통계데이터를 이용한 인터넷 침해사고 추이분석 및 특성에 관한 연구

Pattern Analysis of Internet Accidents using Statistical Process

김현우, 심우철, 김세헌
한국과학기술원 산업공학과
Hyunwoo Kim, Woochul Shim, Sehun Kim
Department of Industrial Engineering, KAIST

Abstract

IT 기술의 발전으로 사회 전반적으로 정보시스템에 대한 의존도가 커지고 있는 가운데 이로 인한 역기능 또한 심화되고 있다. 정보보호 관련 침해사고는 매년 급격히 증가하고 있으며 차츰 지능화, 다양화 및 자동화되는 경향을 보임에 따라 대응이 어려운 실정이다. 하지만, 아직은 침해사고의 정확한 파악조차 힘든 상황으로, 침해사고의 조기탐지 및 분석을 통한 피해 방지 대책이 시급히 요구되고 있다. 본 연구에서는 국내의 해킹·바이러스 통계데이터를 분석하여 미래 침해사고의 추이를 예측하고, 침해유형간의 상호연관성을 분석한다. 이를 위해 국내의 해킹바이러스 통계데이터의 시계열 자료를 이용한 패턴분석을 통해 침해유형별 상관관계를 규명하여 보다 정확한 침해대응의 근거를 제시하고자 한다.

1. 서 론

인터넷의 발달을 통한 정보화 사회의 도래로 국가 사회는 전반적으로 정보시스템에 의존하게 되었다. 그러나, 인터넷은 개방적인 환경으로 인해 해킹, 바이러스 유포, 사이버 범죄에의 이용 등과 같은 정보보호 역기능의 문제를 가지고 있으며, 그 위협은 나날이 증대되고 있는 실정이다. 이에 각 기업 및 기관들은 해킹 및 바이러스의 심각성을 자각하여 인터넷 상의 정보를 보호하고, 자사의 시스템을 보호하기 위한 노력에 큰 관심을 가지게 되었다 [1].

최근에는 보안시스템 도입을 통한 시스템 보호와 함께 네트워크과 시스템에 침입을 시도하는 해킹과 바이러스 등의 정보를 수집하고 분석하는 활동의 중요성이 크게 인식되었다 [2]. 그러나, 매월 수집되는 인터넷 침해와 관련한 데이터의 양은 너무 방대하며, 직접적인 침해대응에 이용하기에도 아직은 데이터에 대한 분석이 부족한 상황이다.

본 연구에서는 국내의 인터넷 침해 관련 통계데이터를 분석하여 미래 침해사고의 추이를 예측하고, 침해유형간의 상호연관성을 분석한다. 그러나, 예측과 관련한 정확한 분석을 하기에는 현재까지 보고된 국내 통계데이터의 일관성이 부족하여, 먼저 국내의 해킹바이러스 통계데이터의 시계열 자료를 이용한 상호상관분석을 통해 침해유형별 상관관계를 규명하고자 한다.

본 논문의 구성은 다음과 같다. 2장은 인터넷 침해사고의 주요 유형들을 설명하고, 이러한 인터넷 침해사고의 통계데이터를 이용한 트렌드 분석 현황을 서술한다. 3장은 통계 기법을 통해 이러한 데이터를 보다 효과적으로

본 연구는 정보통신부 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었음

침해대응에 이용할 수 있는 방법을 제시하며, 4장은 각 인터넷 침해사고의 시계열 자료를 이용하여 패턴과 침해사고간의 상호연관성을 분석한다. 5장은 결론이다.

2. 인터넷 침해사고

2.1 인터넷 침해유형

본 연구에서는 국내 인터넷 침해 관련 데이터를 분석하기 위하여 인터넷침해사고대응지원센터의 통계보고서를 이용한다 [3]. 국내외 시스템과 네트워크를 위협하는 많은 요소들이 통계보고서에 포함되어 있지만, 그 중에서도 정보시스템의 안전에 큰 영향을 끼치면서 최근 3년 동안 일관성 있게 수집되었다고 판단되는 바이러스, 스팸릴레이, 웜, 해킹을 대상으로 하여 패턴과 상호연관성을 분석하고자 한다. 본 연구에 이용되는 각 분석 대상의 유형은 <표 1>에 열거한 바와 같다.

