지능형 디지털 포렌식 도구 및 데이터 간소화 프레임워크에 관한 연구

류정현, 이재동, 석상기, 박종혁*
서울과학기술대학교 컴퓨터공학과
e-mail : {jh.ryu, jdlee731, sksuk, jhpark1}@seoultech.ac.kr

A Study on Intelligent Digital Forensics Tool and Data Reduction Framework

Junghyun Ryu, Jaedong Lee, Jonghyuk Park*
Department of Computer Science and Engineering, Seoul National University of Science and Technology (SeoulTech), Seoul, 01811, REPUBLIC OF KOREA

요 약

범죄수사 과정에서 많은 양의 데이터를 시간 내에 분석하는 것은 성공적인 포렌식의 필수 요소이다. 컴퓨터와 사람 모두에게 있어, 시간과 자원의 제한은 수사 결과에 부정적인 영향을 가져온다. 그러므로 현재 사용되고 있는 다양한 포렌식 도구에는 시간과 자원의 효율적인 사용이 필요하다. 사례기반추론 및 멀티에이전트 시스템과 같은 인공지능 기반의 도구를 통해 디지털 포렌식 수사를 효과적으로 도울 수 있다. 본 논문에서는 인공지능을 활용한 지능형 포렌식 도구 및 프레임워크를 분석하고, 오늘날의 프레임워크의 한계점과 미래에 관해 논의한다. 인공지능 기반 시스템의 목적은 수사에서의 증거를 포함한 데이터를 분석하고 연관성을 밝힘으로서 포렌식 전문가에게 중요한 단서를 제공하고 직접 분석해야하는 데이터의 양을 줄이는 것에 있다. 이러한 인공지능의 활용은 많은 양의 데이터를 수사할 때 사람이 간과할 수 있는 증거들을 연결시켜주는 데에 큰 도움이 된다.

1. 서론

컴퓨터 시스템의 포렌식 과정은 보존, 수집, 분석의 세 단계를 거친다. 범죄 수사에서 디지털 증거는 아동 노동 착취, 문서 위조, 세금 관련 범죄 또는 심지어 테러 등의 몇몇 중대한 범죄의 중요 요소가 될 수 있다. 디지털 저장 매체 기술의 지속적인 성장과 모두의 삶에 존재하는 범용성에 따라 진화하는 저장매체 크기와 확장성에 맞추어 포렌식 수사 또한 진화가 요구된다. 이 문제를 비롯해, 현재의 포렌식 도구들은 엄청난 양의 증거와 데이터들의 연관성을 밝히는 데에 한계가 있다. 그 결과, 디지털 포렌식 전문가들은 많은 시간을 분석에 소비한다. 대부분의 포렌식 도구들은 분산 처리 기능을 갖고 있지 않기 때문에 단순 계산에 많은 시간을 소비한다. 디지털 포렌식 수사과정의 전 단계에서 인공지능이 적절히 적용된다면, 포렌식 전문가가 직접 확인해야 하는 많은 양의 데이터를 간소화하여 반복적인 작업을 줄일 수 있고 디지털 증거 간 관계를 쉽게 분석할 수 있기 때문에 시간과 컴퓨팅 자원의 효율적 운용이 가능하다. 이 논문에서는 현재 연구기관과 기업 및 정부기관에서 사용되고 있는 디지털 포렌식 도구와 인공지능 기반의 차세대 포렌식 도구를 분석하고 디지털 포렌식에서의 인공지능의 가능성에 대해 논의한다.

2장에서 현재의 디지털 포렌식 수사과정의 한계점과 관련연구에 대해 논의한다. 3장에서 지능형 포렌식 도구와 프레임워크를 분석하고 4장에서 결론을 맺는다.

2. 디지털 포렌식 수사

실제 포렌식 전문가들은 어떤 증거가 실제 사건과 관련이 있는지 사전에 파악하는 데에 어려움을 느낀다. 다수의 컴퓨터가 동일한 IP 주소를 공유하는 장소의 경우를 대표적인 예로 들 수 있다. 한 기업 내에서의 다수의 컴퓨터 또는 다른 기기에서 사기 범죄에 대한 증거를 획득할 때도 같은 어려움에 직면한다. 이러한 경우 지능형 포렌식 도구를 사용하면 분석 목표가 되는 기기를 사전 분석함으로써 수집할 기기의 수를 한정하게 되므로 포렌식 수사를 완료하는 데에 걸리는 시간을 크게 단축시킬 수 있다. 그러나 문제는 사전 분석 과정에서 포렌식 전문가를 도울 분석 도구가 부족하다는 점이며, 이는 분석하기 위해 많은 양의 기기를 수집해야하는 결과를 낳고 수집된 기기 중 일부는 수사 전반에 걸친 결과에 도움을 주지 못하여 수사에 드는 시간을 증가시킬 수 있다.

