시퀀스 패턴 마이닝 기법을 적용한 침입탐지 시스템의 경보데이터 패턴분석

신문선*
*안양대학교 교양학부
e-mail:msshin9@anyang.ac.kr

MoonSun Shin*
*Division of Liberal Arts, Anyang University

요 약

침입탐지란 컴퓨터와 네트워크 자원에 대한 유해한 침입 행동을 식별하고 대응하는 과정이다. 점차적으로 시스템에 대한 침입의 유형들이 복잡해지고 전문적으로 이루어지면서 빠르고 정확한 대응을 할 수 있는 시스템이 요구되고 있다. 이에 대용량의 데이터를 분석하여 의미 있는 정보를 추출하는 데이터 마이닝 기법을 적용하여 지능적이고 자동화된 탐지 및 경보데이터 패턴 분석에 이용할 수 있다. 본 논문에서는 경보데이터 패턴 분석을 위해 시퀀스패턴기법을 적용한 경보데이터 마이닝 엔진을 구축한다.구현된 경보데이터 마이닝 시스템은 기존의 시퀀스 패턴 알고리즘인 PrefixSpan 알고리즘을 확장 구현하여 경보데이터의 빈발 경보시퀀스 분석과 빈발 공격시퀀스 분석에 활용할 수 있다.

1. 서론

네트워크 기반 침입탐지 시스템은 침입이나 악의적인 공격에 대하여 많은 양의 경보를 발생시켜 보안 관리자에게 알려준다. 그러나 갈수록 다양해지는 침입에 대해 능동적으로 대처하기에 어려움이 많았다. 따라서 최근 침입 탐지 시스템에 데이터 마이닝 기법을 적용하여 데이터베이스로 구축된 다량의 감사데이터 혹은 경보데이터를 효율적으로 분석하기 위한 연구[6]가 활발히 진행되고 있다.

침입탐지 시스템에서는 패킷을 가지고 미리 정해놓은 규칙들과 비교를 해서 공격을 탐지하게 된다. 또한 침입탐지 시스템에서는 이런 침입에 대해 경보데이터를 생성 하며 과거에 비해 스위칭 기술의 발달과 Bandwidth의 향상 등 네트워크 기술의 발달로 인해 네트워크 상의 트래픽이 증가하게 됨에 따라 생성되는 경보 데이터의 양도 많아지고 있다. 따라서 다량의 데이터에서 유용한 정보를 추출하는 작업인 데이터 마이닝 기법이 다양한 형태로 적용되어 침입탐비 및 감사데이터 분석 등에 활용되고 있다.

일반적인 prefixSpan 알고리즘은 트랜잭션 데이터베이스를 그 대상으로 하고 있으나 경보 데이터는 트랜잭션 데이터와는 다소 다른 특성을 가지므로 경보 데이터의 특성을 고려하여 데이터의 전처리 및 기존의 prefixSpan알고리즘을 확장 설계한 경보데이터 시퀀스패턴 마이너를 설계하고 구현한다. 이 논문에서 구현되는 시퀀스 패턴 마이너는 시퀀스들로 구성 되어 있는 경보데이터들의 빈발한 시나리오를 탐사하여 탐사된 패턴 내에서 침입탐지 시스템에 적용 가능한 유사패턴을 찾아낼 수 있으며 또한 알려진 공격에 대해 지지도를 기반으로 다음 공격에 대한 예측도 가능하다.

2. 관련 연구

콜럼비아 대학의 Wenke Lee[11]는 텔넷 로그데이터, 네트워크 쉘 커멘드 및 Tcpdump와 같은 감사데이터에 연관규칙, Frequent Episodes등의 데이터 마이닝 기법을 적용하였고. M. Joshi[12]는 감사 데이터를 미리 정의된 여러 개의 항목들 중 하나로 매핑 하여 각각의 항목들을 커다란 그룹으로 표현하는 방법인 분류 기법의 적용 및 분류 기법의 정확도를 높이기 위한 bagging, boosting 방식을 연구하였다. 그리고 Park[10]은 시퀀스 패턴 탐사를 이용하여 정상행위의 프로파일을 작성하여 비정상 행위를 탐지해 내는 방법을 연구하였고 Shim[15]은 결정트리를 이용하여 공격들을 정해놓은 몇 개의 카테고리로 구분하는 방법을 연구 하였다.

