감사데이터 분석을 위한 마이닝 시스템 설계 및 구현†

김은희° 문호성 신문선 류근호 김기영*
충북대학교 데이터베이스 연구실
한국전자통신연구원*
{ehkim°, hsmoon, msshin, khryu}@dblab.chungbuk.ac.kr
kykim@etri.re.kr*

Design and Implementation of Mining System for Audit Data Analysis

Eun-Hee Kim° Ho-Sung Moon Moon-Sun Shin Kyun-Ho Ryu Ki-Young Kim*
Chungbuk National University Database Laboratory
Electronics and Telecommunications Research Institute*

요 약

네트워크의 광역화와 새로운 공격 유형의 발생으로 침입 탐지 시스템에서 새로운 시퀀스의 추가나 침입탐지 모델 구축의 수동적인 접근부분이 문제가 되고 있다. 특히 기존의 침입탐지 시스템들은 대량의 네트워크 하부구조를 가진 네트워크 정보를 수집 및 분석 하는데 있어 각각 전담 시스템들이 담당하고 있다. 따라서 침입탐지 시스템에서 증가하는 많은 양의 감사데이터를 분석하여 다양한 공격 유형들에 대해서 능동적으로 대처할 수 있도록 하는 것이 필요하다. 최근, 침입 탐지 시스템에 데이터 마이닝 기법을 적용하여 능동적인 침입탐지시스템을 구축하고자 하는 연구들이 활발히 이루어지고 있다. 이 논문에서는 대량의 감사 데이터를 정확하고 효율적으로 분석하기 위한 마이닝 시스템을 설계하고 구현한다. 감사데이터는 트랜잭션데이터베이스와는 다른 특성을 가지는 데이터이므로 이를 고려한 마이닝 시스템을 설계하였다. 구현된 마이닝 시스템은 연관규칙 기법을 이용하여 감사데이터 속성간의 연관성을 탐사하고, 빈발에피소드 기법을 적용하여 주어진 시간 내에서 상호 연관성 있게 발생한 이벤트들을 모음으로써 연속적인 시간간격 내에서 빈번하게 발생하는 사건들의 발견과 알려진 사건에서 시퀀스의 행동을 예측하거나 기술할 수 있는 규칙을 생성할 수 있다. 감사데이터의 마이닝 결과 생성된 규칙들은 능동적인 보안정책을 구축하는데 활용될 수 있다. 또한 데이터양의 감소로 침입 탐지시간을 최소화하는데도 기여할 것이다.

1. 서 론

최근 네트워크 구성이 복잡해짐에 따라 정책기반의 네트워크 관리기술에 대한 필요성이 증가하고 있으며, 특히 네트워크 보안관리를 위한 새로운 패러다임으로 정책기반의 네트워크 관리 기술이 도입되고 있다. 또한 인터넷 위협에 대응하기 위한 네트워크 전반에 걸친 완벽한 관리 메커니즘이나 침입대응 시스템은 없으므로 인터넷 위협에 대한 네트워크 전반에 걸친 대응 메커니즘과 함께 실제 침입에 대응하기 위한 시스템의 개발도 침입탐지 시스템을 중심으로 활발히 이루어지고 있다. 기존의 침입탐지 시스템 관련 연구[4,5]들을 살펴보자면 대규모의 하부구조를 지닌 네트워크에서의 정보 수집/분석이 각각 전담 시스템에서 수행되는 경우가 많았으며 또한 네트워크 기반 침입탐지 시스템이라 할지라도 갈수록 다양해지는 침입에 대해 능동적으로 대처하기에 어려움이 많았다. 따라서 최근 침입 탐지 시스템에 데이터 마이닝 기법을 적용하여 많은 양의 감사데이터를 효율적으로 분석하거나 자동화된 침입탐지 모델을 구축하는 등의 연구가 활발히 진행되고 있다. 침입 탐지 시스템은 정상 행위의 프로파일이나 공격 기법의 시나리오를 구축하기 위해서는 많은 양의 시스템과 네트워크 감사 데이터를 정확하고 효율적으로 분석해야 한다. 따라서 이 논문에서는 대량의 감사데이터를 효율적으로 분석하기 위한 마이닝 시스템을 설계하고 구현한다. 구현된 마이닝 시스템은 이미 구축된 침입탐지 모델을 갱신하거나 새로운 프로파일을 추가하는데 활용할 수 있으며 이는 침입 탐지 시스템의 효율성을 증가시킬 수 있을 것이다. 감사데이터의 분석을 위해서 먼저 패킷데이터를 전처리하여 마이닝을 할 수 있는 데이터로 저장하는 모듈을 구현하였다. 이 논문의 구성은 2절에서는 관련연구로서 감사데이터의 특성과 침입탐지에서 적용되는 마이닝 기법에 대해서 간략히 설명하고, 3절에서는 이 논문에서 구현한 마이닝 기법과 패킷 데이터 전처리 과정에 대한 프로세서를 설명하고 4절에서는 실제 데이터를 가지고 실험한 결과를 보여준다. 마지막으로 5절에서 결론을 맺는다.

