감사데이터 분석을 위한 데이터 마이닝 기법의 연구

허문행*
*안양대학교 디지털미디어학과
e-mail : moonh@anyang.ac.kr

Research of Data Mining Techniques for the Audit Data Analysis

MoonHeang Huh*
*Dept of Digital Media, Anyang University

요 약

최근 네트워크의 발달로 인해 네트워크 감사데이터의 양이 점점 증가하고 있다. 이렇게 증가하는 네트워크 감사데이터를 능동적이고 효율적으로 분석하기 위한 연구가 진행되고 있다. 하지만 지금까지의 연구들은 특정 침입탐지시스템에 제한되어 데이터 마이닝 기법을 적용하여 감사데이터를 분석하고 침입탐지모델을 구축하는 연구였다. 따라서 이 논문에서는 특정 침입탐지시스템에 의존하지 않고 감사데이터를 효율적으로 분석하여 알려지지 않은 공격패턴이나 규칙들을 발견하여 보안정책 실행시스템에서 활용할 수 있도록 하기위한 감사데이터 분석 마이너를 설계하고 구현하였다. 보안관리자는 구현된 감사데이터 마이너를 이용하여 원하는 정보를 가공 추출하여 고수준의 의미추출에 이용할 수 있다.

1. 서론

기존의 침입탐지 시스템[2] 들은 대규모의 하부구조를 지닌 네트워크에서 정보의 수집 및 분석이 각각 전담 시스템에서 수행되는 경우가 많았다. 네트워크 기반 침입탐지시스템[4]이라 할지라도 갈수록 다양해지는 침입에 대해 능동적으로 대처하기에 어려움이 있다. 침입 탐지 시스템은 정상 행위의 프로파일이나 공격 기법의 시나리오를 구축하기 위해서 많은 양의 시스템과 네트워크 감사 데이터를 정확하고 효율적으로 분석해야 한다. 그래서 최근, 침입 탐지 시스템에서는 데이터 마이닝 기법을 적용하여 많은 양의 감사데이터를 분석하여 침입탐지 모델을 구축[3] 하고 있다. 즉 침입탐지 시스템에서는 침입 탐지 모델을 구축하기 위해서 데이터 마이닝 기법으로 연관 규칙, 패턴 마이닝을 적용하여 빈발한 패턴들을 탐사하고, 분류 기법을 적용하여 RIPPER[5] 라는 규칙 학습 프로그램으로 결과를 생성한다. 이러한 연구들은 특히 오용 탐지 시스템의 경우로 침입을 판별하기 위하여 마이닝 기법 적용하여 감사데이터를 분석하고 침입탐지 모델을 구축하기 때문에 특정시스템에 의존적이다. 그러나 정책기반 보안관리 프레임워크에서는 보안정책 서버 하단 레이어에 보안 정책 실행시스템들이 존재하고 이러한 보안정책 실행 시스템들이 독립적으로 감사데이터를 분석해야 하는 특징이 있기 때문에, 지금까지 연구된 방식으로는 정책기반 보안관리 프레임워크에 적용하기 어렵다. 따라서 이 논문에서는 특정 시스템에 의존적이지 않고 보안정책 실행시스템을 지원하기 위한 감사데이터 분석 마이너를 설계하고 구현한다. 적용된 마이닝 기법은 기존의 마이닝 알고리즘을 확장하여 다차원 데이터 특성을 가진 감사데이터에 적합하도록 항목 제약사항을 추가한다. 구현된 마이너의 수행결과 생성된 규칙들은 보안정책실행시스템에서 새로운 공격유형을 추가하거나 프로파일 추가에 활용하도록 한다. 그리고 구현된 감사데이터 분석 마이너를 평가하고 검토한다.

2. 관련연구

데이터 마이닝 기법을 적용할 도메인으로서 감사데이터[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. 두번째로 감사데이터는 네트워크와 시스템 의미를 가진 정보를 포함하고 있다. 마지막으로 감사데이터는 고속과 고용량의 스트림 데이터이다. 감사 메커니즘들은 모든 네트워크와 시스템 행위를 기록하도록 설

