COBWEB을 사용한 비정상행위도 측정을 지원하는 네트워크기반 침입탐지시스템 설계

이효승*, 원일용*, 이창훈*
*건국대학교 컴퓨터공학과
e-mail : yfduaud,clcc,chlee@konkuk.ac.kr

A Design of Network Based IDS to Report Abnormal Behavior Level using COBWEB

Hyo-Seong Lee*, Il-Yong Won*, Chang-Hun Lee*
*Dept. of Computer Engineering, KonKuk University

요 약

네트워크 기반 침입탐지시스템은 연속적으로 발생하는 패킷의 무손실 축소와 행위패턴을 정확히 모델링 할 수 있는 Event 의 생성이 전체성능을 결정하는 중요한 요인이 된다. 또한 공격이나 비정상 행위의 판별을 위해서는 효과적인 탐지모델의 구축이 필요하다. 본 논문은 네트워크기반에서 패킷을 분석해 비정상행위 수준을 관리자에게 보고하는 시스템의 설계에 관한 논문이다. 속성을 생성하고 선택하는 방법으로는 전문가의 경험을 바탕으로 결정하였고, 탐지모델구축은 COBWEB 클러스터링 기법을 사용하였다. 비정상행위 수준을 결정하기위해 트레이닝 셋에 정상과 비정상의 비율을 두어 클러스터링 이후 탐지모드에서 새로운 온라인 Event 의 비정상 수준을 결정할 수 있게 하였다

1. 서론

침입탐지시스템(IDS: Intrusion Detection System)은 데이터의 생성지나 탐지기법에 따라 다양하게 분류된다[1]. 정상적 행위를 패턴화 하여 비정상 행위(Abnormal)를 탐지하는 IDS 는 주로 통계적 기법이나 데이터마이닝을 사용하여 정상행위를 학습하여 패턴을 형성한다. 이 패턴을 탐지모델이라고 한다. 이러한 IDS 의 설계에서 고려할 사항은 탐지모델을 구축하기위한 정상행동을 정확히 모델링한 학습 데이터를 생성하는 방법과, 생성된 데이터를 바탕으로 탐지모델을 구축하고 탐지를 하는 유용한 학습 알고리즘의 적용이다.

본 논문은 위에서 서술한 고려 사항을 바탕으로 네트워크기반 실시간 비정상행위 IDS 의 설계에 관한 것이다. 제안 시스템은 네트워크 패킷을 원시데이터로 사용하는데, 정상행위를 모델링하고 탐지모델 구축 알고리즘에 적용하기 위해서는 일련의 전처리 과정을 거쳐 유효한 데이터로 재생해야 한다. 이것을 Event 라고 하는데, 본 논문에서는 전문가의 경험[2][3]을 바탕으로 전처리 과정을 설계했다. 또한, 전통적으로 비정상행위를 탐지하기위한 데이터는 정상행위로 이루어진 데이터를 바탕으로 탐지모델을 구축하나, 본 논문에서는 정상과 비정상행위를 함께 적용한 탐지모델을 구축하여 비정상행위레벨(ABL: Abnormal Behavior Level)을 결정하였다. ABL 은 온라인 상에서 발생한 Event 를 탐지한 결과로 관리자에게 보고되어진다. 탐지모델을 구축하고 탐지하는 기법으로는 COBWEB 클러스터링[4] 기법을 사용하였다.

2. 동기 및 관련연구

2.1 동기

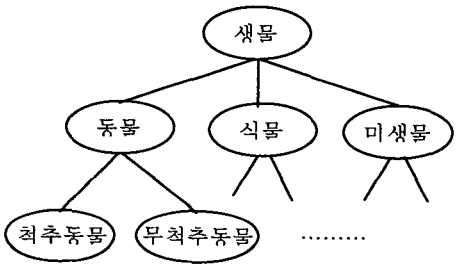
클러스터링은 대용량의 사건들이 기록되어있는 데이터베이스에서 유사 작업군을 탐색하는 기법으로, 네트워크에서 발생하는 대용량의 패킷을 대상으로 행위의 패턴을 분류하는데 적합하다[2]. COBWEB 은 무감독학습 기반의 클러스터링 기법으로 데이터의 속성들이 지닌 값으로 부류 내(Intra-Class)의 유사도와 부류 사이(Inter-Class)의 차이도를 결정하여 유사군의 개념을 형성하며 클러스터링하는 기법이다. COBWEB 을 사용하면 네트워크 Event 의 속성값으로 정상개념과

비정상개념을 형성할 수 있다. 또한, 다른 클러스터링 알고리즘과는 달리 클러스터 결과에 개념을 요약한 정보를 지니고 있어 새로운 Event 에 대한 부류정보를 보고하기에 적합하며, ABL 을 계산하는 데에도 이러한 정보를 이용할 수 있다.

