TTT 타점법을 이용한 웹서버 파일 分布의 厚尾性 분석

정 성 무[†]·이 상 용[††]·장 중 순[†††]
송 재 신[††††]·유 해 영[†††††]·최 경 희[††††††]

요 약

본 논문에서는 TTT 타점법을 이용하여 웹 서버가 서비스하는 파일의 크기에 대한 통계적 분포는 꼬리부분이 두꺼운 분포라는 것을 판단하는 방법을 제시한다. TTT 타점법은 신뢰성 공학에서 사용되는 방법으로써 TTT 통계량 타점결과의 직선성으로 지수분포 여부를 판단하는 방법이다. 본 연구에서 제안하는 방법을 모의실험과 실제 운영중인 웹서버의 자료를 사용하여 실험한 결과, 기존의 방법인 Hill 추정법과 LLCD 타점법에 비하여 후미성을 정확하게 판단하고 있으며, 판단의 효율성 면에서도 그들보다 우수하다는 것을 확인하였다. 특히 제안하는 방법은 기존의 방법이 웹서버의 파일 분포 판정이나 통계학에서의 파레토 분포 판정시 나타날 수 있는 판정의 오류 가능성을 개선할 수 있다는 점도 확인하였다.

A Analysis of Heavy Tailed Distribution for Files in Web Servers Using TTT Plot Technique

Sung-Moo Jung[†] · Sang-Yong Lee[††] · Joong-Soon Jang[†††]
Jae-Shin Song[††††] · Hae-Young Yoo[†††††] · Kyung-Hee Choi[††††††]

ABSTRACT

In this paper, we propose a method of analysis to show the heavy-tailed statistical distribution of file sizes in web servers, using TTT plot technique. TTT plot technique, a well-known method in the area of reliability engineering, determines that a distribution of samples follows a heavy tailed one when their TTT statistical plots are lied on a straight line. We performed an intensive simulation using data gathered from real web servers. The simulation indicates that the proposed method is superior to Hill estimation technique or LLCD plot method in efficiency of data analysis. Moreover, the proposed method eliminates the possible decision error, which Pareto distribution or traditional method might cause.

키워드 : 웹 서버(Web Servers), 부하모델링(Workload Modeling), 후미성 분포(Heavy Tailed Distribution)

1. 서 론

정보통신기술의 급속한 발달과 함께 인터넷의 보급과 이용이 폭발적으로 증가되고 있으나 정보통신 인프라는 서비스나 데이터의 확장 속도를 따라가지 못하고 있어 서비스 응답 지연 등으로 인한 불편이 지속되고 있다[15]. 이에 따라 서비스 응답 지연을 줄이기 위하여 통신망에서의 지연을 줄이는 방법이나 웹 서버의 성능을 개선할 수 있는 방안들이 제시되고 있다. 특히 웹 서버 구조화 및 분산화 기술, 서버의 성능 분석, 프락시 캐싱(proxy caching), 웹 서버의 스케일링(scaling) 등과 같이 웹 서버의 성능 개선 기법 등이 매우 관심있게 연구되고 있다[2, 9].

웹 서버의 성능 개선에 관한 연구에서는 제안하는 기법이 적용된 웹 서버가 적절하며 현실성 있는 웹 부하에 대하여도 타당한 성능을 제공하는 가에 대한 실험이 필요하다. 이때 사용되는 자료로는 실제로 운영되고 있는 웹서버가 생성할 웹 로그와 통계적인 모형 분포로 웹 서버의 부하를 생성하는 도구들을 이용하는 경우가 있다. 웹 로그를 사용하는 경우에는 실제 환경에서 측정된 것이기 때문에 실제 환경에서의 성능을 실험할 수 있다는 장점이 있다. 그러나 신뢰할만한 정도의 실험을 위해서는 로그 확보에 많은 시간이 소요된다. 뿐만 아니라, 여러 가지 서버에 대하여도 해당 기법이 만족할 만한 성능이 나타나고 있는가를 검증하기 위해 다양한 서버로부터 데이터를 수집해야하는 어려움이 있다. 이러한 실측 로그를 이용한 성능 평가 방법의 제약성 때문에, 통계적 모형분포에 따라 웹 서버의 부하를 생성하고 이를 바탕으로 성능을 측정하고 평가하는 방법이

