웹 서버를 공격하는 웜 바이러스의 자기 유사성

정기훈° 송하윤 노삼혁
홍익대학교 컴퓨터공학과, 정보컴퓨터공학부

{khchong, song}@cs.hongik.ac.kr, samhnoh@hongik.ac.kr

Self-Similarity in Worm Attacks on Web Servers

Kihun Chong° Ha Yoon Song Sam H. Noh
Dept. of Computer Engineering, Information and Computer Engineering, Hongik Univ.

요 약

최근 인터넷의 폭발적인 확장에 따라서 웹 서버의 절대 개수가 증가하였으며, 웜 바이러스에 의한 웹 서버의 공격도 빈번해졌다. 바이러스의 무제한 자기 중식이라는 특성에 따라 인터넷상의 공격 패킷의 증가로 인하여 네트워크가 마비되고, 특히 IIS가 DoS공격으로 인하여 많은 피해를 입었다. 이에 따라 웜 바이러스의 공격을 방어하기 위한 많은 방법이 제시되었는데, 본 논문에서는 웜 바이러스의 공격 패턴을 분석하여 웜 바이러스의 공격 패턴의 특성이 어떠한지를 연구하였다. 대부분의 웜 바이러스 공격이 IIS에는 유효하지만 유닉스 시스템에서는 무력한 점을 이용하여, UNIX 시스템에서 구동되는 Apache 웹 서버의 로그 정보를 사용하여 웜 바이러스의 공격 패턴을 분석하였다. 분석 결과 웜 바이러스의 공격 패턴은 Hurst 상수 H에 의한 자기 유사성을 나타내는 것을 알 수 있었다.

1. 서론

몇 년 전부터 인터넷, 특히 World Wide Web (WWW) 사용의 폭발적인 증가와 함께 웹 서버의 수도 크게 증가하였으며, 서비스 역시 다양해졌다. 이와 동시에 인터넷을 이용한 웹 서버의 공격 방법 또한 다양해졌으며, 얼마 전부터는 컴퓨터 바이러스도 웹 서버를 공격하기 시작했다. 컴퓨터 바이러스의 자기 번식 특성에 힘입어 바이러스의 공격 패킷은 순식간에 엄청난 양으로 증가하는 경향이 있으며, 이에 따라 네트워크가 마비되고 웹 서버, 특히 IIS (Internet Information Server)역시 이러한 바이러스의 DoS (Denial of Service)공격과 IIS 자체에 대한 공격으로 인하여 많은 피해를 입었다. 따라서 이러한 웹 서버를 공격하는 바이러스의 패턴을 파악하여 DoS 등의 공격을 막는 것이 매우 중요하게 되었다.

네트워크 트래픽이 자기 유사성(*Self-similarity*)을 가지며 WWW 트래픽이 자기 유사성을 보인다는 것은 잘 알려진 사실이다[1]. 하지만 웹 서버를 공격하는 바이러스의 트래픽 패턴에 대해서는 연구가 거의 없었다. 따라서 본 연구에서는 웹 서버를 공격하는 바이러스의 요청에 대한 분석을 하였으며, 과연 이러한 요청 분포가 자기 유사성의 성격을 나타내는 가에 대한 측정을 하였다. 바이러스의 요청은 웹 서버에 대한 전체 요청의 부분집합이며, 웹 서버에 대한 요청은 자기 유사성을 보이므로 웹 서버를 공격하는 요청이 자기 유사성을 나타내는 지에 대한 여부는 상당히 흥미 있는 연구 주제라고 할 수 있고, 바이러스의 공격을 차단할 수 있는 하나의 방법이 될 수가 있다. 본 논문에서는 바이러스의 일종인 웜(*Worm*) 바이러스가 IIS를 공격하는 요청에 대하여 상세한 측정 및 분석을 하였다. 요청의 분석은 Apache 웹 서버의 로그 정보를 이용하였으며, 측정 결과 웹 서버를 공격하는 웜 바이러스가 나타내는 요청의 분포 역시 Hurst 상수(*Hurst parameter*) H에 의한 자기 유사성[2]을 나타내는 것을 알 수 있었다.

본 논문은 다음과 같은 구성으로 되었다. 우선 제 2절에서 관련 연구에 대하여 언급한다. 3절에서는 자기 유사성과 그에 따른 자기 유사성의 판별을 위한 테스트 방법에 대하여 알아본다. 4절에서는 웜의 특성 및 공격 패턴에 대해서 언급하고, 5절에서는 실제 웹 서버를 공격한 웜의 흔적을 이용하여 데이터를 분석한다. 마지막으로 제 6절에서 결론 및 향후 과제에 대하여 언급한다.

