Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.23, NO.6, Dec. 2013 ISSN 2288-2715(Online)
http://dx.doi.org/10.13089/JKIISC.2013.23.6.1069

사례기반추론기법을 적용한 침해사고 프로파일링 시스템*

한 미 란,[1†] 김 덕 진,[2] 김 휘 강[1‡]
[1]고려대학교 정보보호대학원, [2]한국전자통신연구원 부설연구소

Applying CBR algorithm for cyber infringement profiling system*

Mee Lan Han,[1†] Deok Jin Kim,[2] Huy Kang Kim[1‡]
[1]Graduate School of Information Security, Korea University,
[2]The Attached Institute of ETRI

요 약

최근에 발생하는 웹 사이트 해킹은 기업의 이미지와 평판에 악영향을 끼치는 큰 위협이 되고 있다. 이러한 웹 사이트 변조 행위는 해커의 정치적인 동기나 성향을 반영하기도 하므로, 행위에 대한 분석은 해커나 해커 그룹을 추적하기 위한 결정적인 단서를 제공할 수 있다. 웹 사이트에 남겨진 특정한 메시지나 사진, 음악 등의 흔적들은 해커를 추적하기 위한 단서를 제공할 수 있고, 인코딩 방법과 해커가 남긴 메시지에 사용된 폰트, 트위터나 페이스북 같은 해커의 SNS ID 또한 해커의 정보를 추적하는데 도움을 준다. 본 논문에서는 zone-h.org의 웹 해킹 사례들로부터 특성들을 추출하고, CBR(Case-Based Reasoning) 알고리즘을 적용하여 침해사고 프로파일링 시스템을 구현하였다. 해커의 흔적과 습관에 관한 분석 및 연구는 추후 사이버 수사에 있어 공격 의도를 파악하고 그에 따른 대응책을 마련하는데 있어 IDSS(Investigation Detection Support System)로써 중요한 역할을 기대 할 수 있으리라 본다.

ABSTRACT

Nowadays, web defacement becomes the utmost threat which can harm the target organization's image and reputation. These defacement activities reflect the hacker's political motivation or his tendency. Therefore, the analysis of the hacker's activities can give the decisive clue to pursue criminals. A specific message or photo or music on the defaced web site and the outcome of analysis will be supplying some decisive clues to track down criminals. The encoding method or used fonts of the remained hacker's messages, and hacker's SNS ID such as Twitter or Facebook ID also can help for tracking hackers information. In this paper, we implemented the web defacement analysis system by applying CBR algorithm. The implemented system extracts the features from the web defacement cases on zone-h.org. This paper will be useful to understand the hacker's purpose and to plan countermeasures as a IDSS(Investigation Detection Support System).

Keywords: Cyber Genome, Web Defacement, CBR, Profiling, Incident Response

I. 서 론

특정 기업체나 기관을 지속적으로 공격하는 APT (Advanced Persistent Threat) 공격은 전문적인 팀을 이루어 조직적으로 활동하며 전 세계 정부와 기업의 네트워크에 실제적이고 지속적인 위협을 가한다. 2013년 3월20일과 25일, 26일 연이어 국내에 사이버테러가 발생하였다. 이번 사건은 방송 및 금융사 여섯 군데의 전산 장비 파괴와 대북 및 보수단체 홈페이

접수일(2013년 6월 3일), 수정일(1차: 2013년 10월 1일, 2차: 2013년 10월 28일), 게재확정일(2013년 10월 29일)
* 본 연구는 미래창조과학부 및 정보통신산업진흥원의 IT융합 고급인력과정 지원사업의 연구결과로 수행되었음 (NIPA-2013-H0301-13-3007).
* 본 연구는 ETRI부설국가보안기술연구소의 연구지원으로 수행되었음.
† 주저자, blosst@korea.ac.kr
‡ 교신저자, cenda@korea.ac.kr(Corresponding author)

지의 자료가 삭제되는 사건이었다. 3월 20일 사이버 테러가 발생하기 한 달 전 이미 북한 내부 인터넷 주소(175.45.178.xx)에서 감염PC 원격조작 등 명령 하달을 위한 국내 경유지 접속 흔적을 남긴 바 있고, 공격 경유지 49개 중 22개가 과거 2010년 해킹 때 사용했던 경유지와 동일하였다. 그리고 북한 해커만 고유하게 사용 하는 감염 PC의 식별번호 및 감염 신호 생성코드의 소스프로그램 중 과거와 동일하게 사용한 악성코드도 18종이였다. PC 하드디스크에 'HASTATI' 또는 'PRINCPES' 등 특정 문자열을 남기고, 웹 해킹을 통해 해커들의 메시지를 전달하기도 하였다[1].

이렇게 최근 지속적으로 발생하는 해킹 사건을 통해 우리는 누가 공격을 했는지, 왜 이런 공격을 하였는지 특정 짓는 것이 더 중요해졌다. 사이버 공격을 예방하고 사전에 차단하기 위해 해커의 습성과 특징을 파악하고 앞으로 진행 될 공격 방법과 경로 유추 및 해커의 행동을 실시간으로 분석할 수 있는 기술은 사이버 공격에 대한 효과적인 기술이 될 것이다. 다양한 악성코드 분석과 해킹 사건 분석을 통해 도출되는 해커의 ID, 해커 그룹 식별은 우선적으로 대응해야 할 공격 이해에 도움을 주고, 분석 및 대응에 소요되는 시간을 절약하여 수많은 공격에 효과적으로 대응할 수 있다.

미 국방성 연구 기관 DARPA(Defense Advanced Research Projects Agency)는 Anonymous 해커 단체에 공격을 받은 이후 2010년 1월 사이버게놈(Cyber Genome) 프로젝트 개발 작업에 착수하였다. DARPA의 사이버게놈 프로젝트는 사이버 공격의 경로와 배후를 추적하기 위해 악성코드, 해커, 해커 그룹, 유포지, 공격지 등의 특성을 분석, 분류하는 프로그램이다[2]. 본 연구의 사이버게놈 프로젝트 또한 악성코드의 구조, 제작자, 제작 목적, 공격 사례 등 다양한 특징 정보를 분석하고 공격 배후의 해커 및 해커 그룹을 식별할 수 있는 프로파일링에 목적을 둔 연구이다.

오프라인 수사 시 사용되는 프로파일링 기법 중 하나로 Modus Operandi 라는 방법이 있다[10]. 이는 범죄자가 범죄 행위를 수행하는 행동, 범죄 결과가 발생하는 필수적인 행위를 말하며, 범죄자의 범죄 행위가 사건에 따른 주변의 상황 등에 상호 반응하고 범죄 행위를 수행하면서 학습 진화한다는 이론이다. 아직까지 Modus Operandi 방법을 사이버 범죄에 적용한 경우는 드물다. 본 논문에서는 사례기반추론 기법을 적용하여 침해사고 프로파일링 기법을 제시하고, 1999년부터 최근까지 웹 해킹 사례를 mirror 페이지 형식으로 수집하고 저장하는 zone-h.org 웹 사이트의 데이터를 이용하여 검증해 보았다. zone-h.org 아카이브에 저장되어 있는 웹 해킹 사례를 논문에서 제안하는 Case vector에 따라 데이터베이스화 하고, 신규 웹 해킹 사례를 가중치와 유사도 공식을 적용하여 검증해 보았다. 이와 같은 유사도 분석은 zone-h.org에서 단순 통계 자료로 공유하는 내용과 더불어 해커나 해킹 사건의 관계 파악을 위해 도움을 줄 수 있다.

2장에서는 기존에 연구되어 왔던 CBR 알고리즘과 프로파일링을 통한 오프라인 수사 기법, 그리고 웹 해킹 사례에 관한 연구에 대해 소개하고, 3장에서는 웹 해킹 분석 시스템 구조 및 적용된 알고리즘에 대해 설명한다. 4장에서는 자체 제작한 크롤러를 통해 데이터베이스화한 zone-h.org의 웹 해킹 사례를 본 논문에서 제시하는 사례기반추론 알고리즘에 적용하여 검증한 결과를 보여주고, 3.20 사이버테러에서 도출된 데이터 분석 및 터키 해커 활동 정보에 대한 분석 결과를 보여준다. 마지막으로 5장에서는 결론 및 향후 연구 방향을 제시하였다.

II. 관련 연구

2.1 웹 해킹 사례 분석

해킹 사건과 해커에 관한 연구를 하는데 있어 가장 어려운 점은 충분한 수량의 분석 데이터를 얻는 것이다. zone-h.org의 자동 로봇에 의해 실시간으로 수집되는 각각의 웹 해킹 사례 정보와 mirror 페이지에 내포되어 있는 해커와 해킹 유형과 같은 정보를 통해 해커들의 공격 횟수와 증가율, 정치적 목적을 갖고 공격한 횟수, 특정 국가를 상대로 공격한 횟수 등을 분석할 수 있다[3]. 웹 해킹의 동기는 단순 호기심이 대부분을 차지하지만 민족주의적인 성향이나 종교적인 성향을 보이는 웹 해킹도 20% 이상을 차지한다[4]. 민족적, 종교적 성향이 나아가 국제 정치나 국제 사회로까지 그 목적을 두고 해킹을 시도하는 경우를 우리는 핵티비즘이라고 한다. 해커와 정치행동주의를 뜻하는 액티비즘의 합성어로써 몰래 서버 컴퓨터에 접속하여 소프트웨어를 파괴하여 무력화시키는 행동을 말하며, 정치적인 큰 반향을 일으키기 위한 목적이 크다고 볼 수 있다[5]. 웹 해킹은 다른 해킹 방법에 비해 눈

에 띄는 시각적인 방법을 많이 사용하는 경향이 있다. 글을 통해 자신의 해킹 동기, 목적, 내용을 보여주기도 하지만, 그림이나 동영상, 오디오 파일을 활용하기도 하고, 해커의 E-Mail, SNS, 홈페이지 주소를 남기는 등 하이퍼링크 방법을 많이 사용한다[4].