위에서 기술한 바와 같이 지능형 분석 도구의 필요성을 시사하고 수사 과정에서 더 효율적으로 시간(또는 자원)을 활용할 수 있는 프레임워크를 논하는 것이 이 연구의 초점이다. 포렌식 전문가가 각 기기에 대한 적절한 분석, 증거 대조 및 분석을 통한 연계성 확인, 한정된 자원으로 많은 양의 수집된 데이터를 고려하는 것은 어려운 일이다 [1]. 가장 단순한 경우에서 전문가는 기기의 내용물이 제한적인 단독형 기기를 분석하지만, 컴퓨터는 보통 데이터를 교환할 때 네트워크에 연결된 상태이므로 이는

일반적인 경우가 아니다. 게다가, 이동식 저장장치의 용량이 날로 커져감에 따라 한정된 상황에서의 분석에 관한 데이터의 손실 가능성, 비효율적인 수사 등의 문제는 점점 증가한다. 이런 컴퓨터와 이동식 미디어 장치는 구체적인 교차 분석을 도와주는 도구의 부족함 때문에 개별적으로 분석된다. 지금까지 제시된 문제의 결과로 볼 때, 상당한 양의 잠재적인 증거들이 포렌식 분석 과정에서 분실될 가능성이 있다. 이는 컴퓨터 포렌식의 문제일 뿐만 아니라, 네트워크 사고 대응에 관한 문제이기도 하다.

기존의 디지털 포렌식 수사과정에 있어 전문가의 피로도를 덜어주기 위해 필요한 인공지능의 역할은 크게 중요한 증거를 선별하여 분류하는 것과 분석에 필요한 계산을 분산처리를 통해 수행하는 것이다. 이와 관련해 몇몇 연구들이 존재한다. V.Ganesh는 그의 논문에서 Multi-Agent System(MAS)과 Intelligent Software Agent(ISA) 및 Multi-Agent Digital Investigation toolKit(MADIK)기반의 지능형 분석 도구를 제안하였다 [1]. 이는 디지털 포렌식 분석 과정에서 필요한 기능 별 Agent를 분리하여 분산 처리를 가능하게 하였고 통합적인 해결책을 제시함으로써 전문가에게 효율적인 분석 환경을 제공해 통합적인 해결책을 제시하였다. D.Quick과 K.K.R.Choo는 그의 논문에서 디지털 증거 획득을 위해 분석해야하는 데이터의 양을 줄일 수 있는 새로운 프레임워크를 제안했다 [2]. 이는 물리적 미디어 또는 포렌식 카피를 획득하는 과정에서 데이터를 간소화하고 논리적 증거 컨테이너를 생성하여 포렌식 전문가에게 새로운 데이터 분석 프레임워크를 제공한다.

3. 지능형 디지털 포렌식 도구

디지털 포렌식을 위한 인공지능 연구는 기술의 발전으로 날마다 커지는 저장매체의 용량에 맞추어 데이터 분석 도구를 개선하는 데에 초점이 맞추어져있다. 디지털 포렌식 전문가, 연구자, 분석가는 방대해진 데이터의 양으로 인해 데이터 분석에 어려움을 겪고 있다. 단순 자동화로 해결 할 수 없는 디지털 증거 획득 및 데이터 선별 과정을 사람이 직접 하지 않고 인공지능에 의해 처리할 수 있다면 전문가의 어려움을 해결하고, 오랜 시간이 걸리는 데이터 분석 과정에서의 데이터 손실을 막을 수 있다. 지능형 디지털 포렌식 도구가 되기 위한 요구사항은 다음의 세 가지이다.

-분석을 위한 디지털 증거의 양 축소
-수집된 증거들의 관계 증명
-분석을 위한 컴퓨팅 작업의 분산 처리 수행

이를 만족하는 지능형 포렌식 도구를 개발하여 대용량 저장 매체를 효과적으로 분석하기 위해서, 멀티 에이전트 시스템(Multi-agent system)이라는 개념이 등장한다.

