제 19 회 한국정보처리학회 춘계학술발표대회 논문집 제 10 권 제 1 호 (2003. 5)

통계적 분석을 통한 HTTP 트래픽 모델링 및 분석*

전의수, 김태수, 이광휘
창원대학교 컴퓨터공학과
e-mail : fraser@ce.changwon.ac.kr, {jupi, khlee}@sarim.changwon.ac.kr

HTTP Traffic Modeling and Analysis with Statistical Process

Uie-Soo Jeon, Tae-Soo Kim, Kwang-Hui Lee
Dept. of Computer Engineering, Changwon National University

요 약

통신망을 효율적으로 설계하고 운영하기 위하여 통신망에 대한 구체적인 시뮬레이션이 필요하며 이에 관한 연구가 현재 활발히 이루어지고 있다. 본 논문에서는 통신망 성능 분석을 위한 시뮬레이션 시 필요한 트래픽 생성기의 설계를 위해 실제 트래픽 자료를 수집, 분석하여 HTTP 요구 수준에서 통계적 방법을 통해 확률 분포로 모델링하였다. 기존 연구에서는 응답 크기에 대하여 파레토 분포만을 사용하여 그 특성을 모델링하였지만, 본 연구에서는 지수 분포와 파레토 분포의 혼합으로 모델링할 수 있음을 확인하였다. 또한 응답 크기의 특성은 서버 내 파일 크기의 특성을 그대로 반영하는 것이 아니라 사용자의 웹 문서 요청의 편중화 현상에 영향을 받아 그 특성이 달라질 수 있다는 것을 분석을 통해 확인하였다.

1. 서론

인터넷 사용의 급격한 증가와 다양한 서비스의 출현으로 인하여 데이터 트래픽의 양상이 점점 복잡해지고 있다. 따라서 기존의 경험적 지식만을 이용하여 통신망의 사용량이나 이용자 패턴을 조사하여 통신망을 설계하고 운영하는 것에는 한계가 있다. 이러한 관점에서 현재 트래픽 흐름에 대한 통계적 분석과 더불어 시뮬레이션과 관련된 연구가 활발히 진행중이다.

인터넷 트래픽에는 여러 종류의 응용 서비스에 의해 발생하는 트래픽이 있지만 본 논문에서는 웹 트래픽만을 대상으로 한다.

기존 연구에 따르면, LAN 및 WAN 환경에서 웹 트래픽의 특성이 자기유사(self-similar) 하다는 주장[1,2]도 있고, 이와 상반되게 포아송 과정(poisson process)을 따른다는 주장[3]도 있다. 그러나 웹 트래픽이 어떠한 분포를 따르더라도 대상 네트워크의 특성을 반영하는 파라미터를 결정하기는 쉽지가 않다.

또한 웹 트래픽은 혼잡 제어 등의 영향을 받는 TCP 패킷의 흐름을 기록하여 패킷 수준에서 모델링 될 수 있고, 웹 파일의 크기 분포 등을 이용하여 응용 수준에서 모델링 될 수 있다.

패킷 수준 모델링은 링크상의 흐르는 패킷을 TCPdump[4] 등을 이용하여 캡쳐하고 그 흐름을 분석하여 트래픽의 특성을 모델링하는 것이다. 그러나 패킷간의 시간 간격은 TCP 의 흐름 제어 및 혼잡 제어 알고리즘에 영향을 받는다. 이렇게 수집된 트래픽 특성은 해당 트래픽 수집 시 링크상에 흐르는 배경 트래픽에 따라 TCP 알고리즘의 영향을 받아 매 측정 때마다 그 특성이 달라질 수 있다. 뿐만 아니라 수집된 방대한 양의 패킷 관련 정보들을 조합하고 분석하는데 그 어려움이 따른다[5].

