Análises de linhas temporais contextuais em investigações digitais

Regis Levino de Oliveira[1], Bruno Werneck P. Hoelz[2*]

[1]Departamento de Engenharia Elétrica - Universidade de Brasília
[2]Instituto Nacional de Criminalística – Polícia Federal
Brasília, DF – Brasil

regislevino@yahoo.com.br, werneck.bwph@dpf.gov.br

Abstract. *Timeline analysis is a well-known technique in digital forensics. However, most studies and tools focus on the extraction of timestamps with less emphasis on how to visualize and analyze large volumes of such data. This work proposes a process to generate contextual timelines, where each timestamp is also associated with other four dimensions: place, person, subject and event. A clustering algorithm is then used to generate timelines that contain similar data, and are easier to visualize and interpret.*

Resumo. *A análise de linhas temporais é uma técnica bastante empregada em exames periciais em ambientes computacionais. No entanto, a maioria desses estudos concentra-se nos desafios da extração de registros temporais com menos ênfase em como visualizar e analisar um grande volume desses dados. Este trabalho propõe um processo para gerar linhas temporais contextuais, onde cada rótulo temporal é associado a outras quatro dimensões: local, pessoa, assunto e evento. Um algoritmo de clusterização é então utilizado para gerar linhas temporais com dados similares, que são mais fáceis de visualizar e interpretar.*

1. Introdução

Vestígios digitais podem ser a chave para a elucidação de um fato criminoso, como a existência de uma foto (ex.: comprovante de depósito), a troca de mensagens por e-mail ou aplicativos de mensageria ou o registro de ações realizadas em um sistema computacional (como *logs* de acesso). Para auxiliar na compreensão e análise dessa grande diversidade de vestígios, vários recursos de visualização são empregados, dentre eles o da visualização de linhas temporais. A partir da construção de uma linha do tempo, fica mais fácil analisar eventos em torno de pontos de interesse, como o momento em que um servidor de dados foi acessado indevidamente.

Muitas ferramentas estão disponíveis para a geração e análise de linhas temporais. Ferramentas como Zeitline (Buchholz e Falk, 2005), CyberForensics TimeLab (Olsson e Boldt, 2009) e log2timeline (Guðjónsson, 2010) lidam com o desafio de coletar e apresentar dados temporais de várias fontes. No entanto, uma das maiores dificuldades para a análise das linhas temporais é a grande quantidade de dados, o que requer outras abordagens. Nesse sentido, Chabot et al. (2014) propõem um modelo para a reconstrução de eventos que inclui definições formais das entidades envolvidas em um incidente e

* Os autores agradecem o apoio da Secretaria Nacional de Segurança Pública (SENASP), da Diretoria Técnico-Científica da Polícia Federal e da FINEP (Convênio 01.12.0433.01, Projeto: Defesa Nacional e Segurança Pública) na realização deste trabalho.

operadores que permitem que o conhecimento contido no modelo seja extraído, manipulado e analisado. Já Yu Jin (Jin, 2013) propõe o uso de mapas auto-organizáveis para analisar a relação de registros de atividades no sistema operacional Android.

De acordo com as circunstâncias do caso, pode ser necessário visualizar os eventos temporais segundo o contexto a que estão relacionados, reduzindo o ruído associado ao grande volume de registros temporais. Para isso, este trabalho propõe uma abordagem de contextualização e agrupamento de eventos temporais. Os registros temporais passam por um processo de contextualização, no qual um registro temporal (e seu arquivo de origem) é analisado e associado a entidades previamente definidas nas dimensões pessoa, local, evento e assunto. Posteriormente, algoritmos de agrupamento (clusterização) são aplicados sobre os registros, resultando em linhas temporais contextualizadas segundo essas entidades. Tal abordagem facilita a visualização, identificação e análise de eventos relacionados, que poderiam passar desapercebidos em meio ao grande volume de dados provenientes de diversas fontes, contribuindo, portanto, para uma melhor análise das linhas temporais.

O restante desse artigo está organizado da seguinte forma: a Seção 2 descreve a abordagem proposta e trabalhos correlatos; a Seção 3 apresenta a aplicação da abordagem e os resultados obtidos e, por fim, a Seção 4 apresenta as conclusões e trabalhos futuros.

