경보데이터 패턴 분석을 위한 순차 패턴 마이너 설계 및 구현☆

Design and Implementation of Sequential Pattern Miner to Analyze Alert Data Pattern

신 문 선* 백 우 진**
Moonsun Shin Woojin Paik

요 약

침입탐지란 컴퓨터와 네트워크 자원에 대한 유해한 침입 행동을 식별하고 대응하는 과정이다. 최근 인터넷의 급속한 발달과 함께 침입의 유형들이 복잡해지고 새로운 침입유형의 발생빈도가 높아져 이에 대한 빠르고 정확한 대응이 필요하다. 따라서 이 논문에서는 침입탐지 시스템의 이러한 문제점을 해결하기 위한 한 방안으로 지능적이고 자동화된 탐지를 지원하기 위한 경보데이터 순차 패턴 마이닝 기법을 제안한다. 제안된 순차 패턴 마이닝 기법은 기존의 마이닝 기법 중 prefixSpan 알고리즘을 경보데이터의 특성에 맞게 확장 설계하였다. 이 확장 설계된 순차패턴 마이너는 보안정책 실행시스템의 경보데이터 분석기의 일부분으로 구성된다. 구현된 순차패턴 마이너는 탐사된 패턴 내에서 적용 가능한 침입패턴들을 찾아내어 효율적으로 침입을 탐지하여 보안정책 실행 시스템에서 이를 기반으로 새로운 보안규칙을 생성하고 침입에 대응할 수 있다. 제안된 경보데이터 순차 패턴 마이너를 이용하여 침입의 시퀀스의 행동을 예측하거나 기술하는 규칙들을 생성하므로 침입을 효율적으로 예측하고 대응할 수 있다.

Abstract

Intrusion detection is a process that identifies the attacks and responds to the malicious intrusion actions for the protection of the computer and the network resources. Due to the fast development of the Internet, the types of intrusions become more complex recently and need immediate and correct responses because the frequent occurrences of a new intrusion type rise rapidly. Therefore, to solve these problems of the intrusion detection systems, we propose a sequential pattern miner for analysis of the alert data in order to support intelligent and automatic detection of the intrusion. Sequential pattern mining is one of the methods to find the patterns among the extracted items that are frequent in the fixed sequences. We apply the prefixSpan algorithm to find out the alert sequences. This method can be used to predict the actions of the sequential patterns and to create the rules of the intrusions. In this paper, we propose an extended prefixSpan algorithm which is designed to consider the specific characteristics of the alert data. The extended sequential pattern miner will be used as a part of alert data analyzer of intrusion detection systems. By using the created rules from the sequential pattern miner, the HA(high-level alert analyzer) of PEP(policy enforcement point), usually called IDS, performs the prediction of the sequence behaviors and changing patterns that were not visibly checked.

☞ Key words : alert data, IDS, sequential pattern mining, prefixSpan, 경보데이터, 침입 탐지 시스템, 순차패턴 마이닝, 프리픽스 스팬

1. 서 론

정보화 사회의 활성화와 정보통신 인프라로서 인터넷의 중요성이 급속히 부각되는 반면에, 인터넷으로 인한 여러 가지 역기능들 또한 심각해지고 있다. 네트워크 서비스의 질적인 측면에서 볼 때

* 정 회 원 : 건국대학교 컴퓨터시스템 전공 강의교수
msshin@kku.ac.kr
** 정 회 원 : 건국대학교 컴퓨터시스템학과 부교수
wjpaik@kku..ac.kr(교신저자)
[2008/02/28 투고 - 2008/03/03 심사 -2008/10/13 심사완료]
☆ 이 연구는 학술진흥재단의 기초과학연구 지원에 의하여 수행되었음.(KRF-2006-20D00849)

네트워크 전반에서의 보안 관리의 필요성은 그 중요성이 크게 증대되고 있다. 이와 관련하여 실제 인터넷 위협에 대응하기 위한 시스템의 개발이 침입탐지 시스템(IDS: Intrusion Detection System)을 중심으로 활발히 이루어지고 있다.

침입을 오용탐지와 비정상 사용 탐지로 세분화하고 있으며 이러한 침입을 해석하고 분석할 수 있는 데이터베이스를 구축하고 있다[5]. IDS 연구[1,4,16,17]는 1996년을 기점으로 DARPA를 중심으로 IDS관련된 표준화의 움직임이 시작 되었으며 점차적으로 비정상 사용탐지 방향으로 연구가 진행 중에 있다.

기존의 침입탐지 시스템 관련 연구들[2,4]을 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 감사 데이터 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 빈발에피소드 탐사기법보다 패턴 탐사의 성능이 우수하다.

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 순차패턴 마이너를 설계하고 구현한다. 이 논문에서 구현되는 순차 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.

논문의 구성은 2장에서 관련 연구로서 침입탐지 시스템의 취약점 및 데이터 마이닝기법 중 순차패턴 마이닝에 대하여 기술하고, 3장에서는 순차패턴 마이닝 중 prefixSpan을 확장한 경보데이터 순차 패턴 마이너의 설계를 기술한다. 4장에서는 구현 및 실험환경 그리고 경보데이터를 이용한 실험결과에 대해 기술하며 마지막으로 5장에서는 결론으로 끝을 맺는다.