2.2 사례기반추론(Case-Based Reasoning)

CBR 알고리즘은 유사한 과거 문제의 해결에 기초해서 새로운 문제를 해결하는 과정이다. 현재 발생한 문제가 과거의 사례와 정확하게 일치하지는 않더라도 과거 사례나 경험은 현재 문제에 대한 부분적인 해결책을 제시할 수 있다. 즉, 과거 사례와 지식들을 데이터베이스로 구축하여 새로운 문제가 발생했을 때 기존의 사례 데이터베이스에서 똑같거나 유사한 사례를 선택하여 그 사례가 제시하는 해결책으로 현 발생한 문제에 대한 답을 제시할 수 있다는 것이다. Fig.1.은 CBR 알고리즘의 프로세스 모델로써 네 가지 단계의 절차에 따라 진행된다.

- 검색(Retrieve) : 문제가 주어지면 과거에 해결된 사례들 중 유사한 사례들을 찾는 과정이다.
- 재사용(Reuse) : 이전에 발생했던 유사한 사례들에 대한 해결안을 그대로 제시하거나, 문제에 맞도록 사례를 수정하여 해결안으로 제시하는 과정이다.
- 수정(Revise) : 재사용 과정을 통해 해결안의 타당성을 검증하고, 만약 타당하지 못할 경우 수정하여 개선된 해결안을 제시하는 과정이다.

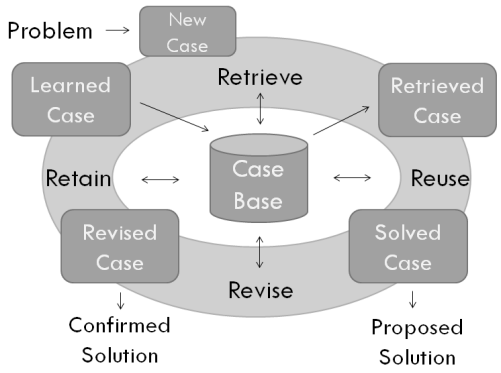

Fig.1. CBR Process Model

- 유지(Retain) : 해결안이 성공적으로 적용되면, 사례와 해결안을 기존 사례 저장소에 추가하는 과정이다.

CBR 알고리즘은 1977년 예일 대학에서 연구가 시작되어 1980년 처음 Koton과 Bareiss에 의해 의학 분야에 적용되었다. 임상의나 의사들은 수많은 경험을 통해 숙련되고, 경험과 더불어 많은 지식도 보유하게 된다. 이것은 새로운 문제 발생 시 그들의 지난 경험을 토대로 문제 해결이 가능하다는 것이며, 새로운 경험들은 또 다시 임상의나 의사들에게 더 풍부하고 새로운 사례(Case)를 경험하도록 하는 계기를 만들어 주어 의학 연구가 더 진보할 수 있었다[6]. CBR 알고리즘은 의학 분야뿐만 아니라 마케팅, 과학, 사회학 등 여러 분야에 광범위하게 적용되고 있다. 범죄 수사 분야도 예외는 아니다. 범죄 수사를 통

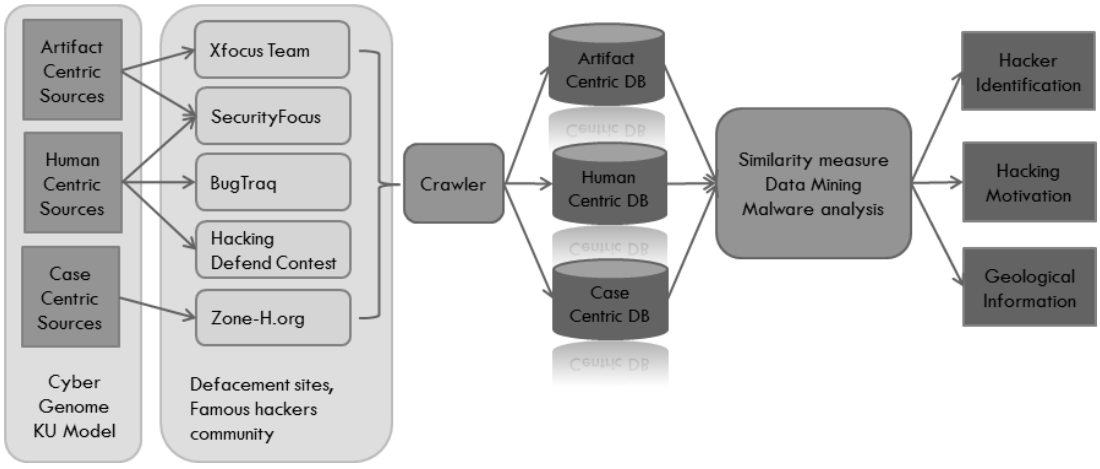

Fig.2. Architecture of overall IDSS

해 수집된 방대한 증거자료들을 효과적으로 분석하기 위해 규칙을 정하여 사례화하는데 적절한 방법이 될 수 있다[7]. 범죄 방법에 대한 자료 수집과 저장 분류는 범죄 특성과 행위 패턴을 식별하는데 도움을 준다. 범죄학에서 설명하는 범죄자 수법은 Existence, Repetitiveness, Consistency 세 가지로 구분이 된다. 첫째로, 범죄자들은 범죄 행위나 범죄 현장의 흔적, 예를 들면 지문이나 발자국 같은 증거들을 숨기려고 하나 항상 무형의 범죄 흔적을 남기게 된다는 것이다. 둘째, 상습범들은 같은 범죄 방법을 반복하는 경향이 있다. 심리학적 관점으로 볼 때 범죄자들은 그들의 목적을 달성할 때까지 지속적으로 같은 방법을 통해 범죄 행위를 하려고 한다. 셋째, 범죄자 개개인은 다른 개성과 특징, 습관이나 이력을 갖고 있으며, 일정한 범죄 행위를 반복함으로써 일관된 범죄자 형태를 갖추게 된다는 것이다. CBR 알고리즘을 통한 범죄 방법과 범죄 행위 패턴 분석 연구는 전반적인 범죄 수사 능력을 향상시키고, 사회 안전과 범죄 근절을 위한 종합적인 수사 개념 체계에 도움을 줄 수 있다[8].

2.3 수사 프로파일링(Investigation Profiling)

프로파일이란 대상자의 아주 현저한 특징들을 간단하고 분명히 인식할 수 있는 심리적, 행동적 특성을 말하며, 프로파일링은 범죄 수사에서 뿐만 아니라 사회 여러 분야에서 개인의 심리를 조사하기 위한 수단으로 많이 사용된다. 그 중에서 범죄 수사에 사용되는 프로파일링을 범죄 프로파일링이라고 하며 사건으로부터 얻는 증거와 단서를 통해 많은 정보를 알아내고 이전에 경험했던 사건을 통해 현재의 사건이 왜 일어났는지 추론하고, 그런 요소들을 통해 범인의 윤곽을 잡아 나간다[9]. 특히, 성범죄인 경우 범죄자의 프로파일 분석이 중요한 단서가 되기도 한다. 성범죄자인 경우 다른 범죄와 다르게 범죄자만의 특별한 취향 및 특성을 갖고 있다. 범행 대상자를 선택하기 위한 범죄자만의 특별한 취향이 반영이 되는데, 우선, 선택된 범죄 피해자의 성별이나 나이가 다르며, 피해자들로부터 신뢰를 얻기 위해 선물을 주거나 사랑의 표현으로 위장하여 접근 한다던가 반대로 강압적 방법으로 약이나 술 같은 도구를 사용하기도 하고 직접적으로 폭력을 행사하기도 한다는 것이다[10].

프로파일 즉, 범죄자의 심리적 행동적 성향을 드러내는 증거나 단서를 통해 범죄의 중요한 양상을 포착하고 범죄 인자를 식별할 수 있으며, 연계된 범죄들에 동일한 인자가 포함되는지 어떤 연관성이 있는지 분석할 수 있다[11][12].