2. Linhas temporais contextuais

Linhas temporais contextuais visam apresentar os dados temporais agrupados segundo outras dimensões que não somente o tempo. A construção de linhas temporais contextuais está dividida em quatro etapas, conforme ilustra a Figura 1. A primeira etapa é a determinação das fontes de evidência das quais serão extraídos os registros temporais. Como discutido anteriormente, registros temporais podem ser provenientes de várias fontes distintas como computadores, *smartphones*, dispositivos de rede, entre outros.

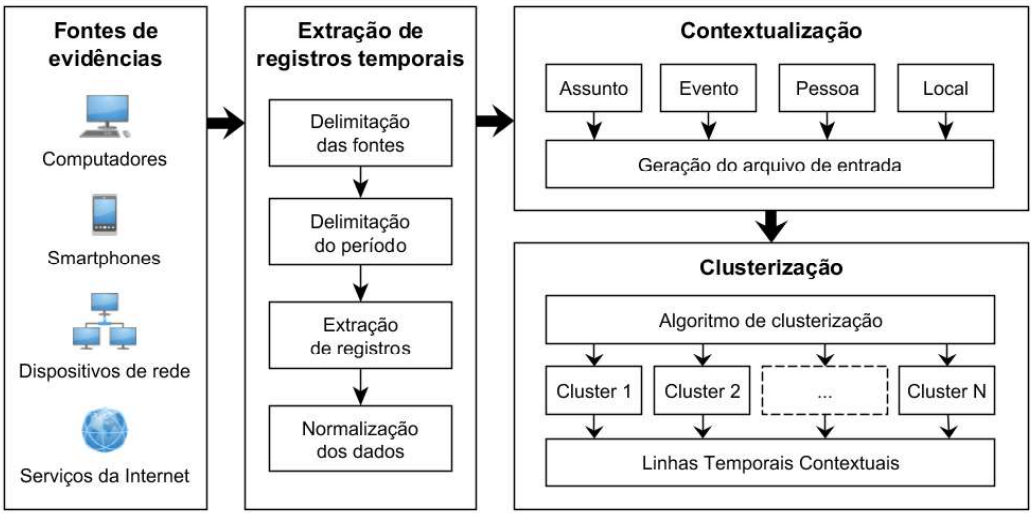

Figura 1 – Arquitetura da proposta

Em seguida, é realizada a extração dos registros temporais. Antes da extração propriamente dita, pode-se delimitar as fontes e o período de interesse, dependendo do incidente investigado. Muitos registros temporais podem não ser interessantes para determinado incidente como, por exemplo, datas de criação, acesso e modificação de

arquivos do sistema operacional. Nessa etapa, assume-se a utilização de ferramentas próprias para extrair os dados, como log2timeline (Guðjónsson, 2010). Por fim, os dados devem ser normalizados, com a uniformização dos formatos de dados gerados, em especial da forma de representação dos rótulos temporais, desvios nos relógios e zonas de tempo diferentes. Após a extração dos registros temporais de interesse, é iniciada a etapa de contextualização. Essa etapa visa vincular um registro temporal a um ou mais contextos. Cada contexto é composto por quatro dimensões: local, evento, pessoa e assunto.

A dimensão do local associa o registro temporal a um lugar no espaço. Por exemplo, o local de ocorrência de um determinado crime ou locais frequentados pelos suspeitos. Para vincular um registro temporal a um determinado local, pode-se utilizar metadados de arquivos de imagens de câmera fotográficas e telefones celulares, coordenadas de GPS extraídas de aplicativos, histórico de navegação de sites de mapas, entre outros. A dimensão evento diz respeito a eventos relevantes para a investigação, ocorridos digitalmente (ex.: troca de uma mensagem) ou no mundo real (ex.: homicídio). Um evento pode ou não ter um momento bem definido no tempo. Nesse caso, os registros temporais podem ser associados ao evento pela proximidade (ex.: logs obtidos após a detecção de uma invasão ao sistema). Registros temporais (eventos de baixo nível) também podem ser vinculados automaticamente a um evento (de alto nível), como sugerido por Hargreaves e Patterson (2012). Por exemplo, a detecção de alteração em vários arquivos do sistema, pode ser vinculado à instalação de um programa malicioso.

