Data Mining based Crime-Dependent Triage in Digital Forensics Analysis

Rosamaria Bertè¹,a, Fabio Marturana¹,b, Gianluigi Me¹,c, Simone Tacconi²,d

¹Department of Computer Science, Systems and Production
University of Tor Vergata, Rome, Italy
²Servizio Polizia Postale e delle Comunicazioni, Rome, Italy
*a rosamariaberte@libero.it, b marturana@libero.it, c me@disp.uniroma2.it, d simone.tacconi@interno.it.

Keywords: Computer Forensics, Data Mining, “Post-mortem” Triage.

Abstract. Over the last few years, law enforcement registered a growing number of crimes related to the worldwide diffusion of high storage capacity low-cost digital devices. As a consequence Computer Forensics, the investigative discipline that aims to find evidence among seized devices is becoming increasingly complex. In this paper, we propose a new approach to digital investigations, based on the application of Data Mining and Knowledge Management theory, which aims to give a theoretical foundation to “post-mortem” Triage. This new practice has the potential utility to speed-up investigations by assigning a priority to each seized device, with a positive impact on the upcoming forensic analysis. The paper shows how the proposed methodology could create intelligence from the extracted data and predict the model’s dependent variable (i.e. the class) and its relation with the independent variables (i.e. system configuration files, installed software, file statistics, browser history and system event log). We identify the class variable with the likelihood that a computer has been used to commit specific crimes such as child pornography, copyright violation, hacking, murder and terrorism. The paper is based on a case study carried out in collaboration with the Servizio Polizia Postale e delle Comunicazioni (the Italian Cybercrime Police Unit).

1. Introduction

Computer Forensics is the application of specific investigative techniques aiming to analyze seized computers and gather evidence for presentation in a court of law. The goal of Computer Forensics is therefore to perform a structured investigation to track computer system and user activity, maintaining a documented chain of evidence.

In 1955, in an article in The Economist, Cyril Northcote Parkinson first wrote that “work expands so as to fill the time available for its completion” (Parkinson’s Law). The article was referring to public administration but today’s corollary to this law might be that “computer forensic examinations expand in proportion to the increase in size of forensic units thus maintaining a significant backlog” [1].

Considering how the large availability of low-cost, sophisticated and heterogeneous digital devices with large storage capacity has contributed to the spread of computer crimes [2], we can
certainly argue that Parkinson’s Law corollary is particularly true dealing with Digital Forensics. A serious problem complained by law enforcement specialists, indeed, is that, considering the effort to analyze hundreds of Terabytes of data, usually only few records of relevant intelligence about the crime under investigation could be extracted. The reasons are basically twofold: on one side, the enormous amount of available data to be processed in the lab and, on the other, the forensic investigator’s habit to look for potential evidence by means of traditional, manually intensive and time-consuming procedures. As a consequence, finding theories and techniques with the aim to narrow the area of interest and deepen the search only where needed is considered a crucial aspect to reverse this negative trend.

In this paper, we propose an application of Data Mining theory to “post mortem” Computer Forensics Triage, inheriting some concepts from a previous research in the field of Mobile Forensics whose effectiveness has already been tested and verified in the field [3,4].

The proposed approach aims to build a priority list among the seized computers, before being processed in the lab, highlighting their relative relevance according to a crime-dependent three-dimensional model of categorization concerning timeline, crime’s features and suspect’s private sphere (habits, skills and interests). This new point of view could represent a complementary activity in digital investigations aiming to identify immediately the most relevant computers.

“Post mortem” Computer Forensics Triage consists of the following four phases: forensic acquisition, feature extraction and normalization, context and priority definition, data classification and triaging. The first one is the classical forensic hard disk image creation. The second is devoted to acquire the whole set of available features (system configuration files, installed software, file statistics, browser history and system event log) from the disk image and normalize them creating a two-dimensional matrix, called complete matrix. The third is in charge of introducing in the model the timeline of interest and the crime-specific features in order to focus the attention only on a part of the aforementioned matrix, called reduced matrix. The fourth is assigned the task of analyzing the reduced matrix’s features and calculating the class variable by means of Data Mining algorithms.

2. Related Work

Over the past few years, Computer Forensics has been supported by several theories and methods developed in order to find evidence quickly on seized computers as far as specific crimes such as murder, child abductions, missing persons, death threats etc. are concerned. In such cases the need for the timely identification, analysis and interpretation of digital evidence is crucial since it could be the difference between life and death for the victim.

As far as Mobile Forensic is concerned, our research group has recently proposed two possible applications of Data Mining based classifications to “post mortem” Triage.

