https://doi.org/10.33778/kcsa.2022.22.2.009

공격키워드 사전 및 TF-IDF를 적용한 침입탐지 정탐률 향상 연구

김 종 관*, 김 명 수**

요 약

최근, 디지털전환의 확대로 사이버공격의 위협에 더욱 더 노출되고 있으며, 각 기관 및 기업은 공격이 유입되는 것을 막기 위해 시그니처 기반의 침입차단시스템을 네트워크 가장 앞단에 운영중에 있다. 그러나, 관련된 ICT시스템에 적절한 서비스를 제공하기 위해 엄격한 차단규칙을 적용할 수 없어 많은 오이벤트가 발생되고, 운영효율이 저하되고 있다. 따라서, 공격탐지 정확도 향상을 위하여 인공지능을 이용한 많은 연구과제가 수행되고 있다. 대부분의 논문은 정해진 연구용 데이터셋을 이용하여 수행하였지만, 실제 네트워크에서는 연구용 학습데이터셋과는 다른 로그를 이용해야만 하기 때문에 실제 시스템에서는 사용사례는 많지 않다. 본 논문에서는 실제 시스템에서 수집한 보안이벤트 로그에 대하여 주요 공격키워드를 분류하고, 주요 키워드별로 가중치를 부과, TF-IDF를 이용하여 유사도 검사를 수행후 실제 공격여부를 판단하는 기법에 대하여 제안하고자 한다.

A Study on Improving Precision Rate in Security Events Using Cyber Attack Dictionary and TF-IDF

Jongkwan Kim*, Myongsoo Kim**

ABSTRACT

As the expansion of digital transformation, we are more exposed to the threat of cyber attacks, and many institution or company is operating a signature-based intrusion prevention system at the forefront of the network to prevent the inflow of attacks. However, in order to provide appropriate services to the related ICT system, strict blocking rules cannot be applied, causing many false events and lowering operational efficiency. Therefore, many research projects using artificial intelligence are being performed to improve attack detection accuracy. Most researches were performed using a specific research data set which cannot be seen in real network, so it was impossible to use in the actual system. In this paper, we propose a technique for classifying major attack keywords in the security event log collected from the actual system, assigning a weight to each key keyword, and then performing a similarity check using TF-IDF to determine whether an actual attack has occurred.

Key words : Security Events, TF-IDF, Similarity, IPS log, Actual network, Precision Rate

접수일(2022년 05월 30일), 수정일(2022년 06월 30일), 게재확정일(2022년 6월 30일)

* 전남대학교 정보보안협동과정(주저자)
** 전력연구원 지능화솔루션연구실(교신저자)

1. 서론

최근 원격근무, 디지털화의 가속화 등으로 정보보안의 중요성은 증가하고 있으며, 사이버공격의 지능화에 따라 인공지능기술을 이용한 다양한 대응기법이 개발중에 있다. 특히, 네트워크 침입차단시스템(NIPS; Network-Based Intrusion Prevention System)의 침입탐지의 정확도 향상을 위한 방향으로 많은 시도가 있으나 실제 운영환경에서 적용된 사례는 아직은 부족한 단계이다.

시그니처(Signature) 기반의 침입차단시스템은 네트워크상 공격트래픽 특징을 분석후 시그니처화 하여 시스템에 입력한 후에 실시간 유입되는 트래픽과 비교하여 일치할 경우 경보발생 또는 자동차단하는 시스템이다. 따라서, 공격특징을 일부 변경하거나 새로운 공격 행위에 대해서는 탐지하는 능력이 떨어지며, 새로운 공격 특징을 탐지하고, 시그니처를 제작하여 업데이트하는데 많은 시간과 노력이 소요된다.

또한, 관련된 ICT시스템의 정상적인 운영을 위해 차단규칙을 엄격히 운영하기 힘들어, 기업의 규모에 따라 하루 발생되는 경보이벤트 건수도 수천~ 수십만 건에 이르며 대부분의 보안관제 담당부서는 인력부족 등으로 인해 수십건 이하만 세부분석을 하여 처리할 수 있다.

따라서, 공격의 탐지 정확도 능력을 향상하여 오탐 이벤트 발생수량을 줄이기 위해 인공지능 기술을 적용하는 방법들이 많이 연구되어왔고 의미있는 결과를 도출중에 있다[1-10]. 인공지능분야가 머신러닝에서 딥러닝으로 발전하며 그에따라 사이버공격의 피해를 미리 탐지하고 제거하는 능력도 향상되고 되고 있다.

하지만, 기존 연구들은 MNIST 등과 같이 가공하고 정형화된 데이터셋을 이용하여 연구를 진행하였는데, 실제 보안시스템과 네트워크에서 발생하는 로그를 이용한 연구는 진행되지 않았다.

본 논문에서는 보안로그를 대상으로 정오탐을 탐지하는 데에 있어서, 실제 네트워크에 적용된 보안장비에서 추출한 로그를 이용해 공격에 많이 사용되는 특정 단어에 가중치를 부여하여 이를 학습하여 유사도 분석을 통해 실제공격 여부를 판단하는 기법을 제시한다. 비슷한 가중치를 갖는 단어가 로그에 많이 분포하고 있다면, 그 로그는 유사하다는 이론으로 접근하여 공격키워드별로 가중치를 비교하여 실제 공격과 공격이 아닌 것을 구분해 내는 방식이다. 또한, 정확도를 높이기 위해, 공격별로 키워드 사전을 미리 만들고, 키워드 사전에는 키워드의 총 출현빈도, 문서 출현빈도를 같이 추출하여, 추출된 단어만을 이용하여 계산을 하므로써, 계산속도와 정확도를 급격히 올릴 수 있었다. 해당 로그의 최종 판단은 유사한 로그를 모두 나열한후, 유사로그의 라벨값(정상:0,공격:1)들을 파악하여 최종적으로 많은 쪽으로 정상, 공격을 판단하게 된다.

