연관 규칙을 이용한 네트워크 기반 침입 탐지 패턴생성 기술

소 진, 이 상훈
국방대학교 국방과학대학원 전산정보학과
jso@kndu.ac.kr, hoony@kndu.ac.kr

Pattern Generation Technique for Network-based Intrusion Detection using Association Rules

Jin Soh, Sang-Hoon Lee
Dept. of Computer Science, Korea National Defense University

요 약

네트워크 기반 컴퓨터 시스템은 현대사회에 있어서 매우 중요한 역할을 담당하고 있기 때문에 이들을 정보 범죄들로부터 안정적이면서 효율적인 환경을 제공하는 것은 매우 중요한 일이다. 현재의 침입탐지 시스템은 네트워크 상에서 지속적으로 처리되는 대량의 패킷에 대하여 탐지속도가 떨어지고, 새로운 침입유형에 대한 대응방법이나 인지능력에도 한계가 있기 때문이다. 따라서 다양한 트래픽 속에서 탐지율을 높이고 탐지속도를 개선하기 위한 방안이 필요하다.

본 논문에서는 침입탐지 능력을 개선하기 위해 먼저, 광범위한 침입항목들에 대한 탐지 적용기술을 학습하고, 데이터 마이닝 기법을 이용하여 침입패턴 인식능력 및 새로운 패턴을 생성하는 적용기술을 제안하고자 한다. 침입 패턴생성을 위해 각 네트워크에 돌아다니는 관련된 패킷 정보와 호스트 세션에 기록되어진 자료를 필터링하고, 각종 로그 화일을 추출하는 프로그램들을 활용하여 침입과 일반적인 행동들을 분류하여 규칙들을 생성하였다. 마이닝 기법으로는 학습된 항목들에 대한 연관 규칙을 찾기 위한 연역적 알고리즘을 이용하였다. 또한, 추출 분석된 자료는 리눅스기반의 환경 하에서 다양하게 모아진 네트워크 로그파일들을 본 논문에서 제안한 방법에 따라 적용한 결과이다.

1. 서 론

최근 컴퓨터 및 네트워크의 환경변화에 따라 기존에 운영되고 있는 침입탐지 시스템(이하 IDS라 칭함)은 새로운 침입 유형에 대한 탐지방법을 개선할 필요성이 요구되고 있다. 현재의 침입탐지 시스템은 네트워크 상에서 지속적으로 처리되는 대량의 패킷에 대하여 탐지속도가 떨어지고, 새로운 침입유형에 대한 대응방법이나 인지능력에도 한계가 있기 때문이다. 따라서 다양한 트래픽 속에서 탐지율을 높이고 탐지속도를 개선하기 위한 방안이 필요하게 된 것이다. 수많은 패킷들로 구성되어 송·수신되는 정보들은 일반 사용자나 불법 접속자 모두 동일한 경로와 패킷을 사용하기 때문에 불법 접속자들의 침입목적과 패킷이 어떤 패킷이고, 또 어떤 유형의 패킷이 침입에 사용되는 지를 알아야만 한다[1].

본 논문에서는 침입탐지능력을 향상시키기 위한 방법으로 대량의 분산된 데이터 집합을 재사용하는 데이터 마이닝 기법을 제안한다. 데이터 마이닝 기법 중에서는 대량의 데이터들의 상관관계를 처리하여 불규칙성속에서 새로운 특징을 추출하는 연관 규칙(association rule) 기법을 이용하였다. 이 기법은 데이터 간의 상호관련을 통한 알려지지 않은 규칙을 발견하는 방법으로 네트워크 데이터에서 공통적으로 관측된 규칙 집합들을 생성하고 보이지 않은 공격에 대한 탐지를 예측할 수 있게 한다. 이렇게 예측을 위한 훈련된 학습규칙들과 새로운 특징패턴들을 기반으로 침입 탐지시스템을 지원함으로서 대량의 트래픽 데이터 속에서 탐지속도를 증가시키고 새로운 침입유형에 대한 대응능력을 향상시킬 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 마이닝에 필요한 침입탐지 데이터의 구성요소들을 제시한다. 3장은 여러 가지 데이터 마이닝 기법들을 간략하게 설명하고 연관 규칙을 이용하여 빈번한 침입과 정상 행동패턴들을 생성할 수 있는지를 설명한다. 4 장은 결론과 미래 연구방향에 대해 제시한다.