III. 방법론

현재 진행 중인 사이버게놈 프로젝트는 인공물 중심 분석과 해커 중심 분석, 그리고 사례 중심 분석 세 가지 측면으로 접근하여 진행되는 연구이며, 본 논문에서는 사이버게놈을 KU Model로 명명하였다. 악성코드에 대한 정보, 해커에 대한 정보, 그리고 사례에 대한 정보가 상호 유기적으로 연동되는 구조이다. 인공물 중심 분석에서는 수집된 악성코드의 특징 추출을 모듈을 통해 진행하고, 악성코드 종류별로 데이터베이스화하여 악성코드 간 유사성을 측정한다. 해커 중심 분석에서는 해킹 관련 커뮤니티 게시물에 남겨진 악성코드 제작자에 대한 활동 정보를 추출하고 분석하며, 사례 중심 분석에서는 과거에 실제로 일어난 웹 해킹 사건을 분석하는 형태로 진행되고 있다. 본 논문에서는 세 가지 측면 중 사례 중심 분석연구에 초점을 맞추어 진행하였고, 사례기반추론알고리즘을 침해사고 프로파일링 시스템에 적용하는 방법을 제안한다.

3.1 Proposed Algorithm

3.1.1 Architecture of overall System

인공물 중심 분석과 해커 중심 분석, 그리고 사례 중심 분석의 세 가지 측면으로 연구되는 사이버게놈 프로젝트(KU Model)는 SecurityFocus, BugTraq, zone-h.org와 같은 사이트에서 악성코드와 해커 정

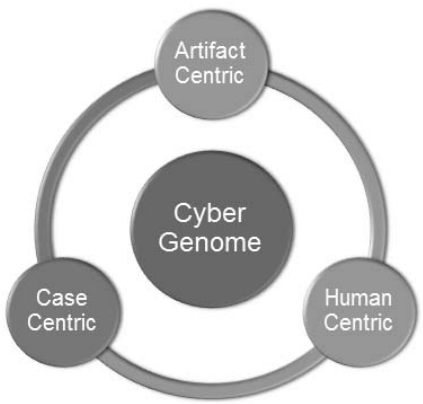

Fig.3. Cyber Genome framework(KU Model)

보, 웹 해킹 사례를 수집하는 Crawler과정을 통해 Artifact DB, Human DB, Case DB로 데이터가 저장된다. 유사도 측정과 데이터 마이닝, 그리고 악성코드 분석 과정을 통해 유의미한 데이터를 추출하고, 클러스터링과 연계과정을 거치며, 최종적으로 해커를 식별하거나 해킹 목적을 유추하고, 공격 지역을 추론하는 과정으로 사이버게놈 연구는 진행된다. Fig.2.와 Fig.3.은 사이버게놈 프로젝트의 전체적인 시스템 구조를 보여준다.

3.1.2 Web Defacement Case Analysis System

zone-h.org 아카이브에 저장되어 있는 웹 해킹 사례를 자체 제작한 크롤러를 통해 수집, 저장, 정제하여 데이터베이스화 하였고, CBR 알고리즘을 적용하여 웹 해킹 분석 시스템을 제작하였다. 웹 해킹 사례들은 zone-h.org에서 작동시키는 자동 로봇에 의해 수집되어 mirror 페이지로 출력되며, 수집된 날짜, Notify, Domain, IP Address, OS, Web server 등의 정보도 함께 보여준다. 웹 해킹한 사례의 mirror 페이지는 해커의 공격 목적이나 성향에 따라 텍스트, 이미지, 동영상, 링크 사이트, SNS, E-mail 등을 이용하여 다양한 메시지를 전한다. Fig.4.는 웹 해킹 사례 분석 시스템의 전체적인 절차를 보여준다.

본 논문에서는 5W1H 원칙(Who, When, Where, What, Why, How)에 따라 웹 해킹 사례의 Case vector를 정의하였다. 각각의 수집된 웹 해킹 mirror 페이지의 Attack Date, Target IP, Target Domain, Target OS, Target Web server, Domain Type, Country code, Encoding, Font 같은 Simple vector를 추출하여 데이터베이스화 한 후 CBR 알고리즘을 적용하여 웹 해킹 사례 분석 시스템을 구성하였다. Fig.5.는 Fig.4.의 절차로 구성된 웹 해킹 사례 분석 시스템이며, Fig.6.은 해커 중심 분석과의 교차 분석 결과로 추출된 해커 ID 정보를 통해 해커의 공격 성향, 공격의 타임라인 정보를 알 수 있는 분석 시스템이다.

3.2 Proposed Algorithm

CBR 알고리즘을 이용하기 위해서는 과거의 사례와 사례들 사이의 유사 정도를 측정하기 위한 유사도 척도가 필요하다. 본 연구에서는 웹 해킹 사례 간 유사도 알고리즘을 정의하여 데이터를 분석하였다.

3.2.1 Case Vector

사례 'A'와 사례 'B' 간의 유사도 정도를 0과 1로 정하였다. 값 0은 사례 'A'와 사례 'B' 가 관계가 없음을 의미하고, 값 1은 사례 'A'와 사례 'B' 가 같은 사례임을 나타낸다. 그리고 0 〈 S 〈1 사이의 값 S는 사례 'A'와 사례 'B'의 유사 정도를 보여준다[14].

Table 1. Similarity Values

사례A와 사례B 유사 정도	의미
0	사례A와 사례B는 유사하지 않다.
0 〈 S 〈 1	사례A와 사례B는 유사하다.
1	사례A와 사례B는 같다.

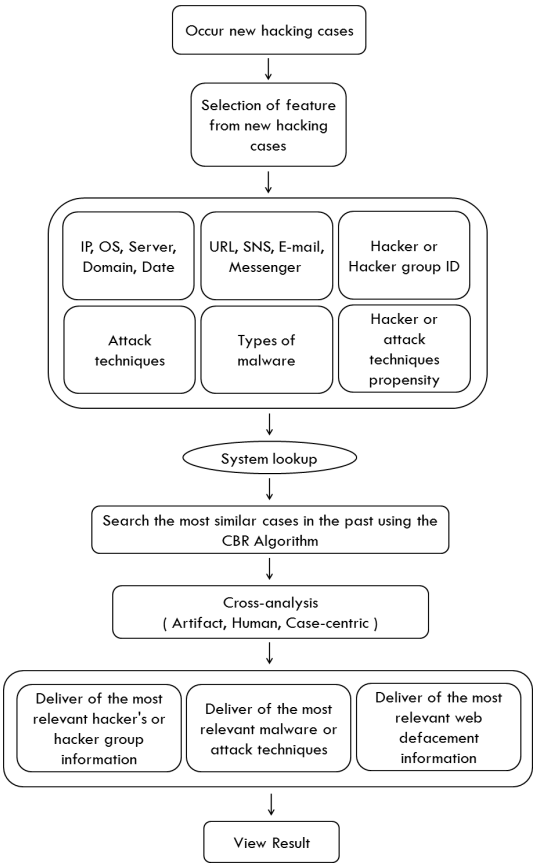

Fig.4. Procedures in analysis system of web defaced case

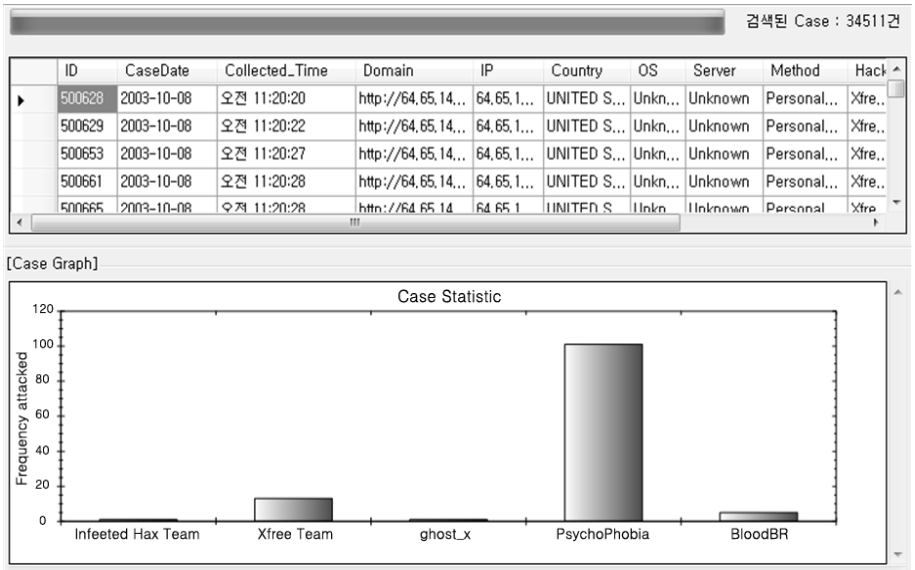

Fig.5. Screenshot of the web defacement analysis system

Fig.6. Screenshot of the Cross-analysis data of Human centric

5W1H 원칙에 따라 정의된 웹 해킹 사례의 Case vector는 Full Version과 Simple Version 으로 나눌 수 있다. Full Version의 Case vector는 zone-h.org 자동 봇에 의해 수집된 날짜, 시스템 사양 등의 정보와 mirror 페이지에서 추출된 정보 모두를 포함한다. 그러나 모든 Case vector를 다 사용하게 되면 연산 양이 많아지고, 불필요하게 겹치는 부분이 존재하게 된다. 일회성 E-mail 계정과 정보가 없거나 존재하지 않는 SNS 사용은 본인을 특정할 수 없는 경우가 많아 분석 시 혼선의 여지가 있다. 때문에 Full Version의 Case vector를 Simple Version Case vector로 다시 정의하여 실험에 적용하였다[14].

