다양한 환경에서의 비정상행위 탐지를 위한 통합 로그 추출 프로그램

신종철, 이종훈, 임선규, 최원섭, 이원석

연세대학교 컴퓨터과학과

Integrated Log Extraction Program for an Anomaly Intrusion Detection in Various Environments

Shin, Jong Cheol, Lee, Jong Hoon, Lim, Seon Kyu, Choi, Won Sub, Lee, Won Suk

Yonsei University

E-mail : ideabell@database.yonsei.ac.kr, saint40@nate.com, capitong@naver.com, icrisis@hotmail.com, leewo@database.yonsei.ac.kr

요 약

최근 정보기술의 발달과 함께 지속적으로 다양해지고 빨라지는 침입 방법에 대처하기 위해 정보를 보호하기 위한 새로운 방법이 요구되고 있는 실정이다. 이를 해결하기 위해 제안된 방법 중 하나가 네트워크 패킷 데이터에 대한 실시간 데이터 스트림 마이닝 알고리즘 기반의 비정상행위 탐지 기법이다. 이는 현재 발생하고 있는 패턴이 기존 패턴과 다를 경우 비정상행위로 간주하고 사용자에게 알려주는 방법으로, 지금까지 없었던 새로운 형태의 침입에도 대처할 수 있는 능동적인 방어법이라고 할 수 있다. 그러나 이 방법에서 네트워크 패킷 데이터 정보만을 통해 얻어낼 수 있는 정보에는 한계가 있다. 따라서, 본 논문에서는 보다 높은 정확도의 비정상행위 판정을 위한 다양한 환경의 로그들을 추출하여 처리에 적합한 형태로 변환하는 전처리 시스템을 제안한다.

1. 서론

최근 정보기술의 발달과 함께 지속적으로 다양해지고 빨라지는 침입 방법에 대처하기 위해 정보를 보호하기 위한 새로운 방법이 요구되고 있는 실정이다. 기존의 방법들은 침입이 진행되는 시간이 길었기 때문에 침입이 일어난 이후에 침입에 대한 정보를 분석하여 대처하는 것이 일반적이었다. 하지만, 최근에 발생하는 대부분의 침입의 유형은 zero-day attack으로 침입이 진행되는 시간이 굉장히 짧아 대처가 이루어지기 전에 모든 공격이 끝나버리기 때문에 기존의 방법으로는 한계가 있는 실정이다. 또한, 과거에 발생하였던 모든 침입 유형을 파악하여 사전에 침입에 대처할 수 있다고 하더라도 새로운 형태의 침입이 발생하였을 경우에는 무방비 상태에 놓이게 될 것이므로 올바른 해결책이 될 수 없다.

이를 해결하기 위해 제안된 방법 중 하나가 네트워크 패킷 데이터에 대한 데이터 마이닝 알고리즘 기반의 비정상행위 탐지 기법[1,2]이다. 이는 사용자의 정상적인 패턴과 현재 발생하고 있는 패턴을 데이터 마이닝법을 이용하여 비교하여 기존 패턴과 다를 경우 비정상행위로 간주하고 사용자에게 알려주는 방법으로, 새로운 형태의 침입에도 대

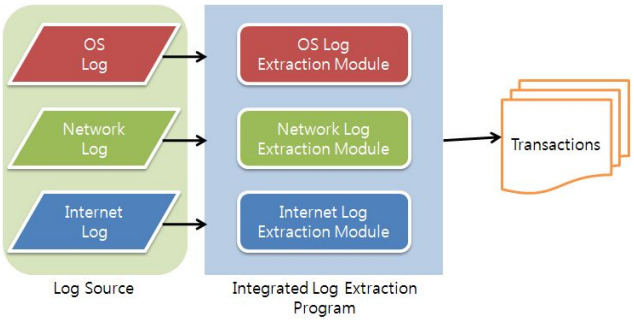
이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 국가지정연구실사업으로 수행된 연구임(No.R0A-2006-000-10225-0)

처할 수 있는 능동적인 방어법이다. 그러나 이 방법은 네트워크 패킷 데이터 정보만을 이용하기 때문에 이를 통해 판단할 수 없는 비정상행위를 탐지할 수 없다는 한계가 있다.