Em uma investigação digital, várias pessoas podem estar envolvidas. Logo, registros temporais provenientes de fontes diversas estarão associados a pessoas distintas. Um registro temporal pode ser associado a uma pessoa não só por ser o usuário principal de um determinado dispositivo, mas também pela interação por meio de registros de chamadas telefônicas, contas de e-mail e outros identificadores associados às pessoas de interesse. A dimensão assunto diz respeito a temas ou palavras-chave relacionadas à investigação. Para isso, devem ser definidos os assuntos segundo o crime investigado, onde cada assunto é composto por um conjunto de palavras-chave. Os registros temporais são então associados a um assunto, se o conteúdo do arquivo possuir uma ou mais dessas palavras. A informação sobre os termos que compõe o assunto pode vir de informações da equipe de investigação, por meio de declarações de vítimas, criminosos em busca do abrandamento da pena (delação premiada) ou por técnicas de processamento de linguagem natural para identificar o assunto a partir do conteúdo dos textos extraídos das fontes de evidências.

Após a definição dos contextos, é realizada a geração do arquivo de entrada que será submetido ao algoritmo de clusterização. Na aprendizagem de máquina, um cluster é o agrupamento de dados de certo conjunto de dados de entrada. Esses agrupamentos são realizados por meio de similaridades (dentro de um mesmo cluster) ou pelas diferenças (entidades de um cluster são diferentes de entidades de outro cluster). É desejável que os pontos em um cluster tenham uma distância pequena um do outro e que pontos em clusters diferentes sejam mais distantes. Como resultado do algoritmo, cada registro é associado a um cluster. Por fim, para cada cluster gerado, ordenam-se os registros por data e hora, obtendo-se assim uma linha temporal contextualizada. Logo, um número N de clusters gera N linhas temporais. De posse das várias linhas temporais, o perito é capaz de analisar atividades correlacionada pelos contextos definidos pelas dimensões de nome, evento, local ou assunto.

3. Aplicação experimental

O objetivo desta seção é exemplificar a aplicação da proposta em um cenário baseado em um caso real de posse e o compartilhamento de material pornográfico contendo crianças ou adolescentes. Nesse caso, o suspeito da conduta descrita, também dava aulas sempre às segundas-feiras em uma escola, não sendo descartada a hipótese de abuso sexual de alunos. Após denúncia e monitoramento do suspeito, foram descobertas outras pessoas associadas às condutas. Nesse cenário, os peritos foram acionados para a busca de evidências digitais no material apreendido na residência dos suspeitos e na escola: computadores, celulares e câmeras digitais. Para representar essas fontes de evidências, foi utilizado um conjunto de arquivos fictícios, simulando 11.500 registros temporais entre as datas de 01/10/2014 à 25/11/2014. Os formatos de data foram normalizados para um formato único (apresentado na Figura 2).

3.1. Contextualização

Para a definição do local, foram utilizadas informações colhidas de metadados EXIF de imagens de câmera fotográfica e telefone celular, vinculação manual por meio de reconhecimento visual (ex.: fotos da casa do suspeito ou da fachada da escola) e informações de localização coletadas por aplicativos. Na ausência de um local mais específico, o registro foi associado ao próprio dispositivo originário (ex.: celular) ou ao local no qual a fonte de evidências se localizava ou ser indeterminada (ex.: uma foto de abuso sem identificação clara do local).

Para a dimensão pessoa, os nomes foram reunidos baseado na extração de informações das trocas de mensagens de aplicativos de chat, nomes repassados pela investigação que seriam suspeitos dos crimes, nomes contidos nos metadados de arquivos (ex.: propriedade autor de um documento de texto).

Na definição do evento, primeiramente, foram definidos eventos relevantes no crime investigado. Por exemplo, a troca de mensagens entre suspeito e vítima. No caso em questão, foi observada uma grande quantidade de fotografias do abuso, downloads de arquivos contendo pornografia infantil e troca de mensagens entre os suspeitos. Por fim, na definição do assunto, foram utilizadas palavras-chave encontradas nos itens de pesquisa para *download* de arquivos de pornografia infantil.

