=104&Itemid=354> Acesso em 05 set. 2015.

[2] Disponível em <http://www.idcbrasil.com.br/releases/news.aspx?id=1777> Acesso em 03 de set. 2015.

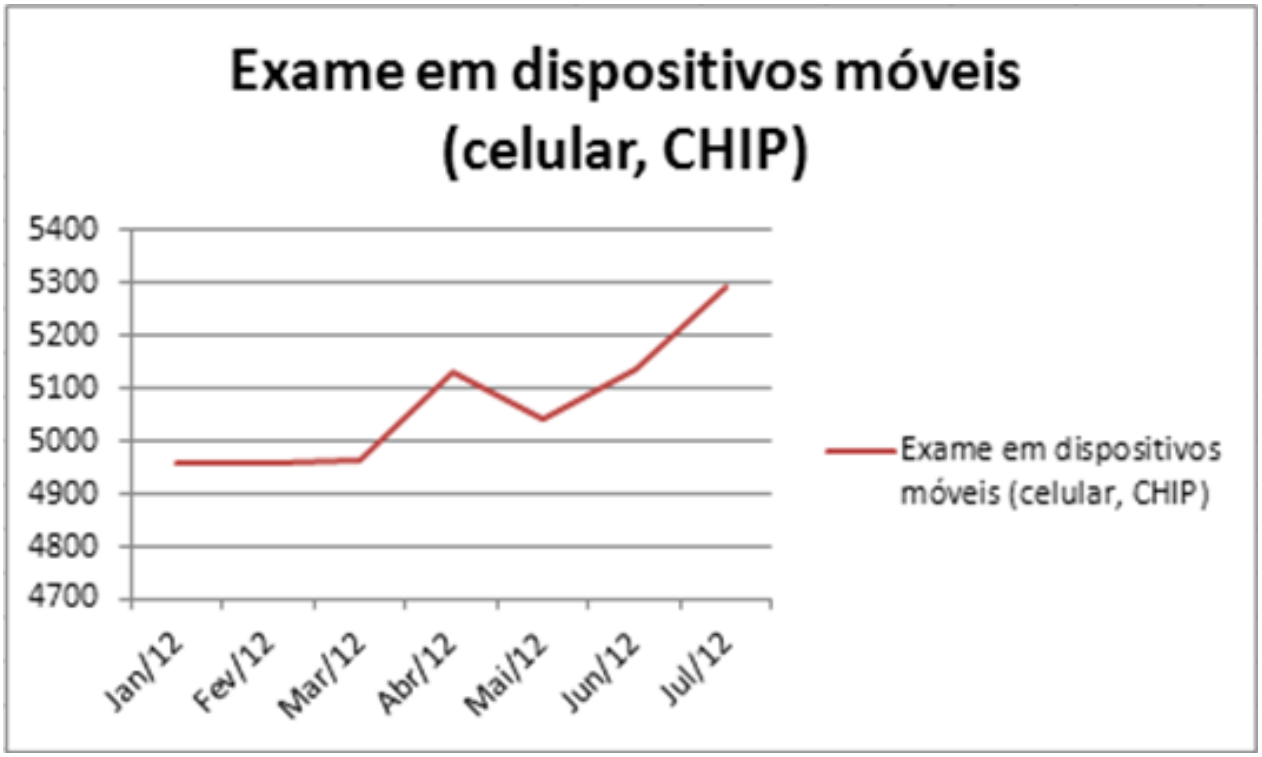


Figura 1.1: Histórico de celulares em estoque.

A Seção de Informática Forense do Instituto de Criminalística do Paraná organizou um gráfico, apresentado na Figura 1.1, no qual se pode observar que a tendência é o crescimento vertiginoso do número de dispositivos móveis a periciar, visto que dispositivos tornem-se cada vez mais portáteis e acompanham a mobilidade do usuário.

Nesse contexto surge a necessidade do desenvolvimento de técnicas computacionais que realizem a tarefa de exploração destes documentos proporcionando a recuperação, visualização e conhecimento de informações contidas nos mesmos.

Processos e ferramentas computacionais que possibilitem a exploração desse volume de dados contido nos laudos periciais de dispositivos móveis, que se mostra cada vez mais crescente, é fundamental no processo de recuperação de informações úteis e que atendam as expectativas dos profissionais da área forense.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver um método que realize o cruzamento de informações contidas nos laudos periciais de dispositivos móveis. Ou seja, se valendo do uso da Ciência da Informação disponibilizar uma ferramenta computacional que auxilie os Serviços de Inteligência e Policiamento Preditivo.

A Figura 1.2 mostra uma visão da interação do método proposto com o exercício da atividade Forense na elaboração de Laudos Periciais de Dispositivos Móveis.

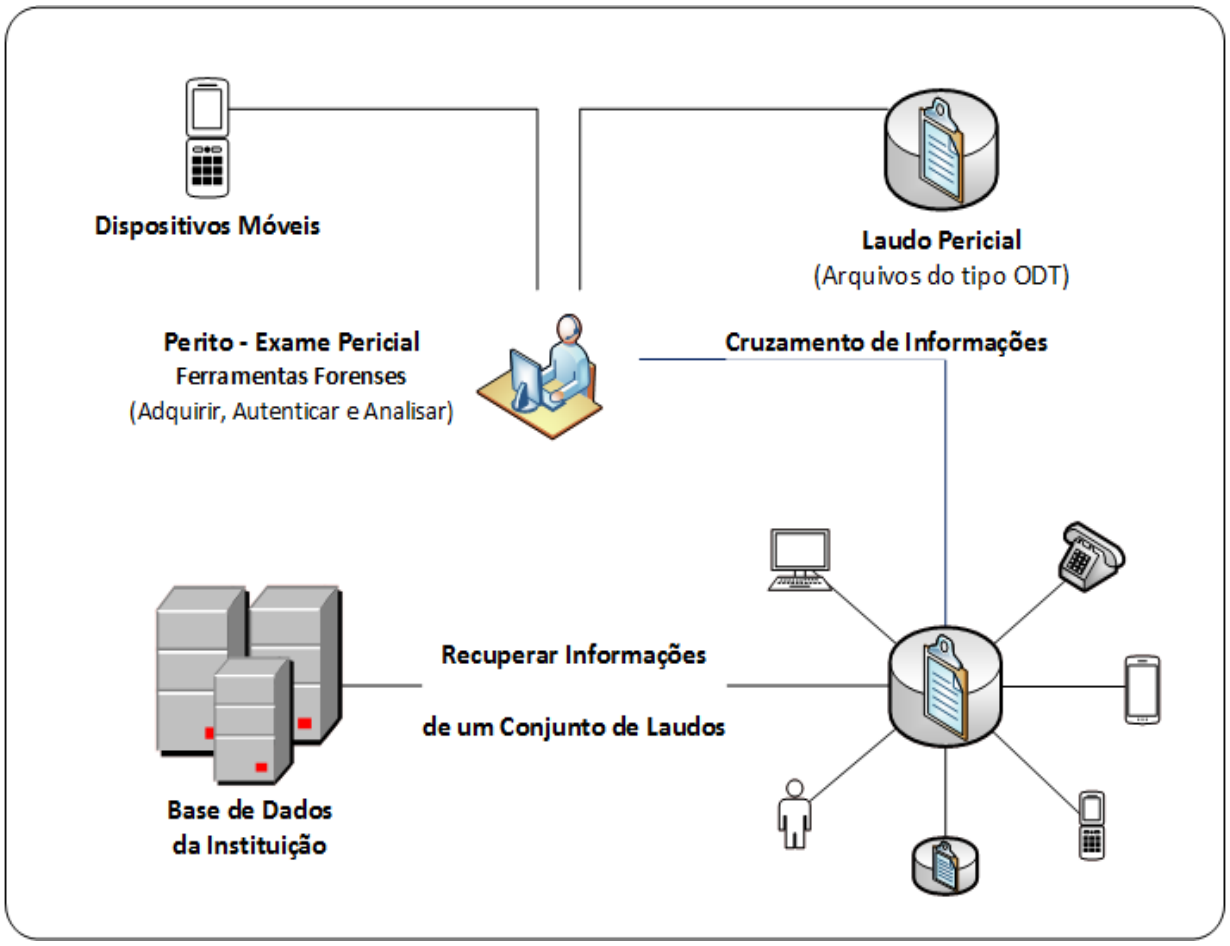


Figura 1.2: Interação entre o Método Proposto e o exercício da atividade Forense.

1.2.2 Objetivos Específicos

- Recuperar registros telefônicos contidos em documentos de laudos periciais de dispositivos móveis;
- Identificar o cruzamento de dados entre os documentos;
- Representar os cruzamentos existentes por meio de grafos.

1.3 Hipótese do Trabalho

A hipótese deste trabalho é:

- "É possível recuperar os registros telefônicos e representar o cruzamento de dados existente entre os laudos periciais de dispositivos móveis utilizando técnicas de recuperação de informações."

1.4 Contribuição

A principal contribuição científica do presente trabalho foi a criação de uma ferramenta computacional que realiza a detecção de cruzamento de registros telefônicos contidos em laudos periciais de dispositivos móveis armazenados em documentos textuais não estruturados.

A ferramenta computacional apresentada neste trabalho é um método que além de Cruzamento de Informações em Laudos Periciais de Dispositivos Móveis contribui para se evitar a subjetividade no exercício da atividade pericial.

O uso da Recuperação de Informações e da Descoberta do Conhecimento proporcionou produção de provas e evidências forenses que até então encontravam-se ocultos em documentos dispersos nas Sessões de Informática Forense.

O trabalho desenvolvimento nessa dissertação gerou, até o momento, a publicação de dois artigos científicos e um registro de software, a saber:

- Artigo: SICReT Sistema de Cruzamento de registros Telefônicos (DECARLI *et al.*, 2013);

- Artigo: Banco de Dados de Laudos Periciais de Dispositivos Móveis (DECARLI *et al.*, 2014);
- Registro de Programa de Computador: Sistema de Cruzamento de Registros Telefônicos - SiCReT. Número do registro no INPI: BR 51 2015 001516 8. Publicada na Revista da Propriedade Industrial - RPI nº 2350, datada de 19/01/2016.

1.5 Organização do Documento

Este trabalho está organizado de forma que o Capítulo 2 apresenta a Fundamentação Teórica necessária para entendimento e desenvolvimento da presente pesquisa e do Método Proposto.

O Capítulo 2 remete o leitor aos princípios e conceitos de Computação Forense, Computação Ubíqua, Coleta de Dados de Dispositivos Móveis e Laudos Periciais Criminais. Apresentando ainda conceitos fundamentais da Recuperação de Informação e Descoberta do Conhecimento.

No Capítulo 3 é feito uma abordagem dos principais trabalhos relacionados que trabalham área de Recuperação de Informações em documentos textuais não estruturados, em especial na área Forense.

O Capítulo 4 apresenta do Método Proposto, fazendo um detalhamento da representação computacional das fases que o compõem. Percorrendo além do Pré-Processamento, as fases de Composição do Conjunto de Documentos, Representação dos Documentos, Processo de Indexação e Apresentação dos Resultados.

O Capítulo 5 mostra os resultados obtidos nos experimentos realizados para comprovação da eficácia do método proposto. Os experimentos foram aplicados em 200 arquivos contendo laudos de perícia criminal em dispositivos móveis do Instituto de Criminalística do Paraná.

O Capítulo 6 finaliza o trabalho apresentando as considerações de conclusão do presente trabalho e trabalhos que podem ser desenvolvidos no futuro.

Capítulo 2

Fundamentação Teórica

Este Capítulo apresenta os principais conceitos e fundamentos teóricos que contribuem para um melhor entendimento do presente trabalho. Inicialmente apresentado é um estudo dos conceitos básicos envolvidos na Computação Forense sob os alicerces da Recuperação da Informação e da Descoberta do Conhecimento.

A Figura 2.1 mostra que as áreas de trabalho e estudo que permeiam a Computação Forense podem contar com a sustentação de sólidos pilares proporcionados pela Ciência da Computação.

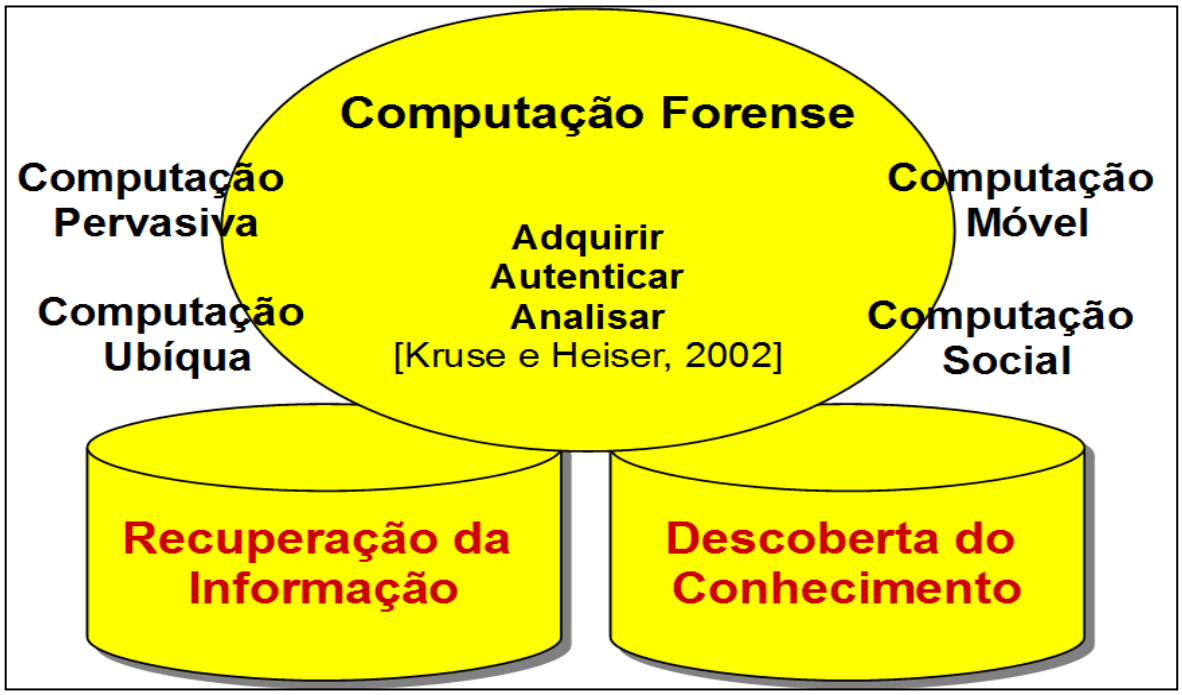


Figura 2.1: Fundamentos Teóricos envolvidos e seu relacionamento.

2.1 Computação Forense

A área denominada de Computação Forense ou Forense Computacional (C*omputer Forensic*) envolve a extração, identificação, preservação e documentação de evidências digitais a partir de dados e informações armazenadas em mídias: magnéticas, ópticas ou eletrônicas (CRAIGER, 2007). Para Michaud, a Computação Forense pode ser definida como uma peça do quebra-cabeça da investigação (MICHAUD, 2001).

Assim, entende-se que Computação Forense é o termo técnico utilizado para definir as perícias realizadas em computadores. A atividade pericial surge da necessidade de se provar a verdade dos fatos, em que se funda a ação ou a defesa. Nessa atividade os peritos fazem análises e aplicam procedimentos específicos com o objetivo de identificar, preservar, analisar e apresentar evidências digitais de modo que estas sejam legalmente aceitas. De acordo com Kruse & Heiser, os métodos e técnicas podem ser resumidos por meio do mnemônico "3A’s", a saber: 1) Adquirir as evidências sem alterar ou danificar o original; 2) Autenticar que as evidências recuperadas são idênticas as originais; 3) Analisar os dados sem que estes sofram modificações (KRUSE; HEISER, 2002).

Na obra "Computer Security - guidelines on cell phone forensics", os autores Jansen & Ayers deixam bem claro que investigações digitais são comparáveis à cenas de crime, visto que técnicas de investigação com base na aplicação da lei têm sido utilizadas para a criação de procedimentos voltados a evidências digitais (JANSEN; AYERS, 2007).

Assim, cabe destacar tal qual Jansen & Ayers, que os Princípios de Probatória consideram a prova digital em dois aspectos, a saber:

- **os componentes físicos**, periféricos e mídia, que podem conter dados;
- **e os dados extraídos** a partir dessas fontes.

Na verdade, os autores estão caracterizando os tipos de evidências a serem coletadas durante o procedimento de produção de provas digitais, a saber:

- **Físicas**: computadores (servidor, *desktops*, *laptops*), HD externos, *pen-drives* (*mp3-player*), CDs, DVDs, celulares, câmeras digitais, jogos e outros;
- **Lógicas ou demonstrativas**: dados, informações, arquivos, textos, imagens, vídeos, músicas, e-mails, entre outros que se encontram armazenados em suportes físico, seja este eletrônico, ótico ou magnético.

Portanto, tais autores sugerem que sejam respeitados quatro princípios ao trabalhar se com evidências digitais, que podem ser resumidos como a seguir:

a) ações realizadas por investigadores/peritos não devem alterar dados contidos em dispositivos digitais ou em mídias de armazenamento que podem posteriormente ser solicitados perante o Juiz;

b) indivíduos que acessam dados originais devem ser competentes para fazê-lo e ter a capacidade de explicar suas ações, visto que tais procedimentos são questionáveis pelas partes ou em juízo;

c) uma cadeia de custódia deve ser estabelecida, bem como o registro de todos os procedimentos realizados deve ser mantido, de maneira que se possa garantir a replicação dos resultados por um terceiro independente, sendo que toda documentação deve ser criada e preservada, documentando-se cada passo investigativo/pericial;

d) a pessoa encarregada da investigação tem a responsabilidade geral de assegurar os procedimentos já mencionados e se os mesmos serão ou foram seguidos em conformidade com os métodos científicos e as leis vigentes.

No tocante à esfera jurídica, o ambiente computacional é regido pelas leis jurídicas vigentes no país, de acordo com o perito criminal Luiz Rodrigo Grochock: "no ambiente computacional são aplicáveis os mesmos valores, as mesmas leis, só que com uma releitura dentro desta nova realidade" (GROCHOCK, 2013).

Dessa forma um crime virtual é a ocorrência de uma prática tipificada como crime no ambiente virtual, e podem ser classificados em dois tipos: Crimes Cibernéticos Abertos ou Exclusivamente Cibernéticos;

O perito demonstra que dentro dessa nova realidade é existe alguns princípios do direito digital a serem considerados no exercício de atividades na área da Computação Forense, a saber:

a) Relações não presenciais;
b) Testemunhas são máquinas;
c) Provas eletrônicas;
d) Fronteiras informacionais e não físicas.

Na data de 30 de novembro de 2012, o Congresso Nacional decretou e sancionou a Lei Nº 12.737[3], também conhecida como Lei Carolina Dieckmann, adicionando no rol de crimes já previstos no Código Penal a tipificação dos delitos informáticos, a saber:

- **Invasão de dispositivo informático**: invasão de computadores, roubo de senhas e de conteúdos de e-mail, e a disseminação de vírus de computador ou códigos maliciosos para roubo de senhas;
- **Interrupção ou perturbação de serviço** telegráfico, telefônico, informático, telemático ou de informação de utilidade pública: a derrubada proposital de sites;
- **Falsificação de documento particular**: falsificar um documento eletrônico em todo ou em parte;
- **Falsificação de cartão**: o uso de dados de cartões de débito e crédito sem autorização do titular.

Considera-se no contexto deste trabalho que a atividade de Perícia Judicial, consiste na atividade pericial onde pessoas especializadas emitem um parecer ao Juiz verificando fatos interessantes à causa do processo judicial (ROSA, 1999). A profª Cinthia O. A. Freitas complementa esse entendimento apresentando a definição de que "a perícia surge da necessidade de se provar a verdade dos fatos" (FREITAS, 2008).

O CPP (Código de Processo Penal) especifica que o Exame de Corpo de Delito e outras perícias serão realizados por perito oficial e será requisitado pela autoridade ao diretor da repartição, juntando-se ao processo o laudo assinado pelos peritos (CPP[4], art.159 e 178). Na falta de perito oficial, o exame pericial será realizado por 2 (duas) pessoas idôneas, portadoras de diploma de curso superior preferencialmente na área específica, dentre as que tiverem habilitação técnica relacionada com a natureza do exame (CPP[5], art.159, §1º). O CPC

[3] Disponível em <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12737.htm >. Acesso em 5 set 2015.
[4] Código de Processo Penal. Disponível em <http://www.planalto.gov.br/ccivil_03/decreto-lei/del3689compilado.htm>. Acesso em 05 de set 2015.
[5] Código de Processo Penal. Disponível em <http://www.planalto.gov.br/ccivil_03/decreto-lei/del3689compilado.htm>. Acesso em 05 de set 2015.

(Código de Processo Civil) menciona que os peritos serão escolhidos entre profissionais de nível universitário, devidamente inscritos no órgão de classe competente (CPC[6], art. 145, §1º).

Dessa forma, a perícia Judicial é produzida dentro do processo judicial e, de acordo com o Fórum, a requisição da atividade pericial pode ser de ordem:

- **Criminal**: por determinação de autoridades policiais no território de suas respectivas circunscrições (CPP, art. 4, §6º, inc. VII);
- **Cível**: por nomeação do Juiz (CPC, art. 421).

O conceito do Método Proposto vem ao encontro e contempla os conceitos, recomendações e aspectos jurídicos que devem ser consideradas no desenvolvimento da atividade de Computação Forense. Objetiva proporcionar uma ferramenta em forma de um Ferramenta Computacional, para os profissionais que exercem essa atividade, em especial os peritos de Institutos de Criminalística, contribuindo assim com a extração, identificação, preservação e documentação das evidências digitais de maneira sistematizada.

2.2 Computação Ubíqua

A ubiqüidade é a propriedade daquilo que está presente em todos os lugares ao mesmo tempo, ou seja, algo onipresente (HOUAISS; VILLAR, 2004).

Registros apontam que o artigo científico *The Computer for the 21st Century* (WEISER, 1991) foi a primeira publicação científica que utilizou o termo "ubíquo" para expressar o uso da informática no cotidiano das pessoas realizado de forma transparente, ou seja, o acesso do usuário ao ambiente computacional, em todo lugar e a todo momento, por meio de dispositivos embutidos nas estruturas básicas e fundamentais da vida do ser humano.

Cabe então caracterizar alguns termos relacionados a este paradigma. Neste sentido, Computação Pervasiva é um termo que se refere à computação embutida (*embedded*) nos aparatos tecnológicos e nos mais variados objetos, sendo diferente dos computadores tradicionais principalmente em relação a sua interface e ao modo mais intuitivo de utilização,

[6] Disponível em <http://www.planalto.gov.br/ccivil_03/leis/l5869compilada.htm>. Acesso em 05 de set 2015.

por meio do tal as tarefas computacionais são realizadas de maneira implícita (WEISER, 1991).

Deste modo, Computação Ubíqua, oriunda do termo em inglês *Ubiquitous Computing* ou *Ubicomp*, e tem por objetivo integrar computadores de forma transparente, aprimorando o mundo real, formulando uma nova forma de pensar para os computadores no mundo e, ainda, considerando que os mesmos sejam inseridos no ambiente natural do ser humano de maneira invisível e, imperceptível (WEISER, 1993).

Greenfield, em sua obra " *Everyware: The dawning age of ubiquitous computing*", escreve sobre uma visão de um novo paradigma da Computação Ubíqua, que ele denomina de *everyware*, por meio do qual o processamento de dados e de informações não estariam em um único objeto, mas sim estariam tanto nos computadores, como em qualquer lugar, inseridos nas estruturas do dia a dia, como casa, carros e dispositivos móveis (GREENFIELD, 2006).

Nessa visão o processamento da informação estaria distribuído no comportamento e na necessidade das pessoas, tratando-se assim de estar diretamente envolvida com a Computação Social, ou também Engenharia Social.

Essas informações, distribuídas nos aparatos tecnológicos e no cotidiano das pessoas, abrem portas para ações que trazem consigo a malícia e a arte de enganar estampada em entendimentos a exemplo de Mitnick e Young & Aitel.

A engenharia social usa a influência e a persuasão para enganar as pessoas e convencê-las de que o engenheiro social é alguém que na verdade não é, ou pela manipulação. Como resultado, o engenheiro social mau intencionado pode aproveitar-se das pessoas para obter as informações com ou sem o uso da tecnologia (MITNICK, 2003).

Ou ainda, na visão de Young & Aitel, trazendo um entendimento de que o objetivo da Engenharia Social é de enganar as pessoas para revelar senhas ou outras informações que comprometa a segurança de um sistema de destino (YOUNG; AITEL, 2004).

Ao mesmo tempo, o entendimento desses conceitos maliciosos são fundamentais e servem de base para a Engenharia Social ser elevada a um patamar usado por profissionais que usam esses princípios da Computação Social de forma benéfica. A exemplo de investigações criminais no exercício da atividade forense computacional. Ou ainda, para trabalhos voltados para Segurança das Informações, preservação e sigilo dos dados no meio digital.

Dessa forma, a Engenharia Social pode ser definida como a arte de manipular as informações de uma determinada pessoa, com vista para a aquisição, análise e aprendizagem de informações ou mesmo para atender a um determinado sistema (STEPHENSON, 2000).

Dentro desse novo paradigma da Computação Ubíqua descrita por Greenfield, vislumbra-se a possibilidade de que os dispositivos sejam cada vez mais sensíveis e compreensíveis às nossas ações. Em outras palavras, o que é proposto é que os objetos que constituem os lugares onde as pessoas se encontram, além de onipresentes, possam ser reativos e até pró-ativos quanto à presença ou ausência de pessoas (GREENFIELD, 2006).

O autor descreve que está ocorrendo uma espécie de colonização do cotidiano pela tecnologia da informação, de modo que a tecnologia será cada vez mais aplicada em tarefas comuns do dia a dia.

Dessa forma a Computação Ubíqua está cada vez mais inserida no cotidiano das pessoas, isso se torna uma realidade com o aprimoramento constante da computação embutida, tornando o ambiente computacional cada vez mais presente em todo lugar a todo momento de nossas vidas.

Esse conceito em que os dispositivos sejam inseridos no ambiente natural do ser humano de maneira invisível para as pessoas, carrega consigo um ambiente computacional que se demonstra cada vez mais rico, poderoso e complexo.