본 논문의 2장에서는 단어빈도를 이용한 보안관제에 관한 기존 연구에 대해 알아본다. 3장에서는 사이버공격의 유형을 분류하고, 4장에서는 관련 사이버공격에 해당하는 공격키워드를 도출하였다. 5장에서는 가중치와 유사도 분석에 대한 기초 실험을 수행하였고, 6장에서는 실험을 통해 제안한 보안관제 방법의 유효성을 검증하였다. 마지막으로 결론 및 향후 연구방향을 논한다.

2. 관련 연구

각각의 보안로그를 하나의 Documentation으로, 보안 로그를 구성하는 공격워드 등을 Term이라고 생각하면, 각 Documentation에서 발생하는 Term 의 빈도 및 유사도를 이용하여 실제 공격로그와 그렇지 않은 로그를 구별해 낼 수 있을 것이다. 그렇다면, 이 문제는 자연어처리의 문제로 변경할 수 있고, 문장내에 사용된 단어가 비슷하다면 그 문장은 유사하다는 기본 가정을 이용하여 문제를 해결할 수 있다. 어떤 단어가 많이 쓰였는가 의 대표적인 모델이 BoW(Bag of Words)이고 Bag란 중복을 제외하지 않은 집합을 말한다. BoW는 단어를 벡터의 열에 할당하고 해당단어의 등장 빈도를 요소로 만든 벡터로, 특정어의 존재나 부재가 주제와 밀접한 연관을 가질것이라는 가정을 바탕으로 한다. 이때 사용하는 것이 TF-IDF 이며, 중요도가 높을 것 같은 특징어에 높은 가중치를 부여하고, 중요도가 낮은 단어에는 작은 가중치를 부여한다. 다음 식과 같이 나타낼 수 있다.

$$W_{x,y} = tf_{x,y} \times log(\frac{N}{df_x}) \quad (1)$$

-Wx,y = Term x within document y
-tfx,y = frequency of x in y
-dfx = number of documents containing x
-N = total number of documents

실제로 TF-IDF는 단어의 중요도와 빈도를 이용하여 탐지율을 높일 수 있고, 입력 트래픽 데이터와 정상 패턴상의 유사도 판별을 수행하여 단어의 빈도, 중요도와 벡터를 이용한 분류 기법을 이용하여 이상탐지를 수행하여 HTTP 공격을 대상으로 연구가 진행되었다[11]. 하지만, 이 연구는 실제 공격이벤트의 페이로드와는 상이한 데이터셋을 이용하였으므로 실제 네트워크에서 사용하는 것은 불가능하다.

또한, TF-IDF 와 BFR 클러스터링 알고리즘을 이용하여 네트워크 유해패킷을 감지 및 식별하는 연구가 시행되었는데, 이 연구에서는 TF-IDF를 가중치를 결정하기 위해 사용하였다[12]. 이 연구는 TF-IDF를 전처리에서 사용하였고, BFR 알고리즘을 이용하여 정탐과 오탐을 분류하였다. 실제로 TF-IDF를 단독 이용하여 침입탐지 이벤트 검출연구도 시행되었는데, HTTP 상태코드를 상황별로 분류하여, 가중치가 높은 순서대로 결과를 도출하여 보안관제사의 분석 우선순위를 제시하는 지표로 활용이 가능하였다[13].

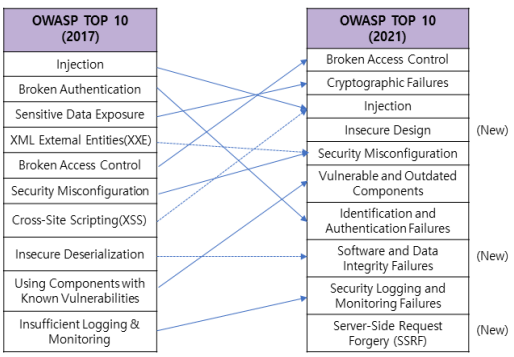
TF-IDF는 단어의 포함여부를 확인하여 유사도를 잘 찾는 장점이 있지만, 비슷한 단어지만 형태가 다른 단어일 경우 유사도를 찾지 못하는 단점이 있어 단어간/문맥간의 유사성을 정확하게 계산하기가 어렵다. 이를 극복하기 위해 분산 유사도의 개념이 소개되었으며, 이는 해당 단어의 주변 단어들을 통해 그 단어가 의미하는 바가 무엇인지 찾는 것이다. 이는 비슷한 분포를 가진 단어들은 비슷한 의미를 갖을 것이라는 가정에서 출발한다. 예를 들어 실제 SQL Injection 공격 로그중 select, union, join, update, concat, by, =, where 등이 같은 페이로드에서 계속 발견된다면 비슷한 분포를 보일 것이고 유사한 의미를 가질 것 이라고 생각할 수 있으며, 많은 로그를 학습하면서 이러한 관계를 특정하다보면 이후에는 예측이 가능할 것이다.

따라서, 본 논문에서는 어떤 단어가 많이 사용되었는가와 어떤 단어가 같이 나타나는가의 조합으로 공격로그를 판단하는 방법을 사용하고자 한다.

실제로 공격로그를 분석한 결과, 각 공격로그는 특정 키워드로 구성되어있는 것을 알게 되었고, 특정 키워드의 조합으로 공격로그를 찾아낼 수 있었다. 각 공격유형별로 단어사전을 만들어 단어별 빈도수와 문서 빈도수를 추출하여, 이를 향후, TF-IDF를 이용해 각 단어별 가중치 도출에 사용하고, 특정 키워드가 동일한 로그에 같이 나타나는 것을 판단하여 공격로그를 분류하고자 한다.