<표 1> 인터넷 침해유형

바이러스	컴퓨터 프로그램이나 메모리에 자신 또는 자신의 변형을 복사해 넣는 악의적인 명령어들로 조합하여 불특정 다수에게 피해를 주기 위한 목적으로 제작된 모든 프로그램 또는 실행 코드
스팸릴레이	스팸 메일 발신자 추적을 어렵게 하기 위하여 타 시스템을 스팸 메일발송에 악용하는 행위
웜	독립적으로 자기복제를 실행하여 번식하는 빠른 전파력을 가진 프로그램 또는 실행코드
해킹	다른 사람의 컴퓨터나 정보시스템에 불법 침입하거나, 정보시스템의 정상적인 기능이나 데이터에 임의적으로 간섭하는 행위

2.2 인터넷 침해사고 분석 현황

인터넷 침해에 대한 분석은 다양한 분야에서 이루어지고 있지만 그 중 인터넷침해사고대응지원센터의 매월 통계보고서는 신고·접수된 침해사고에 대한 통계치를 제시하여 국내 침해사고 통계데이터 분석의 중추적 역할을 담당하고 있다. 이 통계보고서에서는 웜, 바이러스, 해킹, 스팸릴레이 등 각 유형별 침해 현황과 각 월의 특징 및 권고사항을 제시하고 있다. 최근에 들어서는 인터넷 침해에 대한 중요성이 강조되어 추가로 외부 공격을 유인해 현재 벌어지고 있는 침해 상황을 확인할 수 있도록 구성된 가상 네트워크인 허니넷(honey net)을 통한 분석을 제공하고 있으며 분석된 공격 현황을 각 분야별 카테고리로 나누어 상세한 통계치로 제시하고 있다. 하지만, 위에 열거한 분석 방법을 통해서는 현재의 피해 현황과 특징을 알아볼 수는 있지만, 앞으로의 패턴 및 상호연관성을 평가하기에는 부족한 실정이다. 따라서 본 연구에서는 보다 정확한 침해대응에의 근거를 제시하기 위하여 통계 데이터를 이용한 상호상관분석을 수행한다.

3. 상호상관분석

매월 보고되는 인터넷 침해사고 데이터간의 상호연관성을 분석하기 위해서는 시계열 자료간의 비교분석이 필요하다. 일반적으로 뇌파나 통신신호 분석에 주로 이용되는 상호상관분석을 이용하면 두 시계열 자료간에 상호상관을 수학적으로 표현할 수 있다 [4].

상호상관함수는 두 시계열 자료간의 유사성을 시간이 지연됨에 관계없이 판정하는 척도로 사용되거나, 두 시계열 자료가 시간적으로 어느 정도 지연되고 있는가를 알고자 할 때 사용된다. 즉 상호상관함수는 두 시계열 자료 사이의 상관 값을 시간축의 지연을 변수로 갖는 함수로 나타낸 것으로서, 함수 $f(t)$와 $g(t)$가

실수 t에 관한 함수라고 할 때 상호상관함수 $f \star g$는 다음과 같이 정의된다 [5]. 이는 $f(t)$와 $g(t)$가 연속이 아니고 이산함수여도 성립한다.

$$f \star g = \int_{\infty}^{-\infty} \overline{f}(\tau') g(t+\tau')(-d\tau')$$

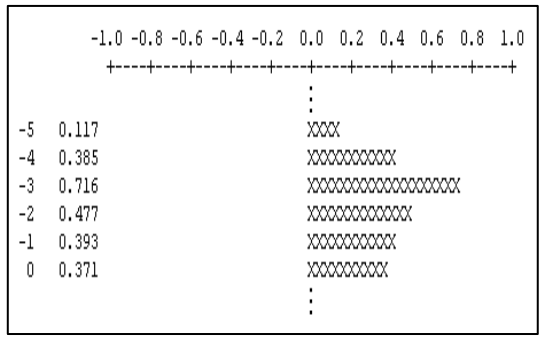
$$= \int_{-\infty}^{\infty} \overline{f}(\tau) g(t+\tau) d\tau.$$

본 연구에서 사용하는 통계데이터는 연속이 아닌 이산형 시계열 자료이므로, 적분(∫)대신 합(∑)으로 대체하여 주면 된다.