2. 관련연구

네트워크 기반 침입탐지 시스템은 네트워크 상에 지나가는 패킷을 분석함으로서 실시간 침입을 탐지하는 시스템이다[1]. 침입탐지 시스템에서는 패킷을 가지고 미리 정해놓은 규칙들과 비교를 해서 공격을 탐지하게 된다. 또한 침입탐지 시스템에서는 이런 침입에 대해 경보 데이터를 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.

콜럼비아 대학의 Wenke Lee[11]는 텔넷 로그데이터, 네트워크 쉘 커멘드 및 Tcpdump와 같은 감사 데이터에 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 순차 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.

데이터 마이닝 기법 중 빈발 에피소드탐사나 연관규칙 탐사, 순차패턴 탐사 등은 모두 특정항목 다음에 어떠한 항목이 뒤이어 발생하는가에 대한 패턴 탐사 기법들이다. 이러한 빈발 항목들에 대한 탐색은 지지도(support)를 기반으로 탐색되어진다. 지지도는 빈발패턴의 빈발항목이 될 수 있는 기준을 제시하는 threshold 값으로 지지도에 따라 빈발패턴의 결과는 달라지게 된다. 따라서 데이터 마이닝의 결과 만들어지는 규칙들은 지지도를 기반으로 하는 신뢰성 규칙들이라 할 수 있다. 즉 생성된 탐사 규칙들은 지지도에 따라서 새로운 지식이 될 수도 있고, 지지도가 높으면 유용한 정보로 활용될 수 없게 되므로 지지도를 결정하는 것은 의사결정에 있어 중요한 요소가 된다. 또한 트랜잭션 데이터베이스를 대상으로 하여 장바구니 분석을 위한 빈발항목 탐사및 연관항목 탐사를 위한 목적으로 전통적인 데이터마이닝기법이 적용되었으며 트랜잭션 데이터베이스는 트랜잭션 아이디(TID)와 항목들에 대한 아이디(OID)로만 구성된 거래데이터베이스이다. 따라서 기존의 데이터마이닝 알고리즘은 경보데이터가 저장된 데이터베이스를 대상으로 적용하기에는 부적합하다. 따라서 기존의 알고리즘을 적용하기 위해서는 경보데이터 전처리 과정과 알고리즘의 확장 설계가 필요하다.

순차패턴 마이닝은 일련의 시퀀스로부터 빈번하게 발생하는 시퀀스들을 찾는 기법[7, 8]이다. 또한 순차패턴 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 순차패턴 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.

침입탐지시스템에서는 침입으로 판명된 트래픽에 대하여 침입과 관련된 정보를 데이터베이스에 저장하게 되고 이를 경보데이터라 한다[14]. 이 경보데이터를 이용하여 관리자는 경보데이터의 연관성 분석, 공격정보 생성, 통계처리 등을 하게 되는데 표1과 같은 객체 속성들을 가지고 있다[17].

이 논문에서는 위와 같은 객체속성들 중 필요한 속성들만을 선택하여 경보데이터 분석을 수행할 것이고 이는 불필요한 다량의 패턴 생성을 감소시키는 역할을 할 것이다. 필요한 속성 선택 후 순차패턴 알고리즘인 prefixSpan을 이용하여 경보데이터의 순차 패턴을 탐사한다. 특히, 경보 데이터의 속성을 고려한 전처리 과정을 거쳐 시퀀스 데이터베이스를 생성하고 빈발 시나리오를 탐사한다.

논문의 효율적인 전개를 위하여 먼저 Prefix, Postfix, Projection Database의 개념을 정리한다. 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 $\beta \sqsubseteq \alpha$. α의 부분시퀀스 α′ $(\alpha' \sqsubseteq \alpha)$은 prefix β에 대한 α의 projection이라 부르고 아래와 같은 경우에만 해당

(표 1) 경보 데이터의 객체 속성

ALERT 개체 필드	내 용
SGSID	경보데이터를 생성하는 SGSs의 ID
ATTACKID	모델링 된 SIGNATURE ID
ATTACKTYPE	공격의 카테고리
DETECTDATE	침입탐지 날짜
DETECTTIME	침입탐지 시간
SRCADDR	근원지 주소
SRCPORT	근원지 포트번호
PROTOCOL	프로토콜
TARGETADDR	목적지 주소
TARGETPORT	목적지 포트
SERVNAME	프로토콜이 TCP일 경우 Well-known Name
ATTACKMSG	공격에 대한 정보제공 메시지
URL	공격에 대한 참조 URL
IMPACT	공격의 심각성
COMPLETION	공격의 성공 여부
TYPE	공격의 종류
ACTCATEGORY	공격에 대한 대응 종류
CONFIDENCE	공격 탐지에 대한 신뢰도
MANUFACTURER	탐지엔진 제조회사
MODEL	탐지엔진의 모델명
VERSION	탐지엔진의 버전
OSTYPE	탐지엔진에 설치된 운영체제 종류
OSVERSION	탐지엔진에 설치된 운영체제 버전
CATEGORY	탐지엔진에 설치된 네트워크 종류
SGSLOCATION	SGS의 설치 위치
SGSNAME	SGS의 호스트 이름
SGSADDRESS	SGS의 IP 주소
SGSNETMASK	SGS의 NETMASK

된다.

