유사도를 이용한 침입 탐지 시스템에서 정상행위 패턴의 확장

정영석, 위규범

아주대학교 정보통신전문대학원

Extension of Normal Behavior Patterns for Intrusion Detection System Using Degree of Similarity

Young-seok Jeong, Kyubum Wee

Graduate School of Information and Communication, Ajou University

요 약

광범위한 인터넷의 발달은 우리의 생활을 윤택하게 해주었지만, 불법적인 침입, 자료 유출 등 범죄도 늘었다. 이에 따라 불법적인 침입을 막는 침입탐지기술도 많이 발전하게 되었다. 침입탐지기술은 크게 오용탐지방법과 비정상적인 행위 탐지 방법으로 나눌 수 있다. 본 논문에서는 비정상적인 행위 탐지 방법의 긍정적 결함을 줄이기 위한 방법으로 유사도 측정 알고리즘을 사용한 방법을 제시하고자 한다.

I. 서론

침입탐지방법은 크게 오용탐지방법과 비정상적인 행위 탐지 방법으로 나누어진다. 오용탐지방법은 이전에 침입의 패턴들을 모아놓고, 이 패턴들과 판정을 원하는 데이터를 비교하는 방법이고, 비정상적인 행위 탐지 방법은 사용자의 정상적인 행위에 대한 데이터를 모아놓은 다음, 사용자가 만약 이 정상행위의 데이터에서 벗어나면 침입이라고 판정을 내리는 방법이다.

정상적인 행위에 대한 데이터를 수집한 후, 입력 데이터와 비교를 해서 만약 정상행위 데이터에 없으면 무조건 침입이라고 판정을 하는 것은 긍정적 결함을 발생시킬 원인을 제공하게 된다. 이에 이러한 결함을 보완하고자 유전자 알고리즘을 이용하여 정상적인 데이터와 유사한 데이터를 생성을 하여 긍정적 결함을 줄이고자 하였다[1]. 또 다른 방향에서는 자연면역시스템을 기반으로 하여 정상적인 데이터로부터 Hamming Distance를 이용하여 정상적인 데이터로부터 일정거리가 벗어나면 이것을 침입패턴이라고 가정하고, 이 침입패턴과 정상패턴을 이용하여 침입탐지를 하는 연구도 있었다[2].

본 논문에서는 우리가 통상적으로 인터넷 검색을 할 때 많이 사용하는 검색사이트에서 사용하는 유사도 측정 알고리즘을 이용하여, 정상일지 모르는 데이터 중 침입이라고 판정된 데이터에 대한 유사도를 측정하여 정상데이터인지 침입데이터인지를 한번 더 확인한다. 이를 침입탐지의 긍정적 결함을 줄일 수 있는 방법으로 본 논문에서 제시한다.

II. 관련연구

2.1 유전자 알고리즘

유전자 알고리즘은 개체군중에서 적합도가 높은 개체가 높은 확률로 살아남아 재생할 수 있게 되며, 이때 교배 및 돌연변이로서 다음 세대의 개체군을 형성하는 것이 기본 개념이다.

유전자 알고리즘을 적용한 침입탐지 연구에서는 Sendmail의 프로세스들의 시스템 호출 궤적을 모아, 이것들에 유전자 알고리즘을 적용하여 정상행위 패턴을 확장한 후에 이전의 정상행위 시스템 호출과 함께 사용하여 침입탐지의 긍정적 결함을 줄이기 위해 사용되었다[1].

탐지대상은 Setuid 프로그램에 해당하는 프로

세스들이다. 그 중에서 Sendmail의 프로세스들을 대상으로 하였다. 이 프로세스들이 수행되는 동안 시스템 호출이 일어나게 되고 이러한 시스템 호출의 모음을 시스템 호출 궤적이라고 부른다. 이 시스템 호출 궤적을 sliding windows 기법을 사용하여 길이 8단위로 나누어서 실험에 사용하였다. 이는 6-9사이의 길이로 sliding windows기법을 사용하여 시스템 호출 궤적을 잘라서 하는 것이 가장 성능과 탐지율이 높은걸로 나타나있기 때문이다[3].

2.2 Hamming Distance

면역 시스템을 바탕으로 하는 침입탐지시스템을 만들기 위해서는 인체내에 병원체가 들어 왔을 때, 만약 이전과 전혀 다른 새로운 병원체이면 새로운 감염 병원체에 대응하는 새로운 탐지자를 만들어 대응하는 1차 대응과 이전에 이미 감염된 적이 있는 병원체가 침입하면 바로 해당 탐지자를 만들어내서 대응하는 2차 대응으로 나누어진다.

면역 시스템바탕 침입탐지시스템의 탐지대상은 특권프로세스의 비정상적인 행위이고, 모델 정의에 따라서 시스템 감사 궤적, 프로세스 행위 궤적 등을 정의를 하였다. 여기서 행위패턴은 시스템 호출 궤적을 의미한다. 주어진 행위 패턴의 Hamming Distance는 행위패턴 시스템 호출 궤적의 모든 행위 패턴들 중 주어진 행위패턴과 가장 유사도가 높은 패턴과의 차이를 의미한다.