2.3 Dispositivos Móveis

A Computação Móvel proporciona a capacidade de mover fisicamente serviços computacionais junto com os usuários, tornando os dispositivos computacionais sempre presentes, permitindo ao ser humano que tenha acesso aos recursos oferecidos por um sistema computacional independentemente da sua localização (WEISER, 1993).

Na Publicação Especial 800-101 do NIST (*National Institute of Standards and Technology*) (JANSEN; AYERS, 2007) os autores sugerem que a chave para o sucesso na análise forense de dispositivos móveis é a compreensão das características de *hardware* e *software* dos telefones celulares. Os dados dos assinantes e suas atividades por meio de celulares são muitas vezes uma fonte valiosa de provas em uma investigação. Portanto, para que a produção de provas possa ser realizada, conta-se com um conjunto básico de

características, obtido a partir da maioria dos celulares, sendo este conjunto comparável entre diferentes aparelhos.

Como exemplos de características que são comuns na maioria dos dispositivos móveis atuais pode ser citado: microprocessador, memória ROM (*Read Only Memory*), memória RAM (*Random Access Memory*), módulo de rádio, processador de sinal digital, alto falante, tela, sistema operacional, bateria, PDAs (*Personal Digital Assistants*), GPS (*Global Positioning System*), câmera, entre outros recursos.

A aquisição de dados a partir de um dispositivo pode ser física ou lógica (JANSEN; AYERS, 2007). A aquisição física tem vantagens sobre a aquisição lógica, uma vez que permite que os arquivos apagados e alguns dados restantes possam ser examinados, por exemplo, na memória não alocada ou em espaço do sistema de arquivos.

É recomendado sempre fazer a aquisição de dados física antes da aquisição lógica. As ferramentas forenses adquirem informações dos dispositivos sem alterar o conteúdo, ou seja, em modo somente de leitura, e em geral geram *hash*, MD5 (*Message-Digest Algorithm 5*[7]) ou SHA (*Secure Hash Algorithm*[8]), que garante a integridade dos dados.

De um modo geral, a função *hash* tem por objetivo identificar univocamente cada conjunto de informações, ou seja, para cada documento criptografado gera-se uma cadeia alfanumérica única, sendo que o procedimento (ou algoritmo) de geração usa o conteúdo do documento para gerar tal cadeia (FREITAS; RICCI, 2012). Assim, se um documento for modificado e novamente criptografado, nunca conterá o mesmo *hash*, pois o conteúdo do documento foi alterado.

Portanto, a simples comparação dos valores dos *hashs* de dois documentos, permite a validação da autenticidade dos mesmos, visto que somente para *hashs* iguais têm-se documentos iguais. Tal característica é muito importante perante um Juiz, sendo que cabe ao perito garantir a integridade das provas digitais por ele coletadas ou a ele confiadas.

2.3.1 Coleta de Dados Periciais

A atividade pericial da análise forense de dispositivos móveis inicia-se na aquisição dos dados digitais a partir dos celulares, sendo que neste trabalho foi estudado os métodos das

[7] RSA Data Security, Inc. MD5 Message Digest Algorithm. Disponível em <http://www.efgh.com/software/md5.htm>. Acesso em 25 de set. de 2015.
[8] NIST: The Secure Hash Algorithm Validation System. Disponível em <http://csrc.nist.gov/groups/STM/cavp/documents/shs/SHAVS.pdf >. Acesso em 25 de set. de 2015.

ferramentas forenses de captura, *hardware* e *software*, utilizadas pelo Instituto de Criminalística do Paraná, a saber: Cellebrite UFED[9] e Microsytemation XRY[10].

Ambos realizam a extração de dados em padrão XML (*eXtensible Markup Language*), padrão este que permite descrever diversos tipos de dados, tendo como objetivo facilitar o compartilhamento de informações.

O *Cellebrite UFED* (*Universal Forensic Extraction Device*) *Touch Ultimate* é uma solução composta por *hardware* e *software* proprietário que permite a extração, decodificação, análise e geração de relatórios avançados em termos tecnológicos dos dados de dispositivos móveis, sendo suportados atualmente 7.900 dispositivos diferentes.

Conta também com uma interface interativa com tela sensível ao toque e um conjunto de cabos com interface USB e RJ45 que realizam a conexão e comunicação com os dispositivos móveis, conforme exemplar mostrado na Figura 2.2.

Figura 2.2: UFED Touch Ultimate.

Fonte: <http://www.militarysystems-tech.com/suppliers/military-mobile-forensics/cellebrite-ltd> Acesso em 26 mar. 2014.

Dentre os principais aplicativos que acompanham a solução vale ressaltar o *Phisical Analyzer* (ferramenta de decodificação, análise e relatórios), o *Phone Detective* (*software* que

[9] Cellebrite UFED. Disponível em <http://www.cellebrite.com/pt/mobile-forensic-products/ufed-touch-ultimate.html>. Acesso em 20 jan. 2014.
[10] Microsytemation XRY . Disponível em < http://www.msab.com/xry/what-is-xry >. Acesso em 04 mar. 2014.

identifica um telefone móvel no início de uma investigação) e o *Reader* (inicialização dos dispositivos em modo somente leitura, permitindo o compartilhamento de informações).

O Instituto de Criminalista conta ainda com o *Microsytemation XRY*, que realiza função semelhante de captura. Em sua versão atual, a 6.7, a solução XRY suporta 10.036 dispositivos móveis, sendo que sua principal característica é essa considerável quantidade de dispositivos suportados, devido a esse fator a ferramenta é utilizada em larga escala na área forense.

Figura 2.3: Microsytemation XRY.

Fonte: < http://www.aresources.pt/produtos.php?sub=39&fam=34> Acesso em 26 mar. 2014.

A Figura 2.3 mostra o *XRY Complete*, que é uma solução do *Microsytemation XRY* composta por aplicativos (*software)* e equipamentos (*hardware)* que permitem aos peritos realizar a extração forense física e lógica de dispositivos móveis.

Por se mostrarem ferramentas bem completas e confiáveis, o Cellebrite UFED e o Microsytemation XRY são as ferramentas de captura de dados de dispositivos móveis utilizadas apenas no Instituto de Criminalística do Paraná, mas sim utilizadas na grande maioria dos Institutos de Criminalística atualmente.

A partir de todo o conjunto de dados extraído pelas ferramentas de captura, pode-se então analisar quais dados permitirão a realização do cruzamento de informações provenientes de celulares distintos. Sabe-se *a priori* que são importantes para o cruzamento os dados contidos nos contatos da agenda, nas chamadas (realizadas e recebidas) e mensagens instantâneas trocadas entre celulares distintos. Alguns modelos de *smartphones* fornecem dados de mensagens eletrônicas (*e-mail*), salas de bate-papo (*chat*), entre outros.

Por se tratar de uma natureza de dados bem diversificada, a captura de dados resulta em um amplo universo de dados armazenados de forma digital em uma enorme diversidade de formas e formatos.

Apesar dos esforços de peritos das Sessões de Computação Forense dos Institutos de Criminalística ainda não existe um padrão de informações que devem ser extraída, e nem um formato padrão de armazenamento.

O documento base de boas práticas na captura de dados em dispositivos móveis seguido por todos os Institutos de Criminalística é o "Computer Security - guidelines on cell phone forensics", divulgado na Publicação Especial 800-101 do NIST. Contudo o próprio documento recomenda que cada Instituto crie sua própria organização e forma de trabalho mais adequadas à sua realidade. Com o intuito de utilizar da melhor forma possível os recursos humanos e tecnológicos disponíveis.

Atendendo essa recomendação o Instituto de Criminalística do Paraná tem sua própria forma de trabalho e manipulação das informações, vem alcançando resultados que o tornaram como referência nacional em se tratando de Laudos Periciais de Dispositivos Móveis.

A Tabela 2.1 exemplifica informações universais, amplamente utilizadas na captura de dados em dispositivos móveis. Aqui demonstradas com base nos dados obtidos na utilização das ferramentas de *hardware* e *software Cellebrite UFED e Microsytemation XRY*.

Tabela 2.1: Informações Gerais da Captura.

Parâmetros
Fabricante selecionado
Modelo selecionado
Fabricante detectado
Modelo detectado
Nome do equipamento
IMEI (*International Mobile Equipment Identity*)
ICCID (*International Circuit Card ID*)
IMSI (*International Mobile Subscriber Identity*)
Endereço Bluetooth
Endereço Wi-Fi
Início da extração

Fim da extração
Data/Hora do telefone
Tipo de conexão
Versão da UFED

2.3.2 O Laudo Pericial

Na esfera da metodologia científica em perícia criminal um relatório é uma exposição gráfica e geral de um assunto, sendo que compreende desde o planejamento, passando por todos os procedimentos, até chegar à conclusão, materialidade e autoria, incluindo-se os processos metodológicos empregados, recursos, equipamentos e ferramentas (REIS, 2011).

Diferente de um trabalho técnico comum, o trabalho pericial elege como elemento essencial de sua estrutura a resposta aos quesitos propostos à perícia, a qual pode ser apresentada através de parecer sucinto, apenas com respostas aos quesitos formulados, ou através da exposição detalhada dos elementos investigados, sua análise e fundamentação das conclusões, além das respostas aos quesitos formulados (ROSA, 1999).

Para Reis, o Laudo Pericial Criminal "é um dos itens mais importantes no estuda da Criminalística, pois é através dele que os exames são expressos e que a prova material do crime é manifestada". Deve apresentar "seu conteúdo exclusivamente voltado para o rigor das Leis Naturais, evitando qualquer relação com as Leis Jurídicas e com as Leis da Consciência" (REIS, 2011).

O CPP prevê que os peritos elaborarão o laudo pericial, onde descreverão minuciosamente o que examinarem, e responderão aos quesitos formulados (CPP, art. 160). Menciona ainda que o laudo pericial será elaborado no prazo máximo de 10 dias, podendo este prazo ser prorrogado, em casos excepcionais, a requerimento dos peritos (CPP, art. 160, §1º). Nas perícias de laboratório, os peritos guardarão material suficiente para a eventualidade de uma nova perícia. Sempre que conveniente, os laudos serão ilustrados com provas fotográficas, ou microfotográficas, desenhos ou esquemas (CPP[11], art. 170).

Uma das tarefas mais árduas enfrentada diariamente pelos peritos é preparar o Laudo Pericial, isto é, expressar em papel, uma opinião formulada por ele relativo a um determinado caso. Esta opinião deve freqüentemente acomodar uma análise complexa (FREITAS, 2008).

[11] Código de Processo Penal. Disponível em <http://www.planalto.gov.br/ccivil_03/decreto-lei/del3689compilado.htm>. Acesso em 05 de set 2015.

A autora apresenta que o laudo pericial deve ser composto por:

- **Relatório:** sumário da perícia e principais ocorrências detectadas;
- **Parte Expositiva:** fundamentação do perito, ou seja, especificação de critérios e métodos utilizados na realização da perícia;
- **Parte Conclusiva:** resposta aos quesitos formulados de forma objetiva, sem considerações sobre o objeto do processo e sem posicionamento pessoal a favor das teses discutidas no processo;
- **Parte Autenticativa:** assinatura e rubrica dos peritos garantindo a validade do documento e a identificação de sua autoria;
- **Anexos:** documento auxiliar contendo o detalhamento de fotografias, mensagens, chamadas, vídeos, localização do dispositivo móvel e demais materiais e conteúdos complementares ao relatório.

O Laudo Pericial é um relatório técnico sobre os exames realizados, no qual o perito discorre sobre todo o trabalho e apresenta o resultado de suas análises (COSTA, 2003).

Laudo Pericial é a expressão final do trabalho de perícia. A Figura 2.4 mostra o fluxo de trabalho envolvido na elaboração desse documento.

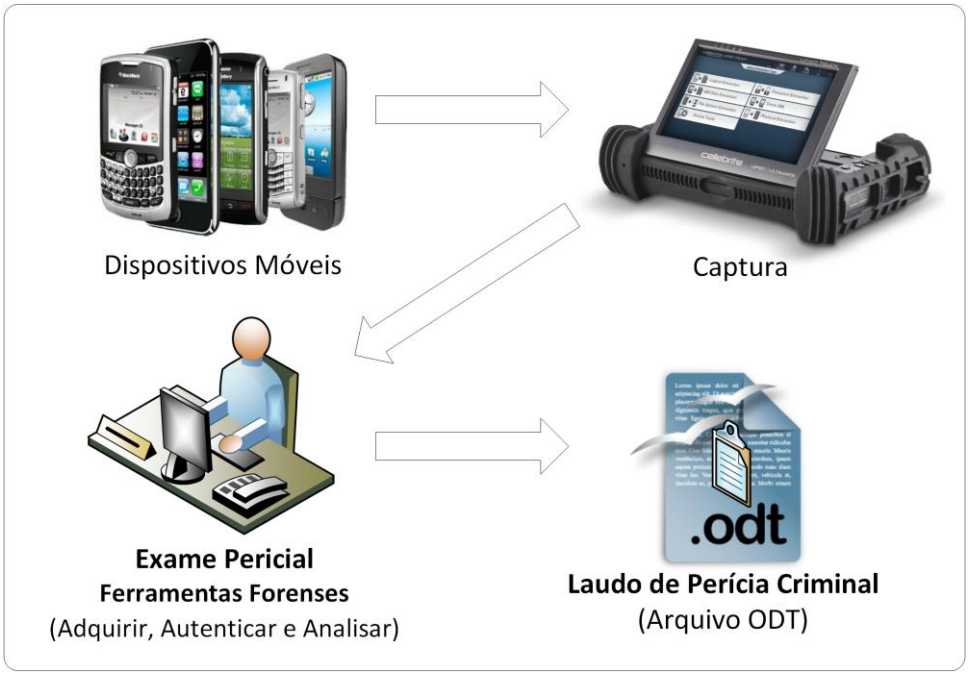


Figura 2.4: Elaboração do Laudo Pericial Criminal.

A Figura 2.4 mostra uma síntese de forma visual das etapas do trabalho realizado pelos peritos na elaboração do Laudo Pericial Criminal no Instituto de Criminalística do Paraná, a saber:

a) Manipulação dos Dispositivos Móveis;
b) Captura de Informações dos Dispositivos Móveis;
c) Árdua tarefa de Exame Pericial;
d) Apresentar os resultados em um Laudo de Perícia Criminal.

O laudo concluído estará sujeito à revisão crítica por via de testemunhos, conferências de pré-julgamento, e talvez um interrogatório rigoroso por vários advogados (FREITAS, 2008).

2.4 Recuperação de Informações

Em 1951, Calvin Mooers criou o termo Recuperação de Informações (*Information* Retrieval) com a definição de que essa área trata dos aspectos intelectuais da descrição da informação e sua especificação para busca, e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação (MOOERS, 1951).

A Recuperação da Informação (RI) trata da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas Web, catálogos *online*, registros estruturados e semi-estruturados, objetos multimídia, etc. A representação e a organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse (BAEZA-YATES; RIBEIRO-NETO, 2013).

De acordo com Baeza-Yates & Ribeiro-Neto (2013) a área de RI pode ser estudada sob dois pontos de vista distintos e complementares: um centrado no computador e o outro centrado no usuário, a saber:

- **A visão centrado no computador**, a RI consiste principalmente na construção de índices eficientes, no processamento de consultas com alto desempenho e no desenvolvimento de algoritmos de ranqueamento, a fim de melhorar os resultados.
- **A visão centrada no usuário**, a RI consiste principalmente em estudar o comportamento do usuário, entender suas principais necessidades e determinar

> como esse entendimento afeta a organização e a operação do sistema de recuperação.

A descrição completa da necessidade do usuário não necessariamente fornece a melhor formulação de consulta para um determinado sistema de RI. Em vez disso, o usuário pode querer primeiro traduzir essa necessidade de informação em uma consulta ou em uma sequência de consultas a serem submetidas ao sistema. Em sua forma mais comum, essa tradução gera uma série de palavras-chave, ou termos de indexação, que sumarizam a necessidade de informação do usuário.

Dada a consulta do usuário, o objetivo maior do sistema de RI é recuperar informações que sejam úteis ou relevantes para o usuário. A ênfase está na recuperação da informação, não na recuperação de dados.

O principal objetivo de um sistema de RI é recuperar todos os documentos que são relevantes à necessidade de informação do usuário e, ao mesmo tempo, recuperar o menor número possível de documentos irrelevantes.

Um ponto importante é que a relevância é um julgamento pessoal que depende da tarefa a ser resolvida e de seu contexto. Por exemplo, a relevância pode mudar com o tempo, à medida que novas informações tornam-se disponíveis, com o local, a resposta mais relevante pode ser a mais próxima, ou até mesmo com o dispositivo, a resposta mais adequada pode ser um documento pequeno que seja mais fácil de ser acessada e visualizada. Nesse sentido, nenhum sistema de RI pode fornecer respostas perfeitas a todos os usuários o tempo todo.

A Figura 2.5, mostra o sequenciamento do Processo de Recuperação de Informação descrito por Salton & MacGill. (SALTON; MACGILL, 1983).

Os sistemas de recuperação de informação devem representar o conteúdo dos documentos e apresentá-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfaçam total ou parcialmente a sua necessidade de informação, formalizada por meio de uma expressão de busca (FERNEDA, 2012).

O processo de representação dos documentos tem por objetivo identificar e descrever cada documento por meio de seu conteúdo (FERNEDA, 2012).

De acordo com Ben Coppin, a Recuperação da Informação envolve casar o texto contido em uma consulta ou um documento com um conjunto de outros documentos.

Frequentemente, essa tarefa gera uma resposta encontrando documentos em um *corpus*[12] de documentos que seja relevante para a consulta de um usuário (COPPIN, 2012).

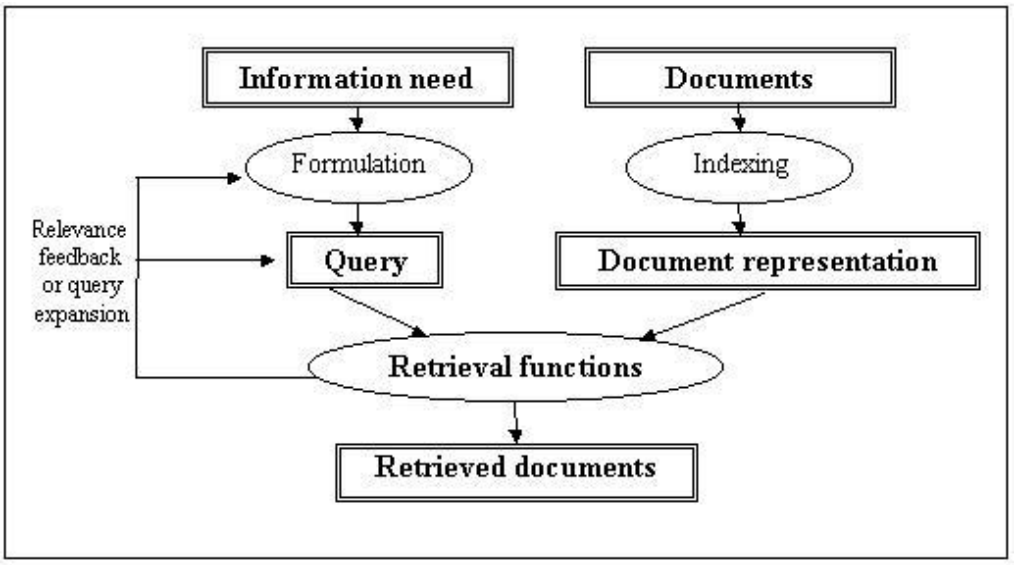


Figura 2.5: Processo de Recuperação de Informação (SALTON; MCGILL, 1983).

A definição de Coppin é reforçada e mais detalhado por Norving & Russell quando afirmam que a Recuperação da Informação é a tarefa de encontrar documentos que são relevantes para a necessidade de um usuário obter informação. Os exemplos mais conhecidos de RI são os mecanismos de busca na World Wide Web. Um usuário Web pode digitar uma consulta em um mecanismo de busca e obter uma lista de páginas relevantes (NORVING; RUSSELL, 2013).

Ainda seguindo o raciocínio de Norving & Russell (2013), um sistema de recuperação de informação pode ser caracterizado por:

- **Um *corpus* de documento**: cada sistema deve decidir o que quer tratar como documento: um parágrafo, uma página ou um texto de várias páginas;
- **Consultas colocadas em linguagem de consulta**: uma consulta especifica o que o usuário quer saber. A linguagem de consulta pode ser apenas uma lista de palavras, especificar um sintagma com palavras que devem ser adjacentes, conter operadores booleanos, operadores não booleanos, entre outros.

[12] *Corpus*: um corpo de texto, geralmente usado em problemas de recuperação de informação (COPPIN, 2012).

- **Um conjunto de resultado**: esse é o subconjunto de documentos que o sistema de RI julga ser relevante para a consulta. Por relevante queremos dizer provável que seja de utilidade para a pessoa que fez a consulta, para a necessidade de informação específica expressa na consulta.
- **Apresentação dos resultados**: isso pode ser tão simples como uma lista ordenada de títulos de documentos ou tão complexo como um mapa de cores de rotação do conjunto de resultados projetado em um espaço tridimensional, processado como uma exibição bidimensional.

2.4.1 Representação de Documentos

Um documento designa os objetos portadores de informação. Um documento pode ser considerado de forma geral como tudo o que representa ou exprime com a ajuda de sinais gráficos (palavras, imagens, diagramas, mapas, figuras, símbolos) um objeto, uma ideia (LE COADIC, 1996).

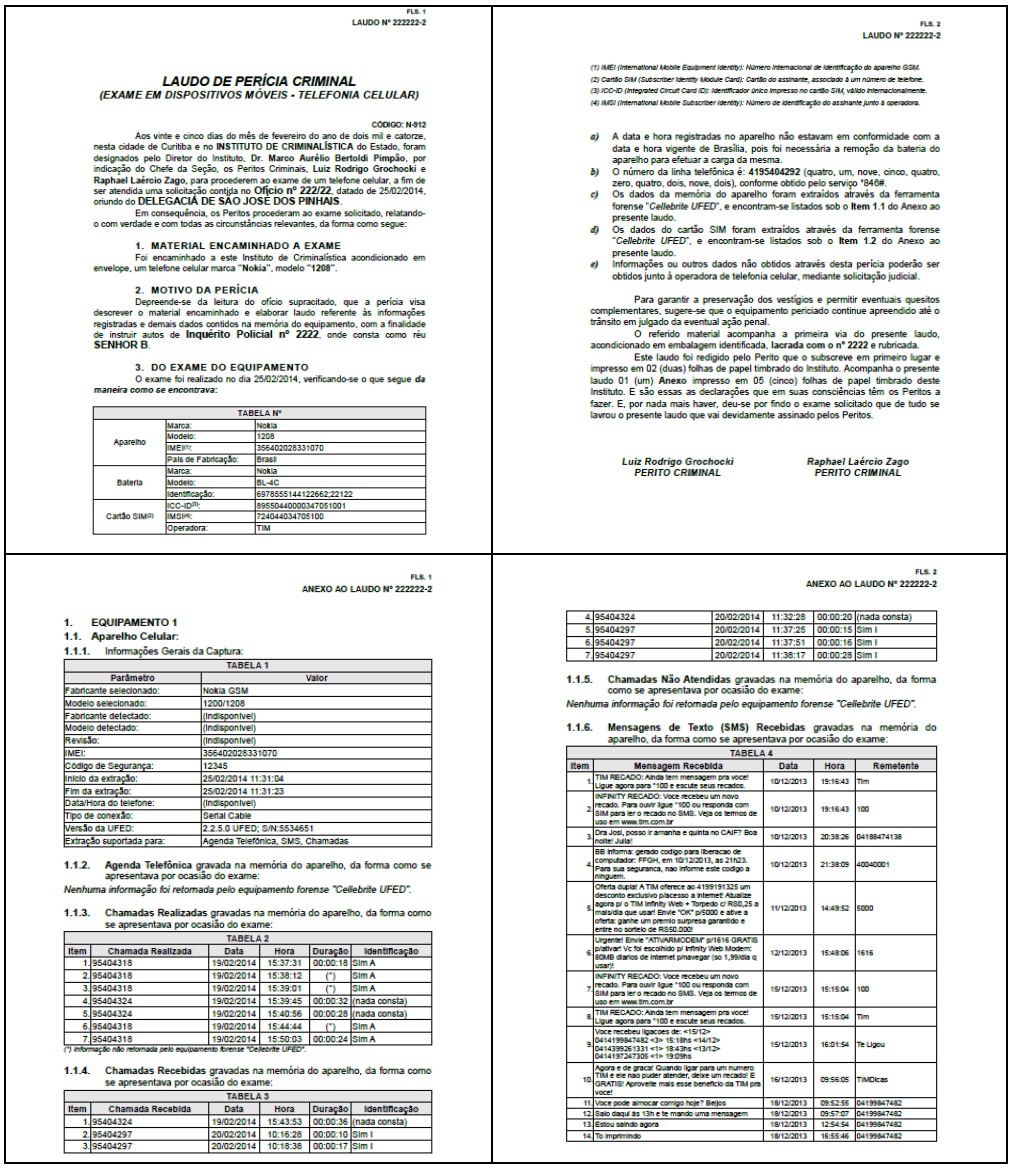
No presente trabalho, considera-se que um documento é uma unidade única de informação textual. A Figura 2.6 mostra um documento de laudo de perícia criminal.

O processo de representação dos documentos tem por objetivo identificar e descrever cada documento por meio de seu conteúdo. Dessa forma, todo documento é descrito por um conjunto de termos, ou ainda palavras-chave, selecionados para representá-lo. Sendo que um termo é qualquer palavra ou conjunto de palavras em um documento.

O conjunto de termos selecionados para representar um documento é denominado de Vocabulário. A Equação 2.1, define a forma geral do Vocabulário de um determinado documento.

$$V = \{k_1, ..., k_t\} \tag{2.1}$$

Sendo que V é o vocabulário, k refere-se aos termos selecionados a representar um determinado documento, e o t representa o número total de termos.



FLS. 1
LAUDO Nº 222222-2

LAUDO DE PERÍCIA CRIMINAL
(EXAME EM DISPOSITIVOS MÓVEIS - TELEFONIA CELULAR)

CÓDIGO: N-912

Aos vinte e cinco dias do mês de fevereiro do ano de dois mil e catorze, nesta cidade de Curitiba e no **INSTITUTO DE CRIMINALÍSTICA** do Estado, foram designados pelo Diretor do Instituto, **Dr. Marco Aurélio Bertoldi Pimpão**, por indicação do Chefe da Seção, os Peritos Criminais, **Luiz Rodrigo Grochocki** e **Raphael Laércio Zago**, para procederem ao exame de um telefone celular, a fim de ser atendida uma solicitação contida no **Ofício nº 222/22**, datado de 25/02/2014, oriundo do **DELEGACIA DE SÃO JOSÉ DOS PINHAIS**.