(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.

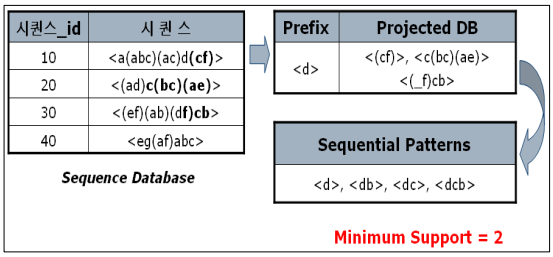
패턴 탐사 방식은 그림 1과 같이 시퀀스 데이터베이스 S를 고려한다. 최소 지지도 min_sup = 2를 가지고 prefixSpan 패턴 탐색방법으로 아래와 같은 단계를 거쳐 패턴을 찾아낼 수 있다.

단계 1 : l-length 순차패턴 검색. 시퀀스 데이터베이스 S를 스캔 하여 min_sup을 만족하는 시퀀스 내에 있는 모든 빈발한 l-length 길이의 아이템들을 찾아낸다.

단계 2 : 탐색공간의 분할. 찾아낸 l-length의 아이템들을 prefix로 선택하여 postfix 및 projected-Database로 분할한다.

단계 3 : 순차 패턴의 부분집합 탐색. 순환적으로 prefix를 확장하여 projected-Database와의 조합으로 순차패턴의 부분집합을 탐색하게 된다.

n-length의 순차패턴이 탐색되면 n-length까지의 패턴이 prefix로 확장되며 (n+1)-legnth의 순차 패턴을 반복적으로 탐사하게 되며 최종 찾아진 순차패턴의 부분집합들은 그림1과 같다.



(그림 1) 시퀀스 데이터베이스와 순차패턴

침입탐지 시스템에서는 이를 이용하여 경보데이터들의 빈발한 시나리오를 탐지하여 정책 및 규칙에 적용시키거나, 일련의 시퀀스로부터 예상되는 다음 공격의 행위를 예측할 수 있다.

경보데이터의 시퀀스를 탐사하기 위한 또 다른 접근 방법인 빈발에피소드 탐사는 순차 패턴 탐사와는 다르게 경보 데이터간 빈발에피소드 탐사를 위한 시퀀스 패턴을 찾기 위해서 WinEPI 알고리즘을 기반으로 하여 주어진 윈도우 크기만큼씩 투플들을 정렬하여 빈발하는 시퀀스 패턴을 탐사한다. 전체 윈도우 테이블 수는 전체 투플의 끝나는 시간에서 시작하는 시간을 뺀 후 주어진 윈도우 시간(즉, 타임단위 x 윈도우크기)을 더하면 전체 윈도우 테이블 수가 계산이 된다. 그리고 투플 간의 상관관계를 고려하도록 참조속성이라는 항목 제약사항을 추가한다.

참조 속성은 어느 한 속성 값이 다른 속성의 값을 참조 할 수 있다는 것이다. 예를 들어, 네트워크 데이터에서 사용자와 행위에 대한 속성이 있다고 하자. 여기서 행위는 사용자에 의해서 얻어지는 속성이다. 그러므로 동일한 사용자에 대해서는 행위 속성에 대한 값을 참조할 수 있다. 빈발에피소드를 탐사한 후 생성된 최종 규칙들을 데이터베이스에 저장한다. 탐사되는 에피소드의 예를 들어 보면,

(Flag=S0, service=http, dst_host=victim), (Flag=S0, service=http, dst_host=victim)

==> (Flag=S0, service=http, dst_host=victim) [0.93,0.03,2]

이 에피소드의 의미는 시간의 93%가 victim 이라는 호스트로 S0 라는 flag를 가지고 http 접속을 2번시도 한 후에, 2초 이내에 같은 패턴으로 접속을 시도하는 이러한 패턴이 전체 3% 발생한다는 것이다.

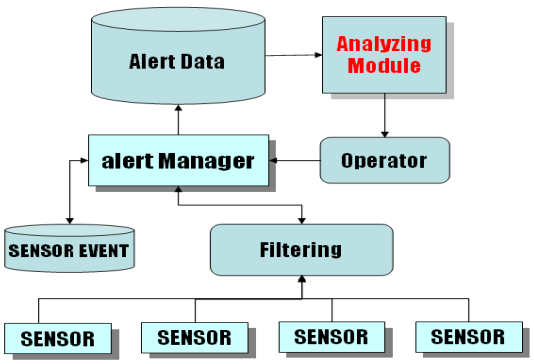
본 논문에서는 먼저 수행된 연구인 빈발에피소드 탐사의 성능 문제를 해결하기 위해서 후보항목을 생성하지 않고 빈발패턴을 탐사하는 순차패턴 탐사기법인 prefixSpan 알고리즘을 적용하였다. 빈발패턴탐사를 위해 지지도를 기반으로 매 단계에서 후보항목을 생성하기 위해 데이터베이스 스캔을 해야 하는 Apriori-Based 방법인 빈발에피소드 탐사는 시간윈도우 내에서 빈발 에피소드를 탐사한다는 점에서는 특정시간내의 공격 패턴을 유추할 수 있다는 장점이 있지만 성능 면에서 크게 문제가 되어 개선이 요구되었다. 따라서 후보 항목 생성을 하지 않는 prefixSpan알고리즘으로 이를 해결할 수 있었다.