이 Hamming Distance를 이용하여 정상행위 패턴에서 거리가 6인 시스템 호출 궤적을 불법적인 행동을 하는 시스템 호출 궤적으로 정하고 이것을 모아서 면역시스템의 2차 대응에 이용하였다[2].

III. 제안시스템

3.1 유사도 측정 알고리즘

유사도 측정 알고리즘은 우리가 일반적으로 접하는 인터넷 검색 사이트에서 많이 사용되는 알고리즘이다. 이 알고리즘은 Information Retrieval(IR)에서 나온 개념이다[4].

3.1.1. Information Retrieval 개요

IR의 기본개념은 그림1과 같다. 사용자가 검색어를 요청하였을 경우 색인데이터에서 검색어에 일치하는 데이터를 검색하여 그에 따른 검색결과를 돌려 보내주는 것이다.

IR의 검색기법 모델은 여러 가지가 있다. 검색어와 인덱싱 데이터와의 정확한 매칭, 즉 Yes/No

그림 1. Information Retrieval

만 답하는 Boolean Model, 이 모델의 단점을 보완하여 단어에 대한 중요도를 설정하여 검색 순위를 부여한 Extended Boolean Model, 검색어와 문서의 유사성 분석을 통한 검색하는 Vector Model등이 있다. 본 논문에서 사용한 알고리즘은 Vector Model에 속한다.

3.1.2. Vector Model

가장 많이 사용되는 모델로 문서와 질의어간의 관계를 고차원 공간에 vector로 표현해서 이 가운데 질의어와 가장 관계가 높은 vector를 선택하는 방법이다. 여기서 vector는 각각의 질의어에 대한 가중치를 고차원 공간상에 표현한 것이다. 다음 예제는 2차원에 vector model를 적용한 예이다.

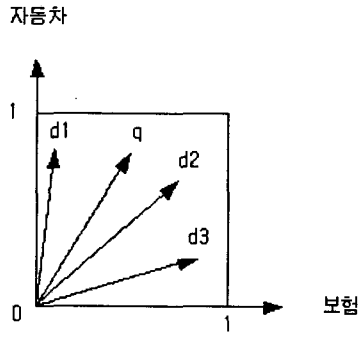

그림 2. Vector Model 예

먼저 문서 d1, d2, d3에 각각의 질의어에 대한 가중치가 설정되어 있어야 된다. 그림3의 질의어 q는 (자동차, 보험)이다. 이 q에 대한 vector가 (0.71, 0.71)로 정해졌을 때, 각각의 d1(0.13, 0.99), d2(0.8, 0.6), d3(0.99, 0.13)이면, 질의어 q와 가장 유사한 것은 d2이므로 d2가 선택된다. 이때 이 유사도를 측정하기 위하여 cosine 계수 알고리즘을 이용한다.

3.1.3 cosine 계수 알고리즘

Vector space model에서 문서와 질의어를 측정하기 위해 사용되는 알고리즘이다(수식 1).

$$\cos(q,d)=\frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} d_i^2}} \quad (1)$$

q : 질의어 d : 문서
i : 매칭 단어의 인덱스
q_i : q 안에서 단어 i의 가중치
d_i : d 안에서 단어 i의 가중치

예를 들어 질의어 q가 프라이드, 보험, 트럭이고 각각 가중치가(0.8, 0.9, 0.7)이라 하자. 문서 d의 대표 단어가 프라이드, 보험, 버스이고 각각의 가중치가(0.6, 0.5, 0.1)이라 하자. 이 질의어와 문서에서 매칭된 단어는 프라이드, 보험이다. 트럭은 문서 d의 대표단어에 없으므로 유사도 계산에서 빠지게 된다. 따라서 cosine 계수 알고리즘에 의해 계산하게 되면 다음 같이 된다(수식 2).

$$\cos(q,d)=$$

$$\frac{0.8*0.6+0.9*0.5}{\sqrt{(0.8)^2+(0.9)^2}*\sqrt{(0.6)^2*(0.5)^2}}=0.42 \quad (2)$$

각 문서의 단어에 가중치를 주는 것은 다음과 같은 방법으로 한다(수식 3).

$tf_{i,j}$: 문서 d_j 안에서 단어 w_i의 발생 빈도수

df_i : 단어 w_i가 있는 문서의 수

N : 전체 문서의 수

$$weight(i,j)=$$

$$\begin{cases} (1+\log(tf_{i,j}))\log(\frac{N}{df_i}) & \text{if } tf_{i,j}\geq 1 \\ 0 & \text{if } tf_{i,j}=0 \end{cases} \quad (3)$$

나온 결과치가 단어 i의 문서 j에서의 가중치이다.

3.2 제안 시스템

침입 탐지의 방법은 오용탐지방법과 비정상적인 행위 탐지 방법이 있다. 본 논문에서는 비정상적인 행위 탐지 방법에 적용을 할 것이다.

