제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

연관규칙을 기반으로한 Web Page 침입탐지 시스템 구현

전홍태*, 윤성대**
*부경대학교 전산정보학과
**부경대학교 전자계산학과
e-mail:htjeon@pufs.ac.kr

Implementation of Web Page IDS(Intrusion Detection System) Based on Association Rule

Hong-Tae Jeon*, Sung-Dae Youn**
*Dept of Computer Science, Pukyong University
**Dept of Computer Science, Pukyong University

요 약

최근에 들어서 Web Pgae 및 서버에 악의적인 사용자들로 하여금 많은 피해가 발생하고 있다. 본 논문에서는 연관규칙을 이용한 침입탐지 시스템을 구현함으로써 해킹 및 부정사용자를 방지하여 시스템의 가용성, 효율성을 높이고 안정적인 운용을 제공한다. 그리고 연관규칙의 신뢰성을 높이기 위하여 가중치 개념을 사용하여 효율적인 침입탐지 시스템 구현을 제시하였다.

1. 서론

컴퓨터와 통신기술의 발달로 사용자에게 다양한 정보와 편리성이 제공된 반면, 컴퓨터 침입 및 범죄로 인한 피해가 증가하고 있다.

침입탐지는 네트워크나 행위 데이터 정보를 분석하여 침입여부를 판정하며 침입탐지의 주요 방식은 오용탐지[1,2]와 비정상행위 탐지[3]가 있다. 오용탐지는 공격행위 패턴을 이용하여 탐지하는 것으로, 알려지지 않은 새로운 공격에는 취약하다. 비정상행위 탐지는 네트워크의 데이터로부터 정상적인 행위 모델을 만들어, 이 정상 모델로부터 얼마나 벗어나 있는가를 찾아내어 탐지하는 것으로 새롭고 알려지지 않은 공격에 유용하다.

일반적인 비정상행위 탐지 모델은 순수 데이터의 학습에 의해 구축된다. 이 모델의 성능은 모델링 방법과 가능한 학습 데이터의 질과 양에 크게 좌우된다. 이 방법은 순수 데이터로 비정상행위 모델을 학습하는 것으로 몇 가지 결점이 있다. 첫째, 순수 데이터를 얻기가 쉽지 않다. 둘째, 불필요한 데이터에 의한 학습은 심각한 결과를 초래한다. 즉 학습 데이터에 침입이 숨겨져 있다면 공격을 정상으로 판정하는 모델이 만들어지게 된다. 셋째, 실시간으로는 학습 대상 데이터가 순수 데이터라고 보장하기 어렵다.

침입자들의 시스템 공격은 초기에는 침입기법이 단순하였지만 정보 통신의 발전과 더불어 시스템 침입기법도 고도화되고 전문적으로 변화해가고 있다. 따라서 이에 대응하는 침입탐지 기법들도 복잡성을 더해가고 있으므로 과거와 같이 각 침입방식에 대한 개별적인 대처 방안으로는 충분한 보안 유지를 기대할 수 없다. 이러한 문제를 해결하기 위해서 자동화된 판정 시스템 개발이 필요하게 되었고, 방대한 양의 감사 자료를 필터링 등의 방법으로 자료의 저장 및 분석에 따른 오버헤드를 최소화시킬 수 있는 기술이 필요하게 되었다. 특히 비정상행위 탐지 모델의 핵심이라 할 수 있는 비정상행위 판정 기술과 관련하여 보안 관련 감사 자료의 수집, 저장, 분석 및 해석 기술에 대한 연구가 활발이 추진 중이다.

현재 발생하는 서버의 오류는 바이러스 및 악성 프로그램에 의한 공격이 대부분이다. Web 서버에 많이 시도 되고 있는 상황이지만 대부분의 침입탐지는 시스템은 전반적인 네트워크나 일반적인 서버(특히 데이터베이스 서버)에 집중적인 탐지 및 오용을 시도 하는 상황이다. 따라서 본 논문에서는 Web서버의 특성에 맞는 침입탐지 시스템을 연관규칙 마이닝과 연계를 하여 시스템 설계 및 구현 방향을 제시한다.

2. 관련연구

비정상행위 탐지 방법으로는 대부분 임계값을 두고 그 값을 기준으로 판정하는 통계적 방법을 이용하고 있다. 특히 데이터 마이닝 기법이 침입탐지에 도입되어, 대규모 감사 데이터베이스에서 그 특징을 추출하여 오용 시그니처나 정상행위 프로파일을 생성하여 침입탐지 판정에 이용되고 있다.