3. 경보데이터 패턴 분석을 위한 순차 패턴 탐사 마이너 설계

이 장에서는 경보데이터 분석을 위하여 순차 패턴 마이너를 기술한다. 순차 패턴 마이닝을 위하여 알고리즘의 전처리 및 마이닝 알고리즘을 확장하고 침입탐지 시스템에 설계된 알고리즘을 추가한다.

3.1 경보데이터 분석 프레임워크

그림 2는 순차 패턴 마이닝을 이용한 침입탐지 시스템의 전체 프레임워크이다. 기존의 침입 탐지 시스템에 마이닝 분석 모듈을 추가한 형태로 구성되어 있는데 탐사된 패턴을 관리자가 판별하여 기존의 경보데이터 분석기에 전달하고 이를 이벤트들의 규칙집합에 저장하여 차후에 침입 탐지를 위한 패턴으로써 사용하게 된다.



(그림 2) 경보데이터 분석 프레임워크

각 구성요소들을 살펴보면 가장 하부에 있는 센서들은 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 순차 패턴들을 탐사하고 탐사된 순차 패턴은 오퍼레이터에게 전달되어 분석과정을 거친 후 경보관리기 모듈을 통해 새로운 침입 규칙으로 생성되어 센서 이벤트에 저장된다.

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

3.2. 전처리와 순차 패턴 마이너

그림 3은 경보데이터 순차 패턴 알고리즘을 보여주고 있다. 최소 지지도가 threshold 값으로 주어지고 경보데이터가 저장된 경보데이터 베이스를 탐사하여 1부터 l-length까지의 순차패턴들의 집합을 구하는 알고리즘이다. 이 알고리즘에서는 경보데이터를 시퀀스 데이터베이스로 변환하는 전처리 단계와 전처리된 시퀀스 데이터베이스의 시퀀스들을 입력값으로 가능한 모든 길이의 순차패턴들의 부분집합을 구하는 단계를 수행한다.

먼저 전처리 단계에서는 전체 데이터베이스에서 기준이 되는 속성과 대상이 되는 속성들이 선택된다. 선택된 속성들 중 기준 속성은 시퀀스 ID로써 사용되고 선택속성은 구분자를 이용해 하나의 아이템으로 변환되어 시퀀스 데이터베이스에 삽입된다.

이렇게 전처리 과정을 통해 시퀀스 ID와 아이템 속성을 갖는 시퀀스 데이터베이스가 생성되면 최소 지지도(min_sup)를 입력 값으로 실제 마이닝 과정을 수행하게 된다.

Algorithm prefixSpan
Input : 경보데이터 데이터베이스 O
최소 지지도 min_sup
Output : 순차 패턴의 최종 규칙들
Method : Call prefixSpan(<> , 0, $S|_{\alpha}$).
subroutine **pefixSpan** (α , 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
   $\alpha' - projected\ DB\ \ S|_{\alpha'}$생성,
   PrefixSpan (α′, $l+1$, $S|_{\alpha'}$)

(그림 3) PrefixSpan Algorithm

첫 번째 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-length의 순차패턴 탐색을 마치게 된다.

두 번째 단계에서 각각의 prefix와 postfix를 대입시키는 작업을 거쳐 2-length의 순차패턴을 탐색하게 되고 탐색된 2-length의 패턴들은 prefix로써 확장되며 postfix에서는 제거된다. 이런 일련의 작업들을 순환적으로 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 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 순차패턴 탐사를 수행하였다.

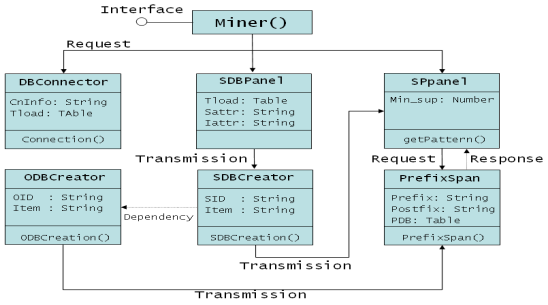
특히 무의미한 순차 패턴의 필터링을 위하여 기준속성이라는 개념을 도입하였다. 기준속성 이라는 것은 시퀀스를 생성하기 위해 그룹화 할 수 있는 트랜잭션 데이터베이스에서의 TID와 같은 속성이고 선택속성은 아이템을 생성하기 위해 선택되는 속성들이다. 이를 이용하여 경보데이터 내에서 필요한 속성들만으로 이루어진 가시적으로 볼 수 없었던 새로운 시퀀스들의 집합을 생성할 수 있으며 패턴 탐색 시 불필요한 패턴들을 탐색하여야 하는 비용을 절감할 수 있다.

