# 연관법칙 마이닝(Association Rule Mining)을 이용한 ANIDS(Advanced Network Based IDS) 설계

정은희* · 이병관**

## ANIDS(Advanced Network Based Intrusion Detection System) Design Using Association Rule Mining

Eun-hee Jeong* · Byung-kwan Lee**


### 요 약

제안한 ANIDS(Advanced Network based IDS)는 네트워크 패킷을 수집하여 연관규칙 마이닝 기법을 이용하여 패킷의 연관성을 분석하고, 연관성이 높은 패킷을 이용해 패턴 그래프를 생성한 후, 생성된 패턴 그래프를 이용해 침입인지를 판단하는 네트워크 기반 침입 탐지 시스템이다. ANIDS는 패킷 수집 및 관리하는 PMM(Packet Management Module), 연관성 있는 패킷들만을 이용해 패턴 그래프를 생성하는 PGGM (Pattern Graph Generate Module), 침입을 탐지하는 IDM(Intrusion Detection Module)으로 구성된다. 특히, PGGM은 Apriori 알고리즘을 이용해 $Sup_{min}$보다 큰 연관규칙의 후보 패킷을 찾은 후, 연관규칙의 신뢰도를 측정하여 최소 신뢰도 $Conf_{min}$보다 큰 연관규칙의 패턴 그래프를 생성한다. ANIDS는 패킷간의 연관성을 분석하여 침입인지를 탐지할 수 있는 패턴 그래프를 사용함으로써, 침입 탐지의 긍정적 결함 오류를 감소시킬 수 있으며, 완벽한 패턴 그래프 패턴이 생성되기 전에, 이미 침입으로 판정된 패턴 그래프 패턴과 비교하여 유사한 패턴 형태를 침입으로 간주하므로 기존의 침입 탐지 시스템에 비해 침입 탐지 속도를 감소시키고 침입 탐지율을 증가시킬 수 있다.



### ABSTRACT

The proposed ANIDS(Advanced Network Intrusion Detection System) which is network-based IDS using Association Rule Mining, collects the packets on the network, analyze the associations of the packets, generates the pattern graph by using the highly associated packets using Association Rule Mining, and detects the intrusion by using the generated pattern graph. ANIDS consists of PMM(Packet Management Module) collecting and managing packets, PGGM(Pattern Graph Generate Module) generating pattern graphs, and IDM(Intrusion Detection Module) detecting intrusions. Specially, PGGM finds the candidate packets of Association Rule large than $Sup_{min}$ using Apriori algorithm, measures the Confidence of Association Rule, and generates pattern graph of association rules large than $Conf_{min}$. ANIDS reduces the false positive by using pattern graph even before finalizing the new pattern graph, the pattern graph which is being generated is compared with the existing one stored in DB. If they are the same, we can estimate it is an intrusion. Therefore, this paper can reduce the speed of intrusion detection and the false positive and increase the detection ratio of intrusion.



키워드

침입탐지시스템, NIDS, 데이터마이닝, 연관규칙, Support, Confidence, 패턴 그래프



* 강원대학교 지역경제학과

** 관동대학교 컴퓨터학과


접수일자 : 2007. 6. 20

## Ⅰ. 서 론

컴퓨터의 급속한 발전과 초고속 통신의 보급으로 수많은 사람들이 네트워크를 이용하여 정보를 접하는 일이 많아졌다. 그로 인해 정보의 가치가 증대되었으나 개인 정보의 누출 위험과 함께 바이러스, 인터넷 웜, 해킹 등의 위험에 노출되고 그에 따른 악영향이 증가하고 있다. 이러한 문제점들을 해결하기 위해 다양한 보안 시스템이 개발되고 있으며, 침입탐지시스템도 그 일부라고 볼 수 있다.

특히, 네트워크 기반 침입 탐지 시스템(Network-based Intrusion Detection System)은 침입내용, 침입시간, 침입유형 등 다양한 침입기법이 기록된 로그정보를 이용하는 호스트 기반 침입탐지 시스템(Host- based Intrusion Detection System)보다 빠르게 탐지할 수 있어서 실시간 처리와 내·외부 침입 흔적에도 탐지가 가능한 역추적 시스템으로 크게 부각되고 있지만, 실시간 침입 차단을 하지 못하는 문제점을 가진다.

본 논문에서는 연관 규칙 마이닝을 이용한 네트워크 기반 침입탐지 시스템인 ANIDS를 제안한다. ANIDS는 수집한 네트워크 패킷간의 연관성을 분석한 후, 연관규칙의 신뢰도를 활용하여 침입 패턴인 패턴 그래프를 생성하고, 생성된 패턴 그래프를 이용해 침입을 탐지하는 시스템이다. 따라서, ANIDS는 연관 규칙의 패턴 그래프를 이용함으로써 침입 탐지시에 잘못된 긍정적 결함 오류 발생을 줄이고 패턴 검출을 용이하게 함으로써 침입탐지 시간을 단축시키고 침입 탐지율을 향상시킬 수 있다.

## Ⅱ. 관련연구

### 2.1 침입탐지시스템

침입 탐지 시스템이란 컴퓨터 시스템의 비정상적인 사용, 오용, 남용 등을 실시간으로 탐지하고, 이에 대응하는 기능을 가진 소프트웨어이다. 침입 탐지 시스템은 분석 자료에 따라 호스트 기반 침입 탐지 시스템과 네트워크 기반 침입 탐지 시스템으로 분류할 수 있다.

