JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 12, pp. 127-134, December 2020
https://doi.org/10.9708/jksci.2020.25.12.127

McDoT: Multi-Channel Domain Tracking Technology for Illegal Domains Collection

Ho-Mook Cho*, JeongYoung Lee**, JaeHoon Jang**, Sang-Yong Choi***

*Principal Researcher, Cyber Security Research Center, KAIST, Daejeon, Korea
**Senior Researcher, APEX ESC, Inchon, Korea
**CEO, APEX ESC, Inchon, Korea
***Assistant Professor, Dept. of Cyber Security, Yeungnam University College, Daegu, Korea

[Abstract]

Recently, Harmful sites, including pornographic videos, drugs, personal information and hacking tool distribution sites, have caused serious social problems. However, due to the nature of the Internet environment where anyone can use it freely, it is difficult to control the user effectively. And the site operator operates by changing the domain to bypass the blockage. Therefore, even once identified sites have low persistence. In this paper, we propose multi-channel domain tracking technology, a technique that can effectively track changes in the domain addresses of harmful sites, including the same or similar content, by tracking changes in these harmful sites. Proposed technology is a technology that can continuously track information in a domain using OSINT technology. We tested and verified that the proposed technology was effective for domain tracking with a 90.4% trace rate (sensing 66 changes out of 73 domains).

▸ **Key words**: Cyber investigation, Domain trace, Illegal Website, OSINT, Detection

[요 약]

음란 동영상, 마약, 개인정보, 해킹 도구 유포사이트 등을 포함하는 유해 사이트는 최근 사회적으로 심각한 문제를 초래하고 있다. 하지만 누구나 자유롭게 사용할 수 있는 인터넷환경의 특성상 접속자를 효과적으로 통제하기 어렵고, 사이트 운영자는 차단을 우회하기 위해 도메인을 변경하면서 운영한다. 따라서, 한번 확인된 사이트라 하더라도 그 지속성은 낮다. 본 논문에서는 이와 같은 유해 사이트의 변화를 추적하여 동일 또는 유사한 콘텐츠를 포함한 유해 사이트의 도메인 주소가 변경되는 것을 효과적으로 추적할 수 있는 기술인 다채널 도메인 추적기술을 제안한다. 제안하는 기술은 OSINT 기술을 이용하여 도메인의 정보를 지속적으로 추적할 수 있는 기술이다. 우리는 실험을 통해 90.4%의 추적률(실험대상 73개의 도메인 중 66개의 변경을 감지)로 제안한 기술이 도메인추적에 효과가 있음을 검증하였다.

▸ **주제어**: 사이버 수사, 도메인 추적, 유해 사이트, 악성코드, 오신트, 탐지

• First Author: Ho-Mook Cho, Corresponding Author: Sang-Yong Choi
*Ho-Mook Cho (chmook79@kaist.ac.kr), Cyber Security Research Center, KAIST
**JeongYoung Lee (mojomoth@apexesc.com), APEX ESC
**JaeHoon Jang (jh.jack@apexesc.com), APEX ESC
***Sang-Yong Choi (csyong95@gmail.com), Dept. of Cyber Security, Yeungnam University College
• Received: 2020. 11. 02, Revised: 2020. 11. 18, Accepted: 2020. 11. 18.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인터넷 및 ICT 기술의 발전에 따라 다양한 서비스가 출연하였으며, 언제 어디서든 모바일 및 컴퓨터를 통해 다양한 정보를 수집, 생산, 유통이 가능하게 되었다. 또한, 현대 사회는 웹사이트, 블로그, SNS 등 사이버상에서 지역, 인종, 신분을 초월한 사회적 관계를 쉽게 맺을 수 있다. 이와 같은 편리함으로 인해 전 세계 인터넷 웹사이트 수는 2020년 10월 현재 18억 개 이상으로 확인된다[1].

인터넷의 발전은 이와 같은 편리성을 증대시킨 반면 음란 동영상, 마약, 개인정보, 해킹 도구 거래 등 유해 사이트를 통한 불법 거래가 사회적으로 심각한 문제를 초래하고 있다. 하지만 접근 조치가 까다로운 인터넷의 특성으로 청소년도 쉽게 접근할 수 있어 문제의 심각성은 더욱 크다고 볼 수 있다. 통계자료에 따르면 국내도 유해 사이트 수가 지속적으로 늘어나고 있으며, '19년 신고 및 적발을 통해 차단된 국내외 유해 사이트는 18만 개 이상에 이른다고 보고되고 있다[2]. 물론 선진국을 포함하여 국제적으로 많은 국가에서 시행하고 있는 유해 사이트 차단 자체가 표현의 자유를 제약한다는 비판이 계속되고 있으며, 유해 사이트 차단과 차단을 회피하기 위한 SNS를 활용하거나, 비개방형 웹사이트를 사용, 지속적인 도메인 변경, 폐쇄 후 재구축, IP 변경 등 회피기술이 동시에 발전하고 있다[3]. 하지만, 지속적으로 증가하는 사이버 범죄로부터 안전한 인터넷환경을 만들기 위해서는 사회적 문제를 일으키는 유해 사이트를 지속적으로 추적하여 차단할 수 있는 기술이 필요하다. 확인된 유해 사이트에 대해서는 SNI차단[4] 등 기존의 차단기술을 활용하여 효과적으로 차단할 수 있지만, 지속적인 변경이 이루어지는 상황에서 유해 사이트의 변경을 효과적으로 추적하여 변경된 도메인을 확보하는 것이 중요하다.