3.2. Clusterização

Para a etapa de clusterização, os dados contextualizados foram inseridos na ferramenta WEKA (Bouckaert et al., 2016). Uma amostra do arquivo de entrada, no formato ARFF, é apresentado na Figura 2. Os atributos são definidos no início do arquivo. Entre eles, a fonte que originou o registro, o rótulo temporal (data) e os dimensões que definem o contexto (assunto, local, evento e nome). Os dados das dimensões são categóricos e os valores válidos são apresentados entre chaves na Figura 2. Os dados foram então submetidos ao algoritmo *Simple K-Means* utilizando a distância euclidiana para formação dos clusters. Registros não contextualizados (cujo valor faltante é representado por um ponto de interrogação) são ignorados. O parâmetro *k,* do número de clusters a serem gerados, depende da quantidade de valores em cada atributo (nome, assunto, local, evento) e da variância desses dados, o que depende do caso em concreto. Logo, o valor de *k* foi definido experimentalmente, variando seu valor até obter um conjunto de clusters satisfatório com k=5.


```
@RELATION caso_pi

@ATTRIBUTE fonte_dos_dados    string
@ATTRIBUTE data     DATE "yyyy/MM/dd HH:mm:ss"
@ATTRIBUTE assunto  {PTHC, PEDO, SEXO}
@ATTRIBUTE local    {escola, notebook1, notebook2, celular, escritorio}
@ATTRIBUTE evento   {Tirar_foto,chat_whatsapp,chat_messenger_facebook,download_arquivos}
@ATTRIBUTE nome     {Rodrigo, Carlos, Lucas, Frederico, Sade}

@DATA
Fonte_10,"2014/10/05 22:27:26",SEXO,celular,chat_messenger_facebook,Rodrigo
Fonte_3,"2014/10/10 09:19:55",?,notebook2,chat_messenger_facebook,Rodrigo
Fonte_4,"2014/10/03 10:13:47",PEDO,celular,chat_messenger_facebook,Carlos
Fonte_8,"2014/10/17 18:39:33",?,notebook1,download_arquivos,Carlos
Fonte_7,"2014/10/15 12:39:02",?,notebook2,?,?
Fonte_6,"2014/10/21 20:37:53",PEDO,notebook2,download_arquivos,Rodrigo
Fonte_10,"2014/10/23 10:31:54",PTHC,escritorio,download_arquivos,Carlos
Fonte_5,"2014/10/11 17:15:31",PTHC,escritorio,download_arquivos,Rodrigo
```

Figura 2 – Arquivo do tipo ARFF utilizado no experimento

Os clusters obtidos são apresentados na Figura 3, com a quantidade de registros em cada cluster e o contexto associado. O cluster 0, por exemplo, tem 5765 registros e o seu contexto é definido como [assunto=sexo, local=escola, evento=Tirar_foto, nome=Rodrigo].

Cluster#				
0	1	2	3	4
(5765.0)	(2163.0)	(1297.0)	(1405.0)	(869.0)
SEXO	PTHC	PEDO	PTHC	SEXO
escola	notebook1	escritorio	notebook2	celular
Tirar_foto	chat_whatsapp	chat_whatsapp	chat_messenger_facebook	chat_whatsapp
Rodrigo	Carlos	Sade	Carlos	Carlos

Figura 3 – Clusters formados após a execução do algoritmo SimpleKMeans

A Figura 4 apresenta as linhas temporais contextuais construídas a partir do clusters gerados. O eixo X apresenta a data em milissegundos (com a data correspondente inserida na figura) e cada linha do eixo Y representa um dos clusters. Cada ponto representa um registro temporal. É possível verificar alguns pontos de maior atividade como nos dias 13/10, 20/10, 27/10, 25/11. No cluster 0, esses pontos foram destacados com um círculo. Observa-se também um tempo de inatividade entre 28/10 e 24/11.