ABL 은 보안관리자에게 이분화된 정보를 알려주는 것이 아니라 온라인에서 발생한 네트워크 Event 의 정상 정도 또는 비정상 정도를 알려줌으로써 보안정책의 적용에 유연성을 제공하게 된다.

2.2 COBWEB

제안한 시스템에서 사용한 침입탐지모델 생성기법은 계층형 개념 군집화(Hierarchical Conceptual Clustering)학습 알고리즘으로 알려져 있는 COBWEB 이다[4][5][6][7]. COBWEB 은 인간이 사물을 분류하는 과정인 점진적 개념 형성(Incremental Concept Formation)을 모델로 하여 개발되었다. 이것은 사물을 하나씩 관찰하여 개념을 형성하면서 하향 분류하는 방법으로, 쉬운 예가 <그림 1>에 나타나있다.



<그림 1> 생물의 분류

위와 같은 클러스터는 인간이 생물을 관찰하는 역사적인 과정을 통해 이루어진 것이다. 사물을 하향분류하기위해 또는 그 과정에서 필연적으로 분류기준이 형성되는데 이것이 바로 부류의 추상화 된 개념정보 이다. 새로운 사물이 관찰되면 자연스럽게 이전에 형성된 분류 기준에 의해서 기존의 부류에 포함시키거나 새로운 부류를 생성하게 된다. COBWEB 의 개념 클러스터링 과정도 이와 유사하다.

COBWEB 은 레코드로 구성된 데이터들을 입력으로 받아, 트리의 형태로 클러스터링을 한다. 트리의 각 노드는 하나의 개념이 되고, 각 노드에는 속성값이 요약되어진 개념정보를 저장하고 있다. 개념 정보는 속성값의 도메인별 확률값이나 평균과 표준편차이다. 속성이 명목형(Nominal) 도메인인 경우에는 확률값이고, 연속형인 경우에는 평균과 표준편차가 된다. 개념정보는 새로운 레코드의 부류를 결정하는데 사용되는 정보가 된다.

COBWEB 은 하나의 레코드를 개념 계층에 통합(incorporate)하고 트리를 성장시키기 위한 4 개의 연산자를 제공한다.

- Incorporate: 레코드를 현재 개념에 통합.
- Create new disjunct: 새로운 개념 노드 형성.
- merge: 개념이 과도하게 분리되어져 있을 때 두개의 개념을 하나로 병합.
- Split: 개념이 과도하게 일반화되어있어 하나의 개념을 다수개의 개념으로 분리.

연산자를 선택하기위해서 Category Utility[8] 평가함수를 사용한다. CU 는 클래스 내부의 멤버들 사이의 유사도(intra-class similarities)가 최대고 다른 클래스들 멤버 사이의 차이가(inter-class differences) 최대인 부류(class)에 최고의 점수(CU score)를 준다. 직관적으로 유사한 속성값을 가진 인스턴스들끼리 클래스를 형성하는 것이 합리적이다.

