자기 유사성을 갖는 World Wide Web 트래픽 소스 모델링에 관한 연구

김동일

동의대학교 정보통신공학과

A Study on the World Wide Web Traffic Source Modeling with Self-Similarity

Dong il Kim

Dongeui University Dept. of Information & Comm

E-mail : dikim@dongeui.ac.kr

요 약

네트워크 트래픽의 연구 동향은 LAN, WAN 및 VBR 비디오 트래픽이 통계학적으로 자기유사과정에 의해 더욱 더 잘 모델링화 된다는 사실을 입증하고 있다.

이것은 기존의 단기간 의존성만을 고려한 포아손 과정에 비해 다른 이론적 특성을 가진다. 즉 광범위한 시계열상에서 aggregation의 정도가 변하더라도 통계학적으로 동일한 특성을 지니게 된다는 것이다.

본 논문에서는 이러한 자기유사과정을 갖는 데이터를 실시간 운영중인 네트워크로부터 측정하여 트래픽을 모델링하고, 비교분석하므로서 다양한 데이터를 지원하는 초고속 네트워크의 성능분석에 적용할 수 있으리라 여겨진다.

ABSTRACT

Traditional queueing analyses are very useful for designing a network's capacity and predicting there performances, however most of the predicted results from the queueing analyses are quite different from the realistic measured performance. And recent empirical studies on LAN, WAN and VBR traffic characteristics have indicated that the models used in the traditional Poisson assumption can't properly predict the real traffic properties due to under estimation of the long range dependence of network traffic and self-similarity

In this paper self-similar characteristics over statistical approaches and real time network traffic measurements are estimated

It is also shown that the self-similar traffic reflects network traffic characteristics by comparing source model.

키워드

self-similar traffic, source modeling, traffic, www traffic

Ⅰ. 서 론

네트워크의 설계 및 성능평가에 앞서 그 네트워크의 트래픽 특성을 이해하는 것이 매우 중요하다 하겠다. 네트워크 트래픽에 관한 최근의 논문들은 LAN, WAN 및 VBR 비디오 트래픽이 통계학적으로 self-similar 과정에 의해 더욱 더 잘 모델링 된다는 것을 설득력 있게 주장해 왔다[1][2][3]. 이것은 기존의 단기간 의존성만을 고려한 Poisson 과정에 비해 매우 다른 이론적 특성을 가진다. 즉, 광범위한 time-scale상에서 aggregation의 정도가 변하더라도 통계학적으로 동일한 특성을 지니게된다는 것이다. 이러한 특징들은 자기유사성(Self-similarity), 장기간 의존성(Long-Range Dependence, Joseph Effect), 무한분산 증후군(Infinite Variance Syndrome, Noah Effect), 천천히 감소하는 분산(Slowly Decaying

Variance) 등으로 대변된다[4][5]. 본 논문에서는 이러한 self-similar 과정의 정의 및 정도에 대한 추정법에 대해 살펴보고, 실시간 운영중인 LAN으로부터 1초 간격으로 약 백만개 이상의 트래픽을 측정하여, 여러 가지 self-similar 특성에 대해 통계학적으로 분석하였다.

II. Self-Similar 정의 및 정도에 대한 추정법

1. Self-Similar 정의

self-similar 확률과정의 일반적인 정의는 다음과 같이 연속시간 변수의 직접 스케일링에 기초한다. 어떠한 실수 $a>0$에 대해, 확률과정 $a^{-H}x(at)$가 $x(t)$와 통계적으로 동일한 특성을 가진다면, 확률과정 $x(t)$는 파라미터 H($0.5<H\leq 1$)를 가지고 통계적으로 self-similar하다. 이러한 관계는 다음의 3가지 조건으로 표현된다.

$$E[x(t)] = E[x(at)] \quad \text{Mean} \tag{1}$$

$$Var[x(t)] = \frac{Var[x(at)]}{a^{2H}} \quad \text{Variance} \tag{2}$$

$$R_x(t,s) = \frac{R_x(at,as)}{a^{2H}} \quad \text{Autocorrelation} \tag{3}$$

Hurst 또는 self-similarity 파라미터 H는 self-similarity의 핵심척도이다. 다시 말하면, H는 통계적인 현상의 지속성(persistence)에 대한 척도이고 확률과정의 장기간 종속에 대한 척도이다. H=0.5의 값은 self-similarity의 부재를 나타내고, H가 1에 가까울수록, 지속성의 정도 또는 장기간의 종속의 정도는 더욱 커진다. 즉, H=0.5에 대하여 과거와 미래의 증가에 대한 상관성이 없어지고, H>0.5에 대하여 지속성의 두드러진 특징을 가진다

2. Self-similarity의 정도에 대한 추정법

본 장에서는, 주어진 실제 데이터의 시계열(time series)이 self-similar한지, 만약 그렇다면 self-similarity의 강도가 얼마나 되는지를 나타내는 Hurst 파라미터를 추정하는 방법 중 본 논문에서는 3가지 방법에 관해서 살펴보겠다.