Em consequência, os Peritos procederam ao exame solicitado, relatando-o com verdade e com todas as circunstâncias relevantes, da forma como segue:

1. MATERIAL ENCAMINHADO A EXAME

Foi encaminhado a este Instituto de Criminalística acondicionado em envelope, um telefone celular marca "**Nokia**", modelo "**1208**".

2. MOTIVO DA PERÍCIA

Depreende-se da leitura do ofício supracitado, que a perícia visa descrever o material encaminhado e elaborar laudo referente às informações registradas e demais dados contidos na memória do equipamento, com a finalidade de instruir autos de **Inquérito Policial nº 2222**, onde consta como réu **SENHOR B**.

3. DO EXAME DO EQUIPAMENTO

O exame foi realizado no dia 25/02/2014, verificando-se o que segue *da maneira como se encontrava*:

TABELA Nº		
Aparelho	Marca:	Nokia
	Modelo:	1208
	IMEI(1):	356402028331070
	País de Fabricação:	Brasil
Bateria	Marca:	Nokia
	Modelo:	BL-4C
	Identificação:	6978555144122662;22122
Cartão SIM(2)	ICC-ID(3):	89550440000347051001
	IMSI(4):	724044034705100
	Operadora:	TIM

FLS. 2
LAUDO Nº 222222-2

(1) IMEI (International Mobile Equipment Identity): Número Internacional de Identificação do aparelho GSM.
(2) Cartão SIM (Subscriber Identity Module Card): Cartão do assinante, associado à um número de telefone.
(3) ICC-ID (Integrated Circuit Card ID): Identificador único impresso no cartão SIM, válido internacionalmente.
(4) IMSI (International Mobile Subscriber Identity): Número de identificação do assinante junto à operadora.

a) A data e hora registradas no aparelho não estavam em conformidade com a data e hora vigente de Brasília, pois foi necessária a remoção da bateria do aparelho para efetuar a carga da mesma.
b) O número da linha telefônica é: **4195404292** (quatro, um, nove, cinco, quatro, zero, quatro, dois, nove, dois), conforme obtido pelo serviço *#06#.
c) Os dados da memória do aparelho foram extraídos através da ferramenta forense "*Cellebrite UFED*", e encontram-se listados sob o **Item 1.1** do Anexo ao presente laudo.
d) Os dados do cartão SIM foram extraídos através da ferramenta forense "*Cellebrite UFED*", e encontram-se listados sob o **Item 1.2** do Anexo ao presente laudo.
e) Informações ou outros dados não obtidos através desta perícia poderão ser obtidos junto à operadora de telefonia celular, mediante solicitação judicial.

Para garantir a preservação dos vestígios e permitir eventuais quesitos complementares, sugere-se que o equipamento periciado continue apreendido até o trânsito em julgado da eventual ação penal.

O referido material acompanha a primeira via do presente laudo, acondicionado em embalagem identificada, **lacrada com o nº 2222** e rubricada.

Este laudo foi redigido pelo Perito que o subscreve em primeiro lugar e impresso em 02 (duas) folhas de papel timbrado do Instituto. Acompanha o presente laudo 01 (um) **Anexo** impresso em 05 (cinco) folhas de papel timbrado deste Instituto. E são essas as declarações que em suas consciências têm os Peritos a fazer. E, por nada mais haver, deu-se por findo o exame solicitado que de tudo se lavrou o presente laudo que vai devidamente assinado pelos Peritos.

Luiz Rodrigo Grochocki
PERITO CRIMINAL

Raphael Laércio Zago
PERITO CRIMINAL

FLS. 1
ANEXO AO LAUDO Nº 222222-2

1. EQUIPAMENTO 1
1.1. Aparelho Celular:
1.1.1. Informações Gerais da Captura:

TABELA 1	
Parâmetro	Valor
Fabricante selecionado:	Nokia GSM
Modelo selecionado:	1200/1208
Fabricante detectado:	(Indisponível)
Modelo detectado:	(Indisponível)
Revisão:	(Indisponível)
IMEI:	356402028331070
Código de Segurança:	12345
Início da extração:	25/02/2014 11:31:04
Fim da extração:	25/02/2014 11:31:23
Data/Hora do telefone:	(Indisponível)
Tipo de conexão:	Serial Cable
Versão da UFED:	2.2.5.0 UFED; S/N:5534651
Extração suportada para:	Agenda Telefônica, SMS, Chamadas

1.1.2. Agenda Telefônica gravada na memória do aparelho, da forma como se apresentava por ocasião do exame:

Nenhuma informação foi retornada pelo equipamento forense "Cellebrite UFED".

1.1.3. Chamadas Realizadas gravadas na memória do aparelho, da forma como se apresentava por ocasião do exame:

TABELA 2					
Item	Chamada Realizada	Data	Hora	Duração	Identificação
1	95404318	19/02/2014	15:37:31	00:00:18	Sim A
2	95404318	19/02/2014	15:38:12	(*)	Sim A
3	95404318	19/02/2014	15:39:01	(*)	Sim A
4	95404324	19/02/2014	15:39:45	00:00:32	(nada consta)
5	95404324	19/02/2014	15:40:56	00:00:28	(nada consta)
6	95404318	19/02/2014	15:44:44	(*)	Sim A
7	95404318	19/02/2014	15:50:03	00:00:24	Sim A

(*) Informação não retornada pelo equipamento forense "Cellebrite UFED".

1.1.4. Chamadas Recebidas gravadas na memória do aparelho, da forma como se apresentava por ocasião do exame:

TABELA 3					
Item	Chamada Recebida	Data	Hora	Duração	Identificação
1	95404324	19/02/2014	15:43:53	00:00:36	(nada consta)
2	95404297	20/02/2014	10:16:28	00:00:10	Sim I
3	95404297	20/02/2014	10:18:38	00:00:17	Sim I

FLS. 2
ANEXO AO LAUDO Nº 222222-2

4	95404324	20/02/2014	11:32:28	00:00:20	(nada consta)
5	95404297	20/02/2014	11:37:25	00:00:15	Sim I
6	95404297	20/02/2014	11:37:51	00:00:16	Sim I
7	95404297	20/02/2014	11:38:17	00:00:28	Sim I

1.1.5. Chamadas Não Atendidas gravadas na memória do aparelho, da forma como se apresentava por ocasião do exame:

Nenhuma informação foi retornada pelo equipamento forense "Cellebrite UFED".

1.1.6. Mensagens de Texto (SMS) Recebidas gravadas na memória do aparelho, da forma como se apresentava por ocasião do exame:

TABELA 4				
Item	Mensagem Recebida	Data	Hora	Remetente
1	TIM RECADO: Ainda tem mensagem pra voce! Ligue agora para *100 e escute seus recados.	10/12/2013	19:16:43	Tim
2	INFINITY RECADO: Voce recebeu um novo recado. Para ouvir ligue *100 ou responda com SIM para ler o recado no SMS. Veja os termos de uso em www.tim.com.br	10/12/2013	19:16:43	100
3	Dra Josi, posso ir amanha e quinta no CAIF? Boa noite! Julia!	10/12/2013	20:38:26	0418847413 8
4	BB informa: gerado codigo para liberacao de computador: FFGH, em 10/12/2013, as 21h23. Para sua seguranca, nao informe este codigo a ninguem.	10/12/2013	21:38:09	40040001
5	Oferta dupla! A TIM oferece ao 4199191325 um desconto exclusivo placebo a Internet! Atualize agora p/ o TIM Infinity Web + Torpedo c/ R$0,25 a mais/dia que usar! Envie "OK" p/5000 e ative a oferta: ganhe um premio surpresa garantido e entre no sorteio de R$50.000!	11/12/2013	14:49:52	5000
6	Urgente! Envie "ATIVARMODEM" p/1616 GRATIS p/ativar! Vc foi escolhido p/ Infinity Web Modem: 80MB diarios de internet pinavegar (so 1,99/dia q usar)!	12/12/2013	15:48:06	1616
7	INFINITY RECADO: Voce recebeu um novo recado. Para ouvir ligue *100 ou responda com SIM para ler o recado no SMS. Veja os termos de uso em www.tim.com.br	15/12/2013	15:15:04	100
8	TIM RECADO: Ainda tem mensagem pra voce! Ligue agora para *100 e escute seus recados.	15/12/2013	15:15:04	Tim
9	Voce recebeu ligacoes de: <1>12> 04141999847482 <3> 15:18hs <14/12> 04143992613 31 <1> 18:43hs <13/12> 04141973473 05 <1> 19:09hs	15/12/2013	16:01:54	Te Ligou
10	Agora e de graca! Quando ligar para um numero TIM e ele nao puder atender, deixe um recado! E GRATIS! Aproveite mais esse beneficio da TIM pra voce!	16/12/2013	09:56:05	TIMDicas
11	Voce pode almocar comigo hoje? Beijos	18/12/2013	09:52:55	04199847482
12	Saio daqui às 13h e te mando uma mensagem	18/12/2013	09:57:07	04199847482
13	Estou saindo agora	18/12/2013	12:54:54	04199847482
14	To imprimindo	18/12/2013	16:55:46	04199847482

Figura 2.6: Laudo de Perícia Criminal.

Durante a coleta de dados, muitas vezes o perito ainda não tem uma noção exata da extensão do problema, e, por isso, pode ocorrer de levantar dados em excesso, ou até mesmo de forma não organizada.

Vale ressaltar que é melhor colher informações em excesso do que em falta, pois, no caso de fenômenos criminais, na maioria das vezes, o local do crime é desfeito após o ocorrido.

Ressalta-se que um "crime não é esclarecido pelo poder da polícia, mas pelo poder da metodologia científica" (REIS, 2011).

Podemos observar que a construção do vocabulário para representar os documentos, naturalmente resulta em estruturas de dados geralmente bem mais compactas do que o conteúdo original, podendo conter apenas as informações relevantes ao processamento do sistema de RI.

2.4.2 Organização de Documentos

As principais técnicas de organização de documentos são:

- **Taxonomias**: trata-se de uma organização hierárquica do conteúdo dos documentos. Onde podemos dividir estes em classes, organizando-as de forma hierárquica.
- **Folksonomias**: escolha livre dos termos, também chamados de *tags* (etiquetas). O conjunto de todas as *tags* cria uma organização coletiva plana, na qual os documentos podem ser encontrados usando uma busca baseada em *tags*.

Essas técnicas computacionais de organização dos documentos nos dão a impressão de serem familiares, isso se deve por se aproximarem da maneira como o ser humano assimila, compreende e visualiza essa atividade em seu cotidiano.

2.4.3 Modelos de RI Clássicas

Um modelo de RI é a especificação formal de três elementos: a representação dos documentos, a representação da necessidade da informação por meio de uma expressão de busca e como estes dois elementos serão comparados, a função de busca (BAEZA-YATES; RIBEIRO-NETO, 2013).

A eficiência de um sistema de RI está diretamente ligada ao modelo que ele utiliza, influenciando diretamente em seu modo de operação. Há três modelos que são considerados os modelos clássicos de RI: Modelo Booleano, Modelo Vetorial e Modelo Probabilístico. São considerados modelos clássicos, pelo fato de que suas principais ideias ainda estão presentes na maioria dos sistemas de RI atuais e nos mecanismos de busca da Web.

No **Modelo Booleano** um documento é representado por um conjunto de termos de indexação que podem ser definidos de forma intelectual, ou seja, manual, por profissionais especializados ou automaticamente, utilizando um algoritmo computacional.

As buscas são formuladas por meio de uma expressão booleana composta por termos ligados através dos operadores lógicos AND, OR e NOT, e apresentam como resultado os conjuntos cuja representação satisfazem às restrições lógicas da expressão de busca. A Figura 2.7, mostra os possíveis resultados desse modelo em forma conjuntos.

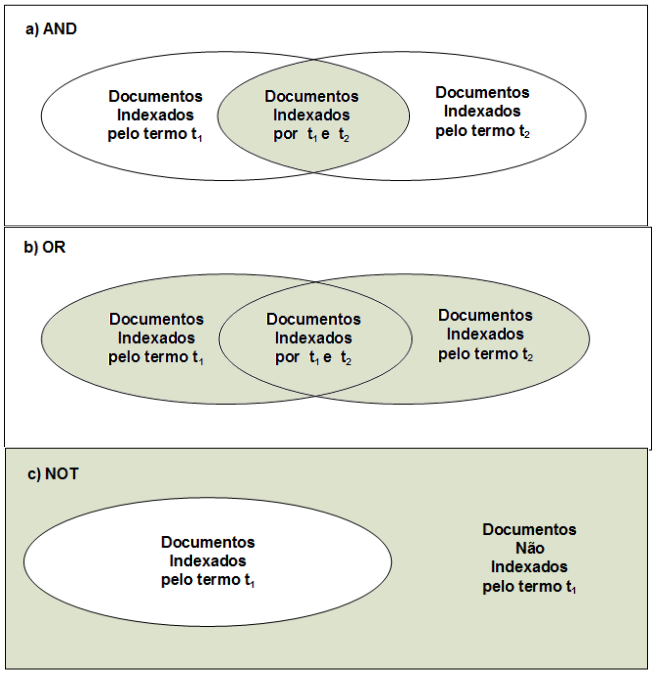


Figura 2.7: Conjuntos que satisfazem às restrições lógicas no Modelo Booleano.

No **Modelo Vetorial** um documento é representado por um vetor onde cada elemento representa o peso, ou relevância, do respectivo termo de indexação para o documento. Cada vetor descreve a posição do documento em um espaço multidimensional, onde cada termo de indexação representa uma dimensão do eixo. Cada elemento do vetor é normalizado de forma a assumir valores entre 0 (zero) e 1 (um). Os pesos mais próximos de 1 (um) indicam termos com maior importância para a descrição dos documentos.

A Figura 2.8, mostra a representação vetorial de um documento com três termos de indexação, atribuído-se os pesos 0.5 ao termo1, 0.4 ao termo2, e 0.3 ao termo3.

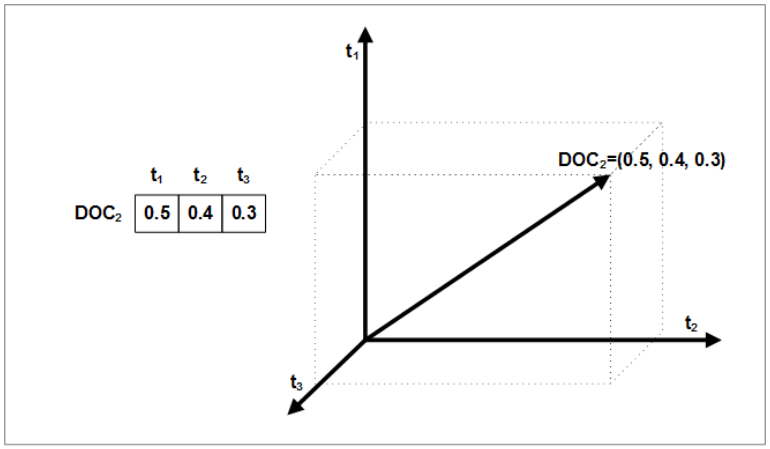


Figura 2.8: Representação vetorial de um documento com três termos.

Nessa estrutura, a utilização da a mesma representação tanto para os documentos como para as expressões de busca permite calcular o grau de similaridade entre os dois documentos ou entre uma expressão de busca e cada um dos documentos. Em um espaço vetorial contendo *N* dimensões, a similaridade (sim) entre um documento d_j e uma expressão de busca *q* pode ser calculada utilizando a Equação 2.1:

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^{N}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \qquad (2.2)$$

Onde $w_{i,j}$ é o peso do i-ésimo termo do documento d_j e $w_{i,q}$ é o peso do iésimo termo da expressão de busca *q*.

O conceito de **Modelo Probabilístico** considera que, dada uma expressão de busca, seu resultado é divido em quatro subconjuntos distintos: o conjunto dos documentos relevantes (Rel), o conjunto dos documentos recuperados (Rec), o conjunto dos documentos relevantes que foram recuperados (RR) e o conjunto dos documentos não relevantes e não recuperados. O conjunto de documentos RR é resultante da interseção dos conjuntos Rel e Rec. A Figura 2.9, mostra a representação probabilística de um conjunto de documentos. Nesse modelo o resultado ideal é que o conjunto RR compreenda todos os documentos relevantes para o usuário, isto é, todo o conjunto Rel.

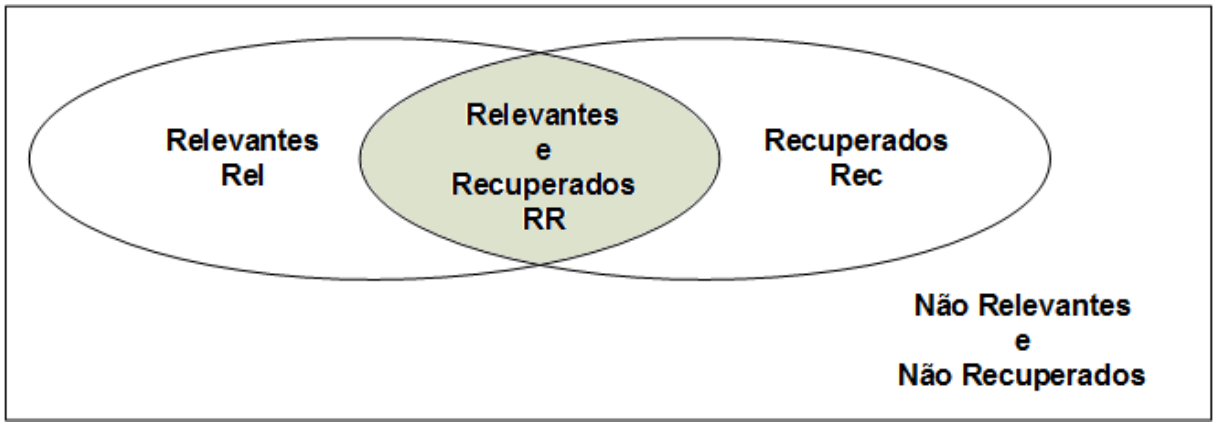


Figura 2.9: Representação probabilística de um conjunto de documentos.

2.4.4 Línguas Naturais

O problema central dos sistemas de Processamento de Línguas Naturais - PLN é a transformação de uma frase de entrada potencialmente ambígua em uma forma não ambígua que possa ser usada internamente por um sistema de computador.

A transposição de uma frase potencialmente ambígua para uma representação interna é conhecida como *parsing* (análise). A palavra parse é derivada do latim: *parse orationis* (parte do discurso).

No PLN, o *parsing* é usualmente um processo de combinar os símbolos de uma frase em um grupo que pode ser substituído por um outro símbolo mais geral. Esse novo símbolo pode, por sua vez, ser combinado em um outro grupo, e assim por diante, até que uma estrutura permitida apareça. (ROSA, 2011).

Existem cinco tipos diferentes de *parsers* (analisadores): casamento de padrões, baseado em gramática, semântico, baseado em conhecimento e analisadores por redes neurais.

2.4.6 Processo de Indexação

Um Índice Invertido é um mecanismo orientado a palavras para a indexação de uma coleção de texto a fim de acelerar a tarefa de busca. A estrutura do índice invertido é composta por dois elementos: o vocabulário e as ocorrências (BAEZA-YATES; RIBEIRO-NETO, 2013).

O vocabulário é o conjunto de termos selecionados para representar o documento, para cada termo do vocabulário o índice armazena os documentos que contém esta palavra. Por essa razão, ele é chamado de índice invertido, pois podemos reconstruir o vocabulário do documento através do índice.

Essa representação simples é largamente utilizada por ser muito rápida, dado que é necessário apenas um acesso ao índice para determinar se um documento contém uma determinada palavra ou não.

Geralmente uma lista de documentos é associada a um determinado termo do índice, representando suas respectivas ocorrências. Dessa forma economiza-se, obtendo uma considerável redução do volume de dados no processo de persistência de dados em uma unidade de armazenamento computacional, quando comparada ao uso de matrizes que exige uma estrutura de dados proporcional ao número de documentos multiplicado pela quantidade total de termos.

A Figura 2.10, mostra o sequenciamento de um processo de indexação que pode ser utilizado na área de RI.

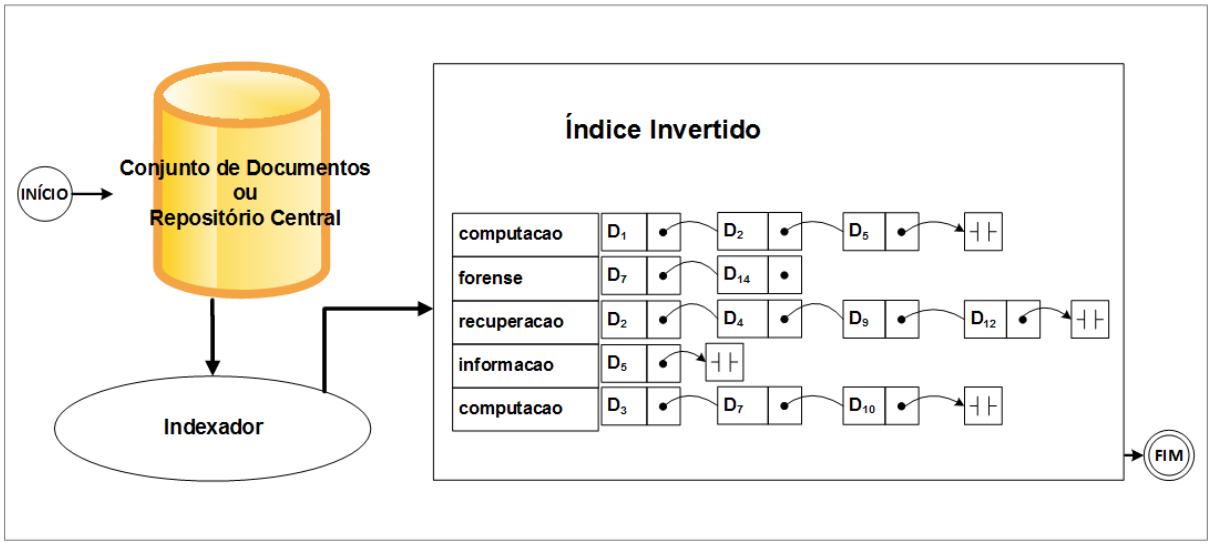


Figura 2.10: Processo de Indexação.

2.5 Descoberta de Conhecimento

Durante a coleta de dados, muitas vezes o perito ainda não tem uma noção exata da extensão do problema, e, por isso, pode ocorrer de levantar dados em excesso, ou até mesmo de forma não organizada.

Vale ressaltar que é melhor colher informações em excesso do que em falta, pois, no caso de fenômenos criminais, na maioria das vezes, o local do crime é desfeito após o ocorrido. Uma premissa da atividade forense é de que um crime não é esclarecido pelo poder da polícia, mas pelo poder da metodologia científica (REIS, 2011).

Todas as informações coletadas podem ser centralizadas em um único Armazém de Dados (*Data Warehousing*) que são banco de dados projetados para tarefas analíticas, que suportam ser alimentados por dados de múltiplas aplicações.

O Armazém de Dados é um processo de armazenamento de dados corporativos de forma que todas as informações da empresa seja integrada em um único repositório a partir do qual o usuário final pode executar consultas, fazer relatórios e realizar análises (SINGH, 1997).

No Armazém de Dados o armazenamento de dados é realizado por uma mistura de tecnologias destinadas à integração efetiva dos bancos de dados operacionais em um ambiente que permite o uso estratégico dos dados (BERSON; SMITH, 1997).

Dentre as tecnologias que podem ser usadas em um Armazém de Dados está incluído o sistema relacional e multidimensional de gerenciamento de banco de dados, arquitetura cliente/servidor, modelagem de metadados, modelagem de repositório, interfaces gráficas e auditoria.

A filosofia da auditoria em tecnologia de informação está calcada em confiança e em controles internos que visam confirmar se os controles internos foram implementados e se são efetivos (IMONIANA, 2005).

A Mineração de Dados, em inglês *Data Mining*, auxilia na obtenção de novas informações e padrões que não poderiam ser encontrados simplesmente pesquisando ou processando dados no Armazém de Dados (ELMASRI; NAVATHE, 2005).

As técnicas de mineração de dados são organizadas para agir sobre bancos de dados com intuito de descobrir padrões, informações e conhecimentos úteis que poderiam, de outra forma, permanecer ignorados e invisíveis em seu meio.

Mineração de Dados é uma parte integrante da Descoberta de Conhecimento em Banco de Dados, também conhecido pelo acrônimo KDD (*Knowledge Discovery in Databases*). KDD é o processo geral de conversão de dados brutos em informações úteis (TAN; STEINBACH; KUMAR, 2009). A descoberta de conhecimento em bancos de dados é composta por seis fases, a saber:

I. Seleção de Dados: levantamento dos itens específicos que representem informações. Como exemplo, no laudo pericial não se necessita armazenar o texto "código do laudo", mas sim os números que o identificam. O trabalho de seleção de dados resultada na estrutura física e definição das tabelas da base de dados;

II. Limpeza: tem por objetivo limpar, eliminar ruídos e inconsistências do conteúdo a ser armazenado, assim como limpar a formatação dos números telefônicos, dados duplicados, garantir a atomicidade da base de dados eliminando registros inconsistentes, tratamento de valores desconhecidos, entre outros;

III. Enriquecimento: incrementa os dados com fontes adicionais de informação, e exemplo de o código de área do número telefônico pode proporcionar o posicionamento geográfico dos contatos. A criação de um novo conjunto de características a partir dos dados originais brutos é conhecida como extração de características (TAN; STEINBACH; KUMAR, 2009);

IV. Transformação ou Codificação: os dados são transformados ou consolidados em formas adequadas para a mineração, trata-se da generalização e normalização dos dados. Pode gerar a redução do volume de dados em casos como, ao invés de armazenar os nomes dos peritos, podemos armazenar o código que os representa;

V. Mineração de Dados: métodos e técnicas inteligentes são aplicados a fim de extrair padrões de dados. O resultado da mineração pode ser descobrir novas regras de associação, padrões seqüenciais ou árvores de classificação;

VI. Apresentação da Informação Descoberta: relatórios, listagens, gráficos e tabelas com o intuito de apresentar os resultados aos usuários.

