As contribuições da Ciência da Informação na perícia em informática no desafio envolvendo a análise de grandes volumes de dados – *Big Data*

José Antônio Milagre
Universidade Estadual Paulista – UNESP, Email: ja.milagre@gmail.com
José Eduardo Santarem Segundo
Universidade de São Paulo – USP, Email: santarem@usp.br

Resumo

A Internet trouxe preocupações para toda a sociedade através do prisma da segurança da informação. Vislumbra-se vulnerabilidades que surgem a partir da utilização de tecnologias. Fraudes e crimes cibernéticos, aumentando de forma crescente, podem explorar estas tecnologias para causar danos sensíveis a empresas e pessoas. A perícia em informática, como a ciência que visa investigar os incidentes cibernéticos e fraudes tem que enfrentar esse cenário, não só de forma reativa mas proativamente, levantando informações que servirão de base para as decisões de inteligência, de modo a proteger a segurança da sociedade no ciberespaço. A questão se agrava quando se trata de Big Data, onde a perícia torna-se complexa e dificultosa. Este artigo apresenta os resultados de uma pesquisa básica, exploratória, realizada mediante levantamento bibliográfico. O objetivo da presente pesquisa foi conceituar a computação forense, apresentar o atual estagio da computação forense aplicada a grandes volumes de dados e avançar, propondo uma análise do problema por meio de elementos e conceitos da Ciência da Informação, o que certamente contribuirá para a construção de soluções eficazes para a análise de grandes volumes de dados que envolvam crimes cibernéticos, fraudes e incidentes.

Palavras-chave: Computação Forense. Investigações digitais. Big data. Ciência da informação. Internet das coisas. Segurança da informação.

1 Introdução

Entende-se como crime cibernético o crime cometido contra ou através das tecnologias de informação ou comunicação. Existem variadas formas de se praticar um crime cibernético, tendo-se em vista o dinamismo da tecnologia da informação. Neste contexto, apresenta-se como indispensável ao investigador digital delinear qual a ferramenta usada pelos atacantes para a ação ilícita que possa caracterizar um crime cibernético. (CAVALCANTE, 2013).

O crescimento do número de dispositivos conectados ou que acessam a *web* fizeram aumentar o número de pontos vulneráveis, o que vem tornando o combate ao crime cibernético extremamente difícil.

Entidades governamentais estão compreendendo a importância do compartilhamento de informações sobre ameaças e ataques, embora pouca evolução exista em termos da concepção de um padrão para compartilhamento de informações.

Em tempos atuais, unidades policiais trocam dados e informações, porém não existe uma unificação para estes dados. Acrescenta-se a isto o fato de que, quando fala-se de ilícitos digitais, muitos são os que não chegam até o conhecimento das autoridades.

Um dos desafios da perícia cibernética é a dificuldade de analisar grandes volumes de dados envolvendo tais atividades. A Internet vem se tornando um local onde uma massiva quantidade de dados é gerada a cada dia. Isto é *Big Data*, que pode ser entendido não apenas

como um conceito abstrato criado pelo Universo da Tecnologia da Informação, mas uma forte tendência da pulsante atividade digital (OHL, 2014).

Os métodos tradicionais de análise de crimes estão desatualizados e não podem ser considerados integralmente na análise de grandes volumes de dados. A perícia em informática *(computer forensics)* sempre teve como escopo apurar a existência de um delito e apontar a autoria do mesmo, delito este cometido por meio ou tendo como alvo a tecnologia da informação.

Porém, a perícia pode ser aprimorada e pode revelar informações sobre cada ameaça cibernética, suas motivações, atores, modelos e técnicas. É preciso considerar a análise de metadados sobre informações ou evidências de incidentes bem como expandir as técnicas e ferramentas no campo da perícia em informática. Neste contexto apresenta-se como relevante o uso dos conceitos da Ciência da Informação para alavancar a organização da informação nas áreas afetas à Computação Forense.

Áreas concentradas de crimes são chamadas de *hotspots* (BEATO *et al.*, 2008). Existe uma série de *hotspots* no ambiente cibernético. Por mineração de dados ou *"mining"* denomina-se o processo de extrair conhecimento de grandes quantidades de dados. Este processo pode ser usado em casos envolvendo potenciais fraudes ou crimes cibernéticos. Neste sentido, apresenta-se o objetivo desta pesquisa que trata de apresentar a computação forense, discutindo o atual estagio da computação forense aplicada a grandes volumes de dados e avançar, propondo uma análise do problema por meio de elementos e conceitos da Ciência da Informação, o que certamente contribuirá para a construção de soluções eficazes para a análise de grandes volumes de dados que envolvam crimes cibernéticos, fraudes e incidentes.

2 Computação forense, desafios atuais e a importância da Ciência da Informação

Brian Carrier (2002) esclarece que a Computação Forense utiliza-se de métodos científicos para a preservação, coleta, validação, identificação, análise, interpretação, documentação e apresentação da evidencia digital derivada de fontes digitais.

A computação forense pode recair sobre um maquina desligada (*dead analysis*) ou sobre um equipamento ligado (*live analysis*). Ainda, a computação forense pode recair sobre redes, coletando-se dados que possam servir para evidências de um incidente.

Sean Mclinden (2010) aborda a temática do que denomina como forense digital preventiva, emprestando da medicina diagnóstica, o *positive predictive value* (PPV), uma forma de antecipar ações ou supostos crimes praticados no campo da informática.