최근 침입 탐지 시스템에 데이터 마이닝 기법을 적용하여 데이터베이스로 구축된 다량의 감사데이터 혹은 경보데이터를 효율적으로 분석하기 위한 연구[6,11,13]가 활발히 진행되고 있다.

이 논문에서는 침입탐지 시스템에서 효율적으로 경보데이터를 분석하고 공격 시퀀스 및 경보시퀀스의 새로운 패턴을 찾아내어 능동적인 대응을 하기 위해 경보데이터 패턴 탐사 마이닝 기법을 제안한다. 논문에서 제안한 경보데이터 패턴 마이너는 prefixSpan 알고리즘[7,8]을 확장 설계한 것으로 경보데이터의 특성에 맞게 설계되었으며 기존의 빈발 에피소드 탐사기법보다 패턴 탐사의 성능이 우수하다.

3. 시퀀스 패턴 마이너 설계및 구현

시퀀스패턴 마이닝은 일련의 시퀀스로부터 빈번하게 발생하는 시퀀스들을 찾는 기법[7, 8]이다. 또한 시퀀스패턴 마이닝 기법 중의 하나인 prefixSpan[8]은 패턴을 탐사하는데 있어 후보 패턴 생성 비용을 줄이기 위해 단계별로 분할된 prefix-Projected 데이터베이스를 구성하여 후보 패턴의 지지도 계산을 위한 탐색공간을 줄이는 시퀀스패턴 탐사 방법이다. 즉, prefixSpan(prefix-projected Sequential pattern mining)은 기존의 Apriori-Based 방법들이 후보 패턴을 만들고 그 후보 패턴이 데이터베이스에 몇 번 나오는가 세느라 시간이 걸리는 단점을 없애기 위해, 후보 패턴을 만들지 않으면서 빈번한 패턴을 찾는 방법이다. prefixSpan은 단계별로 분할된 prefix-Projected 데이터베이스들을 구성하여 후보 패턴들의 지지도 계산을 위한 탐색 공간을 줄이고 시퀀스 데이터베이스에 대한 prefix-projection을 반복적으로 수행하여 패턴을 찾아낸다. prefixSpan은 '모든 빈발 시퀀스들은 빈발한 prefix들의 확장에 의해 발견할 수 있다'는 사실에 기인하여 빈발한 prefix에 대해서만 데이터베이스 projection을 수행한다. 침입탐지 시스템으로 들어오는 네트워크 패킷들을 탐지하여 탐지에 불필요한 부분들을 제거하는 필터링 과정을 거쳐 데이터를 정제한다. 다음 단계로 정제된 데이터는 여러 침입 패턴들이 저장되어 있는 센서 이벤트와 비교하여 침입을 판별하게 되고 침입으로 판별된 패킷들은 경보데이터 스키마 형태로 저장된다. 이 저장된 경보 데이터를 가지고 경보데이터 분석기에서 데이터 마이닝 과정을 수행하여 시퀀스 패턴들을 탐사하고 탐사된 시퀀스 패턴은 오퍼레이터에게 전달되어 분석과정을 거친 후 경보관리기 모듈을 통해 새로운 침입 규칙으로 생성되어 센서 이벤트에 저장된다.

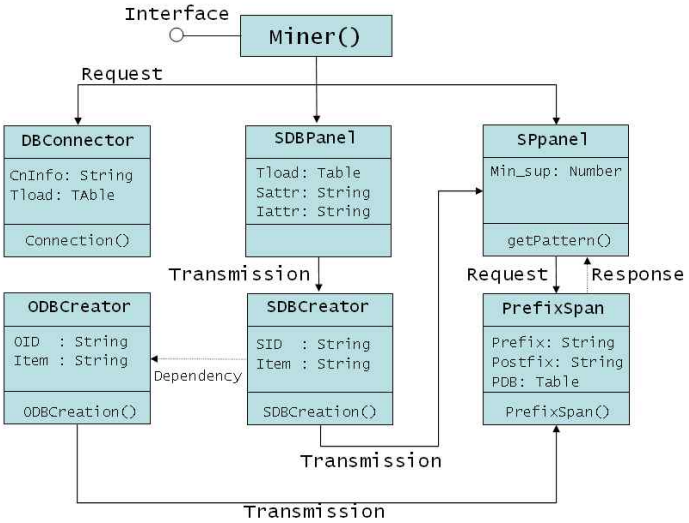
시퀀스 패턴 마이너는 새로운 경보가 발생하게 되면 발생된 경보데이터 패턴들중 저장된 규칙 혹은 시그너쳐와 같은 경보가 시퀀스적으로 발생되었는지를 체크하여 악의적인 공격을 탐사하는 기능을 수행하는 경보데이터 분석기를 지원하게 된다.