Para a Apresentação da Informação ser bem sucedida, é necessário que as informações sejam convertidas em um formato visual de modo que suas características, os relacionamentos entre seus itens de dados ou atributos possam ser analisados e reportados.

O objetivo da visualização é a interpretação da informação visualizada por uma pessoa e a formação de um modelo mental das informações. A principal motivação para o uso da

visualização é que as pessoas posam absorver rapidamente grandes quantidades de informações visuais (TAN; STEINBACH; KUMAR, 2009).

O conhecimento gerado pode ser classificado em dedutivo ou indutivo. Quando se utiliza regras lógicas aplicadas nos dados existentes para gerar novas informações ocorre a formação do conhecimento dedutivo. Já o conhecimento indutivo consiste na descoberta de novas regras e padrões nos dados fornecidos.

Vale mencionar que os possíveis resultados que a descoberta de conhecimento pode alcançar fica entre a confirmação do óbvio, a geração de novos conhecimentos ou simplesmente não gerar novos conhecimentos.

2.6 Considerações Finais

Este capítulo apresentou a fundamentação teórica da Computação Forense e das áreas envolvidas no contexto do processo de Investigação Digital. A Computação Ubíqua e os Dispositivos Móveis estão inseridas diretamente no exercício da atividade pericial.

Na sequência foi apresentado os conceitos fundamentais que devem ser considerados na atividade de Coleta de Dados Periciais, na manipulação das evidências forenses e na elaboração do resultado final de todo o trabalho pericial que é o Laudo Pericial.

O estudo teórico na área da Recuperação da Informação é iniciado com a apresentação de como é feita a Representação Computacional de Documentos, bem como eles podem ser organizados. Avançando com o estudo dos Modelos Clássicos de RI: Modelo Booleano, Modelo Vetorial e Modelo Probabilístico. E finaliza com a definição do Processo de Indexação.

Foi apresentado ainda a definição e as fases envolvidas no processo de Descoberta do Conhecimento. Mostrando que a Mineração de Dados é uma das fases desse processo, sendo que o entendimento e uso de todas as fases são fundamentais para um método ser vitorioso.

O próximo Capítulo apresenta os trabalhos relacionados a essa área de pesquisa e considerados no desenvolvimento do método proposto, servindo de apoio para o desenvolvimento da ferramenta.

Capítulo 3

Trabalhos Relacionados

Neste Capítulo são apresentados trabalhos que se relacionam com o escopo desta pesquisa, ou seja, trabalhos voltados especialmente à aplicação conceitos de Recuperação de Informações em documentos textuais não estruturados e na produção de provas digitais. Os trabalhos estão listados na Tabela 3.1 e encontram-se organizados por 3 sub-temas: Recuperação de Informação (RI), Recuperação de Informação Forense (RIForense) com o objetivo de produção de provas digitais e, finalmente, Indexação, sendo o objetivo principal indexar informações sem detalhar a recuperação ou descoberta de conhecimento.

Tabela 3.1: Trabalhos Relacionados.

Trabalho	RI	RI Forense	Indexação
AL-ZAIDY et al., 2012		X	
ANWAR; ABULAISH, 2014		X	
COELHO et al., 2013	X		
DAGHER; FUNG, 2013		X	
DEAN; GHEMAWAT, 2010			X
DALBEN JR; CLARO, 2011		X	
KUECHLER, 2007	X		
MALLMANN et al., 2010		X	
SCHMIDT et al., 2013	X		
SHRESTHA, 2009			X
ZOBEL; MOFFAT, 2006			X

Para melhor organização deste tópico os trabalhos estudados neste capítulo serão apresentados em três grupos de estudo: Recuperação de Informações em documentos não estruturados, Recuperação de Informações na área Forense e Indexação aplicada na Recuperação de Informações.

3.1 Recuperação de Informações em documentos não estruturados

Em 2007 William Kuechler chama a atenção de profissionais que atuam na área de Recuperação de Informações com a publicação do artigo "*Business Applications of Unstructured*" (KUECHLER, 2007). Voltando o foco da área de RI para o aproveitamentos das informações contidas nos documentos usando aplicações (sistemas) capazes de fazer a aquisição e análise de textos não estruturados, usando o acrônimo UTAA - Unstructured Text Acquisition and Analysis.

Nos estudos apresentados em seu artigo foi constatado que 80% das informações contidas nas organizações são armazenadas em documentos não estruturados. A maior parte dos aplicativos usados nas organizações armazena dados em documentos textuais não estruturados, como memorandos, documentos internos, e-mail, páginas da Web organizacionais e comentários de clientes e de pessoal de serviço internos conforme mostra a Figura 3.1.

UTAA Application	Textual Data Source
Business intelligence	Web, industry blogs, online databases
Customer relationship management	Customer feedback, help desk reports
Regulatory compliance	All internally generated electronic documents
Intellectual property management	Web, copyright and patent databases
Call support (help desk applications)	Call documentation, customer feedback, email, online manuals
Accounts payable/receivable analysis	Invoices, customer and vendor correspondence (used frequently with traditional structured data mining and analysis)
Legal department support	Legal databases, specific streams of organizational communications (such as customer communication, internal email)

Figura 3.1: UTAA Aplication and Textual Data Source (KUECHLER, 2007).

Apontou ainda os três principais motivos das aplicações usadas pelas organizações não serem capazes de processar esses documentos, simplesmente desconsiderando as informações contidas em documentos não estruturados, a saber:

- **Primeiro**: os departamentos de gestão da informação consideravam o valor desses dados como duvidosos, ou seja, não eram informações precisas como as encontradas em campos de documentos estruturados.
- **Segundo**: mesmo as organizações que tinham noção do valor das informações contidas nesses documentos na época ainda não existiam técnicas, ferramentas e meios computacionais para realizar a Recuperação de Informações desses documentos.
- **Terceiro**: o processamento de um grande volume de informações era demasiadamente demorado e tinha um custo monetário considerado longe da realidade da maioria das organizações.

Com o objetivo de superar essas limitações e criar um ambiente que possibilite o desenvolvimento de aplicações UTAA, foi proposto uma arquitetura computacional para o desenvolvimento desse tipo de aplicações como mostra a Figura 3.2.

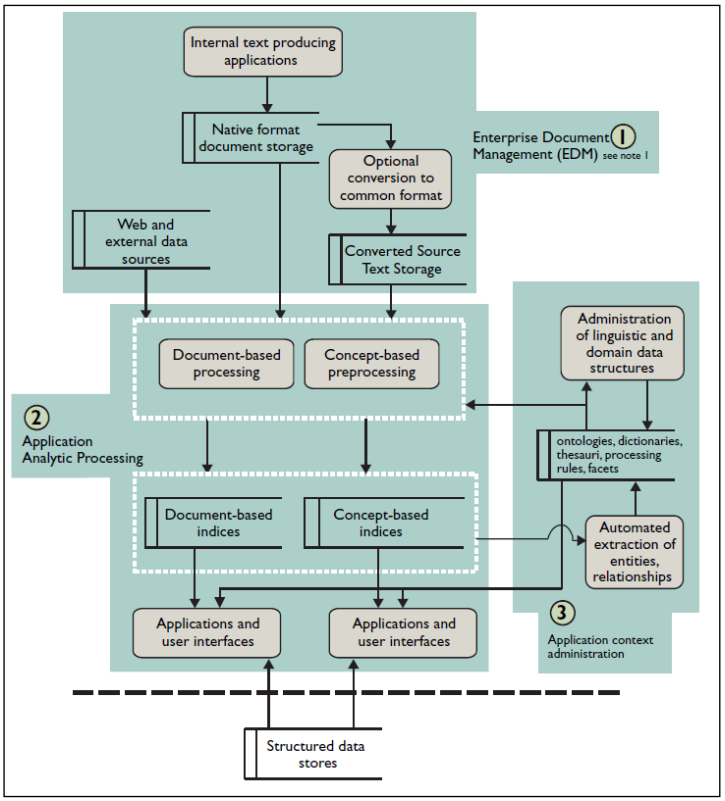


Figura 3.2: General UTAA Archicture (KUECHLER, 2007).

Observa-se que Kuechler (2007) propôs, como ele mesmo denominou, uma espécie de *Middleware* (Camada de Mediação) entre os documentos de texto não estruturado e um contexto de "Aplicação Administrativa" onde os dados são transformados em uma forma estruturada. Ou seja, a ideia proposta é converter documentos não estruturados em documentos estruturados.

Schimit *et al.* (2012) apresentaram um trabalho com o objetivo de identificar e recuperar dados de endereços empresariais em sites da Web (SCHIMIT *et al.*, 2012). Esse ambiente é preenchido com muitos sites (documentos no formato html) contendo informações textuais não estruturadas.

Esse trabalho é realizado por uma sequência de processos aplicados ao conteúdo textual dos sites Web. A Figura 3.3 mostra as etapas de processamento realizado nesse trabalho.

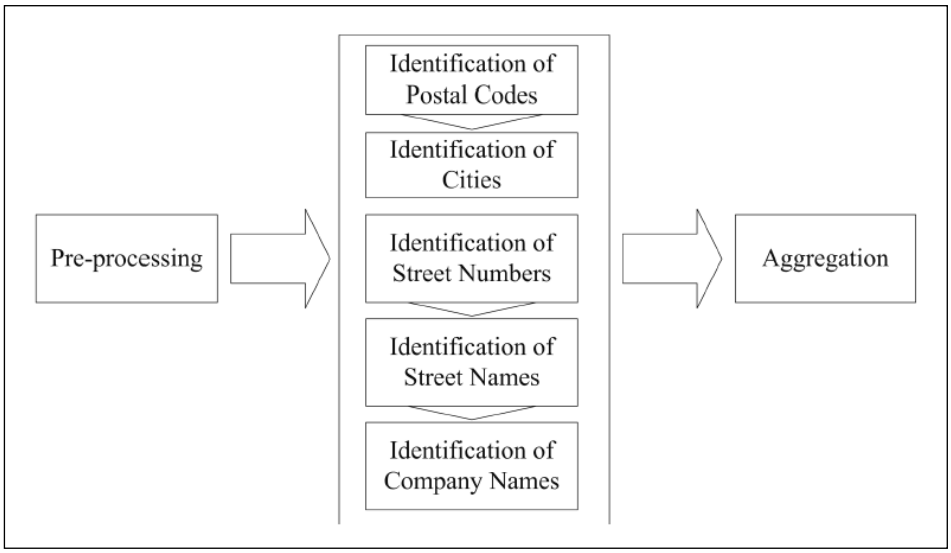


Figura 3.3: Processo de extração de Endereços (SCHIMIT, 2012).

Na etapa de pré-processamento é obtido apenas o conteúdo textual dos documentos html. Para isso é utilizada a ferramenta Apache OpenNLP[13].

A identificação dos atributos código postal, rua, número da rua e nome da empresa são obtidos com o uso de expressões regulares. Como exemplo de expressão regular usada para recuperar o número da rua, tem-se:

- ([0-9]{1,3})([a-zA-Z][0-9]?)?(([+|-])([0-9]{1,3})([azA-Z][0-9]?)?)?.

[13] Apache OpenNLP. Disponível em < https://opennlp.apache.org/ >. Acesso em 15 set. 2015.

As expressões regulares correspondentes aos outros campos não estão disponíveis no artigo publicado.

A identificação do nome das cidades e endereço completo é finalizada com o uso da ferramenta OpenStreetMap[14]. Que é uma ferramenta gratuita e de código aberto que apresenta uma série de recursos para informações geográficas. Consiste na visualização de um mapa sendo enriquecida com informação adicional que varia de nomes de cidades, localizações das caixas de correio, entre outros.

Na fase de agregação dos atributos recuperados é feita uma espécie de combinação entre os campos recuperados e uma verificação de conferência do endereço na ferramenta OpenStreetMap, o resultado obtido no método é lista dos endereços válidos encontrados.

O trabalho de Coelho *et al.* (2013) foi desenvolvido no Hospital Universitário Walter Cantídio da Universidade Federal do Ceará e é parte da pesquisa de Mestrado, apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Federal da Paraíba.

Trata-se de uma pesquisa que tem por objetivo descrever com acuidade os fenômenos relativos à recuperação da informação e a usabilidade da base de dados *Public Medical*, a PUBMED. A Figura 3.4 mostra a página inicial do Sistema de Recuperação de Informações avaliado.

[14] OpenStreetMap. Disponível em < http://www.openstreetmap.org/#map=5/51.500/-0.100 >. Acesso em 15 set. 2015.

Figura 3.4: PUBMED (COELHO *et al*, 2013).

Consiste em um trabalho voltado para a análise da usabilidade da Recuperação de Informações, se propondo a investigar qual o entendimento que os médicos residentes têm sobre o processo de recuperação de informação (COELHO *et al.*, 2013).

A realização dos testes de usabilidade contou com a participação direta dos usuários do sistema no processo de avaliação para detectar problemas de uso do sistema na interação com o usuário.

O intuito de se aplicar um teste de usabilidade com usuários é examinar se o sistema de recuperação está interagindo de forma adequada com os mesmos na realização de tarefas específicas em um contexto das operações de busca por informação na base de dados e apresentação dos resultados.

Uma contribuição que o trabalho "Recuperação da Informação: Estudo da usabilidade na Base de Dados *Public Medical*" apresenta é a comprovação por meio de índices de avaliação da influência exercida pelos elementos de interação sobre os usuários responsáveis pela busca e análise dos resultados da Recuperação de Informações.

Constatou-se que a busca de informações deve apresentar uma forma simples, intuitiva e não deixar margens de dúvidas na hora de interagir com o usuário. Sendo que ao mesmo

tempo, o acesso a recursos e a apresentação dos resultados a exemplo do uso de gráficos e animações não devem ser artefatos meramente decorativos.

Ou seja, é necessário que as informações sejam convertidas em um formato visual de modo que suas características, os relacionamentos entre seus itens de dados ou atributos possam ser analisados e reportados. Criando um ambiente propício para o usuário assimilar o conhecimento gerado.

3.2 Recuperação de Informações na área Forense

Os estudos de trabalhos relacionados na área forense iniciaram pelo trabalho de Mallmann (2010) que propôs um mecanismo para rastreamento de relacionamentos em e-mails que visa encontrar e produzir provas digitais na base de e-mails de um usuário.

Esse mecanismo de rastreamento de relacionamentos em e-mails trata-se de um método que utiliza técnicas de processamento de textos, grafos, métodos de agrupamento e classificação para realizar a classificação de conversações em uma base de e-mails, buscando palavras que definam contextos de mensagens criminosas (MALLMANN *et al.*, 2010).

A Produção de Provas Digitais é dada pelo processamento de 4 fases, a saber:

- Na FASE I, realiza-se um pré-processamento no conteúdo textual dos e-mails, retirando-se elementos existentes no corpo e campo assunto: exclusão de códigos HTML, URLs, símbolos e números. Aplicam-se também as técnicas de case foldering, stop-words e n-grams. Resultando assim, apenas as palavras e termos essenciais para facilitar na determinação das conversações criminosas existentes em uma base de e-mails;
- Na FASE II, o mecanismo agrupa os e-mails pertencentes a uma mesma conversação, ou seja, realiza a extração de conversações da base de e-mails sendo analisada. O vocabulário dos documentos (e-mails) são representados em arquivos de atributos, que são formalizados em formato digital com extensão ARFF (*Attribute-Relation File Format*). A Figura 3.5 mostra os passos da representação dos documentos;

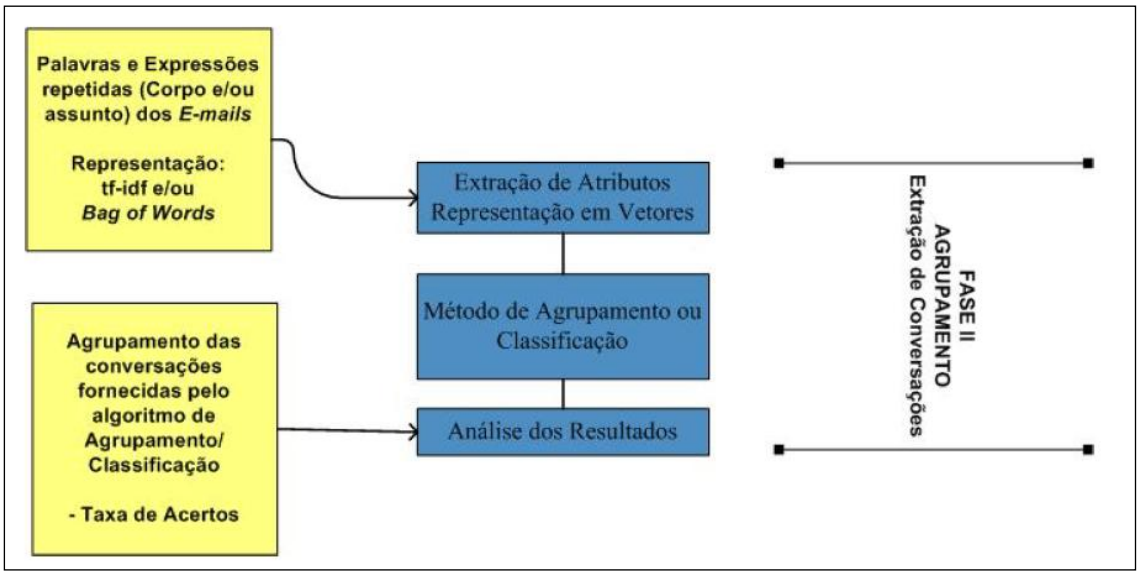


Figura 3.5: Representação dos E-Mails (MALLMANN *et al.*, 2010).

- Na FASE III, todos os e-mails das conversações são classificados com crime, ou não. No caso do trabalho de Mallmann, crimes de assédio moral, ou não. A classificação dos arquivos (teste e treinamento) são submetidos aos métodos de classificação (SVM – kernels Polynomial, Radial e Sigmoid, NB e DT) no uso do software WEKA[15];
- A FASE IV apresenta os resultados alcançados pelo mecanismo. Nesse momento os usuários podem visualizar o Nexo Causal que realize a ligação entre crime e respectivo autor.

A construção de grafos proporciona melhor entendimento de conversações criminosas, rastreamento e visualização das ligações existentes entre os documentos. A Figura 3.6 mostra um Grafo Criminoso gerado nos experimentos do trabalho.

[15] WEKA: Disponível em <http://www.cs.waikato.ac.nz/ml/weka/>. Acesso em 10 set 2015.

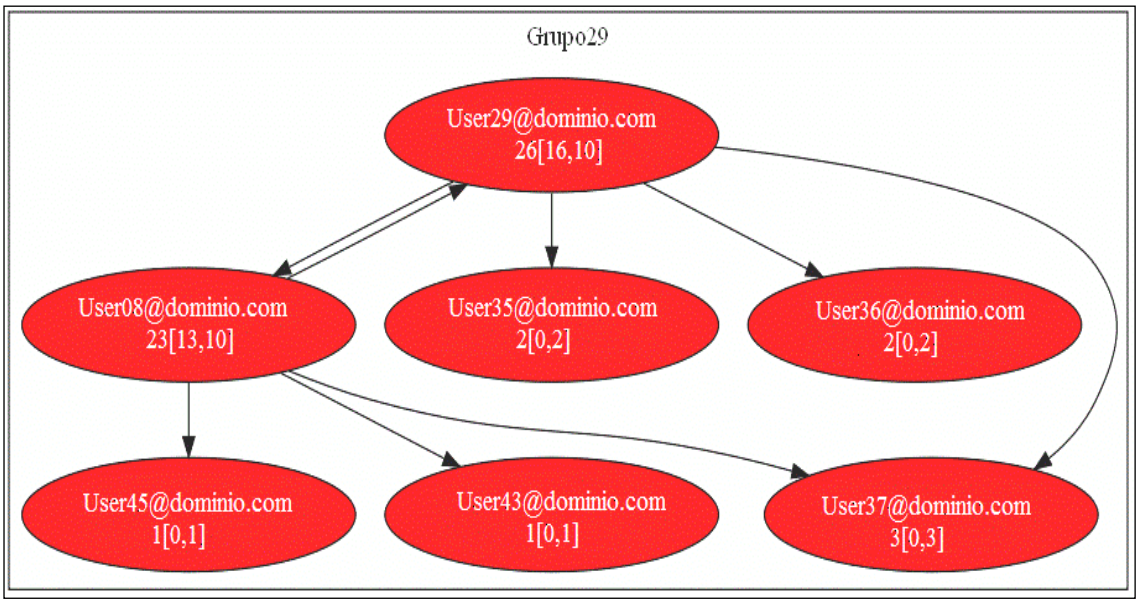


Figura 3.6: Grafo Criminoso (MALLMANN *et al.*, 2010).

Os resultados do trabalho de Mallmann facilitam a construção do nexo causal, fazendo com que um perito consiga com maior agilidade e precisão comprovar a existência de relação entre uma conduta e o suposto infrator sendo investigado.

Em 2011, Dalben Jr & Claro (2011) publicam um artigo de reconhecimento textual de nomes de pessoas e organizações na computação forense com o objetivo deste trabalho de fazer o Reconhecimento de Entidades Mencionadas (REM), ou seja, identificar e classificar entidades (nomes de pessoas, organizações, locais, etc.) contidas em textos não estruturados.

A Figura 3.7 mostra o fluxograma de sequenciamento das etapas de extração e análise aplicadas na tarefa de automatização do REM proposta por Dalben Jr & Claro (2011).

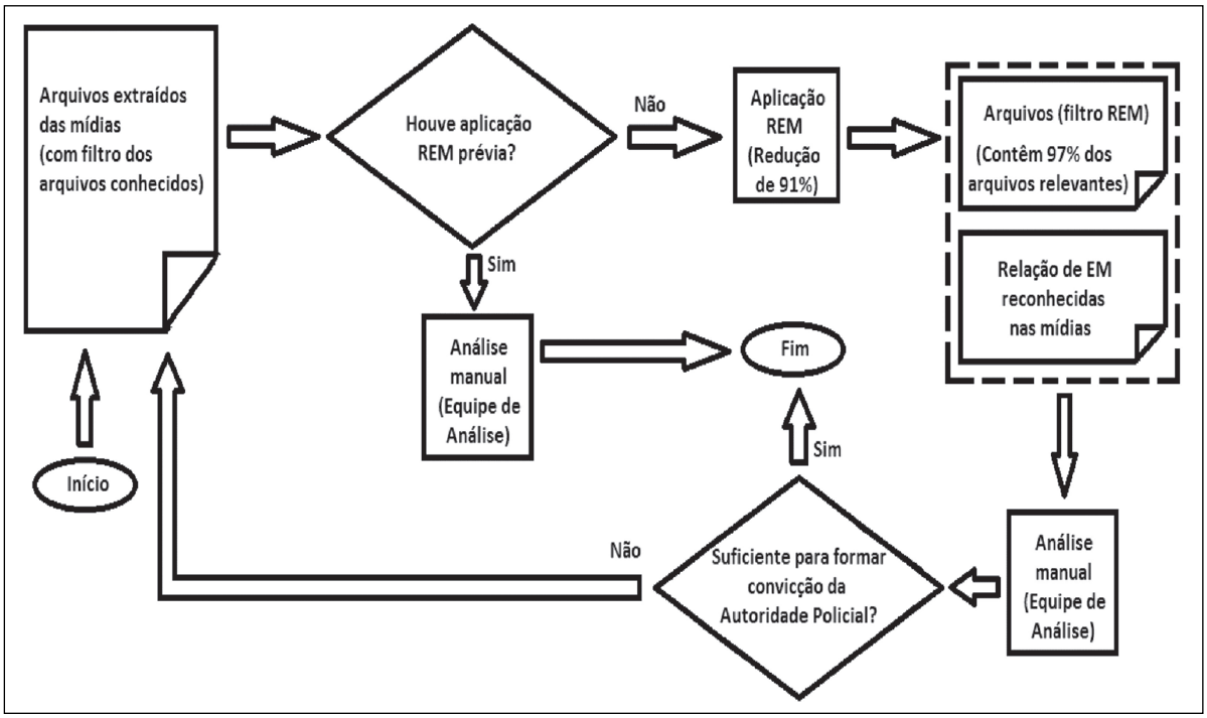


Figura 3.7: Etapas de extração e análise do conteúdo (DALBEN JR; CLARO, 2011).

O REM pode ser caracterizado como um problema de classificação cujo objetivo é atribuir para cada valor de entrada uma classe, identificada por um nome de Entidade Mencionada. Na forma clássica de REM, os valores de entrada são representados pelas palavras ou *tokens* de um texto e a entidade representa a classe ou rótulo do *token* associado (JUNIOR; CLARO, 2011).

Os experimentos realizados para verificar a eficácia desse método foram obtidos com a aplicação do Rembrandt em arquivos contidos em mídias apreendidas em operações da Polícia Federal do Brasil.

Rembrandt é um sistema de REM determinístico, baseado em regras gramaticais manuais e em informações extraídas da Wikipédia. Nesse sistema cada documento lido é submetido a uma sequência de processos de etiquetação sucessivos até alcançar a versão final. O funcionamento do Rembrandt pode ser dividido em três grupos:

1) Divisão dos textos em sentenças e palavras;
2) Classificação das entidades candidatas resultantes da etapa anterior;

3) Repescagem das EMs sem classificação.

Os resultados dos experimentos realizados mostraram que, em média, somente 8,6% dos arquivos do tipo documento contidos nas mídias apreendidas fazem referência a nomes de pessoas ou organizações e que 99,9% dos arquivos julgados relevantes no processo de análise manual estão contidos nesse conjunto de arquivos.

Com isso Junior & Claro (2011) conseguiram comprovar que a utilização do REM no contexto forense minimiza o tempo de análise manual do conteúdo das mídias apreendidas, sem prejuízo à qualidade das informações analisadas.

O artigo " *Mining Criminal Network from Instructured Text Documents*" de Al-Zaidy *et al.* (2012), apresenta um método sistemático para descobrir possíveis redes criminosas e relacionamentos indiretos a partir de uma coleção de documentos de texto obtidos a partir da máquina de um suspeito.

Esse trabalho tem por objetivo gerar uma Rede Criminal considerando dados coletados na análise forense em documentos de dados textuais não estruturados, tais como e-mails, mensagens de chat e documentos de texto. Documentos que muitas vezes contêm informações valiosas sobre as redes sociais dos suspeitos (AL-ZAIDY *et al.*, 2012).