Da análise da literatura dedicada ao tema, percebe-se que a Computação Forense tende a ser uma ciência reativa. Pouco se fala em Computação Forense como análise de comportamento da fraude e do fraudador. O conceito de *"predictive coding"* ainda é absolutamente embrionário, onde submetemos dados a análises de computadores preparados para traçar padrões e comportamentos, antecipando uma possível fraude, ainda não praticada.

Dentre os desafios atuais para a Computação Forense em grandes volumes de informações, descobriu-se:

a) A grande quantidade de dados em formatos diferenciados, textos não estruturados e arquivos multimídia: Coletar informações *web* ou em rede implica estar preparado para lidar com inúmeros formatos e principalmente, para lidar ou interpretar imagens e conteúdos visuais e dados não estruturados;

b) A ausência de profissionais atuando em Computação Forense que compreendam conceitos e técnicas da Ciência da Informação (metadados, ontologias, arquitetura da informação, métodos de indexação, estudo quantitativos e sobre recuperação da informação, entre outros) e que possam desenhar as melhores técnicas para tratamento da informação relevante a uma investigação;

c) A ausência de fontes únicas de dados, sendo que uma perícia envolvendo *Big Data* pode recair sobre toda a Internet, trafego de rede interno, redes sociais, mecanismos de busca ou mesmo sobre arquivos e e-mails armazenados em um dispositivo de armazenamento;

d) Os problemas envolvendo a privacidade, localização dos dados e autorizações legais para a coleta e tratamento de informações;

e) A obsolescência e ausência de ferramentas que possam lidar com grandes quantidades de *datasets*;

f) A volatilidade da informação na Internet, que se não coletada rapidamente pode exaurir-se ou alterar seu contexto a qualquer momento;

g) Os pedaços informacionais distintos nas páginas da Internet, o que demanda um procedimento claro de recuperação, evitando-se ao máximo ruídos ou lixo eletrônico, imprestável a uma investigação;

h) A ausência de cooperação por parte dos provedores de serviços que hospedam ou armazenam os dados gerados por suspeitos, como a ausência de um padrão para interconexão com autoridades e investigadores, para manipulação a tais dados.

O rol citado, meramente exemplificativo, elucida os desafios que a Perícia em Informática ou Computação Forense enfrenta ao se deparar com terabytes de dados e informações, sendo muitas vezes informações dispostas na Internet e em ferramentas de redes sociais de forma desestruturada.

A Ciência da Informação, neste estágio de criminalidade cibernética e limitações dos métodos tradicionais utilizados na Computação Forense, tem papel relevante e passa a ser de importante contribuição no tema, especialmente, no desenvolvimento de modelos, projetos e processos interessantes para a construção de ferramentas e metodologias de investigação e auditoria digital, que sejam mais eficientes e considerem grandes volumes de informações.

O emprego e revisão dos processos de investigação, aplicando-se o aporte de áreas de estudo da Ciência da informação, como protocolos, metadados, ontologias, técnicas de indexação, estudos quantitativos e as tecnologias da *web* semântica podem ser fortes aliados no aprimoramento da atividade da computação forense.

Reis *apud* Paiva (2014) ao esclarecer que um dos objetivos básicos da organização da informação é evitar a sobrecarga informacional, adverte que a oferta excessiva da informação, marcada pela era digital, provoca sensação de angústia. Braga (2010, p.2), consigna que o excesso de informação é a neurose do Século XXI, apresentando como causa da explosão informacional o barateamento dos custos de publicação, transmissão e arquivamento. Esta pode ser a sensação de um perito em computação forense ao analisar um "mundo de dados" apreendido ou disponibilizado para análise.

Wurman (1995) apresenta importante trabalho denominado "ansiedade da informação" onde esclarece o surgimento de profissões que serão fundamentais no objetivo de organizar e gerar conhecimento a partir do mar informacional gerado com a tecnologia da informação.

Atualmente, o acesso nativo às interfaces disponibilizadas pelas ferramentas de redes sociais é dificultoso e pouco interativo. É desafio contornar estas restrições de interfaces, porém sabe-se que a relação governada pela interface deve ser uma relação semântica, caracterizada por significado e expressão, e não por força física. (Johnson, 2001)

A Ciência da Informação tem o papel de auxiliar nos estudos que tenham por objetivo a compreensão da relação entre usuário e os sistemas informacionais (LE COADIC, 1996). Neste cenário, estudos envolvendo a organização e redução da informação, interfaces, interoperabilidade e inteligência coletiva, dentre outros, poderão favorecer a implementações de soluções que façam frente à necessidade de se investigar ocorrências no mundo dos *"petabytes"*.

Não se pode descartar, igualmente, o campo da Netnografia como uma área de contribuição na análise forense de grandes volumes de dados, esta que segundo Bentes *Pinto et al.* (2007), constitui-se em uma abordagem para estudo dos usuários na *Web*, de seus comportamentos. Uma solução em computação forense pode implementar técnicas de netnografia para, dado um determinado contexto, conceber uma ontologia, observar a interação de um usuário e compreender seu significado.