이러한 문제를 극복하기 위해 제안된 논문[3]이 네트워크 로그와 운영체제 로그를 활용한 비정상 행위 탐지법이다. 그러나 다양한 환경에서 발생하는 로그들의 관계를 분석하기 위해서는 동일한 형태로 데이터를 전처리해야 하는데 이 과정에서 소모되는 시간과 노력이 굉장히 크다. 따라서 본 논문에서는 이러한 전처리 단계에서 소모되는 시간과 노력을 최소화하고 기존의 연구[3]에서 보다 높은 정확도를 얻기 위해 네트워크 패킷 로그, 운영체제(Windows Eventlog) 로그 외에도 사용자가 접근하는 인터넷 사이트에 관한 정보인 인터넷접속 기록(Internet Explorer History)이라는 로그를 추가로 수집하여 분석에 유용한 형태로 변환하는 프로그램의 구현 사례를 제시하였다.

2. 본론

본 논문에서 구현한 전체적인 프로그램의 구조는 [그림 1]과 같다.



[그림 1] Integrated Log Extraction Program 구조

Integrated Log Extraction Program은 다양한 환경에서 다양한 형태로 수집된 로그들을 mapping table 등의 요소를 이용하여 다양한 환경에서 발생하는 로그들을 위해 동일한 형태의 트랜잭션으로 변환하는 구조를 가진다. 프로그램은 각각의 로그의 종류마다 스레드의 형태로 추출 모듈이 구성되므로 새로 수집하고자 하는 로그를 유연하게 추가할 수 있는 구조이다.

2.1. 구현환경

본 프로그램은 Windows XP Professional에서 C++를 이용하여 개발되었다. 생성된 트랜잭션을 분석용 서버로 전송하기 위해 데이터의 손실률이 상대적으로 적은 TCP통신을 사용하였으며, File로도 저장하여 비정상 행위 판단에 있어 실시간 데이터 스트림 마이닝 분석뿐만이 아닌 다른 분석방법에도 이용할 수 있도록 하였다.

2.2. Mapping Table

발생되는 로그 데이터 중 가변적인 길이의 문자열 데이터에 대해 처리에 유용한 형태인 고정 길이의 정수형 데이터로 변환해주는 역할을 한다. 예를들면, 인터넷접속 기록 로그에서의 사용자가 방문한 웹페이지 주소나 운영체제 로그에서의 사용자가 접근하는 프로세스의 경로 등이 mapping table을 통해 변환되는 값이다. 웹페이지 주소나 프로세스의 경로는 무한하기 때문에 본 프로그램에서는 사용자가 주로 이용할 만한 것들을 미리 mapping table에 등록해 놓고 이 외의 것들에 대해서는 모두 알려지지 않은 사이트로 보고 특정한 단일 값으로 별도로 처리하였다.

문자열 데이터	정수형 데이터
www.daum.net	2001
media.daum.net	2002
www.naver.com	2003
...	...
알려지지 않은 사이트	2999

[표 1] Mapping Table의 예

2.3. 운영체제 로그 추출 및 변환

운영체제의 로그를 추출하기 위해서 현재 출시되어 있는 개인 PC의 운영체제 중 보급률이 가장 높은 Microsoft사의 Windows XP Professional의 Eventlog의 API를 이용하였다. Eventlog는 Windows 시스템의 프로세스에 대해 발생하는 행위를 기록하는 로그이다.

사용한 Eventlog의 속성에는 윈도우 이벤트의 성공/실패 여부를 판별하는 Event Type(ET), 윈도우 이벤트의 유형을 구분하는 Event Category(EC)와 Event ID(EID), 사용자의 Computer ID(CID), 사용자의 윈도우 시스템 계정에 해당하는 User ID(UID), 사용자의 이벤트로 인해 실행되는 프로세스의 경로 정보(PD)가 포함된다. 이와 같은 정보들은 [표 4]와 같은 스키마를

가지고 [표 2]와 같은 첫째자리 값을 가지는 다섯 자리 정수형 데이터로 변환[그림 2]된다.