3. 침입탐지시스템 구조

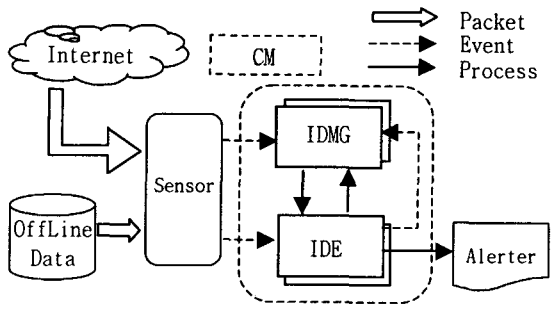
3.1 시스템 전체 구조

<그림 2>는 제안한 시스템의 전체 구성도이다[3][9]. 탐지모델구축단계(학습모드)는 원시 패킷을 dump 하여 생성한 파일을 Sensor 에서 읽어 Event 로 전처리하여 IDMG(Intrusion Detection Model Generator)로 보내 침입탐지모델을 생성한다. 탐지모드에서는 Internet 으로부터 탐지대상 네트워크로 들어오는 패킷을 Sensor 에서 Event 로 가공하여 IDE(Intrusion Detection Engine)로 보내면, IDE 는 IDMG 의 탐지모델을 바탕으로 ABL 을 결정하여 Alerter 에게 보낸다. Sensor 는 프로토콜별로 세분화된 Event 를 형성하여 IDMG 와 IDE 에서 처리하도록 했으며, ABL 결과에 따라 Event 는 다시 IDMG 에게 보내져 탐지모델갱신에 사용된다. 각 구성요소는 네트워크에 분산 배치될 수 있다. 이러한 구성환경 관리를 CM(Configure Manager)을 통해 관리한다. 본 설계에서는 CM 을 별도의 모듈로 두지않았고 환경파일로 설정하였다.

4. 모듈별 관련 기법 및 구현

4.1 Sensor - Event 의 형성

Event 형성의 주 아이디어는 인터넷 프로토콜 설계가 정상 작동만을 고려하여 작성된 것이므로, 정상인 상태에서 발생하는 패킷의 통계값과 비정상 상태에서 발생하는 패킷의 통계는 차이를 보일 것이라는 것이다.



<그림 2> 시스템 전체 구성도

연속적인 다량의 패킷들을 수집 시간 간격, 탐지대상 호스트의 IP 주소 그리고 TCP, UDP, ICMP 등의 프로토콜을 기준으로 하나의 레코드로 Event 화 한다. 이 과정에서 패킷 헤더의 내용을 탐지모델의 구축에 적합하도록 속성별 조합을 통해 새로운 속성을 생성해야 한다. 이 과정은 기계학습이나 데이터마이닝에서 사용하는 속성생성(Feature Generation), 속성선택(Selection)의 기법을 적용해야 하지만 본 논문에서는 전문가의 경험[2][3][9][10]을 바탕으로 적용하였다. 다음 [표 1]은 본 논문에서 제안하는 Event 의 종류와 속성이다. 수집 시간간격은 학습 데이터 집합을 만들어 클러스터링 실험을 반복하여 적당한 시간 간격을 정해야 한다. [표 1]의 Event 는 TCP 와 UDP,ICMP 프로토콜의 두 종류로 구성한 것이다.

종 류	속 성
공 통	- 탐지 네트워크상의 IP 패킷수 - TCP, UDP, ICMP 패킷수 및 비율 - 정상비율 : 0 ≤ r ≤ 1
TCP/IP	- TCP 패킷 비율 : Inbound, Outbound - Connection : SYN, SYN/ACK, ACK 비율 - SYN 을 보낸 출발지 IP 의 수 - FIN, RESET 비율 - TCP 헤더의 플래그 비트 값 - TCP payload 의 길이 총합
UDP,ICMP/IP	- UDP 패킷 비율 : Inbound, Outbound - UDP payload 의 길이 총합 - ICMP 패킷 비율 : Inbound, Outbound

[표 1] Event 의 종류와 속성

탐지모델 구축에 사용되는 Event 를 학습 Event 라 하고 탐지 모드에서 생성되는 Event 를 온라인 Event 라 한다.

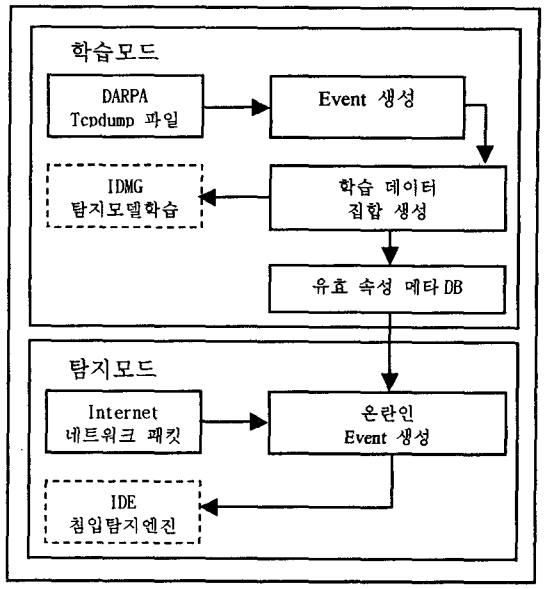
하나의 학습 Event 는 다수의 패킷으로 이루어져있고 그 중에는 정상과 비정상 패킷이 포함되어있을 것이다. 정상비율은 Event 에 포함된 정상 패킷의 비율로 1 에 가까울수록 정상 패킷이 많이 포함된 것이다. 이 속성은 COBWEB 에서 클러스터링 하는데 사용되지 않고 단지 ABL 을 결정하기위해 개념 정보에 저장되는 속성이다. 따라서 온라인 Event 에는 이 속성을 만들지 않는다.