불법적인 데이터 조작이 의심되는 사용자를 감시하기 위해서는 상당한 양의 로그 데이터를 분석해야 하는 부담을 가지게 된다. 따라서 자동화된 로그 데이터의 분류 및 분석을 통한 보안 시스템의 필요성이 대두되고 있다. 최근에 침입 탐지를 위해서 로그 데이터를 지능적이고 자동적으로 분석하는 데이터 마이닝 기법이 이용되고 있다. 데이터 마이닝 기법 중에 연관 규칙 마이닝 방법은 데이터 내에 항목들 간의 연관성을 탐지한다. 연관 규칙 마이닝의 대표적인 방법에는 Apriori, Partition 방법이 있다. Apriori는 사용자가 정의한 최소 지지도를 이용하여 동시에 자주 나타나는 항목(frequent-itemset)들을 정제하고 빈발 항목 집합(frequent-itemset)에서 생성된 규칙들은 신뢰도를 이용하여 정제하는 방식이다. 그리고 Partition 알고리즘은 데이터 집합을 최대 두 번 검색함으로써 데이터 집합에 대한 탐색횟수를 감소시킨다. 이 알고리즘의 기본적인 접근 방법은 분석 대상 데이터 집합을 가용 메모리 공간에 적합한 크기의 블록으로 분할한다. DIC 알고리즘은 서로 다른 길이를 갖는 항목 집합들의 출현 빈도수를 동시에 분석하며 잠재적으로 빈발인 항목들의 출현 빈도수만 래티스(Lattice)에서 관리된다.

침입탐지 시스템 방법에는 JAM, DEMIS가 있다. JAM은 연관 규칙 마이닝과 더불어 frequent-itemset를 이용하여 정상행위 패턴을 생성한다. JAM은 데이터 마이닝 응용프로그램을 평가하는 데에 있어 일반적 접근방법인 메타학습(Meta-Learning)을 채용하고 있으며 분산환경에서의 이식성과 확장성을 제공하는 에이전트 기반 데이터 마이닝 시스템이다.

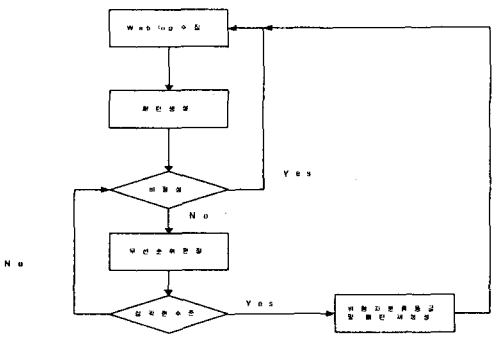
DEMIS 시스템은 데이터베이스 사용자 로그 데이터에 대한 빈발 항목 집합을 정상 행위 패턴으로 추출하여 데이터베이스 보안에 적용하였다. 하지만 DEMIS에서는 SQL 문 단위로 빈발 항목 집합을 생성하기 때문에 의미적인 사용자의 행위 단위인 세션 단위의 모델링이 불가능하다. 따라서 내부 권한 오용자에 대한 효과적인 탐지가 어렵다.

웹 마이닝의 기본적인 개념은 기존의 대용량 데이터베이스 분야에서 지식 추출에 대한 유용성이 검증된 데이터 마이닝 방법을 월드 와이드 웹에 적용하고자 하는 것이다. 즉, 전통적인 데이터 마이닝이 구조화되어 있는 데이터베이스를 연구의 대상으로 하는 반면에 웹 문서 마이닝은 웹을 대상으로 데이터 마이닝 기술을 사용하여 자동으로 월드 와이드 웹으로부터 정보을 발견하고 추출하여 의사결정에 필요한 유용한 정보를 지식 베이스의 형태로 제공하는 것이다.

3. Web log 분석 및 침입탐지 시스템

3.1 침입탐지 시스템 개요

기존의 탐지 시스템은 데이터나 서버 및 네트워크에 집중되어 있었다. 그러나 본 시스템은 Weblog을 이용한 사용자 패턴을 분석하고, 분석한 패턴 데이터를 이용한 비정상행위 탐지 시스템이다. 이시스템은 Weblog을 Web 서버의 로그 데이터를 가지고 패턴생성에 필요한 감사 데이터를 생성한다. 시스템의 개략도는 그림1과 같다.