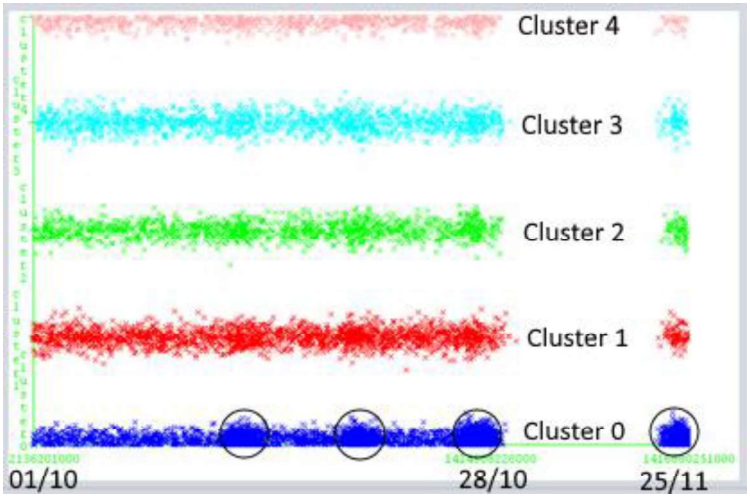

Figura 4 – Linha temporal contextual

Com a análise das linhas temporais contextuais, verificou-se grande atividade entre os interlocutores Rodrigo e Carlos, em especial associada aos eventos

"chat_whatsapp" e "chat_messenger_facebook". O suspeito Rodrigo também teve maior atividade nos dias citados anteriormente. Assim, a linha temporal proveniente do primeiro cluster (Cluster 0) facilita a visualização de evidências associadas ao abuso da vítima na escola exatamente nos dias mencionados, associados às fotos tiradas. Observou-se que as demais linhas permitem uma visualização melhor de mensagens utilizando WhatsApp e Facebook Messenger, usadas no compartilhamento de material pornográfico infantil.

4. Conclusão

A análise de linhas temporais é uma técnica muito empregada em investigações digitais. A grande quantidade de fontes de dados temporais torna essa análise mais complexa, dificultando a adequada visualização das relações entre eventos e da determinação de pontos no tempo que são de interesse da investigação. O presente trabalho demonstrou que essa análise pode ser simplificada pela contextualização e posterior agrupamento (clusterização) dos registros temporais. Como resultado do modelo proposto, obtém-se linhas temporais cujos registros apresentam maior similaridade contextual entre si, reduzindo a interferência de outros registros não relacionados. No experimento proposto, pode-se identificar com mais facilidade os suspeitos com maior interação e os momentos de maior atividade relacionados às condutas investigadas.

Em trabalhos futuros, pretende-se aprimorar o processo de agrupamento com a aplicação de outros algoritmos, a fim de comparar os resultados obtidos para cada um deles. Além disso, pretende-se explorar formas de contextualização baseadas em processamento de linguagem natural, especialmente na dimensão do assunto.

5. Referências

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2016). *WEKA Manual for Version 3-8-0*.

Buchholz, Florian. Falk, Courtney. (2005). *Design and Implementation of Zeitline: a Forensic Timeline Editor*. In: Digital forensic research workshop. https://users.cs.jmu.edu/buchhofp/publications/zeitline_dfrws.ps. Acessado em 10/10/2015.

Charbot, Yoan. Bertaux, Aurélie. Nicolle, Christophe. Kechadi, M-Tahar. (2014). *A complete formalized knowledge representation model for advanced digital forensics timeline analysis.* Digital Investigation 11 (2014) 95-105.

Gudhjonsson K. (2010). *Mastering the super timeline with log2timeline*. SANS Reading Room. http://www.sans.org/reading-room/whitepapers/logging/mastering-super-timeline-log2timeline-33438. Acessado em 20/10/2015.

Hargreaves, Christopher. Patterson, Jonathan. (2012). *An automated timeline reconstruction approach for digital forensic investigations*. Digital Investigation (2012) vol. 9, p. 69–79.

Jin, Yu. (2013). *Timeline analysis for Android-based systems*. Kongens Lyngby 2013 IMM-M.Sc.-2013-42.

Olsson J, Boldt M. (2009). *Computer forensic timeline visualization tool*. Digital Investigation (2009). vol. 6, p.78-87.

