경보데이터 패턴분석을 위한 순차패턴 알고리즘의 구현†

김헌웅, 신문선, 류근호, 장종수*
충북대학교 데이터베이스 연구실, *한국전자통신연구원
e-mail:ghime@dblab.chungbuk.ac.kr
*jsjang@etri.re.kr

Implementation of Sequential Pattern Mining algorithm For Analysis of Alert data.

Hohn Woong, Ghim, Moon Sun, Shin, Keun Ho, Ryu,
Jong Soo, Jang*
Database Laboratory, Chungbuk National University
*Electronics and Telecommunications Research Institute

요 약

침입탐지란 컴퓨터와 네트워크 자원에 대한 유해한 침입 행동을 식별하고 대응하는 과정이다. 점차적으로 시스템에 대한 침입의 유형들이 복잡해지고 전문적으로 이루어지면서 빠르고 정확한 대응을 필요로 하는 시스템이 요구되고 있다. 이에 대용량의 데이터를 분석하여 의미있는 정보를 추출하는 데이터 마이닝 기법을 적용하여 지능적이고 자동화된 탐지 및 경보데이터 분석에 이용할 수 있다. 마이닝 기법중의 하나인 순차 패턴 탐사 방법은 일정한 시퀀스 내의 빈발한 항목을 추출하여 순차적으로 패턴을 탐사하는 방법이며 이를 이용하여 시퀀스의 행동을 예측하거나 기술할 수 있는 규칙들을 생성할 수 있다. 이 논문에서는 대량의 경보 데이터를 효율적으로 분석하고 반복적인 공격 패턴에 능동적인 대응을 위한 방법으로 확장된 순차패턴 알고리즘인 PrefixSpan 알고리즘에 대해 제안하였고 이를 적용하므로써 침입탐지 시스템의 자동화 및 성능의 향상을 얻을수 있다.

1. 서론

정보화 사회의 활성화와 정보통신 인프라로서 인터넷의 중요성이 급속히 부각되는 반면에, 인터넷으로 인한 여러 가지 문제점들 또한 심각해지고 있다. 이러한 상황에서 네트워크 전반에서의 보안의 필요성은 시간이 지날수록 증대되고 있다고 할 수 있다. 이에 실제 인터넷 위협에 대한 시스템의 개발이 침입탐지 시스템(IDS: Intrusion Detection System)을 중심으로 활발히 이루어지고 있다[1,2,3]. 침입을 오용탐지와 비정상 사용 탐지로 세분화 하고 있으며 이러한 침입을 해석하고 분석할 수 있는 데이터베이스를 구축하고 있다. IDS에 대한 연구는 1996년을 기점으로 DARPA를 중심으로 IDS관련된 표준화의 움직임이 시작 되었으며 점차적으로 비정상 사용탐지 방향으로 연구가 진행중에 있다. 기존의 침입탐지 시스템 관련 연구들을 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 데이터베이스로 구축된 다량의 데이터를 효율적으로 분석하기 위한 연구가 활발히 진행되고 있다. 이 논문에서는 침입탐지 시스템에서 효율적으로 경보데이터를 분석하고 공격 시퀀스 및 경보시퀀스의 새로운 패턴을 찾아내어 능동적인 대응을 하기 위해 데이터 마이닝 기법중 PrefixSpan 알고리즘을 확장 설계하였다. 일반적인 PrefixSpan 알고리즘은 트랜잭션 데이터베이스에서의 데이터 마이닝 기법으로 경보 데이터는 트랜잭션 데이터와는 다소 다른 특성을 가지게 되므로 기존의 알고리즘을 경보 데이터의 분석에 적용할 경우 불필요한 많은 양의 규칙들이 발생하게 된다. 이 문제를 해결하기 위해 이 논문에서는 경보 데이터의 특성을 고려하여 데이터의 전처리 및 기존의 알고리즘을 확장 설계 한 PrefixSpan 알고리즘을 설계하고 구현하였다.

이는 경보데이터의 패턴 분석에 유용하게 이용 될 수 있으며 또한 생성된 규칙들은 보안정책을 수행하는 보안정책 서버에서 새로운 보안정책 수립에 이용

† 본 연구는 한국전자통신연구원 보안 게이트웨이 팀 및 한국과학재단 지정 청주대 RRC(정보통신 연구센터)의 지원으로 수행되었음.