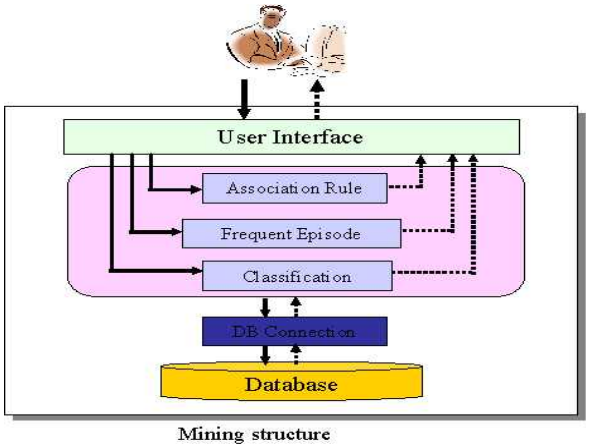
계되어 있기 때문이다. [1]은 감사 데이터로부터 프로그램 행위 데이터를 사용하여 탐지를 하였고, 신경망에서는 시스템 프로그램을 위한 이상 탐지 모델과 오용 탐지 모델을 학습하는데 사용할 수 있다. [2]는 사용자 쉘 명령어 와 이상 탐지를 분석하기 위한 알고리즘을 개발했다. 이러한 알고리즘들은 메타러닝 기법을 적용하여 침입 탐지 모델을 학습 하였다. 하지만 최근에는 네트워크의 광역화와 점점 더 다양해지는 침입 유형에 대해서 능동적으로 대처하기 위해 대량의 데이터를 분석하는 데이터 마이닝 기법을 적용한 연구들이 이루어지고 있다. Wenke가 1998년도에 자동화된 침입 탐지 모델을 구축하기 위해서 침입탐지 모델에 데이터 마이닝 기법을 적용하기 시작 했다. 대표적인 시스템으로 MADAMID (Mining Audit Data for Automated Models for Intrusion Detection)[2]가 있다.

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. 감사데이터 마이닝 시스템 설계

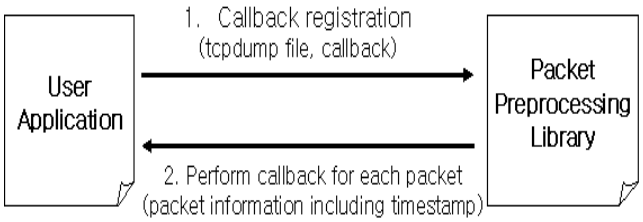
대량의 감사데이터 분석을 위한 마이닝 시스템 구현에 대해서 설명한다. 마이닝 시스템의 전체 아키텍처는 크게 두 부분으로 구분되어 있다. 먼저 패킷 데이터 전처리 과정 부분과 전처리 된 데이터를 마이닝 하는 부분으로 수행된다. [그림1]은 감사데이터 분석마이너의 전체 시스템 구조를 보여준다.



[그림 1] 감사데이터 분석 마이너의 구조

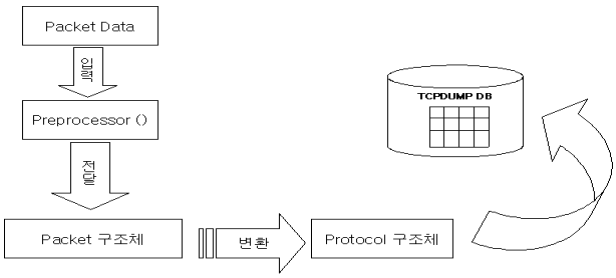
3.1 패킷 데이터 전 처리 프로세서

원시데이터 전 처리 프로세서 흐름도는 [그림 2]에서 보여준다.



[그림 2] 원시 데이터 전 처리 프로세서 흐름도

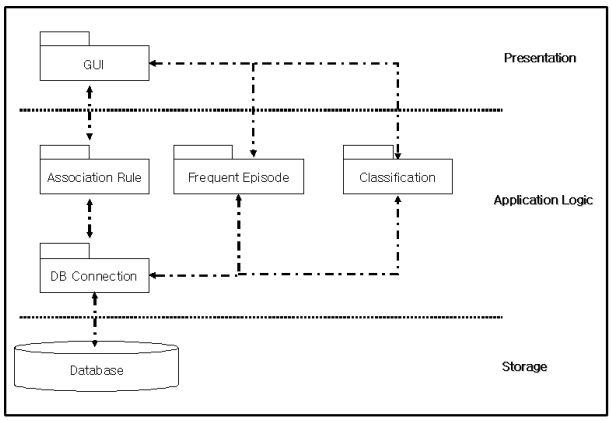
[그림3]에서처럼 Packet Preprocessing Library는 유틸리티 소프트웨어이며, User Application은 패킷 정보를 이용하여 사용 목적에 따라 작업할 일을 정의한다. User Application은 초기화 과정에서 Packet Preprocessing Library로 tcpdump 파일 정보와 그 파일을 이용하여 패킷단위로 수행할 callback function을 등록한다. 이후 Packet preprocessing utility를 수행시키면 라이브러리는 패킷 단위로 user callback function을 호출한다. [그림 3]은 전체적인 원시 데이터 전처리 프로세서 과정을 거쳐 데이터베이스에 저장하는 과정을 보여준다.



[그림 3] 원시 데이터 전 처리 프로세서 과정

3.2 마이닝 시스템 설계

감사데이터 분석 마이너 클래스는 three-layer 구조로 설계된다. [그림 4]는 감사데이터 분석 마이너 클래스의 three-layer를 표현한 것이다.



