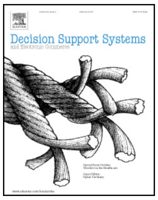
Decision Support Systems 61 (2014) 69–78

Contents lists available at ScienceDirect

Decision Support Systems

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



A case-based reasoning method for locating evidence during digital forensic device triage

Graeme Horsman [1], Christopher Laing [1], Paul Vickers *

Northumbria University, Faculty of Engineering and Environment, Department of Computer Science and Digital Technologies, Pandon Building, Camden Street, Newcastle-upon-Tyne NE2 1XE, UK

ARTICLE INFO

Article history:
Received 1 March 2013
Received in revised form 13 December 2013
Accepted 17 January 2014
Available online 25 January 2014

Keywords:
Digital forensics
Triage
Case based reasoning
Bayesian reasoning
Knowledge reuse

ABSTRACT

The role of triage in digital forensics is disputed, with some practitioners questioning its reliability for identifying evidential data. Although successfully implemented in the field of medicine, triage has not established itself to the same degree in digital forensics. This article presents a novel approach to triage for digital forensics. Case-Based Reasoning Forensic Triager (CBR-FT) is a method for collecting and reusing past digital forensic investigation information in order to highlight likely evidential areas on a suspect operating system, thereby helping an investigator to decide where to search for evidence. The CBR-FT framework is discussed and the results of twenty test triage examinations are presented. CBR-FT has been shown to be a more effective method of triage when compared to a practitioner using a leading commercial application.

1. Introduction

Triage is a technique used in many disciplines, most notably in the field of medicine as a way of prioritising injured or ill patients for treatment [25]. It can be viewed as a way of organising a workload to allow for the efficient allocation of available resources [10]. More recently it has found its way into the Cybersecurity lexicon where it is used to categorise threats [35] allowing an organisation to determine during incident response which events should be dealt with first based on their severity and available resources [11]. When applied to digital forensics (DF), the meaning of triage differs depending on the context in which it is applied but, as Casey [15] suggests, its goal is to speed up an investigation by attempting to identify evidential exhibits and files quicker.

Triage can mean the prioritisation of physical exhibits for investigation (for which we coin the term *high-level triage*) but it can also signify an interrogation of data held on a target digital device (which we call *device triage*, or DT). In addition to known facts about a case, decisions made during high-level triage are also commonly based on a suspected offence type or the physical location of an exhibit at the scene of a crime.

DT involves the identification of evidence on a suspect system from amongst non-case relevant data, whilst allocating as few resources as possible [32]. DT is employed to speed up a DF investigation, attempting to cut down the time it takes to identify evidence and is the focus of this article. Cybercrime and the use of technology to commit crime are on the increase [17]. Bem et al. [9] suggest that this is leading to increased caseloads which, in turn, are causing difficulties in the field of DF.

High tech crime units are experiencing investigation backlogs [42] and DF practitioners are facing increasing pressure to effectively manage their workloads and process their investigations more efficiently. This has led to DF software developers championing their DT tools as a way of increasing investigation efficiency [1,2,19]. Pollitt [33] argues that these tools have fallen short of the requirements needed to deal with current DF cases. DF practitioners have yet to consistently use DT for investigating digital media.

The United Kingdom's Association of Chief Police Officers (ACPO) good practice guide for computer-based electronic evidence [3] provides guiding principles for DF investigations. However, ACPO [4] have been cautious to recommend DT owing to the perception that it carries an increased risk of missing evidential files. We argue that DT has the potential to reduce mounting case backlogs but to do this, DT techniques must be improved.

Apprehension over the use of DT may be due, in part, to a limitation of many current DT applications, namely the use of pre-coded and fixed scripts. Until a vendor releases a software update, these scripts can remain unchanged for months. This means a DF practitioner is limited to using the same evidence gathering script in their DT investigations, even though the way in which a particular offence is committed may have changed. In such a scenario the chance of missed evidence is increased, and reluctance to conduct DT is understandable. Such scripts are frequently derived from an estimate of which data types would be likely to reside on a system for a given offence. Consequently, this approach opens up the DT process to criticism and, arguably, to an increased risk of investigation errors.

Abbreviations: CBR-FT, Case-Based Reasoning Forensic Triager; DF, Digital Forensics; DT, Device Triage; ERR, Evidence Relevance Rating; PRF, Primary Relevance Figure.

* Corresponding author. Tel.: +44 191 243 7614.
E-mail addresses: graeme.horsman@northumbria.ac.uk (G. Horsman), christopher.laing@northumbria.ac.uk (C. Laing), paul.vickers@northumbria.ac.uk (P. Vickers).

[1] Tel.: +44 191 227 4426.

0167-9236/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2014.01.007

DT can be applied at both the scene of a crime (pre-seizure) and within the confines of the forensic laboratory (post-seizure) [10], each with different purposes and consequences. Pre-seizure DT aims to eliminate devices from an investigation and carries the greatest risk of missed evidence as devices of low priority may be omitted from an investigation due to time and resource constraints. Any evidence missed during pre-seizure DT could have serious ramifications, as evidence may be left with a guilty suspect. Even though post-seizure DT occurs in a secure laboratory environment and all the potential evidence is available (providing pre-seizure DT has not occurred first), it is possible that an inadequate DT could still prevent an exhibit from proceeding to a full examination, and thus, failure to find the evidence. Current DT applications have been criticised for lacking the investigative experience needed to extract relevant data from a system [5].