응용 수준 모델링은 패킷 수준 모델링의 단점을 극복하고자 하는 것으로 각 연결이 주고 받는 데이터를 웹 파일 크기 등을 이용하여 언제, 얼마의 크기를 주고 받았는지를 모델링 함으로써 TCP 혼잡 제어 및 흐름 제어 알고리즘에 영향을 받지 않는다. 즉, 패킷들의 도착 시간 간격은 배경 트래픽에 영향을 받아 달라지더라도 응용 수준에서 언제 얼마의 크기가 전송되었는지는 변하지 않는다. [5]에서는 웹 어플리케이션의 성능분석을 위한 트래픽 로드 생성을 위해 응용 수준에서 경험적 트래픽 모델을 제안했으며, 다양한 관점에서 웹 트래픽의 특성을 모델링 했다. 그러나 응용 수준에서 사용자의 웹 액세스 패턴을 모델링함으로써 모델링해야 할 항목들이 많아지고 모델링된 트래픽 로드를 생성시키기 위한 트래픽 생성기 역시 복잡해진다는 단점이 있다. 또한 [6]에서는 웹서버 로그를 이용하는 기법을 제안하고 있다. 이 방법은 자료 수집을 위한 별도의 작업이 필요 없다는 장점이

* 본 논문은 산학협동재단의 2002 년도 학술연구비지원사업으로 수행된 결과의 일부임.

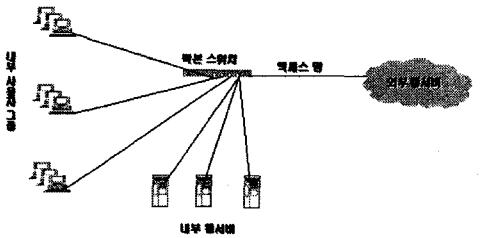
있으나 사용자가 여러 웹 브라우저를 띄워 동시에 다른 서버에 접속 시에는 사용자의 웹 액세스 패턴을 알기가 어렵다. [1]에서는 클라이언트 로그를 이용하는 방법을 제안하고 있다. 다양한 형태의 정보들을 클라이언트 측에 기록되게 할 수는 있으나 이를 지원하기 위해서는 브라우저의 수정 등 부과적인 작업이 필요하다.

따라서 대상 네트워크의 특성을 반영하는 파라미터를 결정할 수 있도록 실제 수집된 트래픽 자료를 이용하여 HTTP 요구 수준에서 웹 트래픽을 모델링한다. 이를 위해 본 논문에서는 서버 로그 분석 기법과 트레이스 분석 기법을 함께 이용한다.

논문의 구성은 다음과 같다. 2 장에서는 실험에 사용된 네트워크 구성을 설명하고, 3 장에서는 웹 트래픽을 모델링하기 위해 각 항목별 분석과 모델링에 관해 다룬다. 그리고 4 장에서는 웹 트래픽의 전반적인 특성을 분석하고, 마지막으로 결론과 향후 연구에 대해서 언급한다.

2. 분석 네트워크 구성

(그림 1)은 실험에 사용된 네트워크의 구성도이다. 네트워크 내 총 컴퓨터 수는 약 300 대 정도이고 그 이용 인원수는 약 700 명 정도로 추산되며, 약 10 대의 웹서버가 현재 운영 중에 있다.



(그림 1) 네트워크의 구성도

내부 사용자와 내부 웹서버 사이의 웹 트래픽, 외부 사용자와 내부 웹서버 사이의 웹 트래픽은 웹 로그를 이용하여 분석할 수 있으나, 내부 사용자와 외부 웹서버 사이의 웹 트래픽은 웹 로그를 이용하여 분석할 수 없으므로 패킷 트레이스를 이용하여 분석하였다. 대상 네트워크 내 모든 웹서버들의 서버 로그를 구하는데 따른 어려움으로 인해 분석 대상을 3 대의 웹서버로 제한하였다(표 1).

(표 1) 서버의 로그 자료 수집

웹서버	기간	요청횟수	총 응답크기(MB)
CE	59 일	350,368	2,193
CSL	78 일	570,040	20,980
HiBrain.Net	7 일	2,599,306	20,027

(표 2) 패킷 트레이스 자료

수집 위치	SUMMIT 48 Switch
기간	24 시간
요청 횟수	591,709
총 응답 크기(MB)	7,517

(표 2)는 패킷 단위로 수집된 데이터를 요청 패킷과 응답 패킷으로 구분하고 단일 요청 및 응답으로 생성된 패킷들은 하나로 묶어서 크기를 산출하여 트래픽 분석에 이용하였다. 수집 기간은 패킷 트레이스의 크기를 고려하여 하루 동안의 데이터로 제한하였다.