2.1 Variance-time plot

self-similar 과정의 m-aggregated 시계열 X(m)에 대해, 분산은 매우 큰 m에 대해 다음을 따른다.

$$Var(x^{(m)}) \sim \frac{Var(x)}{m_\beta} \tag{4}$$

$$\log[Var(x^{(m)})] \sim \log[Var(x)] - \beta\log(m) \tag{5}$$

여기에서, self-similarity 파라미터 $H=1-(\beta/2)$이다. 이것을 log-log 그래프 상에 m에 대한 Var(x(m))을 그리게 되면, 그 점들은 $-\beta$의 경사를 가지는 직선이 나오게 된다. 이 점들의 기울기는 최소 자승 직선 근사(least square line fitting) 법[6]으로 쉽게 구할 수 있다.

2.2. Index of Dispersion for Counts(IDC)

서로 다른 타임 스케일동안의 트래픽에 대한 변이성을 나타내는 척도로서 카운트에 대한 산포지수(index of dispersion)가 사용된다. 주어진 길이 L의 시간 간격에 대해, IDC는 길이 L의 간격 동안 도착수의 분산을 동일한 양의 평균값으로 나눈 것이다.

$$\log[Var(x^{(m)})] \sim \log[Var(x)] - \beta\log(m) \tag{6}$$

$$IDC(L) \sim cL^{\gamma} \tag{7}$$

여기에서 c는 L에 독립인 양의 값을 가지는 상수이고, $\gamma=2H-1$의 값을 가진다. log(L)에 대한 log(IDC(L))을 그리게 되면 경사 γ를 갖는 점근적인 단조증가 직선이 나온다[6].

2.3 Periodogram Method

; X0, X1,.....,Xk-1이 k개의 표본을 갖는 이산시계열일 때, 이 시계열의 DFT는 다음과 같다.

$$\tilde{x_k}(f) = \sum_{m=0}^{k-1} X_m e^{-j2\pi f_m} \tag{8}$$

$\tilde{x_k}(f)$의 크기를 제곱한 것은 주파수 f에서의 에너지를 나타낸다. 만약, 이 에너지를 전체 시간 k로 나누면, 주파수 f에서의 파워의 추정값을 얻을 수 있다.

$$\tilde{p_k}(f) = \frac{1}{k}|\tilde{x_k}(f)|^2 \tag{9}$$

이것을 다르게 표현하면, 다음과 같다.

$$I_N(\omega) = S(\omega) = \frac{1}{2\pi N}\left|\sum_{k=1}^{N} X_k e^{jk\omega}\right|^2 \tag{10}$$

이것을 주기도(periodogram) 또는 강도(intensity) 함수라고 한다. 또한, 스펙트럼 밀도 $S(\omega)$는 자기상관함수 R(k)의 푸리에 변환쌍이기 때문에 자기상관함수를 이용해서 손쉽게 스펙트럼 밀도를 계산할 수 있다.

$$S(\omega) \sim \frac{1}{|\omega|^{\gamma}} \quad as\ \omega \to 0,\ 0<\gamma<1 \tag{11}$$

$$\log[S(\omega)] \sim -\gamma\log[|\omega|] \tag{12}$$

이것은 log-log 그래프 상에서 기울기가 $-\gamma$($\gamma=2H-1$)인 직선이 나온다. 실제로, 이것은 원점 근처에서만 상기(식.11), (식.12)를 만족하기 때문에 ω의 최하위 10%만을 사용한다[6].

III. WWW 트래픽의 측정 및 분석

WWW 트래픽의 self-similarity에 대한 수학적, 통계학적으로 정확한 특성을 연구하기 위해서는 많은 량의 WWW 트래픽 트레이스(trace)가 필요하다[7]. 본 논문에서는 동의대학교의 교내 전산망의 한 서브넷에서 트래픽 측정을 수행하였다. 측정은 HP사의 Internet Advisor를 사용해 1회에 걸쳐 수행하였으며, 각각의 데이터 세트들은 백만개 이상의 표본경로를 갖는다.

