Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

An ontology-based approach for the reconstruction and analysis of digital incidents timelines

Yoan Chabot [a, b, *], Aurélie Bertaux [a], Christophe Nicolle [a], Tahar Kechadi [b]

[a] CheckSem Team, Laboratoire Le2i, UMR CNRS 6306, Faculté des Sciences Mirande, Université de Bourgogne, BP47870, 21078 Dijon, France
[b] School of Computer Science & Informatics, University College Dublin, Belfield, Dublin 4, Ireland

ARTICLE INFO

Article history:
Received 15 March 2015
Received in revised form 13 July 2015
Accepted 14 July 2015
Available online xxxx

Keywords:
Digital forensics
Event reconstruction
Forensic ontology
Knowledge extraction
Ontology population
Timeline analysis

ABSTRACT

Due to the democratisation of new technologies, computer forensics investigators have to deal with volumes of data which are becoming increasingly large and heterogeneous. Indeed, in a single machine, hundred of events occur per minute, produced and logged by the operating system and various software. Therefore, the identification of evidence, and more generally, the reconstruction of past events is a tedious and time-consuming task for the investigators. Our work aims at reconstructing and analysing automatically the events related to a digital incident, while respecting legal requirements. To tackle those three main problems (volume, heterogeneity and legal requirements), we identify seven necessary criteria that an efficient reconstruction tool must meet to address these challenges. This paper introduces an approach based on a three-layered ontology, called *ORD2I*, to represent any digital events. *ORD2I* is associated with a set of operators to analyse the resulting timeline and to ensure the reproducibility of the investigation.

Introduction

Nowadays, digital investigations require the analysis of a large amount of heterogeneous data. The study of these volumes of data is a tedious task which leads to cognitive overload due to the very large amount of information to be processed. To help the investigators, many tools have been developed. Most of them extract unstructured data from various sources without bridging the gap of semantic heterogeneity and without addressing the problem of cognitive overload. To resolve these issues, a promising perspective consists of using a precise and reliable representation allowing to structure data on the one hand, and to standardise their representation on the other hand. A structured and formal knowledge representation has two goals: 1) to build automated processes more easily by making information understandable by machines and 2) to give to investigators an easy way to query, analyse and visualise information. Computer forensics investigations also have to fulfil a set of legal and juridical rules to ensure the admissibility of results in a court. It is particularly necessary to ensure that all evidence presented at a trial are credible and that the methods used to produce evidence are reproducible and did not alter the objects found in the crime scene. Problems of traceability and reproducibility of reasoning are widely discussed in the literature and provenance is particularly relevant to be applied to digital investigations. Indeed, as defined by (Gil and Miles, 2013), the provenance of a resource is a record describing the entities and the processes involved in the creation, dispersion or others activities that affect that resource. The provenance provides a fundamental basis for assessing the authenticity and the truth value of a resource and its reproducibility.

[*] Corresponding author. CheckSem Team, Laboratoire Le2i, UMR CNRS 6306, Faculté des Sciences Mirande, Université de Bourgogne, BP47870, 21078 Dijon, France.
E-mail address: yoan.chabot@hotmail.fr (Y. Chabot).

http://dx.doi.org/10.1016/j.diin.2015.07.005
1742-2876/© 2015 Elsevier Ltd. All rights reserved.

In our work, we propose an innovative digital forensic approach based on a knowledge model allowing to represent accurately a digital incident and all the steps used during an investigation to produce each result. In addition, a full set of operators to manipulate the content of this ontology is proposed. We introduce extraction and instantiation operators to build automatically the knowledge base using digital traces extracted from disk images. Then, automatic analysis operators taking advantage of this ontology are also proposed. In particular, we focus on an operator used to identify potential correlations between events.

This paper is structured as follow: Section 2 gives a comprehensive state of the art of event reconstruction approaches for digital forensics. Section 3 introduces the *SADFC* approach. In particular, the structure of the three-layered ontology is detailed and its various operators are described. Section 4 evaluates the performance of our approach and illustrates its capabilities on a case study.