3.1 멀티 에이전트 시스템(Multi-agent system) 기반의 지능형 포렌식 도구

이 시스템에서 각 에이전트(Agent)는 인공지능 기술이 적용된 지능형 소프트웨어 에이전트(Intelligent Software Agent, ISA)이다. 지능형 소프트웨어 에이전트는 환경과 상호작용하며 정해진 목표를 성취하기 위해 자율적으로 인지하고 행동한다. 지능형 소프트웨어 에이전트가 모여 만들어진 멀티 에이전트 시스템은 각 에이전트가 개별적, 집단적 목표를 달성하기 위해 협력적 혹은 경쟁적으로 환경과 상호작용한다.

Hoelz외 연구자들은 멀티 에이전트 시스템 기반의 디지털 포렌식 도구인 MADIK(Multi-Agent Digital Investigation toolKit)을 그의 논문에서 소개하였다 [3]. 이 시스템은 사건에 대한 디지털 증거를 각각 다른 방향으로 분석하는 지능형 소프트웨어 에이전트(ISA)들의 집합으로 구성되어있다. MADIK에서 각 에이전트는 디지털 포렌식 전문가들의 경험에 기초한 일련의 규칙과 지식 기반을 포함한다. 범죄 수사에 대한 디지털 포렌식 분석과정은 일련의 유사성을 공유하기 때문에 MADIK시스템은 사례기반추론 시스템을 사용하여 어떤 에이전트가 사건에 적합한지 판단한다. 따라서 에이전트는 특정 사건에 더 적합한 방식으로 증거를 추론할 수 있다. 그러므로 전문가에게 증거에 관한 파일의 존재 여부를 더욱 빨리 제공할 수 있다. MADIK은 총 여섯 개의 전문화된 지능형 에이전트를 포함한다. 그 내용은 아래의 표와 같다.

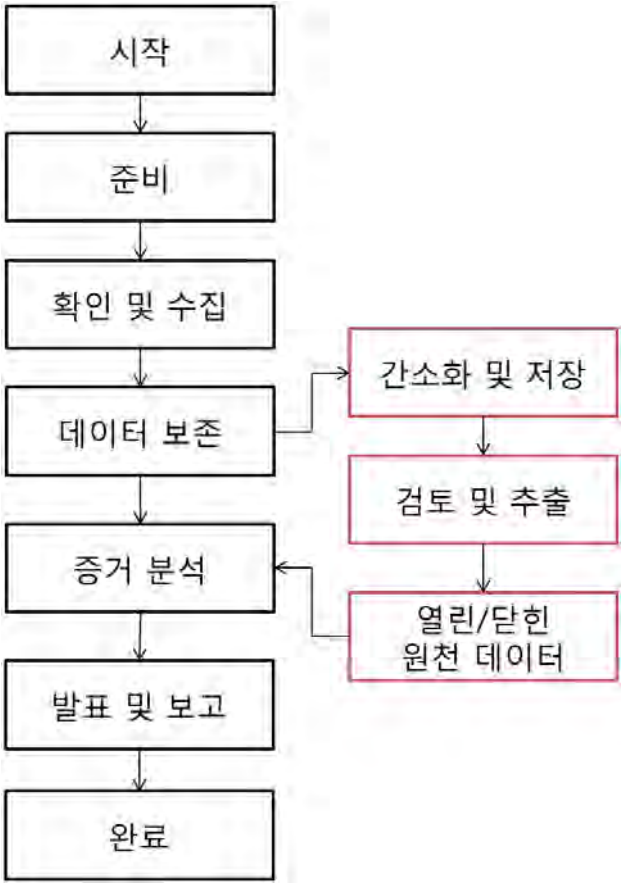
<표 1> MADIK의 Agent 구성

HashSet Agent	중요하다고 판단되는 데이터의 MD5 해시를 지식 기반을 통해 계산 및 비교
FilePath Agent	수사에 도움이 되는 애플리케이션에서 일반적으로 사용되는 폴더의 모음을 판단
FileSignature Agent	파일 헤더를 검사하고 파일 확장자를 검사하며, 일반적으로 사용되는 파일의 정보 저장
TimeLine Agent	시스템, 소프트웨어 설치, 백업 등의 활동 이벤트를 알아내기 위해 타임라인을 검사
Windows Registry Agent	윈도우 레지스트리와 관련된 파일을 검사하고 Time zone 설정, 이동식 장치 정보 등을 추출
Keyword Agent	신용카드, URL, 이메일 주소와 같은 정보를 파일로부터 추출하기 위해 키워드 검색