† 정 회 원 : 한국교육학술정보원 수석연구위원
†† 정 회 원 : 아주대학교 대학원 산업공학과
††† 정 회 원 : 아주대학교 산업정보시스템 공학부 교수
†††† 정 회 원 : 한국교육학술정보원 연구위원
††††† 정 회 원 : 단국대학교 정보 컴퓨터과학부 교수
†††††† 정 회 원 : 아주대학교 정보 및 컴퓨터공학부 교수
논문접수 : 2003년 5월 2일, 심사완료 : 2003년 5월 30일

많은 연구에서 사용되고 있다[1, 3, 13]. 통계적 방법으로 웹 서버의 부하를 생성하기 위해서는 부하 결정 요소를 파악하고 이들이 어떠한 통계적 분포를 따르는지에 대한 연구가 필요하다. 예를 들면, Boston 대학에서 개발된 웹 URL를 생성하는 시스템인 SURGE(Scalable URL Reference Generator)에서는 부하 결정 요소로 서버가 제공하는 파일들의 크기(file size) 및 선호도(popularity), 리퀘스트들의 도착시간(arrival times) 및 임시지역성(temporal locality) 등을 사용하고 있다[13]. 웹 서버의 부하를 결정하는 여러 가지 요소들 중에서 파일들의 크기는 웹 서버의 통신량과 파일 처리 부하에 결정적인 요소로 작용되고 있다. 따라서, 웹서버의 성능 모델링시 파일들의 크기에 관한 정확하고, 효율적인 통계적 분포 분석 과정은 매우 중요한 연구과제의 하나이다.

일반적으로 웹서버에서 파일들의 크기 분포는 여러 연구에서 후미성(厚尾性, 분포의 꼬리부분이 두꺼운 분포, heavy tailed distribution)을 갖는 것으로 나타나고 있다[1, 6, 10, 14]. 파일의 크기 분포가 후미성을 갖는다는 것은 크기가 작은 파일도 많이 서비스되지만 크기가 큰 파일들이 자주 나타나면서 분포의 평균이나 분산이 매우 커지게 되는 경우를 의미한다. 이같이 분포가 후미성을 가지는 경우에는 통계적인 분석이나 표현이 어려워질 뿐만 아니라, 대기행렬 등을 이용한 통계적 분석도 자연히 어려워진다[7, 8]. 이에, 서버가 서비스하는 파일들의 통계적 분포에 따라 웹 서버의 부하 분포를 정의하고, 이에 따르는 웹 요청에 사용되는 파일들의 크기를 생성하기 위해서는 우선적으로 파일들의 크기 분포가 후미성을 갖는 분포인지에 대한 판단이 필요하다.

이상과 같은 배경으로 웹 서버에서 서비스하는 파일 분포의 후미성을 판정하기 위해 여러가지 방법이 제안되고 이용되어 왔다[8, 10, 12, 14] 예를 들면, Hill 추정법과 LLCD (Log Log Complementary Distribution) 타점법, Q-Q 타점법 등이 이에 속한다. Hill 추정법은 지수분포의 모수인 평균의 추정치는 데이터 수가 증가할수록 특정한 값에 수렴한다는 성질을 이용한 것이다. LLCD 타점법이나 Q-Q 타점법은 지수분포의 확률지(probability paper)를 이용한 것이다. 그러나 Hill 추정법은 수렴성 판단이 어려울 뿐만 아니라 파레토 분포가 아닌 경우에도 수렴할 가능성이 있다. 또한 LLCD 타점법이나 Q-Q 타점법의 경우에도 확률지의 특성으로 인하여 분포의 후미성 여부가 잘못 판정될 가능성이 있다.