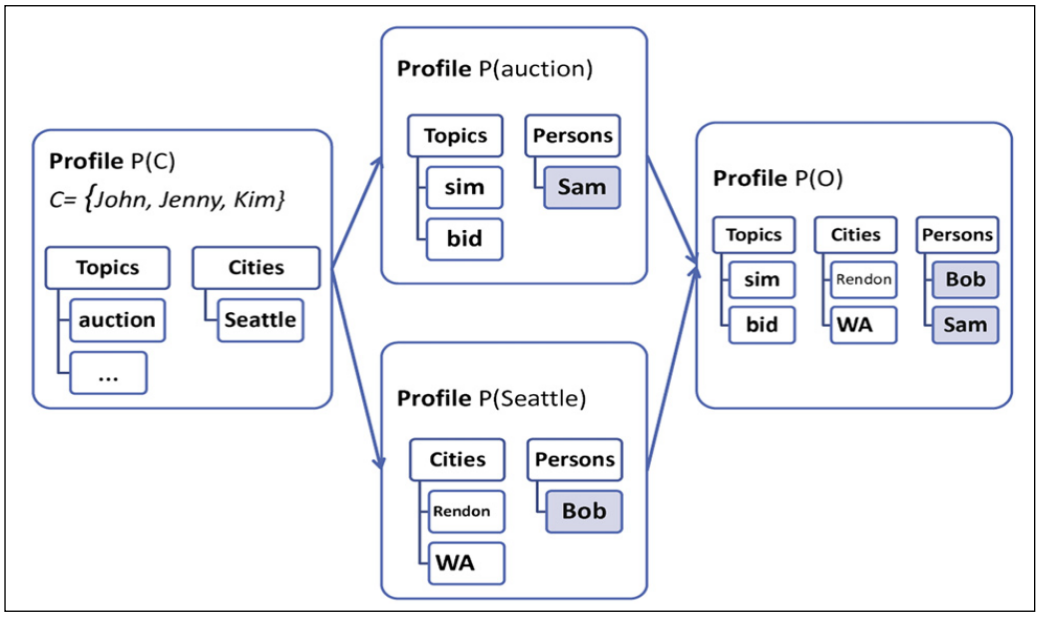


Figura 3.8: Hipótese de Relacionamento Indireto (AL-ZAIDY *et al.*, 2012).

A Figura 3.8 mostra a hipótese de uma possível rede de relacionamento indireta por ter detectado o presença de determinados nomes de pessoas em documentos distintos.

Al-Zaidy *et al.* (2012) mencionam que para realizar esse processo o primeiro passo é ler os documentos de texto investigados e extrair os nomes pessoais deles. A tarefa de extração dos nomes é seguida por um processo de normalização para eliminar nomes duplicados que se referem à mesma pessoa gerando uma estrutura que é denominada de *profile (perfil)*. O próximo passo é descobrir as redes criminosas dos perfis extraídos.

De acordo com seus autores um perfil de um conjunto C, indicado por P(C), é definido por um conjunto de vetores V_{X_1}, V_{X2}, ..., V_{X_n}, onde n indica o número de termos considerados. Cada vetor V_{X_i}, apresenta o comprimento lx_i, onde lx_i é o número de termos de tipo semântico x_i. Esse processo é representado pela Equação 3.1, a saber:

$$V_{X_i} = \begin{bmatrix} t_1, f_{X_i}(t_1) \\ t_2, f_{X_i}(t_2) \\ \dots \\ t_{lx_i}, f_{X_i}(t_{lx_i}) \end{bmatrix} \tag{3.1}$$

Os experimentos de Al-Zaidy *et al.* (2012) foram realizados em uma base de dados denominada de *EnronSmall*. Os resultados obtidos da aplicação do método composta por 24 contas de chat distintas, conseguiu montar 32 perfis e 2 redes criminosas conforme é mostrado na Figura 3.9.

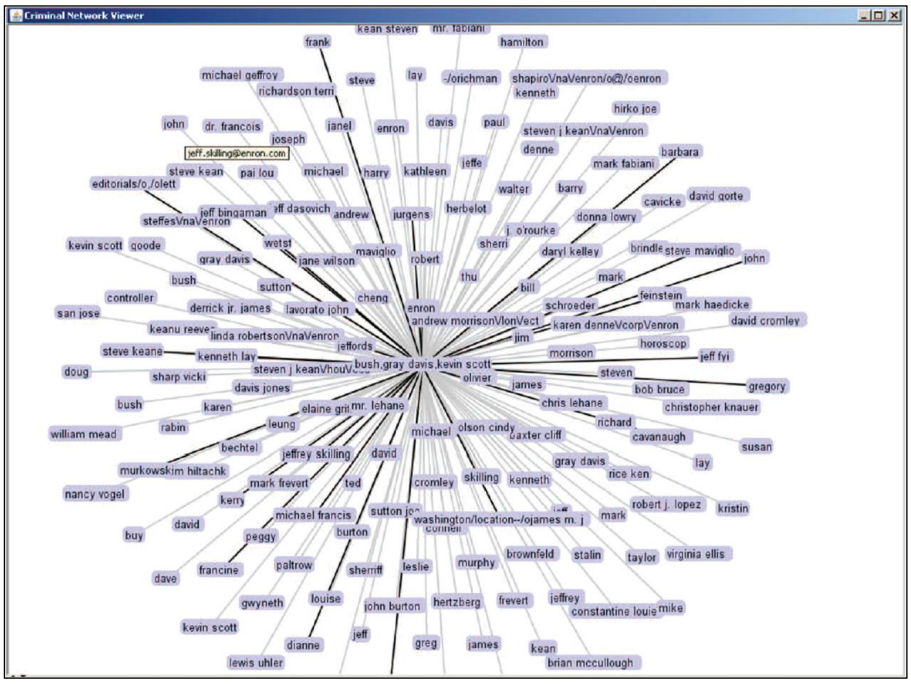


Figura 3.9: Rede Criminosa detectada na base *EnronSmall*.

O centro da Figura 3.9 é composta por dois nós, denotados por S1 e S2, representando o log de duas contas de chat de propriedade de um mesmo suspeito encontradas no computador retido pela Polícia Provincial de Québec.

O artigo "*Subject-based Semantic Document Clustering for Digital Forensic Investigations*", de Dagher e Fung (2013) apresenta um modelo de agrupamento de documentos que pode ser usado como ferramenta de apoio na Polícia Provincial de Québec.

Com o objetivo de permitir que um investigador agrupe documentos de um determinado sujeito, através de assuntos (termos), que se encontrem distribuídos em um computador ou local de armazenamento (DAGHER; FUNG, 2013).

Baseado na semântica textual, o método seleciona termos ou palavras para formar agrupamentos de documentos por similaridade, conforme mostra a Figura 3.10.

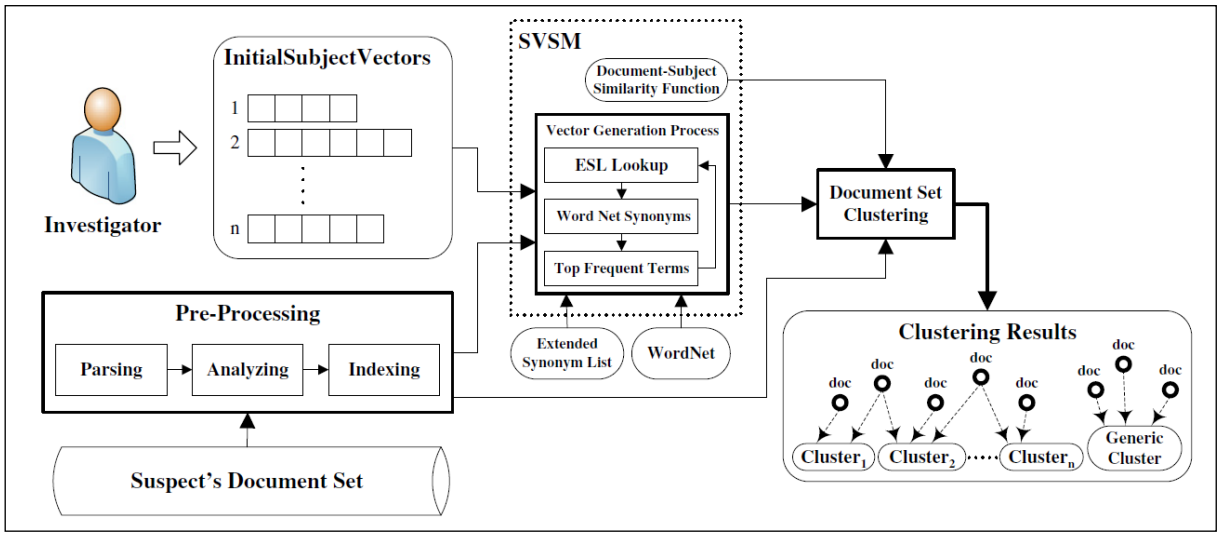


Figura 3.10: Subject-based Semantic Document Clustering (DAGHER; FUNG, 2013).

As principais contribuições deste trabalho é a representação dos documentos forenses estendendo o modelo vetorial e a apresentação de um novo método de agrupamento de documentos por similaridade semântica.

A representação dos documentos forenses é feita estendendo o modelo vetorial com o uso de uma Lista Estendida de Sinônimos (*Extended Synonym List - ESL*) composta por palavras, termos e seus respectivos significados forenses comumente utilizadas na criminalística e que nem sempre são referenciadas em dicionários.

O processo de agrupamento (*clustering*) semântico do resultado obtido no processo da Recuperação de Informações é realizado por meio da aplicação de uma função denominada de *Document-subject Similarity Function* representada pela Equação 3.2.

$$\mathrm{Sim}(\mathrm{d}, \mathrm{S_i}) = \frac{1}{\psi_d} \sum_{t \in \mathrm{T}} \Omega_{t,S_i} \times \Omega_{t,d} \qquad (3.2)$$

Sendo que Ω_{t,S_i} e $\Omega_{t,d}$ correspondem ao tamanho de termos (t) contidos na lista de sinônimos (S) e no documento (d) respectivamente.

Com isso o trabalho apresentado por Dagher & Fung usou a mineração de dados para apresentar um método que pode ser usado no apoio dos serviços de Investigação Digital da Polícia Provincial de Québec.

O artigo "*A social graph based text mining framework for chat log*" de Anwar & Abulaish (2014) apresenta uma proposta de framework para gerar gráficos sociais baseados na mineração de texto com o objetivo de identificar evidências digitais a partir de dados extraídos de logs de bate-papo.

A Recuperação de Informações é usada para descobrir ligações de interesses entre os usuários e seus laços sociais contidos nos dados de conversação de usuários em conversas de bate-papo do aplicativo MSN Messenger.

A sequência de funcionamento do método é mostrada na Figura 3.11. Passando pelas etapas de extração dos dados, normalização, extração do vocabulário, extração de termos chave, identificação dos grupos de usuários e construção do gráfico social.

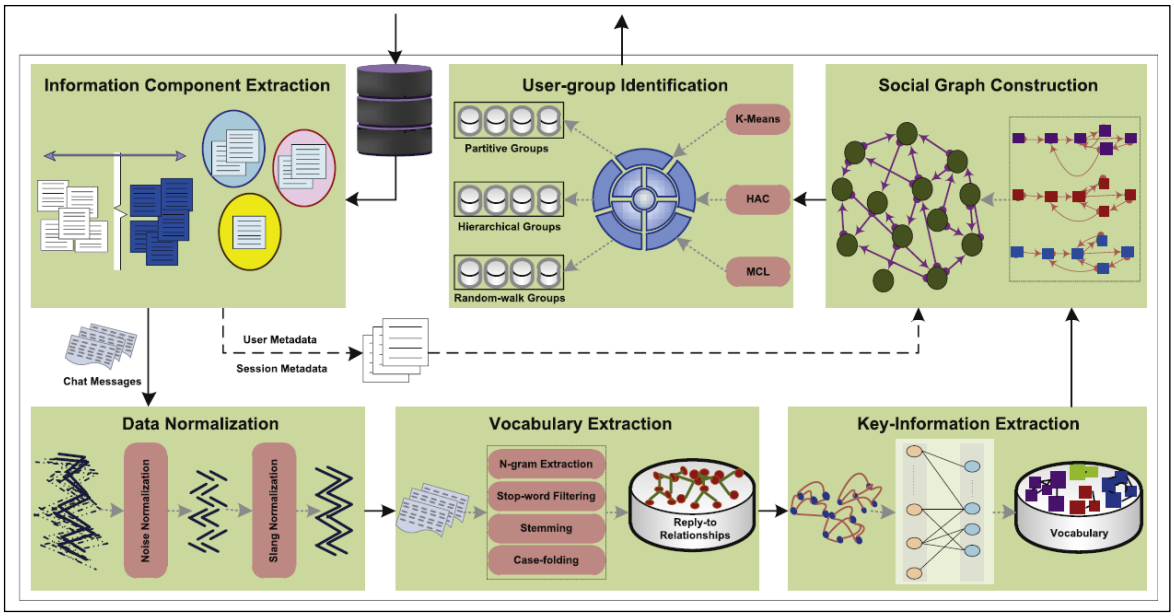


Figura 3.11: Geração de Gráficos Sociais (ANWAR; ABULAISH, 2014).

É apresentado o método *"Hyperlink-Induced Topic Search* - HITS" responsável pela extração de termos chave, o qual é o processo de identificação de termos que apresentam valores representativos entre os termos existentes no vocabulário extraído (ANWAR; ABULAISH, 2014).

A tarefa de extração de termos chave tem como objetivo extrair termos chave de informação a partir de conversas e calcular os valores dos recursos. As informações de chave referem-se a três coisas (vocabulário e termos, os usuários participantes, e sessões de bate-papo) caracterizando dessa forma o destaque dos dados (termos) contidos na conversa.

O método de normalização do algoritmo HITS é obtido aplicando a equação 3.3, a saber.

$$norm_{HS} = \sqrt{\sum_i (HS^t(\tau_i))^2} \tag{3.3}$$

Sendo que $HS^t(\tau_i)$ corresponde ao vetor contendo os escores dos itens vocabulário e termos, os usuários participantes, e sessões de bate-papo, dos termos contidos no vocabulário do documento.

O gráfico social apresentado nos experimentos do método é mostrado na Figura 3.12

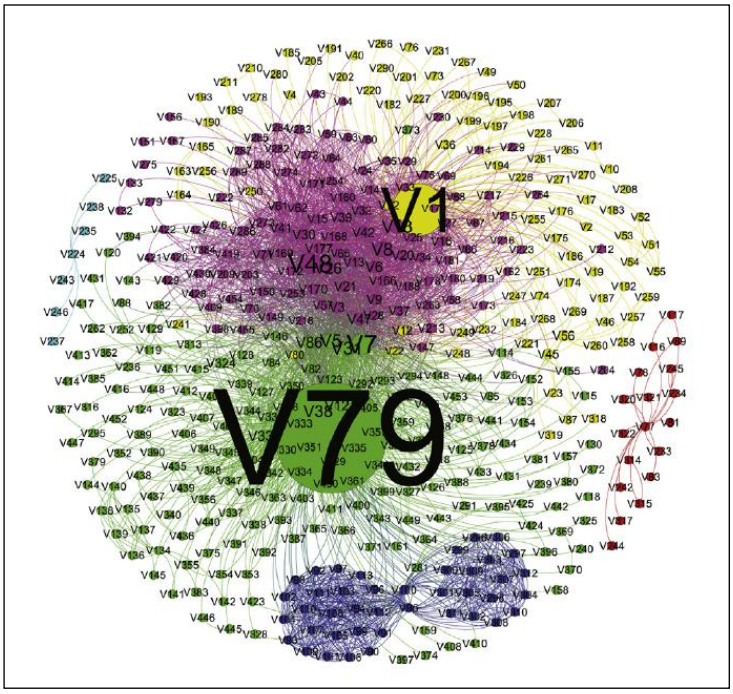


Figura 3.12: Gráfico Social gerado usando o HITS. (ANWAR; ABULAISH, 2014).

O trabalho de Anwar & Abulaish (2014) fez uso dos metadados do aplicativo MSN Messenger para obter informações sobre as interações entre os usuário, e não apenas o conteúdo das mensagens, ou seja, as informações sobre suas conversas textuais.

Entre as dificuldades enfrentadas, os autores mencionam que mensagens de chat geralmente contêm grandes quantidades de ruído e segue estilos informais sem se preocupar muito sobre a ortografia e correção gramatical.

A contribuição científica deste trabalho reside no método de construção do gráfico social explora tanto os metadados quanto o conteúdo das mensagens para modelar os usuários e seus laços usando um grafo ponderado. As relações entre diferentes usuários são extraídos pela sua participação em comum nas sessões de conversa. Cada interesse do usuário é representado por um conjunto de termos chave extraídos do conteúdo da mensagem usando o algoritmo HITS.

3.3 Indexação aplicada na Recuperação de Informações

Nesse tópico foram estudados trabalhos que usam a Indexação como um mecanismo de armazenamento e recuperação de informações. Sendo que o principal propósito da recuperação de informações é ajudar usuários a encontrar informações de seu interesse.

O uso de índices invertidos na recuperação de informações é amplamente difundido e usado desde os primórdios dos estudos nessa área. Com diversos métodos usando essa estrutura de dados como um mecanismo de armazenamento, busca e análise de documentos textuais. Um bom exemplo disso é a publicação "*Complete inverted files for efficient text retrieval and analysis"* (BLUMER *et al.*, 1987).

O atual armazenamento de dados dispersos em ambientes distribuídos e processamento paralelos, mais recente em Big Datas, exigem o emprego de ferramentas computacionais a exemplo de máquinas de busca da Web e outras que indexam terabytes de dados e servem centenas ou milhares de consultas por segundo.

Isso nos remete a um ambiente que exige a manipulação de um grande volume de dados, apresentando uma variabilidade nos formatos de armazenamento das informações. E ainda com uma grande quantidade de usuários acessando as informações ao mesmo tempo, esperando a melhor velocidade de acesso e resposta possível.

Por incrível que possa pareça os Índices Invertidos são estruturas que se demonstram próprias para esse cenário de armazenamento de arquivos em arquiteturas distribuídas, processamentos paralelos e grandes volumes de dados.

Zobel & Moffat (2006), publicaram um amplo estudo do uso de índices invertidos em mecanismos de busca recuperação de informações em grandes volumes de documentos textuais no artigo "*Inverted Files for Text Search Engines*".

O Índice Invertido é um documento completo para armazenar uma base de dados textual. A exemplo de um Índice Invertido com frequência onde cada entrada de um determinado termo *t* é composto por uma frequência *f* e uma lista de pares, que por sua vez consiste de um identificador do documento *d* e sua respectiva frequência do termo no documento. (ZOBEL; MOFFAT, 2006).

term t	f_t	Inverted list for t
and	1	$\langle 6,2\rangle$
big	2	$\langle 2,2\rangle$ $\langle 3,1\rangle$
dark	1	$\langle 6,1\rangle$
did	1	$\langle 4,1\rangle$
gown	1	$\langle 2,1\rangle$
had	1	$\langle 3,1\rangle$
house	2	$\langle 2,1\rangle$ $\langle 3,1\rangle$
in	5	$\langle 1,1\rangle$ $\langle 2,2\rangle$ $\langle 3,1\rangle$ $\langle 5,1\rangle$ $\langle 6,2\rangle$
keep	3	$\langle 1,1\rangle$ $\langle 3,1\rangle$ $\langle 5,1\rangle$
keeper	3	$\langle 1,1\rangle$ $\langle 4,1\rangle$ $\langle 5,1\rangle$
keeps	3	$\langle 1,1\rangle$ $\langle 5,1\rangle$ $\langle 6,1\rangle$
light	1	$\langle 6,1\rangle$
never	1	$\langle 4,1\rangle$
night	3	$\langle 1,1\rangle$ $\langle 4,1\rangle$ $\langle 5,2\rangle$
old	4	$\langle 1,1\rangle$ $\langle 2,2\rangle$ $\langle 3,1\rangle$ $\langle 4,1\rangle$
sleep	1	$\langle 4,1\rangle$
sleeps	1	$\langle 6,1\rangle$
the	6	$\langle 1,3\rangle$ $\langle 2,2\rangle$ $\langle 3,3\rangle$ $\langle 4,1\rangle$ $\langle 5,3\rangle$ $\langle 6,2\rangle$
town	2	$\langle 1,1\rangle$ $\langle 3,1\rangle$
where	1	$\langle 4,1\rangle$

Figura 3.13: Índice Invertido com frequência de termos (ZOBEL; MOFFAT, 2006).

A transformação de um documento textual em um índice invertido naturalmente traz consigo, no mínimo a redução da magnitude máxima do conteúdo armazenado. Ou seja, naturalmente é feito uma espécie de compressão de dados, por exemplo, ao invés de conter N vezes o termo em um documento, o termo consta apenas uma vez com seu respectivo valor de frequência.

O trabalho "*Text Mining with Lucene and Hadoop: Document Clustering With Feature Extraction*" de Dilpesh Shrestha (2009) demonstrou que é possível fazer a Recuperação de Informações e o agrupamento de documentos com o uso da mineração de textos suportando o processamento paralelo, escalonamento de máquinas e os pilares do Big Data (volume, variabilidade e velocidade).

O agrupamento se deu por meio da extensão do algoritmo de k-means na implementação MapReduce do Apache Hadoop[16]. A construção dos índices dos documentos foi obtida com o uso da ferramenta de indexação de documentos textuais Apache Lucene[17].

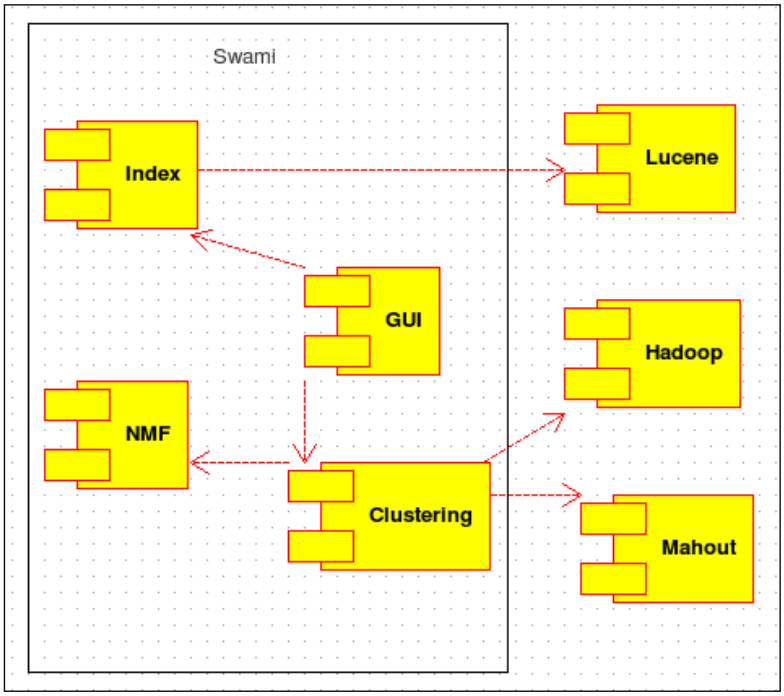


Figura 3.14: Diagrama de Componentes (SHRESTHA, 2009).

O processo de agrupamento (*clustering*) nessa adaptação é realizado por meio da aplicação de uma adaptação do algoritmo de *k*-Means denominada de *Spherical kmeans* representada pela Equação 3.4.

$$L = \sum_{i=1}^{k} \sum_{x_j \in s_i} x_j^T \mu_i \qquad (3.4)$$

Sendo que k são agrupamentos de s_i, i = 1, 2, ..., k. E μ_i são os centróides (pontos centrais) de x_j e s_i.

[16] Apache Hadoop: Disponível em < http://hadoop.apache.org/>. Acesso em 25 set 2015.
[17] Apache Lucene : Disponível em < https://lucene.apache.org/core/>. Acesso em 25 set 2015.

De acordo com a Apache Software Foundation o projeto Apache Hadoop[18] é um framework de desenvolvimento de software de código aberto para computação de forma escalável, confiável e distribuída.

Dessa forma o trabalho Shrestha (2009) conseguiu realizar o agrupamento de documentos textuais não estruturados em uma ambiente Big Data, com o uso das ferramentas Apache Hadoop e Apache Lucene. Proporcionando a Recuperação de Informações em grandes volumes de dados em um ambiente escalável, confiável e de forma distribuída.

O conjunto de ferramentas disponibilizadas pelo Apache Hadoop compõem um framework que permite o processamento distribuído de grandes conjuntos de dados em clusters de computadores utilizando um modelo de programação denominado MapReduce.

Dean & Ghemawat (20010) publicaram o artigo "*MapReduce: a flexible data processing tool*" reafirmando os conceitos da implementação do modelo de desenvolvimento MapReduce.

Além de suporte a índices, sistemas heterogêneos para entrada e saída de arquivos, e dados estruturados e não estruturados. O MapReduce mostrou-se como uma ferramenta extremamente efetiva e eficiente para processamento e geração de dados em larga escala apresentando uma série de benefícios em relação aos Banco de Dados Paralelos (DEAN; GHEMAWAT, 2010).

3.4 Considerações Finais

Este Capítulo apresentou os principais trabalhos relacionados inseridos nas áreas de Recuperação de Informações em documentos não estruturados, Recuperação de Informações na área Forense e Indexação aplicada na Recuperação de Informações.

Vale mencionar que trabalhos de Recuperação de Informações que aplicam técnicas de representação de documentos no modelo vetorial e probabilístico podem apresentar situações problemáticas desses modelos a serem solucionadas, a exemplo da alta dimensionalidade.

O Método Proposto nesta dissertação é focado no desenvolvimento de uma ferramenta voltado à visão centrada no computador, realizando o processo de Recuperação de Informações e Descoberta do Conhecimento em um conjunto de documentos de texto não estruturado.

[18] Apache Hadoop: Disponível em < http://hadoop.apache.org/>. Acesso em 25 set 2015.

Entende-se que a Recuperação de Informações com visão centrada no computador é direcionada na construção de índices eficientes, no processamento de consultas e no desenvolvimento de algoritmos com o intuito de melhorar os resultados.