3. 웹 트래픽 모델링

3.1 HTTP 요구 수준 모델링

HTTP 는 각 연결간의 데이터 전송을 위해 클라이언트의 요청이 서버로 보내지고 이에 상응하는 응답을 클라이언트에게 보내는 단순 요청-응답 메커니즘을 사용한다. 본 논문에서는 이러한 HTTP 의 데이터 전송 메커니즘에 초점을 맞추어 HTTP 요구 수준에서 트래픽을 모델링하였다.

(표 3) HTTP 요구 수준에서의 모델링 항목

항목	단위	설명
요청간 시간 간격	seconds, micro seconds	모든 요청간의 시간 간격
요청 크기	bytes	HTTP 요구 길이
응답 크기	bytes	HTTP 응답 길이

(표 3)은 HTTP 요구 수준에서 모델링 될 항목들이다. 요청간의 시간 간격 항목은 모든 사용자 요청간의 시간 간격을 나타내며, 서버 로그의 경우 초 단위이고 패킷 트레이스인 경우 마이크로초 단위이다. 요청 크기 항목은 클라이언트측에서 서버로의 요청 정보를 보내기 위한 데이터의 크기 항목이고 응답 크기는 서버측에서 클라이언트의 요청에 대한 응답으로 보내는 데이터의 크기 항목이다.

3.2 요청간의 시간 간격 모델링

본 항목의 모델링은 각 내부 웹서버들에 요청된 시간 간격과 내부 사용자들에 의해 액세스된 모든 외부 웹서버들에 대한 요청간의 시간 간격을 대상으로 하였다. 특히 전자의 경우, 내부 사용자와 외부 사용자 그룹으로 나눠 그 특성을 모델링하였다.

(표 4) 내부 사용자의 요청간 시간 간격에 대한 분포 파라미터 추정치

웹서버	적용비율	확률 분포	파라미터	R^2
HiBrain.Net	73%	지수	μ = 0.22	0.83
	23%	파레토	a = 0.99, k = 1.2	0.96
CSL	74%	지수	μ = 0.22	0.87
	26%	파레토	a = 0.77, k = 3.34	0.93
CE	72%	지수	μ = 0.99	0.92
	28%	파레토	a = 0.96, k = 2.37	0.98
외부 웹서버	39%	지수	μ = 0.0035	0.95
	61%	파레토	a = 0.98, k = 0.01	0.96

(표 5) 외부 사용자의 요청간 시간 간격에 대한 분포 파라미터 추정치

웹서버	적용비율	확률 분포	파라미터	R^2
HiBrain.Net	90%	지수	μ = 0.21	0.70
	10%	파레토	a = 2.95, k = 0.47	0.97
CSL	84%	지수	μ = 1.2	0.94
	16%	파레토	a = 1.27, k = 7.6	0.92

CE	76%	지수	μ = 0.42	0.91
	24%	파레토	α = 1.0, k = 1.8	0.96

(표 4,5)는 사용자 요청간의 시간 간격을 확률 분포로 모델링한 것이다. 내부 및 외부 사용자에 의한 3 대의 내부 웹서버들과 내부 사용자에 의한 모든 외부 웹서버들을 대상으로 한 요청 시간 간격이 특정한 단일 분포로는 모델링되지 않았고 두 분포를 사용하여 전 구간을 모델링할 수 있었다. 그 결과, 하위 부분은 지수 분포로, 상위 부분은 파레토 분포로 모델링이 되었으며, 외부 웹서버들을 대상으로 한 항목을 제외한 나머지 모든 항목에서 파레토 분포를 따르는 부분은 총 분포 구간 중 상위 약 10% ~ 28%로 작은 부분을 차지했다. R^2 은 실제 자료들이 얼마나 이론적인 분포에 적합한지를 나타내는 결정 계수로서 0 과 1 사이의 값을 가지며, 1 에 가까울수록 적합하다고 할 수 있다. 이 수치가 HiBrain.Net 의 외부 사용자에 대한 요청 간격 항목의 하위 지수 분포에 대한 부분을 제외한 모든 부분에서 0.8 이상으로 높게 나왔기 때문에 조사된 대부분 서버에서의 요청간의 시간 간격은 지수 분포와 파레토 분포의 혼합으로 잘 모델링된다고 할 수 있다.