할 수 있다. 논문의 구성은 2장에서 관련 연구로서 침입탐지시스템에서의 데이터 마이닝 접근방법에 대하여 기술하고, 3장에서는 순차패턴과 PrefixSpan의 정의 및 분석과 경보데이터의 특성을 고려한 PrefixSpan알고리즘을 제안하고 4장에서는 구현 및 테스트 데이터를 통한 실험 결과를 기술하였다. 마지막으로 5장에서는 결론 및 향후연구로 끝을 맺는다.

2. 관련연구

2.1. 침입 탐지

침입은 컴퓨터가 사용하는 자원에 대하여 무결성(Integrity), 기밀성(Confidentiality), 가용성(Availability)을 저해하는 일련의 행위와, 컴퓨터 시스템의 보안 정책을 파괴하는 행위를 말한다[4]. 이러한 침입에 대한 대비책으로 일반적으로 사용하고 있는것이 방화벽과 침입탐지 시스템이다. 침입탐지 시스템은 방화벽과 같이 네트워크를 통한 외부 침입을 차단하는 단계를 넘어서 침입사실을 감지해 이에 대응하도록 되어 있다. 또한 침입탐지 시스템은 침입방법을 자체적으로 내장하여 침입 행동들을 실시간으로 감지함은 물론 제어할 수 있는 기능들을 제공한다.[5]. 침입 탐지 기술은 오용 탐지 기법(Misuse Detection)과 이상 탐지 기법(Anomaly Detection)으로 분류되는데 오용 탐지는 침입의 여부를 판단하기 위하여 알려진 공격 기법의 패턴이나 시스템의 취약점을 이용하는 방법이다. 이 기법의 가장 큰 약점은 노출되지 않은 공격 기법이나 새로운 공격 기법은 탐지할 수 없다는 점이고 또한 이미 알려진 공격 기법이나 취약점의 코딩과 같은 수동적인 방법을 이용하여야 한다는 문제점이 있다.

이상 탐지는 정상 행위에서 이탈하는 행위를 침입으로 판단하는 기법이다. 정상 사용 패턴을 구축하기 위하여, 이상 탐지 기법에서는 특정 사용자나 프로그램의 CPU 사용량 같은 시스템 속성의 통계를 이용한다. 그렇지만 이러한 시스템의 특성을 선택하기 위하여 경험적인 방법을 이용한다는 문제점이 있으며, 새로운 프로그램이나 사용자가 추가될 경우에 그 프로파일을 새로이 작성해야 하는 오버헤드가 발생한다. 이렇듯 기존 방법에서의 문제점, 즉, 수동적인 접근 부분이나 새로운 시퀀스나 프로파일의 추가 부분에 효율성을 증대시키기 위하여 데이터 마이닝 기법을 이용할 수 있다.

2.2 침입 탐지 시스템에서의 데이터 마이닝

데이터 마이닝이란, 대량의 실제 데이터로부터, 이전에 잘 알려지지 않았지만, 묵시적이고, 잠재적으로 유용한 정보를 추출하는 작업이라고 정의된다[6]. 또한 앞에서 기술된 것처럼 침입 탐지 시스템은 정상 행위의 프로파일이나 공격 기법의 시나리오를 구축하기 위해서 많은 양의 시스템과 네트워크 데이터를 정확하고 효율적으로 분석해야 한다. 이러한 부분에 데이터 마이닝 기법을 도입함으로써 자동화된 침입 탐지 시스템의 구축뿐만 아니라 정책 구축의 효율성과 정확성을 또한 향상시킬 수 있다. 즉, 침입 탐지 시스템의 구축 과정에서 수동적이고 임의적인 요소를 최대한 배제하여 시스템의 구축과 성능의 효율성을 높이기 위해 데이터 마이닝 기법을 이용한다. 데이터 마이닝 작업은 일반적인 규칙을 발견하고 의사 결정 처리를 향상시키기 위해서 과거의 데이터를 얼마나 효율적으로 사용하는가에 관한 문제이다. 대표적인 마이닝 기법중 순차패턴 탐사는 일련의 시퀀스로부터 빈번하게 발생하는 시퀀스들을 찾는 기법이다[7]. 또한 순차패턴 탐사 기법중의 하나인 PrefixSpan은 패턴을 탐사하는데 있어 기존 Apriori-based 알고리즘의 단점이라 할 수 있는 후보 패턴 생성 비용을 줄이기 위해 단계별로 분할된 Prefix-Projected 데이터베이스를 구성하여 후보 패턴의 지지도 계산을 위한 탐색공간을 줄인다[8].

