A Cyber Crime Evidence Collection and MultiAgent Digital Investigation Toolkit

Vaishnavi Ganesh
Assistant Professor
Computer Science And Engineering
Priyadarshini Indira Gandhi College of Engineering,Nagpur,India

Abstract : With the increased use of Internet and information technol-ogy all over the world, there is an increased amount of criminal activities that involve computing and digital data. These digital crimes (e-crimes) impose new challenges on prevention, detection, investigation, and pros-ecution of the corresponding offences. Computer forensics (also known as cyberforensics) is an emerging research area that applies computer in-vestigation and analysis techniques to help detection of these crimes and gathering of digital evidence suitable for presentation in courts. This new area combines the knowledge of information technology, forensics science, and law and gives rise to a number of interesting and challeng-ing problems related to computer security and cryptography that are yet to be solved.To be able to examine large amounts of data in a timely manner in search of important evidence during crime in-vestigations is essential to the success of computer forensic examinations. The limitations in time and resources, both computational and human, have a negative impact in the results obtained. Thus, better use of the resources avail-able are necessary, beyond the capabilities of the currently used forensic tools. Herein, we describe the use of Artificial Intelligence in computer forensics through the development of a multiagent system and case-based reasoning. This sys-tem is composed of specialized intelligent agents that act based on the experts knowledge of the technical domain. Their goal is to analyze and correlate the data contained in the evidences of an investigation and based on its ex-pertise, present the most interesting evidence to the human examiner, thus reducing the amount of data to be person-ally analyzed. The correlation feature helps to find links between evidences that can be easily overlooked by a hu-man expert, specially due to the amount of data involved.

IndexTerms **- computer forensics, artificial intelligence, multiagent sys-tems, digital investigation,digital crimes, forensics technology.**

1.INTRODUCTION

With the increased use of Internet and information technol-ogy all over the world, there is an increased amount of criminal activities that involve computing and digital data. These digital crimes (e-crimes) impose new challenges on prevention, detection, investigation, and pros-ecution of the corresponding offences. Computer forensics (also known as cyberforensics) is an emerging research area that applies computer in-vestigation and analysis techniques to help detection of these crimes and gathering of digital evidence suitable for presentation in courts. This new area combines the knowledge of information technology, forensics science, and law and gives rise to a number of interesting and challeng-ing problems related to computer security and cryptography that are yet to be solved.

Among other issues in collecting evidence from computers, one fundamental difference between paper documents and digital data is that electronic data can be easily copied and modified. A suspect may easily argue that the evidence found in his/her computer was implanted or modified by the law enforcement agency after the computer has been seized by the agency. It is very important to verify the file system integrity of the suspect's computer after it has been seized by the law enforement agency.

Another problem is that there are many different file formats, operating sys-tems and file system structures. Electronic documents can be generated by vari-ous kinds of application programs such as word processors, spreadsheet software, database software, graphic editors, electronic mail systems. The documents can be stored as user files in user directories, or as fake system files in the system directories, or hidden files. Sometimes, evidence can also be found in the deleted files. When a file is deleted, the operation system usually only removes the ref-erences to the file in the file allocation table (FAT). The actual content of the file is still physically stored on the disk until that area has been overwritten by another file. It is a time consuming task to inspect every possible storage area of the whole computer for potentially useful evidence. And it is also not possible to check every file using all available application programs manually. In this project,a cyber crime evidence collection tool will be designed which tries to handle the above problems. Besides the problem of evidence collection, e-crime detection is also very important. Intrusion detection (e.g. detection of distributed denial of service attack) is one of the well-known examples.

To be able to examine large amounts of data in a timely manner in search of important evidence during crime in-vestigations is essential to the success of computer forensic examinations. The limitations in time and resources, both computational and human, have a negative impact in the results obtained. Thus, better use of the resources available are necessary, beyond the capabilities of the currently used forensic tools. Herein, the use of Artficial Intelligence in computer forensics will be described through the development of a multiagent system and case-based reasoning. This system will be composed of specialized intelligent agents that will act based on the experts knowledge of the technical domain. Their goal is to analyze and correlate the data contained in the evidences of an investigation and based on its expertise, present the most interesting evidence to the human examiner, thus reducing the amount of data to be personally analyzed. The correlation feature helps to find links between evidences that can be easily overlooked by a human expert, specially due to the amount of data involved.