이에 본 논문에서는 분포의 후미성을 판단하기 위하여 TTT(Total Time on Test) 타점법[17]을 이용하는 방법을 제안한다. TTT 타점법은 신뢰성 공학에서 고장 분포가 지수분포인가를 확인하기 위하여 사용하는 방법의 하나로써, TTT 통계량 타점결과의 직선성으로 지수분포 여부를 판단하는 방법이다. 제안하는 방법을 기존의 방법인 Hill 추정법과 LLCD 타점법과 비교한 결과, 분포의 후미성을 정확하게 판단하고 있으며, 판단의 효율성 면에서도 그들보다 우수하다는 것을 확인하였다. 제안하는 방법은 기존의 방법이 웹 서버의 파일 분포 판정이나 통계학에서의 파레토 분포 판정시 나타날 수 있는 판정의 오류 가능성을 개선할 수 있다는 점도 확인하였다.

본 논문에서는 제안하는 후미성 판정 방법의 타당성 여부에 대해 모의실험을 통하여 확인하고, 교육용 컨텐츠를 서비스하는 EDUNET 서버, NASA 서버 등을 통하여 실제적인 적용 가능성을 검증하였다. 2장에서는 파일분포의 후미성을 판정하는 여러 가지 방법에 대하여 설명하고 3장에서는 본 연구에서 제안하는 후미성 판정법에 대하여 설명한다. 4장에서는 이에 대한 검증 결과를 제시하고 있다.

2. 후미성의 정의 및 판정방법

2.1 후미성의 정의

후미성을 갖는 분포는 상측 꼬리 부분이 다음과 같은 Power 법칙을 따르는 파레토 분포로 표현된다.

$$P[X > x] \sim x^{-\alpha} \quad 0 \le \alpha \le 2$$

즉, 임의의 상수 c에 대해서 $a(x) \sim b(x)$는 $\lim_{x \to \infty} \frac{a(x)}{b(x)} = c$ 인 분포를 의미한다. 후미성을 갖는 분포의 확률 변수들은 매우 큰 산포를 갖으며, 만약 $\alpha \le 1$이라면 평균과 분산 또한 무한대가 된다.

파일의 크기를 나타내는 확률변수를 X라고 하고, F(x)를 X의 누적분포 함수라고 할 때, X의 분포가 후미성을 갖는다는 것은

$$1 - F(x) \sim x^{-\alpha} L(x), \; x \to \infty \qquad (1)$$

임을 말한다[14]. 여기서 L(x)는 $x > 0$에 대하여

$$\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1$$

을 만족하는 완만한 변화를 갖는 함수(slowly varying function)이다. 이때 L(x)는 다음과 같은 형태가 있을 수 있다.

$$L(x) = \begin{cases} c + o(1) \\ \log x \\ \log(\log(x)) \\ 1/\log(x) \end{cases}$$

따라서 식 (1)의 정의는

$$\lim_{t \to \infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-\alpha}, \quad x > 0$$

로도 표현될 수 있다. 단, $\overline{F} = 1 - F$이다. 여기서 만일 $L(x) = 1$이 되면 $F(x)$는 파레토 분포가 된다.

이와 같이 파일의 크기 분포가 후미성을 갖는 경우에는 일반적으로 잘 알려진 분포, 예를 들어 정규분포나 대수정규분포, 와이블 분포 등으로는 분포의 모델링이 불가능하다.

2.2 후미성과 웹 서버의 부하와의 관계

웹서버가 제공하는 파일들의 크기 분포에 후미성이 있는 경우, 부하에 미치는 영향은 다음과 같이 세 가지를 들 수 있다.

