두꺼운 꼬리 분포를 이용한 수정된 인터넷 트래픽 모델

(A Strategy of Adjusted Internet Traffic Modeling using Heavy-Tailed Distributions)

지 선 수*
(Seon-su Ji)

요 약 인터넷의 사용자가 증가함에 따라 사용자의 인터넷 작업부하 유형을 파악하고 이를 수리적으로 표현하여 웹 사이트에서 송수신되는 트래픽의 형태를 모델링하는데 적합한 통계적 분포를 제시할 필요가 있다. 독립적이고 불규칙적인 트래픽 추세를 고려할 때 어떤 t 시점에서 임의의 외부적 개입이 주어질 경우 파레토 분포를 이용한 수정된 모델을 가지고 트래픽 발생에 어느 정도 영향을 주는지를 알아본다.

핵심주제어 : 개입, 요청구간, 웹 작업부하, 트래픽 추세, 파레토분포

Abstract According to the recent growth of the internet commercialization and differentiated QoS(quality of service), statistical traffic modeling is necessary for forecasting and controlling future network capacity. This paper reviews the essential components in web workloads. And I propose adjusted internet traffic modeling using heavy-tailed distributions and intervention techniques.

Key Words : Intervention, Requested Interval, Web Workloads, Pareto Distribution, Traffic Trends

1. 서 론

오늘날의 인터넷에서 네트워크 자원(resource)은 역동적으로 많은 사람들 사이에 공유하고 개별 이용자에게 QoS(quality of service)를 보장한다. 인터넷은 대부분의 이용자를 위한 통신수단으로 자리매김하였으며 네트워크는 향후 유비쿼터스 시대로 발전하면서 유비쿼터스 컴퓨팅 환경이 구현된 사회는 인간의 생활을 편리하게 해주는 사회로 전환할 것이다. 현재와 같이 단순히 정보를 가공해 전해주는 것이 아니라 창조적으로 만들어 인간의 삶을 편리하게 해주어 미래사회의 중심적 역할을 할 것이다. 즉, 무선 네트워크의 발전으로 '장소와 시간을 초월한 컴퓨팅(anytime anywhere computing) 환경'이 가능하며, 유비쿼터스 사회를 추구하기 위하여 유선·무선·방송 등 정보통신 인프라 간의 구축이 종합적으로 추진된다[7].

인터넷 활용범위의 확대, 접속회선의 광대역화, 고속화 등으로 인터넷 트래픽 특성이 고용량 중심으로 전환되어 향후 트래픽 양의 폭주가 더욱 가속화될 것으로 예상된다. 급속한 인터넷 사용인구의 증가에 비해, 통신 인프라 구축의 상대적인 공급이 부족한 결과 인터넷 혼잡문제가 발생한다. 또한 인터넷 혼잡의 결과로 네트워크의 기동성이 떨어지게 되고, 데이터 손실과 지연 발생 등의 여러

* 강릉대학교 컴퓨터정보공학부 교수

가지 측면에서 부작용이 발생하게 되었다. 이와 같은 혼잡은 네트워크의 효율성을 저하시킬 뿐만 아니라 더 많은 통신 인프라 구축을 위한 예산이 필요하게 되는 결과를 낳게 되었다. 인터넷은 우리에게 필수적인 도구로서 핵심적인 활동공간이 되었으며, 이용자의 요구와 증가추세에 비해 관련 인프라 확장이 부족하여 인터넷 사용자의 다양한 욕구를 만족시키지 못하고 있다. 즉, TCP/IP 네트워크는 예측할 수 없이 늘어나는 인터넷 이용자의 증가로 인한 트래픽의 증가로 심각한 혼잡을 초래하고 있다. 이와 관련하여 네트워크 서비스의 품질 저하 문제는 심각한 수준을 넘고 있다. 그리고 웹 기반의 다양한 애플리케이션의 사용으로 발생하는 무분별한 서비스 혼잡(service congestion)은 핵심 업무에 대한 안정적인 서비스를 보장해 주지 못하고 있는 실정이다. 인터넷 활용도가 증가하면서 안정적이고 효과적인 인터넷 서비스 기반 구축에 대한 요구가 증가되고 있으며, 최적의 네트워크 운용과 관리를 위한 효율적인 트래픽 제어와 대역폭 관리가 매우 중요한 과제로 대두되고 있다[3][7][16].