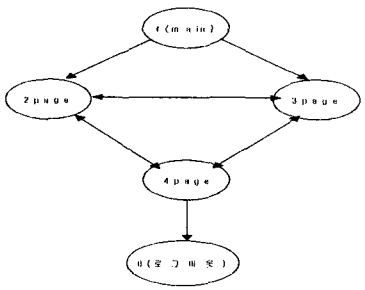
(그림1) 연관규칙을 이용한 침입탐지 개념도

3.2 Weblog 생성 수집

Weblog는 Web 서버에 대한 접근 정보나 접근 페이지 정보로 분류하여 기초 자료로 변환한다. 변환

한 기초자료는 데이터 마이닝 시스템을 이용하여 감사 데이터를 생성한다. 이때 마이닝은 연관규칙 마이닝을 이용하여 접근 정보 및 페이지에 대한 유사도 및 연관성 자료를 생성한다.

일반적인 사용자가의 WebPage 접근경로가 그림 2와 같다고 하면 사용자의 접근경로 정보를 Weblog 자료를 수집한 결과 표2와 같다.



(그림2) Web Page 구성 및 접근경로

<표1> WebPage 접근경로

사용자	접근경로
A(1)	1-2-4-0
A(2)	1-3-4-0
A(3)	1-2-3
A(4)	1-2-4-2-4-0
A(5)	1-3-2-4-0
A(6)	1-2-0

표1의 첫 번째 행의 의미는 A(1)은 사용자A가 1번째 방문한 WebPage 경로를 의미한다. 즉, 1Page(로그인)한후 2Page를 거쳐서 4Page를 접근후 0Page(로그아웃)한 경로를 나타낸 것이다.

3.3 신뢰도, 가중치 및 패턴생성

연관규칙에서 신뢰도(C) 및 지지도(S)를 계산하는 일반적인 수식은 아래와 같다.

X→Y,

$$C = \frac{X\text{를포함한항목중에서 }Y\text{를포함할확률}}{X\text{를포함할확률}}$$

$$S = \frac{X, Y\text{를동시에포함할확률}}{\text{전체항목}}$$

이 수식을 이용하여 표1의 접근경로에 대한 신뢰도를 계산한 것이 표2이다.

표2의 신뢰도 부분의 접근경로 1-2Page 신뢰도는 다음과 같이 산출한다. 표1에서 사용자 A가 6번 Web Server 에 접근하여 1Page 접근한 횟수가 6번이다. 그리고 1Pgae, 2Page 모두를 접근한 경우는 1,3,4,6번 접근때이므로 횟수는 4이다. 그러므로 1pge를 접근하고 2Page를 접근할 신뢰도는 4/6 = 0.66 이다. 이와 같은 방식으로 모든 연관Page의 신뢰도를 계산한 것이 표2의 신뢰도 값이다.

<표2> 일반적인 신뢰도 및 가중치 포함한 신뢰도

접근경로	신뢰도	가중치(신뢰도)
1-2	0.66	0.83
1-3	0.33	0.42
2-3	0.17	0.22
2-4	0.5	0.67
2-0	0.17	0.22
3-2	0.33	0.33
3-4	0.33	0.33
3-0	0	0
4-2	0.2	0.2
4-3	0	0
4-0	0.66	0.8

가중치는 Web 페이지를 하나의 트리로 변형을 시켜서 Main Page를 가장 높은 가중치를 두고 하나의 아래로 내려갈 때마다 가중치를 감소시켰다.

<표3> 가중치 및 접근경로 값

PAGE	가중치(빈도수)	접근경로
1	4	5
2	3	4
3	2	3
4	1	2

표2의 가중치를 포함하는 신뢰도는 1Page에서 2Page로 이동하는 신뢰도는 4*5(가중치)/6*4(가중치) = 0.83 이다. 나머지도 같은 방법으로 계산하였다.

연관규칙에서 최소지지도(minsup(0.4))로 설정하면 위의 사용자의 빈발항목 집합은 가중치를 고려하지 않았을 경우 {1-2,4-0} 이고 가중치를 고려 했을 경우는 빈발항목 집합은 {1-2,1-3,2-4,4-0} 이다.

즉 위의 사용자는 {1,2,3,4,0}의 단위 페이지를 접근 가능하고 가장 많은 접근 경로 패턴은 1-2-4-0의 경로이다.

3.4 비정상행위 탐지 시스템

비정상행위 탐지 시스템은 감사 데이터의 패턴을 이용하여 현재 사용자의 행동이 정상인지 비정상인지를 탐지하는 시스템이다.