MADIK시스템은 일반적으로 사용되는 포렌식 도구들을 완전히 대체할 수 는 없으나, 전문가에게 더 나은 분석 환경을 제공하기 위한 새로운 개념이라고 볼 수 있다. 이러한 지능형 포렌식 도구를 이용하여 현재의 포렌식 도구가 제공하는 단순한 데이터 획득 및 추출을 넘어서 결과물 간 관계를 파악하고 분석할 수 있으며 분산 처리를 통해 컴퓨팅 자원을 더 효율적으로 사용할 수 있다. 또한 사례기반접근은 이전의 사건들을 통해 학습함으로써 더 나은 결과물을 제공할 수 있게 한다.

3.2 데이터 간소화 프레임워크

성공적인 포렌식 수사를 위해서 포렌식 전문가에게 주어지는 데이터는 사전에 간소화되고 정교하게 추출되어야 한다. 이와 관련해서 D.Quick과 K.K.R. Choo는 데이터 축소 및 추출에 관한 새로운 프레임워크를 제안했다 [2]. 이 새로운 프레임워크의 전반적인 과정은 다음의 그림과 같다.



(그림 1) 데이터 간소화 프레임워크

데이터 간소화 프레임워크에는 기존의 디지털 포렌식 프레임워크에서 추가된 세 가지 단계가 있다. 그림 1에서 붉게 표시된 상자가 이에 해당하는데, 주로 디지털 증거로써 잠재력을 가지는 데이터의 양을 줄이는 것에 초점이 맞추어져 있다.

그림 1에서 간소화 및 저장 단계는 데이터에 대한 간소화가 수행되기 때문에 잠재적 디지털 증거가 될 수 있는 데이터가 변경될 수 있으므로 적합한 상황이 아니라면 수행하지 않거나, 혹은 반드시 증거 보존 단계 이후에 수행되어야 한다. 이는 디지털 증거를 법정까지 가져가기 위한 필수적 요건이다. 디지털 증거를 다룰 때, 영국의 Association of Chief Police Officers(ACPO)와 같은 기관이 제공하는 가이드라인이나 기타 에이전시가 세부적인 규제를 하지 않는 경우 원본 데이터가 우연히 변경될 가능성이 있다. 따라서 포렌식 가이드라인은 프레임워크의 전체 과정을 고려할 수 있는 규제를 만들어야 한다. 간소화 프로세스는 증거 보존 과정에 절대 영향을 미쳐서는 안 된다. 디지털 포렌식 전문가는 법정까지의 증거 제출을 위해 반드시 현재의 관행을 따라야 하지만 증거에 관한 법원의 수용만이 포렌식 전문가의 관심사는 아니다. 포렌식 전문가는 수많은 분석 대상 기기와 그에 따르는 분석 과정에 많은 시간을 소비하고 있다 [4]. 디지털 증거로써의 잠재력을 가진 파일은 사전에 선별되고 논리적 증거 컨테이너에 보존된다. 저장 장치를 이미징하는 대신 데이터의 중요도에 초점을 맞추어 분석해야하는 데이터의 크기를 크게 줄일 수 있다. 따라서 간소화 과정은 모든 사건에 대해 모든 저장 장치를 이미징하는 부담을 완화시켜줄 수 있고, 분석 과정 전반에 걸친 소비 시간을 감소시킬 수 있다.

그 다음 단계인 검토 및 추출은 서브셋 데이터를 대상으로 수행된다. 이 단계에서 데이터는 상당 부분 간소화된 상태이기 때문에 처리 과정은 굉장히 빨라진다. 정보 검토는 인터넷 검색 기록 분석, 파일명, 타임라인 검토, 윈도우 레지스트리 분석, 키워드 검색, 해시 분석 및 포렌식 도구를 이용한 기타 통상적인 분석 기술로 구성될 수 있다. 분석에 앞서 데이터를 분류하는 기술은 수년 간 가능해졌지만, 지속적으로 커지는 데이터의 양은 분류에 걸리는 시간 또한 증가시켰다. 전체 포렌식 이미지대신 데이터 서브셋을 분류함으로써, 데이터 처리와 분류에 대한 시간을 상당히 절약 할 수 있다.