호스트 기반 침입 탐지 시스템은 시스템 로그 파일, 처리기 상태 및 시스템 호출 명령어 등의 실행과 같은 호스트 내부의 정보를 침입 탐지에 이용한다. 이 경우 시스템 침입에 성공 및 실패 여부도 탐지가 가능하고 시스템 침입에 성공한 침입자의 활동에 대한 추적도 가능하지만, 너무 시스템에 의존한다는 단점이 있으며, 단일 호스트 기반과 다중 호스트 기반으로 나눌 수 있다.

네트워크 기반 침입 탐지 시스템은 네트워크 패킷 헤더 및 데이터를 분석하거나 패킷 트래픽량 등을 분석하여 침입인지를 판단하며, 네트워크 패킷을 캡쳐하기 위해 네트워크 카드를 무차별 모드(Promiscuous mode)로 설정해 놓아야 한다. 네트워크 기반 침입 탐지 시스템의 패킷 수집 방법은 운영체제마다 다르며, 그 방법은 표 1과 같다[1].

표 1. 패킷 수집 방법
Table. 1 The method of Packet collection

| 운영체제 | 패킷 수집 방법 |
|---|---|
| BSD계열 | BPF(Berkely Packet Filter) |
| SunOS계열 | NIT(Network Interface Tap) |
| Windows계열 | SMS(Systems Management Server) |
| 기타 | Libpcap |

### 2.2 침입탐지시스템의 한계

침입 탐지 시스템은 탐지 위주의 메커니즘 설계로 인해 몇 가지 한계점을 가지고 있다[2].

첫째, 침입 행위가 늘어나면서 네트워크 기반 IDS나 호스트 기반 IDS가 제한된 탐지 능력으로 공격 시도들에 대해 적절하게 구분해 내기가 어려워짐에 따라 False positive와 Miss Detection 문제가 증가한다.

둘째, 네트워크 기반 IDS는 네트워크 상에 있는 패킷들을 감지하지만 차단하지 못하기 때문에, 대부분의 패킷은 네트워크 기반 IDS가 파악하기 전에 이미 공격하고자 하는 목적지에 도달하여 네트워크 기반 IDS가 침입을 판별하기 전에 침입에 성공하므로 실시간 공격을 막을 수 없다.

### 2.3 데이터 마이닝

데이터 마이닝이란 대량의 데이터 사이에 묻혀 있는 패턴을 발견하고 규칙을 추론함으로써, 의사결정을 지원하고 그 효과를 예측하기 위한 기법이다. 즉, 데이터베이스 안으로 깊숙히 침투하여, 데이터 내의 패턴을 발견하고 규칙을 추론한다. 이러한 패턴과 규칙은 의사결정을 지원하고 기업 환경의 변화를 예측하는 데 사용될 수 있다[3,4].

2.3.1 데이터 마이닝 기법

데이터 마이닝을 통해 얻을 수 있는 정보의 형태는 매우 다양하며, 이에 따라 다양한 기법이 존재한다. 가장 대표적인 데이터 마이닝 기법으로는 사건들의 연관성(associations) 탐사, 연속성(sequence)탐사, 분류(classifications)규칙 탐사와 군집 구분(clustering)을 들 수 있다.

■ 연관성(association) 탐사

연관성은 동시에 발생하는 사건 그룹 내에서 사건들 사이에 존재하는 친화성 혹은 패턴을 R:X→Y형식으로 표현한다. 예를 들어 IBM의 '시장바구니 분석(Market Basket Analysis)'을 들 수 있는데, 이것은 슈퍼마켓에서 소비자들이 구입한 물품들의 목록을 분석함으로써 콘칩이 구매되는 경우의 50%는 소비자들이 콜라도 함께 구매한다는 것과 같은 패턴을 발견하는 것이다.

■ 연속성(sequence) 탐사

연관성 탐사의 변형이라고 할 수 있으며, 사건들이 시간적인 관계를 가지는 것을 말한다. 예를 들어 배낭을 구입한 고객은 다음에 텐트를 구입하는 경향이 있는 경우, 배낭의 구입과 텐트의 구입에는 연속성이 존재한다.

■ 분류(classifications) 규칙 탐사

분류 규칙 탐사는 어떤 항목이 속하는 그룹의 특성을 가장 잘 나타낼 수 있는 특징들을 발견하는 것으로, 가장 활발히 연구가 이루어지고 있는 분야이다. 많은 기업들이 안고 있는 공통적인 문제 중의 하나는 단골고객의 이탈이다. 이탈하는 고객들과 그렇지 않은 고객들의 차이를 발견함으로써, 기업은 어떠한 고객이 앞으로 이탈할 가능성이 있는지 예측할 수 있으며, 고객유지 및 유치 전략을 보다 효과적으로 개발할 수 있을 것이다.

■ 군집(clustering) 구분

군집 구분은 분류 규칙 탐사와 관련되어 있으나, 어떠한 그룹도 사전에 정의되어 있지 않다는 점에서 분류 규칙의 탐사와 다르다. 군집 구분은 데이터 내에 존재하는 상이한 그룹을 구분해 내는 기법이다.

2.3.2 연관규칙

연관규칙은 데이터베이스에 있는 아이템간의 관계를 수치적으로 계산하여 어떤 이벤트가 발생되었을 때 아이템들간의 상관관계를 분석하는 방법으로, 침입탐지 시스템에서는 감사데이터들의 속성에 대한 상호 연관성을 결정하는데 사용한다.

연관규칙 R은 조건부와 결과부로 구성되며 항목 집합인 X와 Y에 대하여 'X가 일어나면 Y도 일어난다'는 의미로 다음과 같이 표현한다.

$$R : X \Rightarrow Y \qquad \text{[수식 1]}$$

여기서 X와 Y는 각각 속성들의 집합이고, $X \cap Y = \phi$ 이어야 한다.