2. LITERATURE SURVEY

A number of different approaches have been proposed to deal with the three aspects presented before: (i) reduction in the amount of evidences to be examined, (ii) correlation of evidences and (iii) distribution of forensic examinations computational work.

I. In the various works of [18, 19, 20] we find the idea of the digital evidence bags (DEB). A DEB is a universal con-tainer for digital evidence from any source that allows the provenance to be recorded and continuity to be maintained

II. throughout the life of the investigation. The author suggests the use of intelligent techniques to treat a selective informa-tion capture scenario, using a selective image approach with the DEB. There are important aspects of his work which considers the importance of files during the investigation task,like (i) how to capture and combine the experts knowledge of both domains technical and legal, and (ii) how to be certain that all the relevant information and evidences of other related crimes are captured in the DEB. The answer to this questions is very important in or-der to collect and examine only relevant evidences to the investigation, thus reducing the amount of time required to complete the examinations.

III.

IV. The work of [16] proposes the addition of a case-relevance indicator to the evidences. By their definition, case-relevance would be the property of any piece of information, which is used to measure its ability to answer the investigative `who, what, where, when, why and how' questions in a criminal in-vestigation." The levels of relevance defined by their work go from \Absolutely Irrelevant" to \Probably Case-Relevant". The intelligent agents used in our tool do something simi-lar, although they can sometimes diverge in the relevance given to a file. Such conflict is solved by another agent and ultimately reviewed by the human expert. By considering the case-relevance of an evidence, the expert can focus, for instance, on a subset of files found on a hard drive, tem-porarily ignoring those that are considered irrelevant.

V.

VI. [7] presents two approaches for analyzing large data sets of forensic data called Forensic Feature Extraction (FFE) and Cross-Drive Analysis (CDA). We consider CDA to be the most interesting, since it uses statistical techniques for corre-lating information within a single disk image and across mul-tiple disk images. A recent work by [4] describes a tool called Forensics Automated Correlation Engine (FACE), whose ob-jectives are similar to ours. They also present some scenarios where an increased level of correlation of disparate evidences was achieved.

VII.

VIII. The work of [15] makes the case for Distributed Digi-tal Forensics and present some scenarios where the forensic work can't be performed anymore on a single workstation. They also propose a distributed framework and some perfor-mance results that show the advantages of the distributed approach, which would also enable more sophisticated anal-ysis techniques. The tool presented in [4], unlike ours, does not employ the distribution of processes. In our case, our proposal benefit from the distributed nature of a multia-gent system and we already observed the performance gains, which helped us obtain better computational resource usage and reduce the time required to perform the examination.

IX.

Finally, we also make use of the framework proposed by [1]. They propose a multi-tier, hierarchical framework to guide digital investigations. To us, the most important as-pect of this framework is its objectives-based phases and sub-phases that are applicable to various layers of abstrac-tion, and to which additional layers of detail can be added as needed. These objectives-based phases serve to us as a guide for the definition of the specialized intelligent agents employed in our tool. We also observed the case-based na-ture of the objectives, which means that different types of in-vestigation require different sets of examinations. This gives us the opportunity to apply case-based reasoning (CBR) to the planning of our agents' actions.

3.HYPOTHESIS

The cyber crime evidence collection tool is the DESK (The Digital Evidence Search Kit) which is a general purpose computer foren-sics system focusing on integrity control of the digital data.

3.1 The Framework of DESK

The DESK system will consist of a DESK machine which will be typically a notebook computer with a serial port and a floppy diskette used to start up the suspect's machine (subject machine). The DESK machine will be connected to the subject machine using

a serial (RS-232) cable. There are two software components of DESK: the DESK client that will be installed on the DESK machine; and the DESK server that will be contained on the floppy diskette to be executed by the subject machine. The DESK client will be mainly used to provide a user interface for issuing commands to inspect the subject machine.

The DESK server component, installed on the floppy diskette, will have additional functionalities which will include the followings.