2. 관련 연구

일반적인 네트워크 트래픽이나 WWW 트래픽에 대한 통계학적 연구는 많이 이루어졌다[1][3][4]. 초기에는 LAN, WAN 등의 네트워크 트래픽의 자기 유사성에 대한 연구가 많았으며[3][4][5][6][7][8][9], 이후에는 WWW[1], 비디오 스트림[10], Wavelet[11] 등의 자기 유사성에 대한 연구가 진행되었다. 웹 서버를 공격하는 패턴에 대해서는 많은 연구가 이루어지지 않았으며, 특히 바이러스의 웹 서버 공격에 대한 연구는 거의 이루어지지 않았다. 바이러스에 대해서는 발생일, 특성, 위험 요소 등에 대한 조사와 그에 따른 백신을 제작하는 것이 일반적이며[12][13][14], 최근에 활발해진 웜에 대해서도 공격 방법이나 치료 방법까지만 연구를 하는 것이 보통이다[15][16]. 따라서 웜의 웹 서버 공격 패턴에 대한 연구는 기존의 치료를 위한 바이러스 연구와는 다른 방향의 연구이며, 나아가서는 방어를 위한 방법을 제시할 수 있는 연구라고 할 수 있다.

3. 자기 유사성과 판별을 위한 통계학적 방법

이 절에서는 자기 유사성의 의미, 그리고 자기 유사성의 수학적인 정의 및 표현에 대해서 정리한다.

3.1 자기 유사성의 의미

자기 유사성은 기본적으로 프랙탈[17]의 속성인 자기 유사성과 순환성을 따른다. 자기 유사성은 프랙탈과 같이 부분이 전체와 같은 형태를 띄는데, 특히 통계학 관점에서 바라본 형태를 의미한다. 특히, 시간 t에 대한 시계열(*time process or time series*)형태의 집합을 다루게 되며, 그러한 예를 그림 2에서 보여주고 있다[2]. 그림 2에 나타난 그래프는 네트워크 트래픽을 나타내는데, 시간 단위(*time unit*)의 크기(*scale*)의 변화에 대해서 마치 프랙탈 도형과 비슷한 형태를 보여주고 있다. 이러한 속성을 자기 유사성이라고 한다.

3.2 자기 유사성의 정의

시계열을 다음과 같이 표현하기로 한다.

$X(t)$: 시계열, $t \in Z$ (Z는 자연수)

여기서 $X(t)$ 값으로는 시간 t동안의 패킷 수, 바이트 수, 비트 수

등이 될 수 있다. 또는 시간 t까지의 패킷 수, 바이트 수, 비트 수 등의 누적 값이 될 수도 있다.

본 논문에서는 시계열 X가 다음을 만족하면 X는 상수 H에 의한 자기 유사성을 갖는다고 하겠다.

$$X_t =_d m^{-H} \sum_{i=(t-1)m+1}^{tm} X_i \quad (\text{for all } m \in \mathrm{N})$$

또한, $X^{(m)}$에 대해서 같은 자기 상관 함수(*autocorrelation function*) $r(k)=\frac{E[(X_t-\mu)(X_{t+k}-\mu)]}{\sigma^2}$를 갖는 시계열 X는 상수 H에 의한 자기 유사성을 갖는다고 한다. 이 때, Hurst 상수 H는 다음을 만족해야 한다.

$$\frac{1}{2} < H < 1$$

3.3. 자기 유사성 판별을 위한 통계학적 방법

이 절에서는 자기 유사성을 판별하기 위한 방법론에 대해서 논한다. 이러한 방법으로는 *variance-time plot, R/S plot, Periodogram* 등이 있다[1][3].

3.3.1 Variance-time plot

시계열 $X^{(m)}$에서 큰 값 m에 대한 $X^{(m)}$의 분산 $\mathrm{var}(X^{(m)})$ 값의 분포를 로그-로그 플롯(*log-log plot*)에 나타낸 것을 말한다. 이 플롯에 그려지는 분포의 대략적인 기울기를 β라고 한다면 H는 $H=1-\beta/2$를 만족하며, 이렇게 구한 H가 $\frac{1}{2} < H < 1$를 만족하면 $X^{(m)}$은 상수 H에 의한 자기 유사성을 갖는다고 할 수 있다.