3. 사이버 공격 유형

일반적인 네트워크 구조에서 대부분의 사이버공격은 인터넷망에서 이루어지고 있고, 셀 수 없는 사이버 공격이 네트워크로 유입되고 있다. 따라서, 주요 보안장비는 인터넷망에 설치되어있다. 실제로 모 전력회사 침입차단시스템으로 유입되는 사이버공격은 1일 평균 50만건 정도이고 이중 48만건은 등에서 자동 차단되며, 약 2만건이 보안관제사가 상세 분석해야하는 대상으로 표출되고 있다[14]. 대부분 Payload Header/Body 에 공격 인자를 포함하여 불법수정, 위변조, 시간 변경, 가장, 재연, 부인 등의 공격이다. 그리고, 전송되는 데이터에 멀웨어, 바이러스, 트로이목마, 스파이웨어, 랜섬웨어 등을 포함하여 공격을 시도하는 것을 알 수 있었다. 이렇게 유입되는 사이버공격은 OWASP (the Open Web Application Security Project)에서 발표하는 보안상 영향을 크게 줄 수 있는 공격 10가지로 분류할 수 있다. 그림 1에 OWASP TOP10의 2017년 버전과 2021버전을 도시하였다[15].



(그림 1) 사이버 공격 유형 정의

다양한 공격유형이 있지만, 결국에는 Injection(또는

XSS), Broken Access Control, Cryptographic problem, 비인가접근 등의 공격유형으로 분류할 수 있고, 공격 형태는 시간이 지나면서 급격히 바뀌지 않는 것을 알 수 있다.

대부분의 보안관제시스템은 IPS에서 발생하는 보안 이벤트에 의존하여 관제를 진행중에 있다. IPS에서 발생하는 이벤트 중 룰셋에 일치하는 것은 바로 차단(defense)하지만, 유사한 경우에는 알람(alarm)을 발생하여 보안관제사가 이를 분석해야만 한다. IPS에서 발생하는 이벤트 중 하루 2만여 건의 알람이벤트는 대부분 오탐 및 비공격/식별 불가 데이터이며, 이 중 실제 공격이벤트는 0~2건으로 낮은 비율(0.01%)이다. 하지만 알람이벤트 중에 실제 공격이 포함되어 있으면 치명적이므로 꼭 분석하여 대응하여야 한다. 그러므로 알람이벤트 속에 포함된 진짜 공격을 찾아내기 위해 관제사가 2만 건을 일일이 다 분석해야만 하는 상황이다. 특히 비공격/식별 불가 데이터는 페이로드에 사람이 눈으로 식별할 수 없는 형태로 되어있어서, 보안관제사가 처리할 수 없는 형태의 데이터이고, 이벤트를 클릭해서 보기 전에는 알 수 없는 형태로 되어있다. 따라서 보안관제사는 이러한 식별 불가 데이터를 일일이 열어보고 제외하는 반복적이고 시간 소비적인 일을 어쩔 수 없이 진행해 오고 있었다. 주요 이벤트 유형과 페이로드 구성은 표1과 같다.

표 1은 실제로 모 전력회사의 IPS(Intrusion Protection System)에서 생성되는 공격이벤트를 OWASP Top 10 공격에 맞게 분류하고, 실제 페이로드에 대한 일부예를 나타내었다. 표에서 보는바와 같이 실제 사이버공격은 OWASP TOP 10과 유사한 측면이 있다. File Upload는 Cryptographic Failures, Broken Access Control, XSS는 Injection의 한 유형이며, File Download는 Vulnerable and Outdated Components, Software and Data Integrity Failures의 일종이다. 따라서 전체 OWASP 10을 5개의 공격유형으로 다시 분류하여 공격을 탐지할 수 있다. 이후 실험에서는 5개의 유형으로 공격키워드 도출과 탐지를 시도하도록 한다.

<표 1> 실제 네트워크에서의 공격 유형 및 페이로드 예

(공격) 이벤트명	페이로드(예)
Broken Access Control (46%)	GET /wp-login.php HTTP/1.1 Host:recurit.kepco.co.kr
File Upload (19%)	POst /service/krashrpt.php?kuid='id | wget http://80.82.64.14/richard; curl -O http://80.82.64.14/richard; chmod +x richard; sh richard' HTTP/1.0
Injection (14%)	GET /NewsType.asp?SmallClass=%20unio%select%200,username%2BCHR(124)%2Bpassword,2,3,4,5,6,7,9%20from%20admin%union%20select%20*%20from%news%20where%201=2%20and%20''=' HTTP/1.1 Host:www.kepco.co.kr
Cross Site Scripting (11%)	POST /getl18nResources.do?l=<html><head>...<scrrip%20language="javascropt">function%20redi ret(){top.location.href="http://11.134.1.9/cwp/?SMAC=E8:03:9A:67:9A:67:99:12&SRCIP=10.236.133.123&SPORT=19383&DSTIP=203.248.45.206&DPOrt=80&URL=http://srm.kepco.net"}</scrip>... Host:srm.kepco.net
File Download (10%)	GET /shopper.cgi?newpage=../../../../../etc/ passwd HTTP/1.1 Host:203.248.44.220

4. 공격키워드 추출

모든 SQL Injection 공격구문은 도메인 뒤의 url 내부 파라미터에 작성되는 공격구문이다. union과 select가 각각 다른 로그에 존재할 때는 Injection 공격으로 판단할 수 없고 공격 로그처럼 보이지 않는다. 하지만 두 단어가 같은 로그에 존재할 경우, 확실한 SQL Injection 공격임을 알 수 있으며, 공격을 위해 사용될 위험성이 존재한다. 예를 들어, 1' UNION SELECT 와 같은 싱글쿼터 공격은 공격포인트에 싱글쿼터를 찍어줌으로써 에러 메세지가 출력되는지 확인하고, 출력된다면 '해당 홈페이지가 웹 취약점을 가지고 있다.'라고 공격자가 판단을 하는 구조의 공격이다. SQL Injection 공격은 싱글쿼터 앞에 백슬래쉬를 붙여줌으로써 싱글쿼터를 문자로 인식하게 하여 방어가 가능하지만 오래되고 단순한 공격인 만큼 변형을 통해 다양한 형태의 공격이 계속해서 생성되고 있다. 본 논문에서 사용된 학습데이터는 2019년 12월 19일

