머신러닝 기반 악성 도메인 분석을 위한 핵심 feature 추출 연구

장진호, 임채현, 이태진
호서대학교 컴퓨터공학과
ghdic77@gmail.com, chq112692@gmail.com, kinjecs0@gmail.com

Core Feature Extraction Study for Malicious Domain Analysis Based on Machine Learning

Jang Jin Ho, Lim Chae Hyun, Lee Tae Jin
Hoseo Univ. Computer Engineering

요 약

머신러닝 기술의 발전에 따라 기존에 식별하기 어려웠던 일반 URL과 악성 URL을 구분하려는 시도가 증가하고 있다. 본 논문은 URL의 어휘적 특징을 사용하여 비교적 높은 정확도를 보여주는 모델인 RF(Random Forest), GBT(Gradient Boosting Tree)을 기반으로 어떠한 특징이 악성 URL을 판별하는지 중요한지 판별하는 것을 목표로 한다. 실험결과 Length-based한 특징이 결과를 예측하는데 높은 영향을 끼치는 것을 확인 하였다.

Ⅰ. 서 론

현재 인터넷이 직면한 주요 과제 중 하나는 악성 URL을 통하여 개인정보 탈취, 악성코드 유포 등 사이버 위협이 지속적으로 발생하고 있다. 기존에는 Whitelist/Blacklist, Rule-Based과 같은 방법으로 악성 URL을 탐지하였지만, 이러한 방식은 새로운 유형의 악성 URL를 탐지하는 것에 한계가 있다. 이러한 한계를 극복하기 위해 머신러닝 기법을 이용해 악성 URL을 탐지하는 연구가 진행되고 있다.

악성 URL은 일반 URL과 유사하기 때문에 식별하기 어렵다. 하지만 최근 머신러닝 기술의 발전으로 악성 URL과 일반 URL을 구별하기 위한 시도가 시행되고 있다. 머신러닝 모델이 악성 URL과 일반 URL를 구분하기 위해서는 문자열인 URL의 특징이 추출 되어야 하는데, URL의 다양한 특징을 추출하여 학습을 진행한다.

본 연구에서는 URL의 특징을 추출해 내어 어떠한 매개변수가 머신러닝 모델에 중요한지 중점을 둔다. URL특징의 중요도를 확인하기 위해, 악성 URL 탐지에 높은 정확도를 보여주는 분류 모델인 랜덤 포레스트(RF)과 GBT을 사용하여 비교 분석하였다.

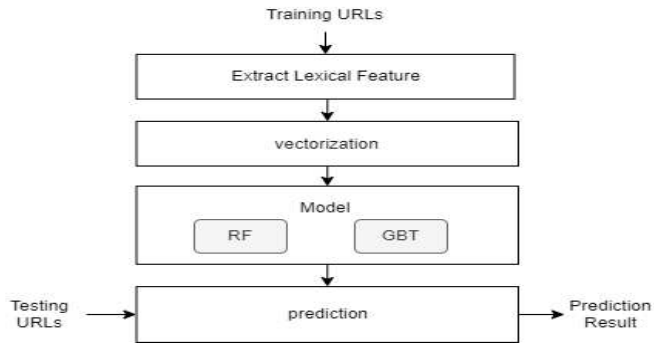
Ⅱ. 관련연구

악성 URL을 탐지하려는 시도는 인터넷이 본격적으로 활성화 되며 지속되어져 왔다. 대표적인 기존 탐지 기법으로 Whitelist./Blakclist, Rule-based, Search Volume이 있다. Whitelist/Blacklist 기반 악성 URL 탐지 방식은 리스트 생성자가 미리 안전하거나 위협이 되는것에 대하여 등록하여 차단하는 것으로 신규위협에 취약하며, 리스트를 생성하는 인력에 의하여 탐지 기준이 정해지기 때문에 정확성의 편차가 떨어진다. Rule-based기반 악성 URL 탐지 방식은 공격 시도 과정에서의 일정한 규칙을 찾아내고, 공격시도를 탐지하여 차단하는 기술로 기존 공격자의 패턴이 있는 공격에 대하여 탐지는 가능하지만, 해당 규칙을 벗어난 변종 공격은 탐지하지 못한다. Search Volume기반 악성 URL탐지 방식은 구글, 네이버 등 포털사이트의 검색엔진을 이용하여 해당 사이트의 URL을 검색하여 포털에서의 URL 검색 양을 확인한다. 비정상적인 URL은 실제로 서비스되지 않고, 검색하는 사람이 적은 특징을 노린 기법이다. 하지만 이 기법의 경우 신규 사이트나 비인기 사이트에 대한 탐지를 하는데 좋지 못한 결과를 보여준다.

URL이라는 정보가 주어졌을 때 머신러닝 모델 학습에 사용 할 수 있는 특징들은 크게 세가지로 나눌 수 있다. 먼저 어휘적 특징으로 특수 문자나 특정 문자, Domain, TLD, HTTPs 등과 같은 URL의 특정 부분에 대한 길이, 존재 여부를 대상으로 할 수 있다. 두 번째로 호스트기반 특징으로 해당 URL에 request하여 response로 받은 HTML, JS 소스코드에서의 리다이렉트 존재여부, 팝업 윈도우 존재여부, <a>태그의 개수 등과 같은 것을 대상으로 할 수 있다. 세 번째로 도메인정보기반 특징으로 Whois에 등록되어있는 도메인 정보인 서비스 기간, 갱신 날짜, 호스팅 업체 등의 정보를 수집 할 수 있으며 검색엔진을 통해 해당 도메인의 검색 볼륨 등을 수집하여 대상으로 할 수 있다. 이러한 URL의 특징 중 어휘적 특징에서 어떠한 특징이 모델 학습에 도움이 되는지 본 연구에서 탐구한다.