오프라인 수사기법에서 사건을 묘사할 수 있는 증거와 단서 개념을 웹 해킹 사례에 접목하여, 해킹 사례에 남겨져 있는 증거와 단서들 간의 유사도를 측정하고, 유사 정도에 따라 동일 공격 범주에 포함하여 분석이 가능하다. 오프라인 수사에서 지문, 발자국, 혈흔이나 머리카락 같은 증거와 범죄자의 말과 행동에 투영되는 어조, 눈빛, 자세 같은 단서는 웹 해킹 사례에서 Attack Date, Target IP, Target Domain, Target OS, Target Web server, Domain Type 같은 증거와 Encoding, Font, Propensity, Message, 해커와 해커 그룹과의 관계 등의 단서로 동일하게 적용하여 본 논문의 Case vector로 사용하였다.

- Attack Date : 비슷한 시기에 일어난 해킹은 유사 해킹, 동일한 공격자에 의한 동일 공격으로 추론할 수 있음.
- Target IP & Domain : 같은 혹은 비슷한 IP대역에서 공격이 왔다면 같은 해커, 같은 해킹 그룹일 확률이 높음.

Table 2. Case vector Version

Full Version	Attack Date, Notify, Target IP, Target Domain, Target OS, Target Web server, Domain Type, Thanks to, Hacker Group, E-mail, SNS, Messenger, Country code, Encoding, Font, Sound, Link site, Message.
Simple Version	Attack Date, Target IP, Target Domain, Target OS, Target Web server, Domain Type, Country code, Encoding, Font

- Target OS & Web server : Case별 유사한 시스템 사양간의 연관성을 추론할 수 있음.
- Domain Type : 공격 목적 혹은 공격 성향을 추론할 수 있음.
- Propensity : 공격 목적 혹은 공격 성향을 추론할 수 있음.
- Country code : 공격 목표를 추론할 수 있음.
- Encoding & Font : 공격자의 지역 정보를 추론할 수 있음.

3.2.2 Algorithm

zone-h.org mirror 페이지 대부분의 사례들은 Attack Date, Target IP, Target Domain, Target OS, Target Web server 정보를 포함하고 있고, Mirror 페이지의 HTML 소스를 통해 Encoding 이나 Font 정보를 추출할 수 있다. Encoding 과 Font 정보는 Attacker의 지역을 특정할 수 있는 결정적인 단서를 제공하는 중요한 Case vector이며, Attack Date, Target IP, Target Domain은 Attacker와 Victim의 관계 정보를 통해 Attacker를 특정할 수 있는 단서가 된다. 첫 번째로 Encoding은 문자나 기호들의 집합을 컴퓨터에서 저장하거나 통신에 사용할 목적으로 부호화하는 방법을 가리킨다. ISO/IEC 8859 계열의 인코딩은 로마 문자를 쓰는 다른 언어에서 ASCII로 처리할 수 없는 추가적인 기호들, 예를 들면, ß(독일어), ñ(에스파냐어), å(스웨덴어와 북유럽 언어) 등의 언어를 표현할 수 있게 해준다. MS 윈도우 문자 집합인 Windows-1250~1258 계열의 인코딩은 라틴 글자를 이용하는 중앙 유럽 언어와 터키어, 발트어, 베트남어 등의 언어를 표현할 수 있게 해준다. 중국어는 GB 계열, 대만어는 HKSCS, 한국어는 EUC-KR이나 ISO-2022-KR을 사용한다[15]. 두 번째로 Font 정보는 HTML 상에 Font family로 적용된다. 해커는 웹 해킹을 통해 메시지 전달 시 영어 외에 해커 본인이 사용하는 언어를 통해 메시지를 전달하기도 한다. HTML 상에 폰트가 적용되지 않을 경우 ASCII로 처리할 수 없는 기호들은 모두 깨져서 보이게 된다[16]. 세 번째로, 2010년 한·일 삼일절 사이버 공격 사건[17]이나 미국의 9·11 테러 12주년을 맞아 해커 단체 어넌고스트(AnonGhost)의 이스라엘에 대한 사이버 공격 예고[18], 시리아 내전이나 이집트의 정국 불안 같은 중동 내 정치적 이슈들로 인한 사이버

Table 3. The measurement formula of similarity

유사도 점수 = (Font × Weight) + (Encoding × Weight) + (IP × Weight) + (Domain × Weight) + (CC × Weight) + (Date × Weight) + (OS × Weight) + (Web server × Weight)

Table 4. The sets of weight for Case vector

	Encoding	Font	IP	Domain	Country code	Date	OS	Web server
영향도	High		Medium				Low	
가중치	9	8	6	5	5	4	2	2
값	0 또는 1	0 또는 1	1	1	0 또는 1	1	0 또는 1	0 또는 1
			0.75	0.75		0.75		
			0.5	0.5		0.5		
			0.25	0.25		0.25		
			0	0		0		

공격 급증[19] 등을 통해서도 알 수 있듯이 Attack Date, IP, Domain은 Attacker를 특정할 수 있는 또 다른 단서가 된다. 본 논문에서는 Attacker를 특정할 수 있는 Vector와 Attacker와 Victim의 관계 정보를 통해 Attacker를 특정할 수 있는 Vector를 구분하여 가중치를 다르게 설정하고 Similarity value를 계산하였다. 전체 유사도 측정 공식은 Table 3.과 같다.

Case vector의 가중치는 세 가지로 구분하여 정의하였다. Case vector가 해커나 해커 집단을 특정짓는 영향의 정도에 따라 High, Medium, Low quality information으로 나누어 가중치를 주었다. Attacker를 특정할 수 있는 Encoding 정보와 Font 정보는 High-quality information으로 정의하고, Attack Date, IP, Domain, CC(Country Code) 정보는 Medium-quality information으로 정의하고, Target OS와 Web Server는 Low-quality information으로 정의하였다. Table 4.는 Case vector에 따른 가중치와 값을 설정한 것이다. Case vector의 값들은 숫자와 문자가 섞여있고, 측정값의 기준이 동일한 것이 아니기 때문에 모든 Case vector의 유사도 측정값을 0과 1로 정규화(Normalization)하되 Attack Date, IP, Domain은 유사 범위를 구간별로 두어, 유사한 정도에 따라 0부터 1사이의 값으로 다르게 설정한다. IP vector는 Net 주소와 Host 주소의 유사 정도에 따라 구분을 하고, Domain은 gTLD(Generic top-level domain)과 ccTLD(Country code top-level domain), 그리고 기관명에 따라 구분한다.

날짜의 유사도를 구하는 것은 다음과 같다. 오프라인 범죄 수사방법과 마찬가지로 공격 날짜 (Attack Date) 특징은 근접한 시기에 발생한 범죄를 유사 범죄로 묶어 분석이 가능하다. 웹 해킹 사례를 1999년부터 2012년 10월까지의 기간으로, 월 단위로 계산하면 154개월로 표현할 수 있다. 만약 검색하고자 하는 사례를 특정 날짜(2004년 1월 1일)로 설정한다고 하자. 특정 날짜와 전후 몇 개월의 차이가 있느냐에 따라 사례에 유사도 점수를 할당하여 유사 값에 따라 내림차순으로 결과를 열거해 준다. Table 5.를 통해 근접한 시기가 얼마냐에 따라 유사도 값을 다르게 설정하여 웹 해킹 사례에 적용하였다.

Table 5. Attack Date's Similarity Measure

기간	유사도 점수
전후 6개월 이내(총 1년)	1
전후 18개월 이내(총 3년)	0.75
전후 30개월 이내(총 5년)	0.5
전후 42개월 이내(총 7년)	0.25
전후 42개월 이상(총 7년 이후)	0

IP와 같이 정성적인 범위를 정량화하기 위한 Vector는 다음과 같은 방법을 적용하였다. IP Address 공간은 8bits로 구성되어 있는 숫자 네 개의 조합으로 세 개의 클래스로 분류 되며, 각각의 클래스는 Net 주소와 Host 주소 부분으로 구분된다.

본 논문에서는 Target IP 특징을 사례 'X'와 사례 'Y' 간의 Net 주소와 Host 주소 숫자가 네 개의 단위 별로 같은지 다른지를 비교하여 유사도 점수를 할당하여 적용하였다. Table 6.에서의 예시와 같이 Net 주소와 Host 주소가 모두 같을 경우 유사를 최고점 1점으로 설정하고, Host 주소가 어디까지 같은지를 비교하여 유사도 점수를 할당하게 된다.

IP Address(사례 X) : A.B.C.D
IP Address(사례 Y) : a.b.c.d

Table 6. Attack IP's Similarity Measure

비교 조건	유사도 점수
모두 다를 경우	0
A와 a가 같을 경우	0.25
A와 a, B와 b가 같을 경우	0.5
A와 a, B와 b, C와 c가 같을 경우	0.75
모두 같을 경우	1

Domain에서 gTLD는 일반 최상위 도메인으로 영리 목적의 기업이나 단체에 사용되는 com, 정부 기관에 사용되는 go, gov 등을 말한다. ccTLD는 국가 코드 최상위 도메인으로 .kr, .cn, .br, .uk 같이 특정 지역을 나타내는 인터넷 도메인에 배당된 고유 부호를 말한다. Table 7.은 Domian의 비교 조건이다.

