사용자 로그의 분석을 통한 실시간 비정상행위 탐지 기술

김명수°, 신종철, 정재명, 고유선, 이원석

연세대학교 컴퓨터과학과

An Anomaly Intrusion Detection Method using Multiple System Log

Kim, Myung Soo°, Shin, Jong Cheol, Jung, Jae Myung, Ko, You Sun, Lee, Won Suk

Yonsei University

E-mail : mskim@database.yonsei.ac.kr, ideabell@database.yonsei.ac.kr, ideale-j@hanmail.net smails2098@hanmail.net, leewo@database.yonsei.ac.kr

요 약

침입의 방법이 점차 치밀해지고 다양해짐에 따라 새로운 방식의 침입 탐지 기법 역시 지속적으로 요구되어진다. 기존의 오용 탐지 방법론은 탐지율은 뛰어나지만 새로운 침입 형태에 대한 대응 능력이 부족하다. 이러한 단점을 보완하고자 등장한 것이 비정상 행위 탐지 방법론이다. 하지만 현재까지의 연구는 네트워크나 서버 OS, 데이터베이스 등 각 개별 분야에 대해서만 진행되고 있어 그 탐지 능력에 한계가 있다. 본 논문에서는 이러한 한계를 극복하고자 사용자의 네트워크 및 운영체제 로그를 통합 하고, 데이터마이닝 기법 중 빈발 패턴 마이닝 기법을 이용한 보다 정확한 비정상 행위 탐지 기술을 제안한다.

1. 서론

침입이란 권한이 없는 사용자가 발생시키는 보안 문제 또는 합법적인 사용자가 권한을 남용하는 문제를 뜻하며 더불어 자원의 기밀성, 무결성 등에 저해되는 행위 집합으로 정의되기도 한다.[1,2]

과거에 주로 사용되었던 침입 방법들이 단순한 패턴과 한정된 범위에 국한되었던 반면 새로운 침입 방법들은 그 방식이 치밀하고 다양하다. 비정상 행위 탐지 방법론은 사용자의 평상시 행위 패턴을 정상 패턴이라 가정하고 이에 벗어나는 행위를 비정상 행위로 탐지해내는 기술이며, 기존의 오용탐지 방법론과 비교하여 알려지지 않은 침입 패턴에 대한 탐지가 가능하다는 장점이 있다. 특히 데이터 마이닝 기법을 이용한 탐지법이 최근 주목받는 추세이다. 하지만 침입 탐지에 대한 이러한 접근은 대부분 네트워크 측면에서만 이루어지고 있으며, 사용자 어플리케이션 등 보다 구체적인 정보의 통합적 사용 없이는 그 탐지력에 한계가 있다.

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업으로 수행된 연구임 (No.R0A-2006-000-10225-0)

본 논문에서는 기존의 네트워크 측에서만 이루어지던 비정상 행위 탐지법을 네트워크와 운영체제가 통합된 환경으로 확장하고, 이와 함께 데이터 마이닝 기법을 이용하여 보다 안정적이고 정확한 탐지율을 가진 비정상행위 탐지 기술을 제안한다.

2. 관련연구

대표적인 침입 탐지 방법론으로는 오용 탐지 방법론(Misuse Detection Model)과 비정상행위 탐지 방법론(Anomaly Detection Model)이 있다. 오용 탐지 방법론은 이미 학습된 패턴만을 사용하기 때문에 새로운 침입 패턴에 대한 탐지가 불가능하다

는 단점이 있다. 비정상행위 탐지 방법론은 이러한 기존의 단점을 보완하는 방법으로서 정상행위로 추정되는 패턴을 미리 생성해두고 현재 사용자의 사용 패턴이 이를 벗어날 경우 비정상행위로 판단한다. 비정상행위 탐지 방법론에는 통계기반의 탐지 방법론[3], 예측 가능한 패턴 생성(Predictive Pattern Generation) 방법론[4], 데이터마이닝(Data Mining) 방법론[5] 등이 있다.