본 논문에서는 지속적으로 변경되는 유해 사이트를 추적하기 위한 기술을 제안한다. 제안하는 기술은 OSINT Framework를 기반으로 하여 유해 사이트 정보를 수집하고, domain tools, Whois 등 공개된 정보 사이트와 연동을 통해 기존 등록자의 유사 사이트를 추적하여, 형태로 클러스터링을 통해 웹사이트를 범주화하는 방법으로 유해 사이트 도메인을 수집하고 추적하는 멀티채널 도메인 추적기술이다.

본 논문의 2장에서는 유해 사이트의 정의와 현황 및 문제점을 살펴보고, 국내외에서 연구되고 있는 유해도메인 판별 및 분류기술을 살펴본다. 3장에서는 제안하는 멀티채널 도메인 추적기술에 관해 설명하고, 4장에서는 실험을 통해 제안하는 추적기술의 도메인 추적에 대한 효과성을 검증하며 5장에서 결론을 맺는다.

II. Preliminaries

1. Related works

1.1 Definition and Problems of Malicious Sites

유해 사이트는 국내의 경우 방송통신심의위원회에서 「방송통신위원회의 설치 및 운영에 관한 법률」 제21조 제4호 또는 제25조 제1항에 따라 지정한 유해성이 짙은 특정 사이트를 말한다. 유해 사이트의 주 대상은 성인사이트나 지나치게 폭력적이거나 혐오적인 사이트, 청소년에게 해를 끼칠 수 있는 사이트 등을 포함하며, 이러한 측면에서 불법도박, 성매매, 음란정보, 국가안보 위해정보, 불법 식약품 관련 정보, 불법촬영물 등 디지털 성범죄정보 등을 모두 포함한다. 2019년 2월 기준 유해사이트로 판별되어 차단된 수는 전체 18만 8천여건이며, 이 중 성매매 · 음란정보 약 8만건(33.4%), 도박정보 약 6.3만건(26.6%), 불법식 · 의약품 정보 약 5만건(20.7%) 등이 포함되어 있다[2].

이와 같은 불법 유해 사이트는 해로운 정보가 통제되지 않고 순식간에 확산되는 특징으로 인해 유해정보에 노출된 사용자의 정서적, 행동적인 악영향을 준다는 것이고, 나아가 이러한 악영향이 사회 전반적으로 문제점을 발생, 증가, 확산시킨다는 것이다. 따라서 유해 사이트는 사이트에서 제공하는 콘텐츠의 종류와 관계없이 공통으로 문제점을 일으킨다.

1.2 Discrimination and Classification Technique of Malicious Domain

유해 사이트를 식별하고 분류하는 기술은 사이트 내 포함된 내용의 문맥을 분석하여 식별하는 기술과 사이트 간의 연결 관계를 분석하여 판별하는 기술, 트래픽에서 유해정보를 추출하는 기술 등 다양한 기술이 있다.

먼저 문맥을 분석하여 유해 사이트를 식별하는 기술로는 비속어 추출기반 웹사이트 접속 제한 시스템이 있다. 이 시스템은 설문 및 인터넷 검색을 통해 실제 사용되는 비속어 5542개를 수집하였고, w-shingling 알고리즘을 사용하여 비속어를 추출하고 사용빈도 및 비속어의 가중치에 따라 웹페이지의 위험도를 계산하여 유해 사이트를 식별한다[5]. 실시간 크롤링을 이용한 유해 사이트 판별 시스템에서는 유/무해 사이트 정보, 사이트 간의 연관 관계 정보 Database를 기준으로 1단계 판별한다. 2단계에서는