[그림 4] 감사데이터 분석 마이너 클래스의 three-layer

[그림 4]에서처럼 상단 레이어 에는 결과를 입출력할 수 있는 사용자 인터페이스 부분이고 중간 레이어 에는 3 개의 감사데이터 분석 마이너들로 구성되어 있다. 마지막으로 하단 레이어 에서는 물리적인 부분으로 저장소인 데이터베이스가 있다.

【연관규칙 마이너】

연관규칙은 항목 집합으로 표현된 트랜잭션에서 각 항목간의 연관성을 반영하는 규칙[1]이다. 연관규칙을 탐사하는 문제는 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 마이닝하여 많은 양의 감사데이터를 효율적으로 분석할 수 있으며 키 속성제약사항에 따라 관심있는 속성들간의 연관성을 분석하며, 불필요한 룰의 생성을 줄일 수 있다.

【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.

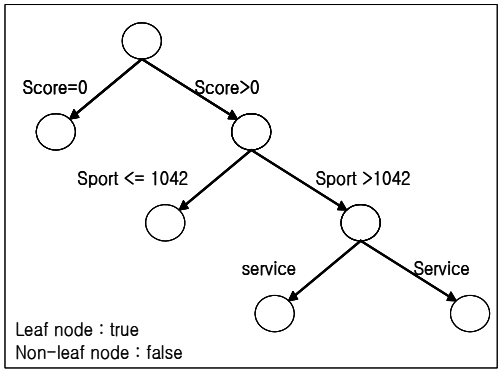
【분류마이너】

분류 마이너를 수행한 결과는 [그림 5,6]와 같이 분포도와 트리 형태로 보여준다. [그림 5]의 분포도에서 보듯이 공격의 80%가 normal 상태로 나타났다. [그림 6]의 결정 트리를 보면 여기서는 테스트 속성이 "score" 로 트리를 결정하기 때문에 score 속성 값에 따라서 분류를 한다.

[그림 5] 분류 분포도

그래서 score=0 이면 true 로서 공격으로 인정되는 회수가 전혀 없었기 때문에 normal이라고 분류하고, score > 0 이면, 일단은 공격을 시도했기 때문에 어

떠한 공격으로 분류를 하게 될 것이다.

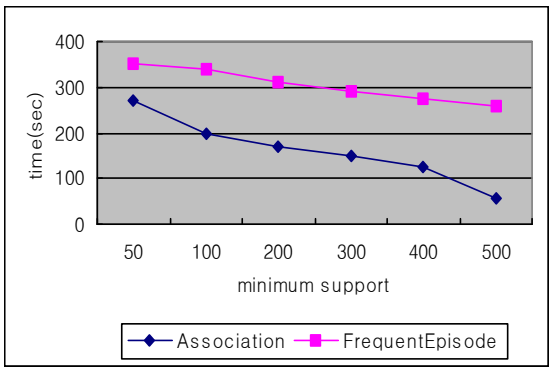


[그림 6] 분류 결정트리

두 번째로 sport 속성을 가지고 비교한다. sport 속성에 대해서도 1042 포트에 대해서 다시 한번 분류가 이루어진다. 그래서 score > 0 이고 sport >=1042 면 일단은 공격으로 의심을 한다.

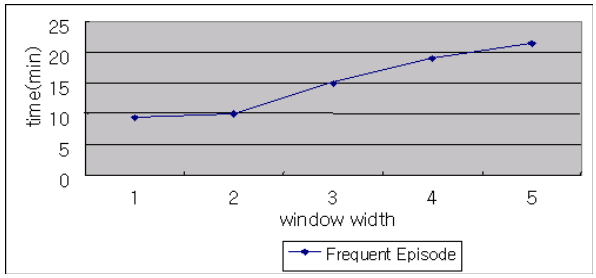
4. 실험

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 32000개의 패킷 데이터를 가지고 실험을 하였다. 최소지지도를 줄여 가면서 성능 평가를 해보았다. 최소지지도는 20%, 15%, 10%, 5% 씩 줄여가면서 실험을 하였다. [그림 7]은 연관 규칙과 빈발에피소드 마이너 수행 시 최소 지지도 변경에 따른 수행 시간을 보여준다.



[그림 7] 최소 지지도 변경에 따른 수행시간

빈발 에피소드에서 윈도우 폭을 증가 시키면서 빈발 에피소드 마이너에 대한 수행시간을 [그림 8]에서 보여준다.



[그림 8] 윈도우 폭에 따른 수행시간

[그림 8]에서 보는 것과 같이 빈발에피소드 마이너는 윈도우 폭을 증가시킴으로서 많은 수의 윈도우 테이블을 생성하기 때문에 시간이 오래 걸린다는 단점이 있었다.

5. 결론

본 논문에서는 침입탐지시스템의 감사데이터를 효율적으로 분석하기 위한 마이닝 시스템을 설계하고 구현하였다. 구현된 마이닝 시스템은 일반적인 트랜잭션데이터베이스에서의 마이닝과는 다른 감사데이터의 특성을 고려하여 지식탐사을 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 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.