3.3.2 R/S plot

시계열 $X^{(m)}$을 R/S값으로 크기변환(*rescale*)한 것을 m에 대한 그래프(*graph*)로 로그-로그 플롯에 나타낸 것을 말한다. 여기서 R과 S는 다음과 같이 나타낸다.

$$S(n)=\sqrt{\mathrm{var}(X^{(m)})}$$
$$R(n)=\left[\max(0,W_1,W_2,\cdots,W_n)-\min(0,W_1,W_2,\cdots,W_n)\right]$$
$$\left(W_k=(X_1+X_2+\cdots+X_k)-k\bar{X}(n),\ k=1,2,\cdots,n\right)$$

이 때, 로그-로그 플롯에 나타난 그래프의 대략적인 기울기 값이 H가 되며, H가 $\frac{1}{2} < H < 1$를 만족하면 $X^{(m)}$은 상수 H에 의한 자기 유사성을 갖는다고 할 수 있다.

3.3.3 Periodogram

주기에 따른 periodogram을 로그-로그 플롯에 나타낸 것을 말한다. Periodogram은 다음의 식을 이용하여 구한다.

$$I(\lambda)=\frac{1}{2\pi N}\left|\sum_{j=1}^{N} X_j e^{ji\lambda}\right|^2$$

여기서, λ는 주기, N은 시리즈의 개수이다. 로그-로그 플롯에 나타나는 분포의 대략적인 기울기를 β라고 한다면 $H=1-\beta/2$를 만족하며, 이렇게 구한 H가 $\frac{1}{2} < H < 1$를 만족하면 $X^{(m)}$은 상수 H에 의한 자기 유사성을 갖는다고 할 수 있다.

4. 웹 서버 공격

WWW의 사용이 활발해짐에 따라 일반 서버의 공격뿐 만 아니라 웹 서버에 대한 공격도 증가하고 있다. 특히, Internet Information Server(IIS)[18]에 대한 공격이 가장 활발하다. IIS는 Microsoft에서 만든 웹 서버이며 MS Windows 운영체제에서 실행하는데, Windows의 취약한 보안 특성 때문에 전세계의 해커들이 상당히 많은 공격을 하고 있다. 보안 업데이트가 자주 이루어지지만 새로운 공격 방법도 빠르게 나타난다.

예전에는 바이러스가 파일에 감염되어 침입한 시스템의 파일 시스템을 무력화 하는 것이 일반적인 공격 방법이었다. 하지만 최근에는 침입한 시스템이 연결된 네트워크의 모든 시스템을 공격하거나 또는 네트워크 그 자체를 무력화하는 등의 더욱 적극적인 공격을 하는 추세로 변하고 있다. 심지어는 웹 서버의 내용까지 변화시켜 해당 웹 서버에 접속하는 모든 클라이언트에게 감염시키거나 웹 서버에 DoS공격을 하기도 한다. 그 대표적인 예로, Windows 시스템을 공격하는 Code Red, Nimda Worm(W32/Nimda worm) 등이 있다. 특히 Code Red보다 나중에 만들어진 Nimda worm 바이러스는 Code Red 바이러스가 생성해 놓은 백 도어(*back door*)까지 검색을 하는 지능적인 면을 보이기도 한다. Nimda worm 바이러스의 공격 패턴은 다음과 같다.

```
① GET /scripts/root.exe?/c+dir
② GET /MSADC/root.exe?/c+dir
③ GET /c/winnt/system32/cmd.exe?/c+dir
④ GET /d/winnt/system32/cmd.exe?/c+dir
⑤ GET /scripts/..%5c../winnt/system32/cmd.exe?/c+dir
   ...
⑥ GET /scripts/..%2f../winnt/system32/cmd.exe?/c+dir
```

□~□는 Code Red II에 의해서 만들어진 백 도어를 검색 시도하는 로그이며, 나머지는 디렉토리 탐색(*directory traversal*) 취약점을 공격하기 위한 로그이다.

5. 측정 및 결과

웜 바이러스가 남긴 로그 데이터는 본교 컴퓨터공학과의 웹 서버를 공격한 것에 대한 로그 정보를 수집한 것으로 사용하였다. 이번 절에서는 데이터의 자세한 정보 및 웹 서버의 시스템 정보에 대하여 알아본 후, 데이터를 분석한 결과에 대해서 설명한다.

5.1 실험 환경

대부분의 웹 서버의 공격은 MS Windows와 IIS에 집중되었으며, Unix의 특성과 Apache 웹 서버[19]의 특성상 Apache 웹 서버에는 무력한 공격이 된다. 따라서 Apache 웹 서버에는 공격에 대한 로그만이 남게 되는데 이러한 로그를 이용하여 웹 서버 공격 정보를 수집하였다. 로그 데이터의 크기는 정제하기 전의 로그 데이터가 343MB, 웜 바이러스의 공격에 대한 로그만 추출해낸 후의 로그 데이터가 13MB이다. 사용한 웹 서버는 Apache 1.3.9이며, 웹 서버가 동작한 시스템은 SUN SPARC station-10이고, 운영체제는 UNIX System V release 4.0 (SunOS 5.5)이다. 로그를 수집한 기간은 웜 바이러스가 본격적으로 공격을 시작하기 시작한 2001년 10월 1일 0시부터 시작하였으며 2002년 2월 28일 24시까지의 로그 정보를 사용하였다.