각 속성집합들에 대한 지지도(support)는 속성들이 집합 X를 포함하고 있는 레코드들의 비율을 의미하며 support(X), support(Y)라고 표현한다. 그리고, 이 규칙의 신뢰도( confidence) 계산식은 수식 2와 같다.

$$confidence(R) = \frac{support(X \cup Y)}{support(X)} \qquad \text{[수식 2]}$$

연관규칙을 탐사하는 가장 대표적인 알고리즘으로 Apriori 알고리즘으로 두 단계로 구성된다[5].

단계 1 : 미리 결정된 최소 지지도 $S_{\min}$을 만족하는 빈발 항목 집합(large item sets)을 찾는다.

단계 2 : 빈발 항목 집합 L에 대한 부분집합 A를 찾는다. 미리 결정된 최소 신뢰도 $C_{\min}$에 대하여 support(L)/support(A) ≥ $C_{\min}$이면, 연관규칙 R:A⇒(L-A)형태의 규칙을 출력한다. 즉, 이 규칙의 지지도는 support(R) = support(L)이며, 신뢰도는 confidence(R) = support(L)/support(A)가 된다.

## Ⅲ. ANIDS 설계

데이터 마이닝을 이용한 네트워크 기반 침입탐지 시스템 ANIDS는 패킷 수집 및 관리하는 PMM(Packet Management Module), 패킷간의 연관성을 분석하여, 패턴 그래프를 생성하는 PGGM( Pattern Graph Generate Module), 침입을 탐지하는 IDM(Intrusion Detection Module)으로 구성된다.

그림 1은 ANIDS의 각 모듈간의 데이터 흐름과 관계를 설명한 것이다

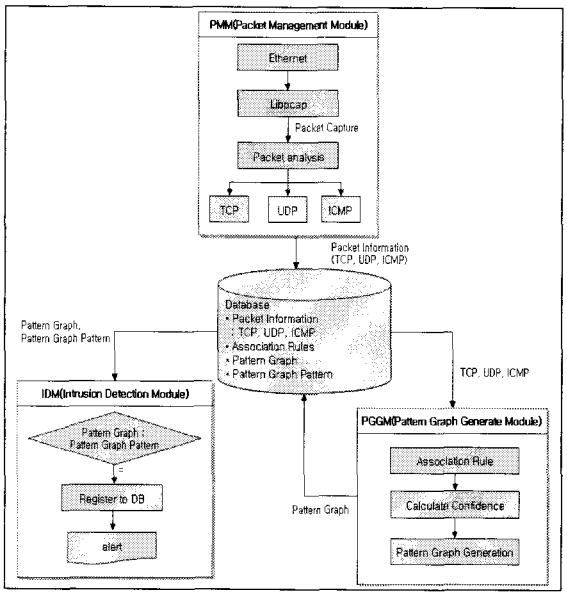



그림 1. ANIDS 시스템
Fig. 1 ANIDS System

### 3.1 PMM(Packet Management Module)

수집된 패킷들은 이더넷 프레임 전체로 받기 때문에 수집된 패킷들을 유형별로 분류를 해야 한다. PMM은 패킷을 유형별로 분류하는 모듈로서 수집된 패킷을 TCP, IP, UDP 등으로 패킷 분류하고, ANIDS에서 사용할 패킷을 축출한 후, 패킷을 ANIDS 시스템에서 요구하는 형식으로 변환시킨다. 이렇게 분류되고 변환된 패킷들은 TCP-t, UDP_t, ICMP_t 테이블에 유형별로 저장되어 패킷 연관성 분석에 사용된다.

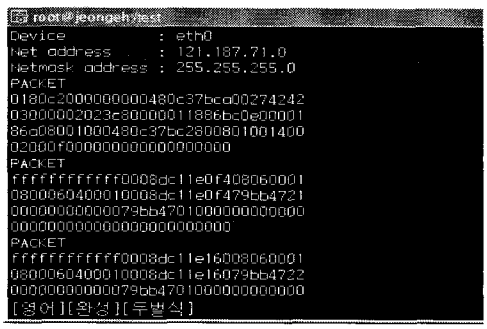



그림 2. PMM에 의해 수집된 패킷 데이터
Fig. 2 Collected Packet data by PMM

그림 2는 PMM에 의해 수집된 패킷들을 보여주는 그림이고, 그림 3은 수집된 패킷을 분석하여 이더넷 헤더의 포인터로 넘긴 다음 IP 프로토콜일 경우 IP 헤더의 포인터로 넘긴 후 각각의 프로토콜별로 분리하는 알고리즘이다[5].

```
packet_analysis(packet){
 packet_h: packet header
 h_length : length of header
 tcp_h : TCP header
 udp_h : UDP header
 icmp_h : ICMP header

 if (packet_h == TCP){ //TCP 프로토콜일 때
  tcp_h = packet_h;
  tcp_data = extract(packet, tcp);
        :
 }
 else if (packet_h ==UDP){ //UDP 프로토콜일 때
  udp_h = packet_h;
  udp_data = extract(packet, udp);
        :
 }
 else (packet_h == ICMP){ //ICMP 프로토콜일 때
  icmp_h = packet_h;
  icmp_data = extract(packet, icmp);
        :
 }
}
```

그림 3. 패킷 분석 알고리즘
Fig. 3 Packet Analysis Algorithm

그림 4는 PMM에서 생성된 패킷 정보들을 이용해 패킷간의 연관성을 분석하는 PAA(packet association analysis) 알고리즘으로 발신자 IP, 목적지 IP와 포트번호, 발생시간의 관련성을 분석하여 연관성 있는 패킷들을 분리하여 발생시간 순으로 정렬하여 데이터베이스에 저장한다.