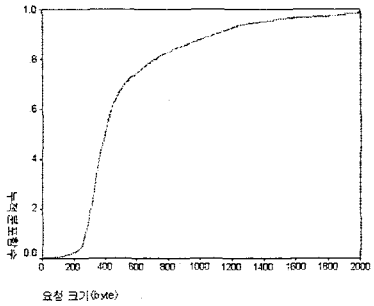
3.3 요청 크기 모델링

클라이언트가 웹서버로부터 데이터를 받기 위해 데이터 요청을 해당 웹서버로 보내게 된다. 이러한 요청을 위한 패킷 내의 정보로는 데이터 전송을 위해 사용되는 해당 프로토콜간의 통신을 위한 정보가 포함되며, 그 크기는 프로토콜별로 고정적이다. 또한 해당 클라이언트가 웹서버 내에 있는 웹 문서 등을 요청하는 경우 요청된 데이터의 웹서버 내 위치를 나타내는 정보가 있다.

그러나 클라이언트가 항상 웹서버로 요청 정보만을 보내는 것이 아니라 게시판에 글을 작성하거나 파일을 첨부하여 올릴 수도 있다. 이렇게 클라이언트 측에서 웹서버로의 업로드(upload)가 발생하면 패킷들의 크기가 자료에 따라 아주 커질 수 있다. 이러한 클라이언트에 의해 웹서버로의 업로드는 데이터 등의 요청은 아니지만 본 연구에서는 요청 크기 항목에 포함하여 모델링하였다.

모든 외부 웹서버에 대한 사용자의 웹 트래픽을 모델링하기 위해 사용하는 패킷 트레이스 자료를 가지고 요청 크기를 모델링하였다. 수집한 패킷 트레이스 자료에 따르면 요청 크기가 전체 웹 트래픽에서 차지하는 비중이 5% 이하로 작은 비중을 차지하는 것으로 나타났다. 또한 그 특성이 특정 사용자 집단에 따라 독립적이지 않고 어느 정도 일반적인 특성이 있는 것으로 보이기 때문에 개별 웹서버를 대상으로 한 웹 트래픽 생성 부분에서 사용하여도 무방할 것이다.

(그림 2)는 요청 크기에 대한 누적분포함수 그래프를 나타낸 것이다. 약 1,200 바이트 이하의 요청 크기가 대부분을 차지하며, 그 이상에서 그래프가 긴 꼬리를 가지며 계속 이어지는 것을 볼 수 있다. 이러한 특성은 클라이언트 측에서 웹서버로의 업로드에 의해 발생하는 트래픽의 부분으로 짐작된다. 그리고 약 250 바이트 에서 약 500 바이트 사이에서 누적 확률값이 급격히 상승하는 것을 보아 이 구간에 많은 요청이 집중되어 있음을 알 수 있다. 그러나 TCPdump 를 이용한 패킷 트레이스 자료에서는 프로토콜 헤더에 관한 정보 밖에 볼 수 없기 때문에 정확한 원인을 알 수 없으며, 원인 분석을 위해서는 추가적인 정보 수집이 필요하다.



(그림 2) 요청 크기에 대한 누적분포함수 그래프

(표 6)은 요청 크기를 확률 분포를 사용하여 모델링한 것이며, 그 결과 로그노말 분포로 가장 적합하게 모델링되었다. 비록 적합도를 나타내는 결정계수의 수치가 0.76 으로 낮게 산출되었지만 요청 크기가 전체 웹 트래픽에 미치는 영향이 다른 항목들에 비해 미비하다. 그러므로 시뮬레이션 시 트래픽 로드를 생성하기 위해 이 분포를 사용하여도 무방하다고 생각한다.