다양한 애플리케이션들이 생성해 내는 다양한 유형의 트래픽들은 아직도 최대노력(best effort)의 단일 서비스 모델에 의해 서비스 되고 있으므로, 다양한 유형의 트래픽들을 그 특성에 맞게 처리함으로써 QoS를 보장하기 위한 많은 연구들이 진행되고 있다. 또한 인터넷이 발전되면서 현재 대부분의 정보가 인터넷에서 서비스 및 유통되는 형태로 바뀌어가고 있는 추세이다. 인터넷이 급속도로 성장함에 따라 네트워크의 효율적인 운용을 위한 연구 및 개발이 활발히 진행되고 있다. 그러므로 인터넷에서의 서비스 품질과 트래픽 용량을 검증하고, 트래픽이 발생하는 형태를 분석하여 네트워크 수행능력분석에서 다루기 쉬운 접근법을 제안할 필요가 있다. 또한 네트워크 트래픽 추세를 잘 반영하지 못하기 때문에 접속지연이 발생할 수 있으므로 트래픽 추세분석을 통해 수정된 통계적 모델을 제시할 필요가 있다.

인터넷 트래픽 측정 및 추세분석 기술은 지난 수십 년간 주로 미국을 중심으로 연구되어지고 있으며, 국내의 경우 가장 발전된 인터넷 네트워크 인프라를 구축하고 있음에도 불구하고 인터넷 트래픽 측정과 추세분석에 투자하는 노력이 상대적으로 약하지만 현재 몇몇 학자들에 의해 초기연구가 진행되는 상황이다. 웹 기반의 고품질 콘텐츠 서비스를 제공하기 위해서 인프라 차원에서 근본적으로 제공되어야 될 트래픽 측정 및 분석기술로 SLA(service level agreement) 모니터링, 네트워크 이상 징후 탐지, 대응 및 예방, 사용량 기반 인터넷 요금부과 등에 활용되는 매우 기본적인 사항이다[7][8].

이 논문에서는 트래픽을 표현하는 방법에 대해 조사하고 실질적으로 유용한 트래픽의 추세를 추정하는 모델을 구성한다. 또한 인터넷의 사용자가 증가함에 따라 사용자의 인터넷 작업부하 유형을 파악하고, 이를 수리적으로 표현하여 웹 사이트에서 송수신되는 트래픽의 형태를 모델링하고 적합한 통계적 모형을 제시할 필요가 있다. 논문에서 독립적이고 불규칙적인 트래픽 추세를 고려할 때 어떤 t 시점에서 임의의 외부적 개입(공휴일 및 시간주기, 통신정책변화(차등요금부과), 혼잡발생, ISP 업체의 전략변화 등)이 트래픽 발생에 어느 정도 반영되는지를 알아본다. 이는 인터넷의 수행능력, 서버의 용량, Proxy 용량, 네트워크 용량산정 등을 할 때 매우 중요한 자료로 이용될 수밖에 없다[1][3]. 일반적으로 인터넷 수행능력(performance)을 측정하는 도구로서 데이터를 어디서 구하고, 분석하고자 하는 수치(통계치)는 어떤 것, 자료를 구하는 간격, 데이터 추정방법을 이용한다. 2장에서는 웹의 트래픽 특성과 관련 연구에 대해 조사하고, 3장에서는 파레토와 와이블 분포를 이용한 변형된 통계적 모델을 제시한다. 4장에서의 모의실험 결과를 가지고 5장에서 결론을 제시한다.

2. 웹의 트래픽 특성

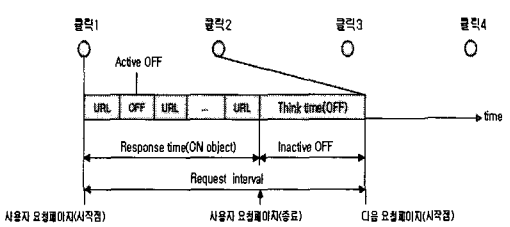
인터넷 트래픽 모델에 대한 연구는 Leland, Taqqu, Willinger, Wilson의 세미나 발표에서부터 유래되며, Paxson, Floyd는 LAN과 WAN 모두에서 획득되는 트래픽 추적이 LRD(long range dependence) 특성을 보여준다는 것을 제시하였다. 또한 On, Off 구간이 두꺼운 꼬리분포를 따라 생