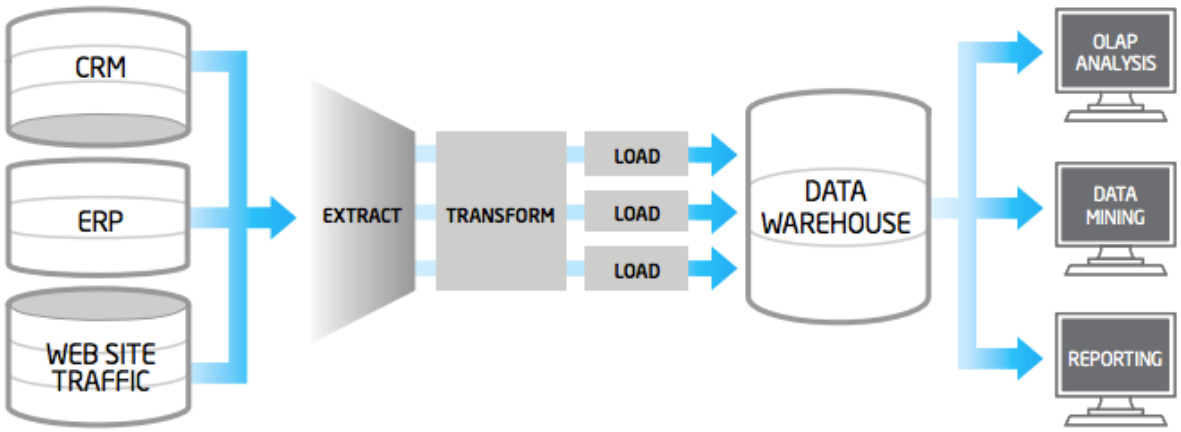
3 Estado da arte da pesquisa envolvendo a computação forense em Big Data

São recentes as pesquisas que tratam da computação forense em ambientes com grandes volumes de dados. Beebe e Clark (2005), abordam a necessidade da integração da mineração de dados com a Computação Forense, o que proporcionaria redução de custos em uma atividade forense, redução de recursos humanos e tempo.

As ferramentas de computação forense existentes, de certo modo, subutilizam o poder de processamento dos computadores. Do mesmo modo, evidencia-se que a mineração de dados, importante na coleta e redução de grandes volumes de dados, incorpora uma série de disciplinas dentre as quais cita-se a ciência da computação e ciência de informação.

No mundo, algumas consultorias já apresentam o conceito de "*Extract, Transform & Load"(ETL)* um processo para extração de dados em diversas fontes e consequente ajustes às necessidades analíticas de uma pessoa, em um *data wareshouse* – Figura 1 (INTEL, 2010):

Figura 1 - Processo ETL



Fonte: (INTEL, 2013)

No entanto, não foram identificadas aplicações ETL moduladas ou voltadas para a Computação Forense. Uma proposta plausível de análise de dados envolvendo fraudes e crimes na Internet, com o uso de *data mining*, o que incrementaria a efetividade da computação forense, seria converter todos os artefatos *web* para texto plano, após, classificá-los como eventos, e na sequência, a utilização de diversas técnicas de mineração de dados para se extrair informação e conhecimento. Tal técnica não alcançaria sites protegidos e imagens que hoje podem ser utilizadas para prática de crimes e troca de informações confidenciais por meio de esteganografia, arte de se esconder uma informação sobre ou dentro de outro artefato informacional.

Hosseinkhani *et. al* (2012) estuda também as técnicas de *data mining* aplicadas à detecção de informações suspeitas, considerando *data mining* como o processo de descoberta, extração e análise de padrões compreensíveis, estrutura, modelos e regras para grandes quantidades de dados. Fica claro que a mineração de dados cresce como uma técnica para detecção do crime, principalmente na tecnologia da informação. Para o pesquisador em análise, uma ferramenta *"crawler"* (programa capturador) poderia capturar páginas de tempos em tempos. Após isso, uma ajuste com *"parsers"* (programas organizadores) e a mineração ocorreria, por meio de interfaces de pesquisa. Tal princípio é muito utilizado em aplicações que hoje desenvolvem a gestão da reputação online onde, após organizados os dados, fazem a leitura e detectam se um conteúdo é positivo, negativo, ofensivo ou neutro em relação a alguma pessoa física ou jurídica.

Do mesmo modo, Perner (2003) e Djeraba *et al.* (2007) possuem contribuições importantes no campo da extração de conteúdos multimídia da Internet.

Por sua vez, Vashisht *et al.* (2013) propõe uma análise de detecção do crime cibernético com base na análise do comportamento do usuário, o que é uma contribuição significativa, pois como é cediço, muitos crimes informáticos permanecem impunes pelo anonimato, tendo em vista a falta de legislação e cooperação dos provedores responsáveis ou que servem de base para a prática do crime, no fornecimento de dados que apontem a autoria de um delito. O comportamento poderia revelar a autoria e este processo pode ser automatizado para análise de grandes volumes de dados.

Em que pesem os estudos acima identificados, não se encontram estudos que relacionem a Ciência da Informação com Computação Forense, tampouco que considerem suas áreas centrais e campos às atividades de recuperação e tratamento de dados relativos a incidentes, fraudes e crimes.

A adoção de soluções isoladas, sem um modelo ou melhores práticas, pode culminar na geração de evidências ilícitas, não reproduzíveis ou mesmo inadmissíveis em um processo legal. Ademais, informações precisas e relevantes podem ser simplesmente desconsideradas.

Avaliada a literatura identificada sobre o assunto, passa-se a descrever as melhorias e contribuições que a Ciência da Informação pode conceber a embasar novas soluções sobre a temática abordada.

4 Contribuições da Ciência da Informação para aprimoramento dos sistemas de análise de Big Data

Compreender a informação em uma investigação digital é atitude preliminar e necessária. A Ciência da Informação é a disciplina que investiga as propriedades e o comportamento informacional (BORKO, 1968).