IV. 적용 및 실험

4.1 사례기반추론 알고리즘 적용

본 논문에서는 유사도를 기반으로 하여 해킹 공격에 대한 추론이 가능하다는 것을 입증하기 위하여, zone-h.org 웹 해킹 사례 중 해커 순위 50위권 내의 해커들의 해킹사고사례를 랜덤하게 선택하여 사례기반추론 알고리즘을 적용하였다. 알고리즘의 유사도 값과 정확도는 랜덤하게 선택된 해커 사례에 반복적으로 적용하여 값을 도출하였다. Table 8.은 랜덤하게 선택된 해커들의 신규 데이터 Case vector 이며, Fig.7.~Fig.10.은 Table 8.의 랜덤하게 선택된 해커들의 Case vector 값을 사례기반추론 알고리즘에 적용하여 도출된 값을 보여준다. 유사도 값은 본 논문에서 제안하는 알고리즘을 적용한 값이며, 정확도는 유사도 값을 100%로 환산한 값이다. 네 개의 해커 중 Islamic Ghosts Team은 신규 사례가 기존의 사례와 유사함을 보여주는 정확도가 85% 이상이 되는 것을 확인할 수 있다.

Table 7. Attack Domain's Similarity Measure

비교 조건	유사도 점수
모두 다를 경우	0
gTLD와 ccTLD, 기관명 중 한 가지가 같을 경우	0.25
gTLD와 ccTLD 가 같을 경우	0.5
기관명이 같고 gTLD와 ccTLD 중 한 가지가 같을 경우	0.75
모두 같을 경우	1

4.2 3.20 사이버테러 데이터 분석

2013년 발생했던 3.20 사이버테러 사건의 당시 해킹 주체들은 LG U+ 그룹웨어 홈페이지와 KBS 영문 홈페이지를 웹 해킹 하였다. LG U+ 웹 해킹 사례에 사용된 이미지는 3 Calaveras라는 이미지로, 주로 유럽권에서 이용된다. 그리고 웹 해킹 이후 남기고 간 메시지에 사용된 Encoding 은 서유럽 언어

Table 8. Case vector of Hacker's new data

Hacker Name	Fatal Error	Barbaros-DZ	S4t4n1c_S0uls	Islamic Ghosts Team
Date	2012-09-04	2012-12-21	2003-07-27	2012-12-29
IP	200.20.0.21	61.155.161.19	212.71.68.71	62.240.36.45
Domain	www.dadj.uff.br	www.cngl.gov.cn	gjoevik.toyota.no	agpc.gov.ly
Country Code	브라질	중국	노르웨이	리비아
OS	Linux	Win 2003	Linux	Linux
Web server	Apache	IIS/6.0	Unknown	Apache
Encoding	windows-1252	Unknown	iso-8859-1	windows-1256
Font	Courier New	Segoe Print	Courier, mono, Courier New	Comic sans ms

에서 사용되는 인코딩 값(windows-1252)으로 해커가 서유럽권 해커일 수 있다고 유추해볼 수 있다[20]. 이렇듯 웹 해킹에 사용된 이미지, 남기고 간 메시지, 해커 그룹 같은 흔적들은 기존의 해커 프로파일링 소스와 비교 분석이 가능하다. zone-h.org에서 크롤링한 웹 해킹 사례들과 3.20 사이버테러에서 나타난 흔적과의 연관성은 발견하지 못하였으나, 만약 해커 프로파일링 소스가 충분하다면 LG U+ 홈페이지를 해킹한 'Whois' 팀이 이전에 활동하던 해커 그룹인지 아닌지와 활동 시기는 언제이며, 어떤 목적으로 공격을 행하였는지에 대한 유추가 가능할 것이다. KBS 영문 홈페이지에 남겨진 메시지 'HASTATI'는 로마군 보병대의 3개 대열 중 맨 앞에 서는 부대를 의미한다. 해커가 남긴 'HASTATI'라는 단어만으로도 3.20 사이버테러가 일시적인 공격이 아닌 2차, 3차 공격이 가능한 것임을 의심할 수 있다.

Table 9. 3.20 South Korea cyber terror information

HTML 소스 내 데이터 정보	
Encoding	windows-1252
Font family	Tahoma
IP	101.106.25.105 175.45.178.xx
URL, E-mail	APTM4st3r@whois.com dbM4st3r@whois.com d3sign3r@whois.com vacc1nm45t3r@whois.com r3cycl3r@whois.com s3ll3r@whois.com
Message	who is whois? Hacked By Whois Team Hacked By HASTATI
기타	Ie4, ns6, JavaScript1.2

Date	Domain	IP	CC	System	Server	Encoding	Font
2012년 6월 23일	moodle.dc.ufscar.br	200.18.99.98	브라질	Linux	Apache	windows-1252	Courier New
2011년 9월 11일	www.nigeria.gov.ng	72.37.212.66	미국	Linux	Apache	iso-8859-1	Ebrima
2006년 4월 21일	www.legislativo.cristalina.go.gov.br	200.199.232.132	브라질	Win 2000	IIS/5.0	iso-8859-1	Courier New, Courier, mono
2009년 12월 18일	srvportal.cidadania.am.gov.br	200.242.43.143	브라질	Linux	Apache	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2012년 7월 28일	fotogrametri.ogm.gov.tr	88.255.50.36	터키	Win 2008	Apache	windows-1252	Courier New
2009년 6월 4일	www.agenciaminas.mg.gov.br	200.198.22.74	브라질	Linux	Apache	iso-8859-1	Arial
2008년 11월 21일	www.prefeiturazacarias.com.br	200.192.137.48	브라질	Linux	Apache	iso-8859-1	Unknown
2011년 4월 19일	www.sepesca.itajai.sc.gov.br	187.44.99.70	브라질	Linux	Apache	windows-1252	Cambria Math,Calibri, Tahoma
2010년 5월 15일	www.arapora.mg.gov.br	189.90.130.20	브라질	Linux	Apache	iso-8859-1	arial
2008년 12월 19일	intramed.uol.com.br	200.221.9.43	브라질	Win 2003	IIS/6.0	windows-1252	Verdana
2008년 1월 18일	www.uba.mg.gov.br	200.157.178.144	브라질	Win 2003	IIS/6.0	windows-1252	Arial, impact
2008년 5월 25일	www.brazip.com.br	74.55.26.226	브라질	Win 2003	IIS/6.0	windows-1252	Franklin Gothic Demi, Fixedsys
2009년 9월 20일	comune.abetone.pt.it	81.31.157.12	이탈리아	Linux	Apache	iso-8859-1	Fixedsys
2007년 11월 15일	www.bynrsgs.gov.cn	202.99.232.41	중국	Win 2003	IIS/6.0	windows-1252	Verdana, WST_Span
2006년 11월 15일	eptv.globo.com/promocao	201.7.183.211	Unknown	Win 2003	IIS/6.0	iso-8859-1	Arial, Helvetica, sans-serif

Date	IP	Domain	CC	OS	Server	Encoding	Font	Similarity	Accuracy
1	0.25	0.25	1	1	1	1	1	32.75	79.88%
1	0	0	0	1	1	1	1	25	60.98%
0	0.25	0.25	1	0	0	1	1	24.75	60.37%
0.75	0.25	0.25	1	1	1	1	0	23.75	57.93%
1	0	0	0	0	1	1	1	23	56.10%
0.5	0.25	0.25	1	1	1	1	0	22.75	55.49%
0.5	0.25	0.25	1	1	1	1	0	22.75	55.49%
0.75	0	0.25	1	1	1	1	0	22.25	54.27%
0.75	0	0.25	1	1	1	1	0	22.25	54.27%
0.5	0.25	0.25	1	0	0	1	0	18.75	45.73%
0.5	0.25	0.25	1	0	0	1	0	18.75	45.73%
0.5	0	0.25	1	0	0	1	0	17.25	42.07%
0.75	0	0	0	1	1	1	0	16	39.02%
0.5	0	0	0	0	0	1	0	11	26.83%
0.25	0	0	0	0	0	1	0	10	24.39%

Fig.7. Similarity and accuracy of hacker named 〈Fatal Error〉

Date	Domain	IP	CC	System	Server	Encoding	Font
2012년 8월 3일	www.jiantouji.gov.cn	61.156.45.13	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 6월 25일	wap.qichun.gov.cn	61.184.131.10	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 3월 21일	minjian.huzhou.gov.cn	61.175.225.133	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 11월 19일	zjj.tx.gov.cn	220.191.222.10	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 10월 22일	hnhh.hnforestry.gov.cn	222.240.131.223	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 9월 6일	jgsw.huaihua.gov.cn	220.169.97.6	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 8월 29일	www.yfzf.gov.cn/dz.htm	219.159.107.165	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 8월 13일	qys.yuanjiang.gov.cn	220.170.157.111	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 7월 28일	www.sgcin.gov.cn	60.215.129.71	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 7월 23일	www.wh.qj.gov.cn	220.164.174.9	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 7월 11일	gzw.huaihua.gov.cn	220.169.97.6	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 6월 30일	gsj.xzjw.gov.cn	58.218.171.2	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 6월 9일	cgw.huishan.gov.cn	218.90.157.36	중국	Win 2003	IIS/6.0	Unknown	Segoe Print
2012년 10월 3일	jh.yclgb.gov.cn	221.231.124.23	중국	Win 2008	IIS/7.0	Unknown	Segoe Print
2012년 7월 2일	www.suilin.gov.cn		중국	Win 2003	IIS/6.0	gb2312	Unknown