웹 크롤러를 사용해 해당 사이트의 텍스트 정보를 추출하고 후처리를 거쳐 keyword로 사용할 수 있는 단어 집합을 추출한다. 이후 판별 model의 input으로 사용된다. 3단계에서는 추출된 단어 집합으로 사전 수집 Dataset으로 생성된 판별모델에 의해 유해 여부를 판별한다[6]. 또 다른 방법으로는 기존의 유해 사이트 수집 방법의 문제를 해결하기 위해 실증 데이터를 기반으로 화이트리스트를 생성하고 생성 알고리즘을 모듈화로 구현하였다. 초기 화이트 데이터베이스에 의해 유해판정을 받은 웹사이트로 후보군을 만들고 각 후보의 자식 중 유해 사이트가 존재하는지 검색한다. 후보의 자식 중 유해 사이트가 존재한다면 아무런 작업 없이 다음 후보로 넘어가고, 존재하지 않는다면 후보를 화이트리스트로 판별 후 화이트 데이터베이스에 저장한다. 유해 데이터베이스에 속한 각 웹사이트의 상위 10개 단어를 TF-IDF 벡터화 모델을 통하여 추출하고 사전기반 검색을 통하여 결과가 참이면 유해로 분류하며, 데이터베이스를 확장한다[7]. 성인사이트를 분류하는 방법으로는 웹페이지에서 추출된 이미지들을 Open NSFW를 통해 성인물일 확률을 얻고 벡터화하여 SVM을 통해 해당 웹 페이지가 성인 콘텐츠를 포함하고 있는지를 분류하는 방법이 연구되었다. 이 연구는 실험 및 성능 평가를 Precision 및 Recall로서 진행하였으며, 이때 성인사이트에 대한 분류에 대해서는 88.66%, 84.85%가 일반 웹사이트에 대한 분류에 대해서는 92.86%, 94.77%의 성능을 보였다[8]. Z Liu[9]는 DNS 트래픽의 특성상 데이터가 어느 한 도메인에 몰리는 만큼 불균형을 이룰 수밖에 없어 이러한 불균형 트래픽에서 유해 도메인을 검출하는 방법론을 연구하였다. Xiang Tian[10]은 라지스케일 비디오 트래픽에서 불법 도메인을 검출하는 VegaStar 시스템을 제안하였다. 이 시스템은 VegaStar를 사용하여 비디오 트래픽에서 도메인 네임을 추출 후 분석, 불법 도메인을 검출하는 방식으로 5백만 개의 URL을 분석하였다.

Kyle Soska[11]는 악성 웹사이트들에 대한 단순 검출도 중요하지만, 만약 악성 웹사이트로 변질될 가능성이 있는 웹사이트들을 미리 솎아낼 수 있다면 예방 차원에서 큰 역할을 할 것이라는 점에 착안하여 다양한 데이터마이닝 기술들을 활용하여 미래에 악성 웹사이트로 변질될 가능성이 있는 웹사이트들을 검출하는 기법을 제안하였다. 이 연구에서는 1년간의 기간 동안 약 44만 개의 웹사이트를 분석하여, 개발된 검출기가 어느 정도 정확성을 가짐을 확인할 수 있었다.

이처럼 유해 사이트를 식별하고 분류하기 위한 연구는 다양하게 진행되었으나, 한번 확인된 유해 사이트가 변경되었을 경우 기존의 데이터와 연관분석을 통해 변화를 추적하는 기술은 아직 연구되지 않고 있다. 하지만 유해 사이트 제작자들은 유해 사이트 차단기술에 대응하기 위해 우회할 수 있는 다양한 기술을 적용하고 있으며, 이와 같은 기술에 효과적으로 대응하기 위한 연구가 필요하다.

1.3 Utilizing Technology for Domain Tracking

본 연구에서는 효과적인 유해도메인 추적을 위해 오픈소스를 이용하였다. 먼저 도메인 추적기를 설계하면서 방대한 수의 웹사이트 및 도메인을 효과적으로 처리하기 위해 클라우드 플랫폼을 활용하였다. 연구에서 활용한 클라우드 플랫폼은 구글 클라우드[12]로 구글의 데이터 센터 인프라를 기반으로 스토리지, 네트워킹, 빅데이터, 머신러닝 등의 서비스를 제공하는 글로벌 클라우드이며, 단순 웹사이트에서부터 복잡한 애플리케이션에 이르는 일련의 프로그램을 빌드하기 위한 환경을 제공한다. 또한, 웹사이트 내의 문자열을 분석하기 위해 Komoran[13], Noir[14] 등의 형태소 분석기를 사용하였다. Komoran은 Korean Morphological Analyzer의 약자로 Java로 구현한 한국어 형태소 분석기이며, 파이썬 환경에서 쉽게 사용이 가능하다. Nori는 루신 프로젝트에서 공식제공하는 한글 형태소 분석기로 일래스틱서치에서 공식적으로 배포되었다. 그리고 자료수집과 관리, 시각화를 위해 ELK[15]를 활용하였다. ELK의 구성 요소인 일래스틱서치(Elasticsearch)는 아파치 루씬을 기반으로 하는 검색엔진이며, 정형 및 비정형의 데이터를 위한 분산형 검색 및 분석 엔진을 지원하고 뛰어난 검색 능력과 대규모 분산 시스템을 구축할 수 있는 기능을 제공한다. 주요 특징으로는 오픈소스라는 점과 실시간 분석, 전문(Full Text)검색, Restful API 지원 등이 있다. ELK의 또 다른 구성 요소인 키바나(Kibana) 일래스틱서치를 위한 시각화 및 관리 도구로서 실시간 히스토그램, 선 그래프, 파이 차트, 지도 등을 제공하며 사용자가 자신의 데이터를 기반으로 사용자 정의한 동적 인포그래픽을 만들 수 있는 캔버스, 위치기반 정보 데이터를 시각화하기 위한 일라스틱 맵 같은 고급 애플리케이션을 지원한다. 개발 언어로는 구글에서 개발한 GO Language[16]를 사용하였다. GO는 2009년 구글에서 개발한 프로그래밍 언어로 가비지 컬렉션 기능이 있고, 병행성을 지원하는 컴파일 언어이다. 구문이 C와 비슷하지만 메모리 보안, 쓰레기 수집, 구조 타이핑, CSP 스타일 병행성을 제공한다.