2. 관련연구

2.1 감사데이터

데이터 마이닝 기법을 적용할 도메인으로서 감사 데이터[3]의 특징들에 대해서 간단히 살펴 보고자 한다. 첫째 감사 데이터는 바이너리 형태이고 비구조화된 형태로 시간에 의존적인 원시 데이터이다. 데이터 마이닝을 위해서는 먼저 가능한 형태인 ASCII 형태로 전처리 과정이 이루어 져야 한다. 두 번째로 감사데이터는 네트워크와 시스템 의미를 가진 정보를 포함하고 있다. 마지막으로 감사데이터는 고속과 고용량의 스트림 데이터이다. 감사

† 이 연구는 한국전자통신연구원의 보안게이트웨이 연구팀의 위탁과제로 수행된 것임.

메커니즘들은 모든 네트워크와 시스템 행위를 기록하도록 설계되어 있기 때문이다.

2.2 데이터 마이닝 기법

데이터 마이닝 기법 들 중에서 감사 데이터 마이닝을 위해 유용한 기법[2]들을 소개한다.

● 연관 규칙

대량의 감사데이터로부터 적당한 특성이나 패턴 탐사를 위한 속성간의 연관성을 추출하기 위해 유용하다.

● 빈발 에피소드

시간 기반 네트워크와 시스템 행위들의 빈발한 패턴들은 시간과 통계 측정을 합하기 위해 중요하게 제공된다.

● 분류

침입 탐지에서 이상적인 애플리케이션은 사용자나 프로그램에 대한 감사 데이터가 "정상" 과 "비정상"인지를 충분히 수집하게 될 것이다.

3. 마이닝 시스템 설계

대량의 감사데이터 분석을 위한 마이닝 시스템 구현에 대해서 설명한다. 마이닝 시스템의 전체 아키텍처는 그림 3.1 과 같다.

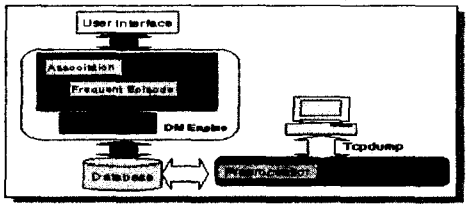

그림 3.1 마이닝 시스템 아키텍처

위 그림에서처럼 전체 아키텍처는 크게 두 부분으로 구분되어 있다. 먼저 패킷 데이터 전처리 과정 부분과 전처리 된 데이터를 마이닝 하는 부분으로 수행된다.

3.1 패킷 데이터 전 처리 프로세서

패킷 데이터 전처리 프로세서는 정보를 전달하는 유틸리티 소프트웨어인 TCPDUMP를 통해서 만들어진 원시 패킷 파일을 입력하여 패킷에 대한 시간정보와 패킷정보로 변환시켜 준다.

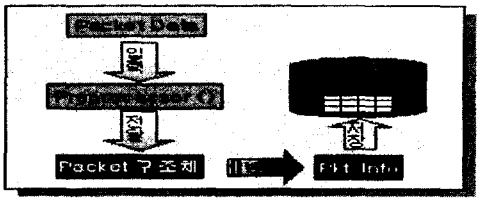

그림 3.2 패킷 전 처리 과정

전 처리 프로세서 수행 과정은 위 그림 3.2처럼 원시 패킷 데이터를 입력받아 ASCII 형태로 변환 후 데이터베이스에 저장되는데, 이때 각각의 프로토콜별로 분류하여 각 테이블에 저장된다.