성된다면 결과적으로 모여진 트래픽은 LRD 동작을 갖는 점근적인 자기 유사성을 보인다는 것을 증명하였다[9][13]. 무한분산 서비스 모델 M/G/∞ 대기는 Kurnz, Makowski 등에 의해 연구되어 졌다. 종합적인 트래픽을 생성하기 위해 Ridi, Gouse, Ribeiro, Baraniuk는 다양한 차원분열도형 모델을 이용하였다. Horvath, Telek은 척도불변을 이용하였으며, Robert, Bouder는 점근적인 LRD 정의를 이용하였다. Grosslauser, Bolot는 어떤 한계로 트래픽의 긴 구간 상관관계가 시스템의 수행능력에 영향을 줄 수 없음을 보였으며, 상관관계가 수행능력 연구에서 신뢰 결과를 유도하는 한계로 모델을 이용하는 방법을 제안하였다[1][6][14]. Crovella, Bestavros는 네트워크 트래픽의 자기유사성을 발생시키는 매카니즘을 웹 기록크기의 분포, 캐싱 효과와 파일전송의 사용자 선택, 이용자의 대기시간, 지역네트워크에서 전송되는 중첩(superposition)을 기반으로 전송시간과 대기시간의 합리성 검사를 하였다[5]. Muscariello, Mellia, Meo, Marsan은 종단 라우터에서 측정된 추적과 MMPP(Markovian modulated Poisson process) 모델에 의해 만들어진 트래픽의 대기 수행 능력 비교를 통해 트래픽 모델을 개발하였다. 이때 새션 도착안에서 패킷 흐름은 포아송 과정을 따르고, 새션안에서 흐름의 수와 흐름안에서 패킷의 수는 기하분포를 따른다. 5가지 모수에 의해 전송률과 도착률을 바탕으로 4단계 계층적 모델을 이용하였다. SDPP(θ) /(state dependent Poisson Process)의 성공 발생율 θ는 시스템 내부의 상태에 따라 달라진다. 예를 들어 M/M/s(고객도착과정이 PP(λ) /고객을 서비스 하는데 소요되는 시간들이 iid Exponential(μ) /server의 수)의 서비스 과정에서 시스템에 있는 고객수가 s 명 이하이면 $\Theta=n\mu$ 이고, 고객수가 s명 이상이면 $\Theta=s\mu$ 이다. 그러나 MMPP 의 성공 발생율은 시스템과 무관한 외부의 상태에 따라 달라진다는 것을 제시하였다[8][11].

웹을 사용하는 이용자의 작업부하 및 트래픽 특성을 고려할 때 두 가지 범주 즉, 이용자 동등성(user equivalent)과 분포 모델(distribution model)로 고려된다. BarFord와 Crovella는 경험적인 측면으로 볼 때 네트워크에서의 작업부하와 관련하여 사용되는 통계량으로는 서버의 파일크기분포, 요청크기분포, 관련된 파일 선호도, 내장된 파일 참조, 참조의 시간적 지역성, 개별 이용자의 유휴시간 등이 이용됨을 제시하였다. 또한 웹 객체에서 휴무시간을 대기시간(Off Time)과 활동중 대기시간(Active Off Time)의 두 종류를 제시하였다[3][16].

웹에서 작동되는 트래픽은 정상적으로 크기와 도착간격과 같은 매우 다양한 요구를 하며, 자기유사성을 가지는 비정상적인 특성을 가진다. 트래픽에서의 자기 유사성은 네트워크 수행에서 유의적인 부정적 충돌을 가질 수 있음을 지적할 수 있다. 이는 기존의 단일 통계분포를 이용하여 설명하기가 어려우며, 포아송 모델이 트래픽의 주요한 특성을 가지고 있는 것을 이용한다. 예를 들어 지수분포의 무기억성 특성을 나타내지 않는 두꺼운 꼬리 분포는 모아진 트래픽의 자기 유사성 동작을 유도할 수 있다. 자기 유사성의 성질을 이용하여 트래픽의 도착 유형의 모델을 세울 수 있으며, 이와 관련된 파레토 분포를 적용하는 기법을 연구할 필요가 있다.

일반적으로 이용자 동등성은 인터넷에서 이용자가 획득하려는 웹 파일정보를 위한 요청과 대기사이에 무한반복 되는 단일 공정으로 표현할 수 있다. 따라서 요청과 대기시간을 기초로 한 분포를 나타낼 수 있으며, 실제적인 웹 이용자의 특성이 있는 상관관계를 보여준다[12][13].



<그림 1> 작업부하에 따라 구분되는 요청구간 형태

<그림 1>을 참고하면 요청한 메시지가 모두 전송된 시점을 기준으로 요청시간, 대기시간으로 구분할 수 있으며, 요청시간에는 Active On과 Off로 나누어 설명된다. 반응시간(On Object)은 사용자의 요청시간부터 서비스가 완료되는 시간을 의미하며, Active Off 시간은 하나의 웹 객체의 컴포넌트 전송사이의 시간에 대응된다. 이것은 웹 브라우

저를 분석하는 웹 파일에 의해 소비되는 과정시간과 새로운 TCP 연결을 시작하기 위한 준비하는 시간에 대응된다. Inactive Off 구간은 사용자가 수신한 정보를 읽는 시간을 의미한다.