3.2.1 대상

본 논문에서는 특권권한을 가지고 동작하는 프로세스들의 시스템 호출 궤적을 대상으로 한다. 즉, setuid 프로세스들이 그 대상이 되고, 그 프로세스들의 시스템 호출을 기본 데이터로 한다. 이 시스템 호출을 모으기 위해 먼저 사용자의 정상행위로부터 정상적인 시스템 호출 궤적을 추출(프로파일)한다[그림 3].

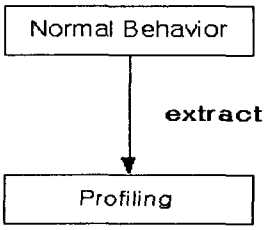

그림 3. 프로파일

프로파일이 끝난 데이터를 정상 데이터로 가정을 한 후 이것을 sliding windows 방법 기법으로 시스템 호출 궤적을 잘라 모아놓은 것을 정상행위 판단을 위한 기본 데이터로 삼는다.

3.2.2 Cosine Measure 적용

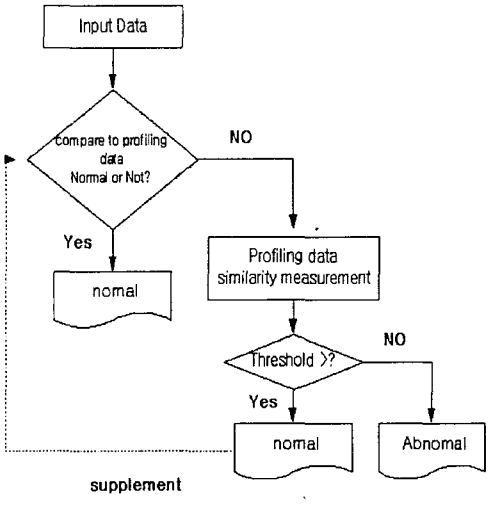

그림 4. 전제 흐름도

모여진 시스템 호출 궤적과 비교 하고자하는 시스템 호출을 비교하여 모여진 정상 시스템 호출 궤적에 속해 있으면 정상 시스템 호출이라 판정을

하고, 없으면 비정상이라고 판정하게 된다[그림4].

만약 정상적인 시스템 호출이 정상 시스템 호출 궤적에 속해있지 않으면 침입이라고 판정되게 된다. 이것을 긍정적 오류라고 한다. 본 논문에서는 이렇게 비정상으로 판정된 시스템 호출을 유사도 측정 알고리즘인 cosine 계수 알고리즘을 적용하여 정상적인 시스템 호출 궤적과 유사도를 분석하게 된다. 유사도를 비교하기 위해 다음과 같은 정의를 한다.

정의(1) 문서 = 정상 시스템 호출 궤적

정의(2) 단어 = 2~3개의 시스템 호출의 묶음

단어를 시스템 호출을 2~3개 묶은 이유는 cosine 계수 알고리즘은 단어의 순서에 대한 의미 없이 단순히 단어가 있으면 매칭되었음을 판정한다. 그러나 시스템 호출 궤적과 비교하는 시스템 호출은 시스템 호출 순서를 고려해야 된다. 이런 차이를 프로그램에서 해결하기 위해 시스템 호출을 2~3개로 묶었다.

consine 계수 알고리즘을 사용하기 위해서 먼저 입력 데이터의 가중치를 산정한다. 가중치 산정 방법은 수식 3를 통하여 한다. 정상 시스템 호출 궤적들과 각각의 유사도를 측정한 후, 전체 유사도 가운데 가장 높은 유사도와 임계치 비교 판정을 한다. 임계치는 여러 차례의 실험을 통해서 조정한다. 입력 데이터가 이 임계치 이상으로 나오는 경우에는 입력 데이터를 정상적인 시스템 호출 궤적에 추가하여 보강한다.

IV. 향후 계획

유사도 측정 알고리즘을 이용하여 정상적인 시스템 호출 궤적을 보완하여 긍정적 결함을 줄이는 방법을 모색하였다. 본 논문에서 제안한 방법의 실험 데이터는 뉴 멕시코 대학에서 모아놓은 sendmail, ftp, inetd등의 정상 행위 패턴을 가지고 실험을 할 것이다[5]. 프로그램은 C++을 사용하여 구현할 예정이며, 공격 패턴을 수집하여 실험함으로서 제안되는 방법의 부정적 결함율에 대해서도 조사하여, 본 논문에서 제시한 침입 탐지방법의 타당성을 보일 것이다.

참고문헌

[1] 김기택, "유전자 알고리즘을 적용한 침입 탐지 방안 연구", 아주대학교 정보통신전문대학원 석사학위 논문, 2001년 8월.

[2] 이종성, 채수환, "특권 프로세스의 시스템 호출 추적을 사용하는 침입탐지시스템 설계 : 면역 시스템 접근", KISA '99 정보보호 우수논문집, 1999.

[3] C. Warrender, S.Forrest, and B. Pearlmutter. "Detecting intrusion using system call: Alternative data models", Proceedings of the 1999 IEEE Symposium on Security and Privacy, May 1999.

[4] C. D. Manning, H. Schütze, "Foundations of statistical natural language processing", MIT press, 2000.

[5]http://www.cs.unm.edu/~immsec/systemcalls.htm.