3.2 마이닝 시스템 설계

이 논문에서는 항목간의 연관성 분석과 시퀀스 패턴에 대한 분석을 위해서 연관규칙 과 빈발 에피소드 기법을 적용하여 설계 및 구현하였다.

3.2.1 연관규칙 마이너

연관규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙[1]이다. 연관규칙을 탐사하는 문제는 기본적으로 미리 결정된 최소 지지도 이상의 트랜잭션 지지도를 갖는 항목집합들의 모든 집합들인 빈발항목집합(large itemset)을 찾아내어 연관규칙을 생성하는 단계로 이루어진다. 또한 관심도가 높은 속성들에 대해서 마이닝을 수행하도록 하기 위해서 이 논문에서는 키 속성제약사항(Axis Attribute)의 개념을 도입하여 관심 속성들에 대한 마이닝을 수행토록 하였으며, 이런 속성들은 사용자가 임의로 선택할 수 있도록 하였다.

확장된 알고리즘에서는 크게 3단계로 나뉘어 마이닝을 수행한다.

● 빈발항목 집합을 생성하는 단계

선택된 항목들 중에서 최소 지지도를 만족하는 항목들만을 추출하여 빈발항목 집합을 생성.

● 연관 규칙을 생성하는 단계

빈발항목들간의 최소지지도를 가지고 최소 신뢰도를 계산하여 연관규칙을 생성.

● 최종 룰 생성 단계

이전 단계에서 만들어진 룰에 대해서 최소 신뢰도 를 만족하는 최종 룰, 즉 Conf(R) ≥ min_conf 만을 생성하여 룰 테이블에 저장하게 된다.

위와 같이 3단계로 속성들간의 연관성을 마이닝 하여 많은 양의 감사데이터를 효율적으로 분석할 수 있으며 키 속성제약사항에 따라 관심있는 속성들간의 연관성을 분석하며, 불필요한 룰의 생성을 줄일 수 있다.

3.2.2 빈발에피소드 마이너

감사데이터 특성상 속성들 간의 상관관계 보다는 튜플들 간의 상관관계를 고려하고, 키 속성제약사항(Axis Attribute)을 적용함으로써 후보항목 생성 시 관심 있는 속성만을 고려할 수 있다.

빈발 에피소드는 4 단계로 나뉘어 마이닝을 수행한다.

● 윈도우 시간 별 항목 생성 단계

선택된 속성들로 이루어진 투플들에 대해서 주어진 time window에 의해서 투플 들을 정렬 즉, End_time - Start_time + window_width, 하여 후보 항목 생성을 위해 윈도우 시간 별로 테이블을 생성.

● 후보 에피소드 생성 단계

윈도우 시간별로 정렬된 테이블을 가지고 후보 에피소드 집합을 생성.

● 빈발 에피소드 생성 단계

생성된 후보 에피소드 집합에서 최소빈발도를 만족하는 에피소드들을 추출하여 빈발한 에피소드집합 생성

● 최종 에피소드 생성 단계

생성된 빈발 에피소드들로부터 최소 신뢰도를 만족하는 빈발 에피소드를 생성해 낸다.

위의 단계로 마이닝을 수행함으로써 규칙 생성 시 불필요한 에피소드 항목들이 많아지는 것을 감소시킬 수 있다.

4. 실험

이 절에서는 앞에서 설계한 마이닝 시스템을 가지고 실제 데이터를 입력하여 실험한 결과를 보여준다.
실험은 Window2000에 메모리 256 머신을 사용하였고, 서버측에서는 Linux에 DBMS 로서 Oracle 8.1.7을 사용하였다. 실험 데이터는 DARPA 데이터 1주1일 데이터와 실제 TCPDUMP를 통해서 생성한 500개의 패킷 데이터를 가지고 실험을 하였다.

4.1 연관규칙 마이너

연관규칙 마이너를 수행하기 위해서는 최소 지지도와 최소신뢰도를 미리 결정을 해주어야 한다. 이 실험에서는 최소 지지도를 60%와 최소 신뢰도 60%를 주었다. 항목간의 연관성을 고려하여 Source_IP, Source_Port, Destination_IP, Destination_Port, Service, Flag 항목들을 선택하였다. 실험결과는 표 4.1과 같은 연관규칙이 생성되었다.

표 4.1 생성된 연관규칙

Rule : sport(80), service(smtp) ==> flag(SF), [7,100] [support, confidence]
의미 : source port (80) 이고 service (smtp)가 전체 데이터중에서 70%이상 나타나고 이들 항목다음에 flag(SF)가 오는 항목이 100% 나타난다.