The first one [3] concerned a methodological approach to Mobile Forensics in order to identify the most interesting mobiles from an investigative point of view by predicting the device owner’s usage profile. The proposed methodology can help investigators split up relevant and less important aspects of the case under investigation by assigning a priority to every involved device, person and crime. By means of a quick memory search on the whole set of seized devices it is possible to create a list of phones, ordered by probative value, which require additional processing at Forensic Lab.

The second study [4], extensively discussed with Italian law enforcement cybercrime specialists, was based on Mobile Forensics Triaging and self-knowledge algorithms for mobiles classification and concerned a viable methodology to determine the likelihood that a mobile phone has been used to commit child pornography.

In 2006 a research group proposed a new methodology called Cyber Forensic Field Triage Process Model (CFFTPM) [5], which deals with “live” on-site or field activities for providing the identification, analysis and interpretation of digital evidence in a short time frame, without the
requirement of taking the system back to the lab for an in-depth examination or acquiring a complete forensic image. The proposed methodology, although entailing a real risk of exhibit pollution, is justified by the need to provide investigative leads quickly in time critical situations.

Recently a new trend combining computer forensic principles and Data Mining statistical approach & knowledge discovery process is taking hold in the research community. According to recently published papers, the proposed model allows to extract, store and analyze digital devices data with forensically sound methods that could be used in a court of law [6].

Finally an important research was conducted about the application of Data Mining theory to digital text analysis, based on clustering text mining techniques for investigational purposes. The work addressed text clustering for forensics analysis based on a dynamic, adaptive clustering model to arrange unstructured documents into content-based homogeneous groups [7,8].

3. Proposed process model

This paragraph describes the proposed four-phases model allegedly applied on a set of data extracted from forensic images of seized hard disks and regarding crimes such as child pornography, copyright violation, hacking, murder and terrorism.

The whole process is carried out in a forensic lab in parallel with the traditional forensic procedures (acquisition, retrieval and analysis) [9,10,11] and is summarized in the following figure:

![Post-mortem triaging model](image)

**Fig.1 – “Post-mortem” triaging model**

The *first stage* of the process is, therefore, the *forensic acquisition*. According to NIST [12], indeed, the first step in a Computer Forensics investigation is to make a disk image in order to preserve digital evidence integrity and guarantee the analysis repeatability.

The *second stage* of our workflow, called *feature extraction and normalization*, is in charge of extracting relevant data from disk images.

This stage of the process is inspired by official best practices concerning computer incident response which strongly suggests to:

- Look at the last date of change on critical files;
- Examine configuration and start-up files;
- Look for hacking tools (password crackers, copies of passwords, etc.);
- Examine the password file for unauthorized accounts;
- Search for keywords appropriate to the incident;
- Search for hidden areas, slack space, and cache;
- Look for changes to files, critical file deletions, and unknown new files;
- Collect a list of all e-mail addresses, FTP sites and URLs, visited from the computer.

We concentrate on the following computer usage parameters: *system configuration files*, *installed software*, *file statistics*, *browser history* and *system event log* which represent the independent variable of the model. Occurrences of each parameter, for instance the number of deleted images or installed hacking tools, the presence of a system log file or the number of visited URL etc. is counted and summarized within a two-dimensional matrix called *complete matrix*.

During drive image analysis, we suppose to collect the whole set of *features* concerning user’s
habits, technical skills and interests (*suspect’s private sphere*).

We propose to associate user's habits with the following set of parameters:
- percentage of modified files, ordered by time slot (morning, afternoon, evening, night);
- internet connections, ordered by time slot;
- monthly login frequency;
- system utilization ordered by time slot.

With regards to user's technical skills, we consider:
- system configuration files;
- system log settings;

With regards to user’s interests we focus our attention on:
- stored and deleted files statistics (audio, video, images, documents, executable);
- installed applications statistics;
- visited URLs statistics.

The *third stage* of our workflow is called *context and priority definition* since it introduces in the model the *timeline* of interest (i.e. the time frame during which the crime happened and thus we suppose to find more evidence) and the *crime-specific features* (i.e. the crime related fingerprint such as the presence of child images in a child pornography case or illegally downloaded software or movies in a copyright violation). After the aforementioned three stages we will be able to create the following data structures (*complete* and *reduced matrix*) which summarizes the set of parameters that could be processed by the model:

![Fig.2 - complete and reduced matrix](image)

The feature nickname is indicated in the leftmost column while the other columns show a set of instance samples of retrieved hard disk images.

The *fourth stage* of our workflow is called *data classification and triaging* and it is in charge of elaborating the reduced matrix in order to provide the final classification of the input data. According to our experience, this phase could be based on a collection of machine learning algorithms for data mining tasks such as Waikato Environment for Knowledge Analysis [13].