즉, 패킷 정보를 종류별로 저장하고 있는 TCP_t, UDP_t, ICMP_t 테이블에 저장되어 있는 패킷을 순차적으로 읽은 후, 각각의 테이블에 저장되어 있는 또 다른 패킷들과 비교하여 연관성을 찾아 연관성 정보 테이블인 TCP_at, UDP_at, ICMP_at 테이블에 저장한다.

```
Packet_association_analysis(p_type){
 sip : source IP;
 dip : destination IP;
 dport : destination PORT number;
 timestamp : packet capture time;
 p_type : TCP, UDP, ICMP;
 cnt : total packet count;
  while(1){
  n_packet = read(p_type);
  if (n_packet == NULL) break;
  for(i=1 ; i<cnt ; i++){
    o_packet = read(p_type);
    gen_group(sip);
    gen_group(dip);
    gen_group(dport);
  }
}
```

그림 4. PAA 알고리즘
Fig. 4 PAA algorithm

### 3.2 PGGM(Pattern Graph Generate Module)

PGGM은 패킷 유형(TCP, UDP, ICMP), 발신지 IP, 도착지 IP별, timestamp로 분류하여 연관성 있는 패킷에 연관규칙을 적용하여 신뢰성이 높은 패킷들을 묶어서 패턴 그래프를 생성하는 모듈이다.

PGGM에 의해 생성된 패턴 그래프는 새로운 패턴으로 데이터베이스에 저장되어 ANIDS 시스템에 의해 침입 탐지에 이용되는데, 특히 침입 탐지시에 잘못된 긍정적 결함 오류 발생을 줄이고 패턴 검출을 용이하게 함으로써 침입탐지 시간을 단축시키고 침입 탐지율을 향상시킬 수 있다.

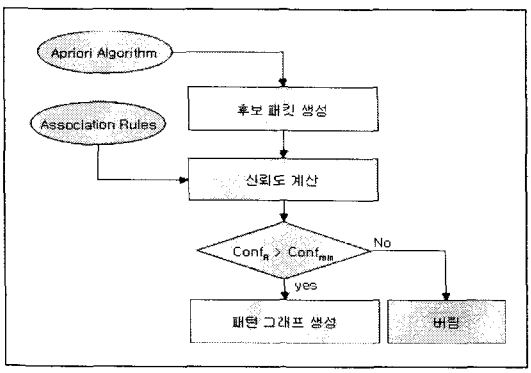



그림 5. PGGM 흐름도
Fig. 5 The flowchart of PGGM

PGGM 모듈의 처리 단계는 그림 5와 같으며, 각 과정을 처리하는 알고리즘은 그림 6은 Apriori 알고리즘을 이용해 패턴 그래프를 생성할 후보 패킷 생성 알고리즘이다[6].

```
Frequent_item() {
L1:frequent 1-itemsets;
k = 2:k represents the pass number;
Smin:minimum of support;
Ck:New candidate of size k generated from Lk-1;
Lk=All candidate in Ck with minimum support;
 while(Lk-1 ≠ φ){
  Ck = Lk-1 × Lk-1;
  For all transactions t ∈ D do begin
   Ck = subset(Ck, t);
   For all candidate c ∈ Ck do begin
     c.count = c.count + 1;
   end
  end
  k=k+1;
 }
return(Lk)//Lk = {c ∈ Ck | support(c) ≧ Smin};
}
```

그림 6. 패턴 그래프를 생성할 후보 패킷 생성 알고리즘
Fig. 6 Candidate packet generation Algorithm to create pattern graph

그림 7은 각 규칙들에 대한 신뢰도를 계산하는 알고리즘이다[6].

```
Cal_Confidence(){
 Lk:All candidate in Ck with minimum support;
 Sk:support of Lk;
 Rk:Rules of K-th, X∪Y ⇒ X;
 Confk:confidence of Rk ;
 while (Rk ≠NULL)
  Confk = Sk(X∪Y)/Sk(X) ;
}
```

그림 7. 연관규칙 신뢰도 알고리즘
Fig. 7 Association Rule Confidence Algorithm

그림 8은 최소 신뢰도를 만족하는 패킷들만 모아서 연관성 규칙에 맞게 패턴 그래프를 생성하는 알고리즘이다.

```
Pattern_graph() {
Lk : All candidate in Ck with minimum support;
Sk : support of Lk;
Rk : Rules of K-th, X∪Y ⇒ X;
Confk : confidence of Rk;
Confmin : minimum of confidence;
 k=0;
 while(1) {
  read(Rk);
  if (Confk < Confmin) break;
  gen_graph(Rk);
  k = k+1;
 }
 return(p_graph);
}
```

그림 8. 패턴 그래프 생성 알고리즘
Fig. 8 Pattern graph create Algorithm

예를 들어, 표 2는 목적지 ip가 같은 패킷(항목)을 트랜잭션별로 구분하여 정리한 것이다. 표2의 트랜잭션과 항목을 이용해 최소 지지도 $S_{min}$=0.4라고 할 때, PGGM 모듈을 적용시켜 패턴 그래프를 생성해보면 다음과 같다.

표 2. 트랜잭션과 항목
Table. 2 Transaction and item

| 트랜잭션 | 항목(패킷) |
|---|---|
| 1 | b, c, g |
| 2 | a, b, d, e, f |
| 3 | a, b, c, g |
| 4 | b, c, e, f |
| 5 | b, c, e, f, g |

먼저, k값에 따라 후보 빈발항목 집합을 만들고, 그에 따른 빈발항목 집합을 계산한다.