침입 탐지 시스템에서는 이를 이용하여 자주 반복되는 시퀀스 패턴을 탐지하고 규칙에 적용시키거나, 일련의 시퀀스로부터 예상되는 다음 공격의 행위를 예측할 수 있다.

3. 경보데이터 분석을 위한 PrefixSpan 알고리즘

3.1. PrefixSpan의 정의

PrefixSpan(Prefix-projected Sequential pattern mining)은 기존의 Apriori-Based 방법들이 후보 패턴을 만들고 그 후보 패턴에 데이터베이스에 몇 번 나오는가 세느라 시간이 걸리는 단점을 없애기 위해, 후보 패턴을 만들지 않으면서 빈번한 패턴을 찾는 방법이다[8]. PrefixSpan은 단계별로 분할된 Prefix-Projected 데이터베이스들을 구성하여 후보 패턴들의 지지도 계산을 위한 탐색 공간을 줄인다. 즉 시퀀스 데이터베이스에 대한 prefix-projection을 반복적으로 수행하는 것이다. PrefixSpan은 '모든 빈발 시퀀스들은 빈발한 prefix들의 확장에 의해 발견할 수 있다'는 사실에 기인하여 빈발한 prefix에 대해서만 데이터베이스 projection을 수행한다.

3.1.1 Prefix, Postfix and Projection Database

시퀀스 $\alpha = \langle e_1, e_2, \cdots, e_n \rangle$으로 주어졌을 때, 다음과 같은 조건을 만족하는 $\beta = \langle e'_1, e'_2, \cdots, e'_m \rangle$ $(m \leq n)$를 α의 prefix라고 하고, 아래와 같은 경우에만 해당된다.

(1) $e'_i = e_i$ for $(i \leq m-1)$ (2) $e'_m \subseteq e_m$ (3) $(e_m - e'_m)$안에 있는 모든 아이템들은 알파벳 상으로 e_m'안에 있는 아이템들의 뒤에 있다.

β는 α의 부분시퀀스이다 즉 $\beta \sqsubset \alpha$. α의 부분시퀀스 α' ($\alpha' \sqsubset \alpha$)은 prefix β에 대한 α의 projection이라 부르고 아래와 같은 경우에만 해당된다.
(1) α'은 prefix β를 갖는다. (2) α'의 적합한 super-시퀀스 α''이 존재하지 않는다. 즉 $\alpha' \sqsubset \alpha''$, $\alpha' \neq \alpha''$.

시퀀스 $\gamma = \langle e''_m e_{m+1}, \cdots, e_n \rangle$를 prefix β에 대한 α의 postfix라고 부르고 $e''_m = (e_m - e'_m)$인 경우에만 해당되며 $\alpha = \beta \cdot \gamma$와 같이 표시할 수 있다.

3.1.2 Pattern Search 방법

그림 1과 같이 시퀀스 데이터베이스 S가 있다. 최소 지지도 min_sup =2를 가지고 PrefixSpan 패턴 탐색 방법으로 아래와 같은 단계를 거쳐 패턴을 찾아낼 수 있다.

단계 1 : *1-length* 순차패턴 검색. 시퀀스 데이터베이스 S를 스캔하여 *min_sup*을 만족하는 시퀀스 내에 있는 모든 빈발한 *1-length* 길이의 아이템들을 찾아낸다.

단계 2 : 탐색공간의 분할. 찾아낸 *1-length*의 아이템들을 prefix로 선택하여 postfix 및 projected-Database로 분할한다.

단계 3 : 순차 패턴의 부분집합 탐색. 순환적으로 prefix를 확장하여 project-Database와의 조합으로 순차패턴의 부분집합을 탐색하게 된다. *n-length*의 순차패턴이 탐색되면 *n-length*까지의 패턴이 prefix로 확장되며 *(n+1)-legnth*의 순차 패턴을 반복적으로 탐사하게 되며 최종 찾아진 순차패턴의 부분집합들은 그림 1과 같다.