또한 상대적으로 많은 시간이 소요되는 빈발 에피소드 탐사에서의 효율성을 개선할 수 있는 순차패턴 탐사기법이다.

3.3 클래스 다이어그램

경보데이터 패턴 분석 마이너의 클래스 다이어그램은 그림 4와 같다. 먼저 그림 4에서 보여주는 DBConnector 클래스는 접속정보를 입력받아 Connection()함수를 실행하게 되고 실행 후에는 기본적으로 전체 테이블 리스트를 로딩하게 된다. 두 번째로 SDBpanel을 호출하게 되는데 이는 기준속성과 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 순차적인 번호인 OID(일련번호)와 아이템(ITEM)들로 구성되어있는 테이블로서 prefixSpan 알고리즘 내에서 항목들 간의 비교를 위해 사용된다.



(그림 4) 경보데이터 순차 패턴 마이너의 Class Diagram

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

저장된 최종 규칙들은 그림 2의 프레임워크 중 경보분석기(Alert Manager)에서 새로운 경보 데이터 패턴을 참조하여 악의적인 공격을 유추하는데 활용된다.

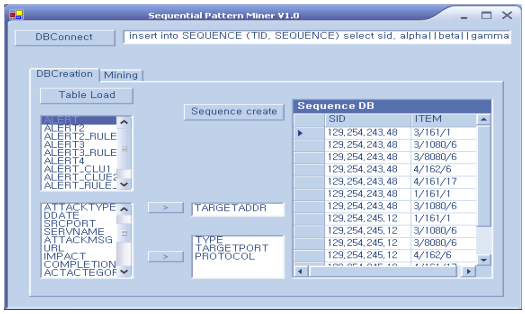
4. 구현 및 실험 평가

이 장에서는 3장에서 설계된 경보데이터 순차패턴 마이너의 구현에 대하여 기술한다. 경보데이터 패턴 마이너의 구현 환경과 구현결과 그리고 기존의 빈발에피소드 알고리즘과의 비교 실험 및 평가를 기술한다.

4.1 구현 환경 및 구현 결과

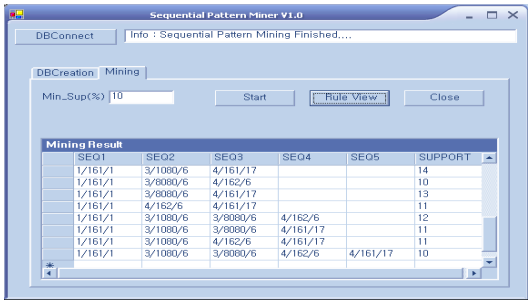
우선 경보데이터를 저장하는 데이터베이스로는 Oracle 8i를 사용하였고, 순차 패턴 탐사 마이닝 연산인 Miner() 모듈은 C#으로 구현하였다. 또한 경보데이터 순차 패턴 탐사 모듈인 SPMiner의 User Interface는 C#에서 제공되는 컨트롤을 활용하여 구현하였다.

SPMiner는 그림 5와 같은 사용자 화면을 가지며 테이블 로딩, 마이닝 대상 속성 선택, 시퀀스 데이터베이스 생성 등의 주요 전처리 기능을 위한 컴포넌트들을 포함한다.



(그림 5) 접속 및 시퀀스 데이터베이스 생성 패널

SPMiner에서의 순차 패턴 탐사는 threshold값인 최소 지지도를 입력받아 전처리 단계에서 생성된 시퀀스 데이터베이스를 대상으로 prefixSpan() 함수를 호출하여 순차패턴 탐사를 수행하게 된다. 그림 6은 지지도 입력 및 실제 탐사과정 수행 후의 결과 화면을 나타내고 있다.

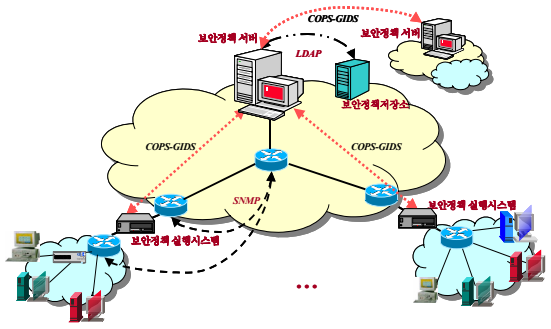


(그림 6) 결과 화면

4.2 실험 배경

구현된 경보데이터 순차 패턴 마이너에 대한 실험은 정책기반 네트워크 보안관리 프레임워크 구조를 기반으로 하는 테스트 베드환경에서 생성되는 경보데이터를 대상으로 이루어 졌다.

