악성코드 행위기반 유사도 측정 기법 연구

A Study on the Similarity Measurement Method using Malware Behavior

저자
(Authors)

강홍구, 김경한, 유대훈, 최보민, 박준형
Kang Hong-Koo, Kim Kyung-Han, Yoo Dae-Hoon, Choi Bo-Min, Park Jun-Hyung

출처
(Source)

한국통신학회 학술대회논문집 , 2017.11, 697-698 (2 pages)
Proceedings of Symposium of the Korean Institute of communications and Information Sciences , 2017.11, 697-698 (2 pages)

발행처
(Publisher)

한국통신학회
Korea Institute Of Communication Sciences

URL

http://www.dbpia.co.kr/Article/NODE07284938

APA Style

강홍구, 김경한, 유대훈, 최보민, 박준형 (2017). 악성코드 행위기반 유사도 측정 기법 연구. 한국통신학회 학술대회논문집, 697-698.

악성코드 행위기반 유사도 측정 기법 연구

강홍구, 김경한, 유대훈, 최보민, 박준형
한국인터넷진흥원
{redball, kookie, ydh83, bmchoi, junpark}@kisa.or.kr

A Study on the Similarity Measurement Method using Malware Behavior

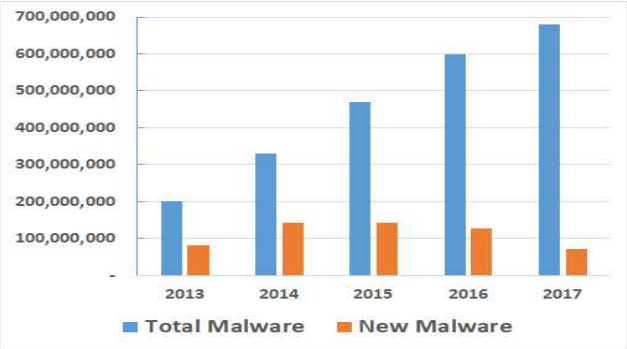
Kang Hong-Koo, Kim Kyung-Han, Yoo Dae-Hoon, Choi Bo-Min, Park Jun-Hyung
Korea Internet & Security Agency

요 약

최근 유사/변종 악성코드가 급증하고 있는 환경에서 악성코드 공격의 신속한 대응을 위한 악성코드 분류 기술이 요구되고 있다. 이러한 악성코드 분류를 위한 다양한 유사도 측정 기법에 대한 연구가 활발히 이루어지고 있다. 본 논문에서는 기존의 유사도 측정 기법을 소개하고 단점을 보완하여 악성코드 행위정보를 기반으로 유사도를 측정할 수 있는 기법을 제안하였다. 제안하는 기법은 악성코드가 실행되는 과정에서 호출되는 API 시퀀스를 벡터화하고, 공간상의 벡터간 거리와 각도를 이용하여 부채꼴 면적을 유사도 측정에 사용한다. 실험을 통해 기존 기법과의 유사도 측정결과를 비교하고 제안하는 기법이 악성코드 분류에 효용성이 있음을 증명하였다.

Ⅰ. 서론

대표적인 보안제품 테스트 기관인 AV-TEST에 따르면, 2016년에 약 6억 개의 악성코드가 발생하였고, 2017년 9월에만 약 6.8억 개의 악성코드가 발생한 것으로 보고되고 있다. 반면에 신종 악성코드 발생량은 2015년에 약 1.4억 개에서 2016년에 약 1.25억 개로 줄어들었으며, 2017년 9월에는 약 7천개로 감소폭이 커지고 있다.[1]



(그림 1) 신종 악성코드 발생량(AV-TEST, 2017.9월)

신종 악성코드 발생량이 줄어들고 있으나 여전히 악성코드 발생량이 급증하는 원인으로 전문가들은 변종 악성코드의 증가에 주목하고 있다. 공격자는 악성코드가 보안장비에서 탐지되지 않도록 기존 악성코드를 변형한 변종 악성코드를 많이 이용한다. 이는 악성코드를 새로이 제작하는 것보다 변종 악성코드를 제작하는 것이 비용대비 유리하기 때문으로 추정된다. 또한, 악성코드를 자동으로 제작하는 도구가 등장하면서 전문성 있는 해커가 아니더라도 손쉽게 다양한 악성코드를 제작할 수 있는 점도 변종 악성코드 증가의 주요 원인으로 꼽히고 있다.[2]