경보데이터로부터 유용하다고 여겨지는 정보를 찾아내기 위해 데이터 마이닝 기법을 적용하는데 있어 기존의 알고리즘을 이용할 경우 기준 속성의 모호성 및 불필요한 속성들마저도 마이닝 과정에 포함시켜 비용의 증가를 초래할 수 있다. 따라서 본 논문에서는 전처리 및 확장된 알고리즘을 제안하여 경보데이터 시퀀스패턴 탐사를 수행하였다.

특히 무의미한 시퀀스 패턴의 필터링을 위하여 기준속성이라는 개념을 도입하였다. 기준속성 이라는 것은 시퀀스를 생성하기 위해 그룹화 할 수 있는 트랜잭션 데이터베이스에서의 TID와 같은 속성이고 선택속성은 아이템을 생성하기 위해 선택되는 속성들이다. 이를 이용하여 경보데이터 내에서 필요한 속성들만으로 이루어진 가시적으로 볼 수 없었던 새로운 시퀀스들의 집합을 생성할 수 있으며 패턴 탐색 시 불필요한 패턴들을 탐색하여야 하는 비용을 절감할 수 있다. 또한 상대적으로 많은 시간이 소요되는 빈발 에피소드 탐사에서의 효율성을 개선할 수 있는 시퀀스패턴 탐사기법이다.

[그림 1]은 시퀀스 패턴 마이너의 클래스 다이어그램이다. DBConnector 클래스는 접속정보를 입력받아 Connection()함수를 실행하게 되고 실행 후에는 기본적으로 전체 테이블 리스트를 로딩하게 된다. 두 번째로 SDBpanel을 호출하게 되는데 이는 기준 속성과 선택 속성을 선택하기 위한 인터페이스로써 사용자가 선택한 값들을 SDBCreator로 전달하게 되고 SDBCreator 클래스에서는 시퀀스 데이터베이스와 시퀀스들의 순서만을 저장해 두는 Order Database를 생성하게 된다.

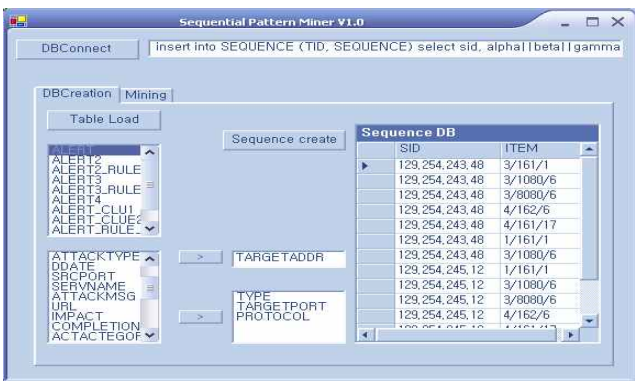
Order Database는 시퀀스 데이터베이스의 아이템들 중 유니크한 값들만을 추출하여 시퀀스적인 번호인 OID(일련번호)와 아이템(ITEM)들로 구성되어있는 테이블로서 prefixSpan 알고리즘 내에서 항목들 간의 비교를 위해 사용된다.



[그림 1] 경보데이터 시퀀스 패턴 마이너의 Class Diagram

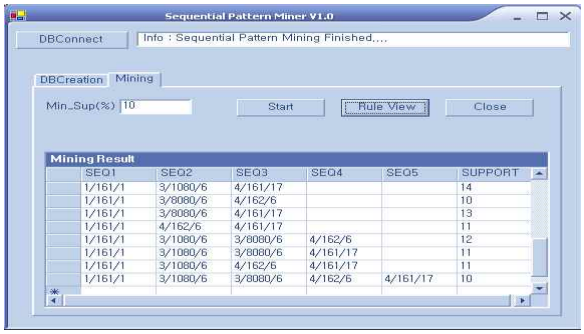
이렇게 생성된 시퀀스 데이터베이스는 SPpanel로 전달되고 SPpanel에서는 사용자 입력값인 최소 지지도를 가지고 알고리즘이 정의되어 있는 prefixSpan클래스를 호출하여 시퀀스패턴을 탐사한 후 생성된 최종 규칙들을 데이터베이스에 저장하고 SPpanel을 통해서 결과를 출력하게 된다.