5.2 측정 결과

그림 2는 time unit이 1, 10, 100, 1000초 일 때의 데이터 분포를 나타내고 있다. x축은 시간의 진행을 나타내며, y축은 해당 시간의 공격 회수를 나타낸다. 그림 2에서 보는 것과 같이 각 공격의 분포는 다른 시간 단위에 대해서 유사한 모양을 나타내는 것을 알 수 있다. 이러한 결과는 웹 서버 공격 패턴도 자기 유사성을 갖는다는 것을 보여준다.

그림 3의 왼쪽 그래프는 variance-time plot을 보여준다. 점선이 variance값을 나타낸 그래프이며, 실선은 최소 제곱법을 이용하여 구한 근사 직선이다. 측정 결과 로그-로그 플롯에서의 β값은 0.65로 나타났으며, 따라서 H값은 0.67이 된다.

그림 3의 가운데 그래프는 R/S plot을 보여준다. R/S plot을 살펴보면, x축의 n은 $X^{(m)}$의 m과 같은 값이며 많이 그려진 점들이 r/s값을 나타낸다. 분포한 점들의 위 아래로 있는 직선은 로그-로그 플롯에서의 기울기가 각각 0.5, 1인 직선이며, 따라서 H값이 0.5와 1사이에 있다는 것을 보여준다. 실제로 최소 제곱법을 이용하여 구한 H값은 0.57이었다.

그림 3의 오른쪽은 periodogram을 그래프로 그린 것이다. 전체적으로 찍혀 있는 점들이 주기에 따른 periodogram값이며, 점들 사이로

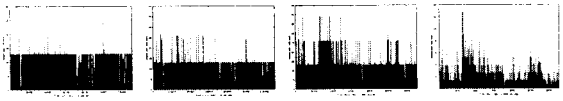

그림 2. Virus의 web server 공격 request분포

그어진 직선은 점들에 대한 최소제곱법을 이용하여 그린 근사 직선이다. 로그-로그 플롯에서의 직선의 기울기 β는 0.11으로 나타났으며, 따라서 H값은 0.94가 된다. 결론적으로, 위의 세 가지 방법에서 모두 H가 $\frac{1}{2}<H<1$를 만족하므로 웜 바이러스의 공격 패턴은 상수 H에 의한 자기 유사성을 갖는다고 말할 수 있다.

6. 결론 및 향후 연구 과제

본 논문에서는 웹 서버를 공격하는 웜 바이러스가 나타내는 요청의 분포에 대한 분석을 하였다. 대부분의 웜 바이러스가 Windows 운영체제의 웹 서버인 IIS를 공격하기 때문에, 전혀 다른 플랫폼인 UNIX 운영체제에서는 같은 공격이 통하지 않으며, 따라서 UNIX 운영체제에서 사용하는 웹 서버에는 전혀 지장을 주지 않는다. 본 논문에서는 UNIX 운영체제에서 사용하는 Apache 웹 서버가 남긴 로그를 이용하여 웜 바이러스의 흔적을 찾아내었으며, 이를 이용하여 웜 바이러스가 어떤 패턴으로 공격을 하는지에 대한 정보를 추출하였다. 추출한 정보를 이용하여, 웜 바이러스의 요청 분포에 대한 통계학적인 조사를 하였다. 자기 유사성 여부에 대한 조사를 위하여 Variance-time plot, R/S plot, periodogram 등의 방법을 사용하였으며, 측정 결과 웜 바이러스가 나타내는 요청의 분포는 Hurst 상수 H에 의한 자기 유사성을 갖는 것으로 나타났다. 이로써 웜 바이러스의 공격 여부를 판단할 수 있게 되었으며, H값을 이용하여 이에 따른 대비를 할 수 있게 되었다.

향후 연구 과제로 실제 웹 서버를 공격하는 웜 바이러스를 H값을 이용하여 실시간으로 가려내는 것이 가능한지에 대한 연구와 웜 바이러스간의 구별 여부에 대한 연구, 그리고 웜 바이러스를 이용한 공격 이외의 다른 방법으로 웹 서버를 공격하는 것에 대한 통계학적인 연구를 들 수 있다.