In order to improve DT a greater emphasis must be placed on achieving higher precision in the identification of relevant evidence. One possible way to achieve this involves the reuse of knowledge from past DF investigations in order to establish where evidence is commonly found. Patterns of suspect activity can help to isolate key areas of a system to be targeted during DT. Reusing investigation data in this way would allow DF practitioners to extract data from a system based on the probability of evidence being present in particular locations. This could transform the current approach of guessing where evidence may be located into an informed decision based on where evidence has been regularly found in the past.

This paper presents our Case-Based Reasoning Forensic Triager (CBR-FT) which uses patterns of behaviour to detect evidential activity in DT. Through the use of past DF case results we demonstrate the tool's ability to target evidential files. We offer CBR-FT as a method of implementing knowledge reuse for the benefit of DT. It also goes some way to answering Pollitt's call for triage to be treated "as a formal process that can be measured for efficiency and efficacy" [33].

Current approaches to triage are discussed in Section 2. Sections 3–7 introduce the new CBR-FT framework and its functionality. The results of twenty triage examinations are presented in relation to the offence of fraud (Section 8). The precision and recall of CBR-FT during DT are discussed and compared to EnCase Portable [21], a commercial DF DT application.

2. Approaches to triage

Research often focuses on the development of basic frameworks that highlight general approaches to high-level triage [38]. There is little published research in the area of DT and even when techniques are presented their functionality often remains insufficiently tested as they are rarely used in actual DF DT investigations [28]. Commercial applications for DT do exist (*e.g.*, [1,2,19]) but share the same fundamental weaknesses. Using predefined scripts, an investigator executes the application and data is automatically collected for review. These scripts are often coded to search for and retrieve specific evidential artefact types. This can lead to very large quantities of data being recovered which the DF practitioner must then interpret.

In addition, commercial approaches to DT frequently use hash set analysis and keyword searching to identify evidential files without the need for a practitioner to view file content. Keyword searching is a process of looking for relevant alphanumeric strings which may be contained within evidential files on a system [7]. Hash analysis involves the comparison of two file hash values in order to establish a match [39]. A file hash value is generated by an algorithm which produces a character string which is unique to a given binary file. The MD5, SHA-1, and SHA-256 hash functions are all commonly used in DF analysis. Except in very exceptional circumstances [41], two files with matching hash values will contain exactly the same binary data.

Using a hash set (a collection of hash values from known evidential files) to perform a hash analysis of a target system has the potential to lead to the identification of evidential files. This process is often used in the identification of indecent child images [16] through the use of the Child Exploitation and Online Protection (CEOP) hash sets. Hashing can be used not only to identify relevant files but also to filter out known non-evidential files (*e.g.*, standard operating system files) which could reduce the overall time needed to carry out an examination. However, as discussed below, both hashing and keyword searching approaches can limit the effectiveness of DT because they are too restrictive, leading to a failure to identify digital evidence.

2.1. Limitations of hash analysis for DT

During hash analysis, should any aspect of a target file be altered, (*e.g.*, altering one pixel in a picture) the file's hash value would change even though the target file is essentially the same, thereby rendering hash analysis ineffective. In addition, hash sets must contain hash values of files known to be evidential. What constitutes evidential value in one case may not in another, especially in crimes such as fraud. This is unlike offences involving indecent images, for example, where a single file can be evidential regardless of the system on which it is found. For example, a hash value from one particular corporate financial file is unlikely to be of value when dealing with an investigation from another company. In reality, such files would not exist in both scenarios as these files maintain company specific data and make hashing ineffective. Creating a hash set of files from company A would therefore be unlikely to highlight files found in company B.

A limitation of hashing is that it is defeated when a copy of a file is altered slightly (*e.g.*, by cropping a photograph). Kornblum's [27] piecewise hashing approach uses a "context triggered rolling hash" to highlight known files which have been modified or amended slightly but relies on *a priori* knowledge of the modified files. Perceptual hashing is one way to combat these limitations as it makes judgements about the human perceptual similarity of files rather than by comparing their binary representations. Perceptual hashing offers some flexibility as the user can identify files which maintain a certain level of similarity as opposed to an exact match [26]. However, perceptual hashing maintains processing overheads which would increase the overall length of the DT process, thereby negatively affecting the efficiency of the investigation.

Therefore, hashing techniques are only helpful in limited DT scenarios. Hashing (both normal and perceptual) works well for certain file types, particularly picture files (as they often remain unedited by a user) and therefore lends itself to particular offence types concerning these types of files. However, there are many other crime types (*e.g.*, fraud) which do not involve the types of file for which hashing is well suited and require, instead, semantic analysis of file content.

2.2. Limitations of keyword searching for DT

Keyword searching also raises issues for DT. First, keyword searching can take a considerable amount of time which works against the goal of triage which is to prioritise cases as quickly as possible. Second, defining keyword dictionaries can prove problematic because, in many DT investigations, the surrounding circumstances of the case may not yet be fully known, making key terms difficult to identify. Third, the key terms identified are subjective, being based upon the experience of the investigator, leading to varying degrees of success. It must be noted, however, that techniques to automatically generate key term dictionaries do exist [36] including ontological structures which create and maintain domain related keyword knowledge bases [20]. Although such techniques have the potential to be manipulated and applied in a DT investigation, the difficulty remains in automatically generating a key term database on a subject (the case under investigation) about which little is known at the time.

Finally, compound or compressed files may possess internal structures which cannot be easily identified through a simple binary keyword search. An example includes the latest .docx files used in Microsoft