첫째, 분포의 꼬리 부분에 위치하는 크기가 큰 파일들은 저장공간이나 네트워크 트래픽 부하에 결정적인 영향을 줄 수 있다. 1998년 월드컵 웹서버가 전송한 유일한 파일들의 분포를 살펴보면 1KB 미만의 크기를 갖는 파일들의 개수는 전체 파일의 10.4%인 것으로 나타났다. 그러나, 이들이 차지하는 총 저장용량은 0.4%밖에는 되지 않으며, 전송된 데이터의 총량도 5.8%에 지나지 않는다. 반면에 64KB 이상의 크기를 갖는 파일들의 개수는 전체 파일의 0.4%임에도 불구하고, 저장용량은 50.7%를 차지하고 있다. 또한, 총 요청수 중 이 파일들의 요청수는 0.1%임에도 불구하고 이들에 의한 전송량은 전체의 21%를 차지하게 된다. 이러한 현상을 미루어 볼 때 파일 분포의 꼬리 부분에 존재하는 파일들은 웹서버의 저장 용량과 전송 효율에 미치는 영향은 결정적이라 할 수 있다[10].

둘째, 캐싱 효율을 감소시킬 수 있다. 웹 자원에 대한 캐싱은 다른 분야의 캐싱과 달리 파일 단위로 이루어진다. 또한 파일의 크기가 아주 작은 파일부터 큰 파일까지 다양한 종류의 파일들이 캐시에서 처리된다. 캐시의 크기가 한정되어 있는 경우 크기가 큰 파일이 캐시에 들어가면 크기가 작은 여러 개의 파일들이 캐시에서 제거되며, 이로 인해 캐시의 히트율이 감소될 수밖에 없다. 이러한 문제를 개선하기 위하여 Markatos의 메인 메모리 캐싱에서는 크기를 고려한 캐싱 전략을 제안한 바 있다[4]. 이 연구에서는 캐싱 대상 파일을 다양한 크기로 변경하여 실험한 결과, 적정한 크기의 임계치(threshold) 이내에 포함되는 파일들만을 캐싱하는 것이 보다 효과적임을 보여주고 있다. 따라서 파일 분포의 꼬리 부분에 위치하는 파일들을 무작위로 캐싱할 경우 캐시의 효율성에 악 영향을 주며 궁극적으로 서비스 효율을 감소시키게 된다.

셋째, 파일들이 서비스 되다가 중단되는 현상이 빈번하게 발생하여 서비스에 대한 신뢰도가 떨어질 수 있다. 웹 서버에서 처리하는 대다수의 파일들이 20KB 미만으로 분포되어 있을 경우, 이들은 대개 1~2초 이내에 서비스가 완료된다. 그러나, 꼬리 부분에 분포하는 파일들이 서비스되기 위해서는 그 이상의 시간이 소요된다. 이 경우 사용자들은 정상적인 서비스가 이루어지지 못하고 있다고 판단하여 서비스를 강제적으로 중단하거나 시간 초과(time out)로 정상적인 서비스가 불가능할 수 있다. John Dilly의 연구 결과에서는, 서비스되는 파일들의 평균 크기가 18KB인 웹 서비스 시스템에서 파일의 크기가 37KB~64KB인 이미지 파일들의 25% 이상은 서비스가 종료되지 않은 상태로 파일 서비스가 중단된다는 사실을 밝힌 바 있다[6].

한편, 후미성이 있는 분포에 의한 통계적인 영향을 살펴보면, 이러한 분포는 평균이나 분산이 무한대가 되어 통계적인 분석이나 표현이 어려울 뿐 아니라, 대기행렬 등의 분석도 힘들게 된다.

이상과 같은 배경으로 웹서버에서 서비스하는 파일들의 크기 분포에 후미성이 있는가를 판단하는 것은 정확한 부하 생성시 반드시 고려하여야 할 요소의 하나이다.

2.3 후미성 판정법 고찰

종래의 관련 연구에서 통계 분포의 후미성을 판정하기 위하여 Hill 추정법, LLCD 타점법, Q-Q 타점법 등이 주로 사용되어 왔다. 이들을 살펴보면 다음과 같다.