우선 경보데이터를 저장하는 데이터베이스로는 Oracle 8i를 사용하였고, 시퀀스 패턴 탐사 마이닝 연산인 Miner() 모듈은 C#으로 구현하였다. 또한 경보데이터 시퀀스 패턴 탐사 모듈인 SPMiner의 User Interface는 C#에서 제공되는 컨트롤을 활용하여 구현하였다. SPMiner는 [그림 2]와 같은 사용자 화면을 가지며 테이블 로딩, 마이닝 대상 속성 선택, 시퀀스 데이터베이스 생성 등의 주요 전처리 기능을 위한 컴포넌트들을 포함한다.



[그림 2] 접속 및 시퀀스 데이터베이스 생성 패널

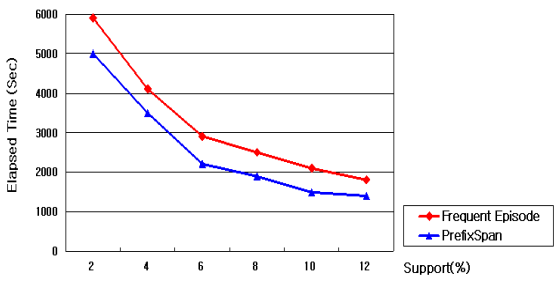
SPMiner에서의 시퀀스 패턴 탐사는 threshold값인 최소 지지도를 입력받아 전처리 단계에서 생성된 시퀀스 데이터베이스를 대상으로 prefixSpan() 함수를 호출하여 시퀀스패턴 탐사를 수행하게 된다. [그림 3]은 지지도 입력 및 실제 탐사과정 수행 후의 결과 화면을 나타내고 있다.



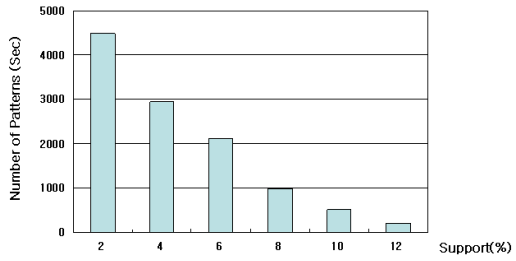
[그림 3] 결과 화면

4. 구현및 실험

구현된 경보데이터 시퀀스 패턴 마이너에 대한 실험은 세 가지로 수행 되었다. 첫 번째는 찾아낸 시퀀스 패턴의 집합들이 침입탐지 시스템에 적용 가능한 패턴인지를 분석하고, 두 번째로 패턴의 길이에 따른 수행시간 및 지지도 변화에 따른 패턴 탐사의 효율성에 대해서 실험한다. 기존에 구현 된 빈발 에피소드(Frequent Episode)와의 비교실험을 통해서 SPMiner이 성능이 뛰어나다는 결과를 얻을 수 있었다.탐사한 패턴의 효율성은 지지도를 기반으로 신뢰할 수 있다. 최소 지지도를 만족하는 항목들만이 빈발한 시퀀스로 탐사되기 때문에 최소 지지도와 같은 임계치 변화와 수행에 미치는 영향 관계를 분석하였다. 실험결과 [그림 4]와 같이 지지도를 2%씩 증가시킴에 따라 빈발 이동 패턴 탐사에 소요되는 시간이 감소함을 알 수 있었고, [그림 5]와 같이 지지도 변화에 따른 패턴 수의 변화도 확인 할 수 있었다. 지지도는 탐사될 패턴의 수에 영향을 미치는 중요한 요소이다. 이러한 임계치는 반복적인 패턴 탐사를 통하여 데이터의 규모 및 응용환경의 규모에 따라 적합한 수치를 설정할 수 있다.