- k=1일 때,

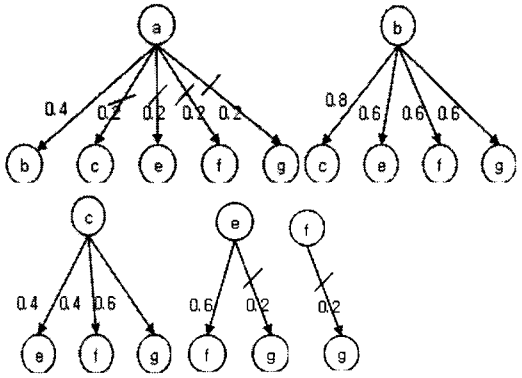
$C_1$=[{a},{b},{c},{d},{e},{f},{g}]의 후보 빈발 집단에서 최소 지지도 0.4를 마족하는 빈발 항목 집단을 만든다.

$L_1$=[{a},{b},{c},{e},{f},{g}]

- k=2일 때,

$C_2$
=[{a,b},{a,c},{a,e},{a,f},{a,g},{b,c},{b,e},{b,f},{b,g},{c,e},{c,f},{c,g},{e,f},{e,g},{f,g}]의 후보 빈발 집단에서 최소지지도 0.4를 만족하는 빈발 항목 집단을 만든다.

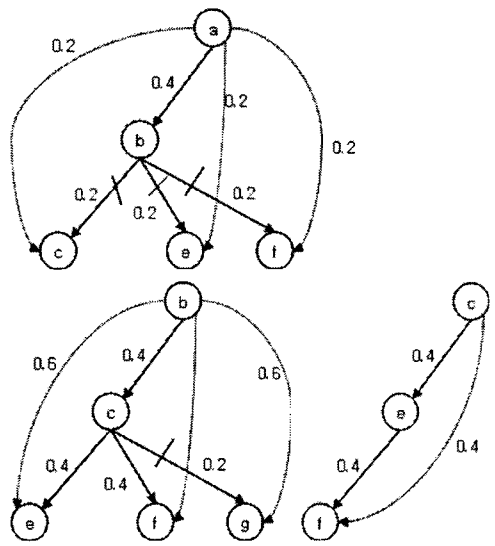
$L_2$
=[{a,b},{b,c},{b,e},{b,f},{b,g},{c,e},{c,f},{c,g},{e,f}]




- k=3일 때,

$C_3$=[{a,b,c},{a,b,e},{a,b,f},{b,c,e},{b,c,f},{b,c,g},{b,e,f},{c,e,f}]의 후보 빈발 집단에서 최소지지도 0.4를 만족하는 빈발 항목 집단을 만든다.

$L_3$=[{b,c,e},{b,c,f},{b,c,g},{b,e,f},{c,e,f}]

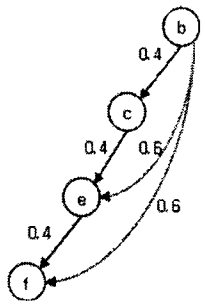



- k=4일 때,

$C_4$=[{b,c,e,f}]의 후보 빈발 집단에서 최소지지도 0.4

를 만족하는 빈발 항목 집단을 만든다.

$L_4$=[{b,c,e,f}]




- k=5일 때,

$C_5 = \phi, L_5 = \phi$

연관규칙 R: X⇒Y에서 X={항목 b와 c가 발생한다}, Y={항목 g가 발생한다}로 가정 한다면, 연관규칙 R:X={b,c}⇒Y={g}가 된다.

Y에 해당되는 g항목에 대한 지지도(support)를 계산하면, support(Y)=4/5=0.8가 되고, support(X)= support({b,c})=4/5=0.8, support(R)=support({b,c,g})= 3/5 = 0.6이 된다.

따라서, confidence(R)= support(XUY)/support(X) = support({b,c,g}) / support({b,c}) = 0.6/0.8 = 0.75가 된다. 최소 신뢰도 $Conf_{\min}$=0.5라고 한다면, 연관규칙 R:X={b,c}⇒Y={g}은 침입으로 간주하는 패턴이 되므로 각 항목의 b, c, g를 시간항목별로 구분하여 패턴 그래프를 생성하면 그림 9와 같은 형태의 패턴 그래프가 된다.

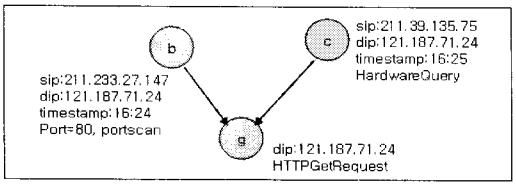



그림 9. 패턴 그래프 예
Fig. 9 Pattern graph example

### 3.3 IDM(Intrusion Detection Module)

IDM은 PGGM에 의해 수집된 패킷간의 연관성을 분석하여 새롭게 생성된 패턴 그래프 패턴을 침입패턴(이미 정의되어 있는 침입유형의 패턴 그래프)과 비교하여 침입인지를 탐지하는 모듈이다.

IDM 처리 과정 흐름은 그림 10과 같으며 단계별로 살펴보면 다음과 같다.