BarFord와 Crovella는 일반적 기법을 이용하여 웹 사이트에서 흐르는 트래픽 흐름은 일반적으로 6가지의 통계적 특성을 고려할 것을 제시하였다 [1][3].

[파일크기] 서버가 가지는 파일크기의 분포는 꼬리가 두꺼운(heavy tailed) 특성을 가지며, 서버에서 네트워크로 전송되는 파일의 크기가 매우 다양하기 때문에 서버의 파일 시스템이 여러 가지 파일 크기를 다루어야 함을 의미한다.

[요청크기] 경험적으로 볼 때 요청크기의 자료집합은 두꺼운 꼬리분포를 보이며 사용자가 서버에 요청한 크기와 파일크기의 설정사이의 차이는 서버에서의 파일시스템에 저장된 것과 서버로부터 네트워크로 전송된 것과의 분포에서 차이에 대응된다.

[선호도] 사용자에 의해 서버에 요청한 파일의 요청횟수를 의미하며, 파일에 대한 선호도는 지프(zipf) 법칙을 따른다. 즉, 최적의 선호도에서 선호도까지 순서가 된다면 파일의 참조수(p)는 순위(r)에 역비례 하는 경향이 있다는 것이다.

$$p=k \cdot r^{-1} \qquad (1)$$

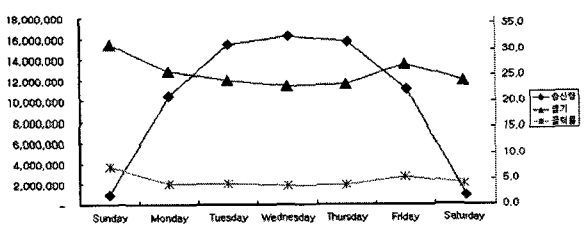
여기에서 k는 양의 정수를 의미한다. 참고로 대부분의 요청되는 파일들은 특별한 경우를 제외하고는 매우 작은 선호도를 갖고 있음을 가정하며 실제적으로도 그렇다.

[내장형참조] 웹 객체의 구조를 파악하기 위해 매우 중요한 개념으로서 하나의 페이지에 포함된 파일을 의미하며, 웹 객체에서 내장형 참조의 수를 특성화 하는데 이용된다. 일반적으로 내장형 참조의 수는 내장형 참조사이의 활동중대기시간이 비교적 짧다.

[시간적지역성] 시간적 지역성이 나타날 때 캐싱 효과가 유의적으로 증가하기 때문에 매우 중요하며, 참조된 기억장소의 일정부분이 그 이후에도 계속 참조될 가능성이 높음을 의미한다.

[대기시간] 웹 서버가 요청을 받아들이기 위한 필요한 조치를 할 때 사용자의 대기(사고)시간을 의미한다. 참고로 Inactive Off 시간의 정확한 모델은 개별적인 웹 사이트에서 요청의 폭주 특성을 얻기 위해 필요하며, Active Off 시간의 적절한 특성은 웹 객체의 전송을 복사하기위해 필요하다.

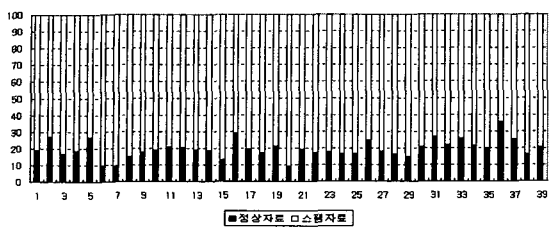
일반적으로 트래픽의 증가 및 감소량은 오전에 일과가 시작되는 시점에서 폭발적인 장비의 접속과 함께 트래픽 증가가 급속히 시작된다. 퇴근 시간이 지나 어느 정도 시간이 흐른 뒤에는 네트워크에 접속돼 있던 웹 객체가 급속하게 사라지기 시작하면서 트래픽 감소가 시작된다. 보편적인 방법으로 이메일 송수신을 통해 트래픽 증감 추세를 분석할 수 있다. 예를 들어 EROI(2005)의 e-Mail Marketing 분석 보고서[18]에 나타난 이메일 송신량, 열기 및 클릭율을 기본으로 서버에서 클라이언트로 이동되는 트래픽 양을 참조할 수 있다. 보고서에 의하면 이메일 송수신량이 가장 많을 때는 주중 3일이며, 사용자에 의해 열려지는 메일은 금요일과 일요일이 최다이며, 이러한 통계량은 고정된 추세가 아니라 시시각각 변할 수 있다는 것이다. 또한 전체적으로 볼 때 열기 및 클릭율은 비교적 고른 분포를 보이고 있다. 실질적으로 트래픽은 트래픽 분산의 유의성이 넓은 범위의 시간대에 나타난다.