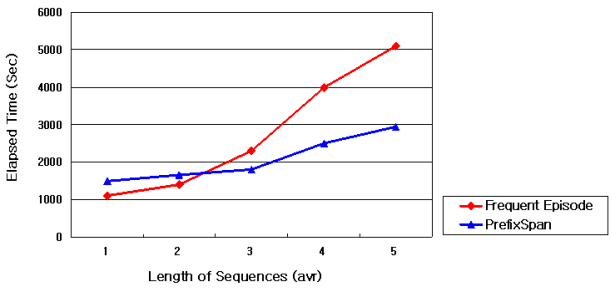
[그림 4] 지지도 증가에 따른 수행시간



[그림 5] 지지도 증가에 따른 패턴의 개수

시퀀스 길이 증가에 따른 패턴 탐사의 효율성은 시간 간격의 크기의 변화를 이용하여 평균 시퀀스의 길이를 판단하여 평가 할 수 있다. 그러나 제안된 prefixSpan 알고리즘은 시간 간격 윈도우 내에서 빈발 패턴을 탐사하는 빈발에피소드와는 달리 전체 시퀀스에서 빈발한 모든 패턴을 찾아내는 방법이다. 따라서 이 실험에서는 전체 데이터 집합에서 일정 시간 간격으로 데이터를 추출하여 시퀀스의 평균 길이를 측정한 후 시퀀스의 길이 변화에 따른 수행시간의 관계에 대해 실험하였다.

실험결과 [그림 6]과 같이 시퀀스의 길이 증가에 따라 빈발에피소드 기법과의 수행시간의 차이를 확인할 수 있다. 작은 길이의 시퀀스에서는 빈발에피소드의 수행 시간이 prefixSpan 알고리즘에 비해 빠른 수행 성능을 보이고 있으나 점점 시퀀스의 길이가 길어짐에 따라 prefixSpan 알고리즘 쪽의 효율이 좋다는 것을 확인 할 수 있다.



[그림6] 시퀀스 증가에 따른 수행시간

5. 결론

이 논문에서는 침입 탐지 시스템의 경보데이터 패턴 분석을 위해 경보데이터 시퀀스 패턴 마이너를 설계하고 구현하였다. 기존의 빈발 에피소드 탐사에서의 시간 성능을 개선하기 위해서 후보항목 생성을 하지 않는 시퀀스 패턴 탐사를 위하여 prefixSpan을 확장 적용하였다. 제안된 시퀀스패턴 마이너가 전체의 시퀀스를 대상으로 패턴 탐사를 수행하므로 빈발 에피소드 탐사와 비교하여 볼 때 긴 길이의 시퀀스에 대한 성능이 우수함을 확인 하였다.

참고문헌

[1] D. Anderson : Next-generation intrusion detection expert system(NIDES). Technical Report SRI-CLS-95-07 (1995)

[2] James Cannady : A Comparative Analysis of Current Intrusion Detection Technologies. http://iw.gtri.gatech.edu/papers/ids_rev.html (1998)

[3] M.S. Shin, K.H. Ryu : Data mining methods for alert correlation analysis. IJCIS to be appear (2003)

[4] R. Heady : The Architecture of a Network Level Intrusion Detection System. Technical Report. University of New Mexico, Department of computer Science (1990)

[5] D. Denning : An Intrusion Detection Model. IEEE Trans.Softw.Eng.,13(2), (1987)

[6] Usama M. Fayyad et al. : Advances in knowledge discovery and data mining. MIT Press (1996)

[7] Rakesh Agrawal, Ramakrishnan Srikant : Mining Sequential Patterns. ICDE (1995)

[8] J. Pei, J. Han : PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. ICDE'01 (2001)

[9] M.J. Lee, M.S. Shin, K.H. Ryu : Design and Implementation of Alert Analyzer with Data Mining Engine. IDEAL'03 (2003)

[10] J. C. Park, B. N Noh : Intrusion Detection Method Using the PrefixSpan Algorithm. KIPS'03 (2003)

[11] W. Lee, S. Stolfo : Data Mining Approach for Intrusion Detection. UESNIX'98 (1998)

[12] M. Joshi, et al : Predicting Rare Classes : Can Boosting Make Any Weak Learner Strong ACM SIGKDD'02 (2002)