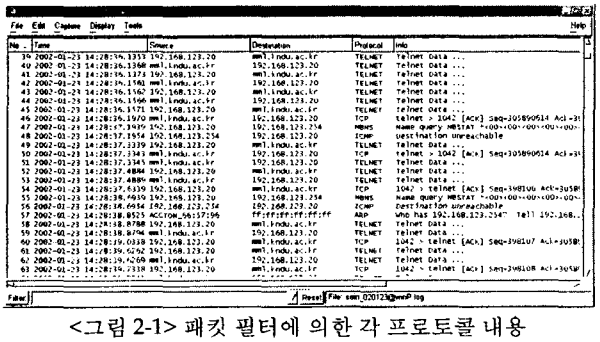
2. 침입 탐지를 위한 데이터

침입탐지시스템을 개발하는데 있어서 가장 중요한 것은 현재 네트워크의 명확한 데이터들 중에서 특징을 가지고 있는 패턴규칙집합의 초기화작업이다. 차후 이 규칙은 추가, 삭제되고 수정되어지면서 정확성을 유지하고 학습되는데 이러한 변화과정에서 추출되는 규칙집합들은 초기 규칙집합과는 분명히 차이가 있다는 점이다. 또한, 데이터 레코드는 많은 속성을 가진다. TCPDUMP 레벨에서 추출된 데이터 형태로는 <표 1>과 같이 Ping 정보, TCP 및 ICMP 패킷에 대한 정보에 대한 자원으로 규칙을 생성한다.

<표 1> 패킷 필터에 의한 추출된 레코드 내용[2]

식별자	내용 및 의미
ipoption	IP 옵션필드 코드값
content	패킷의 payload 내용 패턴
offset	패턴매칭이 시작되는 주소값(hex)
depth	패턴매칭이 끝나는 주소값(hex)
ttl	IP 헤더의 TTL 필드 값
tos	IP 헤더의 TOS 필드 값
id	IP 헤더의 특정값에 대한 fragment) ID 필드 값
fragbits	IP 헤더의 fragmentation bits
dsize	패킷 값에 대한 payload 크기
flags	어떤 값들에 대한 TCP flags
seq	TCP sequence number 필드 값
ack	TCP acknowledgement 필드 값
itype	ICMP type 필드 값
icode	ICMP code 필드 값
icmp_id	ICMP ECHO ID 필드 값
icmp_seq	ICMP ECHO sequence number 값
session	주어진 세션에 대한 응용층 정보 자료
resp	적극적 반응(knock down connections, etc)
react	적극적 반응(block web sites)
uricontent	URI 부분에서 패킷패턴 정보
ip_proto	IP 해더의 프로토콜 값
sameip	만약 원천지 IP와 목적지 IP가 같은지를 결정

그밖에 패킷에서 얻을 수 있는 정보는 소스 및 목적지 IP주소, Port 번호, Date/Time, 전송프로토콜(TCP,UDP,ICMP,etc), 트래픽 기간(Traffice Duration) 등이다. 패킷 필터 추출 프로그램으로는 Tcpdump, Tcptrace, libpcap 등이 있다. <그림 2-1>과 같이 패킷 필터에서 생성한 각 프로토콜 서비스에 대한 항목들을 살펴보면 다음과 같다.



<그림 2-1> 패킷 필터에 의한 각 프로토콜 내용

3. 연관 규칙를 이용한 패턴 생성

3.1 연관규칙 정의

규칙을 공급하는 센서(sensor) 역할을 하는 연관 규칙은 일반적으로 패킷과 연결상태(connection)레벨에서 처리하는 네트워크 트래픽에 대한 관계를 생성하는 기능을 한다. 주로 데이터베이스 테이블로부터 다중 특징(속성들:Attributes)에 대한 상호관계들을 찾는데 이용된다. 주어진 레코드 집합에서 각 레코드는 항목들의 집합(항목집합:A set of item)으로 표현되며, 연관규칙은 다음과 같이 표현한다. $X \rightarrow Y, [s, c]$. X와 Y는 항목집합이고 $X \cap Y = \emptyset, s = support\ (X \cup Y)$ 으로 계산되며 s는 support 규칙이다. 그리고 $c = \frac{support(X \cup Y)}{support(X)}$ 으로 계산되며, c는 confidence 규칙이다[4].