Ao examinador forense é preciso investigar se está lidando com dados não estruturados ou estruturados e principalmente, cabe compreender as forças que governam o

fluxo da informação. Por meio da Ciência da Informação, será possível aplicar as melhores técnicas de organização, classificação e indexação aos grandes volumes de dados.

Um dado pode ser tornar um artefato, que pode se tornar uma informação, que pode ser tornar uma evidência, mas sobretudo, que precisa ser documentado.

Ao examinador forense é interessante que antes de coletar, armazenar, processar, ou analisar, saiba que está exercendo a documentação, concebendo processos e relatórios para a perfeita recuperação dos dados coletados, com finalidades de servirem de prova em um caso administrativo ou judicial. Esta tarefa imprescinde da organização da informação.

Destaca-se igualmente que o processo de Computação Forense, para que tenha validade, deve ser documentado. Ocorre que, no caso envolvendo grandes volumes de dados, poderá ser a máquina o agente que analisa as informações de forma automatizada, logo, deverá ser criado um processo de documentação de cada tarefa executada pelo sistema de análise forense, valorando a prova coletada.

Seja na segurança da informação ou na Computação Forense, **a arquitetura dos sistemas e dados** imprescinde do olhar da Ciência da Informação. Os profissionais da área podem aplicar as teorias e técnicas da ciência para melhorar sistemas de manipulação da informação. (BORKO, 1968)

Em se tratando de investigação cibernética em ambientes *web* ou que envolvam múltiplos usuários, é possível categorizar comportamentos de usuários, definir ontologias de atividades suspeitas atinentes a determinados crimes, identificar padrões que possam avaliar intenções, competências, potenciais atividades criminosas ou envolvimento com delitos.

O uso de **ontologias** no contexto de *web* semântica pode ser implementada em aplicações forenses para organizar dados para detecção da fraude humana no espaço virtual. É possível, por exemplo, definir ontologias relativas a perfis falsos, fraudes empresariais, e outros delitos, associando a expressões e comportamentos que poderiam indicar atividades suspeitas. Estas atividades serão categorizadas e lançadas em base de suspeição, proporcionando uma significativa redução de dados e ganho na interpretação de cenários, permitindo ao perito realizar associações.

Gruber (1993, p. 1) define ontologia como "uma especificação explícita de uma conceitualização". Santarem Segundo e Coneglian (2015, p. 227), complementam o conceito de ontologias, no contexto de aplicação, dizendo que "[...] entende-se as ontologias como: artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de: classes, propriedades, relações, restrições, axiomas e instâncias".

O **processamento de linguagem natural** pode ser utilizado para detectar fraudes. Técnicas para interpretação automatizada de linguagem humana, uso de folksonomia e vocabulários controlados podem ser aplicadas em plataformas forenses, o que permitirão a capacidade de detectar termos usados por criminosos digitais.

A utilização de **metadados** sobre registros de atividades ou artefatos em rede ou *web* é também tema basilar para aprimoramento dos sistemas de investigação que considerem grandes volumes de dados. A possibilidade de descrever objetos, postagens, comentários de acordo com o processamento considerando ontologias e processamento de linguagem natural, pode facilitar a rastreabilidade de ofensas na rede, sendo possível inclusive apurar a origem e nível de proliferação na rede de uma atitude criminosa.

Alves (2010) indica que "a necessidade de representar as informações em diversas áreas do conhecimento, em distintos domínios, fez com que surgissem variados tipos de padrões de metadados. Tais padrões apresentam características específicas, apresentando diferentes tipos de estruturas, desde estruturas mais simples, até estruturas complexas, com um grande número de elementos".

Os métodos de classificação e indexação também devem ser considerados quando o escopo é organizar a informação relativa a incidentes de segurança. A Ciência da Informação poderá, de acordo com as características dos crimes cibernéticos, estabelecer grupos e gerar uma classificação, bem como relacionamentos entre os grupos.

Considerar-se-á também a questão da **interoperabilidade**, visto que ainda é um grande desafio para a apuração de crimes eletrônicos a dificuldade de conexão entre sistemas de detecção de incidentes ou mesmo na geração de um padrão adaptável aos provedores de conteúdo e serviços, que registram os acessos às aplicações e conexão na Internet. A construção de um padrão para armazenamento e comunicação destas informações é fundamental para o fortalecimento da segurança da informação, possibilitando que autoridades de aplicação de lei e empresas possam acessar e trocar dados sobre acessos, criações de contas e usos maléficos de serviços, em ambiente protegido, com velocidade.

Deste modo, conceitos e técnicas advindas da Ciência da Informação são fundamentais para o processo de reconhecimento e tratamento da informação, não só textual, mas em formato imagem ou multimídia. São raras as ferramentas forenses disponíveis no mercado que realmente compreendem o que as imagens significam ou querem dizer.

Por fim, não se cogita de qualquer solução ou técnica de análise forense de grandes volumes de dados, que não considere o ***data mining***, o processo para explorar grandes volumes para identificar padrões, associações ou sequenciais que extraiam conhecimento e novas informações. A escolha do melhor algoritmo ou a construção de técnicas de identificação, coleta, preservação, análise e representação da informação relativas a fraudes e golpes na Internet será atribuição do profissional de Ciência da Informação, que pode incluir métodos de **gestão do conhecimento** e/ou **inteligência competitiva**, de forma a arquitetar o cenário informacional para que os peritos possam desempenhar as atividades de coleta e análise de evidências.

5 Tecnologias e processos para computação forense em grandes volumes de dados

Passa-se a investigar técnicas e ferramentas que possam ser utilizadas em casos envolvendo incidentes, fraudes ou crimes eletrônicos, em repositórios de grandes volumes de dados.