4. 패턴 및 분석 결과

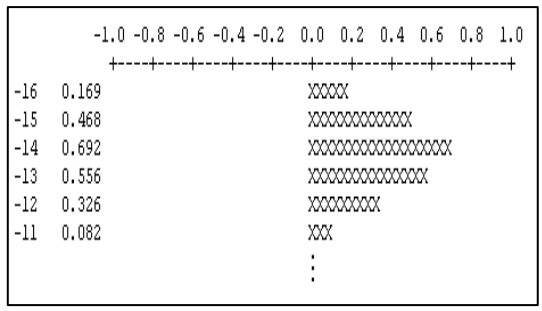
본 연구에서는 해킹, 바이러스, 웜, 스팸릴레이 간의 상관관계를 분석하였다. 매월 보고되는 이들 침해유형의 시계열 데이터를 이용하면 각 유형의 패턴을 얻을 수 있고, 두 시계열 데이터를 상호상관함수로 분석하면 각 유형간의 상관관계 결과를 얻을 수 있다. 분석에 이용한 데이터는 인터넷침해사고대응지원센터의 2002년 1월부터 2004년 12월까지의 통계보고서에 수록된 네 가지 침해유형의 데이터이다. 침해유형간의 상관관계 결과를 얻어내기 위한 도구로 MINITAB™의 MINITAB Release 13.1 Software에 Cross Correlation Function을 이용하여 총 6번의 상호상관분석을 시행한 결과 상호상관함수 값이 큰 다음의 두 결과를 얻을 수 있었다. 상호상관함수 값은 두 개의 시계열 자료가 완벽히 일치할 때 1의 값을 가지며, 아무런 상관이 없는 경우 0의 값을 가진다.

<그림 1>은 해킹과 바이러스 두 개의 시계열 데이터에 대한 상호상관함수의 결과 중 일부로 바이러스 시계열 데이터와 해킹의 시계열 데이터간에 3개월의 시간 지연이 있을 때에 0.716인 높은 상호상관함수 값을 가지는 결과를 볼 수 있다.



<그림 1> 해킹·바이러스 상호상관분석

<그림 2>는 바이러스와 스팸릴레이 두 개의 시계열 데이터에 대한 상호상관함수의 결과 중 일부로 바이러스 시계열 데이터와 스팸릴레이의 시계열 데이터간에 14개월의 시간 지연이 있을 때 0.692로 상호상관함수 값이 비교적 높은 값을 보였다.

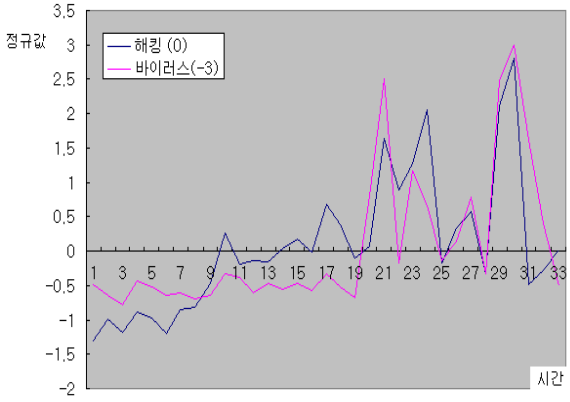


<그림 2> 바이러스·스팸릴레이 상호상관분석

상관관계가 크다고 볼 수 있는 두 데이터를 보다 자세히 분석하기 위해서는 두 데이터간의 패턴을 분석해 볼 필요가 있다. 그러나, 각 시계열 데이터간의 패턴을 단순비교하기에는 데이터간의 평균과 양의 차이가 있으므로, 이의 패턴을 정확히 비교분석하기 위해서는 정규화가 필요하다.