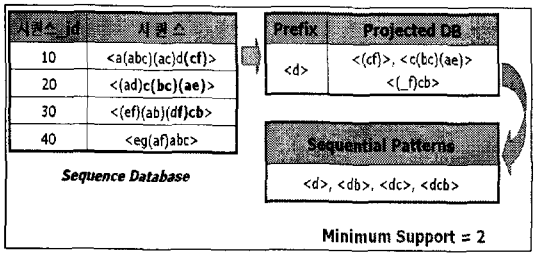

그림 1 시퀀스 데이터베이스와 순차패턴

3.2. 알고리즘의 전처리 및 확장

표 1과 같은 알고리즘은 경보데이터와 최소 지지도가 주어져 있을 때 1부터 *l-length*까지의 순차패턴들의 집합을 구하는 알고리즘이다. 이 알고리즘에서는 경보데이터를 시퀀스 데이터베이스로 변환하는 전처리 단계와 전처리 된 시퀀스 데이터베이스의 시퀀스들을 입력값으로 가능한 모든 길이의 순차패턴들의 부분집합을 구하는 단계로 작업한다. 먼저 전처리 단계에서는 전체 데이터베이스에서 기준이 되는 속성과 대상이 되는 속성들을 사용자 임의로

Input : 오리지널 데이터베이스 O,
최소 지지도 min_sup
Output : 순차 패턴의 최종 집합들
Method : prefixSpan 함수 호출
Subroutine PrefixSpan $(\alpha, l, S|_\alpha)$
Parameters:
α: *SequencialPattern* l: *length of* α
$S|_\alpha$: α − *projected DB*
Method:
1. Original Database O를 스캔
2. 기준 속성과 선택 속성으로 시퀀스 데이터베이스 S 생성
3. $S|_\alpha$를 1회 스캔, 아이템들의 집합 b을 찾음
4. 각각의 빈발 아이템 b를 α에 추가,
Output : α' 생성
5. 각각의 α'에서
α' − *projected DB* $S|_{\alpha'}$생성,
PrefixSpan $(\alpha', l+1, S|_{\alpha'})$

표 1 PrefixSpan Algorithm

선택하게 된다 선택된 속성들중 기준 속성은 시퀀스 ID로써 사용되고 선택속성은 구분자를 이용해 하나의 아이템으로 변환되어 시퀀스 데이터베이스에 삽입 된다. 이렇게 전처리 과정을 통해 시퀀스 ID와 아이템 속성을 갖는 시퀀스 데이터베이스가 생성되면 최소 지지도(min_sup)를 입력값으로 실제 마이닝 과정을 수행하게 되는데 3.1.2절에서와의 방법으로 패턴을 탐색하게 된다. 첫번째 단계에서 가장 작은 단위의 길이인 아이템 하나로 이루어진 시퀀스들을 찾아내고 이 시퀀스들의 빈발도를 카운트하여 빈발도에 만족하는 1-length의 시퀀스들을 prefix로 설정한다. 전체 시퀀스에서 prefix로 선택된 아이템들을 제외한 나머지 시퀀스들은 postfix로써 projected-database에 저장이 되면서 *1-length*의 순차패턴 탐색을 마치게 된다. 다음 단계로써 각각의 prefix와 postfix를 대입시키는 작업을 거쳐 최소지지도를 만족하는 *2-length*의 순차패턴을 탐색하게 되고 탐색된 *2-length*의 패턴들은 prefix로써 확장되며 postfix에서는 제거된다. 이런 일련의 작업들을 거치게 되면 prefix는 점점 길이가 늘어나게 되고 postfix의 길이는 점점 줄어들게 되며 이 작업을 반복하면서 최종적으로는 l-length까지의 모든 빈발한 시퀀스들의 집합을 찾아내게 되는 것이다. 경보데이터로부터 유용하다고 여겨지는 정보를 찾아내기 위해 데이터 마이닝을 적용하는데 있어 기존의 알고리즘을 이용할 경우 기준 속성의 모호성 및 불필요한 속성들 마저도 마이닝 과정에 포함시켜 비용의 증가를 초래할 수 있다. 따라서 이 논문에서는 기준 속

성 및 선택속성을 이용한 전처리 및 확장된 알고리즘을 제시한다. 기준속성이라는것은 시퀀스를 생성하기 위해 그룹화 할 수 있는 트랜잭션 데이터베이스에서의 TID와 같은 속성이고 선택속성은 아이템을 생성하기 위해 선택되는 속성들이다. 이를 이용하여 경보데이터내에서 필요한 속성들만으로 이루어진 가시적으로 볼 수 없었던 새로운 시퀀스들의 집합을 생성할 수 있으며 패턴 탐색시 불필요한 패턴들을 탐색하여야 하는 비용을 절감할 수 있다.