본 논문에서 제안하는 시스템은 사용자가 Web 페

이지에 접근할 때부터 탐지시스템이 작동한다. 접근경로 정보가 감사 데이터에 존재하기 때문에 비정상적인 로그인 이나 접근할수 없는 경로를 강제적으로 접근하는 사용자는 바로 탐지시스템에 포착이 가능하고, 사용자 정상행위 정보를 감시시스템이 가지고 있으므로 비정상적인 행동을 하는 사용자도 파악이 가능하다. 그리고 해당 사용자의 정보와 Web Page 접근정보 및 접근정보의 신뢰도도 탐지 시스템이 보유하고 있으므로 전혀 새로운 접근 패턴을 사용자가 생성하면 부정사용자로 판단하여 부정사용자도 판별할수 있다.

탐지시스템의 작동 알고리즘은 그림3과 같다.

```
Procedure IdsDetection(UserId,PageInfo)
  If (login) then
    For Web_Page_access
      Web_page_insert() //access info write
    if(web_page_normal_check()=true)
    then //normal login  Check
       if(web_mining_ids_check() = true)
          then //priority base ids
       else
          break;
       end if
    else
      break;
    end if
  Next
  End if
End Procedure
```

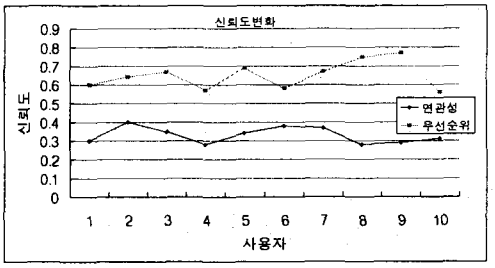
(그림3) 탐지시스템의 작동 알고리즘

4. 실험결과 및 분석

실험을 위해 데이터베이스는 NT 기반의 Oracle 8.05을 사용하였고 Web log 수집을 위하여 IIS6.0 web 서버를 이용하여 log 자료를 수집하였다.

그림4는 연관성과 우선순위에 의한 시스템의 신뢰도 변화이다. 우선순위를 가미한 시스템이 Support(X∪Y)의 값이 크기 때문에 더 좋은 신뢰도를 나타내고 있다. 그리고 WebPage의 접근정보 및 구성정보도 감사데이터로 활용하기 때문에 침입탐지 성능도 우수하다.

사용자의 접근경로 정보를 수집하는 시스템을 채용하기 때문에 패턴 매칭도 일반적인 매칭 방법보다도 우수한 결과를 나타낸다.



(그림4) 신뢰도 변화

5. 결론 및 향후 연구과제

본 논문에서는 Web 서버 중심의 비정상행위 시스템을 제안하여 불법 침입자 뿐만아니라 비정상행위의 사용자도 탐지 가능한 시스템을 제안하였다. 이를 위하여 패턴을 생성하고 Web의 특성에 맞은 접근 방식 및 접근경로 값과 페이지 및 접근성을 가미한 우선순위를 산정하여 정상행위 패턴의 신뢰도를 향상시키는 방법을 제안하였다. 이런 향상된 신뢰도를 이용하여 더 좋은 탐지 성능을 나타내는 비정상행위 탐지 시스템을 생성하였다. 또, 접근정보를 누적하여 관리함으로써 일정시간이 지난후 새로운 패턴으로 변경하여 변화하는 사용자 및 시스템에 능동적으로 대처할 수 있는 시스템을 제안하였다.

향후 연구과제로는 web 서버 뿐만 아니라 일반적인 서버와 시스템 그리고 네트워크에도 우선순위 정보를 이용한 탐지 시스템을 적용하는 것이 효율적이라고 생각하고 이분야의 연구가 더 진행되어야 할 것이다.

참고문헌

[1] W.Lee and S.Stolfo, "Data Mining Approaches for Intrusion Dection" In Proc. of the 7th USENIX Security Symposium, San Antonio, Texax, January, 1998

[2] W.Lee, S.Stolfo and P.K.Chan "Learning Patterns from Unix Process Execution Traces for Intrusin Dection." Proc. AAAI-97 Work. On AI Methods in Fraud and Risk Management 1997

[3] S.Stolfo, A.L. Prodromidis, S.Tselepis,W.Lee, D.Fan, P.K. Chan, "JAM:JavaAgents for Meta-Learning over Distributed Database," Proc. KDD-97 and AAAI97 Work. On AI Methods in Fraud and Risk Management.