통계기반의 탐지 방법론은 기존에 습득한 패턴들을 통계적인 공식을 이용하여 프로파일을 만들고 이 프로파일을 통해 비정상행위를 판단하는 방법이다. 예측 가능한 패턴 생성 방법론은 이미 발생했던 이벤트들을 기반으로 미래에 발생할 이벤트를 예측하여 침입을 탐지하는 방법이다. 데이터마이닝 방법론은 데이터로부터 패턴정보를 추출하여 이를 이용하는 방법으로서 연관규칙(Association Rules), 빈발항목집합(Frequent Patterns), 군집화(Clustering) 등의 알고리즘이 사용될 수 있다.

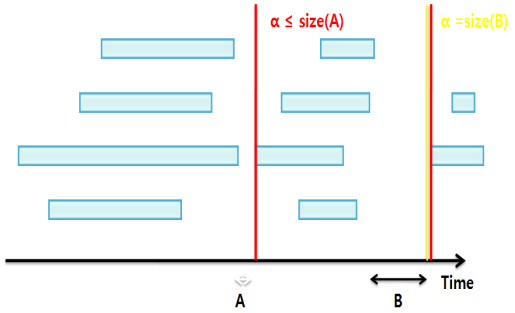
기존의 비정상행위 탐지 방법론에 관한 연구는 패킷들 간의 연관관계를 이용하여 비정상 행위를 탐지하는 방법[6]이나, 데이터베이스 질의를 대상으로 비정상 행위를 탐지하는 방법[7] 등과 같이 각 독립된 분야만을 그 대상으로 해왔다. 본 논문에서는 운영체제와 네트워크 로그를 동시에 고려한 환경에서 데이터마이닝 기법 중 빈발항목집합 알고리즘을 적용하여, 보다 안정적이고 정확한 탐지율을 가진 침입탐지 기술을 제안한다.

3. 시스템 모델링

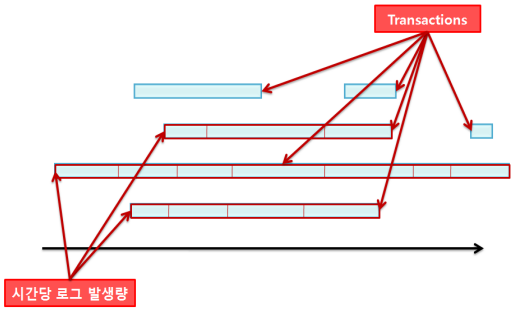
본 논문은 정상행위 패턴 생성 및 비정상행위 판별을 위해 운영체제의 이벤트 로그(Windows Event Log)와 네트워크 패킷 로그(TCP/IP packet Header)를 복합적으로 사용하여 비정상행위 탐지를 위한 정상행위패턴을 생성한다. 실시간으로 생성되는 각 운영체제 로그와 네트워크 로그는 트랜잭션의 단위로 분석에 사용되며, 이러한 트랜잭션은 시간당 로그 발생량이나 프로세스를 기준으로 구분된다. 하나의 프로세스가 수행되는 시간이 일정하지 않고 시간당 로그 발생량도 규칙적이지 않으므로 트랜잭션의 길이는 매우 유동적이다. 패턴 매칭은 시간, 시간당 로그 발생량, 프로세스 종류 등에 따라 정의되는 비교단위집합(short-term)을 기준으로 수행되며, 정상행위로 분류되는 트랜잭션의 경우 새로운 패턴을 생성하기 위한 패턴 생성자의 입력값이 된다.