부터 2020년 2월 5일까지 모전력회사 IPS 13대에서 추출한 보안이벤트 로그이다. 총 18만건을 추출하였는데 대부분이 Broken Access Control에 해당하는 공격이어서 다른 공격들에 대한 충분한 데이터를 확보할 수 없었다. 따라서, 충분한 학습데이터 확보를 위해 보안전문업체에서 보유하고 있는 공격 및 정상데이터를 추가하여 전체 학습데이터를 구성하였다. 표 2에 이를 나타내었다.

<표 2> 라벨링된 학습데이터 구성

구분		한전	보안업체	총계
Broken Access Control	공격	155,308	0	155,308
	정상	12,433	137,700	150,133
File Upload	공격	4,818	152,183	157,001
	정상	2,314	142,208	144,522
Injection	공격	1,018	148,982	150,000
	정상	121	154,998	155,119
Cross Site Scripting	공격	252	163,198	163,450
	정상	37	141,452	141,489
File Download	공격	2,084	152,006	154,090
	정상	2,304	147,171	149,475
총계		180,689	1,339,898	1,520,587

실제 IPS 에서 추출한 로그는 아래와 같은 형태로 구성되는데, 아래 로그는 SQL Injection에 대한 공격 로그이다.

/blog/postlist.asp?‥=union%20select%20from%20admin%20WHERE%20cal_date=&b_id=jesus100%20AND%201=1

여기에서 추출되는 공격 키워드는 union, select, from, admin, where, id, and 등이다. 공격키워드는 숫자, 공백 등을 제외하고 / 이후의 단어 등 공격에 직접 사용되지 않은 URL 등을 제외한 모든 단어를 표 2에 제시된 총 150만 건에 대한 로그로 부터 키워드로 추출하였다. OWASP cheat sheet를 참조하여 실제로 공격로그를 대상으로 빈도 계산을 하여 공격키워드를 추출 할 수 있었다. 실제 공격별 키워드는 표3과 같다. 공격키워드는 표3보다 더 많이 추출되었지만, 대표적인 키워드를 나타내었다. 실환경에서 추출한 IPS로그는 표1에서 예시로 나타내었고, 공격유형별로 그 형태가 매우 다양하다. 따라서, 위의 SQL Injection 로그를 주요 키워드만 분리하여 아래와 같이 학습용 데이터로 다시 구성하였다.

"union select from admin WHERE id AND 1(공격)"

<표 3> 공격 유형별 키워드(예)

Broken Access Control	File Upload	Injection	Cross Site Scripting	File Download
MyAdmin	importuser	select	innerHTML	etc
root	shell	address	onload	vhosts
Authorization	php	cmd	eval	opt
master	GET	decode	alert	grub
administrator	H4x0r	content	XSS	conf
Basic	exec	cmdshell	script	proc
admin	asp	DEFAULT	onmouseover	mkuser
administrators	POST	cookie	find	default
ZmEu	awen	database	navigator	root
wp	passthru	echo	Hello	config
manager	jsp	DELAY	svg	wsconfig
superuser	HTTP	dbms	onsubmit	usr
python	PerlKit	char	top	passwd
saedit	system	dba	vibrate	ini
webmaster	title	EXE	fuzzelement	var
weblogic	upfile	document	onfocus	group
mysql	popen	exec	source	gz
config	cdx	eval	document	inetpub
root	VBScript	and	onblur	hosts
asdf	anyloginеval	where	toString	bashrc
scan	war	file	domain	recycle
funcspecs	upload	shell	INJECTX	motd
qwer	c99madshell	getmapping	onclick	

5. 가중치 및 유사도 분석

5.1 가중치 알고리즘

가중치 알고리즘은 문서의 특징을 추출하거나 분류하는데 쓰인다. 이를 이용하여 보안이벤트가 실제 공격인지 아닌지 여부를 파악할 수 있다. 단어 가중치 계산방법에는 다양한 알고리즘이 있고, 특정 단어가 로그분류라는 과업에 얼마나 중요한 역할을 하는지 수치화하는 걸 목적으로 한다.

DF(Document Frequency) 는 w라는 단어가 몇 개의 문서에 등장했는지 빈도를 나타내고, AC(Accuracy) 는 w라는 단어가 긍정적인 문서(정상로그)에 나타난 빈도, w가 부정적인 문서(공격로그)에 나타낸 빈도 간 차이이다. AccR(Accuracy Ratio) 은 긍정적인 문서(정상로그)가 주어졌을 때 w라는 단어가 등장할 조건부 확률로 가중치를 부여하는 방식이고, PR(Probability Ratio)은 AccR과 유사하나 긍/부정(정상/공격) 조건부확률 차이의 절대값 대신 두 값 사이의 비율을 계산하는 방법이다. OddR(Odds Ratio)

은 임의의 사건 A가 발생하지 않을 확률 대비 일어날 확률 사이의 비율이다. OddR이 클수록 긍정 범주 판별에 유용한 단어라는 의미이다. OddN (Odds ratio Numerator)은 계산 효율성을 목적으로 OddR에서 분자 부분만 떼어낸 식이다. F1-Measure 는 F1이 클수록 긍정 범주 판별에 유용한 단어이다. Information Gain은 정보이론에서 Entropy란 불확실성 또는 혼잡도의 척도로 쓰인다. 만약 어떤 데이터의 범주는 2개인데, 전체 관측치의 절반이 한 범주이고 나머지 절반이 다른 범주라면 엔트로피는 최대값(1)을 가진다. 반대로 모든 관측치가 하나의 범주로 구성돼 있다면 엔트로피는 최소값(0)이 된다. 정보이득이란 특정조건과 비교한 에트로피 간 차이를 의미한다.