Tais técnicas podem ser utilizados para tanto para a Computação Forense, onde o escopo é reconstruir é apurar a autoria e materialidade de um crime, como para o *e-discovery*, onde o escopo é revisar grandes volumes de dados, ou mesmo para o *analytics,* que objetiva analisar grandes bases de dados. Neste contexto, o *data mining* apresenta-se como processo fundamental.

É preciso fazer uma distinção entre *data mining* e *web mining*. Enquanto no primeiro a coleta de informações se dá em centros de armazenamentos como os *Data Warehouses*, na *web mining*, os dados são colhidos mediante o rastreamento de dados em páginas *web* (*crawlers*).

Como salienta Hosseinkhani (2012, p. 2-3), páginas na Internet são diferentes de dados ou documentos de texto convencionais pois possuem âncoras, *hyperlinks* e demais elementos que não existem em documentos tradicionais. Páginas da Internet não são simples parágrafos, mas são estruturadas, podendo ser simples parágrafos, metadados, corpo, código, dentre outros. Nem todos os blocos são importantes para uma investigação digital e é necessário fazer a separação.

Acrescente-se que pessoas podem se passar por outras na rede, onde estamos diante dos perfis falsos ou *fake profiles*. O modelo tradicional de mineração de dados simplesmente classificam padrões em dados estruturados, como classificação, predição, análise de associação ou de cluster.

Na Computação Forense, é possível extrair métricas de softwares, bancos de dados, características na programação, variáveis e outros indicadores para se buscar, por exemplo, códigos potencialmente ofensivos e pessoas responsáveis por tais códigos ou mesmo a aproximação da sua localização geográfica.

Em investigação de crimes cibernéticos, apurar a autoria é algo complexo, considerando que a ameaça vem de longe ou mesmo age no anonimato. Dentre as principais técnicas de *data mining* de crime temos a clusterização, mineração de regra de associação, classificação e mineração de padrão sequencial, as quais podem ter relevância em uma auditoria em informática, em uma de suas especificidades (Hosseinkhani, 2013), como se apresenta:

- Técnicas de clusterização: Ocorre o agrupamento de dados em classes de características similares. Na sua modalidade *"link analysis"* é possível identificar padrões e transações similares, sendo possível associá-las a grupos criminosos específicos ou a localizações similares;

- Mineração de regras de associação: Nesta técnica é possível detectar a frequência de um registro em um banco de dados, e neste sentido, identificar a frequência, por exemplo, de um ataque ou tentativa de ataque, o que pode ser utilizado como uma regra para auditoria ou no sistema de segurança da informação.

- Detecção de desvio: Basicamente consiste em identificar dados que diferem visivelmente dos demais dados. Importante padrão que deve ser utilizado na análise de logs e principalmente nos casos de intrusão de redes e fraudes, na análise de pacotes da rede. Também é chamada de *"outlier detection"*. Pode ser útil na detecção de perfis falsos nas ferramentas de redes sociais.

- Comparação de *strings*: Método muito comum na Computação Forense, consiste em comparar campos textuais de registros distintos o que pode identificar, em casos de crimes cibernéticos, a existência de determinado conteúdo em disco, rede ou mesmo fraudes, tentativas de invasão e golpes ou frases suspeitas.

Igualmente, a categorização é técnica que pode ser considerada na Computação Forense. A categorização de textos aplicada a computação forense em *big data* pode ser dar, por exemplo, envolvendo palavras positivas, negativas, combinações, dentre outras, e ser associada também a uma análise sintática.

Assim como a filtragem de textos sobre sintomas pode ser utilizada para definir um possível diagnóstico, a filtragem de textos de um usuário pode ajudar a apuração da autoria de determinada fraude ou associar alguém a uma prática ilícita ou ainda, apresentar um possível sentimento do investigado.

Ainda no âmbito da mineração de dados, neste ponto alinhada a inteligência artificial, importante destacar que através da clusterização, poderá ocorrer a busca por características comuns em grupos. Pode-se utilizar, igualmente, a extração de textos que comparados a uma base de palavras pode indicar um conceito, também previamente definido.

Esta proposta elucida um conceito denominado *knowledge-discovery in text (KDT),* onde é possível extrair conhecimento de dados com aplicação em diversas áreas como médica, jurídica, concorrencial, policial, inteligência competitiva, dentre outras. A técnica

ajuda o usuário a descobrir informações até então desconhecidas, por isso está afeta a inteligência artificial.

Utilizando conceitos e linguística, o processamento de linguagem natural permite aproveitar o máximo conteúdo do texto, extraindo entidades (ARANHA; PASSOS, 2006).

Esta subárea da inteligência artificial pode ser aplicada à Computação Forense na conversão de linguagem de banco de dados de computadores ou de ferramentas de redes sociais, envolvendo segurança da informação e incidentes, em uma linguagem compreensível, o que irá gerar velocidade na resposta a um incidente.

Do mesmo modo, integrante do conceito de inteligência artificial, as redes neurais crescem em aplicação para a detecção de fraudes, principalmente envolvendo bancos e uso de cartão de crédito. Fraudes em cartão e transações financeiras custam 2,3 bilhões no Brasil (ROSA, 2014).

Com as redes neurais torna-se possível aprender com as fraudes, que são cadastradas, catalogadas, e pontuadas, onde tal pontuação é usada para comparar a similaridade entre um golpe e outro, o que pode gerar conhecimento sobre local, autoria, responsáveis pela fraude, dentre outros dados relevantes. Logo, pode-se afirmar que tais redes aprendem com a experiência, assim como os seres humanos.