<그림 2> 송신량, 열기, 클릭율의 주별 비교

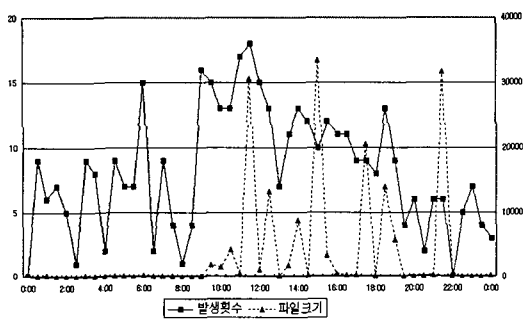
인터넷에서 송수신되는 트래픽 양의 추세는 외적요인에 매우 민감하게 작동되지만 전체적으로 다음 4가지로 구별할 수 있다. 즉, 작업부하가 한 시점으로 집중될 때, 작업부하가 서서히 증가할 때, 전반적으로 작업 부하량이 많을 때, 작업 부하량이 비교적 적을 때 등으로 나누어 볼 수 있다 [2][18]. 현실적으로 1등만이 존재하는 승자독식적인 특성이 강한 웹 사이트에서 대용량의 트래픽을 발생시키는 적은 소수의 웹 사이트와 극소량의 트

래픽을 발생시키거나 트래픽 발생자체가 거의 없는 대다수의 웹 사이트들이 존재한다[16][19].



<그림 3> 정상 및 스팸자료 비율(%) 분포

<그림3>과 <그림4>는 정보통신연구LAB의 메일서버에서 40일간 송수신된 자료를 분석한 것으로 여기에서도 파레토 법칙이 적용됨을 확인할 수 있다. 트래픽을 발생시키는 시점과 발생크기는 특정 시점에 집중되어 있음을 확인할 수 있다.



<그림 4> 시간대별 트래픽 발생횟수와 크기분포

일반적으로 인터넷에서 네트워크 효과를 통해서 규모가 큰 플레이어는 많은 신규 고객들을 블랙홀처럼 쉽게 끌어 들인다는 점, 원본이 완성된 후부터 저비용 고효율적인 생산비로 콘텐츠를 대량생산 할 수 있으며, 낮은 생산비조차 미리 확보된 거대한 기존 사용자 층에게 분산함으로써 강력한 원가 경쟁력을 갖게 된다는 점, 엄청난 대규모의 선행투자 비용에 있어서 후발 ISP 업체가 쫓아오는 것이 매우 힘들어 진다는 점, 대부분 반복학습을 통해서 습득된 네티즌에게 지식상품들은 한 번 습관화되면 계속해서 그 회사 제품을 쓰게 될 가능성이 커지게 된다는 고객학습효과 등[19]으로 경쟁력 있는 소수의 웹 사이트가 웹 트래픽 발생량의 80% 이상을 차지한다는 것이다.

3. 통계적 모델

네트워크를 관리하고 유지하기위해 가장 중요한 것은 네트워크 트래픽의 특징을 이해하는 것이다. 그러므로 현재 네트워크의 트래픽을 측정하고, 통계적인 모델을 세워 결과를 이용하여 트래픽의 특징과 자기유사성을 판단하는 것이 매우 주요하다. 고속 대용량 트래픽 측정시 모든 패킷을 수집하는 것은 기술적으로나 효율성 면에서 긍정적이지 않다는 이유로 통계적 혹은 기타 방법을 사용하여 전체 패킷 중 일부를 수집하는데 필요한 샘플링/필터링 방법 및 프로토콜을 표준화하고 있다. 웹 트래픽의 경우 파레토 모델로서, 폭주(burst) 기간의 편차가 매우 큰 것을 볼 수 있다. 통계적 모델을 개발할 때 일반적으로 단순성과 쉽게 이해될 수 있는 트래픽 측정으로 모수를 매핑 시키도록 적용시킨다. 따라서 BarFord와 Crovella가 제안한 것을 참고로[3] 6가지의 통계적 특성을 고려할 때 관련된 모델을 다음과 같이 제시할 수 있다.