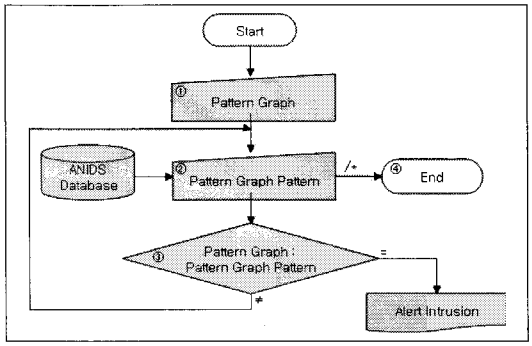



그림 10. IDM 흐름도
Fig. 10 The flowchart of IDM

[1 단계] IDM은 PGGM에 의해 새롭게 생성되어 데이터베이스에 저장된 패턴 그래프를 읽는다.

[단계 2] IDM은 이미 침입으로 판정되어 데이터베이스에 저장된 패턴 그래프 패턴을 차례대로 읽은 후 단계 3으로 이동하고, 패턴 그래프 패턴이 더 이상 없으면 단계 4로 이동한다.

[단계 3] 단계 1의 패턴 그래프와 단계 2의 패턴 그래프 패턴과 비교하여 같이 않으면 단계 2로 이동하고, 같으면 시스템 관리자에게 침입을 경고한다.

[단계 4] 침입 판정 단계를 종료한다.

## Ⅳ. ANIDS 데이터베이스

ANIDS 데이터베이스는 PMM에 의해 분류된 패킷 정보, PAAM에 의해 분석된 패킷간의 연관성, PGGM에 의해 생성된 패턴 그래프와 침입 유형으로 판정된 패턴 그래프 패턴을 저장한다.

그림 11은 각각의 테이블간의 관계를 설명한 것이다. ANIDS에서 사용되는 모든 패킷들의 저장되어 있는 captured_packet_info. 테이블, 패킷 유형별로 분류하여 저장한 TCP_t, UDP_t, ICMP_t 테이블, 연관성 있는 패킷들만 모아서 저장한 TCP_at, UDP_at, ICMP_at 테이블, 패턴 그래프가 저장되어 있는 packet_graph 테이블, 기존의 침입형 패턴 그래프가 저장되어 있는 pattern_graph_pattern 테이블, 실험에 사용할 연관 규칙이 저장되어 있

는 Assciation_rules 테이블이 있으며, 각각의 테이블은 packet_id로 조인되어 있다. 따라서, pattern_graph 테이블에서 패킷에 대한 좀 더 자세한 정보를 원한다면 captured_packet _info 테이블을 연계하여 패킷 정보를 검색할 수 있다.

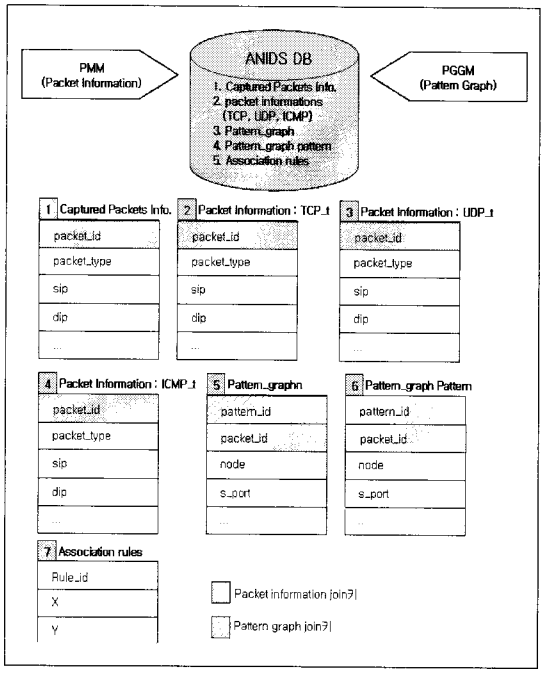



그림 11. ANIDS 데이터베이스
Fig. 11 ANIDS database

표 3은 TCP_t, UDP_t, ICMP_t 테이블의 형식으로 PMM에 의해 TCP, UDP 등의 유형별로 분류된 패킷들이 이 테이블에 저장된다.

표 3. TCP_t 테이블 형식
Table. 3 TCP_t table format

| Field | Field Type | File size | 설명 |
|---|---|---|---|
| packet_id | int | 11 | 패킷 고유 id |
| packet_type | varchar | 15 | 패킷 유형 |
| sip | varchar | 15 | 발신지 IP |
| s_port | varchar | 10 | 발신지 포트번호 |
| d_ip | varchar | 15 | 도착지 IP |
| d_port | varchar | 10 | 도착지 포트번호 |
| timestamp | varchar | 20 | 캡쳐시간 |
| data | text | 16 | 내용 |

저장된 패킷 정보들은 PGGM에서 패킷간의 연관성을 분석하는 정보로 이용되며, 연관성 있는 패킷들을 모아서 TCP_at, UDP_at, ICMP_at 테이블에 저장한다. 따라서, TCP_t,UDP_t, ICMP_t 테이블은 TCP_at, UDP_at, ICMP_at 테이블과 동일한 형식을 갖는다.

PGGM은 TCP_at, UDP_at, ICMP_at의 패킷 정보를 패킷 유형, 발신지 IP, 도착지 IP별, timestamp로 분류하여 Apriori 알고리즘을 이용해 지지도 $Sup_{\min}$보다 큰 패킷만을 모아 연관 규칙의 신뢰도를 계산하여 $Conf_{\min}$보다 큰 연관 규칙의 패킷만을 묶어서 패턴 그래프를 생성한다. 패턴 그래프가 침입일 경우에 pattern_graph_pattern 테이블에 저장하므로, pattern_graph와 pattern_graph_pattern 테이블의 형식은 동일하다.

표 4는 생성된 패턴 그래프를 저장하는 테이블 형식이다.