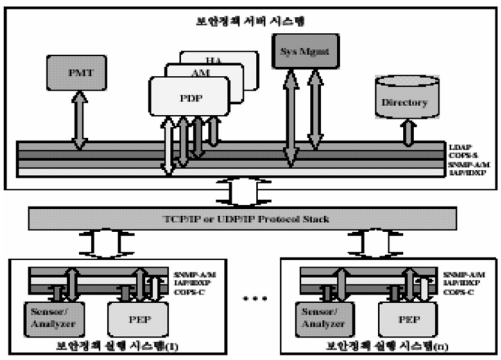
정책기반 네트워크 보안구조(Policy-Based Network Management for Network Security: NS-PBNM)는 네트워크 보안을 위한 정책기반의 네트워크 관리 기법으로서 정책기반 네트워크 보안구조를 지칭한다. 개념적 구성도는 그림 7과 같다.



(그림 7) 보안정책기반 네트워크 관리의 개념구성도

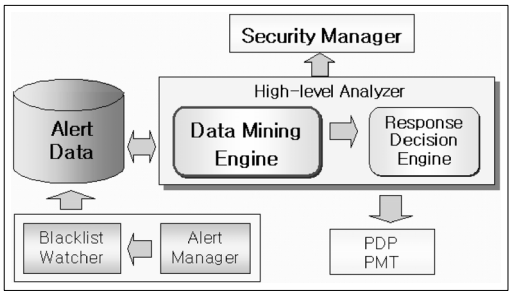
정책기반 네트워크 보안관리의 프레임워크의 구성요소는 보안 정책을 생성하고 관리하는 보안정책 서버의 구성요소인 PMT(Policy Management Tool), 보안 규칙에 따라 보안 행위를 결정하는 PDP(Policy Decision Point), 보안 규칙을 저장하는 보안정책 저장소 PR(Policy Repository)과 보안 행위를 수행하는 보안정책 실행시스템인 PEP(Policy Enforcement Point)와 PDP와 PEP간의 보안 정책 전달을 위한 통신 프로토콜[16,17]로 구성된다. 네트워크 보안 정책을 위한 프레임은 정책기반 네트워크 보안 구조의 계층적인 구성을 가지며 적어도 두개의 계층으로 구성한다. 하나는 관리 계층에 해당하는 보안 정책 서버 시스템이며 다른 하나는 실행 계층에 해당하는 접속점에서의 해킹 트래픽 감지 및 대응을 위한 침입탐지 기반의 보안정책 실행 시스템이다. 보안 정책 서버 시스템은 크게 PMT 블록과 PDP 블록, 보안정책 실행 시스템으로

부터 전달된 경보를 처리하는 AM(Alert Manager)과 HA(High-level Analyzer)블록과 PR을 위한 디렉토리로 구성된다. 보안 정책 실행 시스템은 네트워크 접속점에서 입력 패킷에 대한 탐지와 분석을 제공하는 Sensor/Analyzer 블록과 보안정책 실행기능을 제공하는 PEP 블록으로 구성된다. 그림 8은 정책기반 네트워크 보안관리의 구성요소와 상호간의 관계를 나타나고 있다.



(그림 8) 정책기반 네트워크 보안관리 프레임워크

특히 구현된 순차패턴 마이너는 HA(High-level Analyzer)블록의 데이터 마이닝 엔진의 한 구성요소로서 새로운 보안정책이나 네트워크 상태 진단을 위한 정책 결정에 그 결과를 이용하게 된다. 그림 9는 경보분석기와 다른 구성 요소들간의 관계를 나타내고 있다. 경보데이터 순차패턴 마이너의 실험을 위해서 위 정책기반 네트워크 보안관리 프레임워크의 테스트베드에서 생성된 경보데이터를 이용하였다.



(그림 9) 경보데이터 분석기 구성도와 관계

4.3 실험 및 평가

구현된 경보데이터 순차 패턴 마이너에 대한 실험은 세 가지로 수행 되었다. 첫 번째는 찾아낸 순차 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.

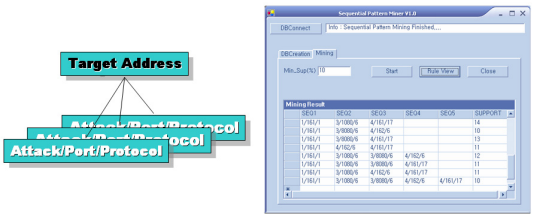
(표 2) 평가 항목

구 분	평가 항목
패턴분석	□ 경보데이터 패턴의 적용 가능성
순차 패턴 탐사	□ 지지도 변화에 따른 패턴 탐사 효율성
	□ 시퀀스 길이 증가에 따른 탐사 효율성

실험데이터는 한국 전자통신연구원에서 개발한 네트워크 기반 침입탐지 시스템(SGS: Security Gateway System)에서 수집한 데이터 중 13,000여개의 테스트 데이터를 임의로 추출하였고 패턴 탐사를 위한 선택 속성들은 아래와 같다.