State of the art

Event reconstruction is a complex process because of three main issues: the large volume of data, the heterogeneity of information due to the use of a large number of sources and the legal requirements that the results have to meet. This section aims to review the existing solutions described in the literature to reconstruct and analyse past events. The quality and the relevance of nine reconstruction approaches are reviewed based on the previous three issues and their limitations. These answers are then synthesised and used to guide the development of our approach to ensure that the three main issues of event reconstruction are taken into account.

Data volume

Currently, investigators are facing huge data volumes on a digital crime scene. The growth of the data size is caused by several factors. Digital devices are more and more present in our daily lives. This increases the number of devices owned by each person and therefore the number of devices found on the crime scenes. In addition, the frequency of use and the increase of storage capacity of the digital devices have caused the increase in the quantity of data stored by each device. The very large amount of data makes the analysis very complex and tedious, even causing cognitive overload. Information that is potentially relevant to reach the objectives of an investigation is diluted in the amount of data, making the investigation difficult. For example, the Plaso toolbox (Gudhjonsson, 2010), which produces timeline from hard disk image, can identify thousands of events from a wide range of sources (Apache logs, Skype conversations, Google Chrome history, Windows event logs, etc.) from an image of only a few gigabytes. In conclusion, an event reconstruction approach should be able to efficiently process information and retrieve and visualise the results in a clear and intuitive way.

In the literature, a large number of approaches provide tools to automatically extract the information and populate a central storage constituting the timeline. ECF (Chen et al., 2003) is an approach aiming to extract, store in a database and manipulate events from heterogeneous sources. It introduces a set of extractors to collect events and store them in a database, which quickly generates a temporal ordered sequence of events. These automatic extractors, a widely used concept, can also generate the timeline as in FORE (Schatz et al., 2004), FACE (Case et al., 2008), CyberForensic TimeLab (Olsson and Boldt, 2009), Plaso and PyDFT (Hargreaves and Patterson, 2012). However, in some approaches including (Gladyshev and Patel, 2004) and (James et al., 2010), the lack of automation seems difficult to address and they present very high complexity (combinatorial explosion).

The automated extraction of events from a large number of sources leads to the creation of large timeline that is difficult to read, interpret and analyse. To assist the investigators during this phase, event reconstruction approaches should provide analysis tools to carry out all or part of the reasoning and visualisation of the data in a clear and intuitive way. The use of a textual representation or a database is not suited to build high-performance analysis process. In approaches, such as ECF, the reasoning capabilities are limited by the data structure used for the storage of information.

ECF stores data in a database consisting of a table of common information about events and tables of information specific to each type of event. One of the objectives of ECF is to provide a canonical form for representing events uniformly regardless of their sources. Adopting a generic information level and a specialised information level is also used in CybOX (Barnum, 2011). This conceptual separation allows to introduce a canonical representation of events that facilitates information processing (analysis tasks for example) while preserving the specificities of each event. However, the use of a database does not allow to take full advantage of this feature. Unlike ontology, databases do not enable to explicitly represent semantic of data which constrains the understanding of data by the analysis algorithms. The FORE approach (Schatz et al., 2004) proposes a correlation tool based on rules to identify causal relationships (the event A causes the event B if the event A should happen before the event B) between events. Despite the relevance of this tool, the use of a rule-based system requires the definitions of the rules. The construction of such a set of rules is a tedious task and it cannot take into account all cases. Event reconstruction approaches must implement algorithms that can adapt to any kind of situations even those unknown by the investigators. It is, therefore, necessary to develop analysis tools not relying on the rules defined by the user. In (Gladyshev and Patel, 2004), the reconstruction process can be seen as a process of finding the sequence of transitions that satisfy the constraints imposed by the evidence. The authors try to perform the event reconstruction by representing the behaviour of the system as a finite state machine. Scenarii that do not match evidence collected are then removed. After reducing the number of potential scenarii, a backtracking algorithm is used to extract all possible scenarii. However, the lack of automation does not allow to handle complex cases. Indeed, the investigation of a single computer may involve several processes such as web browsers,

Please cite this article in press as: Chabot Y, et al., An ontology-based approach for the reconstruction and analysis of digital incidents timelines, Digital Investigation (2015), http://dx.doi.org/10.1016/j.diin.2015.07.005