Chi-Squared Statistic 은 통계량이 클수록 '최고'라는 단어가 긍정, 부정 극성 분리에 중요한 term이라는 이야기이다. BNS(Bi-Normal Separation)은 w가 존재하는 긍정로그 및 부정로그 개수를 계산하여 BNS score를 계산하는 것이다.

이와 같이 10개 가중치 알고리즘에 대하여 실험해 본 결과, DF와 AC는 직관적으로 키워드에 대한 중요도를 알수 있었고, AccR, PR, OddR, OddN, F1은 조건부에 대한 등장빈도가 너무 낮아 보안 로그에는 사용할 수 없음을 알 수 있었다. Information Gain, Chi-Squared Statistic, BNS 는 계산식에 0이 들어가는 경우가 많아 무한대로 수렴하는 경우가 많았다. 결국 DF와 AC를 사용하는 것이 보안로그에는 좋은데, 이 둘을 합쳐놓은 것이 TF-IDF라고 할 수 있다.

5.2 유사도 알고리즘

분석대상 로그에 등장하는 단어들의 빈도를 세서 이를 벡터형태로 바꿀 수 있고, 그 벡터들간의 거리(유사도)를 잴 수 있으며, 이렇게 구한 거리는 상호간의 의미를 내포한다. 유사도(Similarity)란 비슷한 정도를 나타내는 지표를 뜻한다. 하지만, '비슷하다'는 단어의 어감에서도 알 수 있듯 굉장히 주관적인 지표이다. 이를 정량화하는 노력이 필요한데, 문서간 유사도를 측정하는 지표는 다수 제안되었지만, 대체로 단어(word, term)수준의 방법론 들이다. 두 문서에 겹치는 단어가 많을수 록 유사도가 높다는 결과를 내놓는 식이다. 단어 수준의 유사도 측정은 (1) 문서길이 (2) 동시 등장 단어 (3) 흔한/희귀한 단어 (4) 출현빈도 등을 어떻게 처리하는지에 따라 다양한 방법론이 있다. 이와 관련한 측정 지표는 7개 정도 있는데 이를 살펴보도록 한다.

<표 4> 유사도 알고리즘 계산을 위한 표

문서1 \ 문서2	존재하는 단어 수	존재하지 않는 단어 수
존재하는 단어 수	a_{ij}	b_{ij}
존재하지 않는 단어 수	c_{ij}	d_{ij}(전체 단어 중 문서 i, j 모두 존재하지 않는 단어의 수)

-aij : 두문서에 모두 등장한 단어
-bij : Doc1에 등장했는데 Doc2에는 등장하지 않은 단어
-cij : Doc1에 등장하지 않지만, Doc2에 등장한 단어
-dij : 두 문서에 모두 등장하지 않는 단어
-xik : i번째 문서에 k번째 단어가 몇 번 등장했는지 빈도

Common Features model은 i번째 문서와 j번째 문서에 동시에 등장한 단어수를 전체단어수로 나누어 구한다. 보통 전체 문서에 등장하는 단어수가 10만개에 육박하기 때문에 dij 가 매우 크다. 따라서 이처럼 계산하는 유사도는 대체로 0에 가까운 작은 값을 지니고, 아래의 계산식으로 나타낼수 있다.

$$S_{ij} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}} \quad (2)$$

Ratio model 은 common features model에서 dij를 빼고 계산하면 된다. Simple matching coefficient 는 Common features model의 식에서 분자에 dij를 반영하면 되는데 대부분 구한 값이 큰 경향이 있고 아래 식으로 나타낸다.

$$S_{ij} = \frac{a_{ij} + d_{ij}}{a_{ij} + b_{ij} + c_{ij} + d_{ij}} \quad (3)$$

Jaccard similarity 는 Ratio model 과 본질적으로 유사하고, 아래 식으로 구할 수 있다.

$$Sij = \frac{\sum_k \min(xik, xjk)}{\sum_k \max(xik, xjk)} \quad (4)$$

Overlap similarity는 아래 식 5와 같이 구할 수 있다.

$$Sij = \frac{\sum_k \min(xik, xjk)}{\min(\sum_k xik, \sum_k xjk)} \quad (5)$$

Cosine Similarity는 각각의 문서에 해당하는 열을 벡터로 놓고 두 벡터를 내적하게 되면 두 벡터가 이루는 각도(유사도)가 되는데 아래 식 6으로 구할 수 있다.

$$Sij = \frac{\sum_k (xik \times xjk)}{\sqrt{(\sum_k x^2 ik), (\sum_k x^2 jk)}} \quad (6)$$

Euclidean Similarity는 유사도 계산을 위하여 가장 간단하게 사용할 수 있고 다른 유사도 방식과 다르게 결과값이 얼마나 유사한지에 대한 수치가 아닌, 비교 기준에서 얼마나 멀리 떨어져 있는가를 의미한다. 아래 식 7로 나타낼 수 있다.