표 4. pattern_graph 테이블 형식
Table. 4 pattern_graph table format

| Field | Field Type | File size | 설명 |
|---|---|---|---|
| pattern_id | int | 11 | 패턴 그래프 고유 번호 |
| packet_id | int | 11 | 패킷 고유 번호 |
| node | int | 11 | 항목 번호 |
| s_port | varchar | 10 | 발신지 포트번호 |
| d_ip | varchar | 15 | 도착지 IP |
| d_port | varchar | 10 | 도착지 포트번호 |
| timestamp | varchar | 20 | 캡쳐시간 |
| data | text | 16 | 내용 |
| l_point | int | 11 | 연결되는 왼쪽노드 값 |
| r_point | int | 11 | 연결되는 오른쪽노드 값 |

본 논문에 사용한 연관규칙은 R:X⇒Y형태로 표현되며, 표 5는 X와 Y에 대한 규칙 표현식을 저장한 테이블이다. 연관규칙 표현은 미리 정의된 형식으로 실험이 진행되었으며 실험이 진행되는 동안에 추가되지 않도록 하였다.

표 5. Association Rules 테이블 형식
Table. 5 Association Rules table format

| Field | Field Type | File size | 설명 |
|---|---|---|---|
| rule_id | int | 11 | 연관규칙 고유 번호 |
| X | varchar | 20 | 규칙 X 값 |
| Y | varchar | 20 | 규칙 Y 값 |

## V. ANIDS 시뮬레이션

제안된 ANIDS 실험은 실험을 위해 준비한 패턴 그래프 패턴을 데이터베이스에 미리 저장한 후 새로 생성된 패턴 그래프와 비교하여 침입인지를 판단하는지를 실험하였으며, ANIDS 시뮬레이션은 한컴 리눅스 Professional 운영체제, 2.39GHz Pentium IV, 512MB RAM 환경에서 실행되었다.

실험에 이용된 패턴 그래프는 깊이에 따라 ANIDS가 침입 탐지률이 다르게 나타났으며, 패턴 그래프 깊이가 2일 경우에 가장 높게 나타났다.

### 5.1 패킷 종류에 따른 분석

표 6는 ANIDS에서 24시간 동안 패킷 필터링 규칙에 의해 수집된 패킷들이다. 캡쳐된 패킷들 중에서 패턴 그래프에 필요한 정보인 TCP, UDP, ICMP를 분류하여 패턴 그래프를 자원으로 이용하였다.

표 6. 캡쳐된 패킷 현황
Table. 6 The status of Captured Packet

| Packet Type | 24 hours | 패턴 그래프 |
|---|---|---|
| TCP | 199,611 | ○ |
| UDP | 54,450 | ○ |
| ICMP | 622 | ○ |
| ARP | 162,574 | × |
| NetBios | 655 | × |
| IPX | 6,530 | × |
| other | 77,456 | × |
| Total | 501,898 | |

### 5.2 패턴 그래프를 이용한 분석

ANIDS는 캡쳐된 패킷들을 중에서 연관규칙의 $Conf_{min}$보다 큰 신뢰성을 갖는 연관규칙의 패킷들만 축출하여 연관규칙의 패턴 그래프를 생성한다. 표 7은 $Conf_{min}$보다 큰 신뢰성을 갖는 패킷들의 수를 나타낸 것이다. 연관규칙의 기본 길이는 2,3,4로 하였는데, 연관규칙의 길이가 결국은 패턴 그래프의 깊이가 된다.

표 7. $Conf_{min}$보다 패킷 현황
Table. 7 The status of Packet large than $Conf_{min}$

| Packet Type \ 패턴 그래프 깊이 | 2 | 3 | 4 | Total |
|---|---|---|---|---|
| TCP | 50 | 25 | 15 | 90 |
| UDP | 42 | 24 | 13 | 79 |
| ICMP | 12 | 10 | 8 | 30 |

표 8은 생성된 패턴 그래프에서 침입으로 탐지된 패턴 그래프의 개수와 IDM에 의해 침입으로 탐지된 패턴 그래프의 수를 설명한 것이다.

표 8. 침입 탐지 패턴 그래프 현황
Table. 8 The status of intrusion detection Packet graph

| 패킷 유형 \ 패턴 그래프 깊이 | 2 | | 3 | | 4 | |
|---|---|---|---|---|---|---|
| TCP | 25 | 20 | 18 | 14 | 12 | 8 |
| UDP | 20 | 18 | 16 | 10 | 10 | 6 |
| ICMP | 9 | 7 | 7 | 5 | 6 | 4 |

그림 12는 생성된 패턴 그래프와 패턴 그래프 패턴과 비교하여 침입인지를 탐지한 비율을 그래프의 깊이 별로 나타낸 것이다. 그래프 깊이가 2일 경우에 침입인지를 가장 잘 탐지 했으며, 깊이가 커질 수록 침임 탐지 시간이 증가하고 침입 탐지율이 감소되는 것을 볼 수 있다. 이것은 여러 노드를 거치면서 처리해야 할 자료가 많아지면서 탐지율이 떨어지는 것을 알 수 있다.

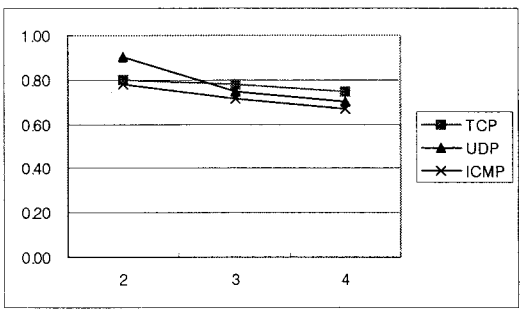



그림 12. 패턴 그래프 탐지 성공률
Fig. 12 The success rate of Pattern graph detection

### 5.3 긍정적 결함(False Positive) 분석

ANIDS는 트래픽이 많은 네트워크 상에서 침입을 탐지하므로 침입 탐지 오류율 증가로 인해 ANIDS의 침입

탐지 성능이 저하될 수 있다. 따라서 본 논문에서는 침입 탐지 오류율을 감소시키기 위해, 패턴 그래프 패턴을 이용하였다.