1.To start up the subject machine: Usually the file (e.g. system files) in the subject machine will be modified if it is booted up by its own operating system.

2.To lock the subject machine: This is to protect the subject machine from any accidental corruption by the interrupts of the machine. This step is very important as it can ensure that the contents found on the subject machine cannot be modified, thus ensures the integrity integrity of the subject ma-chine while various forensic operations are being performed.

3.To provide a simple user interface for simple search operations: The user interface is much less sophisticated than that of the DESK client running on the notebook due to the storage limitations of floppy diskettes.

There will be two main operations of DESK: keyword search and file system integrity checker.

Keyword Search: A pre-defined text pattern file which contains important key-words that can be specific to a particular case, in Chinese and/or English, to be searched for on the subject machine, is used for three different types of search, namely physical search, logical search and deleted file search. Physical search per-forms a search of the patterns in each physical sector of the subject machine's storage system. E-crime evidence stored purposely in unused sectors can be dis-covered. Logical seach, on the other hand, makes use of the information about the file system, so patterns stored across different sectors can be located. Deleted file search will try to locate the deleted file provided it is not yet overwritten by another file and perform the pattern search on these files.

File System Integrity Checker: There are two functions in this checker. Firstly, it is to ensure the integrity of the file system of the subject machine. We compute a hash value of the whole file system (e.g. a hard disk) of the subject machine. By recording this hard disk hash value properly, the law enforcement agency can easily prove that the content of the hard disk has not been modified after the machine has been captured by the agency. Also, in order to reduce the possibility of causing accidental damage to the hard disk, usually exact copies of disks (also called clone images) are made for the subsequent analysis. The hash values of the clone images and the original hard disk can be compared to show that the clone images are exactly the same as the original hard disk.

Secondly, the suspect may store some crime evidence in standard files of com-mon software applications (e.g. freecell.exe). A hash value database that contains fingerprints (hash values) of all files in a standard software distribution are used to compare with the hash values of the corresponding files in the subject ma-chine.

Frequently, at real computer forensic examinations, ex-perts can't determine beforehand which evidences will turn out to be the most relevant to the investigation of a crime. Consider the example of a cybercaf or any other scenario where several computers appear to share the same IP ad-dress. The traces will often lead to the cybercaf and not to any particular machine. The same difficulty can be faced while collecting evidences of fraud in companies with sev-eral machines and users. In these examples a pre-analysis of the suspect machines would limit the number of machines to collect, reducing the time required to complete the forensic examinations. The problem is the lack of intelligent tools to help forensic experts with the pre-analysis phase, which results in the collection of a large number of machines to be examined, some of which will not contribute to the overall result of the investigation and will just increase the time needed to complete the examinations.

To find the need for intelligent tools and to better em-ploy computational resources during forensic examinations is the main focus of this research.

3.2 MultiAgent Digital Investigation toolKit (MADIK)

In the proposed work, the use of the MultiAgent Digital Investigation toolKit (MADIK), a multiagent system will be proposed to assist the computer forensics expert on its examinations. The system will be composed of a set of ISAs that perform different analysis on the digital evidence related to a case on a distributed manner.

In MADIK, each ISA contains a set of rules and a knowl-edge base, both based on the experience of the expert on a certain kind of investigation. Since the examination of digital evidence in crime investigations share similarities, MADIK uses CBR to determine which agents are better employed in which kind of investigation. This also allows the agents to reason about the evidences in a way

that is more adequate to the special case in question. As an exam-ple, we can cite the use of hash sets in a child exploitation case. The ISA will use first the hash sets related to child exploitation, thus giving the examiner a quicker feedback on the presence of such files in a piece of evidence. At the moment, the MADIK will have six specialized intelligent agents implemented:

i.HashSetAgent calculates the MD5 hash from a file and compares it with its knowledge base, which contains sets of files known to be ignorable (e.g. system files) or important (e.g. contraband files like child exploita-tion imagery). We might cite that some of these hash sets contain more than 10 million hash values, from dif-ferent softwares, as cited in [12]. Also we might note that the bigger the hash set the longer the comparison takes.

ii.FilePathAgent keeps on it's knowledge base a collec-tion of folders which are commonly used by several ap-plication which may be of interest to the investigation like P2P (peer-to-peer), VoIP and instant messaging applications.