$$\sum_{k=1}^{n} (T_k - N_k)^2 \quad (7)$$

- T : 학습 Document의 단어 가중치 집합
- N : 새로운 Document의 단어 가중치 집합

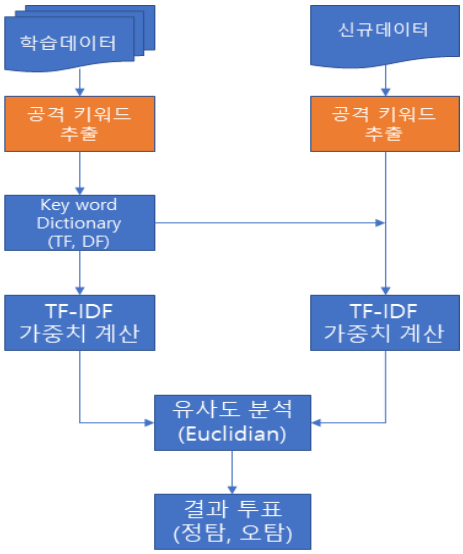
총 7개의 유사도 알고리즘에 대하여 실험해 본 결과 정상로그를 학습하여 이 로그와 정상로그, 공격로그와 유사도를 비교하여 아래 표 5의 결과를 얻게 되었다. 정상로그와 공격로그에 대한 유사도의 차이를 비교해 보면 유클리디안 모델이 가장 뚜렷한 차이를 가지는 것을 알 수 있다.

<표 5> 유사도 알고리즘 비교

모델 \ 유사도	정상	공격	차이 (절대값)
Common Features Model	0.023	0.009	0.014
Ratio Model	0.328	0.187	0.141
Simple Matching Coefficient	0.952	0.934	0.018
Jaccard Similarity	0.653	0.652	0.001
Overlap Coefficient	0.894	0.868	0.026
Cosine	0.952	0.913	0.039
Euclidean	16.093	61.073	**44.98**

6. 실험 및 결과

IPS 로그를 의미있는 word 형태로 분류하고, 각각의 word 빈도(TF)와 전체 학습데이터셋에서 특정 word가 얼마나 자주 등장하는 지를 의미하는 역문서 빈도(IDF)를 곱하여 유사도를 계산한다. 즉, 다른 로그에는 등장하지 않지만 특정 로그에서만 자주 등장하는 word를 찾아 로그 내 word의 가중치를 계산하는 방법이다. 표2의 학습 데이터셋에서 유사도가 높은 로그를 발견하여 라벨링을 확인후, 라벨링이 공격이라고 되어있으면 공격으로 표현한다. 여러개의 유사 로그를 발견하였다면, 과반수 이상의 결과를 채택한다. 이를 순서도로 나타내면 아래 그림 2와 같다.



(그림 2) 전체 프로세스

유사도 측정을 위해서는 먼저 문서를 수치화하는 과정이 필요하다. 이 과정에서 활용되는 것이 바로 TF-IDF 이다. 특정 Document에 등장하는 단어들이 해당 Document에서 얼마나 중요한지를 나타내는 통계적 수치이고, 이 값이 높은 단어일수록 Document에서 중요한 단어라고 판단할 수 있다. 이 값은 TF(단어의 빈도)와 IDF(문서 빈도의 역수)의 곱으로 계산된다. 다시 말해 이 TF-IDF값을 활용하여 문서를 수치화한후 유사도를 측정하는 식으로 문서 간 유사도

를 계산한다고 할 수 있다.

공격로그의 페이로드 단어를 분석해 보면, 수십개의 키워드가 존재하는데, 각 키워드 들은 전체 로그를 대상으로 학습하면 TF, IDF 값을 찾아낼수 있고, 가중치가 비슷한 로그들이 존재하는 것이 가장 유사한 로그로 구분될 수 있는 것이다.

공격키워드에 사용된 학습데이터는 표2에서 언급한 실제 네트워크 취득 로그와 보안전문업체에서 보유하고 있는 로그를 혼합하여 정탐/오탐 비율을 약 50:50으로 구성하였다.

6.1 단어사전 생성

학습용 데이터를 이용하여 단어빈도와 문서빈도로 구성된 단어사전을 생성한다. 아래 로그는 실제 Injection 공격의 페이로드로 공격키워드만을 추출하여 키워드로 구성된 단어사전의 예이다.

<표 6> 단어사전(예)

```
select: {'freq': 6132, 'doc_freq': 1022}
address: {'freq': 5110, 'doc_freq': 1022}
cmd : {'freq': 4156, 'doc_freq': 1090}
decode : {'freq': 4088, 'doc_freq': 1022}
content: {'freq': 4088, 'doc_freq': 1022}
cmdshell: {'freq': 4088, 'doc_freq': 1022}
default : {'freq': 4088, 'doc_freq': 1022}
cookie : {'freq': 2934, 'doc_freq': 978}
database : {'freq': 2367, 'doc_freq': 1345}
```

* freq : 단어 빈도, doc_freq : 단어가 나온 문서수

6.2 가중치 계산

가중치는 문서의 특징을 추출하거나 분류하는데 쓰인다. 이를 이용하여 보안이벤트가 실제 공격인지 아닌지 여부를 파악할 수 있다. 우리의 목적은 단어 정보만으로 범주(공격, 정상)를 예측하는 것이다. 다시 말해 보안이벤트 로그를 공격/정상으로 분류할 때 어떤 단어가 중요한 역할을 하는지 알고 싶은 것이다. 위에서 생성한 단어사전을 이용하여 하나의 로그에 대하여 각각 단어들에 대한 TF-IDF 가중치를 각각 미리 계산하여 저장해 놓는다. 아래 표는 이를 이용한 하나의 로그에 대한 결과값이다. 로그 각각에 대하여 단어 가중치를 갖게 되고 마지막에는 이 로그가 실제 공격인지, 아니면 정상인지를 나타내는 Label을 입력하게 된다.