위 표 4.1에서처럼 source_port 와 service 그리고 flag는 서로 연관성을 가지고 있다고 할 수 있다.

4.2 빈발에피소드 마이너

빈발에피소드 마이너를 수행하기 위해서는 최소 빈발도, 최소신뢰도, 타임윈도우 크기등을 미리 결정을 해주어야 한다. 타임 윈도우 크기는 주어진 윈도우 크기별로 테이블을 정렬하는데 사용된다. 이 실험에서는 최소 빈발도 10%와 최소 신뢰도 60%, 타임윈도우 크기는 3을 주었다. 주어진 시간내에 유사한 패턴들이 빈발한 사건으로 발견되는지에 대해 알아보기 위해서 Source_IP, Source_Port, Destination_IP, Destination_Port, Flag Service 항목을 선택하였다.

표 4.2 생성된 빈발 에피소드 규칙

Rule : 1 203.255.74.102 210.155.167.10 9156 21 tcp 50 ==> 7 203.255.71.10 210.155.167.10 9158 21 tcp 50 [10,60,10] [frequent, confidence, sec]
의미 : 1 203.255.74.102 210.155.167.10 9156 21 tcp 50 ==> 7 203.255.71.10 210.155.167.10 9158 21 tcp 50 패턴이 10초 동안 전체 데이터 중에서 10회이상 발생하며 60%이상을 만족한다.

위 표4.2에서처럼 같은 destination_ip 로 21번 port를 이용하여 tcp 로 접근하는 패턴이 자주 발생한다는 것을 알 수 있다.

위 실험을 통해서 생성된 규칙들은 침입탐지 시스템에 적용하여 자동화와 성능향상에 도움을 줄 수 있으며 능동적인 정책모델 구축에도 활용될 수 있다.

5. 결론

대량의 데이터로부터 유용한 정보를 추출하거나 데이의 양을 감소시키는 기능을 제공하는 마이닝 기법을 침입탐지시스템에 적용하여 감사데이터를 효율적으로 분석하기 위한 마이닝 시스템을 설계하고 구현하였다. 먼저 패킷 로우데이터를 전처리하여 마이닝을 수행 할 수 있는 데이터로 저장하는 모듈을 구현하였으며 연관규칙마이너와 빈발마이너를 구현하여 속성간 연관규칙이나 빈발 시퀀스를 탐사하였다. 구현된 마이닝 시스템은 일반적인 트랜잭션데이터베이스에서의 마이닝과는 다른 감사데이터의 특성을 고려한 마이닝을 수행할 수 있도록 하였으며 연관규칙 기법을 이용하여 감사데이터 속성간의 연관성을 탐사하고, 빈발에피소드 기법을 적용하여 주어진 시간 내에서 상호 연관성 있게 발생한 이벤트들을 모음으로써 연속적인 시간간격 내에서 빈번하게 발생하는 사건들의 발견과 알려진 사건에서 시퀀스의 행동을 예측하거나 기술할 수 있는 규칙을 생성할 수 있다.

향후 생성된 연관규칙과 빈발시퀀스를 도메인 지식으로 하는 분류 마이너를 구현하여 감사데이터의 정상/공격여부를 판단할 수 있는 침입탐지 시스템을 위한 통합 데이터 마이닝 시스템 개발을 계속 수행할 것이다.

참고문헌

[1] R.Agrawal, T.Imielinski, and A.Swami, Mining association rules between sets of items in large databases. In Processings of the ACM SIGMOD Conference on Management of Data, pages 207-216, 1993.

[2] Wenke Lee, Salvatore J. Stolfo, Data Mining Approaches for Intrusion Detection, In Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, January 1998.

[3] Wenke Lee, Salvatore J. Stolfo,and K.W.Mok , Mining audit data to build introduction detection models. In Proceedings of the 4th Internalional conference on Knowledge Discovery and Data Mining, New York, NY, August 1998. AAAI Press.

[4] D. Anderson, T. Frivold, A. Valdes, Next-generation intrusion detection expert system(NIDES), Technical Report SRI-CLS-95-07, May 1995.

[5] James Cannady, Jay Harrell, A Comparative Analysis of Current Intrusion Detection Technologies, Feb. 1998.