X와 Y는 규칙 항목집합으로서 데이터 베이스 속에 속성들(attributes)의 집합이 된다. 백분율 s %과 c %은 상투적으로 규칙들에 대해 support과 confidence로 표현한다. support 규칙은 X와 Y가 동시에 포함하는 레코드들의 비율을 말하고 confidence 규칙은 Y를 포함하는 X의 레코드들의 비율을 말한다. 보통 미리 정의된 데이터 집합은 지속적으로 변화된다.

3.2 연역적 알고리즘(Apriori Algorithm)[5]

생성된 행동 패턴은 각 항목 레코드 데이터베이스에 저장하여 규칙을 생성하기 위한 항목집합들의 패턴의 수를 찾는다. 이러한 과정에서 지원(Support)규칙의 최소 값과 신뢰(Confidence)규칙을 계산한다. 즉 빈번하게 발생하는 항목집합들에 대한 행동패턴의 수를 백분율로 나타내는 지원 및 신뢰규칙을 생성하여 저장한다.

물론, 모든 규칙은 관련된 항목집합끼리 연관시켜야한다. 다음 나타낸 <표 2>는 대량의 Unix 명령어 항목집합에 대한 연관규칙 생성 및 절차 예를 보여준다.

$vi \rightarrow last$ 규칙은 지원규칙 $Support(\{ vi, last \}) = 50\%$, Confidence = $Support(\{ vi, last \}) / Support(\{ vi \}) = 66\%$ 으로 계산된다. 즉, 계산결과에 의거하여 $\{ vi \}$를 선택하면 $\{ last \}$ 명령을 사용한다는 행동패턴 지원율이 50%이면 신뢰성은 66%라고 정의한다는 의미이다.

<표 2> 지원 및 신뢰규칙을 생성하는 절차

'#	항목집합	항목집합*	지원 규칙율
1	{ *vi* , *time*, *last* }	{ *vi* }	75%
2	{ *vi* , *last* }	{ *time* }	50%
3	{ *vi* , *rlogin* }	{ *last* }	50%
4	{ *time* , *ls*, *history* }	{ *vi* , *last* }	50%

다음은 대량의 데이터 항목집합에서 빈번한 항목집합을 생성하는 절차는 다음과 같다.

<표 3> 연역적 알고리즘(Aprioir Algorithm)

```
procedure Apriori Algorithm()  begin
  L1 := {frequent 1-itemsets};
  for ( k := 2; Lk-1 0; k++ ) do {
      Ck= apriori-gen(Lk-1) ; // new candidates
  for all transactions t in the dataset do {
      for all candidates c Ck contained in t do
            c:count++  }
      Lk = { c Ck | c:count >= min-support }
  }
  Answer := k Lk
end
```

알고리즘의 목적은 빈번한 항목집합을 찾기 위해 사용된다. 실행 절차는 다음과 같다. 첫 번째 단계에서, 빈번한 1-항목집합(초기 항목집합:L1) 결정하기 위하여 단순히 항목사건들의 갯수를 계산한다. 다음단계, 즉 단계 k는 2부터 증가하며 구성된다. 단계 k에서는 우선, (k-1)번째 단계에서 찾아낸 빈번한 항목집합 Lk-1은 후보 항목집합 Ck을 생성하는데 사용한다.

이 함수의 역할은 첫번째, Lk-1과 정렬된 첫번째 k-2 항목의 동일한 조건을 가진 Lk-1을조인한다. 두 번째, Lk-1에 없는 (k-1)부분집합을 가진 조인된 결과로부터 모든 항목집합들을 삭제하고 Ck을 생성한다.