3.1 논리적 행위 구분 모델링

제안하는 방법에서는 사용자의 PC들을 대상으로 하여 특정 이벤트의 결과로 생성되는 운영체제 로그들과 패킷 단위로 생성되는 네트워크 로그들로부터 정상 행위 패턴을 도출하여 실시간으로 발생하는 사용자 로그와의 비정상행위도를 계산한다. 또한, 정상 행위 패턴을 도출하기 위해 기존의 데이터 마이닝 기법 중 빈발항목집합 마이닝 기법을 사용한다. 사용자 로그 데이터는 논리적으로 하나의 행위를 나타낼 것으로 추정되는 로그 집합으로 트랜잭션을 생성하며, 시간당 로그 발생량, 수행되는 프로세스의 종류 등을 기준으로 구분 된다.



(a) 시간당 로그 발생량 기준 트랜잭션 구분



(b) 프로세스 기준 트랜잭션 구분

[그림 1] 트랜잭션 구분 기준 예시

시간당 로그 발생량을 기준으로 트랜잭션을 구분하는 경우 사용자가 하나의 행위를 마치고 다음 행위를 시작하기 전에 항상 일정한 간격으로 운영체제 이벤트 발생이 줄어드는 시점이 존재한다고

가정한다. 이 경우 정상 행위 패턴에 각 프로세스 간 동시 발생 가능성이 반영된다는 장점이 있다. 수행되는 프로세스의 종류를 기준으로 트랜잭션을 구분하는 경우 각각의 행위에 대하여 수행되는 프로세스가 다르다는 점을 이용한다. 이 경우 빈발하게 발생하지 않는 프로세스에 대한 고려가 가능하다는 장점이 있다.

3.2 정상행위 패턴 생성 모델링

[표 1]은 예제 로그 데이터이다. 먼저 각 로그의 속성 값에 대한 지지도를 모두 구하면 후보1-항목 집합이 된다. 이때 지지도는 전체 트랜잭션 중 특정 속성의 값이 등장한 트랜잭션의 개수이며, 예제

[표 1] 예제 로그 데이터 및 빈발1-항목 집합

트랜잭션ID	PID	사용자ID	event ID	종류
T1	p1	u1	e1	k1
	p2	u1	e2	k1
	p1	u1	e1	k1
T2	p3	u2	e3	k1
	p4	u3	e4	k1
T3	p4	u3	e1	k1
	p3	u3	e2	k2
	p2	u3	e1	k1
T4	p1	u1	e2	k2
	p1	u1	e1	k1

Data	Support
p1	2
p2	2
p3	2
p4	2
u1	2
u3	2
e1	3
e2	3
k1	4
k2	2

의 경우 최소지지도를 2라고 가정한다. 이후 이들 중 최소 지지도 이상의 지지도를 갖는 속성 값들로 이루어진 빈발1-항목 집합을 생성하고 이를 이용하여 빈발2-항목 집합을 순차적으로 생성한다. 같은 방법으로 더 이상 항목이 생기지 않을 때까지 빈발 k-1-항목으로부터 빈발k-항목을 생성한다. 이렇게 생성된 모든 빈발항목집합들은 빈발로그집합을 구하기 위해 사용된다. 빈발항목집합이 하나의 로그에서 같이 발생하는 항목들의 집합인 반면 빈발로그집합은 하나의 트랜잭션에 같이 나타나는 로그들의 특징들의 집합으로서 이때 로그의 특징은 빈발항목집합의 결과가 된다. 생성된 빈발항목집합들이 빈발1-로그집합이 되고 식별자를 통해 구분된다. 빈발항목집합을 생성할 때와 마찬가지로 더 이상 새로운 로그집합이 생성되지 않을 때까지 빈발 k-로그집합을 생성하게 되면 이것을 최대빈발로그집합(MLL: Maximal Large Log-set)이라 정의한다.