- 기준속성 : TARGETADDR(목적지 주소)
- 선택속성 : TYPE(공격유형)/TARGETPORT(목적지 포트)/PROTOCOL(프로토콜)
- 최소 지지도 : 10

실험 데이터를 기반으로 순차패턴탐사 마이너의 실행 결과는 다음과 같다. 그림 10은 위와 같은 속성들로 데이터 전처리를 했을 때의 형태를 가시적으로 나타낸 것이고 순차패턴 탐사를 통해 찾아낸 실험 결과이며 표 3은 탐색된 순차 패턴 결과의 일부이다.



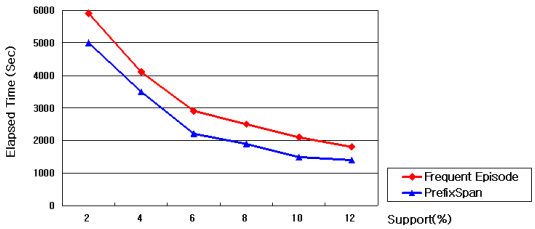
(그림 10) 동일 목적지 주소에 대한 순차패턴 탐사 결과

(표 3) 순차 패턴 탐사 결과

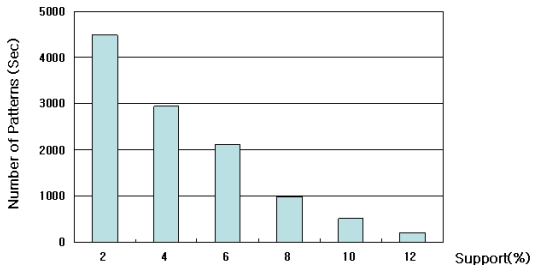
[공격유형/목적지포트/프로토콜]	
1) 1/161/1 ⇒ 3/1080/6 ⇒ 4/161/17	(14%)
2) **1/161/1 ⇒ 3/1080/6** ⇒ 3/8080/6 ⇒ 4/162/6 ⇒ **4/161/17**	(10%)

표3에서와 같이 찾아낸 패턴들을 분석해 보자면, 1)의 경우 1번, 3번, 4번 공격의 형태가 전체 데이터 집합 중 14%를 차지하고 있다. 이를 보아서는 탐사된 패턴의 유용성에 대해서는 언급할 수 없지만 2)의 경우를 보면 1번, 3번, 3번, 4번, 4번의 순서로 공격이 진행되고 있고 1)과 비교하여 볼 때 공격 시나리오의 유사성을 발견할 수 있다. 네트워크 상에서의 공격이라는 것은 단일 공격 보다는 시나리오에 의한 다중 공격들이 충분히 더 의미가 있다. 또한 공격자들은 자신의 공격 시나리오를 감추기 위하여 시나리오 중간에 무의미한 행동을 집어넣는다. 만약 1)과 같은 순차 패턴이 공격 시나리오라고 가정한다면 공격자는 자신의 공격을 침입탐지 시스템이 잡아낼 수 없도록 중간에 무의미한 행동들을 추가한 2)와 같은 형태의 시나리오로써 공격을 시도하게 되는 것이다.

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 11과 같이 지지도를 2%씩 증가 시킴에 따라 빈발 이동 패턴 탐사에 소요되는 시간이 감소함을 알 수 있었고, 그림 12와 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 11) 지지도 증가에 따른 수행시간

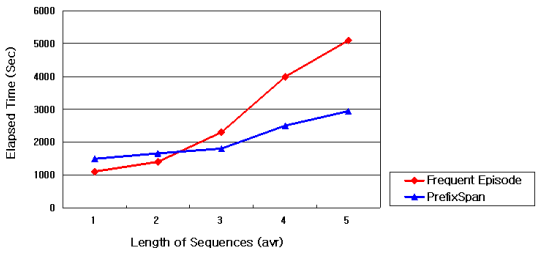


(그림 12) 지지도 증가에 따른 패턴의 개수

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 13과 같이 시퀀스의 길이 증가에 따라 빈발에피소드 기법과의 수행시간의 차이를 확인할 수 있다. 작은 길이의 시퀀스에서는 빈발에피소드의 수행 시간이 prefixSpan 알고리즘에 비

해 빠른 수행 성능을 보이고 있으나 점점 시퀀스의 길이가 길어짐에 따라 prefixSpan 알고리즘 쪽의 효율이 좋다는 것을 확인 할 수 있다.



(그림 13) 시퀀스 증가에 따른 수행시간

이는 빈발에피소드 알고리즘이 단계별 후보 항목의 생성을 통해 지지도를 비교한 후 불필요한 부분을 버리는 방식이고, prefixSpan은 최초 최소 지지도를 만족하는 항목들을 모두 만들어 놓은 후 후보항목의 생성과정 없이 다음 단계의 시퀀스를 탐사하는 방식이기 때문에 작은 길이의 시퀀스에서는 많은 항목들의 생성으로 수행시간이 타 알고리즘에 비해 떨어지지만 길이가 긴 시퀀스를 탐사할수록 후보항목 생성에 필요한 수행시간에 대한 이득을 볼 수 있는 것이다.