Event 의 성격에 따라 침입을 탐지하는 영역이 결정된다. [표 1]과 같이 Event 를 구성하여 탐지모델을 구축하면 탐지영역은 DoS 공격이나 Port Scan 을 탐지하는데 적합하게 된다. 즉, Event 를 형성하는 의도에 따라 침입탐지 영역을 정할 수 있다.

<그림 3>은 Sensord 의 내부 절차를 표현한 것이다. 점선의 사각형은 Sensor 부가 아닌 IDMG 와 IDE 부분이다.

4.2 침입탐지모델의 형성 및 ABL 결정

침입탐지모델을 생성하는 IDMG 의 핵심 알고리즘은 COBWEB 이다. COBWEB 은 개체의 속성값을 바탕으로 개념을 분류하므로, 정상과 비정상이 혼합된 Event 를 클러스터링 한 결과는 완전히 정상인 개념, 정상과 비정상이 혼합된 개념 또는 완전히 비정상인 개념의 부류를 형성할 것이다.



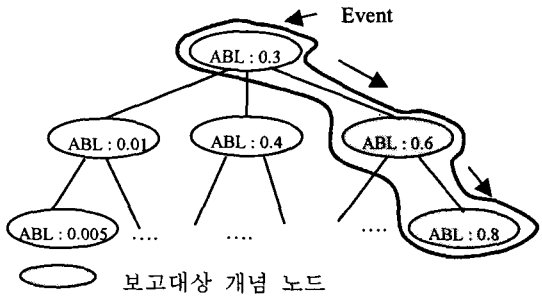
<그림 3> Sensor 부의 구성도

Event 의 정상비율은 개념노드를 형성해가면서 Event 가 포함된 노드에 값을 누적해 간다. 학습이 끝나면 개념 노드 당 포함된 Event 의 총수와 정상비율의 누적합으로 비정상행위의 레벨을 구할 수 있다.

ABL = 1 - (정상비율누적합 / 노드의 총 Event 수)

ABL 을 결정하는데 탐지모델 트리의 일정 깊이 이하 노드는 사실상 무의미하다. 따라서 트리 형성 후 노드에 포함된 총 Event 수의 임계치를 정해 임계치 이하의 개념노드는 탐지에 제외시켜야 한다.

IDE(침입탐지엔진)에서는 온라인 Event 가 입력되면 학습된 탐지모델 트리와 COBWEB 의 예측알고리즘을 사용하여 학습과 동일한 방식으로 분류를 한다. 분류결과 탐지 Event 의 개념노드가 결정되면 그 노드의 ABL 을 Alerter 모듈에서 관리자에게 알려주는데, 일정 임계값 이하의 ABL 은 보고하지 않도록 설정할 수 있다. <그림 4>는 ABL Alert 의 개념도이다.



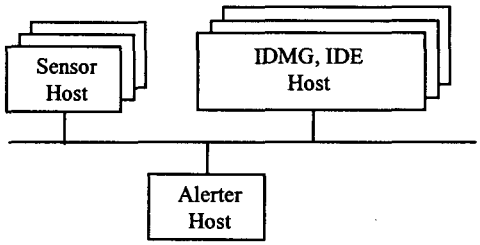
<그림 4> ABL Alert 개념도

위 그림처럼 Event가 분류되면 해당 레벨별 개념의 모든 ABL을 보고하여 관리자가 분석할 수 있도록 한다.

5. 구현환경 및 실험데이터

5.1 구현 환경

제안한 IDS의 각 구성부는 네트워크에 분산 배치하여 개발하였다. Sensor 부는 Linux 환경에서 개발되었으며 탐지 대상 네트워크의 DMZ이나 호스트에 직접 배치가 가능하다. IDMG와 IDE는 Windows2000에서 개발되었으며 Alerter 또한 Windows2000에서 개발 배치되었다. 각 부분간의 통신은 Socket 통신을 한다. Sensor부의 위치는 네트워크상의 어느 위치에 있어도 무관하나 구현은 OS와 Event 종류에 종속적이다. <그림 5>는 제안한 시스템의 네트워크 구성도이다.