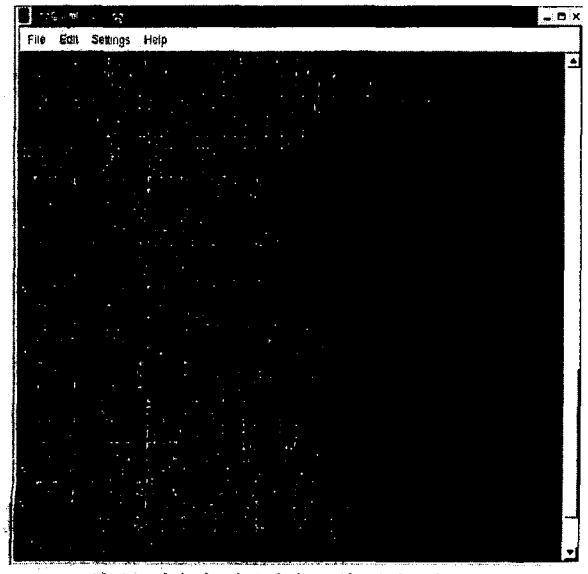
각 처리마다, 해쉬트리 자료 구조를 이용하여 Ck에서 각 처리를 포함한 후보들을 결정하고 그 후보집합들의 수를 증가시킨다. 마지막 단계에서, Ck는 어떤 후보가 빈번한지를 결정하기 위해서 검사한 후, Lk를 생성한다. 이 알고리듬은 Lk가 공집합이 되면 종료한다. 자세한 알고리즘은 <표 3>에 기술하였다.

3.3 패턴 규칙 생성

모든 규칙은 패킷형태의 조합으로 나타낸다. 예를 들면, ICMP, TCP, UDP 등의 조합으로 표현될 수 있다. 이는 패킷 크기, 소스와 목적지 포트 개념을 포함한다. 잘 알려진 포트일 경우 계층적 관리에 따라 저장된다. 일반적으로 각 일련의 연속적인 패턴(sequential pattern)의 규칙집합을 구별하는 방법은 그 속성 중에서 가장 식별하기 쉽고 정확하고 정밀한 속성이나 값에 따라 규칙을 설정할 수 있다.

<그림 1>과 <그림 2>에서는 연관 규칙을 이용하여 데이터 로그파일(logfile)을 입력을 받아 연역적 알고리즘에 따라 규칙을 생성(Generating Rules)한 결과를 나타낸다.

<그림 1> 패턴 규칙 프로그램 실행 화면



<그림 2> 사용자 행동패턴 규칙 생성(예: x.rule).

4. 결론 및 향후 연구

연관 규칙을 이용한 네트워크 패킷 분석과 침입 유형에 따른 실시간 네트워크 기반 탐지 기법들을 대해 연구하였다. 리눅스 시스템 환경 하에서 TCPDUMP를 이용하여 IDS시스템 로그파일을 수집하였으며, 침입에 대한 탐지를 어떻게 하면 실시간으로 빠르게 인식할 수 있는지에 대한 연구로 데이터 마이닝 기법의 연관 규칙을 적용하였다. 또한 새로운 침입유형을 탐지하기 위한 방법과 알고리즘을 제시하였다.

새로운 공격패턴과 예측 불가능한 네트워크 트래픽 패턴에 대한 연구는 자동화 생성을 목적으로 대량의 네트워크 패킷 데이터들 속에서 패턴 식별자와 값을 분석한다. 향후 이러한 데이터들의 속성을 광범위하게 분석하여 학습에 따른 패턴 분류 알고리즘을 연구하고, 이를 통해 새로운 규칙들을 생성하여 침입탐지의 대한 적중율을 높이고, 좀더 빠른 탐지속도 개선을 통하여 시스템 효율성 증대에 대한 연구가 요구된다.

참고문헌

[1] 이 경하 외, "네트워크 패킷 정보를 기반으로 한 보안 관리", 한국정보과학회 논문지, Vol.25, No.12, pp.1405-1412, Dec. 1998

[2] http://www.snort.org/

[3] http://www.tcpdump.org/

[4] Intrusion Detection Systems, http://www.cerias.purdue.edu/coast/intrusion-detection/ids.html

[5] Eric Bloedorn, Alan D. Christiansen,외 "Data Mining for Network Intrusion Detection: How to Get Started," The MITRE Corporation, In http://www.afcea.org/pastevents/db2001/Bloedorn_files/ frame.htm, 2001

[6] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Models ," IEEE Symposium on Security and Privacy, In http://citeseer.nj.nec.com/154973.html, 1999

[7] Karuna Pande Joshi, "Analysis of Data Mining Algorithms." In http://userpages.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm#apriori, 1997