4. Methodology algorithmic foundations

We assume that our model’s dependent variable, called *class*, is the likelihood that a computer has been used to commit the crimes of *child pornography, copyright violation, hacking, murder* and *terrorism*. The goal is therefore to assign a *class* to each inspected exhibit, a sort of relative score used to create an ordered list of items. To meet the model requirements, during the *context and priority definition* phase, it is necessary to gather the whole set of features extracted during the previous phase, creating a context-dependent features subset, correlated with the crime under observation/evaluation and the specific timeframe i.e. the time the crime occurred. In other words, the goal is to determine the presence of crime’s digital fingerprints in the exhibit. Talking about child pornography, for example, there is a high likelihood to find private pictures and videos on the sized computer, while in case of copyright violation the presence of a large number of hard disk
stored music files and films is a relevance indicator from an investigative point of view. In other cases such as hacking crime we could probably find installed firewalls and system or hacking applications while, in case of terrorism, we will search the computer for documents concerning the criminal conduct. Finally, considering murder specific digital fingerprints, we will probably find specific log events or chat/instant messaging stored sessions, social networks visited URLs or received/sent emails etc. The outcome of context and priority definition phase is therefore the contextualization of the complete matrix which will be deprived of unnecessary items and provided as input to the next classification phase (i.e. reduced matrix).

The Data Mining supervised classifier requires first a trainingset, i.e. a collection of crime-dependent representative patterns with a known class, in order to train the underlying algorithm. Once trained the classifier will elaborate the collection of real patterns.

The first step of data classification and triaging is, therefore, the collection of a consistent set of representative exhibits with a known class concerning the crimes that we want to classify (i.e. trainingset creation). This is a crucial activity since, if performed by inexperienced analysts, it could negatively influence the whole learning process. With a well-formed trainingset at our disposal, it now is possible to train a classifier by adopting, for instance, the iterative and predictive method called 10 folds cross-validation [13] which splits the trainingset into ten approximately equal partitions, each in turn used for testing and the remainder for training. The procedure is repeated ten times so that, in the end, every instance has been used exactly once for testing.

It is also possible to use 10 folds cross-validation output to evaluate classifiers’ effectiveness (i.e. the ratio between performance and acquired knowledge) calculating performance indicators such as Precision, Recall and F-Measure [13].

5. Conclusions

This paper deals with a new methodological approach to Computer Forensics, the branch of Digital Forensics concerning evidence extracted from digital devices according to well defined and standardized methods and with probative value in court.

The research questions that we addressed were the following: “which are the potential benefits of applying Data Mining and Machine Learning algorithms to the actual computer forensic techniques? Can this methods eliminate bottlenecks and increase investigation efficiency?”.

To answer the questions we interviewed experts and law enforcement, in order to assess their investigation procedures and we realized that the actual forensics lab data analysis process, based on a rigid 4-steps workflow (HD Image creation, Extraction, Analysis and Reporting) implies that, in each case, terabytes of seized data must be searched to isolate a single evidence.

In this paper we propose a new model which redefines the aforementioned forensics workflow, introducing the concept of “post mortem” Triage. Based on Machine Learning algorithms and data analysis to identify crime-dependent activity patterns not discernible through a manual review process, Triage aims indeed at giving a priority to each inspected computer based upon the likelihood that a computer has been used to commit one of the following crimes: child pornography, copyright violation, hacking, murder and terrorism. We identify this likelihood with the model’s dependent variable under observation and we call it class.

The class is calculated upon a set of independent variables based on system configuration files, installed software, file statistics, browser history and system event log.

In particular, the proposed crime-dependent categorization model, concerning timeline, crime’s features and suspect’s private sphere (habits, skills and interests) consists of the following four phases: forensic acquisition, feature extraction and normalization, context and priority definition, data classification and triaging. The latter two phases are the most important since, during the third one we are able to introduce in the model the timeline of interest and the crime-specific features creating the, so called, reduced matrix while in the fourth we analyze the matrix and calculate the class variable by means of classifiers, assigning a relative score to each analyzed exhibit.
6. Future Work

Our research relates to the theoretical foundation of Computer Forensics “post mortem” Triage and draw the guidelines for future implementations.

The proposed methodology could be implemented with specific digital investigations support tools allowing the automated disk categorization and the creation of a user activities’ timeline, based on the WEKA java library downloadable at [14]. It is important to highlight that classification task’s success depends on the number of “real” observations and sized data that will be used to build the trainingset which is the core of the knowledge discovery process. The higher is the number of analyzed hard disk images, indeed, the better the knowledge discovery process works.

References