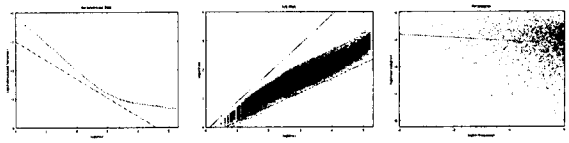

그림 3. Variance-time plot, R/S plot, Periodogram

References

[1] Mark E. Crovella and Azer Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, pp.835-846, 1997

[2] Kihong Park and Walter Willinger, "Self-Similar Network Traffic: An Overview," *Self-Similar Network Traffic and Performance Evaluation*, p.1, John Wiley and Sons, New York, 2000

[3] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson, "On the self-similar nature of Ethernet traffic (extend version)," *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp.1-15, 1994

[4] Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson, "Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, pp.71-86, 1997

[5] Vern Paxson, "Fast, Approximate Synthesis of Fractional Gaussian Noise for Generating Self-Similar Network Traffic," *Computer Communication Review*, Vol. 27, No. 5, pp.5-18, 1997

[6] Boris Tsybakov and Nicolas D. Georganas, "On Self-Similar Traffic in ATM Queues: Definitions, Overview Probability Bound, and Cell Delay Distribution," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 3, pp.397-409, 1997

[7] Kihong Park, Gitae Kim, and Mark E. Crovella, "On the Effect of Traffic Self-similarity on Network Performance," *Proceedings of the 1997 SPIE International Conference on Performance and Control of Network Systems*, pp.296-310, 1997

[8] Kihong Park, Gitae Kim, and Mark E. Crovella, "The Protocol Stack and Its Modulating Effect on Self-Similar Traffic," *Self-Similar Network Traffic and Performance Evaluation*, p.349, John Wiley and Sons, New York, 2000

[9] Kihong Park and Tsunyi Tuan, "Performance Evaluation of Multiple Time Scale TCP Under Self-Similar Traffic Conditions," *ACM Transactions on Modeling and Computer Simulation*, Vol. 10, No. 2, pp.152-177, 2000

[10] Mark W. Garrett and Walter Willinger, "Analysis, Modeling and Generation of Self-Similar VBR Video Traffic," *Proceedings of ACM SIGCOMM 94*, pp.269-280, 1994

[11] Patrice Abry and Darryl Veitch, "Wavelet Analysis of Long Range Dependent Traffic," *IEEE Transactions on Information Theory*, Vol. 44, No. 2, pp.2-15, 1998

[12] KISA, Cyber 118, `http://www.cyber118.or.kr`

[13] Ahnlab, Inc., `http://www.ahnlab.com`

[14] Hauri, Inc., `http://www.hauri.co.kr`

[15] Korea Information Security Agency, `http://www.kisa.or.kr`

[16] KISA, Coordination Center, `http://www.certcc.or.kr`

[17] Benoit B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, New York, 1983

[18] Microsoft Corporation, Internet Information Services Features, `http://www.microsoft.com/windows2000/server/evaluation/features/web.asp`

[19] Apache Software Foundation, The Apache web server, `http://www.apache.org`

[20] Murad S. Taqqu, Vadim Teverovsky, and Walter Willinger, "Estimators for long-range dependence: an empirical study," *Fractals*, vol. 3, no. 4, pp.785-798, 1995

[21] Stefano Giordano, Riccardo Pannocchia, and Franco Russo, "Estimation of the Hurst Parameter: Analysis of the Burstiness of Self-Similar Traffic Models," *IEEE International Conference on Communication Systems ICCS '96*, Vol. 3, pp.34.2.1-34.2.6, 1996

[22] O. Rose, "Estimation of the Hurst Parameter of Long-Range Dependent Time Series," *Technical Report #137*, Institute of Computer Science, University of Würzburg, 1996

[23] Murad S. Taqqu and Vadim Teverovsky, *Robustness of Whittle-type Estimators for Time Series with Long-Range Dependence*, Preprint, 1997

[24] Murad S. Taqqu and Vadim Teverovsky, "On Estimating the Intensity of Long-Range Dependence in Finite and Infinite Variance Time Series," *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Birkhauser, Boston, 1998

[25] Sergio Ledesma and Derong Liu, "Synthesis of Fractional Gaussian Noise Using Linear Approximation for Generating Self-Similar Network Traffic," *Computer Communication Review, a publication of ACM SIGCOMM*, Vol. 30, No. 2, 2000

[26] Carlos Velasco and Peter M. Robinson, "Whittle Pseudo-Maximum Likelihood Estimation of Nonstationary Time Series," *Journal of the American Statistical Association*, Vol. 95, No. 452, pp.1229-1243, 2000