이렇듯 유사/변종 악성코드가 급증하고 있는 환경에서 악성코드 공격의 신속한 대응을 위한 악성코드 분류 기술이 요구되고 있다. 이러한 악성코드 분류를 위한 다양한 유사도 측정 기법에 대한 연구가 활발히 이루어지고 있다. [2]에서는 악성코드가 호출하는 API에 대해 2-gram 시퀀스와 출현빈도를 기반으로 벡터를 생성하고 코사인(Cosine) 유사도 분석을 통해 유사도를 측정하였다. [3]에서는 악성코드의 정적 분석을 기반으로 함수 리스트 호출빈도에 대한 유사도 측정에 코사인 유사도 기법을 사용하였다. [4]에서는 코사인 유사도 기법의 한계를 제시하고 이를 보완하기 위해 공간상의 거리차이와 면적을 이용한 유사도 측정 기법(TS-SS)을 제안하였다.

본 논문에서는 기존의 유사도 측정 기법을 소개하고 단점을 보완하여 악성코드 행위정보를 기반으로 유사도를 측정할 수 있는 기법을 제안하였다. 제안하는 기법은 악성코드가 실행되는 과정에서 호출되는 API를 추출하여 2-gram 시퀀스와 호출빈도를 벡터로 생성하고, 공간상의 벡터간 거리와 각도를 이용하여 부채꼴 면적을 유사도 측정에 사용한다. 실험을 통해 기존 기법과의 유사도 측정결과를 비교하고 제안하는 기법이 악성코드 분류에 효용성이 있음을 증명하였다.

Ⅱ. 악성코드 행위기반 유사도 측정 기법

악성코드 간 유사도를 측정하는 기법으로 코사인 유사도가 있다. 악성코드에서 추출된 정보를 벡터로 생성했을 때, 코사인 유사도는 벡터 간의 각도를 측정하는 기법이다. 코사인 유사도는 (수식 1)과 같다. 코사인 유사도는 벡터의 개수 차이가 큰 경우에도 0과 1 사이의 값으로 유사도를 측정할 수 있는 편리성이 있다.

$$\text{코사인 유사도}(A,B) = \cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|}$$

(수식 1) 코사인 유사도 측정 수식

그러나 코사인 유사도 측정 기법은 두 개의 벡터 간의 각도(θ) 차이에 초점을 두고 있으나 벡터 간 거리(ED: Euclidean Distance) 차이를 적절히 반영하지 못하는 한계를 가지고 있다. 예를 들어, 두 개 벡터 간의 각도 차이가 동일하더라도 거리 차이가 큰 경우나 벡터 간 거리 차이가 동일하더라도 각도 차이에 따라 유사도가 크게 달라진다. 이러한 한계를 보완하

기 위해 공간상의 벡터 간 거리와 면적(Area)을 이용한 유사도 측정 기법(TS-SS)이 제안되었다.[4] TS-SS 유사도 측정 기법은 두 벡터 간 각도와 거리로 만들어지는 삼각형 면적(TS)과 부채꼴 면적(SS)을 이용하여 유사도를 측정한다. TS-SS 유사도는 (수식 2)와 같다.

$$\text{if } ED(A,B) = \sqrt{\sum_{n=1}^{k}(A(n)-B(n))^2} \text{ and}$$

$$MD(A,B) = \left| \sqrt{\sum_{n=1}^{k}A_n^2} - \sqrt{\sum_{n=1}^{k}B_n^2} \right| \ then$$

$$TS-SS(A,B) = \frac{|A| \cdot |B| \cdot \sin(\theta') \cdot \theta' \cdot \pi \cdot (ED(A,B)+MD(A,B))^2}{720}$$

(수식 2) TS-SS 유사도 측정 수식

그러나 TS-SS는 벡터 간 각도가 0인 경우에는 면적 측정이 불가능하여 임의의 각도를 넣어야 하는 보정이 필요하고, 공간상의 삼각형과 부채꼴 면적의 곱으로 유사도를 측정하므로 유사도 값이 차이가 크게 나타나며 계산식이 다소 복잡하다.