Iii.FileSignatureAgent examines the file headers (the first 8 bytes of the file), to determine if they match the file extension. If someone changes the file exten-sion in order to hide the true purpose of the file, this will be detected by this agent. It also keeps a list of commonly used prefixes and file names, such as the ones used by digital cameras.

iv.TimelineAgent examines dates of creation, access and modification to determine events like system and soft-ware installation, backups, web browser usage and other activities, some which can be relevant to the investi-gation.

v.WindowsRegistryAgent examines files related to the windows registry and extracts valuable information such as system installation date, time zone configuration, removable media information and others.

vi.KeywordAgent searches for keywords and uses regular expressions to extract information from files such as credit card numbers, URLs or e-mail addresses.

The MADIK is not a replacement for commonly used forensic tools like AccessData Forensic ToolKit or Guidance EnCase. The proposed agents are a reduced set that al-lows for many rules to be conceived and many examinations to be carried over, as a proof of concept. New agents can be conceived by encapsulating the functionality of existing tools and scripts such as Foundstone's Galleta or Volatile Systems' Volatility. This works main contributions, are the definition of an automated architecture where specialized agents can analyze and correlate findings beyond the simple acquisition and extraction of data provided by the current tools, with the added benefit of seeking the better use of computational resources through distribution. The case-based approach also provides a way to improve the agents results over time, by learning from previous cases, as suggested in [3].

If an unexperienced examiner tries to compare every hash set he has available against every single file, the process will take too long and the results will not be much better than those obtained by an experienced examiner who chooses the most likely hash sets so he can have quick but yet effective results.

To better understand how the system works we will ex-plain MADIK's operation processes. The strategic manager receives different cases to perform the forensic analysis. Ac-cording to the organization's priorities, the strategic man-ager defines the order of execution and amount of resources (number of computers) for each case. A tactical manager is them assigned to one specific case which can contain several evidences, like a number of hard drives. The tactical man-ager defines the priority of its evidences and distributes them to the available operational managers, which are limited by the resources available to that case. The operational man-ager will employ the necessary specialized agents to perform the different tasks it deems important to examine a piece of evidence.

During the specialists' execution, they will insert their conclusions and remarks in the blackboard. Each entry con-tains the agent's recommendation, an user-friendly descrip-tion and the time taken to examine the file, for benchmark-ing purposes. There are three distinct levels of recommenda-tion:

i. ignore - the strongest recommendation to ignore a file, indicating its unimportant according to the agent,

ii. alert - strongly recommends the selection of a file, and

iii.inform - this recommendation is an intermediate value, which contains information to help the human reviewer to decide whether to select that file or not. There can be an additional sign (+ or -) representing an ignore or alert bias, respectively.

4 OBJECTIVES OF THE WORK

1.Computer forensics research is an important area in ap-plying security and computer knowledge to build a better society.

2.The main goal is to present a tool to help experts during spe-cialized forensic examinations in order to obtain significantly better results when compared to those obtained by the cur-rently used tools considering three aspects: (i) reduction of routine and repetitive

analysis while also reducing the amount of evidence that must be personally reviewed by the expert, (ii) correlation of evidences, (iii) distribution of processes. With this, human and computational resources can be applied more efficiently.

3.To find the need for intelligent tools and to better em-ploy computational resources during forensic examinations is the main focus of this research.

4. There are two design objectives of the tool DESK:-.
i.One of the objectives is to ensure the validity and reli-ability of the digital evidence. Once it has been proved that the tool has been used properly and in compliance with the Evidence Ordinance [10], the digital evidence found in the suspect's computer can be presented and used in courts for prosecution.
ii.Another objective is to provide an efficient and automatic search function to search for digital contents that can be used as evidence for the e-crime.