Não se tem conhecimento de redes neurais aplicadas à computação forense em grandes volumes de dados, porém as redes hoje utilizadas para detecção de fraudes podem ser preditivas diante de um futuro crime ou de uma situação que mereça uma ação preventiva.

Um modelo computacional identificado na pesquisa é o *Map-Reduce*, desenhado para processar grandes volumes de dados, desenvolvido pelo Google e que apresenta um modelo de computação distribuído.

Aplicações para computação forense e investigação digital poderão considerar o padrão *Map-Reduce*, o que tornará os programas preparados para lidar com grandes quantidades de dados. O modelo vem para fazer frente a limitação de processamento de grandes volumes de informações.

Na investigação da fraudes, também em grandes repositórios de dados, *red flags* são consideradas um conjunto de circunstâncias que não são usuais em determinada atividade. É importante lembrar que as *red flags* não indicam precisamente culpa ou inocência (DINAPOLI, 2011). A técnica já é comum na disciplina envolvendo detecção da fraude e poderá integrar sistemas de Computação Forense.

Em uma ferramenta envolvendo computação forense preventiva em grandes volumes de dados, as *red flags* devem ser implementadas nas aplicações e assim proporcionar ao investigador conhecimento ou indícios de uma fraude, como por exemplo:

- Alteração no estilo de vida dos colaboradores: novas aquisições, carros importados, etc;
- Problemas com crédito e débito;
- Mudanças de comportamento: uso de jogatinas, drogas, álcool, etc, sites acessados, etc.

Como se percebe, a adoção de *red flags* é sinalização importante para o sucesso de um processo de análise de grandes volumes de dados, na investigação de fraudes e crimes informáticos.

Embora esteja afeta a técnica de *"data mining",* o algoritmo K-Means também merece destaque por ser uma interessante técnica de clusterização, já citada neste trabalho. Poucas são as produções técnicas que estudam o uso das áreas da Ciência da Informação na

Computação Forense, mas grande parte da produção existente trata do algorítimo em questão, quando o desafio envolve grande volume de dados.

Utiliza-se o algoritmo K-Means para separar dados por categoria. A ideia do K-Means é fornecer uma classificação de informações de acordo com os próprios dados. (PICHILIANI, 2006, p.1). O algoritmo trabalha sem nenhuma classificação humana, sem a formação de uma tabela pré-existente, o que o enquadra no conceito de algoritmo de mineração de dados não supervisionado.

Destaque no algoritmo é que o investigador indica o numero de classes (clusteres) que deseja e o sistema, mapeando os dados, calculando a distância entre eles (comumente a euclidiana), inicia a classificação, inclusive, apurando ou refinando a classe de acordo com a média dos valores de cada atributo.

Na Computação Forense, sobretudo na análise de dados em ferramentas de redes sociais, tal aplicação é de suma relevância, eis que pode predizer atividades criminosas ou ofensivas. Como visto, muitas são as abordagens relacionadas a Ciência da Informação que devem ser consideradas na construção de soluções que pretendam encontrar significado na análise grandes volumes de dados, na atividade de Computação Forense, considerando que não foram identificados modelos, padrões ou guias concebidos até o término da pesquisa.

6 Considerações finais

O crime eletrônico é crescente e muitos indícios podem ser identificados na Internet ou em grandes volumes de dados, hoje espalhados por diversas fontes, em formato desordenado, sendo provado com a pesquisa que o *big data* está mudando o cenário a Computação Forense, que já não atende o cenário com suas técnicas. (CÁRDENAS *et al*. , 2012).

Verificou-se que a Computação Forense ou Perícia em Informática é ciência que tem objetivo de investigar incidentes e fraudes do ambiente informático. Demonstrado que o *big data* oferece potencial informacional para a Computação Forense, mas que hoje o grande desafio é a extração de conhecimento deste volume de dados.

Não foram identificados propostas ou modelos para exercício da Computação Forense em grandes volumes de dados, embora a comunidade científica já reconheça os desafios. Arrolou-se na pesquisa os desafios para a atividade da Computação Forense em grandes volumes de dados, dentre eles, a grande quantidade de dados em formatos diferenciados e não estruturados, ausência de fontes únicas de dados e a volatilidade da informação.

Apresentou-se áreas e campos da Ciência da Informação com potencial para auxílio à construção de propostas e guias para Computação Forense em *big data*, como a web semântica, o trabalho colaborativo, a arquitetura da informação, ontologias, e os conceitos envolvendo a netnografia e inteligência artificial.

No levantamento bibliográfico realizado, verificou-se que grande parte das fontes sobre Computação Forense em grandes volumes de dados tratam de *data mining*, como Beebe e Clark (2005) e Hosseinkani *et. al*. (2014). Identificamos contribuições para a computação forense envolvendo o conceito de "extração, transformação e carregamento" (ETL) e a extração de conteúdos multimídia da Internet, propostas por Perner (2003) e Djeraba (2007). Identificada igualmente uma proposta para análise e detecção de crimes cibernéticos com base no comportamento do usuário (Vashisht *et al.*, 2013)

Explorou-se algumas técnicas afetas ao *data mining* e inteligência artificial que podem agregar na construção de propostas, práticas e sistemas de identificação de fraudes em grandes volumes de dados mais efetivos. Evidenciou-se assim técnicas e tecnologias úteis à

investigação digital em grandes volumes de dados seja em *"data mining"* ou *"web mining"*, algumas, já utilizadas em detecções de fraudes.