Eventlog 속성	Item의 첫째자리 값
Event Type	1
Event Category	2
Event ID	3
Computer ID	4
User ID	5
Process Directory	6

[표 2] 운영체제 로그의 속성별 매핑 값

문자열 데이터	정수형 데이터
성공감사	10000
실패감사	10001
...	...
개체 액세스	20000
...	...
[EventID]	30[EventID]
HOME_PC	40000
...	...
알려지지 않은 컴퓨터	49999
Administrator	50000
SYSTEM	50001
Shin	50002
...	...
알려지지 않은 ID	59999
\explorer.exe	60000
\IEXPLORE.EXE	60001
\notepad.exe	60002
...	...
알려지지 않은 프로세스	69999

[표 3] 운영체제 로그의 Mapping Table 일부

TID	ET	EC	EID	CID	UID	PD

[표 4] 운영체제 로그의 스키마

```
200910222228    10000   20000   30567   49999   59999   60002
200910222228    10000   20000   30562   49999   59999   60000
200910222229    10000   20000   30560   49999   50001   60003
200910222229    10000   20000   30567   49999   50001   60003
```

[그림 2] 변환된 운영체제 로그의 예

2.4. 네트워크 패킷 로그 추출 및 변환

분석에 이용하기 위한 로그의 표본으로 본 프로그램에서는 네트워크 패킷의 Header를 추출하였다. Header에는 각 Packet의 최종 목적지와 Spoofing 공격에서 사용되는 순서번호(Sequence Number)와 같은 정보가 들어있어 Packet의 data보다 보안상 중요한 데이터로 취급되기 때문이다.

네트워크 패킷 로그는 WinPcap Library를 이용하여 추출되었다. WinPcap은 NPF(Netgroup Packet Filter) 드라이버를 통해 들어오는 패킷을 User Level에서 받아 응용프로그램에서 사용할 수 있도록 한다.

현재까지 본 프로그램에서 수집할 수 있는 Protocol은 Mac Address Header, IPv4 Header, TCP / UDP Header이다. 이외에도 ARP 등의 Protocol은 사용자의 선택에 따라 추가로 수집할 수 있도록 구현하였다.

사용한 패킷 로그의 속성으로는 Protocol(PRO), Source IP(SIP), Destination IP(DIP), Source Port(SPort), Destination Port(DPort)가 있다. 이와 같은 정보들은 [표 7]과 같은 스키마를 가지고 [표 5]과 같은 첫째자리 값을 가지는 여섯자리 정수형 데이터로 변환[그림 3]된다. 단, IP 값의 경우 길이가 매우 길기 때문에 ntohs() 함수를 이용하여 호스트 오더 방식으로 변환하여 출력하였다.

패킷 로그 속성	Item의 첫째자리 값
Protocol	1
Source Port	2
Destination Port	3

[표 5] 네트워크 패킷 로그의 속성별 매핑 값

문자열 데이터	정수형 데이터
TCP	100006
UDP	100017
...	...
알려지지 않은 프로토콜	199999
[SPort]	2[SPort]
[DPort]	3[DPort]

[표 6] 네트워크 패킷 로그의 Mapping Table 일부

TID	PRO	SIP	DIP	SPort	DPort

[표 7] 네트워크 패킷 로그의 스키마

```
200910222147    100006  3554410256      2776922391      252286  319101
200910222147    100006  2776922391      3554410256      219101  352286
200910222147    100006  2776922391      3554410256      219101  352286
200910222147    100006  2776922391      3554410256      219101  352286
200910222147    100006  2776922391      3554410256      219101  352286
```

[그림 3] 변환된 네트워크 패킷 로그의 예

2.5. 인터넷접속 기록 로그 추출 및 변환

인터넷접속 기록 로그의 추출을 위해 현재 웹브라우저 중 가장 널리 보급되어 있는 Internet Explorer의 History를 로그로 이용하였다. 대부분의 PC 이용자들이 Internet Explorer를 통해 인터넷을 이용하고 있기 때문에 운영체제 로그를 통한

사용자의 시스템 사용 패턴이나 네트워크 패킷 로그를 통한 사용자의 네트워크 사용 패턴 이외에도 인터넷접속 기록 로그는 사용자의 비정상행위를 판단하는데에 있어 중요한 역할을 하게 된다. Internet Explorer의 History는 IUrlHistoryStg Interface를 이용하여 추출하였고, 이로부터 인터넷 접속 기록 로그를 얻어내었다.