5.CONCLUSION

This paper described an application of AI in computer forensics and the latest results obtained with the use of the MADIK, a MAS to assist the experts during computer foren-sic examinations. The system has been tested using real data with four specialized agents. Its results show that the application of a MAS in the forensic examination of com-puters is an interesting approach to improve the usage of computational resources available and reduce the time re-quired to conduct the examination through the reduction in the volume of evidence to be examined. The intelligent and autonomous agents that compose the system seek to incor-porate the experts knowledge to perform the analysis of a great volume of data, giving the expert a subset that is more likely an important evidence to the investigation.

The combination of the reduction in the volume of evi-dences to be examined by the expert and the reduction in the execution times obtained with the distributed process-ing of the evidences already show the potential of the tool and the productivity gains it can offer to computer foren-sic experts and to investigators that face an ever increasing volume of digital evidence.

REFERENCES

[1] Nicole Beebe and Jan Guynes Clark. A hierarchical, objectives-based framework for the digital investigations process. Digital Investigation, 2(2):147{167, 2005.

[2] Fabio Luigi Bellifemine, Giovanni Caire, and Dominic Greenwood. Developing Multi-Agent Systems with JADE. Wiley Series in Agent Technology, Sussex, England, 2007. ISBN 978-0-470-05747-6.

[3] D. Bruschi and M. Monga. How to reuse knowledge about forensic investigations, 2004.

[4]Andrew Case, Andrew Cristina, Lodovico Marziale, Golden G. Richard, and Vassil Roussev. Face: Automated digital evidence discovery and correlation. Digital Investigation, 5(Supplement 1):S65{S75,September 2008.

[5]Daniel D Corkill. Collaborating Software: Blackboard and Multi-Agent Systems & the Future. In Proceedings of the International Lisp Conference, New York, USA, October 2003.

[6]Mark d'Inverno and Michael Luck. Understanding Agent Systems. Springer Series in Agent Technology, Berlin, Germany, 2nd revised and extended edition, 2004. ISBN 3-540-40700-6.

[7] Simson L. Gar¯nkel. Forensic feature extraction and cross-drive analysis. Digital Investigation, 3S:S71{S81, 2006.

[8] Michael N. Huhns and Munindar P. Singh, editors. Readings in Agents. Morgan Kaufmann, San Francico, USA, 1998. ISBN 1-55860-495-2.

[9]Telecom Italia Lab (TILAB). Java Agent DEvelopment framework - JADE. Online. http://jade.tilab.com.

[10] V. Jagannathan, R. Dodhiawala, and L.S. Baum, editors. Blackboard Architectures and Applications.Academic Press, Orlando, FL, USA, 1989.

[11]George F. Luger. Arti¯cial Intelligence: Structures and Strategies for Complex Problem Solving. Addison-Wesley, USA, 4th edition, 2002. ISBN 0-201-64866-0.

[12] Steve Mead. Unique ¯le identi¯cation in the national software reference library. Digital Investigation,3(3):138{150, 2006.

[13]H. Penny Nii. Blackboard systems, part one: The blackboard model of problem solving and the evolution of blackboard architectures. AI Magazine, 7(2):38{53, 1986.

[14]S. Pinson and P. MoraÄ³tis. An intelligent distributed system for strategic decision making. Group Decision and Negotiation, 6:77{108, 1996.

[15]Vassil Roussev and Golden G. Richard III. Breaking the performance wall: The case for distributed digital forensics. In Digital Forensic Research Workshop - DFRWS, 2004.

[16]Gong Ruibin and Mathias Gaertner. Case-relevance information investigation: Binding computer intelligence to the current computer forensic framework. International Journal of Digital Evidence, 4(1), 2005.

[17]Stuart J. Russell and Peter Norvig. Arti¯cial Intelligence: A Modern Approach. Prentice-Hall, USA, 2nd edition, 2002. ISBN 0-13-790395-2.

[18]Philip Turner. Uni¯cation of digital evidence from disparate sources (digital evidence bags). In Digital Forensic Research Workshop - DFRWS, 2005.

[19]Philip Turner. Selective and intelligent imaging using digital evidence bags. Digital Investigation, 3(Supplement-1):59{64, 2006.

[20]Philip Turner. Applying a forensic approach to incident response, network investigation and system administration using digital evidence bags. Digital Investigation, 4(1):30{35, 2007.