[표 2] 최대 빈발 로그 집합 (MLL)

로그집합	지지도	로그집합	지지도
{(e1,k1),(e2),(k1)}	2	{(p2,k1),(e1,k1)}	2
{(p1,u1,e1,k1),(p1,u1)}	2	{(e1,k1),(e1,k1)}	2
{(p1,u1,e1,k1),(e2)}	2	{(e1,k1),(e2,k2)}	2
{(p4,u3,k1),(p3)}	2	{(k1),(e2)}	3
{(p4,u3,k1),(k1)}	2	{(k1),(k1)}	3

3.3 비정상행위 판별 모델링

사용자 입력으로부터 생성되는 실시간 로그 데이터와 앞서 생성한 최대 빈발 패턴 테이블과 비교하여 비정상행위여부를 판별할 수 있다. 이때 비정상행위를 판별하는 로그 데이터의 단위를 비교단위집합이라고 하며, 각 비교단위집합은 고정 시간, 로그 발생량, 수행 프로세스 등의 기준으로 구분되어진다.

패턴 매칭 알고리즘은 기존의 패킷간 연관 관계를 이용한 네트워크 비정상행위 탐지에서 사용되었던 방식을 사용한다. 최대빈발로그집합 MLL과 속성-지지도 집합 ESS(Element-Support Set), 비교단위집합(short-term) S를 통하여 최대빈발패턴 테이블을 이용한 정상행위도 RNA를 정의하고, 비교단위집합 S의 각 빈발 항목과의 일치도를 나타내는 ANA를 정의한다. 이러한 정의들을 토대로 비교단위집합 S에 대한 비정상행위도를 도출할 수 있으며, 기존의 네트워크 로그만을 통한 분석에 운영체제 로그를 추가적으로 고려하기 위하여 각 RNA 및 ANA 수치를 가중치를 적용하여 합한다. 다음은 운영체제 로그와 네트워크 로그를 동시 적용한 정상행위도 RNA 및 비교단위집합과 빈발 항목과의 일치도 ANA이다.

$$RNA(os\&network) = \alpha \cdot RNA(os) + \beta \cdot RNA(network)$$
$$ANA(os\&network) = \alpha \cdot ANA(os) + \beta \cdot ANA(network)$$
$$(\text{where } \alpha+\beta = 1)$$

4. 실험 및 평가

제안하는 방법의 가능성을 보이기 위해 다양한 사용자 그룹으로부터 직접 수집한 로그들을 이용하여 비정상행위 탐지를 위한 운영체제 및 네트워크 로그의 통합적 RNA 및 ANA를 도출하였다. 실험을 위한 로그 데이터 중 운영체제 로그 데이

터에는 timestamp, 사용자, 개체 이미지, 프로세스 이미지, 엑세스 종류 등이 포함된다. 이 중 개체 이미지는 프로세스의 목적이 되는 이미지 파일의 경로를 포함한 이름을 의미하며, 프로세스 이미지는 프로세스의 이미지 파일의 경로를 포함한 이름을 뜻한다. 엑세스 종류란 프로세스가 목적이 되는 파일에 수행한 행위이다. 또한 네트워크 로그 데이터에는 timestamp, IP 주소, 포트번호, 패킷 크기, 프로토콜 등에 대한 정보가 포함된다.

실험은 운영체제 이벤트 로그와 네트워크 로그를 동시에 사용한 경우와, 운영체제 이벤트 로그, 혹은 네트워크 로그 한가지만을 사용한 경우 각각에 대한 RNA 및 ANA 수치를 계산함으로써 그 차이를 비교하는 방식으로 진행하였다. 또한, 각 트랜잭션의 구분은 시간당 로그 발생량을 기준으로 계산되었다.

	Network					
	short-term	1	2	3	4	5
inner1	2min	17.823002	17.823002	17.823002	17.74756	17.674612
	5min	17.823002	17.674612			
inner2	2min	17.21996	17.295086	17.21996	17.139755	17.706697
	5min	17.21996	17.139755			
outer1	2min	13.33	13.33	13.33	13.33	13.33
	5min	13.33	13.33			
outer2	2min	13.33	13.33	13.33	13.33	13.33
	5min	13.33	13.33			