Apresentou-se assim as contribuições para o aprimoramento dos sistemas de análise de *big data.* Uma solução identificada para perícia em web foi tratar seu conteúdo como eventos, os associando aos seus respectivos responsáveis, níveis de participação, data, hora, dentre outros metadados disponíveis. A análise semântica de eventos aplicada a informações sobre de condutas na *web* podem fornecer elementos rápidos detecção de crimes, o que demoraria anos para com base em métodos convencionais e análises humanas. Do mesmo modo pode-se construir ontologias relativa a crimes praticados nas redes associando categorias a termos e objetos suspeitos.

Foi possível concluir que antes de se cogitar em Computação Forense ou mineração de dados é necessário refletir sobre um padrão, melhor prática ou metodologia que possa ser adaptada e derivada, servindo de base para soluções futuras neste campo e que considere a arquitetura da informação em grandes volumes de dados.

Identificou-se claramente que, com a aplicação de campos ou áreas da Ciência da Informação à Computação Forense, envolvendo grandes volumes de dados, ocorrerá um ganho perceptível no desenvolvimento de práticas e aplicações e mesmo na profundidade e eficiência das perícias, que poderão envolver detecções proativas de incidentes, registro e descrição constante das atividades e objetos na *web* e principalmente, prever futuros eventos com agilidade, revelando informações ocultas e gerando base de conhecimento para antecipações e tomada de decisões estratégicas no domínio da Computação Forense.

The contributions of Information Science on computer forensics in the challenge involving the analysis of large data volumes - Big Data

Abstract

The Internet has brought concerns to society through the prism of information security. Vulnerabilities in sight that arise from the use of technologies. Cyber frauds and crimes, increasing in the world, can exploit these technologies to harm sensitive to companies and individuals. The computer forensics, as the science that seeks to investigate cyber-incidents and fraud must face this scenario, not just reactively but proactively, getting information for clarification and detection of incidents so as to contribute to the security of society and reducing impunity in relation to offensive conduct in cyberspace. The issue is compounded when it comes to big data, where expertise becomes complex and troublesome. This article presents the results of a basic research, exploratory, conducted through literature review. The aim of this research was to conceptualize forensic computing, present the current stage of computer forensics applied to large volumes of data and advance, proposing an analysis of the problem and presenting the important contribution of the Information Science in Computer Forensics activity through their areas of study, which will undoubtedly contribute to building effective solutions for the analysis of large volumes of data involving cybercrime, fraud and incidents.

Keywords: *Computer Forensics. Digital investigations. Big data. Information science. Internet of things. Information security.*

Referências

ALVES, R. C. V. **Metadados como elementos do processo de catalogação**. 2010. Tese (Doutorado em Ciência da Informação) -Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2010.

ARANHA, Christian; PASSOS, Emanuel. A tecnologia de mineração de textos. Disponível em: <www.spell.org.br/documentos/download/26518> Acesso em: 25 fev. 2016

BEATO, Cláudio et al. **Crime e estratégias de policiamento em espaços urbanos.** Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0011-52582008000300005> Acesso em: 16 fev. 2016.

BEEBE, Nicole; CLARK, J. **A hierarchical objectives-based framework for the digital investigations process, to appear in Digital Investigation**, 2005. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.2406&rep=rep1&type=pdf> Acesso em: 16 fev. 2016

BEEBE, Nicole; CLARK, J. **Dealing with Terabyte Data Sets in Digital Investigations.** IFIP International Conference on Digital Forensics. ISSN 1571-5736. vol. 194, pp. 3-16. Disponível em: < http://link.springer.com/chapter/10.1007%2F0-387-31163-7_1> Acesso em: 16 fev. 2016

BENTES PINTO, Virginia. et al. **"Netnografia": uma abordagem para estudos de usuários no ciberespaço.** 9º Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, 2007, Açores-Portugal. Anais do 9º Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas. Lisboa: APBAD, 2007.p. 79-95. Disponível em: http://www.repositorio.ufc.br/bitstream/riufc/11579/1/2007_eve_fmpbezerra.pdf Acesso em: 16 fev. 2016.

BRAGA, Ryon. **O Excesso de Informação – A Neurose do Século XXI.** Disponível em:< http://www.mettodo.com.br/pdf/O%20Excesso%20de%20Informacao.pdf> Acesso em: 16 fev. 2016.

BORKO, H. **Information Science: What is it?** American Documentation, v.19, n.1, p.3-5, Jan. 1968. (Tradução Livre)

CÁRDENAS, Alvaro; MANADHATA; Pratyusa K., RAJAN, Sreeranga P. **Big Data Analytics for Security.** Disponível em: <http://www.utdallas.edu/~alvaro.cardenas/papers/IEEESnP.pdf> Acesso em: 13. Dez 2015.

CARRIER, Brian. **Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layer.** International Journal of Digital Evidence. Vol. 1. Issue 4. Winter 2003. Disponível em: < http://digital4nzics.com/Student%20Library/Defining%20Digital%20Forensic%20Examination%20and%20Analysis%20Tools%20Using%20Abstraction%20Layers.pdf> Acesso em: 16 fev. 2016

CAVALCANTE, Waldek F. **Crimes Cibernéticos: noções básicas de investigação e ameaças na internet.** Disponível em: <http://jus.com.br/artigos/25743/crimes-ciberneticos> . Acesso em: 22 fev. 2016.

DINAPOLI, T. **Red Flags for Fraud.** Disponível em:< http://www.osc.state.ny.us/localgov/pubs/red_flags_fraud.pdf> Acesso em: 16 fev. 2016.