사용자가 서버에 요청에서부터 사용자가 클라이언트로 모든 웹 객체를 가져올 때까지의 트래픽 전송의 전 과정에 트래픽 분포를 적합시키는 것은 일반적으로 좋은 방법이 아니다. 그러나 요청된 파일(트래픽)의 크기 및 내장된 참조, 대기시간은 오른쪽이 두꺼운 꼬리 분포를 갖으며, 로그정규분포보다는 파레토 분포를 적용시키는 것이 일반적인 방법이다. 인터넷 사용자들에게 공휴일 및 시간주기, 통신정책변화(차등요금부과), 혼잡발생 등과 같은 외부사건의 개입으로 인한 영향을 받아 트래픽 발생량에 일부 영향을 줄 수 있다. 따라서 두 가지 개입형태 즉, 충격함수(pulse function), 단계함수(step function)를 포함하는 수정된 통계량을 제안한다. 파레토 분포의 확률밀도함수(pdf)와 누적분포함수(cdf)는 다음과 같이 각각 표시할 수 있다[3][17].

$$f(x) = \frac{\alpha k^{\alpha}}{x^{\alpha+1}}, \quad x > 0, k > 0 \tag{2}$$

$$\mathrm{F}(x) = 1 - (\frac{k}{x})^{\alpha} \tag{3}$$

여기에서 $\alpha > 0$는 안정지수(stability index)이

며, $\alpha < 1$ 일 때 무한 평균과 분산을 가진다. $1 \leq \alpha < 2$ 일 때 무한분산만을 갖으며, 트래픽의 유사성 특성을 보인다. 또한 평균과 분산을 다음과 같이 각각 표현할 수 있다[3].

$$E(X) = \frac{\alpha}{\alpha - 1} \cdot k \ , \ \alpha > 1$$

$$V(X) = \frac{\alpha k^2}{(\alpha-1)^2(\alpha-2)} \ , \ \alpha > 2 \qquad (4)$$

파레토 분포의 CDF를 참고하고, 균등분포로부터 얻은 확률변수를 이용하여 파레토 분포의 확률변량(random variates)은 (6)식을 이용하여 구할 수 있다.

$$X_p = \frac{k}{(1-U)^{1/\alpha}} \ , \quad U \sim U(0,1) \qquad (5)$$

$$X_{pareto} = B_o + B_1 X_{p_{(t-1)}}(\alpha, k) + B_2 I_t^{(T)} + \eta_t \qquad (6)$$

여기에서 B_0와 B_2는 임의의 상수이다. B_1은 시점 t에서 주어지는 후진연산자이며, η_t는 백색잡음으로서 $N(0, \sigma^2)$을 따르며, 여기에서는 무시한다. T는 개입시점을 나타내며, $I_t^{(T)}$는 개입변수로서 단계함수를 적용할 경우 다음 (7)식을 이용한다.

$$I_t^{(T)} = \begin{cases} 0, & t < T \\ 1, & t \geq T \end{cases} \qquad (7)$$

충격함수를 적용할 경우 다음 (8)식을 이용한다.

$$I_t^{(T)} = \begin{cases} 0, & t \neq T \\ 1, & t = T \end{cases} \qquad (8)$$

파레토 분포의 특성을 이용하여 트래픽의 파일크기, 요청크기, 대기시간(Inactive Off), 내장형 참조 등을 표현할 수 있다[3][12].

지수분포의 변형으로 임의성이 다소 낮은 자료에 적합한 것으로 와이블 분포를 들 수 있다. Active Off 시간의 분포는 지수분포에서 변형된 와이블 분포를 갖으며 확률밀도함수와 누적분포함수는 다음과 같이 각각 표시할 수 있다.

$$f(x) = \frac{bx^{b-1}}{a^b} e^{-(x/a)^b} \qquad (9)$$

$$F(x) = 1 - e^{-(x/a)^b} \qquad (10)$$

여기에서 $x > 0, a > 0, b > 0$ 이다. 와이블 분포의 CDF를 참고하고, (5)와 (6)식과 비슷한 방법으로 균등분포로부터 얻은 확률변수를 이용하여 와이블 분포의 확률변량은 (12)식을 이용하여 구할 수 있다.

$$X_w = a[-\ln(1-U)]^{1/b} \ , \ U \sim U(0,1) \qquad (11)$$

$$X_{weibull} = B_o + B_1 X_{w_{(t-1)}}(a,b) + B_2 I_t^{(T)} + \eta_t \qquad (12)$$

와이블 분포를 이용하여 활동 중 대기시간 등을 표현할 수 있다. 여기에서 a는 형태모수, b는 위치모수를 나타낸다.