마지막으로, 열린/닫힌 데이터 단계는 이전 단계에서 얻어진 정보를 활용한다. 이전 단계의 검토에서 얻어진 정보는 열린/닫힌 데이터와 같은 다른 정보를 얻는데 사용할 수 있다. 닫힌 데이터는 기밀성이 요구되는 내부 보고서나 다른 정보 자산 등을 포함할 수 있다. 열린 데이터는 SNS나 블로그와 같은 공개적인 인터넷에서 수집된 정보들을 포함한다. 전 단계인 검토 혹은 이 단계에서 얻어진 정보는 다음 단계인 증거 분석 단계의 자료로 제공되고 수사와 관련되거나 증거로서의 가치가 있는 정보로서 지식 기반을 개선시키는데 사용된다.

4. 결론

본 논문에서는 오늘날의 디지털 포렌식의 한계점과 배경에 대해 논하고, 이를 극복하기 위한 지능형 포렌식 도구와 프레임워크를 분석한다. 갈수록 저장 매체의 크기가

방대해지는 현재의 컴퓨팅 환경에서 성공적인 디지털 포렌식 수사가 이루어지기 위해서는 포렌식 전문가를 지원할 수 있는 지능형 도구의 적용과 이를 위한 프레임워크의 개선이 필요하다.

이 논문에서 언급한 지능형 포렌식 도구와 데이터 간소화 프레임워크는 저장매체 크기의 성장에 따라 발생하는 디지털 포렌식 수사의 어려움을 극복하기 위한 대표적인 시도이다. 이는 포렌식 전문가가 반복적으로 수행해야 하는 일련의 작업을 단순화하고 분석 대상인 기기 및 데이터를 간소화하여 시간(또는 자원)활용에 효율성을 가져오며, 데이터 손실의 가능성 또한 낮출 수 있다.

현존하는 인공지능 기반의 포렌식 도구들은 포렌식 수사의 전체 과정을 고려하고 있지 않다. 따라서 수사 전반에 걸쳐 포렌식 전문가를 효과적으로 도울 수 있는 지능형 포렌식 도구의 개발이 이루어져야한다. 개선된 프레임워크는 그 과정이 포렌식 수사에서 데이터의 무결성, 연계보관성을 해칠 수 있으므로 관련 법 및 규제가 반드시 함께 마련되어야 한다.

Acknowledgenment

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(No. 2016R1A2B4011069).

참고문헌

[1] Ganesh Vaishnavi. "Artificial Intelligence Applied to Computer Forensics." International Journal 5.5 (2017).

[2] Quick Darren, and Kim-Kwang Raymond Choo. "Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive." (2014).

[3] Hoelz, Bruno WP, et al. "Madik: A collaborative multi-agent toolkit to computer forensics." OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg, 2008.

[4] Quick, Darren, and Kim-Kwang Raymond Choo. "Big forensic data reduction: digital forensic images and electronic evidence." Cluster Computing 19.2 (2016): 723-740.

[5] Ali, Azliza Mohd, and Plamen Angelov. "Applying Computational Intelligence to Community Policing and Forensic Investigations." Community Policing-A European Perspective. Springer International Publishing, 2017. 231-246.

[6] Quick, Darren. "Digital forensic data and intelligence: Using data reduction to enable intelligence analysis." Journal of the Australian Institute of Professional Intelligence Officers 23.2 (2015): 18.

[7] Tassone, Christopher FR, Ben Martini, and Kim Kwang Raymond Choo. "Visualizing digital forensic datasets: a proof of concept." Journal of forensic sciences (2017).

[8] Lillis, David, et al. "Current Challenges and Future Research Areas for Digital Forensic Investigation." arXiv preprint arXiv:1604.03850 (2016).

[9] Chauhan, Pranay, and Pratosh Bansal. "Emphasizing on Various Security Issues in Cloud Forensic Framework." Indian Journal of Science and Technology 8.1 (2017).

[10] Harper, Jack. "Artificial intelligence methods for difficult forensic fingerprint collection." U.S. Patent No. 9,342,732. 17 May 2016.