<그림 5> 네트워크 구성도

5.2 실험 데이터의 생성

제안한 시스템의 실험을 위해 학습 원시데이터는 DARPA 산하 MIT Lincoln Lab.[11]에서 제공하는 Solaris tcpdump 트레이닝 데이터를 사용한다. DARPA는 1998년부터 2000년까지 Solaris 기반 BSM Audit 데이터와 tcpdump 데이터를 IDS 개발과 평가를 위한 데이터로 제공하고있다. 본 시스템은 1998년 트레이닝 데이터 중 week1 수요일 데이터와 week2 금요일의 tcpdump 데이터를 사용하였다. 이틀간의 데이터에는 정상 패킷과 smurf와 syslog와 같은 서비스거부공격으로 이루어진 비정상 패킷이 혼합되어있다. 탐지 데이터도 동일한 시스템에서 작성된 DARPA 데이터를 사용한다.

6. 결론 및 향후과제

제안한 시스템이 기존시스템과 구별되는 특징은 다음과 같다.

① 침입탐지모델의 유지보수에 용이 : 대부분의 IDS는 탐지모델을 갱신하기위해서 종전의 학습데이터와 새로운 학습데이터를 통합해 재 학습을 하여야 하나, 본 시스템은 COBWEB을 사용함으로써 새로운 학습데이터만으로도 재 구축이 가능하다.

② ABL 경고 방식 : 비정상행위 레벨을 관리자에게 보고하는 방식의 이점은 탐지실패(False Negative)나 오용탐지(False Positive)를 줄여주고 관리자에게 정확한 분석의 기회를 제공한다.

③ Event 확장의 유연성 : Event의 종류에 상관없이 IDMG와 IDE는 동일한 적용이 가능하고, Event의 추가에 따라 탐지영역의 확대도 용이하다.

현재 본 시스템의 프로토타입이 구현 중에 있다. 완성도는 약 80%이며 모듈별로는 이미 완성되어 부분 테스트를 실시하고있다. 향후 연구과제는 다음과 같다.

① Event 생성의 자동화 알고리즘 적용 연구(속성 생성과 속성 선택 알고리즘)

② ABL을 계산하는 타 알고리즘에 관한 연구

③ Event 종류별로 보고되는 ABL을 통합(Information Fusion)하는 알고리즘 연구.

참고문헌

[1] Häkan Kvarnström, "A survey of commercial tools for intrusion detection", Techinal report 99-8 Dept. of Computer engineering Chalmers University of Technology Göteborg, Sweden, 1999

[2] 한국정보보호진흥원, "정보통신기반구조 보호기술 개발 최종 보고서", 한국정보보호진흥원 보고서,2001

[3] 이정현, "네트워크 기반 비정상 행위에 대한 다계층 침입 탐지 시스템 설계 및 구현", 석사학위논문, 건국대학교 컴퓨터 공학과, 2001

[4] Fisher, D. H., "Knowledge acquisition via incremental conceptual clustering", Doctoral dissertation, Dept. of Information & Computer Science, University of California, Irvine, 1987

[5] Fisher, D.H., "Iterative Optimization and Simplification of Hierarchical Clusterings.", Technical Report CS-95-01, Vanderbilt University, Nashville TN., 1995

[6] Kathleen McKusick, Kevin Thompson, "COBWEB/3: A Portable Implementation", Technical Report FIA-90-6-18-2, AI Research Branch, NASA Ames Research Center, 1990

[7] Ian H. Witten, Eibe Frank, "Data Mining", Morgan Kaufmann, 2000

[8] Gluck, M., & Corter, J., "Information, Uncertainty and the utility of categories", Proceedings of the Seventh Annual Conference of the Cognitive Science Society(pp. 283-287). Irvine, CA: Lawrence Erlbaum., 1985

[9] Stephan Northcutt, Judy Novak, & Donald McLachlan, "Network Intrusion Detection An Analyst's Handbook", 2nd Ed. New Riders, 2000

[10] W. Richard Stevens, "TCP/IP Illustrated, Vol.1, The Protocols", Addison-Wesley, 1994

[11] http://www.ll.mit.edu/IST/ideval/index.html