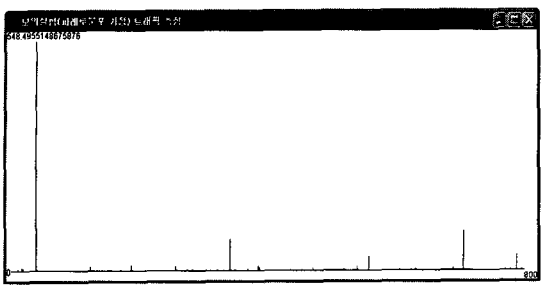
4. 모의실험

이 논문에서는 JAVA를 이용하여 모의실험 프로그램을 구성하였다. 트래픽이 발생되는 시간 구간(time interval)은 지수분포에 의해 생성된 확률변량을 이용하여 결정한다고 가정하고 1단위 시간*확률변량 간격으로 확률변량을 발생시키도록 하였다. 모수 B_0, B_1, B_2는 임의의 상수로서 Box-Jenkins 기법을 이용한 AR(1) 모형을 통하여 구한다. 5,000번의 모의실험을 통해 개입을 고려하지 않은 것에 비해 개입이 주어졌을 경우 트래픽 발생량(크기)의 변동 폭을 다음과 같이 확인하였다.

개입형태	
단계함수	충격함수
±8.367%	±6.316%

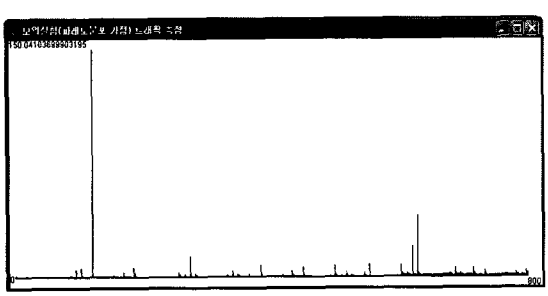
수정된 트래픽 발생모델에 개입(단계함수)이 주어질 때 파레토 분포를 고려한 상태에서 ±8.367%의 변동 폭을 확인할 수 있다. 충격함수보다 단계함수일 경우 좀 더 많은 개입효과가 있음을 확인하였다.

비슷한 방법으로 B_0=2.1095, B_1=0.0703, B_2=-0.000067, a=0.31, b=1.0, 외부(단계함수) 개입이 있고 와이블 분포를 고려할 경우 (12)식을 적용한다면 ±2.776%의 활동 중 대기시간변동 폭을 확인할 수 있다.

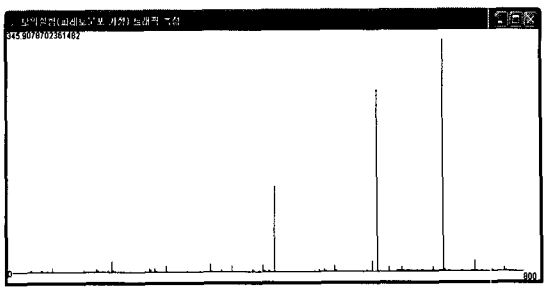


<그림 5> 제안된 식 (6)을 이용하여 나타낸 트래픽 양(α=1.9, k=1.0, $I_t^{(T)}$=0.0)

<그림 5>는 B_0=10.394, B_1=0.0407, B_2=-0.90308, α=1.9, k=1.0, 개입이 없을 경우 트래픽 양을 나타낸 것이다. <그림 6>은 임의의 시간 T에서 외부(단계함수)개입이 존재하고 동일한 조건에서 트래픽 양을 나타낸 것이다. 개입이 주어진 시점($I_t^{(T=600)}$)부터 트래픽 변화가 나타남을 확인할 수 있다. <그림 7>은 외부(충격함수)개입이 존재할 경우 동일한 조건에서 트래픽 양을 나타낸 것이다. 개입이 주어진 시점부터 일정부분($I_t^{(T=600)}$)에 걸쳐 트래픽 증감이 나타남을 확인할 수 있다.



<그림 6> 제안된 식 (6)을 이용하여 나타낸 트래픽 양(α=1.9, k=1.0, $I_t^{(T=600)}$ 개입(단계함수) 적용)



<그림 7> 제안된 식 (6)을 이용하여 나타낸 트래픽 양(α=1.9, k=1.0, $I_t^{(T=600)}$ 개입(충격함수) 적용)

외부 사건의 개입으로 인하여 트래픽 증감의 영향이 나타나는 것을 볼 때 통신정책 변화 등으로 혼잡을 어느 정도 줄일 수 있음을 구체적인 수치값으로 확인할 수 있다.