O próximo Capítulo apresenta com detalhes o método proposto nesta dissertação, incluindo detalhamento computacional de todas as fases. Além do pré-processamento, é apresentado as fases de composição do conjunto de documentos, representação dos documentos, processo de indexação e por fim a fase de apresentação dos resultados obtidos.

Capítulo 4

Método Proposto

O visível crescimento do volume de informações gerado no planeta, mencionado no Capítulo 1, apresenta um forte reflexo em todas as áreas de conhecimento e ramos profissionais, e em especial para profissionais que atuam diretamente na manipulação desses dados.

Na área de Computação Forense, as atividades de exame pericial e elaboração de laudos, exigem que nada seja desprezado. Sendo essencial a consideração de toda e qualquer informação disponível, a fim de coletar o que for necessário para a elaboração de laudos de perícia criminal. Evidências forenses consistentes normalmente estão armazenadas em documentos não estruturados, podendo conter anexos de diversas formas e formatos de arquivo.

Neste contexto surge a necessidade do desenvolvimento de técnicas computacionais para a exploração dados proporcionando a recuperação, visualização e conhecimento de informações contidas nos mesmos. Processos que permitam a exploração desse grande volume de dados se tornam cada vez mais fundamentais no processo de recuperação de informações.

Atualmente existe uma considerável quantidade de sistemas de recuperação de informações. Contudo um sistema de RI nunca irá atender a todas as necessidades de todos os usuários, deixando em aberto lacunas a serem preenchidas nessa área de conhecimento.

O Sistema de Cruzamento de Registros Telefônicos - SiCReT, surgiu da necessidade das Seções de Informática Forense utilizarem um procedimento computacional capaz de realizar o cruzamento de informações contidas nos laudos periciais de dispositivos móveis.

O SiCReT pode ser definido como uma ferramenta computacional capaz de realizar o processo de recuperação, detecção e visualização de cruzamentos existentes entre termos de interesse de um determinado conjunto de laudos periciais de dispositivos móveis.

A Figura 4.1 mostra o sequenciamento no método proposto. Pode-se observar que o processo é iniciado com a transformação de um conjunto de arquivos a ser processado em um conjunto de documentos textuais. Na sequência é criado o vocabulário dos respectivos documentos, fazendo a representação computacional utilizando os termos de interesse. O processo de indexação resulta em uma estrutura de dados de índice invertido, possibilitando a detecção e representação visual dos cruzamentos gerados.

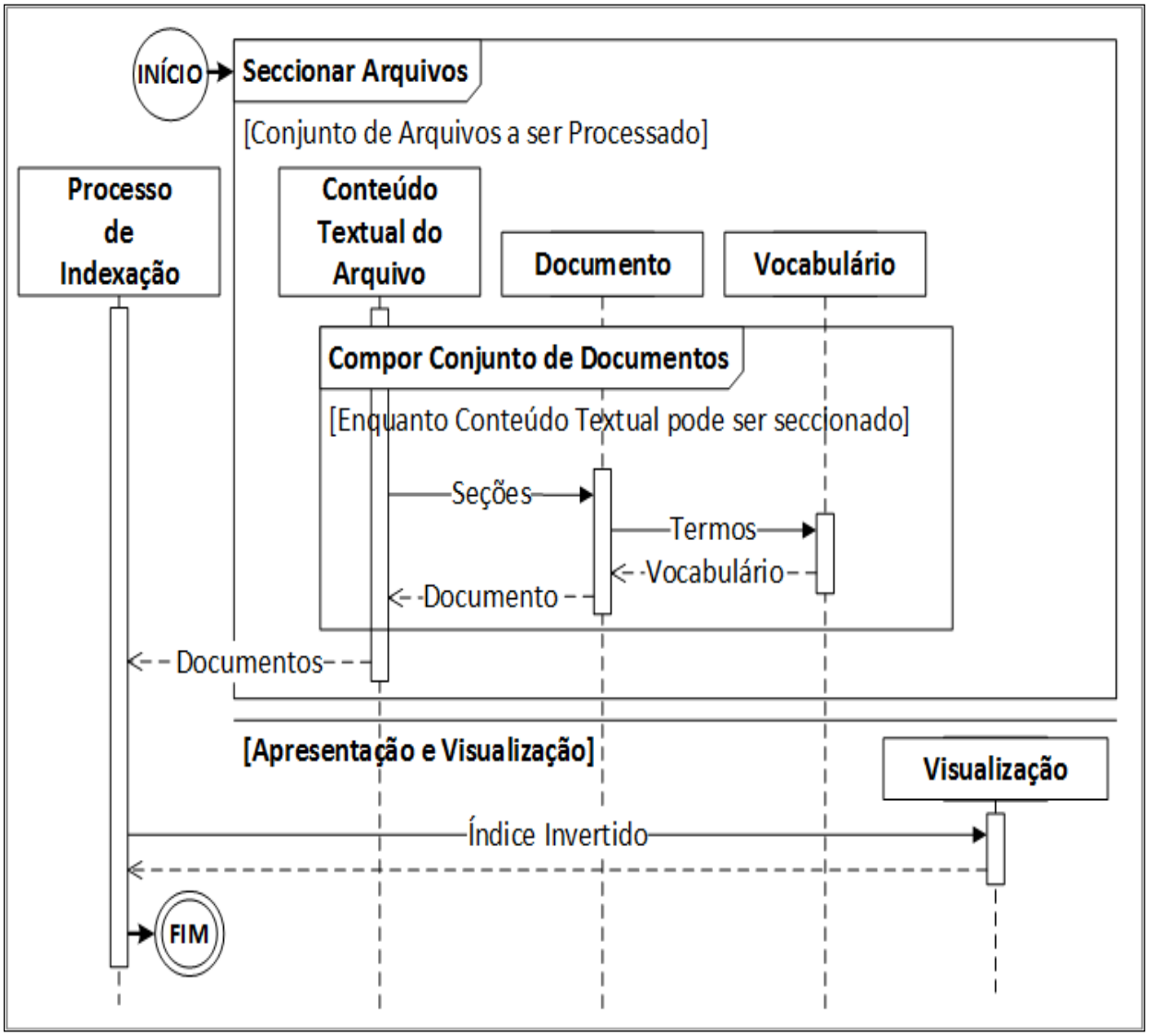


Figura 4.1: Sequenciamento para Aplicação do Método.

A Figura 4.2 mostra a Visão Geral do Método Proposto, que evolve as fases de Composição do Conjunto de Documentos, Representação dos Documentos, Processo de Indexação e Apresentação e Visualização das Informações.

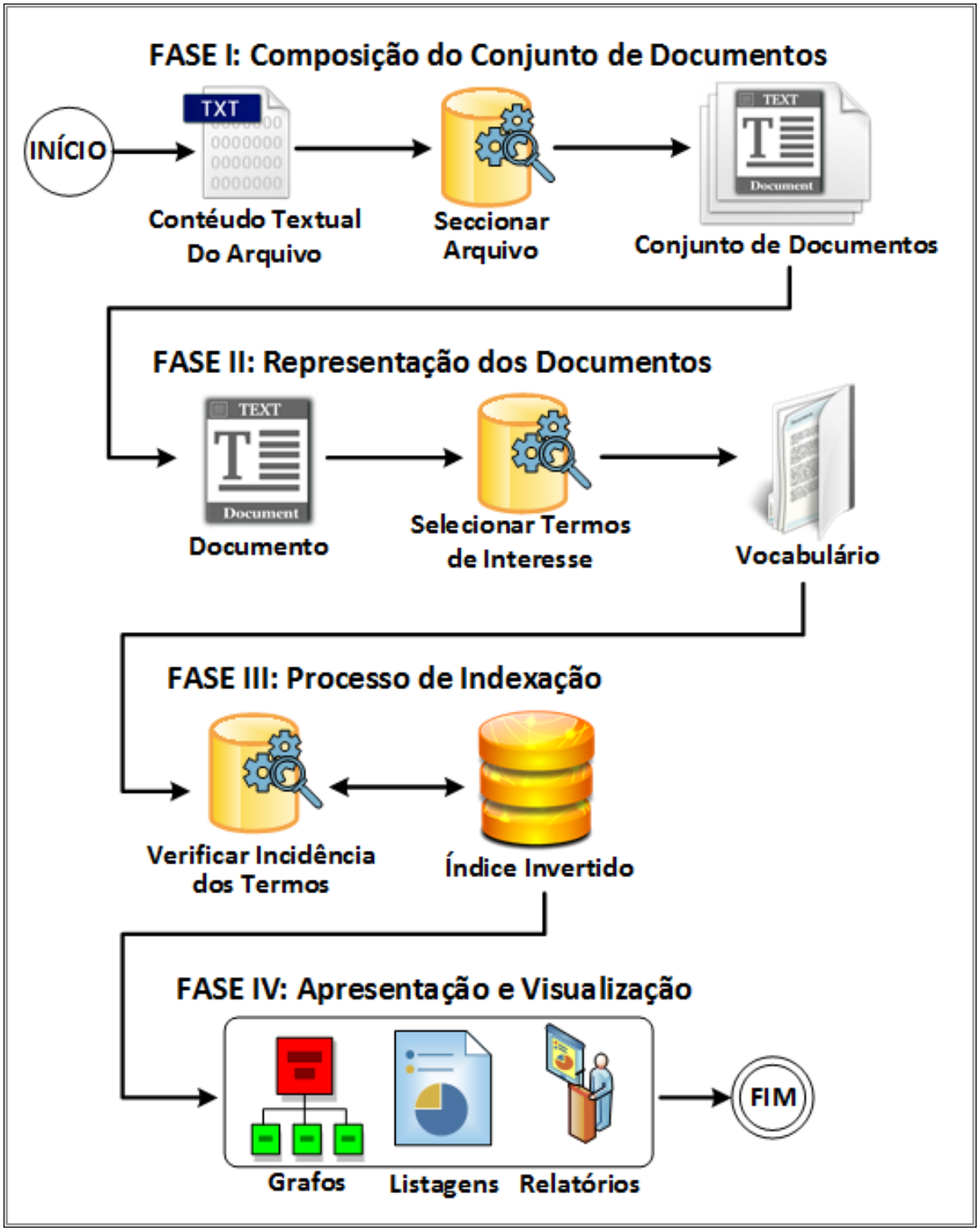


Figura 4.2: Visão Geral das Fases.

O presente método foi desenvolvido com o foco voltado à visão centrada no computador, realizando o processo de Recuperação de Informações e Descoberta do Conhecimento em um conjunto de documentos de texto não estruturado.

Os parâmetros de entrada do SiCReT tornam essa ferramenta bem flexível às necessidades de seus usuários, a saber:

- **Arquivos**: conjunto de arquivos a ser considerado no processamento;
- **Indexadores**: estruturas de dados utilizadas para extrair os termos de interesse contido no conteúdo textual dos documentos;
- **Seccionadores**: estruturas de dados responsáveis por realizar o seccionamento dos arquivos em um conjunto de documentos. Esse parâmetro é opcional.

Os Indexadores e Seccionadores são estruturas de dados que possuem um atributo específico para o armazenamento de uma Expressão Regular (*Regular Expressions*[19]).

A recuperação de informações no método proposto se dá por meio do processamento de um conjunto de indexadores sobre os arquivos a serem analisados. Com a informação desses dois parâmetros de entrada (arquivos e indexadores), o método é capaz de recuperar os termos de interesse, assim como detectar e representar graficamente as interseções entre os documentos analisados.

Vale mencionar que essa ferramenta computacional suporta o seccionamento de arquivos. Com esse recurso é possível trabalhar com situações específicas, a exemplo do cruzamento de informações entre os equipamentos contidos em um conjunto de laudos periciais armazenados em um único arquivo.

O resultado do processamento do SiCReT é a geração de uma estrutura de dados em forma de Índice Invertido, ideal para o processamento incremental e a verificação de interseção de informações.

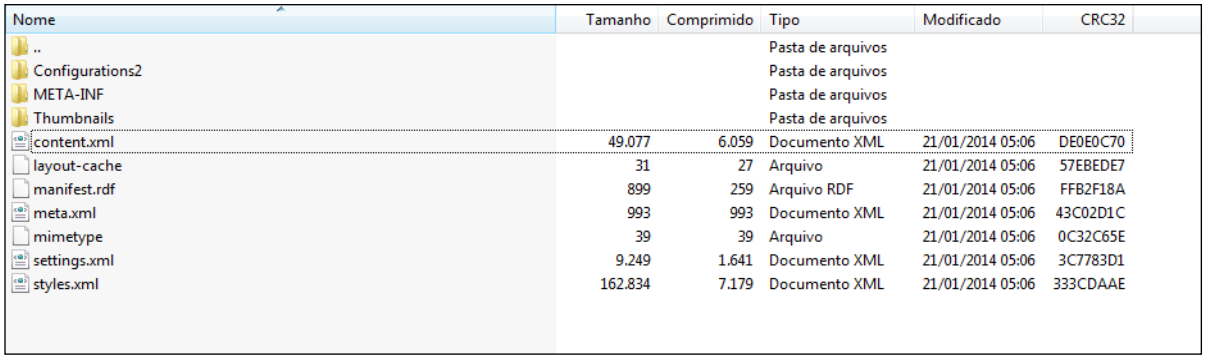
A visualização dos resultados consiste na transformação dos registros contidos no Índice Invertido gerado de forma a apresentar os cruzamentos detectados entre os termos de interesse recuperados.

[19] Regular Expressions. Disponível em < http://pubs.opengroup.org/onlinepubs/007908799/xbd/re.html > Acesso em 10 set. 2015.

4.1 Fase 0: Pré-Processamento - Extração do Conteúdo Textual

Atualmente a base de dados de laudos periciais de dispositivos móveis da Seção de Informática Forense do Instituto de Criminalística do Paraná da Polícia Científica do Paraná - Curitiba encontra-se armazenada em formato ODT (*OpenDocument Text*) que é o formato de documento do *Open Office*.

O fato de que os arquivos dos Laudos de Perícia Criminal se encontram armazenados no formato ODT, nos apresenta um cenário onde a estrutura dos arquivos de entrada para o processamento de dados seja semelhante a um *container*, conforme visualização na Figura 4.3.



Nome	Tamanho	Comprimido	Tipo	Modificado	CRC32
..			Pasta de arquivos		
Configurations2			Pasta de arquivos		
META-INF			Pasta de arquivos		
Thumbnails			Pasta de arquivos		
content.xml	49.077	6.059	Documento XML	21/01/2014 05:06	DE0E0C70
layout-cache	31	27	Arquivo	21/01/2014 05:06	57EBEDE7
manifest.rdf	899	259	Arquivo RDF	21/01/2014 05:06	FFB2F18A
meta.xml	993	993	Documento XML	21/01/2014 05:06	43C02D1C
mimetype	39	39	Arquivo	21/01/2014 05:06	0C32C65E
settings.xml	9.249	1.641	Documento XML	21/01/2014 05:06	3C7783D1
styles.xml	162.834	7.179	Documento XML	21/01/2014 05:06	333CDAAE

Figura 4.3: Estrutura de um arquivo de Laudo Pericial no formato ODT.

O SiCReT pode ser aplicado em qualquer documento textual, incluindo documentos estruturados, semi-estruturados e documentos não estruturados.

A fase de Pré-Processamento é responsável por realizar a extração do conteúdo textual dos arquivos fornecidos como parâmetro na entrada de dados do método. Possibilitando o suporte às formas e formatos de arquivos que armazenam documentos textuais existentes na atualidade, a exemplo de formatos como: HTML, XHTML, XML, *Microsoft Office document formats*, *OpenDocument Format*, PDF, EPUB, RTF e *Text formats*.

Ao final do pré-processamento, será fornecida para a primeira fase do método apenas o conteúdo textual contido no conjunto de arquivos fornecidos na entrada para processamento.

4.2 Fase I: Composição do Conjunto de Documentos

Essa fase considera o fato de que um determinado arquivo possa conter um ou mais laudos periciais de dispositivos móveis, aqui denominados de documentos. Ou seja, cada laudo pericial é considerado um documento a fim de representação computacional.

O Processo de Composição de Documentos verifica se foi informado um conjunto de seccionadores. Os seccionadores proporcionam ao método um meio de detectar e definir o seccionamento do conteúdo textual do arquivo de entrada em um Conjunto de Documentos.

A Figura 4.4 mostra a sequência de aplicação da primeira etapa do método proposto. Nesta fase, o conteúdo textual dos arquivos é analisado no sentido de verificar se ele corresponde a um único documento ou um conjunto de documentos.

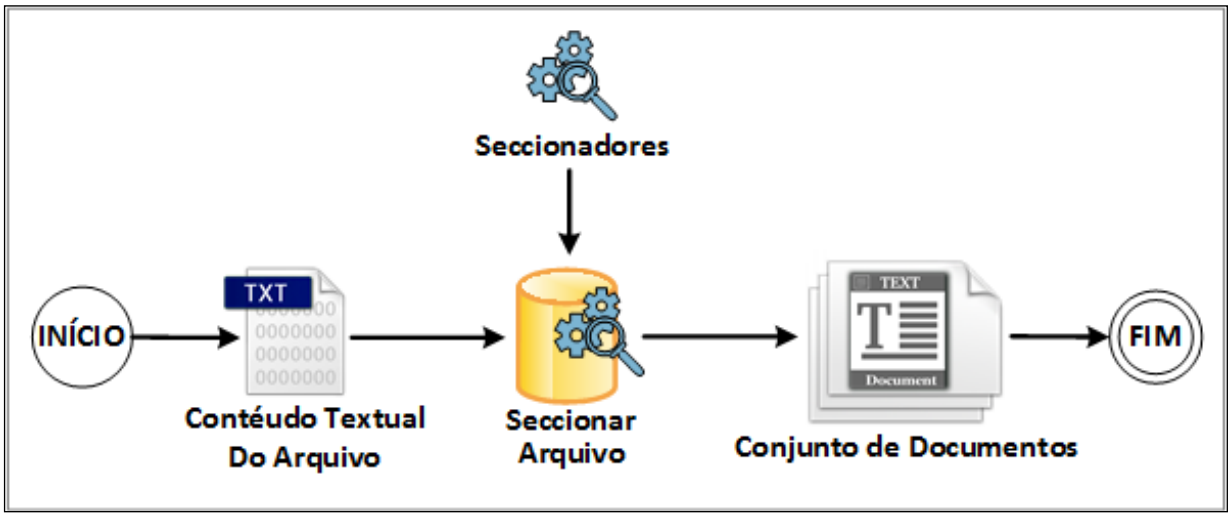


Figura 4.4: SiCReT - Fase I.

O processo de composição do conjunto de documentos pode ser formalmente representado por:

$$C = \bigcup_{\forall d \in T} \{section(d, \alpha)\} \quad (4.1)$$

Sendo que:

C é o conjunto de documentos;

T corresponde ao conteúdo textual do arquivo;

d são os documentos que pertençam ao conteúdo textual do arquivo;

α corresponde ao conjunto de seccionadores dos documentos;

section(.,.) é a função que secciona o conteúdo textual do arquivo em um conjunto de documentos.

Parâmetros de Entrada exemplificando o uso de seccionadores, a saber:

- Seccionamento por Equipamentos: "^equipamento\\s*\\d{1,3}";
- Seccionamento por Cartão SIM: "^cartao\\s*sim\\s*";
- Seccionamento por Cartão SIM-Avulso: "^cartao\\s*sim\\s*-?avulso:";
- Seccionamento por Aparelho Celular: "^aparelho\\s*celular\\s*";

Por ser um parâmetro de entrada opcional, um arquivo vai gerar nesse processo obrigatoriamente no mínimo um documento contento todo o seu conteúdo textual.

Caso não seja encontrado resultados que satisfaçam os seccionadores α que definem o início das seções dos documentos, cada arquivo será considerado automaticamente como um único documento para fins de retorno da função *section(.,.)*.

4.3 Fase II: Representação dos Documentos

A representação dos documentos textuais no SiCReT é dada por meio de um conjunto de termos resultantes da aplicação de indexadores baseados em expressões regulares, que deve ser informada ao método como um de seus parâmetros de entrada.

O conjunto dos termos válidos extraídos do documento proporciona a composição do vocabulário do documento conforme mostra a Figura 4.5.

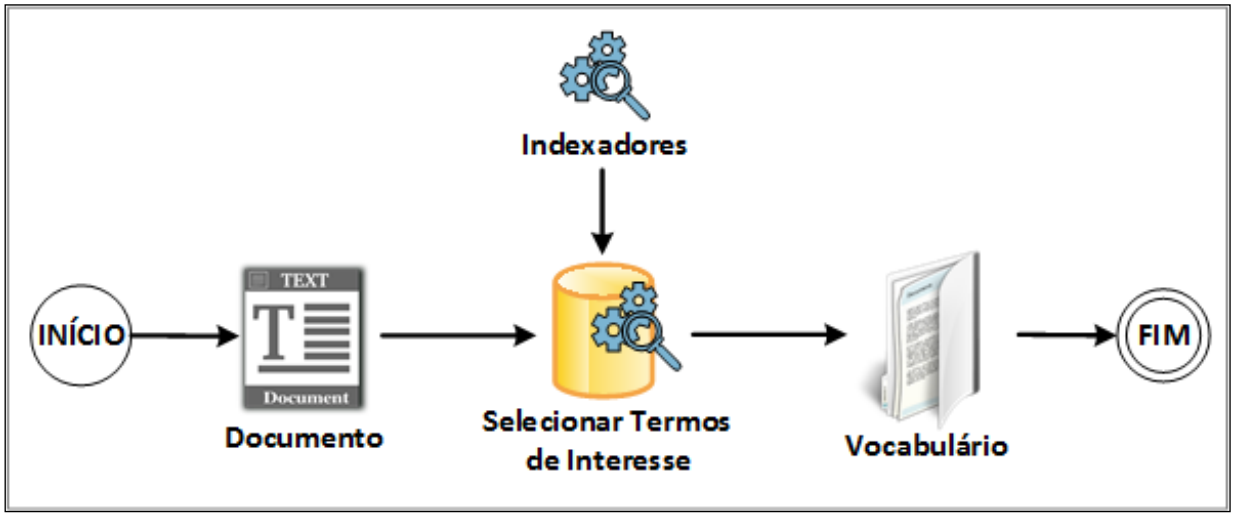


Figura 4.5: SiCReT - Fase II.

A obtenção do vocabulário de cada documento pode ser formalmente representada por:

$$V = \bigcup_{\forall t \in D} \begin{cases} \{t\}, \text{caso representa}(t, \beta) = \text{VERDADEIRO} \\ \emptyset, \text{caso contrário} \end{cases} \quad (4.2)$$

Sendo que:

V é o vocabulário do documento;

D é o documento;

t são os termos que pertençam ao Documento;

β corresponde ao conjunto de indexadores informado para seleção dos termos de composição do vocabulário;

representa(..) é a função que verifica se o termo pode fazer parte do conjunto de termos que representarão o documento, desde que se enquadrem nas condições especificadas no conjunto de indexadores β.

Parâmetros de entrada exemplificando o uso de Indexadores que podem ser utilizados para a recuperação de termos contendo registros telefônicos, a saber:

- Padrão Internacional: "\\D?\\d{1,2}\\D{1,3}\\d{1,3}\\D{1,3}\\d{3,5}[-]{1,3}\\d{3,5}";
- Padrão Nacional: "\\D{1,2}?\\d{1,3}\\D{1,3}\\d{3,5}[-]{1,3}\\d{3,5}";
- Padrão Nacional Reduzido: "\\d{3,5}[-]{1,3}\\d{3,5}";

4.4 Fase III: Processo de Indexação

Nesse momento, realiza-se o cruzamento das informações propriamente dito. Nesta etapa, é gerada uma estrutura de dados por meio da unificação de todos os termos que compõem os vocabulários, representando os seus respectivos documentos. A Figura 4.6 mostra a sequência de aplicação deste processo.

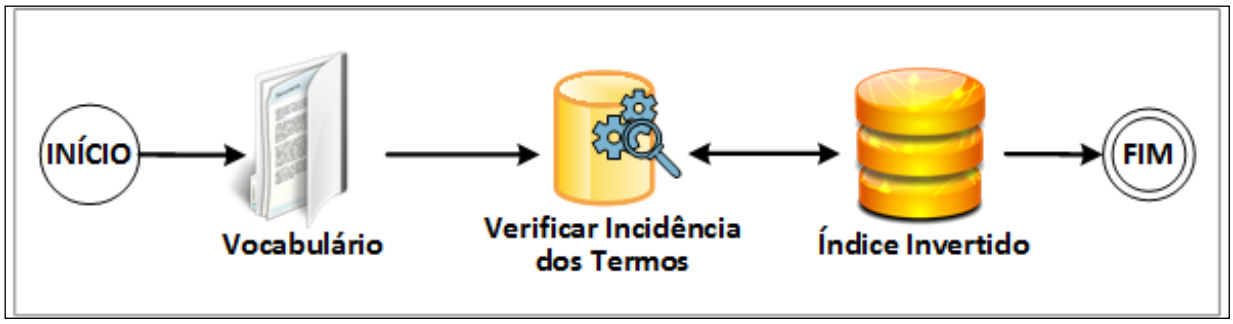


Figura 4.6: SiCReT - Fase III.

Nessa estrutura de dados, os termos são individualizados e, conforme é verificada a incidência dos termos nos vocabulários, os respectivos documentos passam a fazer parte do conjunto de documentos referenciados pelo termo que está sendo analisado.

A estrutura de dados resultante deste processo é denominada de Índice Invertido. Na qual as chaves são os termos e os valores associados a cada chave pode ser um documento ou um conjunto de documentos que contêm a incidência da chave referenciada. A Figura 4.7 mostra um exemplo de estrutura de dados formada por um Índice Invertido.