3.1. WWW 트래픽 측정

WWW 트래픽 측정은 산학관의 한 서브넷(203.241.205.)에서 수행하였으며, 측정기간은 2001년 10월 2일~10월 13일 걸쳐서 각각 백만개이상의 트래픽 표본을 수집하였다. 데이터 세트는 자기발견적인(heuristic) 방법에 의한 self-similarity를 고찰하기 위해, WWW 패킷, 바이트의 트레이스들을 수집하였다. 트래픽 샘플링 간격은 각각 1초 단위로 수행하였다. 표 1은 비록 전체망에 대한 트래픽 수집은 아니지만, WWW 트래픽의 self-similar 특성에 대해 분석하기에는 충분할 것이다. 표 1에는 각 트래픽 측정의 데이터 세트들에 대한 설명을 요약해 놓았다.

표 1. 분석에 사용된 WWW 트래픽 세트에 대한 설명

WWW 트래픽 측정 트레이스		
시작시간 (2001년 10월2일)	OCT2001.PKT	패킷트레이스
종료시간 (2001년 10월 13일)	OCT2001.BYT	바이트트레이스

3.2 OCT2001.PKT 데이터 세트에 대한 분석결과

그림 1은 OCT2001.PKT 데이터 세트의 전체 트래픽 트레이스를 보여준다. 그리고, 그림 2, 3 ,4는 각각 variance-time, IDC, periodogram plot을 나타내고, 분석결과 아주 높은 self-similarity를 나타내었다. 표 2는 2001년 패킷 데이터 세트에 대한 각 self-similar 파라미터 추정값들을 나타낸다. 표 2에서 알 수 있듯이, V-T plot과 IDC plot의 경우 거의 비슷한 값이 나왔지만, periodogram법의 경우 추정구간에 따라 다소 차이가 남을 알 수 있다.

3.3 OCT2001.BYT 데이터 세트에 대한 분석결과

바이트수에대한 트래픽 트레이스에서도 상당히 버스트함을 볼 수 있었고, 패킷수에 비해 약간 더 높은 Hurst 파라미터 추정값이 계산되었다. 바이트 수에 대해 높은 self-similarity를 갖는다는 것은 전통적인 큐잉이론에 기초해 계산된 필요 버퍼량보다 실제의 트래픽에서는 더 많은 버퍼를 요구한다는 것을 의미한다. 그림 5, 6 ,7는 각각 variance-time, IDC, periodogram plot을 나타낸다. 표 3은 2001년 바이트 데이터 세트에 대한 각 self-similar 파라미터 추정값들을 나타낸다.

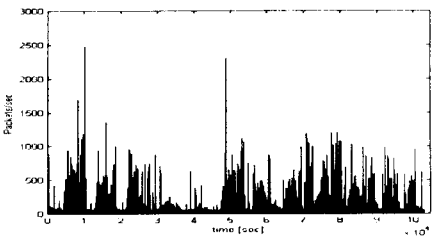

그림 1. OCT2001.PKT의 트래픽 트레이스

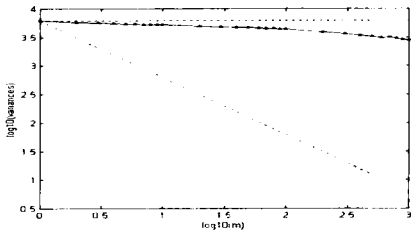

그림 2. OCT2001.PKT에 대한 variance-time plot

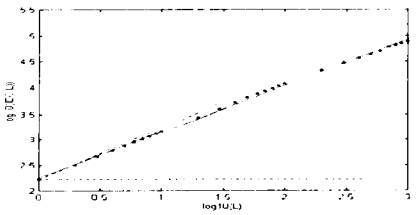

그림 3. OCT2001.PKT에 대한 IDC plot

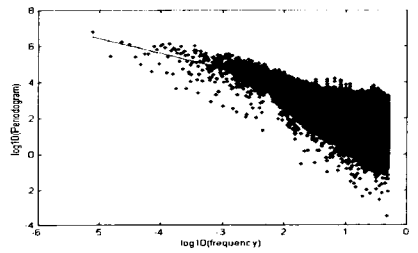

그림 4. OCT2001.PKT에 대한 periodogram plot

표 2. OCT2001.PKT에 대한 Hurst 파라미터 추정값

Variance-time plot		IDC plot		Periodogram plot	
β	H	γ	H	γ	H
0.1116	0.9442	0.8885	0.9442	0.8386	0.9193

표 3. OCT2001.BYT에 대한 Hurst 파라미터 추정값

Variance-time plot		IDC plot		Periodogram plot	
β	H	γ	H	γ	H
0.0964	0.9518	0.9037	0.9518	0.8852	0.9426