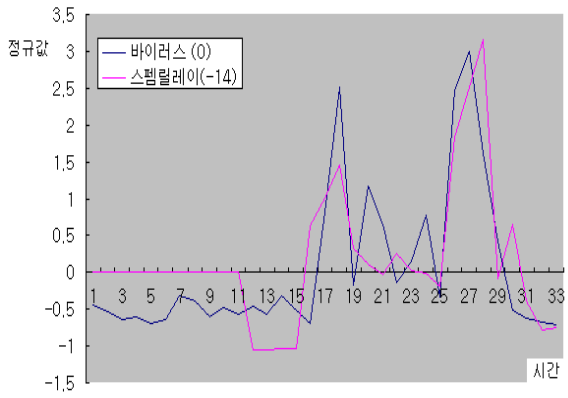
<그림 3>은 해킹과 바이러스 두 개의 시계열 데이터를 정규화 하여 그래프로 나타낸 것이다. 상호상관분석의 결과에서 나타난 것처럼 그래프의 추세 및 패턴이 바이러스의 시계열 데이터를 3개월 뒤로 미루었을 때 해킹과

유사함을 보인다. 이는 해킹의 발생빈도는 3개월 전의 바이러스의 발생빈도를 통하여 추이변화를 예측할 수 있음을 의미한다.



<그림 3> 해킹·바이러스 시계열 분석

<그림 4>는 바이러스와 스펨릴레이 두 개의 시계열 데이터를 정규화 하여 그래프로 나타낸 것으로 그래프의 추세 및 패턴이 스펨릴레이의 시계열 데이터를 14개월 뒤로 미루었을 때 해킹과 유사함을 보인다. 이는 바이러스의 발생빈도는 14개월 전의 스펨릴레이의 발생빈도를 통하여 추이변화를 예측할 수 있음을 의미한다. 그러나, 상호상관함수의 값이 크게 나타났음에도 불구하고 비교를 위한 시간지연이 너무 크므로, 바이러스와 스팸릴레이 사이의 상관관계를 명확히 규명하기엔 36개월의 데이터로는 무리가 있다.



<그림 4> 바이러스·스팸릴레이 시계열 분석

5. 결 론

최근 급속히 늘어나는 인터넷 침해사고에 대응하기 위해서는 인터넷 침해사고 통계데이터를 이용하여 침해유형들의 특성과 침해유형들 사이의 상관관계를 명확하게 파악한 뒤 이를 침해대응의 근거로 제시할 필요가 있다.

본 연구의 분석 결과에서 각각 해킹과 바이러스, 바이러스와 스펨릴레이는 서로 상호상관함수 값이 높게 나타났으며, 특히 해킹과 바이러스는 짧은 시간지연의 차이를 두고 유사한 추세를 나타내어 해킹에 대한 대응을 위해서는 바이러스의 발생 추이에 더욱 관심을 가져야 할 필요가 있다는 결론을 얻었다. 그러나, 인터넷 침해사고의 원인은 아직까지 정확히 규명하기 어려우므로 침해사고의 단순 추이 분석만으로는 적절한 침해대응을 하기가 어렵다. 따라서 보다 정확하고 폭넓은 통계데이터를 다양한 분석기법에 활용할 수 있다면, 침해사고의 정확한 예측모형 개발까지도 가능할 것이다.

참 고 문 헌

[1] 서동일, “최근 사이버 공격기술 및 정보보호 기술전망”, Digital Administration, 제92호, 2003.6.

[2] 윤영태, 류재철, 박상서, 박춘식, “Global Incident Trends 분석기술 동향”, 인터넷정보학회지, 제5권, 제1호, 2004.

[3] 인터넷침해사고대응지원센터, 인터넷 침해사고 동향 및 분석 월보, 2002.1-2004.12.

[4] 김철기, “코드분할 다중통신용 코드의 상호상관 특성”, 정보보호학회지, 제2권, 제2호, 1992.3.

[5] Bracewell, R. "Pentagram Notation for Cross Correlation." *The Fourier Transform and Its Applications, 3rd ed.* New York: McGraw-Hill, pp. 46 and 243, 1999.