DJERABA, C. O.; ZAIANE, R.; SIMOFF, S. **Mining Multimedia and Complex Data.** New York: Springer, 2003.

DOMINIQUE, B; TOM, K. **Internet Best Current Practice. Guidelines for Evidence Collection and Archiving.RFC 3327.** Disponível em: <http://www.ietf.org/rfc/rfc3227.txt>. Acesso em: 24 fev. 2016.

Extract, Transform, and Load Big Data with Apache Hadoop. Intel. 2010 Disponível em: < https://software.intel.com/sites/default/files/article/402274/etl-big-data-with-hadoop.pdf> Acesso em: 10 fev. 2016

FARMER, D.; VENEMA, W. **Perícia Forense Computacional: Como investigar e esclarecer ocorrências no mundo cibernético.** São Paulo: Pearson/Prentice Hall, 2006.

GIARDELLI, Gil. **Os dados nunca mentem. Bem-vindos à era do big data.** Disponível em:< http://exame.abril.com.br/rede-de-blogs/pessoas-do-seculo-21/2013/05/03/os-dados-nunca-mentem-bem-vindos-a-era-do-big-data/> Acesso em: 16 fev. 2016.

GRUBER, T. R. **Toward principles for the design of ontologies used for knowledge sharing**. Knowledge Systems Laboratory, Stanford University, 1993. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.6200 >. Acesso em: 12 dez. 2015.

HOSSEINKHANI, J.; CHUPRAT, S.; TAHERDOOST, H.; SHAHRAKI MOGHADDAM, Amin. **Propose a Framework for Criminal Mining by Web Structure and Content Mining.** International Journal of Advanced Computer Science and Information Technology (IJACSIT), Helvetic Editions. Vol 1. n. 1. Disponível em: < http://elvedit.com/journals/IJACSIT/wp-content/uploads/2012/12/Propose-a-Framework-for-Criminal-Mining-by-Web-Structure-and-Content-Mining.pdf> Acesso em: 26 fev. 2016

JOHNSON, Steven**. Cultura da Interface: como o computador transforma nossa maneira de criar e comunicar.** Rio de Janeiro. Ed. Jorge Zahar, 2001. Disponível em:< http://www.niett.unirio.br/public/upload/ff79006eb77a377f5a7500dc5e672255.pdf> Acesso em: 16 fev. 2016

LE COADIC, Y. F. **A ciência da informação**. Brasília: Briquet de Lemos/Livros, 1996.

McLINDEN, Sean. **Positive predictive value and digital forensics.** Disponível em: <http://forensicfocus.blogspot.com.br/2010/05/positive-predictive-value-and-digital.html> Acesso em: 16 fev. 2016.

MORANDINI, Marcelo. **Critérios e Requisitos para Avaliação da Usabilidade de Interfaces em Groupware–CSCW.** Disponível em: <http://www.dca.fee.unicamp.br/courses/IA368F/1s1998/Monografias/morandini.html>. Acesso em: 16 fev. 2016.

OHL, Rodolfo. **Big Data: Como analisar informações com qualidade. Disponível em: <**http://corporate.canaltech.com.br/coluna/big-data/Big-Data-como-analisar-informacoes-com-qualidade/> Acesso em: 16 fev. 2016

PAIVA, Rodrigo Oliveira de. **Um olhar para a arquitetura da informação no ciberespaço.** DataGramaZero – Revista de Informação. vol. 15, n. 5, out. 2014. Disponível em:< http://www.dgz.org.br/out14/Art_05.htm> Acesso em: 16 fev. 2016.

PERNER, Petra. **Data Mining on Multimedia Data.** Germany, Springer, 2002.

PICHILIANI, Mauro. **Data Mining na Prática: Algoritimo K-Means.** Disponível em: < http://www.devmedia.com.br/data-mining-na-pratica-algoritmo-k-means/4584> Acesso em: 15 fev. 2016.

ROSA, João Luiz. **Brasil perde R$ 2,3 bi com fraudes em transações financeiras em 2013.** Disponível em: <http://www.valor.com.br/financas/3502148/brasil-perde-r-23-bi-com-fraudes-em-transacoes-financeiras-em-2013> Acesso em: 26 fev. 2016

ROSENFELD, Louis; MORVILLE, Peter. **Information Architecture for the World Wide Web.** Beijing, O'Reilly, 1998.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Tecnologias da web semântica aplicadas a organização do conhecimento: padrão SKOS para construção e uso de vocabulários controlados descentralizados. In: José Augusto Chaves Guimarães; Vera Dodebei. (Org.). **Organização do Conhecimento e Diversidade Cultural**. 1ed.Marília: Fundepe, v. 3, p. 224-233, 2015.

VASHISHT, Sheveta; KAUR, Manveer; RICHA, Sapra; MANDEEP, Sigh. Detecting Cyber Crime by Analyzing Users Data. Int. J. Computer Technology & Applications, vol. 3 (3), 1029-1033. Disponível em: < http://www.ijcta.com/documents/volumes/vol3issue3/ijcta2012030327.pdf> Acesso em: 16 fev. 2016.

WURMAN, Richard Saul. **Information Architects. The Design of Information to Improve, Clarify and Facilitate the Process of Communication.** Disponível em: <http://buch.archinform.net/isbn/3-85709-458-3.htm>. Acesso em: 16 fev. 2015

WURMAN, Richard Saul. **Ansiedade de Informação 2.** São Paulo: Editora de Cultura, 2005