RNA

	Windows					
	short-term	1	2	3	4	5
inner1	2min	13.323805	13.323805	13.323805	13.323805	13.323805
	5min	13.323805	13.323805			
inner2	2min	13.323805	13.323805	13.323805	13.323805	13.323805
	5min	13.323805	13.323805			
outer1	2min	7.690001	7.690001	7.690001	7.690001	7.690001
	5min	7.690001	7.690001			
outer2	2min	7.690001	7.690001	7.690001	7.690001	7.690001
	5min	7.690001	7.690001			

[그림 2] 네트워크 및 윈도우 로그의 RNA 수치

RNA

	Windows&Network					
	short-term	1	2	3	4	5
inner1	2min	15.5734	15.5734	15.5734	15.53568	15.49921
	5min	15.5734	15.49921			
inner2	2min	15.27188	15.30945	15.27188	15.23178	15.51525
	5min	15.27188	15.23178			
outer1	2min	10.51	10.51	10.51	10.51	10.51
	5min	10.51	10.51			
outer2	2min	10.51	10.51	10.51	10.51	10.51
	5min	10.51	10.51			

ANA

	Windows&Network					
	short-term	1	2	3	4	5
inner1	2min	2.96456	2.913473	2.887148	2.881416	2.881931
	5min	2.925602	2.883561			
inner2	2min	2.683436	2.692613	2.626886	3.24133	3.508211
	5min	2.662558	3.124026			
outer1	2min	3.265597	3.265264	3.240114	3.265533	3.265723
	5min	3.262539	3.258008			
outer2	2min	3.266093	3.266353	3.265926	3.257564	3.26531
	5min	3.266093	3.262498			

[그림 3] 통합된 로그의 RNA 및 ANA 수치

5. 결론

비정상행위 탐지에 관한 기존의 연구들은 주로 네트워크 기반의 단조로운 정보만을 바탕으로 수행되어 왔다. 하지만 한 가지 분야만을 통해 얻어낼 수 있는 정보는 그 종류가 한정되어 있을 뿐 아니라 사용자와 높은 연관성을 보장하기 힘들다. 이를 해결하기 위해 본 논문에서는 운영체제의 사용자 로그 정보와 네트워크 패킷 로그 정보를 함께 사용하여 보다 구체적인 사용자 행위패턴을 바탕으로 정상행위패턴을 생성할 수 있는 방법을 제안하였으며, 이로써 비정상행위 탐지를 위한 더 나은 정확도를 부여하고자 하였다.

구체적인 실제 데이터에 대한 알고리즘의 적용 및 효율성 확인이 추후 선행되어져야 할 연구과제이며, 보다 정교한 정상행위도 계산법과 더욱 다양한 사용자 정보의 통합 사용 역시 앞으로 연구되어져야 할 부분이다.

[참고문헌]

[1] B. Mukherjee, T. L. Heberlein, and K. N. Kevitt, "Network intrusion Detection", IEEE Network, 8(3):26-41, May/June 1994.

[2] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The Architecture of a Network Level Intrusion Detection System", Technical Report, Computer Science Department, University of New Maxico, August 1990.

[3] H.S Javitz, A. Valdes, "The SRI IDES Statistical Anomaly Detector", In Proceedings of the 1991 IEEE Symposium on Research in Security and Privacy, May 1991.

[4] Henry S. Teng, Kaihu Chen, and Stephen C. Lu, "Security Audit Trail Analysis Using Inductively Generated Predictive Rules", In Proceedings of the sixth conference on Artificial intelligence applications, Santa Barbara, California, United States, p. 24-29, January 1990.

[5] Jiawei Han, Micheline Kamber, "Data Mining : Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[6] 오상현, 이원석, "패킷간 연관 관계를 이용한 네트워크 비정상행위 탐지", 한국정보보호학회논문지, 제12권 제5호, 2002.

[7] 박정호, 오상현, 이원석, "데이터베이스 시스템에서 연관 규칙 탐사 기법을 이용한 비정상 행위 탐지", 정보처리학회논문지, 제9-C, 제6호, 2002.