Date	IP	Domain	CC	OS	Server	Encoding	Font	Similarity	Accuracy
1	0.25	0.5	1	1	1	0	1	25	60.98%
1	0.25	0.5	1	1	1	0	1	25	60.98%
1	0.25	0.5	1	1	1	0	1	25	60.98%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	1	1	0	1	23.5	57.32%
1	0	0.5	1	0	0	0	1	19.5	47.56%
1	0	0.5	1	1	1	0	0	15.5	37.80%

Fig.8. Similarity and accuracy of hacker named 〈Barbaros-DZ〉

Date	Domain	IP	CC	System	Server	Encoding	Font
2002년 7월 18일	www.daewoo.es	212.31.45.154	스페인	Win NT9x	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2003년 1월 25일	www.doh.gov.tw	203.65.100.154	대만	Win 2000	Unknown	iso-8859-1	Courier New, Courier, mono
2002년 6월 10일	www.nask.navy.mil	160.128.241.15	Unknown	Windows	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2002년 5월 31일	www.em.gov.bc.ca	142.36.87.104	캐나다	Win NT9x	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2002년 5월 15일	dialupip14.shore.co.monmouth.nj.us	199.233.80.23	미국	Windows	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2002년 5월 7일	lrc.doe.state.de.us	167.21.203.171	미국	Windows	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2002년 4월 17일	wdc.govt.nz	203.109.230.1	뉴질랜드	FreeBSD	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2001년 12월 29일	www.co.carver.mn.us		미국	Windows	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2001년 11월 2일	www.dot.state.sc.us		미국	Windows	Unknown	iso-8859-1	Verdana, Arial, Helvetica, sans-serif
2002년 12월 9일	www.ihemsys.gob.mx	148.223.153.94	멕시코	Win 2000	Unknown	iso-8859-1	Verdana
2002년 9월 1일	msgate.pstripes.osd.mil	208.242.80.9	Unknown	Win NT9x	Unknown	iso-8859-1	Tahoma
2002년 9월 30일	webmail.devon.gov.uk	159.15.11.223	영국	Win NT9x	Unknown	Unknown	Times New Roman, Times, serif
2002년 7월 29일	www.ttkayseri.telekom.gov.tr	212.175.39.57	터키	Win 2000	Unknown	Unknown	Tahoma
2002년 9월 13일	www.tupc.tcg.gov.tw/default.asp	163.29.37.164	대만	Win NT9x	Unknown	Unknown	Tahoma
2002년 8월 9일	www.asics.co.kr	203.234.222.225	한국	Win NT9x	Unknown	Unknown	Tahoma

Date	IP	Domain	CC	OS	Server	Encoding	Font	Similarity	Accuracy
0.75	0.25	0	0	0	0	1	1	21.5	52.44%
1	0	0	0	0	0	1	1	21	51.22%
0.75	0	0	0	0	0	1	1	20	48.78%
0.75	0	0	0	0	0	1	1	20	48.78%
0.75	0	0	0	0	0	1	1	20	48.78%
0.75	0	0	0	0	0	1	1	20	48.78%
0.75	0	0	0	0	0	1	1	20	48.78%
0.75	0	0	0	0	0	1	1	20	48.78%
0.75	0	0	0	0	0	1	1	20	48.78%
1	0	0	0	0	0	1	0	13	31.71%
1	0	0	0	0	0	1	0	13	31.71%
1	0	0	0	0	0	0	1	12	29.27%
1	0.25	0	0	0	0	0	0	5.5	13.41%
1	0	0	0	0	0	0	0	4	9.76%
1	0	0	0	0	0	0	0	4	9.76%

Fig.9. Similarity and accuracy of hacker named 〈S4t4n1c_S0uls〉

Date	Domain	IP	CC	System	Server	Encoding	Font
2009년 11월 3일	www.cbl.gov.ly	62.240.36.40	리비아	Linux	Apache	windows-1256	comic sans ms,sans-serif
2011년 10월 15일	www.jnpc.gov.ly	62.240.36.45	리비아	Linux	Apache	windows-1256	Unknown
2009년 4월 4일	bakreu.go.th	61.47.40.41	태국	Linux	Apache	windows-1256	Comic Sans MS, Bradley Hand ITC
2010년 7월 31일	libyan-embassy.co.uk	62.240.36.52	영국	Linux	Apache	windows-1256	Unknown
2010년 12월 26일	www.mof.gov.ly/site/index.php	74.50.3.9	리비아	Win 2003	IIS/6.0	windows-1256	Monotype Koufi
2010년 3월 28일	hctqe.gov.sd/ar/	72.52.208.200	수단	Linux	Apache	windows-1256	Tahoma
2010년 1월 21일	www.carc.gov.jo/index_ar.php	72.9.148.232	요르단	Linux	Apache	windows-1256	Tahoma, Century Gothic
2010년 1월 9일	www.cairomoe.gov.eg/vb/	212.103.160.154	이집트	Linux	Apache	windows-1256	Impact
2010년 10월 28일	cspd.gov.jo/index.php	193.188.66.122	요르단	Solaris 9/10	Apache	windows-1256	Californian FB
2010년 6월 20일	bappekab.jemberkab.go.id	202.80.113.21	인도네시아	FreeBSD	Apache	UTF-8	Comic Sans MS
2011년 9월 4일	www.plaibang.go.th/person.php	203.147.62.86	태국	Linux	Apache	windows-874	Arial, Helvetica, sans-serif
2010년 9월 18일	npcih.gov.pk	66.40.7.9	파키스탄	Linux	Apache	windows-1252	Arial Narrow
2011년 12월 16일	www.adnancy.cef.fr/x.html	195.214.243.10	프랑스	Linux	Apache	ISO-8859-1	Times New Roman
2011년 8월 11일	www.kia.co.zw	196.44.177.66	잠바브웨이	Linux	Apache	ISO-8859-1	Andalus
2009년 6월 9일	www.dwr.go.th	202.129.59.68	태국	Win 2003	IIS/6.0	windows-874	Tahoma

Date	IP	Domain	CC	OS	Server	Encoding	Font	Similarity	Accuracy
0.5	0.75	0.5	1	1	1	1	1	35	85.37%
0.75	1	0.5	1	1	1	1	0	29.5	71.95%
0.5	0	0.25	0	1	1	1	1	24.25	59.15%
0.75	0.75	0	0	1	1	1	0	20.5	50.00%
0.75	0	0.5	1	0	0	1	0	19.5	47.56%
0.75	0	0.25	0	1	1	1	0	17.25	42.07%
0.75	0	0.25	0	1	1	1	0	17.25	42.07%
0.75	0	0.25	0	1	1	1	0	17.25	42.07%
0.75	0	0.25	0	0	1	1	0	15.25	37.20%
0.75	0	0.25	0	0	1	0	1	14.25	34.76%
0.75	0	0.25	0	1	1	0	0	8.25	20.12%
0.75	0	0.25	0	1	1	0	0	8.25	20.12%
0.75	0	0	0	1	1	0	0	7	17.07%
0.75	0	0	0	1	1	0	0	7	17.07%
0.5	0	0.25	0	0	0	0	0	3.25	7.93%