<표 7> 가중치와 Label을 포함한 로그

```
{"tf_idf_weight": {"select": 0.007868774263794794, "address":
0.035238271765110116, "cmd": 0.006465473071127079, "decode":
0.007868774263794794, "content": 0.008871310399600348, "cmdshell":
0.015737548527589587, "default": 0.02360632279138438, "cookie":
0.007868774263794794, "database": 0.025070680016573088, "edho":
0.007564817867629225, "delay": 0.025070680016573088, "dbms":
0.007868774263794794, "char": 0.007868774263794794, "dba":
0.007868774263794794, "exe": 0.007868774263794794, "document":
0.007868774263794794, "exec": 0.007868774263794794, "eval":
0.007868774263794794}, # 키워드별 가중치 정보
 "label ":1,# 레이블(정답지 1-공격, 0-정상)
```

6.3 유사도 분석

이를 종합하면 분석대상 로그 내에 등장하는 단어들의 빈도가중치를 계산하여 실제 로그간의 거리(유사도)를 잴 수 있으며, 이렇게 구한 거리는 언어학적인 의미를 내포한다. 유사도 계산을 위하여 정상과 비공격을 가장 확실히 구별할 수 있는 유클리디안 디스턴스(식 7)를 활용하여 로그간 유사도를 계산하였다.

이를 이용하여 유사도를 계산하면, 각 로그에 대한 유사도를 구할 수 있다. 이를 표 8에 나타내었다.

<표 8> 유사도 계산

```
{"tf_idf_weight": {"select": 0.007868774263794794, "address":
0.035238271765110116, "cmd": 0.006465473071127079, "decode":
0.007868774263794794, "content": 0.008871310399600348, "cmdshell":
0.015737548527589587, "default": 0.02360632279138438, "cookie":
0.007868774263794794, "database": 0.025070680016573088, "edho":
0.007564817867629225, "delay": 0.025070680016573088, "dbms":
0.007868774263794794, "char": 0.007868774263794794, "dba":
0.007868774263794794, "exe": 0.007868774263794794, "document":
0.007868774263794794, "exec": 0.007868774263794794, "eval":
0.007868774263794794}, # 해당 Document의 가중치 정보
 "label ":1,# 레이블(정답지 1-공격, 0-정상),
 "similarity": 0.0009754693056655417
```

6.4 최종 결과 도출(voting)

분석할 신규 로그와 유사도 분석이 끝난 로그중에 유사도 차이가 0.001 이하인 로그들을 추출하여 분석할 신규로그에 대한 공격/정상을 판별하게 된다. 0.001의 임계치는 3번의 실험을 통해 도출해 내었는데, 아래 표와 같은 결과를 얻었다.

<표 9> 임계치 도출 결과

임계치	일치건수 (1/2/3회차)	유사도 측정불가 건수	Accuracy (정확도)
0.1	940/960/1,100	0/0/0	47/48/55%
0.01	940/960/1,100	0/0/0	47/48/55%
0.001	2,000/2,000/2,000	0/0/0	100/100/100
0.0001	1,520/1,460/1,520	480/540/480	76/73/76

정상, 공격의 일치를 수동으로 확인해야 하므로 150만건의 전체 데이터셋에서 정상 1천건, 공격 1천건에 대한 로그를 추출하여 수동으로 비교하여 결과를 도출하였다.

표 9에서 유사도 측정불가 건수는 유사도 임계치 이하인 건수가 존재하지 않는다는 의미이다. 임계치가 높을 경우(0.1, 0.01)는 학습로그 전체가 유사도 정보에 많이 포함되어 TF-IDF의 의미가 없는 경우이고, 임계치가 낮을 경우(0.0001)는 학습로그에서 유사도 정보가 존재하지 않아 측정 불가 상태가 발생되는 경우이다.

유사도 측정결과 1개의 로그가 존재할 경우에는 Label을 확인하여 1이면 공격으로 0이면 정상패킷으로 분류한다. 하지만, 1개 이상의 로그가 존재할 경우에는 Label 의 1의 개수와 0의 개수를 counting 하여 많은 쪽으로 로그를 분류하며, 동수가 나왔을 경우에는 판단불가로 보안관제사에게 결과를 통보해 준다.

데이터셋을 8:2로 학습 및 검증셋트로 구분후, 위에 설명한 프로세스를 구현한 결과를 표10에 나타내었다.

<표 10> 학습데이터를 이용한 결과

	Accuracy	Recall	Precision
Broken Access Control	100	100	100
File Upload	99.94	99.98	99.90
Injection	99.94	100	99.88
Cross Site Scripting	100	100	100
File Download	99.88	99.99	99.78
평 균	99.95	99.99	99.91

전체 정확도는 99.95%를 달성하였으며, 특히 미탐과 관련된 Recall은 99.99%의 정확도를 나타내었다. 이를 실제 시스템에 적용하여 실제 알람데이터를 대상으로 실계통에 적용해 보았다. 실계통 적용기간은 총 4개월(2020.9.1.~12.31)이며, 1일 평균 2만건을 대상으로 실증시험을 진행하였고, 키워드 최적화 등을 통하여 표11과 같은 결과를 취득하였다.

<표 11> 실제 로그를 이용한 결과

	Accuracy	Recall	Precision
Broken Access Control	100	100	100
File Upload	99.88	100	99.75
Injection	99.96	100	99.92
Cross Site Scripting	100	100	100
File Download	99.91	100	99.82
평 균	99.95	100	99.90

실제 시스템에서 검증한 결과 정확도는 학습데이터를 이용한 결과와 같은 99.95%의 평균치를 기록하였다. 탐지 이벤트 중 TP(정탐) 이벤트들은 Broken Access Control(94%), File Upload(16%), File Download(15%), Cross Site Scripting (5%) 순으로 많았으며 탐지된 공격들은 자동화공격툴인(취약점 스캔 도구) ZGrab(70%)에 의해 발생한 공격이벤트가 가장 많았다.