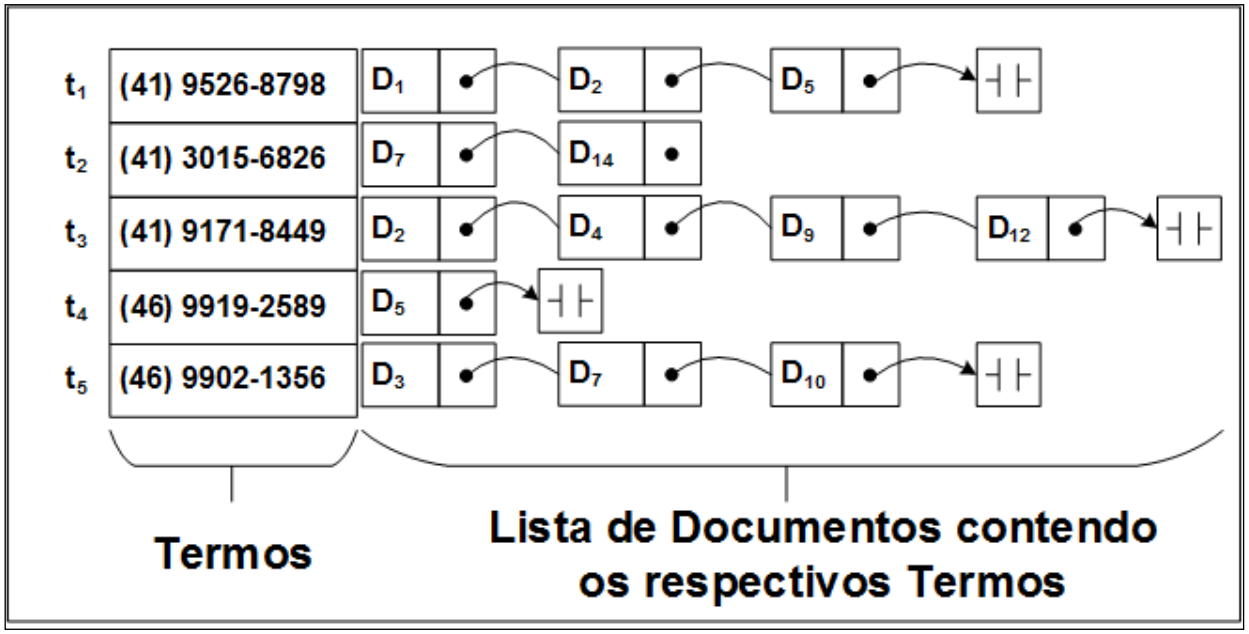


Figura 4.7: SiCReT - Índice Invertido.

O processamento de geração do Índice Invertido adotado no SiCReT pode ser formalmente representada por:

$$\mathrm{I} = \left[\left\langle t_{\mathrm{i}}; L_{\mathrm{i}} \right\rangle\right]_{\forall t_i \in \mathrm{V}} \tag{4.3}$$

$$\mathrm{L_i} = \bigcup_{\forall doc_i \in \mathrm{D}} \begin{cases} \{\mathrm{doc_i}\}, caso\, t_i \in doc_i \\ \emptyset, caso\, contrário \end{cases} \tag{4.4}$$

Sendo que:

I é o processo de indexação que resulta na composição do índice invertido;

t_i são os termos que pertençam ao vocabulário que representa o documento;

L_i é a lista de documentos, ou seja, documento ou conjunto de documentos atribuída como valor de um determinado termo t_i;

doc_i é o *iésimo* documento que pertence ao conjunto de documentos do processamento.

Na estrutura de dados do Índice Invertido observa-se que cada conjunto de documentos atribuído ao valor de um termo significa que existe um cruzamento de informações, ou seja, o está contido em dois ou mais documentos. Podemos afirma nesses casos que o termo é responsável pela interseção entre os respectivos documentos.

Já nos casos onde o termo tem incidência em apenas um único documento, o cruzamento de informações é inexistente. Pois o termo referenciado está contido em apenas um documento.

4.5 Fase IV: Apresentação e Visualização

A fase de apresentação e visualização dos cruzamentos é realizada com o uso das informações contidas no Índice Invertido gerado na fase anterior. Esse processo é responsável por apresentar aos usuários os resultados obtidos pelo método proposto.

Como já mencionado no Capítulo 2, a representação visual, em especial em forma de grafos, permite que os usuários possam absorver rapidamente grandes quantidades de informações e construir mapas mentais das informações recuperadas. Na Figura 4.8 é mostrado a sequência de aplicação deste processo.

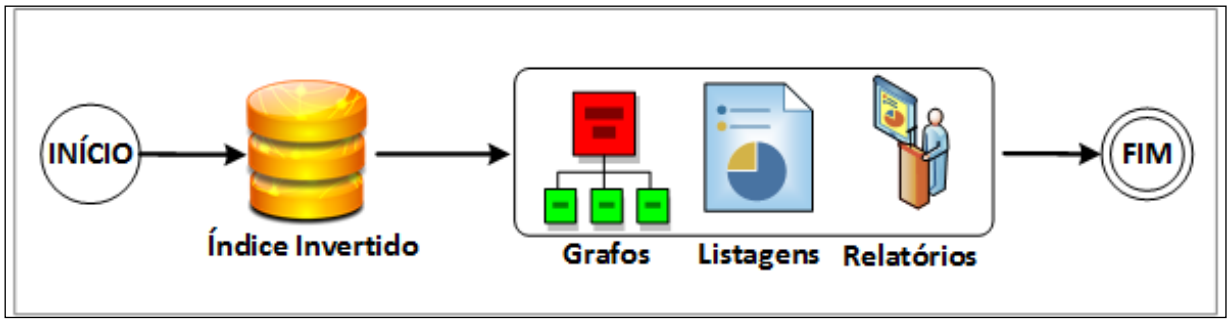


Figura 4.8: SiCReT - Fase IV.

A representação visual dos cruzamentos facilita a identificação de cruzamentos diretos e indiretos entre as interseções de informações detectadas. Proporcionando dessa forma maior capacidade de interpretação das informações visualizadas, e ainda a criação de modelos mentais das informações por parte dos usuários.

4.6 Considerações Finais

Este Capítulo apresentou a concepção do método Sistema de Cruzamento de Registros Telefônicos - SiCReT. Foi feito um detalhamento e a representação computacional de cada uma de suas quatro fases: Composição do Conjunto de Documentos, Representação dos Documentos, Processo de Indexação e Apresentação das Informações.

O SiCReT é uma ferramenta computacional capaz de realizar o processo de recuperação, detecção e visualização de cruzamentos existentes entre termos de interesse de um determinado conjunto de laudos periciais de dispositivos móveis. Atendendo uma demanda existente na área de Recuperação de Informações e Descoberta do Conhecimento apresentada pelas Seções de Informática Forense.

O próximo Capítulo apresentará os experimentos realizados com o intuito de comprovar a eficácia do método proposto na base de dados do Instituto de Criminalística do Paraná.

Capítulo 5

Experimentos Realizados e Análise de Resultados

Com o intuito de realizar a abordagem das principais situações que os peritos se depararam no exercício da atividade pericial, foi celebrado um contrato de parceria de trabalho entre a PUC-PR e o Instituto de Criminalística do Paraná.

Um dos benefícios proporcionados por essa parceria foi a autorização do Diretor Geral da Instituição de acesso a utilização da base de dados de laudos periciais de dispositivos móveis da Seção de Computação Forense do IC-PR, conforme referência mencionada nos Anexos I e II.

Possibilitando dessa forma o acesso a dados reais para realização da Prova dos Conceitos de aplicação do método proposto neste trabalho.

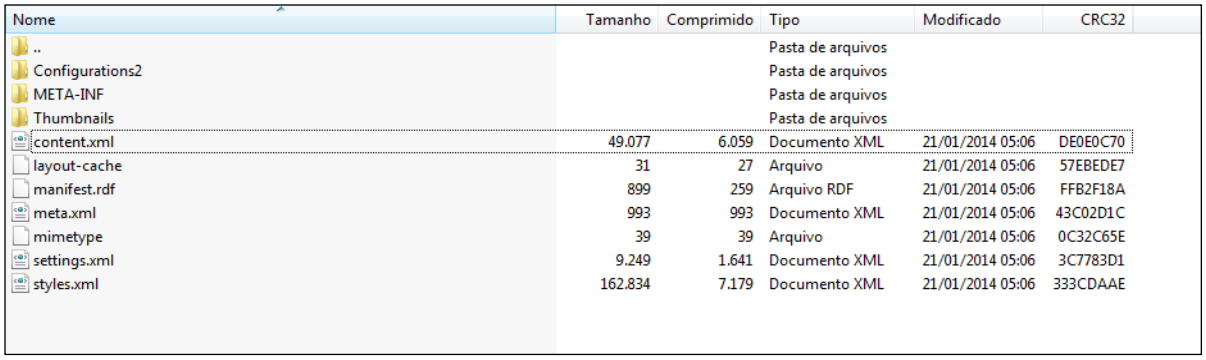
5.1 Base de Dados do Experimento

Os experimentos realizados no presente trabalho utilizaram uma base de dados composta por 200 arquivos no formato ODT. Esses arquivos possuem informações referente aos laudos periciais, anexos de laudos e arquivos complementares.

A base de dados de laudos periciais de dispositivos móveis da Seção de Informática Forense do Instituto de Criminalística do Paraná da Polícia Científica do Paraná - Curitiba encontra-se armazenada em formato ODT (*OpenDocument Text*[20]) que é o formato de documento utilizado por padrão pelo editor de textos Writer do LibreOffice.

[20] OASIS OpenDocument Essentials: Using OASIS OpenDocument XML. Disponível em < http://books.evc-cit.info/OD_Essentials.pdf>. Acesso em: 23 de fevereiro de 2014.

O fato de que os arquivos dos Laudos de Perícia Criminal se encontram armazenados no formato ODT, nos apresenta um cenário onde a estrutura dos arquivos de entrada para o processamento de dados seja semelhante a um *container*, conforme visualização na Figura 4.3.



Nome	Tamanho	Comprimido	Tipo	Modificado	CRC32
..			Pasta de arquivos		
Configurations2			Pasta de arquivos		
META-INF			Pasta de arquivos		
Thumbnails			Pasta de arquivos		
content.xml	49.077	6.059	Documento XML	21/01/2014 05:06	DE0E0C70
layout-cache	31	27	Arquivo	21/01/2014 05:06	57EBEDE7
manifest.rdf	899	259	Arquivo RDF	21/01/2014 05:06	FFB2F18A
meta.xml	993	993	Documento XML	21/01/2014 05:06	43C02D1C
mimetype	39	39	Arquivo	21/01/2014 05:06	0C32C65E
settings.xml	9.249	1.641	Documento XML	21/01/2014 05:06	3C7783D1
styles.xml	162.834	7.179	Documento XML	21/01/2014 05:06	333CDAAE

Figura 4.3: Estrutura de um arquivo de Laudo Pericial no formato ODT.

Para dimensionar o tamanho dos arquivos, utilizou-se o valor de retorno da função WordCount proposta por Dean & Ghemawat, em artigo publicado a respeito dos conceitos utilizados na implementação da ferramenta MaReduce (DEAN; GHEMAWAT, 2010).

A Tabela 5.1, mostra os principais itens de dimensionamento da Base de Dados e do Processo de Indexação.

Tabela 5.1: Características da Base de Dados e do Índice Invertido.

Descrição do Item	Valores
Total de Arquivos	200
Termos (detectados na WordCount)	516.592
Total de Termos Relevantes	7.913
Total de Termos Recuperados	8.725
Ocorrência dos Termos Recuperados	46.034

Para melhor entendimentos dos resultados obtidos nos experimentos considera-se nesse trabalho, a saber:

- **Termos**: parte de dados expressiva detectada em um documento, a exemplo de palavras, artigos, registros telefônicos, entre outros;

- **Termos Relevantes**: termos de interesse recuperados, ou seja, termos correspondentes a registros telefônicos;
- **Termos Recuperados**: conjunto composto por termos relevantes recuperados somados a termos que não correspondem a registros telefônicos, ou seja, "falsos positivos".

Os valores gerados nos experimentos, são os resultados obtidos da aplicação do método SiCReT na base de dados mencionada, tendo as expressões regulares de seccionadores e indexadores como parâmetros de entrada, da seguinte forma:

- **Seccionadores**
  - Equipamentos: "^equipamento\\s*\\d{1,3}";
  - Cartão SIM: "^cartao\\s*sim\\s*";
  - Cartão SIM-Avulso: "^cartao\\s*sim\\s*-?avulso:";
  - Aparelho Celular: "^aparelho\\s*celular\\s*";
- **Indexadores**
  - Padrão Internacional: "\\D?\\d{1,2}\\D{1,3}\\d{1,3}\\D{1,3}\\d{3,5}[-]{1,3}\\d{3,5}";
  - Padrão Nacional: "\\D{1,2}?\\d{1,3}\\D{1,3}\\d{3,5}[-]{1,3}\\d{3,5}";
  - Padrão Nacional Reduzido: "\\d{3,5}[-]{1,3}\\d{3,5}";
  - Padrão Compacto: "[\\d#]{7,16}";

O "Padrão Compacto" foi aplicado nos arquivos de anexo, apenas nas seções de contatos da agenda, chamadas e mensagens, e ainda no arquivos de laudo, apenas na seção onde é especificada a linha telefônica.

Informando ao método SiCReT um conjunto de arquivos a serem processados e os parâmetros de seccionamento e indexação, o processo de análise está pronto para ser iniciado sobre o conteúdo textual contido nos arquivos.

5.2 Experimentos Realizados na Base de Dados Real

5.2.1 Cruzamento de Registros Telefônicos entre os Arquivos

No processo de indexação obteve-se um total de 8.725 termos recuperados, sendo 7.913 termos relevantes (registros telefônicos), gerando 281 interseções de registros telefônicos encontradas entre os arquivos na aplicação do método proposto. A Tabela 5.2, mostra os valores obtidos no processamento.

Tabela 5.2: Processo de Indexação entre os Arquivos.

Descrição do Item	Valores
Total de Arquivos	200
Total de Termos Recuperados	8.725
Termos Relevantes Recuperados	7.913
Total de Interseções Detectadas	281
Cruzamentos entre Laudo e Anexo	21
Cruzamentos entre Laudos Distintos	7
Total de Estudo de Casos	28

A Figura 5.1 mostra de forma visual os cruzamentos detectados. Que podem ser divididos em dois grupos, a saber: cruzamentos de informações entre o arquivo de laudo e seu respectivo anexo, e outros tipos de cruzamentos.

Tendo em mente que os cruzamentos entre os laudos e seus respectivos anexos não tem, nesse momento, interesse ao exercício da atividade pericial. Podemos eliminar esse tipo de relacionamento (21 relacionamentos), criando assim, uma forma mais sintética de visualização conforme mostra a Figura 5.2.

Dessa forma a Figura 5.2 reúne apenas as representações dos cruzamentos de interesse ao trabalho investigativo dos peritos do IC-PR. Como pode ser observado, resultou em 7 grafos de cruzamento de dados que indicam relacionamentos entre laudos periciais de dispositivos móveis distintos.

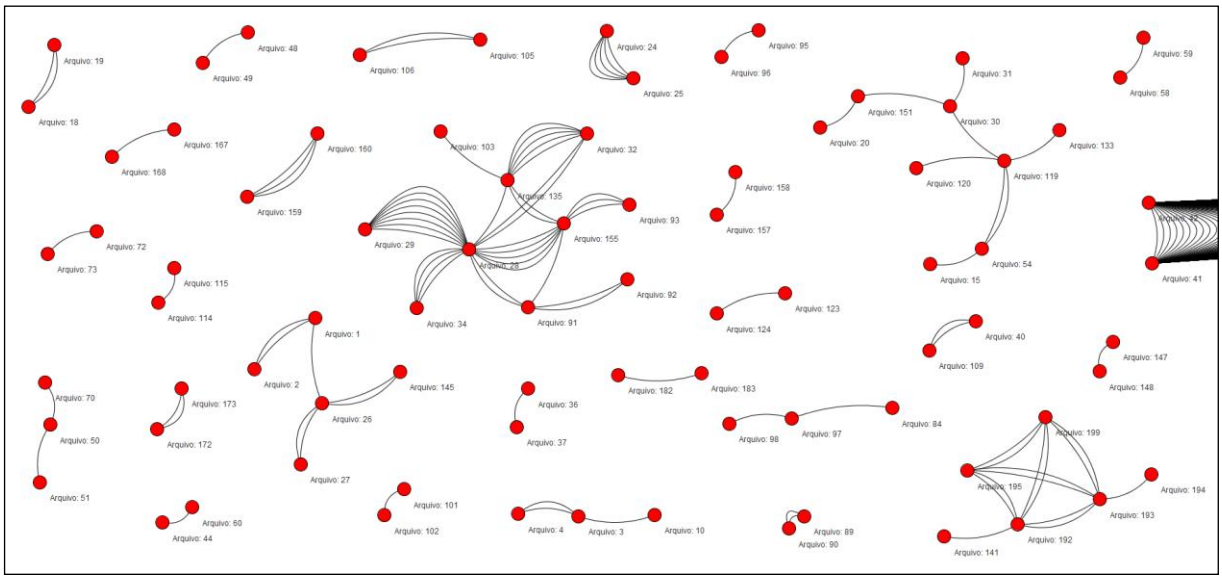


Figura 5.1: Cruzamento de Registros Telefônicos entre Arquivos.

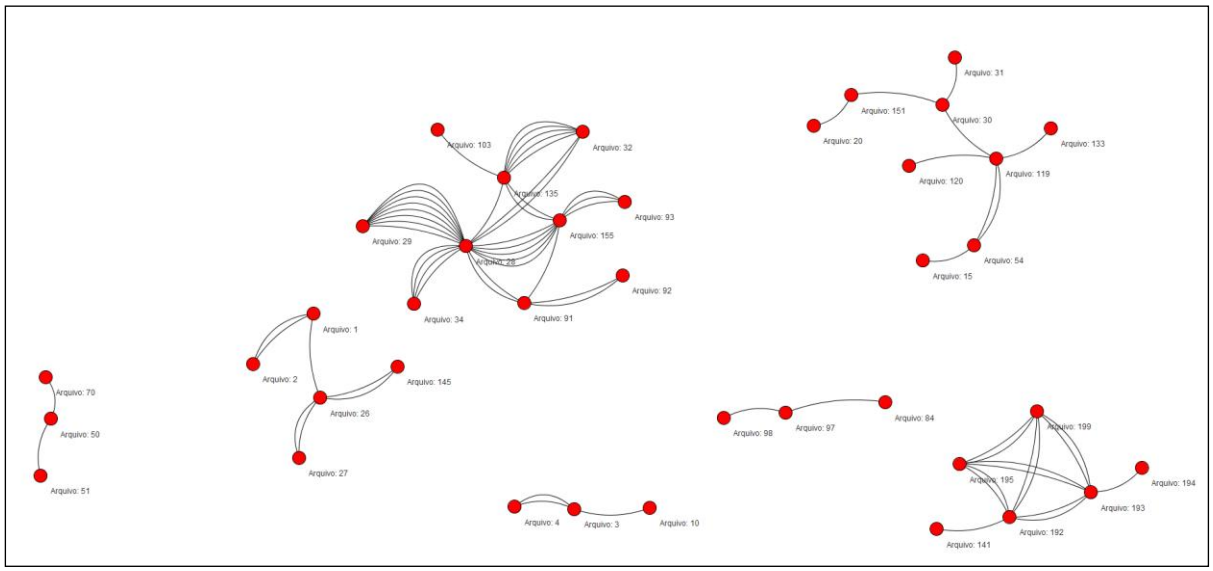


Figura 5.2: Cruzamento de Registros Telefônicos entre Arquivos - Sintético.

5.2.2 Cruzamento de Dados da Raiz dos Termos

Entende-se por raiz de um termo a sua forma mais reduzida possível. Nesse experimento foi especificado que a raiz é representada pelos 8 dígitos base dos registros telefônicos. Podemos exemplificar que a raiz do registro telefônico +55 (41) 9526-8798 ou (41) 9526-8798, é 95268798.

A Tabela 5.3 mostra os valores obtidos no processamento do método proposto, em relação às interseções encontradas na raiz dos registros telefônicos entre os arquivos da base de dados.

Tabela 5.3: Processo de Indexação (Raiz) entre os Arquivos.

Descrição do Item	Valores
Total de Arquivos	200
Total de Termos Recuperados	8.725
Total de Termos Recuperados (Raiz)	7.917
Total de Interseções Detectadas	329
Cruzamentos entre Laudo e Anexo	35
Cruzamentos entre Laudos Distintos	7
Total de Estudo de Casos	42

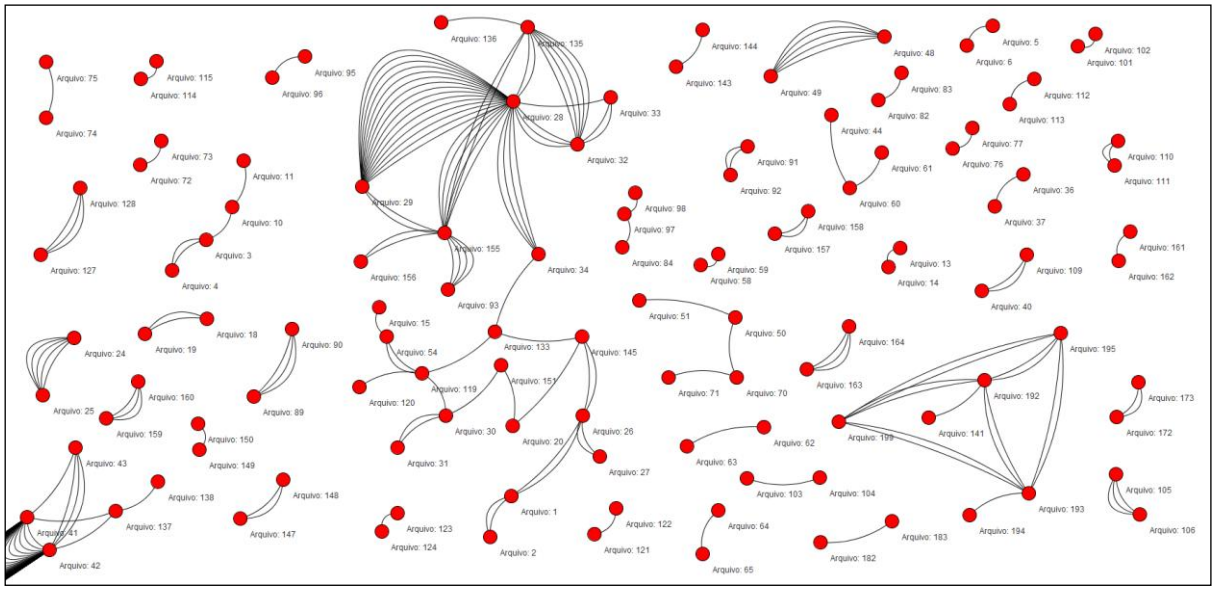


Figura 5.3: Cruzamento da Raiz de Registros Telefônicos entre Arquivos.

Observa-se que os resultados mostrados na Figura 5.3 apresentam um significativo aumento no trabalho de vértices e arestas nos grafos que representam os cruzamentos, em relação aos cruzamentos mostrados na Figura 5.1. Resultando ainda, no aumento no número total de interseções detectadas de 281, para 329.

Eliminando os cruzamentos entre os laudos e seus respectivos anexos obtemos o resultado sintético que mostra a Figura 5.4.

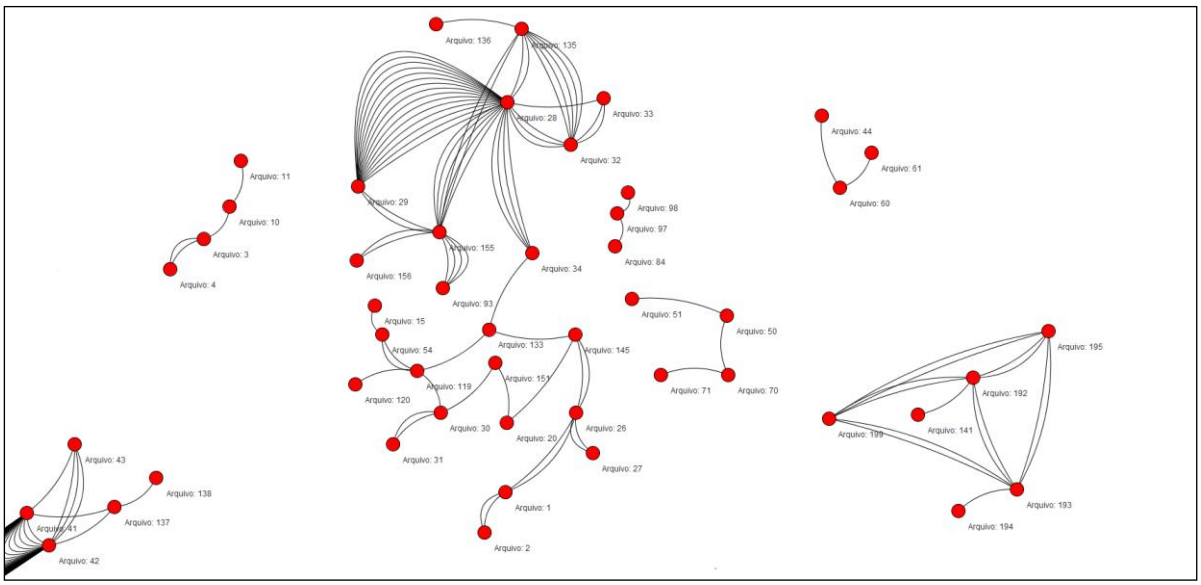


Figura 5.4: Cruzamento da Raiz de Registros Telefônicos entre Arquivos - Sintético.

5.2.2 Cruzamento de Informações com Seccionamento de Arquivos

O cruzamento de informações entre partes seccionadas de um arquivo também traz consigo uma considerável forma de apoio no exercício da atividade pericial. Contribuindo em situações onde é interessante, por exemplo, visualizar os relacionamentos entre os equipamentos contidos em um laudo pericial.

Dentre os 200 arquivos contidos na base de dados, 36 deles se enquadraram nas especificações da expressões regulares dos seccionadores de documentos usadas como parâmetros de entrada. Resultando em um total geral de 402 documentos seccionados. A Tabela 5.4 e a Figura 5.5 mostram os resultados obtidos no seccionamento do ArquivoX.

Tabela 5.4: Seccionamento do ArquivoX.