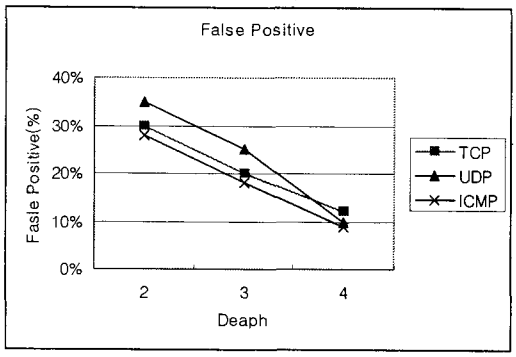



그림 13. 패턴 그래프 깊이에 따른 False Positive
Fig. 13 The False Positive according to Pattern graph depth

실험을 통해 패킷단위로만 침입 탐지를 했을 때와 패턴 그래프 패턴을 이용해 침입 탐지를 했을 때를 비교해 보니, 패턴 그래프 패턴을 이용했을 때 긍정적 결함(False Positive) 비율이 낮아진 것을 알 수 있었다.

## VI. 결 론

인터넷의 급속한 성장으로 현재 우리 생활속에서 빼놓을 수 없는 하나의 정보 매체로 자리매김을 하고 있으나, 전산망 침입 행위, 전자기록 위조 및 변조, 바이러스 유포 등 많은 위협들이 증가하고 있으며, 특히 인터넷의 웜이나 서비스 거부 공격 등으로 공공기관 이나 기업에 큰 피해를 주고 있다. 이러한 네트워크 공격으로 인한 피해를 줄이기 위해 침입 탐지 시스템이 많이 사용되고 있다.

본 논문에서 제안한 ANIDS는 패킷 수집 및 패킷 유형별로 분류하는 PMM(Packet Management Module), 패킷간의 연관성을 분석하고, 연관 규칙 마이닝을 이용해 연관규칙의 신뢰도를 계산하여 최소 신뢰도를 만족하는 연관규칙의 패턴 그래프를 생성하는 PGGM(Pattern Graph Generate Module) 그리고 패턴 그래프를 이용해 침입을 탐지하는 IDM(Intrusion Detection Module)로 구성된다.

ANIDS는 패킷간의 연관성을 분석하여 침입인지를 탐지할 수 있는 패턴 그래프 패턴을 사용함으로써, 침입 탐지의 긍정적 결함 오류를 감소시킬 수 있으며, 완벽한 패턴 그래프 패턴이 생성되기 전에, 이미 침입으로 판정된 패턴 그래프 패턴과 비교하여 유사한 패턴 형태를 침입으로 간주함으로써 기존의 침입 탐지 시스템에 비해 침입 탐지 속도를 감소시키고 침입 탐지율을 증가시킬 수 있다.

## 참고문헌

[1] 황현욱, 김민수, 노봉남, "감사로그 상관관계를 통한 호스트 기반의 침입탐지시스템", 정보보호학회논문지, 제 13권 제 3호, pp. 81~90, 2003.6

[2] James Cannady and Jay harell, "A Comparatice Analysis of Current Intrusion Detection Technologies", February, 1998

[3] Eric Boledan, Alan D. Christiansen, William Hill, Clement Skorupka, Lisa M. Talbot, Jonathan Tivel, "Data Mining for Network Intrusion Detection : How to Get Started", The MITRE Corporation, http://www.mitre.org/work/tech_papers/tech_papers_01/bloedorn_datamining/bloedorn_datamining.pdf

[4] 이상훈, 소진, "데이터 마이닝 기반 침입탐지 패턴 알고리즘의 설계 및 구현", 정보처리학회 논문지 Vol.6, No.10-C, pp.717~726, 2003. 9.

[5] 노광민, "리눅스에서 pcap library를 사용하여 패킷을 잡아보기 v0.3 2000. 09. 14., 리눅스 한글 문서 프로젝트

[6] Jyothna R., Nayak and Diane J. Cook, "Approximate Association Rule Mining", http://ranger.uta.edu/~cook/pubs/flairsj01.pdf,

[7] http://www.mic.go.kr/index.jsp

[8] Martin Roesch, "Snort - Lightweight Intrusion Detection for Network"

[9] Christopher Kruegel, Tomas Toth and ClemensKerer, "Decentralized Event Correlation for Intrusion Detection", 2002. 4.

## 저자소개


**정 은 희(Eun-Hee J대ng)**

1991년 강릉대학교 통계학과 이학사
1998년 관동대학교 전자계산공학과 공학석사
2003년 관동대학교 전자계산공학과 공학박사
2003년 9월~현재 강원대학교 지역경제학과 조교수
※관심분야: 네트워크 보안, 전자상거래, 웹 프로그래밍

**이 병 관(Byung-Kwan Lee)**

1975년 부산대학교 기계설계학과 학사
1986년 중앙대학교 전자계산공학과 석사
1990년 중앙대학교 전자계산공학과 박사
1988년 3월~현재 관동대학교 컴퓨터학과 교수
※관심분야: 네트워크 보안, 전자상거래, 컴퓨터 네트워크