ZGrab은 ZMap과 연동하여 취약점 스캔을 수행하는 애플리케이션 계층의 오픈소스 도구이다. 각종 프로토콜을 지원하며 TLS(Transport Layer Security) 연결도 지원한다. 취약점을 찾아 보완하기 위한 용도로 개발되었지만, 이 도구를 악용하면 취약점을 스캔하여 정보수집 후 다음 공격으로 전개할 수 있다.

탐지 이벤트 중 FP(오탐) 이벤트들은 DNS SRV Type Query Flood 4.27%를 제외한 대부분 이벤트가 (탐지명에 @가 포함된) 상위기관 배포 룰셋에 의해 탐지된 이벤트이며 이 룰셋의 패턴이 우연히 일치하여 탐지된 이벤트였다. 이 이벤트들은 내부 사용자가 유튜브나 신문 서비스 이용 시 광고를 위한 카운터 등에서 패턴이 우연히 일치하거나 정해진 임계 값을 넘는 패턴을 보여서 탐지된 이벤트였다.

7. 결론

본 논문에서는 오탐률 감소와 정탐률을 높이기 위해 TF-IDF 알고리즘 기반의 프로세스를 도입하였다. 과거 연구들은 키워드를 이미지화하여 패턴을 비교하는 복잡함을 가지고 있었으며, 기존 텍스트기반 유사도 분석 알고리즘 들은 정해진 데이터셋을 이용하여 실험함에 따라 실제 데이터에는 사용할 수 없는 단점이 있었다. 본 연구에서는 실제 IPS에서 발생된 로그를 이용하여 학습에 사용하였고, 공격에 많이 사용된 키워드를 추출하여 가중치를 부여하고, 새로 분석할 로그가 있으면, 기존 로그에서 가중치가 유사한 로그를 추출하여 그 로그가 실제로 공격 또는 정상인지를 판별하여 정오탐을 도출하는 프로세스를 제시하였다. 전체적으로 평균 정탐율 99.95%를 달성하였다. 본 논문에서 제시한 방법을 다른 알고리즘과 앙상블하여 사용한다면 정탐율은 더 향상되리라 생각된다.

참고문헌

[1] 윤영근, 최인혁, 구자빈, 손주암, 오태근. “자기애자 손상평가를 위한 머신러닝 기법의 적용.” 대한전기학회 학술대회 논문집, pp. 88-89, 2019.

[2] WANG Wei-Hong, LV Yin-Jun, CHEN Hui-Bing, FANG Zhao-Lin, “A Static Malicious Javascript Detection Using SVM.” In proceedings of the 2nd International Conference on Computer Science and Electronics Engineering(ICCSEE), 2013.

[3] H. Kim, J.H. Huh, “Detecting DNS-poisoning-based phishing attacks from their network performance characteristics”, Electronics Letters, vol. 47, no.11, pp. 656-658, 2011.

[4] Y Liao, VR Vemuri, “Use of K-Nearest Neighbor classifier for intrusion detection” Computer&Security, vol 21, no.5, pp. 439-448, 2002.

[5] 양환석, “머신러닝을 이용한 APT 공격탐지 기법에 관한 연구”, 한국융합보안학회 융합보안논문지 제21권 제5호, pp. 21-27, 2021.

[6] 김도형, 이상근, 정순기, “이상금융거래 탐지시스템(FDS)을 위한 딥러닝 모델의 설계 및 구현”, 한국융합보안학회 융합보안논문지 제21권 제5호, pp. 69-78, 2021.

[7] 변성현, 김영원, 고관섭, 이수진, “CNN기반 악성코드 탐지에서 이미지 형식이 탐지성능과 자원 사용에 미치는 영향 분석”, 한국융합보안학회 융합보안논문지 제21권 제4호, pp. 59-68, 2021.

[8] 안병욱, 이중찬, 최재성, 박원형, “머신러닝과 딥러닝을 활용한 악성 패킷 탐지 기술 연구”, 한국융합보안학회 융합보안논문지 제21권 제4호, pp. 109-115, 2021.

[9] 김남욱, 이동규, 엄정호, “지능형 사이버 공격 경로 분석 방법에 관한 연구”, 한국융합보안학회 융합보안논문지 제21권 제1호, pp. 93-100, 2021.

[10] 권현, 박상준, 김용철, “딥뉴럴네트워크상에 신속한 오인식 샘플 생성 공격”, 한국융합보안학회 융합보안논문지 제20권 제2호, pp. 111-121, 2020.

[11] Mohsen Kakavand, etc. ”A Text Mining-Based Anomaly Detection Model in Network Security“, Vol14, No 40G(2014):Global Journal Of Computer Science and Technology, 2015.

[12] WesamS. Bhaya, etc. ”Anomaly Detection System for Internet Traffic bsedon TF-IDF and BFR Clustering Algorithms“, International Journal of Engineering & Technology, 8(1.5), pp 131-137, 2019.

[13] Hyoseok Kim, etc. “A Validation of Effectiveness for Intrusion Detection Events Using TF-IDF”, Journal of the Korea Institute of Information Security & Cryptology, Vol.28, pp.1489-1497, 2018

[14] Choi S, Jang M, Kim M (2020), A Study on AI algorithms to Improve Precision Rate in a Managed Security Service, Trans Korean Inst Electrical Engineering,

pp 1046-1052, https://doi.org/10.5370/KIEE 2020.69.7.1046.
[15] OWASP, “OWASP Top 10 -2001”, http://owasp.org/Top10/

[저 자 소 개]

김 종 관 (Jongkwan Kim)
2000년 2월 충주대학교 공학사
2014년 2월 고려대학교 사이버보안학과 공학석사
2022년 2월 전남대학교 정보보안협동 공학박사수료
email : keterkim@gmail.com

김 명 수 (Myoungsoo Kim)
1994년 명지대 공학사
1996년 명지대 공학석사
2011년 Penn’ State University 공학박사
1996~ 한전전력연구원 책임연구원
email : myongsoo.kim@kepco.co.kr