4. 실험 및 평가

실험에 사용된 데이터는 2002년 11월 IDS에서 수집한 경보 데이터로써 이 중 1,300여개를 실험에 이용했으며 17가지의 데이터 속성중에서 전처리 작업을 위한 기준속성, 선택속성 및 최소 지지도는 다음과 같다.

- 기준속성 : Destination Address
- 선택속성 : AttackType/DestinationPort/Protocol
- 최소 지지도 : 10(%)

그림 2는 실험 결과의 내용이고, 표 2는 탐색된 순차 패턴 결과의 일부이다. 이를 통해 목적지 주소

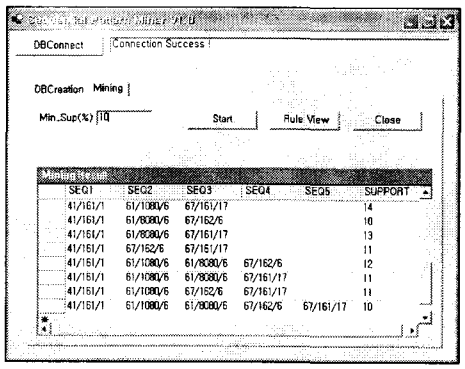

그림 2 순차패턴 마이너 실험 결과

를 기준 속성으로 선택했을 때 공격 유형 변화 패턴, 공격 포트 변화 패턴등의 사실을 지지도를 기반으로 예측 할 수 있다.

[공격유형/목적지포트/프로토콜]
41/161/1 ⇒ 61/1080/6 ⇒ 67/161/17 (14 %)
41/161/1 ⇒ 61/1080/6 ⇒ 61/8080/6 ⇒ 67/162/6 ⇒ 67/161/17 (10 %)

표 2 순차 패턴 탐사 결과

5. 결론 및 향후연구

데이터 마이닝 기법은 대량의 데이터베이스에 유용하게 자료를 추출할 수 있는 방법을 제공한다. 이러한 장점을 이용하여 데이터 마이닝 기법을 침입탐지 시스템에 적용함으로써 경보데이터의 시퀀스 패턴 생성과 효율적인 패턴 분석, 침입탐지 시스템의 자동화 및 성능 향상을 얻을 수 있다. 이 논문에서는 시퀀스 패턴을 탐색하기 위한 데이터 마이닝 알고리즘인 PrefixSpan을 적용한 경보데이터 마이닝 기법을 제안하였다. 경보 데이터라는 데이터 속성을 고려하여 데이터의 속성을 선택하여 이를 이용해 새로운 시퀀스를 생성하는 데이터 전처리 과정에 대하여 제시하였고 이렇게 생성된 시퀀스들의 집합 내에서 순차 패턴 탐사를 하여 가시적으로 볼 수 없었던 시퀀스의 변화 패턴 및 시퀀스의 행동을 예측 가능한 순차 패턴 마이너를 설계, 구현 하였다. 현재 구현된 알고리즘이 전체의 시퀀스를 대상으로 패턴 탐사를 수행하므로 기타 다른 알고리즘에 비해 긴 길이의 시퀀스에 대한 성능이 좋다고 할 수 있으나 짧은 길이의 시퀀스에 대한 검색 시간은 그렇지 못하다. 여기에 시간 연산자, 즉 또하나의 시간 제약사항을 설계 및 추가함으로써 짧은 길이의 시퀀스에서도 높은 성능을 보일 수 있는 확장된 순차패턴 알고리즘의 구현이 현재 수행중에 있다.

참고문헌

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

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

[3] M.S. Shin, "Data mining methods for alert correlation analysis", IJCIS 2003 to be appear

[4] R. Heady, "The Architecture of a Network Level Intrusion Detection System", Technical report, University of New Mexico, Department of computer Science, Aug. 1990.

[5] D. Denning, "An Intrusion Detection Model", IEEE Trans.Softw.Eng.,13(2), Feb. 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, April 2001.

[9] M.J. Lee, "Design and Implementation of Alert Analyzer with Data Mining Engine", IDEAL'03, March 2003.

[10] M.S. Shin, "Applying Data Mining Techniques to Analyze Alert Data", APWeb'03, April 2003