5. 결 론

웹 트래픽의 지속적인 추적 및 분석과 그에 따른 콘텐츠 구성의 관리는 매우 중요하다. 트래픽 모델은 실제 트래픽 흐름의 동작을 예측하는데 이용되는 중요한 확률과정으로서 가장 이상적인 트래픽 모델은 근원적인 트래픽의 적절한 수정된 통계적 성질을 이용하여 나타낼 수 있음을 보였다. 즉, 인터넷 트래픽의 단순함과 정확한 모델을 구별하기가 쉽지 않다는 현실적인 제약임에도 불구하고 외부적 사건개입으로 트래픽 증가를 줄일 수 있는 가능성을 수정된 트래픽 발생 모델을 통해 확인할 수 있다. 파레토 분포와 와이블 분포를 동시에 적용하여 개입을 주는 혼합모델은 향후 연구되어져야 할 부분이다.

참 고 문 헌

[1] G. R. Addie, T. D. Neame and M. Zukerman (2002), "Performance Evaluation of a Queue Fed by a Poisson Pareto Burst Process", Computer Networks, Vol 40, No. 3, pp. 377-397.

[2] A. T. Andersen and B. F. Nielsen(1998), "Markovian Approach for Modeling Packet Traffic with Long-Range Dependence", IEEE Journal on Selected Areas in Communications, Vol. 16, No. 5.

[3] P. Barford and M. Crovella (1998), "Generating Representative Web Workloads for Network and Server Performance Evaluation", Proceeding of ACM SIGMETRICS conference '98.

[4] E. P. Box, George, G. M. Jenkins and G. C. Reinsel(1994), "Time Series Analysis, Forecasting and Control", 3rd Ed. Prentice Hall.

[5] M. E. Crovella and A. Bestavros(1995), "Explaining World Wide Web Traffic Self-Similarity", Technical Report TR-95-015 Boston University.

[6] M. Grosslauser, and J. Bolot(1999), "On the Relevance of Long Range Dependencies in Networking Traffic", IEEE/ACM Transactions on Networking, Vol. 7, No. 5, pp. 629-640.

[7] S. S. Ji(2006), "Congestion Pricing Function of Internet Differentiated Services for Social Benefit", KIISC, Vol. 11, No. 2, pp. 9-17.

[8] H. Joel Trussell, A. Nilsson, Mo-Yuen Chow and C. Trivedi(2005), "A Statistical Method for Classification of Internet Traffic Flows", CACC, North Carolina State University.

[9] W. E. Leland, W. Willinger, M. S. Taqqu and D. V. Wilson(1994), "On the Self-similar Nature of Ethernet Traffic(Extended Version)", IEEE/ACM Transactions on Networking, Vol. 2, pp. 1-15.

[10] A. I. McLeod and E. R. Vingilis(2005), "Power Computations for Intervention Analysis" Technometrics, Vol. 47, No 2, pp. 174-180.

[11] L. Muscariello, M. Mellia, M. Meo and M. Ajmone Marsan(2004), "An MMPP-Based Hierarchical Model of Internet Traffic", http://www.tlc-networks.polito.it/muscariello/papers, Torino, Italy.

[12] António Nogueira, Paulo Salvador, Rui T. Valadas and António Pacheco(2003), "Modeling Self-similar Traffic through Markov Modulated Poisson Processes over Multiple Time Scales", High Speed Networks and Multimedia Communications.

[13] V. Paxson and S. Floyd(1995), "Wide-Area Traffic: The Failure of Poisson Modeling", IEEE/ACM Transactions on Networking.

[14] R. H. Riedi, M. S. Crouse, V. J. Ribeiro and R. G. Baraniu(1998), "A Multifractal Wavelet Model with Application to Network Traffic", IEEE Transactions on Information Theory, Vol. 45, No. 4, pp. 992-1018.

[15] N. Vlajic(2003), "JAVA Socket Programming", University of Ottawa, USA.

[16] M. Zukerman, T. Neame and R. G. Addie(2003), "Internet Traffic Modeling and Future Technology Implication", Proc. of IEEE InfoCom 2003, S. Francisco, CA, USA.

[17] "On generating self-similar traffic using pseudo-Pareto distribution", [Online] Available http://wwwcsif.cs.ucdavis.edu

[18] Inc eROI(E-mail return-on-investment), "Email Statistics by List Size", [Online] Available http://www.eroi.com

[19] "Internet Traffic Reports by Opnix", [Online] Available http://www.internettrafficreport.com/

지 선 수 (Seon-Su Ji)

- 정회원
- 1984년 충남대학교 계산통계학과(학사)
- 1986년 중앙대학교 응용통계학과(석사)
- 1993년 중앙대학교 응용통계학과(박사)
- 2006년 명지대학교 컴퓨터공학과(박사수료)
- 원주대학 컴퓨터정보관리과 교수
- (현) 강릉대학교 컴퓨터정보공학부 교수
- 관심분야 : 혼잡제어, 정보보안(암호학), 이미지 프로세싱