(표 6) 요청 크기의 분포 파라미터 추정치

확률 분포	파라미터	R^2
로그 노말	μ = 321, σ = 322	0.76

3.4 응답 크기 모델링

웹 트래픽의 응답 크기에는 HTML 문서 및 멀티미디어 파일들과 같은 웹서버 내에서 크기가 정해져 있는 정적인 파일들과 CGI 로 구성된 게시판 액세스 등으로 인해 발생하는 동적인 자료의 전송에 의한 트래픽도 포함된다.

응답 크기의 통계적 성질은 웹서버 내의 파일들의 특성을 따르며[1], 사용자 그룹에 따라 그 특성이 다르지 않을 것으로 가정하여 본 항목은 사용자들의 요청에 대한 응답 크기를 내부 및 외부 사용자의 요청에 의한 응답 크기로 나누지 않고 웹서버별로 응답 크기를 모델링하였다. 분석되는 웹서버들은 먼저 대상 네트워크 내에 존재하는 3 대의 웹서버들과 내부 사용자들의 요청에 의한 외부 웹서버들의 총체적인 응답 크기를 외부 웹서버 항목으로 정의하여 총 4 가지로 나누었다.

(표 7)은 웹서버들에 대한 응답 크기의 특성을 확률 분포를 사용하여 모델링한 결과를 나타낸 것이다. CSL 과 CE 그리고 외부 웹서버들의 응답 크기는 하위 부분은 지수 분포로 그리고 상위 부분은 파레토 분포

로 잘 모델링되었다. 그러나 HiBrain.Net 의 경우 본 연구에서 조사된 다른 서버들과 [1,5]과는 달리 파레토 분포로는 모델링이 되지 않았고 전 구간에서 지수 분포로 더욱더 적절히 모델링되었다.

(표 7) 응답 크기의 분포 파라미터 추정치

웹서버	적용비율	확률분포	파라미터	R^2
HiBrain.Net	100%	지수	μ = 15,621	0.98
CSL	30%	지수	μ = 676	0.85
	70%	파레토	a = 0.94, k = 2,117	0.94
CE	53%	지수	μ = 526	0.95
	47%	파레토	a = 1.53, k = 2,903	0.82
외부 웹서버	70%	지수	μ = 406	0.97
	30%	파레토	a = 1.13, k = 1,800	0.99

4. 웹 트래픽 특성 분석

액세스 망에서의 웹 트래픽 특성을 좌우하는 응답 크기의 특성이 서버 내 파일들의 크기 특성에서 기인한다고 주장하였다[1].

(표 8) 서버 내 파일 크기들의 분포 파라미터 추정치

웹서버	확률 분포	파라미터	R^2
CSL	파레토	a = 1.09	0.98
CE	파레토	a = 0.81	0.96
HiBrain.Net	정규	μ = 39,872, σ = 10,486	0.99

(표 8)은 서버 내의 파일 크기들을 확률 분포를 사용하여 가장 적합한 분포와 파라미터를 산출한 것이다. 각 서버 내 파일 크기들의 분포가 CSL 과 CE 의 경우, 응답 크기에서와 같이 파레토 분포로 잘 모델링이 되었다. 하지만 HiBrain.Net 의 경우 응답 크기에서와 같이 지수 분포로 모델링되지 않고 형태가 완전히 다른 정규 분포로 모델링되었다. 이렇게 HiBrain.Net 의 경우처럼 응답 크기와 서버 내 파일 크기의 특성이 다르다는 것은 사용자의 파일에 대한 요청이 특정 파일들에 편중적으로 이루어진다는 것을 알 수 있다.