III. The Proposed Scheme

1. System Architecture

도메인 추적 시스템의 전체적인 구조는 Fig. 1과 같다. 시스템은 GO 모듈과 일래스틱서치, 키바나, 구글 클라우드와 연결되어 있다. 제안하는 시스템 엔트리 포인트는 “DomainTracker”, “KeywordTracker” 모듈을 실행하며 구글 클라우드 플랫폼의 클라우드 스케줄러를 통해 시스템 실행을 제어하고 데이터는 일래스틱서치에 저장, 인터페이스는 키바나를 사용한다.

2. Module Structure

2.1 Domain Analyzer

도메인 분석기(Domain Analyzer)는 Fig. 2와 같이 Site Extractor, Site Classifier, Site Dissector 등 3가지 실행함수를 통해 구현된다. Site Extractor는 웹사이트에 접속하고 DOM(Document Object Model)을 추출, 링크와 텍스트를 분석한다. 분석한 내용을 일래스틱서치에 저장하고 자식 도메인을 생성하여 새로운 Seed Domain을 확보한다.

Site Classifier는 저장된 DOM에서 추출한 텍스트를 가져와서 구글 클라우드의 자연어 처리 모듈과 연동하여 텍스트를 분류한다. 이 과정에서 불법 사이트와 일반 사이트를 구분한다. 분류하는 방법은 먼저 확보된 텍스를 기반으로 분속 모듈에 라벨링 된 키워드를 매핑하여 TF-IDF[17]으로 Fig.3과 같이 키워드 분류점수를 산정한다. 이후 엘라스틱 서치의 Nori형태소 분석기로 키워드를 추출, 데이터 색인/역색인 기능으로 불법 사이트와 연관관계를 추출한다.