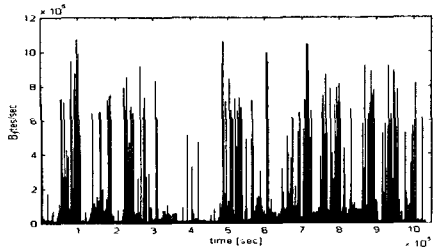

그림 5. OCT2001.BYT의 트래픽 트레이스

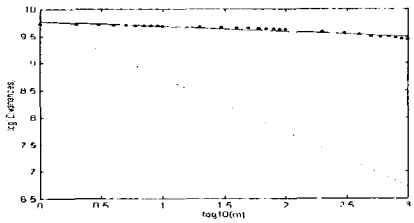

그림 6. OCT2001.BYT에 대한 Variance-time plot

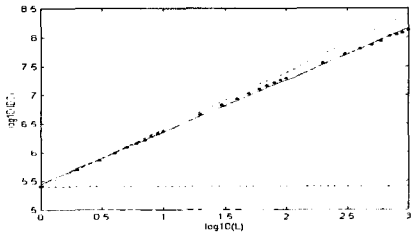

그림 7. OCT2001.BYT에 대한 IDC plot

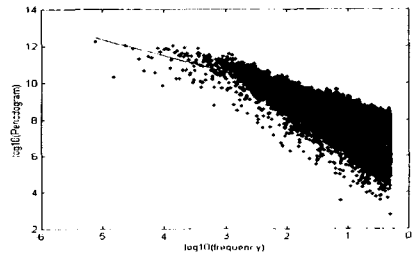

그림 8. OCT2001.BYT에 대한 Periodogram plot

IV. 결 론

초고속, 광대역 통신망에서 트래픽 특성을 이해하는 것은 네트워크의 성능평가에 핵심이 된다. 기존의 트래픽 모델들은 패킷 도착과정이 단기간 의존성을 갖는 분포라고 가정하였고, 이러한 모델들은 광범위하게 사용되어 왔으며, 이들의 해석적인 용이성 때문에 큐잉이론 및 성능분석가들에게 많이 사용되어왔다. 그러나, 최근 몇 년 동안 이러한 소스들의 많은 멀티플렉싱이 측정된 트래픽과 일치하지 않는다는 것이 판명되었고, 실제 트래픽의 LRD 특성 및 self-similar 특성을 정확하게 반영하지 않은 Poisson이나 다른 모델들을 사용해 네트워크 트래픽을 모델링 하는 것은 시뮬레이션 및 분석에서 평균 패킷 지연이나 최대 큐 사이즈와 같은 성능척도를 상당히 과소평가 하는 결과를 낳았다. 따라서 본 논문에서는 실제 운영중인 네트워크에서 WWW 트래픽을 수집하여 다양한 방법으로 버스트의 강도를 측정한 결과 아주 버스트한 트래픽의 성질을 가지고 있음을 판명하였고 실제 WWW 트래픽에 강한 self-similar한 특성이 존재한다는 사실을 보였다. 따라서 이러한 분석은 다양한 데이터를 지원하는 초고속 네트워크의 성능분석에 중요한 요소로 적용할 수 있으리라 여겨진다.

참고문헌

[1] Leland, W., Taqqu, M.,Willinger,W.,Wilson,D. "On the Self-similar Nature of Ethernet Traffic(Extended Version)", IEEE/ACM Transaction on Networking, Feb,1994.

[2] J.Beran, R. Sherman, M.S. Taqqu and W. Willinger, "Long-Range Dependence in Variable Bit Rate Video Traffic", IEEE Transaction on Communications 43, No.2/3/4, pp 1566-1579, 1995.

[3] Willinger,W.,Wilson,D.,Taqqu,M. "Self-similar Traffic Modeling for Highspeed Networks", ConneXions,Nov,1994.

[4] Peyton Z. Peebles,JR. Probability, "Random Variables, and Random Signal Principles", McGraw Hill, p134-198,1993

[5] J.H. Mathews, Kurtis D. Fink, "Numerical Methods Using Matlab", 3rd Edition, Prentice Hall, pp 253-278, 1999.

[6] 김창호, 김동일 외, "트래픽에서의 장기간 의존성 및 Self-similar 특성", 하계통신학회 Proc., p463-467, 1999.

[7] 김창호, 김동일 외, "Ethernet 트래픽의 장기간 의존성 및 Self-Similar 트래픽 소스 모델링에 관한 연구', Telecomm. Review, vol.11, No.6, 2001