본 논문에서는 악성코드가 실행되는 과정에서 호출되는 API를 추출하여 2-gram 시퀀스와 호출빈도를 벡터로 생성하고, 공간상의 벡터 간 거리와 각도를 이용하여 부채꼴 면적을 유사도 측정에 사용한다. 제안하는 유사도 측정 수식은 (수식 3)과 같다. 제안하는 유사도 측정 기법은 벡터 간의 각도 차이가 0인 경우에는 거리 차이로만 유사도를 계산하므로 벡터 간 각도를 보정할 필요가 없고, 그 이외에는 거리를 반지름으로 하는 부채꼴 면적만을 이용하므로 각도에 따른 유사도를 측정할 수 있다.

$$\text{if } \cos(\theta) = 0 \ then$$

$$\text{제안 유사도}(A,B) = ED(A,B) = \sqrt{\sum_{n=1}^{k}(A(n)-B(n))^2}$$

$$else$$

$$\text{제안 유사도}(A,B) = \frac{\theta' \cdot \pi \cdot ED(A,B)^2}{360}$$

(수식 3) 제안하는 유사도 측정 수식

Ⅲ. 실험결과

실험에서는 기존의 유사도 측정 기법과 제안하는 기법의 유사도 측정 결과를 비교하였다. 실험에서는 실제로 유포되었던 악성코드 샘플 중 백신 진단명을 갖는 샘플로 구성하였고, 객관성을 높이기 위해 국내외 주요 백신 3곳(V3, Kaspersky, Mcafee)의 진단명이 동일한 악성코드 샘플 10개를 이용하였다. (표 1)은 해당 악성코드 샘플을 보여준다.

(표 1) 동일 백신 진단명 기준 악성코드 샘플

악성코드 샘플	백신 진단명 (V3, Kaspersky, Mcafee)
1~4	PUP/Win32.BrowseFox, AdWare.Win32.BrowseFox.dild, BrowseFox
5	PUP/Win32.DownloadAdmin, Trojan.Gen.2, PUP-XAR-QT
6~10	PUP/Win32.Agent, AdWare.Win32.Generic, Artemis

실험결과, 두 벡터 간 거리와 각도의 차이가 거의 없는 경우에는 코사인, TS-SS, 제안기법 모두 유사도가 동일하게 측정되었다. 그러나 벡터 간 각도가 유사하지만 거리 차이가 다소 있는 경우에는 TS-SS와 제안기법이 코사인 유사도 보다 정확도가 높은 것으로 나타났다. 또한, 벡터 간 각도와 거리 차이가 크지 않은 경우에 TS-SS는 유사도 값의 차이가 큰 반면에 제안기법은 유사도 값의 차이가 크게 나타나지 않았다.

(표 2) 악성코드 유사도 측정 기법 비교결과

비교대상	벡터간 거리(ED)	벡터간 각도(θ)	유사도 측정 기법		
			코사인	TS-SS	제안기법
1-2	0	0	1	0	0
1-3	0	0	1	0	0
1-4	0	0	1	0	0
1-5	14.07	0.21	0.98	7392.68	1.21
1-6	16.37	0.22	0.97	10783.35	1.79
1-7	18.22	0.28	0.96	12998.67	2.64
1-8	18.76	0.29	0.96	13704.54	2.91
1-9	19.85	0.32	0.95	15232.84	3.50
1-10	98.30	0.18	0.98	1246035.87	61.11

Ⅳ. 결론 및 향후연구

최근 급증하고 있는 악성코드 공격을 신속하게 대응하기 위해 유사/변종 악성코드를 자동으로 분류하는 기술이 요구되고 있으며, 이를 위한 다양한 유사도 측정 기법에 대한 연구가 활발하다. 본 논문에서는 기존의 유사도 측정 기법을 소개하고 단점을 보완하여 악성코드 행위정보를 기반으로 유사도를 측정할 수 있는 기법을 제안하였다.

제안하는 기법은 악성코드가 실행되는 과정에서 호출되는 API 시퀀스를 벡터화하고, 공간상의 벡터간 거리와 각도를 이용하여 부채꼴 면적을 유사도 측정에 사용하였다. 실험을 통해 제안하는 악성코드 유사도 측정 기법이 악성코드 분류에 효용성이 있음을 증명하였다. 향후 연구에서는 다양한 악성코드 정보를 대상으로 주요 Featrue를 선별하고, 유사도 측정 알고지름을 개선할 예정이다.

ACKNOWLEDGMENT

이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2016-0-00081, 악성코드 全 생명주기 통합 프로파일링 및 공격그룹 식별 기술 개발)

참 고 문 헌

[1] AV-TEST, https://www.av-test.org/en/statistics/malware/

[2] 강홍구, 유대훈, 최보민, “행위기반 악성코드 프로파일링 시스템 프로토타입,” 한국정보처리학회 춘계학술대회, 2017

[3] Karnik, A., Goswami, S., and Guha, R. “Detecting Obfuscated Viruses Using Cosine Similarity Analysis,” AMS, 2007.

[4] Arash, H., and Michael, J. D. "A Hybrid Geometric Approach for Measuring Similarity Level Among Documents and Document Clustering," IEEE BigDataService, 2017.