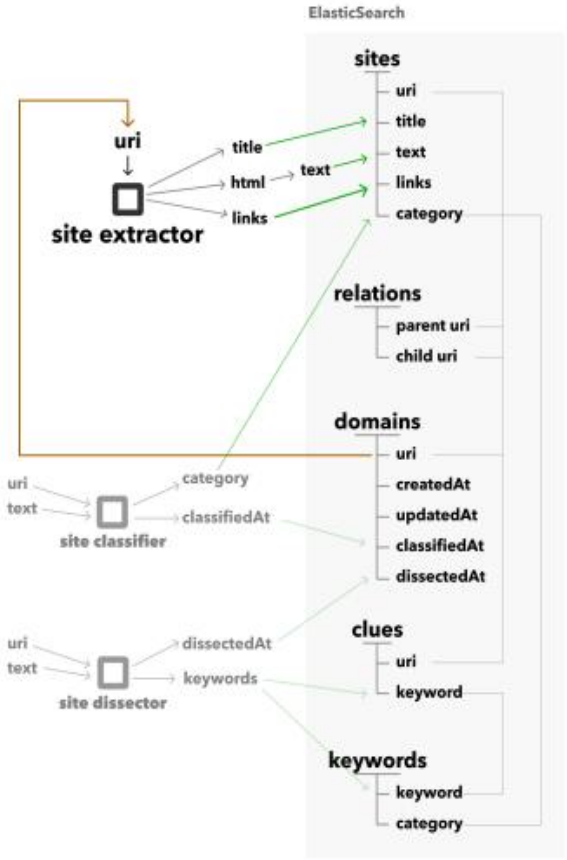


Fig. 2. Site Extractor

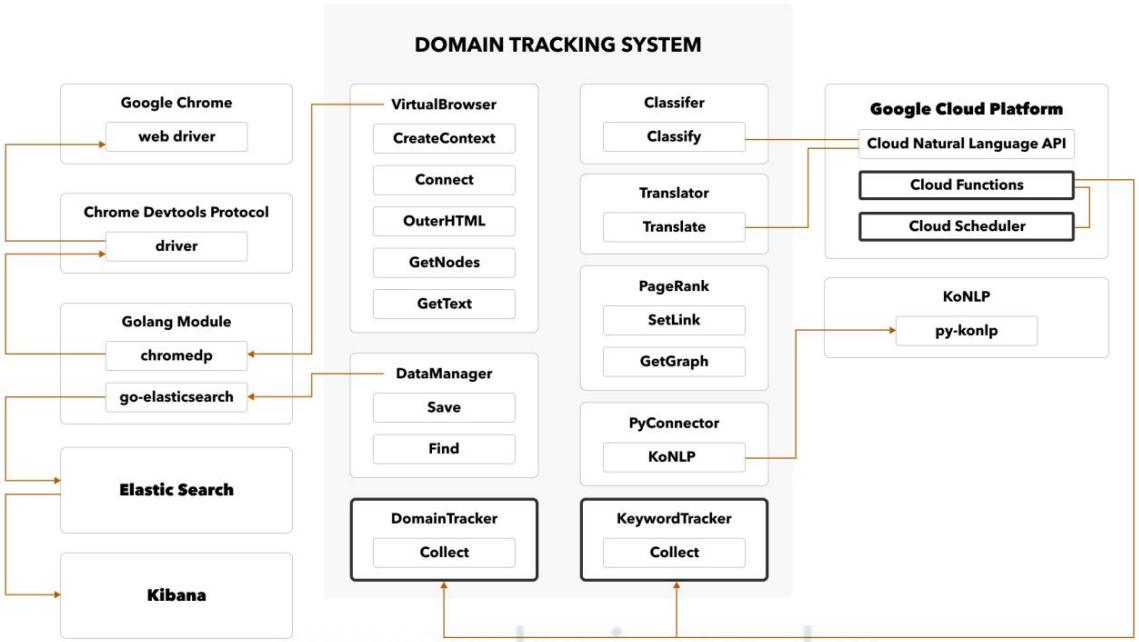


Fig. 1. System Architecture

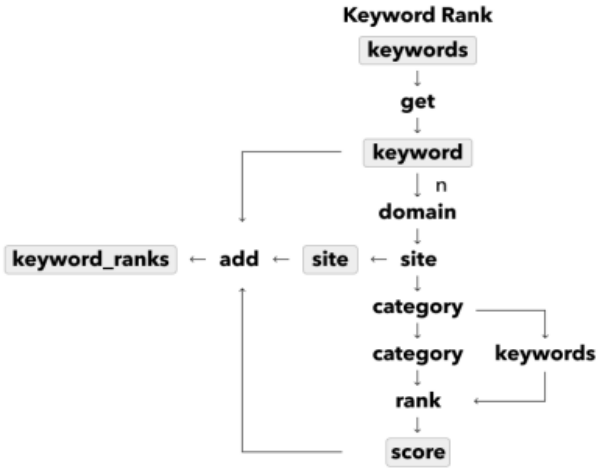


Fig. 3. Keyword Analyzer

Site Dissector는 저장된 DOM에서 추출한 텍스트를 구문 분석하여 명사를 추출, 키워드로 저장한다. 각 키워드는 앞서 분류된 카테고리와 매핑한다. 즉 본 시스템 아키텍쳐에서 도메인 추적을 위한 알고리즘으로는 TF-IDF와 PageRank알고리즘을 사용하며, 구글 클라우드 플랫폼 문서 분류기를 사용하여 웹사이트를 분류하고, 일래스틱서치 데이터베이스와 파이썬 형태소 분석기 등을 활용하여 문서 키워드를 추출하고, 불법 키워드 색인을 추출하게 된다.

2.2 Domain Tracker

도메인 추적기(Domain Tracker)에서 사용하는 도메인 추적기술은 크게 도메인 주소 패턴 추적, 사이버 범죄 키워드 추적, 레퍼런스 도메인 역추적, OSINT 도메인 정보 추적 등 4가지이다. 먼저 도메인 주소 패턴 추적은 도메인의 숫자 패턴을 예측하여 도메인 구조의 “Top-Level”을 추적하는 방법이다. 일반적으로 불법 도메인은 Fig. 4 와 같이 기존 도메인에서 도메인 내의 숫자 또는 Top-Level을 변경 후 재개방하는 경우가 많기 때문이다.

Fig. 4. Domain Address Tracking

사이버 범죄 키워드 추적은 Fig. 5와 같이 수집 및 분류 라벨링을 통해 범죄 키워드를 추출하고 추출한 키워드를 다양한 채널을 통해 수집된 웹사이트의 텍스트와 비교하는 방법이다.

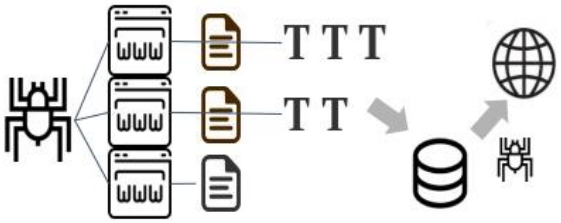


Fig. 5. Keyword Tracking

레퍼런스 도메인 역추적은 Fig. 6과 같이 불법 사이트의 운영 패턴을 분석하여 SNS 또는 불법 포털을 역추적하는 방법이다.

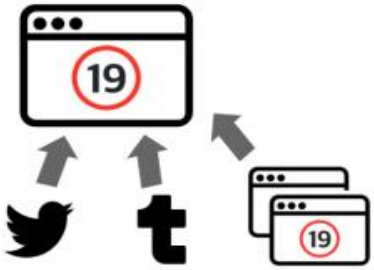


Fig. 6. Reference Domain Tracking

종합적으로는 이와 같은 방법은 OSINT의 범위 내에 포함되며, 이러한 정보를 활용하여 도메인 정보를 추적하고 프로파일링한다.

IV. Evaluation and Result

실험은 불법 도메인에 대한 식별이 가능한지 아닌지를 확인하는 도메인 분석과 분석된 불법 도메인의 변경 여부를 식별하여 추적하는 도메인 추적 성능에 대해 실험을 진행하였다. 실험 환경은 2.3GHz 8Core I9 CPU와 16Gbyte RAM이 설치된 노트북 컴퓨터를 활용하였다.

먼저 도메인 식별 여부를 확인한 결과 단순 키워드인 “Adult”로만 분류한 기준으로 30분 동안 100개의 Seed 도메인을 수집하였으며, 그중 성인사이트로 분류된 개수는 29개로 확인된다. 실험에 활용한 초기 시작 도메인은 https://darkgg6.com, https://podo11.com, https://avsee11.tv 등 3개이며, Fig. 7과 같이 이 3개의 부모 도메인에서 각각 5개, 32개, 3개 등 총 40개의 자식 도메인을 추출하였다.