위와 같은 실험 및 평가에 대해 요약하자면 첫째, 지지도의 변화에 따라 수행시간에 많은 영향을 미친다는 것을 알 수 있다. 지지도가 증가할수록 탐사되는 패턴의 수도 또한 늘어나고 패턴 생성에 걸리는 시간이 추가되기 때문에 당연히 수행시간도 늘어나게 되는 것이다. 둘째, 시퀀스의 길이 변화에 대하여 시퀀스의 길이 증가는 순차패턴 탐사 수행시간과 밀접한 관계가 있지만 prefixSpan 알고리즘은 긴 길이의 패턴 탐사에 있어 나은 성능을 보인다는 것이다. 그렇지만 짧은 길이의 시퀀스를 탐사하는데 있어서는 수행 시간이 기존의 기법에 비해 오래 걸린다는 문제점이 있다. 따라서 결론적으로 이 문제점을 해결하고 탐사 비용을 줄이기 위해서는 시간 제약 사항의 추가 및 적절한 임계치 설정이 요구된다.

또한 구현된 프로그램의 데이터 저장 방식이 데이터베이스와 하드웨어에 의존적인 형태이기 때문에 수행 성능의 효율을 높이기 위해서는 인덱싱과 같은 트리를 이용한 자료 저장 구조의 연구가 필요하다.

5. 결 론

경보데이터의 순차 패턴 생성과 효율적인 패턴 분석을 통해 침입탐지 시스템의 자동화 및 성능 향상을 위한 방안으로 이 논문에서는 경보데이터 순차 패턴 마이너를 설계하고 구현하였다. 기존의 빈발 에피소드 탐사에서의 시간 성능을 개선하기 위해서 후보항목 생성을 하지 않는 순차 패턴 탐사를 위하여 prefixSpan을 확장 적용하였다. 또한 경보 데이터의 특성을 고려하여, 데이터의 속성을 선택하고 이를 이용해 새로운 시퀀스를 생성하는 데이터 전처리 과정을 수행하였고 생성된 시퀀스들의 집합 내에서 순차 패턴 탐사를 하여 가시적으로 볼 수 없었던 시퀀스의 변화 패턴 및 시퀀스의 행동 예측이 가능한 순차 패턴 탐사 마이너를 설계하였다. 구현된 경보데이터 패턴 마이너의 성능실험을 위하여 네트워크 기반 침입탐지시스템인 SGS(Security Gateway System)의 테스트베드에서 수집한 경보 데이터를 이용하여 성능을 평가 분석하였다. 제안된 순차패턴 마이너가 전체의 시퀀스를 대상으로 패턴 탐사를 수행하므로 빈발 에피소드 탐사와 비교하여 볼 때 긴 길이의 시퀀스에 대한 성능이 우수함을 확인 하였다. 짧은 길이의 시퀀스에 대한 검색 시간은 다른 알고리즘에 비하여 성능이 낮음도 확인 하였다. 그러나 이 성능 저하 원인은 전체 시퀀스에 대한 Projected DB의 과도한 생성으로 인한 부하 이므로 시간 윈도우, 즉 또 하나의 시간 제약사항을 설계 및 추가함으로써 해결할 수 있으므로 이를 추가하는 연구를 계속 수행할 것이다.

참 고 문 헌

[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)

[13] M.S. Shin, K.H. Ryu : Applying Data Mining Techniques to Analyze Alert Data. APWeb'03 (2003)

[14] S. K. Park, J. O. Kim, J. S. Jang : Alert Data Processing for heterogeneous Intrusion Detection Systems in Secure Network Framework. KIPS'03 (2003)

[15] C. J. Shim, C. H. Lee : Research on False Positive Alert reduction using pattern matching technique. KIPS'03 (2003)

[16] D. Curry, H. Debar : Intrusion Detection message exchange format data model and extensible markup language(xml) Document type definition. Internet Draft, draft-ietf-idwg-idmef-xml-03.txt (2001)

[17] 박상길, 김진오, 장종수, "보안네트워크 프레임워크에서 이기종의 침입 탐지 시스템 연동을 위한 경보데이터 처리", 제19회 한국정보처리학회 춘계학술발표대회논문집, 제10권 제1호, pp.2169-2172.

[18] Ho Sung Moon, Eun Hee Kim, Moon Sun Shin, Keun Ho Ryu, Jinoh Kim, "Implementation of Security Policy Server's Alert Analyzer," ICIS2002.

◐ 저 자 소 개 ◑

신 문 선

1988년 충북대학교 전산통계학과 졸업(학사)
1997년 충북대학교 대학원 전자계산학과 졸업(석사)
2004년 충북대학교 대학원 전자계산학과 졸업(박사)
2005～2008 건국대학교 컴퓨터시스템 전공 강의교수
관심분야 : 데이터베이스, 데이터마이닝, 보안 etc.
E-mail : msshin@kku.ac.kr

백 우 진

1988년 연세대학교 토목공학과 졸업(학사)
2000년 Syracuse University School of Information Studies 졸업(박사)
2004～현재 건국대학교 컴퓨터시스템학과 부교수
관심분야 : 인간컴퓨터상호작용, 자연어처리, 정보검색/추출
E-mail : wjpaik@kku..ac.kr