사용한 인터넷접속 기록 로그의 속성으로는 Protocol(PRO), 사이트 주소(Domain), 세부 사이트 주소(SDomain)가 있다. 이와 같은 정보들은 [표 10]과 같은 스키마를 가지고 [표 8]과 같은 첫째자리 값을 가지는 네자리 정수형 데이터로 변환[그림 4]된다.

인터넷접속 기록 로그 속성	Item의 첫째자리 값
Protocol	1
Domain	2
Sub Domain	3

[표 8] 인터넷접속 기록 로그의 속성별 매핑 값

문자열 데이터	정수형 데이터
http	1001
ftp	1002
...	...
알려지지 않은 프로토콜	1009
www.daum.net	2001
media.daum.net	2002
www.naver.com	2003
news.naver.com	2004
mail2.naver.com	2005
...	...
알려지지 않은 사이트	2999
(null)	3001
...	...
/index.nhn	3004
...	...
알려지지 않은 사이트	3999

[표 9] 인터넷접속 기록 로그의 Mapping Table 일부

TID	PRO	Domain	SDomain

[표 10] 인터넷접속 기록 로그의 스키마

```
200910221302    1001    2005    3004
200910221302    1001    2003    3001
200910221302    1001    2003    3001
200910221302    1001    2006    3005
200910221302    1001    2001    3001
```

[그림 4] 변환된 인터넷접속 기록 로그의 예

2.6 병합 분석

앞서 동일한 형태로 변환된 로그들은 TID라는 공통된 속성을 통해 비정상행위 분석에 사용할 로그를 선택하여 해당 로그들을 병합함으로써 통합된 환경에서의 비정상행위 분석이 가능해진다.

200910222228	10000	20000	30562		
49999	59999	60001	1001	2001	3001

익명의 PC(59999)를 사용하는 익명의 사용자(49999)는 인터넷 익스플로러라는 웹 브라우저(60001)의 개체에 대한 접근(20000, 30562)을 통해 http 프로토콜(1001)로 www.daum.net(2001)이라는 웹사이트에 접근하는 것을 2009년 10월 22일 22시 28분에 성공적으로 수행(10000)하였다.

[그림 5] 병합된 로그 트랜잭션의 예

운영체제 로그와 인터넷 접속 로그가 병합된 트랜잭션의 경우 하나의 로그만을 단독으로 사용했을 때에 비해 훨씬 더 많은 정보를 비정상행위 판단에 활용할 수 있게 된다.

3. 결론

호스트 PC에서의 사용자의 행위 분석을 가능하게 하는 운영체제 로그와 인터넷접속 기록 로그를 네트워크 패킷 로그와 함께 전처리하는 본 프로그램을 통해 네트워크 패킷 정보만을 가지고 이루어지던 기존의 비정상행위 분석의 한계를 어느 정도 극복할 수 있는 환경이 만들어졌다. 따라서, 본 프로그램을 통해 전처리된 형태로 추출된 로그 데이터를 이용하여 보다 효과적인 비정상행위 탐지에 사용할 수 있으며, 데이터들은 동일한 형태로 변환되고 트랜잭션화 되어 있기 때문에 여러가지 데이터 마이닝의 방법을 이용한 비정상 행위 분석기술들에서도 별도의 처리 과정을 거치지 않고 간단하고 유연하게 적용할 수 있다.

그러나 앞선 문자열 데이터의 매핑 방법들은 지극히 직관적인 방법이다. 문자열 데이터의 개수는 대부분의 로그에서 무한하게 나타나기 때문에 좀 더 효과적인 매핑 방법이 제시되어야 할 것이다.

[참고문헌]

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

[2] 패킷간 연관 관계를 이용한 네트워크 비정상행

위 탐지, 오상현, 이원석, 한국정보보호학회논문지, 제12권 제5호, 2002.
[3] 사용자 로그의 분석을 통한 실시간 비정상행위 탐지 기술, 김명수, 신종철, 정재명, 고유선, 이원석, 2009 한국IT서비스학회 춘계학술대회, 2009.