수집된 도메인을 기반으로 하여 도메인 탐색, 웹사이트 분류 모듈을 실행한 결과 도메인 탐색 모듈은 각 웹사이트

Root Address	Leaf Address	Category	Count
seed domain	https://darkgg6.com	Adult	
seed domain	https://podo11.com	Adult	
seed domain	https://avsee11.tv	Adult	
			3
https://darkgg6.com	https://darkgg7.com		
https://darkgg6.com	http://wb-tt.com		
https://darkgg6.com	http://nb-we.com		
https://darkgg6.com	http://wjdwlgp2360.cafe24.com		
https://darkgg6.com	http://hkck24.top		
			5
https://podo11.com	https://use.fontawesome.com		
https://podo11.com	https://twitter.com		
https://podo11.com	https://xn--oj4boq74kw5c61n.com		
https://podo11.com	https://mexppp.com		
https://podo11.com	https://xn--9g3ba863cba.com		
https://podo11.com	https://jusomoya.com		
https://podo11.com	https://www.sakuraherb.com		
https://podo11.com	https://bbabam1.com		
https://podo11.com	https://mcjsport.com		
https://podo11.com	https://xn--939av9a78cnwvizi.com		
https://podo11.com	https://opqa50.com		
https://podo11.com	http://hangeul.naver.com		
			32
https://avsee11.tv	http://hol-01.com		
https://avsee11.tv	http://placa070.com		
https://avsee11.tv	http://mex-av3.com		
			3

Fig. 7. Domain Extractor

의 링크를 순서대로 추가하게 되어 단일 프로세스상에서 Fig. 8과 같이 총 1,717개의 새로운 도메인을 수집하였다.

수집한 도메인에 대해 약 30분간 불법 유해 사이트 여부를 분석한 결과 같이 제안하는 Fig. 9와 같이 시스템은 29개의 웹사이트를 불법 유해 사이트로 식별하였다. 실제 해당 사이트를 접속하여 분석한 결과 시스템이 식별한 불법 유해 사이트가 실제 불법 유해 사이트임을 확인할 수 있었다.

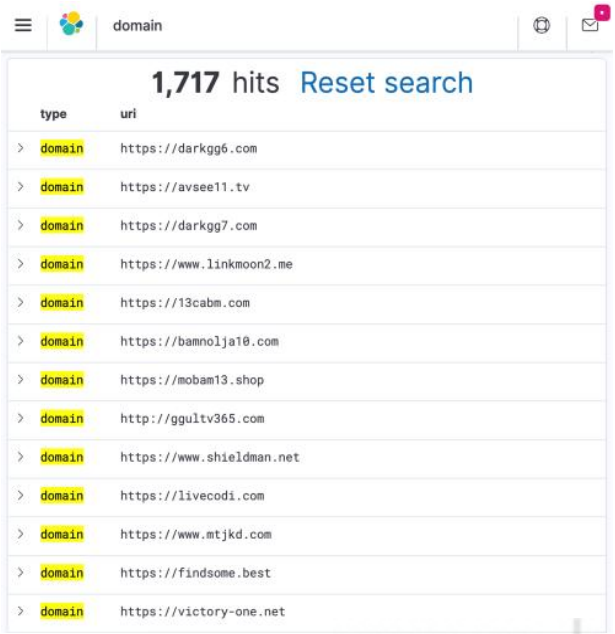


Fig. 8. Collected Domain

	Total Sites	100
	Adult Sites	29
Identifier	Web Address	Category
_SK0I3UBD75OP9phdV1y	http://smartfile.co.kr	Unknown
_yKgI3UBD75OP9phA1ck	http://hkck24.top	Unknown
-3gGJHUBX7W8q4Xez_Ji	https://www.sexnori8.me	Adult
-CK0I3UBD75OP9phPV2c	http://www.via999.me	Unknown
-CKII3UBD75OP9phO1y9	https://t.me	Unknown
-HgGJHUBX7W8q4XeefKe	https://torrenttip8.com	Unknown
0yKzI3UBD75OP9phoF1Z	http://wb-tt.com	Unknown
2CKzI3UBD75OP9phz13f	https://il119.com	Unknown
2iLYI3UBD75OP9phVV-M	https://guide-page.dothome.co.kr	Unknown
2yK0I3UBD75OP9phBF0Y	https://twitter.com	Unknown
3HgKJHUBX7W8q4XeHPP7	https://www.sorabam3.me	Unknown
3ngNJHUBX7W8q4XelvQ5	https://gop-1.com	Unknown
4HgDJHUBX7W8q4XeEfFw	https://yagong14.com	Adult
4yLOI3UBD75OP9ph-F7E	https://jusomoya.com	Adult
53gNJHUBX7W8q4Xeb_T1	http://solomv02.com	Unknown
53j-I3UBX7W8q4XegO8S	http://cp45.767hk.com	Unknown