Fig.10. Similarity and accuracy of hacker named 〈Islamic Ghosts〉

Fig.11. LG U+ Groupware homepage's case

Fig.12. KBS English homepage's case

4.3 터키 해커 그룹 및 연계 분석 시나리오

TurkguvenLigi.info 해커 그룹은 2008년부터 2013년 최근까지 활동한 터키 해커 그룹이다[21]. zone-h.org 아카이브에 저장되어 있는 이들 해커 그룹 데이터를 Simple Version Case vector와 Full Version의 Case vector 중 메시지 정보와 해커나 해커 그룹 간의 관계정보까지 추가하여 분석하면 좀 더 명확하게 이들이 어느 국가 혹은 어느 민족성을 띄는 그룹이고, 어떤 목적을 위해 웹 해킹을 하는지 의도를 파악할 수 있다. Attacker의 지역을 특정하는데 중요한 역할을 하는 인코딩 정보를 보면, 터키어는 라틴문자를 토대로 약간의 변형을 가한 터키어 로마자로써, 아제르바이잔어와 카자흐어와 유사성을 띈다. 그리고 "Ağaç(나무)" 나 "Ateş(불)" 처럼 알파벳만으로 표현되지 않는 문자들이 존재하기 때문에, 웹 해킹 시 Encoding으로 windows-1254[22]와 iso-8859-1[23]을 사용하게 된다. windows-1254와 iso-8859-1는 터키어를 쓰기 위해 마이크로 소프트에서 사용되는 코드페이지이다. 이들은 또한 웹 해킹 사이트에 "TurkguvenLigi" 와 "4 Sept. We TurkGuvenligi declare this day as World Hackers Day" 라는 메시지를 남기기도 하였으며, " muslims", " israel", "palestinian" 같은 민족이나 국가를 뜻하는 단어를 많이 사용하기도 하였고, "genocide(학살)", "Sunar(선물)", "Babalar(아버지)", "pentagon(펜타곤)", "saldırmak (공격)", "basınında(압박)" 같은 공격적 성향의 단어를 남기기도 하였다. TurkguvenLigi.info 해커 그룹은 웹 해킹된 사이트에 "Thanks to NetDeviZ" 라고 문구를 남겼다. NetDeviZ 해커 그룹은 2008년 2월부터 2012년 7월까지 꾸준히 활동해 온 터키 해커 그룹이다. TurkguvenLigi.info 해커 그룹과 NetDeviZ 해커 그룹의 활동 시기가 유사한 점과 웹 해킹 사이트에 "Thanks to NetDeviZ" 문구를 남기는 것 등을 보았을 때 NetDeviZ 해커 그룹이 터키 출신의 해커 그룹일 수 있다고 유추 가능하다[24].

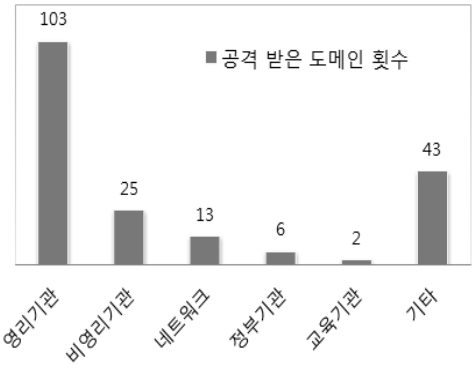

Fig.13. Attack domain of hacker group named 〈TurkguvenLigi.info 〉

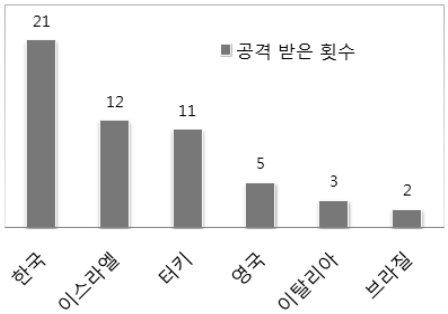

Fig.14. Attack Nation of hacker group named 〈TurkguvenLigi.info 〉

Table 10. Similarity of Turkish hacker group

	사례1	사례2	사례3	사례1 vs 사례2	사례1 vs 사례3	사례2 vs 사례3
Attack Date	2010년 2월 26일	2010년 9월 25일	2006년 5월 10일	1×4=4	0.5×4=2	0.5×4=2
Target Domain	tg.devturkler.com	www.destinyschild.com	www.aulavirtual.misiones.gov.ar	0.25×5=1.25	0×5=0	0×5=0
Target IP	174.121.56.105	174.143.100.66	200.45.71.35	0.25×6 =1.5	0×6=0	0×6=0
Target OS	Linux	Linux	Win 2003	1×2=2	0×2=0	0×2=0
Target Server	Apache	Apache	IIS/6.0	1×2=2	0×2=0	0×2=0
Encoding	ks_c_5601-1987	ks_c_5601-1987	Unknown	1×9=9	0×9=0	0×9=0
Font	Times New Roman, sans-serif	Arial, Helvetica, sans-serif	Times New Roman	1×8=8	1×8=8	1×8=8
유사도 점수 (유사도 최고 점수 41)				27.75	10	10

Table 10.은 TurkguvenLigi.info 해커 그룹 사례들의 유사도 비교와 TurkguvenLigi.info 해커 그룹과 무관한 사례와의 유사도를 비교한 결과를 보여준다. 사례1과 사례2는 27.75로 유사도 값이 높은 반면에 사례1과 사례3, 사례2와 사례3은 상대적으로 낮은 것을 알 수 있다. Fig.13.은 TurkguvenLigi.info 해커 그룹이 공격한 도메인 수치이다. 영리기관을 통해 해커들의 메시지를 알리는 경우가 가장 많았다. Fig.14.는 공격한 국가의 수치이다. 192번의 해킹 공격 중 한국이 가장 많으며, 해커가 전하려는 정치적, 민족적 메시지를 선전하기 위해 한국을 선택했다. 수집된 데이터 상의 해킹 공격 시기와 비교해 볼 때, 12번의 이스라엘 공격은 2010년 5월 이스라엘 군이 가자지구로 향하던 구호선을 공격해 터키인 9명이 숨진 이후 이스라엘과 갈등이 심화되었던 사건과 무관하지 않다.

zone-h.org에서는 해커들의 공격의도를 아래 7가지로 분류하여 공유하고 있다. 본 논문에서는 몇 가지 공격 의도를 더 추가하였다. zone-h.org 웹 해킹 데이터를 자체 분석한 결과 Patriotism(애국)을 Racism(민족주의)과 Religion(종교)로 세분화 시킬 수 있었으며, Criminals demonstrating(범죄 시연), diversion to cover compromise(교란 작전), Money(돈)을 요구하는 경우도 존재하였으며, 해커 그룹의 Propagation(선전)을 위한 경우도 존재했다.

Table 11. Various attack Reason(zone-h.org)

Attack Reason
I just want to be the best defacer
Heh...Just for fun!
As a challenge
Political reasons
Patriotism
Revenge against that website
Not available

Fig.15.는 사이버게놈의 세 가지 모듈인 인공물 중심 분석과 해커 중심 분석, 그리고 사례 중심 분석을 통해 추출할 수 있는 정보들이 어떤 방식으로 연계되어 분석되는지 보여준다. 악성코드 샘플이 인공물 중심 분석으로 입력되어 들어가면, 해당 악성코드에서 추출된 해커 정보나 해커의 습성을 보여주는 정보들은 해커 중심 분석과 연계되어 해커 커뮤니티에서 자주 사용하는 사용자 ID나 소속되어 있는 해커 그룹, 그리고 주요 관심사 같은 해커의 활동 정보들을 강화시킨다. 악성코드 메타 데이터의 컴파일러, Packer, String, 분류 등의 정보를 이용하여 도출된 해커 중심 분석 자료는 실제 해킹 사례의 정보들과 연계되고 분석되어 해커 혹은 해커 그룹을 특정 지을 수 있는 정보가 된다.