이러한 분석을 통해 CSL 과 CE 에서는 응답 크기의 특성과 서버 내 파일 크기의 특성이 일치함으로써 사용자의 웹 파일 액세스 패턴에 의해 응답 크기의 특성이 영향을 받지 않지만, HiBrain.Net 의 경우 직접적으로 그 영향을 받는 것으로 보아 서버 내 파일 크기와 응답 크기의 특성이 서버에 따라서는 다르게 나타날 수 있다는 것을 알 수 있다.

HiBrain.Net 서버 내 파일 크기의 성질이 CSL 과 CE 와는 달리 자기유사한 특성이 없는 정규 분포와 같이 모델링이 되는 이유를 살펴보면, 3 대의 웹서버에서 주로 게시판을 사용하여 사이트를 구성하였지만 CSL 과 CE 서버에서는 게시판에 사용자들이 첨부하여 업로드할 수 있는 파일의 최대 크기를 제한하지 않는 반면, HiBrain.Net 의 경우 최대 크기를 5MB 로 제한하였다. 그리고 또한 HiBrain.Net 서버의 경우 멀티미디어 관련 파일 등과 같이 큰 파일들이 거의 존재하지 않음으로써 파일 크기들의 범위가 작고 이로 인해 분산 또한 작다. 결국 이러한 원인에 의해 서버 내 파일 크기의 분포가 자기유사한 특성이 없는 정규 분포로 모델링되고, 이러한 특성이 사용자의 웹 파일 액세스 패턴에 영향을 받아 응답 크기 항목이 지수 분포의 특성을 가지게 됨을 알 수 있다.

5. 결론 및 향후연구

본 논문에서는 통신망 성능 분석을 위한 시뮬레이션 시 필요한 트래픽 생성기의 설계를 위해 실제 트래픽 자료를 수집, 분석하여 HTTP 요구 수준에서 통계적 방법을 통해 확률 분포로 모델링하였다.

응답 크기는 전체적인 웹 트래픽 특성에 미치는 영향이 큰 만큼 정확한 분석을 통해 그 특성을 파악하는 것이 중요하다. 기존 연구에서는 단일분포(파레토 분포)만을 사용하여 그 특성을 모델링하였지만, 본 연구에서는 파레토 분포를 따르지 않는 하위 부분은 지수 분포로 적합하게 모델링이 된다는 것을 분석을 통해 알 수 있었다. 그러므로 응답 크기는 지수 분포와 파레토 분포의 혼합으로 적합하게 트래픽 로드를 생성할 수 있다. 또한 응답 크기의 특성은 서버 내 파일 크기의 특성을 그대로 반영하는 것이 아니라 사용자의 웹 문서 요청의 편중화 현상에 영향을 받아 그 특성이 달라질 수 있다는 것을 분석을 통해 알 수 있었다.

향후 연구로는 본 연구에서와 같이 웹 트래픽이 지수 분포와 파레토 분포의 혼합으로 생성되지만, 두 분포의 비율 및 파라미터 값은 웹서버에 따라 다름으로 인하여 이러한 파라미터를 빠르고 정확하게 파악할 수 있는 자동화 도구의 개발이 필요하며, 또한 네트워크상의 전체적인 트래픽 로드를 생성하기 위해 웹 트래픽 외 다른 응용의 트래픽들의 특성 분석이 요구된다.

참고문헌

[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. 835-840, Dec. 1997.

[2] S. Uhlig and O. Bonaventure, " Understanding the Long-Term Self-similarity of Internet Traffic" , QOFIS2001, Portugal, pp. 286□ 298, Sep. 2001.

[3] Jin Cao et al., " Internet Traffic Tends To Poisson and Independent as the Load Increases" , Bell Labs. Technical Report, Murray Hill, 2001.

[4] Lawrence Berkeley National Laboratory, Berkley *Packet Filtering Tool*, *http://www.tcpdump.org/*

[5] Bruce A. Mah, " An Empirical Model of HTTP Network Traffic" , in Proceedings of IEEE INFOCOM, Japan, pp. 592-600, Apr. 1997.

[6] Martin F. Arlitt and Carey L. Williamson, " Web Server Workload Characterization: The Search for Invariants" , in Proceedings of the ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems, pp. 126□ 137, May 1996.