Ⅲ. 제안모델

본 연구에서 제안하는 머신러닝 기반 악성 URL 감지 시스템은 URL의 어휘적 특징을 추출하여 RF와 GBT모델을 통해 구성되었다. 제안되는 시스템은 URL의 특징을 추출하는 모듈, 추출한 데이터를 벡터화 하는 모듈, 모델을 생성하는 모듈, 악성 URL을 예측하는 모듈로 구성 된다. 모델을 생성하기 위한 Training Set과 성능측정을 위한 Testing Set은 각각 80%, 20% 비율로 구성하였다.



[그림 1] Malicious URL prediction system

URL만으로 악성 URL를 판별하는 것을 목표로 하여 URL의 어휘적 특징을 사용하였는데, [표 2]에서 길이기반, 개수기반, 존재기반으로 분류되어 연구에 사용한 20개의 어휘적 특징을 볼 수 있다. Training Set은 URL의 어휘적 특징으로 추출하고 벡터화 되어 Random Forest Tree(RF)와 Gradient Boosted Tree(GBT) 모델을 통해 학습하게 된다. 학습이 끝난 모델에 Testing Set도 어휘적 특징으로 추출하고 벡터화 하여 결과를 예측하게 된다.

IV. 시험결과

4.1 데이터셋

본 연구에서는 Kaggle[1]에서 제공하는 URL 데이터셋을 사용하였다. 데이터셋은 [표 1]과 같이 4개의 라벨로 구성 되어 있다.

benign	defacement	phishing	malware	Total
428,103	96,457	94,111	32,520	651,191

[표 1] 데이터셋 정보

4.2 악성 URL 탐지 결과

Category	Lexical features
Length-based	URL, hostName, Path, Query 길이
Count-based	".", "-", "@", "_", "%", "&", "#", 숫자 개수, hostName의 "-" 개수, subdomain Level, path Level, Query 개수
Existence-based	"~"가 존재하는가, https인가, ipaddress형태인가, "//"가 존재하는가

[표 2] 연구에 사용한 Lexical feature Category

URL의 어휘적 특징을 추출한 결과를 3가지 카테고리로 나눌 수 있다. URL길이와 같은 길이기반 특징, URL에 포함된 특수문자의 개수와 같은 개수기반 특징, URL이 https로 구성되어 있는가와 같은 bool형태의 결과를 반환하는 존재 기반 특징이다.

머신러닝 모델을 이용한 실험 결과 RF모델은 90.71%의 정확도를 보여주었으며, GBT모델은 87.63%의 정확도를 보였다. recall 측정 결과 RF모델은 91%, GBT모델은 87%을 보여주며 RF가 더 좋은 성능을 보였다.

Feature	Random Forest	Gradient Boost
hostNameLength	32.8	29.0
pathLevel	7.9	1.1
charNum	7.1	1.7
urlLength	5.8	6.0
pathLength	4.8	2.3
numDash	3.8	0.8
subdomainLevel	3.3	2.4
dotNum	1.9	0.6
isHttps	1.4	1.8
queryLength	1.4	0.4

[표 3] Feature별 중요도 (TOP10)

[표 3]에서 두 모델 중 중요도가 높았던 Feature 상위 10개를 뽑았다. Length-based인 hostNameLength, urlLength와 같은 URL의 특정 구간의 길이가 높은 중요도를 나타내는 것을 알 수 있다. Count-basded인 pathLevel, charNum 등도 준수한 중요도를 나타내는걸 볼 수 있다. bool 타입인 Existence-based는 numeric타입인 다른 두 카테고리보다 현저하게 낮은 중요도를 나타냈다.

V. 결론

웹 시장이 커짐에 따라 악성 URL을 악용한 개인정보 탈취, 악성코드 유포 등 사이버 위협이 지속적으로 발생하고 있다. 이러한 사이버 위협에 대하여 악성 URL을 구분 할 수 있는 정보보안 시스템이 요구되어 지고 있다. 최근 활발히 연구가 진행되는 머신러닝 기반 악성 URL탐지 기법에서 본 연구에서는 악성 URL 예측을 위해 특징 추출이 용이한 어휘적 특징 중 어떠한 부류의 특징이 머신러닝 모델에 적합하고 중요도가 높은지 연구하였다.

기존 연구를 기반으로 악성 URL의 특징 중 머신러닝 모델의 중요도가 높은 특징이 될 수 있는 추가적인 특징을 확보하고, 예측 정확도와 처리속도를 개선 할 수 있게 하고자 한다.

ACKNOWLEDGMENT

본 연구는 2022년 과학기술정보통신부와 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2019-0-01834)

참 고 문 헌

[1] Malicious URLs dataset https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset

[2] Chen Hajaj, Nitay Hason, Amit Dvir "Less Is More. Robust and Novel Features for MaliciousDomain Detection" pp.2-8, 2022

[3] Rakesh Verma, Avisha Das "What's in a URL Fast Feature Extraction and MaliciousURL Detection" pp.2-6, 2017

[4] Cho Do Xuan, Hoa Dinh Nguyen "Malicious URL Detection based on Machine Learning" pp.150-153, 2020