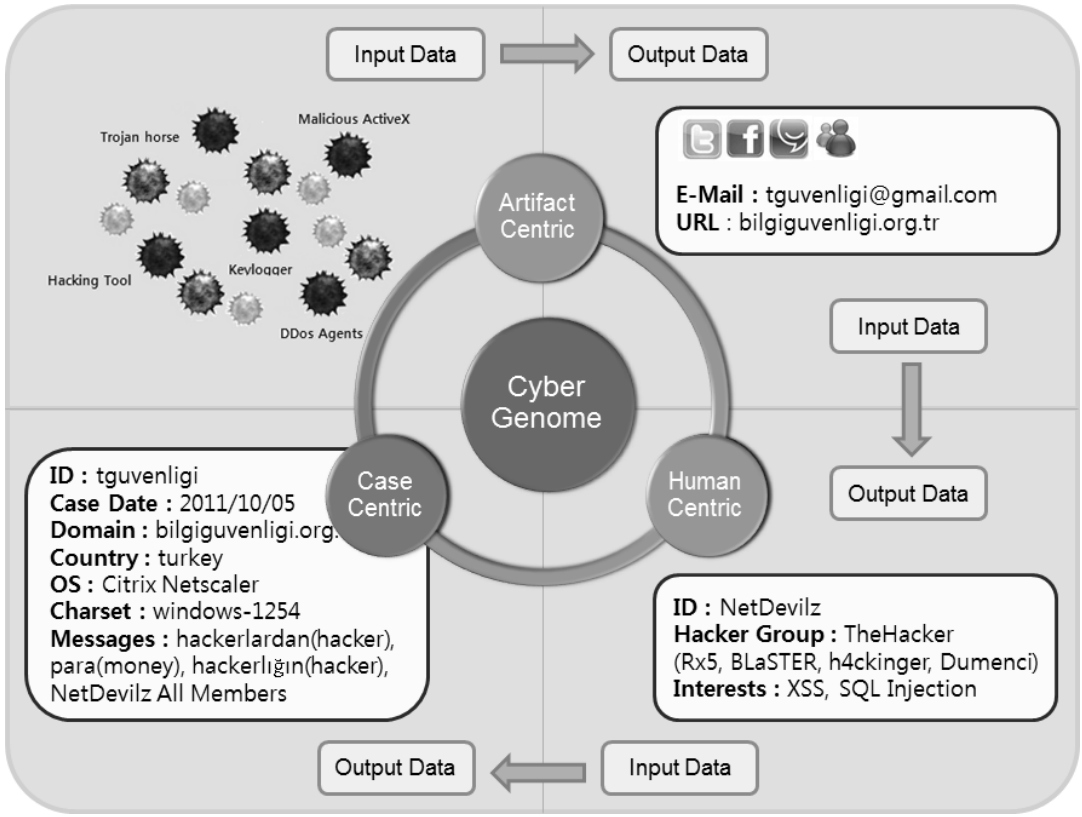

Fig.15. Cross-analysis system's Output

V. 결론

본 논문에서 제안된 알고리즘은 해킹 사고 사례에 대하여 유사도를 기반으로 하여 사례기반추론을 하는 첫 시스템으로써 의미를 갖는다. 해킹 사례에서 추출할 수 있는 유의미한 특성들은 CBR알고리즘을 적용하여 과거 사례와 어느 정도 유사한지 확인할 수 있다. 사이버게놈 프로젝트(KU Model)의 세 가지 모듈 중 인공물 중심 분석 시스템을 통해 악성코드와의 연관 정보를 구체화 할 수 있고, 해커 중심 분석 시스템을 통해서는 해커 활동 정보를 구체화 할 수 있다. zone-h.org에서 공유하는 공격 국가, 공격한 국가 웹 사이트의 시스템 정보뿐만 아니라 해커 정보의 정확성을 높이기 위해 폰트 정보, 인코딩 정보, 해커와 해커 그룹과의 관계 정보, 페이스북과 트위터 계정 정보 같은 휴먼 정보도 추가하여 분석하였다. 이는 해커들의 웹 해킹 시 남기게 되는 흔적과 습관 정보로써 사이버 공격의 의도가 돈을 위한 것인지, 아니면 개인정보 탈취 때문인지, 아니면 정치적, 종교적, 민족적 투쟁 때문인지를 파악할 수 있고, 그에 따른 대응책 마련에 도움을 준다. 본 논문에서 분석한 해킹 사례는 모든 해킹 공격 방법을 다룬 것은 아니며, 웹 해킹에 편중되어 있어 어느 정도의 한계는 존재한다. 결과의 정확도를 높이고자 하는 목적보다는 Security Intelligence 로서 디지털 포렌식 프로파일링과 사이버 수사 정보 차원에서 활용되는 목적이 더 크다고 할 수 있다. 추후 연구에서는 XSS(Cross Site Scripting), SQL Injection, Input Validation, Cook-ie stealing, Session id 해킹, Command injection과 같은 웹 해킹의 다양화된 공격방법에 대한 자료를 수집 분석하고, 웹 해킹 이외의 다른 공격 정보도 함께 수집 분석할 예정이다. 그리고 텍스트 마이닝(Text Mining)과 감정 탐지(Mood Detection) 방법을 통해 자동적으로 해커의 성향을 분류하는 연구도 진행할 예정이다. 이와 같은 자료 수집과 분석 작업은 사건에 연루된 해커 및 해커 그룹 수사 진행 시 IDSS(Investigation Decision Support System)으로써의 역할을 기대할 수 있으리라 본다.

References

[1] The civilian government military joint team, '3.20 Cyber Terror' mid-term report. http://www.msip.go.kr/www/brd/m_211/view.do?seq=28&srchFr=&srchTo=&srchWord=&srchTp=&multi_itm_seq=0&itm_seq_1=0&itm_seq_2=0&company_cd=&company_nm=&page=66

[2] DARPA(Defense Advanced Research Projects Agency). http://www.darpa.mil/Our_Work/I2O/Programs/Cyber_Defense_(Cyber_Genome).aspx

[3] M. Warren and S. Leitch, "Hacker Taggers: A new type of hackers," Information Systems Frontiers, pp. 425-431, Sep. 2010.

[4] H. Woo, Y. Kim and J. Dominick, "Hackers: Militants or Merry Pranksters? A Content Analysis of Defaced Web Pages," Media Psychology, vol. 6, no. 1, pp. 63-82, Feb. 2004.

[5] M. Milone, "Hacktivism: Securing the National Infrastructure," Knowledge, Technology & Policy, vol. 16, no. 1, pp. 75-103, Mar. 2003.

[6] S. Begum, M.U. Ahmed, P. Funk, Ning Xiong, and M. Folke, "Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 41, no. 4, pp. 421-434, Jul. 2011.

[7] Z. Yin, Y. Gao and B. Chen, "On Development of Supplementary Criminal Analysis System Based on CBR and Ontology," Computer Application and System Modeling (ICCASM), 2010 International Conference on, pp. V14-653 - V14-655, Oct. 2010.

[8] C.C. Chang and K.H. Hua, "Applying Case-Based Reasoning and Expert Systems to Coastal Patrol Crime Investigation in Taiwan," Intelligence and Security Informatics, pp. 161-170, Jun. 2008.

[9] A.J. Pinizzotto and N.J. Finkel, "Criminal Personality Profiling," Law and Human Behavior, vol. 14, no. 3, pp. 215-233, Jun. 1990.

[10] K.L. Kaufman, D.R. Hilliker and E. L. Daleiden, "Subgroup Differences in the Modus Operandi of Adolescent Sexual Offender," Child Maltreat, pp. 17-24, Feb. 1996.

[11] R.R. Hazelwooda and J.I. Warren, "Linkage analysis: modus operandi, ritual, and signature in serial sexual crime," Aggression and Violent Behavior, vol. 9, no. 3, pp. 307-318, May-Jun. 2004.

[12] C. Bennell and N.J. Jones, "Between a ROC and a hard place: a method for linking serial burglaries by modus operandi," Journal of Investigative Psychology and Offender Profiling, vol. 2, no. 1, pp. 23-41, Jan. 2005.

[13] B. Leclerc, E. Beauregard and J. Proulx, "Modus Operandi and Situational Aspects in Adolescent Sexual Offenses Against Children: A Further Examination," International Journal of Offender Therapy and Comparative Criminology, vol. 52, no. 1, pp. 46-61, Feb. 2008.

[14] H.K. Kim, K.H. Im and S.C. Park, "DSS for Computer Security Incident Response applying CBR and collaborative response," Expert Systems with Applications, vol. 37, no. 1, pp. 852-870, Jan. 2010.

[15] Character Encoding from Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Character_encoding

[16] Typeface from Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Typeface

〔17〕 Cyber war blooming between Korea and japan on March 1, 2010 from Wikipedia, the free encyclopedia. http://ko.wikipedia.org/wiki/2010%EB%85%84_%ED%95%9C%C2%B7%EC%9D%BC_%EC%82%BC%EC%9D%BC%EC%A0%88_%EC%82%AC%EC%9D%B4%EB%B2%84_%EA%B3%B5%EA%B2%A9_%EC%82%AC%EA%B1%B4
〔18〕 Hacking groups calling for 9/11 cyber attacks against Israel, U.S. http://www.jta.org/2013/09/10/news-opinion/united-states/groups-call-for-cyber-attacks-against-israel-u-s-on-9-11
〔19〕 Syria, Egypt crises spur escalation of ME cyber attacks. http://www.itp.net/594742-syria-egypt-crises-spur-escalation-of-me-cyber-attacks
〔20〕 Unknown group calling itself Whois Team 's html source, they attacked South Korea 's thriving Internet community. http://mlbpark.donga.com/mbs/articleV.php?mbsC=bullpen&mbsIdx=2106425
〔21〕 Turkish hacking group defaces UPS, TheRegister, Acer, Telegraph, Vodafone on zone-h.org. http://www.zone-h.org/news/id/4741
〔22〕 Windows-1524 from Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Windows-1254
〔23〕 ISO/IEC 8859-1 from Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Iso_8859-1
〔24〕 ICANN and IANA domains hijacked by Turkish on zone-h.org.crackershttp://www.zone-h.org/news/id/4695

〈저자소개〉

한 미 란 (Mee Lan Han) 학생회원
2002년 2월: 동덕여자대학교 컴퓨터공학 학사
2004년 5월~2012년 3월: NEXON 해외사업 개발본부
2012년 3월~현재: 고려대학교 정보보호대학원 석사과정
〈관심분야〉 온라인게임 보안, 네트워크 보안, 데이터 마이닝, 시각화, 빅데이터

사 진

김 덕 진 (Deok Jin Kim) 정회원
2004년 2월: 인하대학교 컴퓨터공학 학사
2006년 2월: 포항공과대학교 컴퓨터공학 석사
2006년 1월~2007년 9월: LG 전자 연구원
2007년 10월~2010년 1월: 한국전자통신연구원 연구원
2010년 2월~현재: 한국전자통신연구원 부설연구소 선임연구원
〈관심분야〉 네트워크 보안, 침입탐지시스템, 악성코드 동적분석

김 휘 강 (Huy Kang Kim) 종신회원
1998년 2월: KAIST 산업경영학과 학사
2000년 2월: KAIST 산업공학과 석사
2009년 2월: KAIST 산업및시스템공학과 박사
2004년 5월~2010년 2월: NC소프트 정보보안실장, Technical Director
2010년 3월~현재: 고려대학교 정보보호대학원 조교수
〈관심분야〉 온라인게임 보안, 네트워크 보안, 네트워크 포렌직, 침입탐지시스템, 봇넷탐지