Descrição do Item	Valores
Total de Arquivos	1
Total de Documentos (seccionados)	7
WordCount	11.226
Termos - Registros Telefônicos	378
Termos - Registros Telefônicos (Raiz)	332
Ocorrências de Registros Telefônicos	1.526
Total de Interseções Detectadas	52
Total de Interseções Detectadas (Raiz)	51

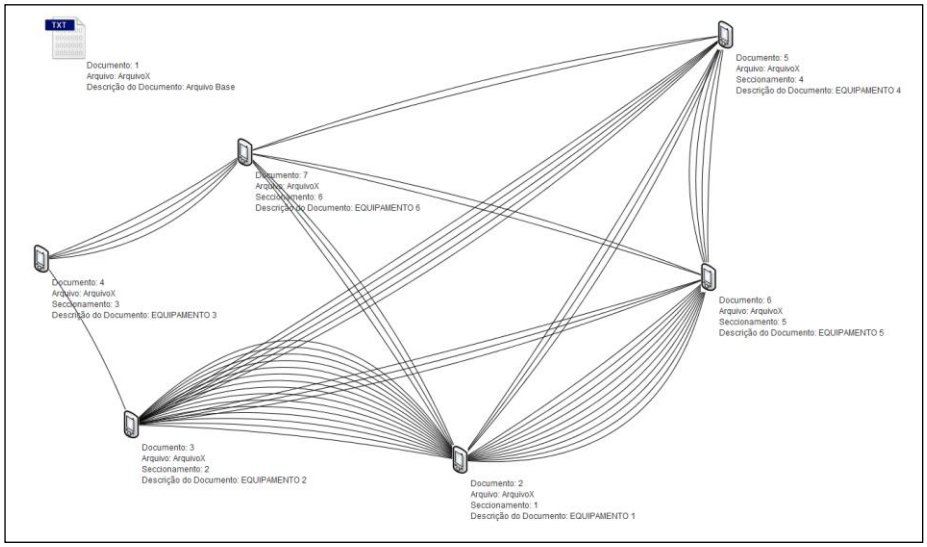


Figura 5.5: Seccionamento do ArquivoX.

Outro seccionamento de arquivo é mostrado na Tabela 5.5 e a Figura 5.6 apresenta os resultados obtidos a partir do seccionamento do ArquivoX2.

Tabela 5.5: Seccionamento do ArquivoX2.

Descrição do Item	Valores
Total de Arquivos	1
Total de Documentos (seccionados)	5
WordCount	11.226
Termos - Registros Telefônicos	378
Termos - Registros Telefônicos (Raiz)	332
Ocorrências de Registros Telefônicos	1.526
Total de Interseções Detectadas	52
Total de Interseções Detectadas (Raiz)	51

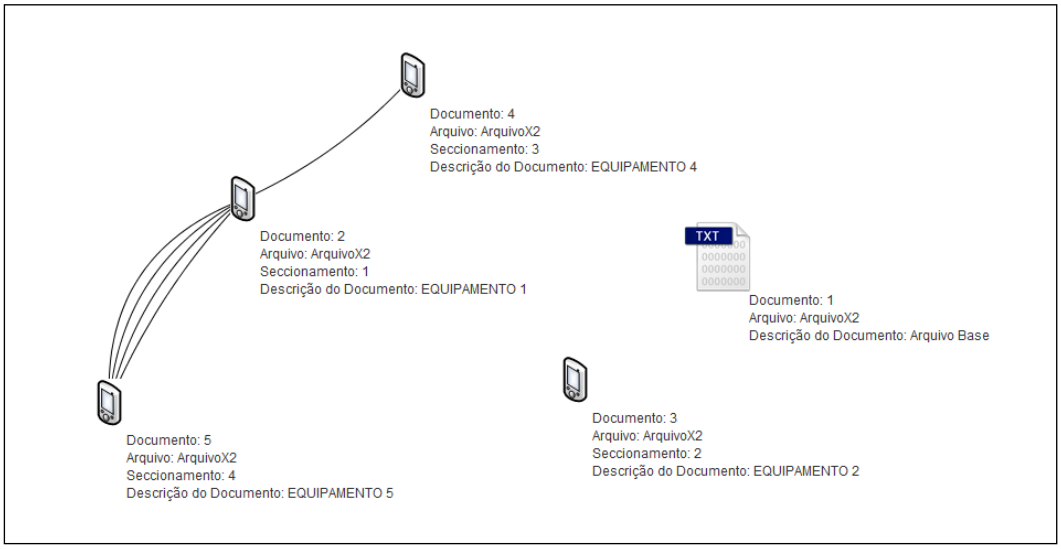


Figura 5.6: Seccionamento do ArquivoX2.

5.3 Análise dos Resultados

A base de dados do IC-PR utilizada nos experimentos deste trabalho, composta por 200 arquivos de Laudos Periciais de Dispositivos Móveis, possui um total de 7.913 termos relevantes a serem recuperados. Ou seja, registros telefônicos que devem ser recuperados.

A aplicação do método SiCReT nos 200 arquivos, utilizando os parâmetros de entrada mencionados no Capítulo 4.1, obteve-se um total de 8.725 termos recuperados no processamento dos experimentos.

Dentro desse total de termos recuperados, detectou-se 812 termos "falsos positivos", ou seja, foram recuperados porém não eram termos de interesse dos usuários, a exemplo de números de protocolos de atendimento, número de cupom fiscal, entre outros.

O fato de que as expressões regulares usadas nos indexadores serem bem genéricas pode ter contribuído para a ocorrência de termos "falso negativo", ou seja, ter reconhecido registros que não correspondiam a números telefônicos. Por outro lado contatou-se que, nos experimentos realizados, as informações relevantes foram recuperados com sucesso.

Dessa forma o cálculo da precisão do sistema de recuperação de informações pode ser representado com a seguinte equação 5.1, a saber:

$$P = \frac{|R \cap A|}{|A|} \tag{5.1}$$

Sendo que:

P é a precisão do sistema de recuperação de informações;

R é o total de termos relevantes;

A é o total de termos recuperados pelo sistema de recuperação;

R ∩ A é o total de termos da interseção dos conjuntos R e A;

Quando aplicada o cálculo da precisão nos resultados obtidos nos experimentos do presente trabalho resulta em uma precisão de 0,91, ou seja, 91% de precisão na recuperação da informação.

Vale mencionar que a recuperação dos falsos positivos, 812 termos, não influenciaram no cruzamento das informações, pelo fato de que não ocorreu uma interseção desses termos entre os documentos analisados.

O conhecimento gerado foi apresentado em forma de grafos. Permitindo meios a seus usuários absorverem rapidamente grandes quantidades de informações, detectar cruzamentos e construir mapas mentais das informações recuperadas.

Um exemplo disso pode ser observado nas imagens contidas na Figura 5.1 e a Figura 5.3 que mostram todos os cruzamentos de registros telefônicos contidos nos laudos periciais analisados. Contudo alguns cruzamentos contidos nessas imagens eram formados por termos contidos entre um laudo e seu respectivo anexo, ou seja, um cruzamento de informações que não representa uma natureza criminosa.

Já as imagens contidas na Figura 5.2 e a Figura 5.4 mostram de forma visual a os cruzamento de informações que ocorre apenas entre laudos periciais distintos. O cruzamento entre esses laudos é de grande interesse por representar novas evidências forenses, fluxo de informações e cruzamentos de dados que até então estavam ocultos nas Sessões de Informática Forense. Esses resultados apresentados nunca puderam ser alcançados ou mesmo visualizados antes da existência do SiCReT, dessa forma, foi gerado um conhecimento que ninguém havia observado até então.

As imagens da Figura 5.5 e da Figura 5.6 mostraram que o Cruzamento de Informações utilizando o Seccionamento de Arquivos pode gerar com precisão a detecção de cruzamento de informações contidas em diversos laudos periciais armazenados em um único arquivo.

5.4 Considerações Finais

Este Capítulo apresentou os resultados gerados na aplicação do SiCReT em 200 arquivos da base de dados de laudos periciais de dispositivos móveis da Seção de Informática Forense do Instituto de Criminalística do Paraná da Polícia Científica do Paraná - Curitiba.

A parceria celebrada entre a PUC-PR e o Instituto de Criminalística do Paraná possibilitou o acesso aos dados reais para comprovar a eficácia do Método Proposto no apoio do exercício de atividades de Policiamento Preditivo e Serviços de Inteligência.

O próximo capítulo apresentará as conclusões obtidas deste trabalho, assim como os possíveis trabalhos futuros.

Capítulo 6

Conclusão

A Recuperação de Informações em documentos de texto não estruturado, assim como todos os Sistemas de RI, possuem o objetivo de fazer com que o usuário encontre a informação que está precisando rapidamente, de modo que este usuário não necessite analisar todas as informações existentes na base de informações.

O SiCReT é um método com visão centrada no computador, realizando o processo de Recuperação de Informações e Descoberta do Conhecimento em um conjunto de documentos de texto não estruturado. Realiza a Recuperação de Informações direcionada na construção de índices eficientes, no processamento de consultas e no desenvolvimento de algoritmos.

O presente trabalho apresentou uma ferramenta computacional com o propósito de atender uma lacuna que está em aberto nessa área nas Sessões de Computação Forense dos Institutos de Criminalística.

Com o intuito de fornecer meios de processamento e otimização das atividades realizadas por peritos, o trabalho apresentou um método para Cruzamento de Registros Telefônicos a partir de dados extraídos de laudos periciais de dispositivos móveis.

A aplicação dessa ferramenta em 200 arquivos de Laudos Periciais de Dispositivos Móveis da base de dados real do IC-PR, no formato ODT, utilizando os parâmetros de entrada mencionados no Capítulo 5.1, obteve um total de 8.725 termos recuperados.

O sistema de Recuperação de Informações do método SiCReT obteve uma precisão de 0,91, ou seja, 91%. Contudo os 812 termos falsos positivos, nesse experimente, não influenciaram no cruzamento das informações existentes.

Os resultados obtidos nos experimentos nos permite afirmar que a aplicação do método proposto na base de dados especificada no presente trabalho conseguiu detectar as

intersecções previstas entre os documentos através dos termos de interesse contidos nos mesmos, ou seja, os registros telefônicos.

Foi observado que a visualização dos resultados gerados, especialmente em formato de grafos, permitiu aos usuários analisar rapidamente grandes quantidades de informações detectando e visualizando os cruzamento de informações de interesse entre laudos distintos.

Vale mencionar que resultados obtidos nos experimentos realizados, a exemplo da imagens contidas na Figura 5.2 e a Figura 5.4, apresentaram resultados de grande interesse por representar novas evidências forenses, fluxo de informações e cruzamentos de dados que até então estavam ocultos nas Sessões de Informática Forense.

Os resultados gerados a partir de documentos dispersos nas Sessões de Computação Forense mostrou que o trabalho pode contribuir fornecendo uma ferramenta de apoio aos Serviços de Inteligência e Policiamento Preditivo, evitando a subjetividade no exercício da atividade pericial e proporcionando a produção de provas e evidências forenses que até então não tinham sido visualizados antes da existência do SiCReT.

6.1 Trabalhos Futuros

Segue algumas sugestões de trabalhos futuros que podem ser realizados, a saber:

- **Cruzamento entre Arquivos contidos nos Laudos**: pode ser exemplificado sua aplicação na realização do cruzamento entre imagens de pedofilia contidas em laudos de dispositivos móveis distintos;
- **Cruzamento de Informações do SISBALA** (Sistema de Indexação Balística): aplicar o SiCReT com as adaptações necessárias para realizar o cruzamento de informações armazenadas no SISBALA;
- **Aplicar o SiCReT em um Big Data**: ou seja, em uma estrutura centralizada entre todos os laudos periciais do país, proporcionando o processamento em grandes volumes de dados a verificação de laudos em tempo real. Quando o SiCReT for aplicado utilizando essa tecnologia, no momento em que o laudo percial estiver sendo elaborado pelos peritos, as informações armazenadas nas Seções de Informática Forense do País podem ser consideradas no cruzamento de dados.

Referências Bibliográficas

AL-ZAIDY, R.; FUNG, B. C.; YOUSSEF, A. M.; FORTIN, F.. *Mining criminal networks from unstructured text documents. Digital Investigation*, 2012, v. 8, n. 3, p. 147-160.

ANATEL, A. *N. de T. Brasil alcança 268,44 milhões de acessos móveis em agosto de 2013*. 2013. Disponível em: <http://www.anatel.gov.br/>. Acesso em: 07 de outubro de 2013.

ANATEL, A. *N. de T. Julho de 2015 fecha com 281,45 milhões de acessos móveis*. 2015. Disponível em: <http://www.anatel.gov.br/>. Acesso em: 25 de setembro de 2015.

ANWAR, Tarique; ABULAISH, Muhammad. *A social graph based text mining framework for chat log investigation. Digital Investigation*. Elsevier, 2014, v. 11, p 349-362.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca. Bookman Editora, 2013. 590p.

COELHO, Odete Máyra Mesquita; PINTO, Virgínia Bentes; DE SOUSA, Marckson Roberto Ferreira. *Recuperação da informação: estudo da usabilidade na base de dados Public Medical (PUBMED)*. Pesquisa Brasileira em Ciência da Informação e Biblioteconomia, 2013, v8, n 1, p 40-50.

COPPIN, Ben. *Inteligência Artificial*. Rio de Janeiro, Editora LTC, 2010.

BERRY, Michael J. A.. *Data Mining Techniques: for Marketing, Sales, and Customer Support*. John Wiley & Sons Inc., 1997. 454p.

BERSON, Alex; SMITH, Stephen J.. *Data Warehousing, Data Mining & OLAP*. McGraw-Hill, 1997. 612p.

BLANCO, Roi; BARREIRO, Alvaro. *Probabilistic static pruning of inverted files*. ACM Transactions on Information Systems (TOIS), 2010, v. 28, n. 1, a. 1, 33p.

BLUMER, A.; BLUMER, J.; HAUSSLER, D.; MCCONNELL, R.; EHRENFEUCHT, A.. *Complete inverted files for efficient text retrieval and analysis*. Journal of the ACM (JACM), 1987, v. 34, a. 3, p. 578-595.

COSTA, M. A. S. L. *Computação Forense*. Campinas, SP: Millenium, 2003. 246 p.

CRAIGER, J.P. Computer forensics procedures and methods. To appear in H. Bigdoli (Ed.), Handbook of Information Security. John Wiley & Sons, 2007.

DAGHER, Gaby G.; FUNG, Benjamin CM. *Subject-based semantic document clustering for digital forensic investigations*. Data & Knowledge Engineering, 2013, v. 86, p. 224-241.

DATE, C. J.. *Introdução a Sistemas de Banco de Dados*. Rio de Janeiro, RJ: Editora Campus, 2004. 805p.

DEAN, J.; GHEMAWAT, S. *MapReduce: a flexible data processing tool*. Communications of the ACM, 2010, 53(1), 72-77.

DECARLI, A.; FREITAS, C. O. A. ; GROCHOCKI, L. R. ; VRUBEL, A. ; ZAGO, R. L. . *SICReT Sistema de Cruzamento de registros Telefônicos*. In: XIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg), 2013, Manaus. Anais do XIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg). Manaus: Sociedade Brasileira de Computação, 2013. v. 1. p. 527536.

DECARLI, A.; FREITAS, C. O. A. ; GROCHOCKI, L. R. ; PARAISO, E. C. ; GROKOSKI1, C. L. . *Banco de Dados de Laudos Periciais de Dispositivos Móveis*. In: XIV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg), 2014, Belo Horizonte. XIV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais. Belo Horizonte: Sociedade Brasileira de Computação, 2014. v. 1. p. 559571.

EISENBERG, J. David. *OASIS OpenDocument Essentials: Using OASIS OpenDocument XML, Firends of OpenDocument Inc*, 303p. 2005. Disponível em: < http://books.evc-cit.info/OD_Essentials.pdf>. Acesso em: 23 de fevereiro de 2014.

ELSMASRI, Ramez; NAVATHE, Shamkant B.. Sistemas de Banco de Dados. São Paulo, SP: Pearson Education, 2005. 724p.

FERNEDA, E.. *Introdução aos modelos computacionais de recuperação de informação*. Rio de Janeiro, Editora Ciência Moderna, 2012.

JANSEN, W.; AYERS, R. *Computer Security - guidelines on cell phone forensics, NIST - Special Publication 800-101*, 104p. 2007. Disponível em: <http://csrc.nist.gov/publications/nistpubs/800-101/SP800-101.pdf>. Acesso em: 07 de abril de 2013.

DALBEN JR, Osvaldo Dalben; CLARO, Daniela Barreiro. *Uma análise do reconhecimento textual de nomes de pessoas e organizações na computação forense*. Brazília, ICOFCS, 2011, v. 6, p. 7-15.

FIGUEIREDO, T. *Sisbala: A solução para as armas ainda não periciadas no país*. Revista: Perícia Federal, APCF, XIII, n. 29, p. 14 - 15, 2012.

FREITAS, C. O. A. *Módulo Temático: Perícias e Laudos Técnicos*. Curitiba, PR: PUC-PR, 2008. 62 p.

GANTZ, J.; REINSEL, D.. *The digital universe in 2020: Big Data, nigger digital shadows, and biggest growth in the far east*. IDC, 2012.

GREENFIELD, A. *Everyware: The dawning age of ubiquitous computing*. Berkeley: New Riders, 2006.

GROCHOCKI, Luiz Rodrigo: *Escola de Governo: Palestra Nº 51: Aspectos Jurídicos do uso de Computadores - Internet*. 2013. Disponível em: <http://webcast.pr.gov.br/escoladegoverno/historico.php?evt_id=52>. Acesso em 26/02/2014.

GROTH, Robert. *Data Mining: A Hands-On Approach for Business Professionals*. Prentice Hall PTR, 1997. 264p.

HOUAISS, A.; VILLAR, M. d. S. *Minidicionário Houaiss da língua portuguesa*. Objetiva, 2004.

IMONIANA, J. O. *Auditoria de Sistemas de Informação*. Editora Atlas, 2005.

JANSEN, W.; AYERS, R. *Computer Security - guidelines on cell phone forensics, NIST - Special Publication 800-101*, 104p. 2007. Disponível em: <http://csrc.nist.gov/publications/nistpubs/800-101/SP800-101.pdf>. Acesso em: 07 de abril de 2013.

KRUSE, W.G.; HEISER, J.G. *Computer forensics: incident response essentials*. Indianapolis: Addison-Wesley, 2002.

KUECHLER, William L. *Business applications of unstructured text*. Communications of the ACM, 2007, v. 50, p. 86-93.

LAFORE, R. *Estruturas de Dados e Algoritmos em Java*. Editora Ciência Moderna, 2004. 702p.

LE COADIC, Yves-François. *A Ciência da Informação*. Lemos Informações e Comunicação Ltda, 1996. 115p.

MALLMANN, J.; FREITAS, C. O. A.; SANTIN, A. O. *Produção de provas digitais a partir de rastreamento em relacionamentos por e-mails*. X Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais - SBSeg, v. 2010, 2010.

McLAUGHLIN, B. *Java and XML*. Mike Loukides, 2000. 479p.

MICHAUD, D.J. Adventures in computer science. SANS Institute, 2001.

MITNICK, K., SIMON, W., WOZNIAK, S., *Ataques de Hackers: Controlando o Fator Humano na Segurança da Informação*, São Paulo, SP: Makron Books, 2003.

MOOERS, C.N. *Zatocoding applied to mechanical organization of knowledge*. American Documentation, v. 2, p. 20-32, 1951.

NORVIG, Peter., RUSSELL, Stuart. *Inteligência Artificial*. 3ª Edição. Vol. 1. Elsevier Brasil, 2013.

REIS, A. B. d. *Metodologia científica em perícia criminal*. Campinas, SP: Millenium, 2011. 262p.

REZENDE, D. A. *Sistemas de Informação e as Decisões Gerenciais na Era da Internet*. Editora Atlas, 2007.

REZENDE, SOLANGE O.; RICARDO M. MARCACINI; MARIA F. MOURA. *O uso da mineração de textos para extração e organização não supervisionada de conhecimento.* Revista de Sistemas de Informação da FSMA, v. 7, 2011, p. 7-21.

RIBEIRO, C. J. S. *Diretrizes para o projeto de portais de informação: uma proposta interdisciplinar baseada na Análise de Domínio e Arquitetura da Informação*. Diss. Tese (Doutorado em Ciência da Informação) – Convênio UFF/IBICT, Rio de Janeiro, 2008.

RICCI, H.C.; FREITAS, C.O.A. *OS Títulos de Crédito Eletrônicos e sua (In)Compatibilidade com os Princípios do Direito Cambial: por uma mudança de paradigma frente aos Documentos Eletrônicos*. Revista Jurídica CESUMAR. Mestrado, v. 12, p. 439-461, 2012.

ROSA, M. V. F. *Perícia Judicial - Teoria e Prática*. Sérgio Antônio Fabris Editor, 1999.

SALTON, Gerard; MCGILL, Michael J.. *Introduction to modern information retrieval.* McGraw-Hill Book Co., New York, 1983.

SCHMIDT, S.; MANSCHITZ, S.; RENSING, C.; STEINMETZ, R.. *Extraction of Address Data from Unstructured Text using Free Knowledge Resources*. 13th International Conference on Knowledge Management and Knowledge Technologies. ACM, 2013. 8p.

SHRESTHA, Dilpesh. Text mining with Lucene and Hadoop: Document clustering with feature extraction. Wakhok University, 2009, 51p.

SILVA, A. *Sistema de gerenciamento de informações periciais*. In: XXI Congresso Nacional de Criminalística. 2011.

SIMÃO, A. *et al. Aquisição de evidências digitais em smartphones android*. 2011.

SINGH, Harry S.. *Data Warehousing: Concepts, Technologies, Implementations, and Management*. Prince Hall PTR, 1997. 332p.

SUBRAMANIYASWAMY, V.; VIJAYAKUMAR, V.; LOGESH, R.; INDRAGANDHI, V.. *Unstructured Data Analysis on Big Data Using Map Reduce*. Procedia Computer Science, 2015, v. 50, p 456-465.

STEPHENSON, PETER. *Investigating computer-related crime : handbook for corporate investigators*. CRC Press LLC, 2000. 295p.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. Rio de Janeiro, RJ: Editora Ciência Moderna, 2009. 900p.

VRUBEL, A. *Using xslt filters to improve productivity and quality on cell phone forensics*. In: 6th International Conference on Forensic Computer Science (ICoFCS2011). 2011. p. 132 - 136.

VOLPI NETO, A.; FREITAS, C.O.A. *Forense computacional*. Jornal do Estado - Coluna Atualidades Legais, 2008. Disponível em: <http://www.jornaldoestado.com.br>.

WEISER, M. *The computer for the 21st century. Scientific American*, v. 265, n. 3, p. 94 -104, 1991. Disponível em: <https://www.cs.cmu.edu/afs/cs/Web/People/jasonh/courses/ubicompsp2007/papers/02-weiser-computer-21st-century.pdf>. Acesso em: 07 de outubro de 2013.

WEISER, M. *Some computer science issues in ubiquitous computing. Communications of the ACM*, v. 265, n. 3, p. 137 - 143, 1993. Disponível em: <http://www.ubiq.com/hypertext/weiser/UbiCACM.html>. Acesso em: 07 de outubro de 2013.

YOUNG, Susan; AITEL, Dave. *The hacker's handbook : the strategy behind breaking into and defending Networks*. CRC Press LLC, 2004. 849p.

ZOBEL, Justin; MOFFAT, Alistair. *Inverted files for text search engines*. ACM computing surveys (CSUR), 2006, v. 38, n. 2, a. 6., 56p.

Anexo I

Ao Diretor Geral do Instituto de Criminalística do Paraná
Dr. Hemerson Bertassoni Alves
Nesta Capital

Inicialmente, cabe explicar que desde 2013 o Instituto de Criminalística do Paraná (IC-PR) e a Pontifícia Universidade Católica do Paraná (PUCPR) estão trabalhando conjuntamente no desenvolvimento do Sistema de Cruzamento de Registros Telefônicos - SiCReT, haja visto as publicações já realizadas em conjunto, a saber:

- DECARLI, Alonso; GROKOSKI, Cícero; PARAISO, Emerson Cabrera ; GROCHOCKI, Luiz Rodrigo; FREITAS, Cinthia Obladen de Almendra . Banco de Dados de Laudos Periciais de Dispositivos Móveis. In: XIV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg'2014), 2014, Belo Horizonte. Anais do XIV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais. Porto Alegre: Sociedade Brasileira de Computação SBC, Manaus, AM, 2014. v. 1. p. 559-571;
- GROCHOCKI, Luiz Rodrigo; VRUBEL, Alexandre; ZAGO, Rafael; DECARLI, Alonso; FREITAS, Cinthia Obladen de Almendra. SICReT - Sistema de Cruzamento de registros Telefônicos. In: XIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg'2014), 2013, Manaus. Anais do XIII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg). Porto Alegre: Sociedade Brasileira de Computação, Belo Horizonte, MG, 2013. v. 1. p. 527-536.

Para realização deste projeto foi alocado o Mestrando Alonso Decarli, junto ao Programa de Pós-Graduação em Informática (PPGIa) da PUCPR, sob minha orientação e com bolsa Institucional CAPES (**EDITAL 04/2014 – Bolsa Capes MESTRADO**).

Neste contexto, urge a realização de experimento científico para prova de conceito e validação do sistema SiCReT e, portanto, solicita-se a utilização de base de dados de laudos periciais de dispositivos móveis da Seção de Computação Forense do IC-PR. Esta atividade será coordenada no Instituto de Criminalística do Paraná pelo Perito Criminal **Msc. Luiz Rodrigo Grochocki**, sendo que todos os cuidados em relação à proteção e sigilo dos dados serão devidamente atendidos.

Curitiba, 09 de Dezembro de 2014.

Profa. Dra. Cinthia O. de A. Freitas – Professor Titular da PUCPR
Coordenadora e Orientadora do Projeto

Anexo II

SECRETARIA DE ESTADO DA SEGURANÇA PÚBLICA
POLÍCIA CIENTÍFICA
INSTITUTO DE CRIMINALÍSTICA

Ofício nº 473/GAB/2014-IC

Curitiba, 09 de Dezembro de 2014.

Prezada Senhora:

Conforme solicitação requerida no ofício encaminhado por e-mail e datado de 04 de Dezembro de 2014, autorizo o aluno **ALONSO DECARLI** a ter acesso ao banco de dados de laudos periciais da Seção de Computação Forense, sob orientação e supervisão do Perito Criminal **Luiz Rodrigo Grochocki**.

Atenciosamente,

Hemerson Bertassoni Alves
Diretor do Instituto de Criminalística

A Ilustríssima Senhora
Profa. Dra. Cinthia O. de A. Freitas
Coordenadora e Orientadora do Projeto
Curitiba - Paraná

Pabx (041) 3281-5500 - Sol Laudo (041) 3281-5571 - Fax (041) 3281-5577 - Gab (041) 3281-5509
Site: www.pr.gov.br/ic E-Mail: iccwbgab@ic.pr.gov.br