Fig. 9. Domain Classification

이러한 방법으로 수집한 도메인 100개에 대해 도메인 추적 기능을 실험하였다. 실험결과 전체 Fig. 10과 같이 100개의 도메인 중 접속이 가능한 도메인은 42개였으며, 도메인 변경이 감지된 도메인은 47개, 변경이 감지되지 않은 도메인은 19개였다. 이 중 접속이 불가능하였던 58개 도메인이 변경된 것은 31개로 확인되었다. 또한, 애초에 접속이 가능했던 도메인의 변경은 16개로 확인되었다. 도메인 변경을 추적한 방법은 OSINT 방법의 하나인 구글 레퍼런스와 트위터 레퍼런스를 활용하였다. 변경된 도메인 47개 중 변경의 주요 요인은 도메인의 숫자 패턴을 변경한 방법이 31건으로 다수를 차지하고 있으며 전체주소를 변경하였거나, VPN을 사용하는 등 기타 방법이 확인되었다.

Total	100	Domain Change	47
Accessable	42		
Domain Address	Accessable	Domain Change	Change Content
https://ubang20.com/bbs/board.php?bo_table=east	O	X	
https://kongcafe5.site/bbs/board.php?bo_table=kr	O	X	
https://www.redgochu12.me/	X	https://www.redgochu15.xyz/	Number 15
https://hjtime13.com/bbs/board.php?bo_table=kor_mo	O	https://hjtime35.com/bbs/board.php?bo_table=kor_movie	Number 35
https://www.yazaral19.net/bbs/board.php?bo_table=kc	X		
https://www.yabuni7.men/bbs/board.php?bo_table=plar	X	https://yabuni01.com/	Number 01
https://yadong9.com/Article/index/id/249	X	https://yadong16.net/article/index/id/258	Number 16
https://sorabada3.com/bbs/new.php?gr_id=kor&view=v	X	https://sorabada25.com/bbs/new.php?gr_id=kor&view=w	Number 25
https://ktx.mango15.org/bbs/board.php?bo_table=010	X		
https://namja30.com/bbs/board.php?bo_table=kor_mo	O	https://bame34.net/	Full Address Change
http://sexmoa1.net/board/kr_video	O	X	
https://avpop50.com/bbs/livevideo2/%ED%95%9C%E/	O	https://avpop56.com/bbs/livevideo2	Number 56
https://avpop50.com/bbs/xjaqmffj/%ED%95%9C%EA%	O	https://avpop56.com/bbs/xjaqmffj	Number 56
https://ddaltime40.com/bbs/board.php?bo_table=kor	O	X	
https://ddaltime40.com/bbs/board.php?bo_table=best	O	X	
https://hujking11.com/bbs/board.php?bo_table=av01&	O	X	
https://ddrblr16.com/bbs/board.php?bo_table=kor	O	https://ddrblr23.com/bbs/board.php?bo_table=kor	Number 23
https://www.miso95.com/bbs/board.php?bo_table=m0	O	X	
https://mingkyaa28.com/bbs/board.php?bo_table=gall	X	https://mingkymm.com/pc_list.php?cate=mov_board16	Full Address Change
https://avmany4.com/bbs/board.php?bo_table=korean	X		
https://bananatv5.com/bbs/board.php?bo_table=data1	X	https://ddaljabi1.com/	Full Address Change
https://yasekma5.org/bbs/board.php?bo_table=korea	O	X	
https://dogjoa5.net/bbs/board.php?bo_table=avko	X	https://dogjoa33.net/bbs/board.php?bo_table=avko	Number 33
https://boomboom5.site/bbs/main.php?gid=movie	O	X	
https://ibiza4.site/bbs/main.php?gid=movie	X		
https://ibiza4.site/bbs/board.php?bo_table=hdyadong	X		

Fig. 10. Domain Tracking Result Sample

Table 1. Domain Tracking Result

	Test Result			Total
Total Test Domain	100			
Accessible	Change (a)	Unchanged (b)	Unknon (c)	42
	16	19	7	
Inaccessible	Change (d)	Unchanged or Unknown (e)		58
	31	27		
Tracking Rate	{(a+b+d) / (a+b+c+d)}×100 = 66/73×100 = 90.4%			

실험결과 Table. 1과 같이 100개의 실험 도메인 중 접속이 애초에 불가능하였던 58개의 도메인 중 변경 여부가 확인되지 않은 27개의 도메인을 제외한 73개의 도메인에 대해 변경 여부를 추적할 수 있었던 도메인의 수는 66개(변경 47개, 변경되지 않음 19개)로 변경 추적률은 수치적으로 90.4%로 확인되었다.

V. Conclusions

본 논문에서는 불법 사이트 차단을 우회하는 다양한 방법에 대해 대응하기 위해 불법 사이트를 추적하는 기술을 제안하였다. 최근 불법 사이트는 도메인 변경, 폐쇄 후 재개방 등 다양한 방법으로 운영을 지속하고 있다. 제안하는 기술은 도메인 주소 패턴 추적, 사이버 범죄 키워드 추적, 레퍼런스 도메인 역추적, OSINT 도메인 정보 추적 등 4가지 추적기술을 사용하여 불법 사이트를 식별하고 불법 사이트의 주소가 변경되는 것을 추적하여 사이버 범죄 예방 및 불법 사이트로 인한 사회적 피해를 예방하기 위한 기술이다. 또한, 제안한 방법의 효과를 검증한 결과 도메인 추적률은 90.4%를 보여 실제 변경된 도메인을 잘 추적하고 있음이 확인되었다. 본 논문에서 제안한 기술을 활용하면 국가기관 등에서 사이버 범죄를 예방하기 위한 불법 유해 사이트 차단의 효과가 증대될 것으로 기대된다. 향후 본 연구를 지속적으로 확장하여 더욱 다양한 형태의 불법 사이트 변경 기술을 분석, 대응할 수 있는 연구를 지속적으로 할 계획이다. 또한, 개발한 시스템을 할용하여 장기간 지속적인 실험을 통해 개발한 시스템의 한계점을 지속적으로 보완하고, 불법도메인 변경의 유형과 동향을 상시적으로 분석하는 플랫폼을 구축하여 보다 안전한 사이버 세상을 만들기 위해 노력할 것이다.

ACKNOWLEDGEMENT

This work was supported by the Supreme Prosecutors' Office ("A Study on Domain Tracking Technology for Cyber Crime Investigation").

REFERENCES

[1] Internetlivestats *https://www.internetlivestats.com/*

[2] Unlawful sites, enhanced blockage with blackouts, *http://news.knue.ac.kr/news/articleView.html?idxno=2197*

[3] 2019 Internet censorship controversy, *https://zdnet.co.kr/view/?no=20190214091551*

[4] Server Name Indication, https://namu.wiki/w/SNI

[5] Kim Jong Woo, Lee Sun Jeong, “Developing a Connection Restrictions Filtering System for Websites based on Swear Words Extraction”, Journal of KIISE, Vol. 46, No. 12, pp. 1272-1278, 2019, 10.5626/JOK.2019.46.12.1272

[6] SukYoon Kang, JooYoung Cho, GaHyun Joo, YountGu Lee, “Harmful Website Detection System Using Real-time Web Crawling”, Korea Computer Congress 2018, pp. 1904-1906, Jul. 2018.

[7] BoungJin Kim, SangJun Lee, “Improvement of Methods for Discriminating Harmful Web Sites by using Link Relations between Web Sites and Constructing Whitelist”, KIISE Transactions on Computing Practices, Vol. 25, No. 10, pp. 506-510, 2019, 10.5626/KTCP.2019.25.10.506

[8] KwangSu Shin, JinHa Song, HongHo Nang, “An Adult Web Site Classification Method using Analysis of Multiple Images in Web Page”, Korea Computer Congress 2017, pp. 868-870, Dec, 2017.

[9] LIU, Zhenyan, et al. An imbalanced malicious domains detection method based on passive dns traffic analysis. Security and Communication Networks, 2018, 2018.

[10] TIAN, Xiang, et al. VegaStar: An Illegal Domain Detection System on Large-Scale Video Traffic. In: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). IEEE, 2018. p. 783-789.

[11] SOSKA, Kyle; CHRISTIN, Nicolas. Automatically detecting vulnerable websites before they turn malicious. In: 23rd {USENIX} Security Symposium ({USENIX} Security 14). 2014. p. 625-640.

[12] Google Cloud Platform, *https://console.cloud.google.com/getting-started?hl=ko&pli=1*

[13] KOMORAN, *https://github.com/shineware/KOMORAN*

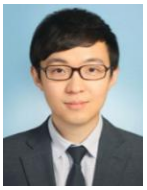
[14] Korean Analysis Plugin, *https://www.elastic.co/guide/en/elasticsearch/plugins/current/analysis-nori.html*
[15] ELK, *https://www.elastic.co/kr/*
[16] GO, *https://golang.org/*
[17] TF-IDF, *https://ko.wikipedia.org/wiki/Tf-idf*

Authors

Ho-Mook Cho received his M.S. degree in Information Security from Ajou University in 2006, and Ph.d degree in Interdisciplinary of Information Security from Chonnam National University in 2018,

Dr. Cho is a principal researcher in KAIST Cyber Security Research Center, KAIST, Daejeon, Korea. His research interests are in web security, malware analysis, XAI.

JeongYoung Lee recievied his B.S. degree in Computer Science from Kookmin University in 2012. JeongYoung Lee is a senior researcher at APEX ESC Inc and Zipida Inc now. research interests are big data collection & automation, AI analysis, and visualization systems.



JaeHoon Jang received his B.S. degree in Computer Science from Seokyeong University in 2013, and M.S. degree from Sungkyunkwan University EMBA in 2019. JaeHoon Jang established APEX ESC Inc. in 2016.

He is the founder and current CEO of Zipida Inc. from 2019. He is interested in Business Development, AI, Big Data Visualization, and Information Security.

Sang-Yong Choi received his B.S. degree in Mathematics and M.S. degree in Computer Science, both from Hannam University in 2000 and 2003, and Ph.d degree in Interdisciplinary of Information Security from Chonnam National University in 2014, Dr. Choi is a assistant professor at the Dept. of Cyber Security in Yeungnam University College, Daegu, Korea. His research interests are in web security, network security and cloud computing security.