2.3.1 Hill 추정법

Hill 추정법은 파레토 분포의 특성을 이용하여 분포의 후미성을 판정하는 것이다. 우선 파일들의 크기를 $X_1, \cdots, X_n$이라고 하고, $X_{(1)}, \cdots, X_{(n)}$을 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$을 만족하는 순서통계량(order statistic)이라고 하자. 여기서 Hill 추정치 $\hat{\theta}_{n,k}$는 식 (2)와 같이 표현된다.

$$\hat{\theta}_{n,k} = \frac{1}{(n-k)} \sum_{i=k+1}^{n} \ln\left(\frac{X_{(i)}}{X_{(k)}}\right) \tag{2}$$

이때 $X_1, \cdots, X_n$이 $X_{(k)}$ 이후에서 파레토 분포를 갖는 경우 Hill 추정치 $\hat{\theta}_{n,k}$는 특정한 값으로 수렴하게 된다[5].

이 방법은 파일들의 크기 분포에 대한 Hill 추정치가 특정한 값으로 수렴하게 되면 이 분포는 모수 α를 갖으며, 후미성을 갖는 분포라고 판단하는 방법이다. 따라서 이 방법은 매우 간단하고 적용하기 쉬운 장점이 있는 반면, 후미성을 갖지 않는 경우에도 추정치가 수렴할 가능성이 있다는 단점이 있다.

2.3.2 LLCD 타점법

LLCD 타점법 또한 파레토 분포의 특성을 이용하여 분포의 후미성을 판정하는 것이다. 파레토 분포의 경우

$$\left\{\left(\ln\left(1 - \frac{i}{n-k-1}\right), \ln X_{(i)}\right), k < i \le n\right\} \tag{3}$$

을 타점한 그래프는 직선이 된다는 사실을 이용한 방법이다. 파레토 분포의 경우는 식 (1)에서 $L(x)$가 1이므로,

$$\ln P\{\ln X > x\} = -\alpha x \tag{4}$$

가 성립함을 이용한 것이다[14]. 지수분포의 적합성을 그래프로 검정하는 방법으로 지수확률지가 제안되었는데, LLCD 타점법도 같은 것이라 할 수 있다. 즉, LLCD 타점을 한 경우 꼬리 부분이 선형에 근접한다는 사실을 이용하는 것이다.

2.3.3 Q-Q 타점법

Q-Q 타점법은 LLCD 타접법과 유사한 방법이다. 파레토 분포의 경우

$$\left\{\left(-\ln(1-\frac{i}{n-k}),\ \ln X_{(i)}\right),\ k < i \le n\right\} \qquad (5)$$

을 타점한 그래프는 근사적으로 $\frac{1}{\alpha}$의 기울기를 갖는 직선이 된다는 사실을 응용하는 방법이다[12]. 이 방법에서 n이 충분히 크면 LLCD 타점법과 동일한 특성을 갖는다.

3. TTT 타점법을 이용한 후미성 판정법

3.1 TTT 타점법과 후미성

앞에서 설명한 방법인 Hill 추정법이나 LLCD 타점법 등은 모두 수렴성을 이용하는 것이다. 이러한 방법들은 매우 간단하고 적용하기 쉬운 장점이 있다. 그러나 Hill 추정법은 수렴성의 판단이 어려울 뿐만 아니라 파레토 분포가 아닌 경우에도 수렴할 가능성이 있다는 단점이 있다. 또한 LLCD 타점법이나 Q-Q 타점법의 경우에도 확률지의 특성으로 인하여 파레토 분포가 아닌 경우에도 그렇다고 판단될 가능성이 있다. 이에 본 논문에서는 분포의 후미성을 판단하기 위하여 TTT 타점법[17]을 이용하는 방법을 제안한다.

TTT 타점법은 신뢰성 공학에서 고장분포가 지수분포인가를 확인하는데 사용하는 방법이다. 즉, 지수분포의 특성인 Durbin 변환의 지수분포성을 이용하는 것으로서, TTT 통계량 타점결과의 직선성으로 지수분포 여부를 판단하는 방법이다.

$Y_1,\cdots,Y_n$이 평균 θ를 갖으며, 지수분포를 따르는 확률표본이라고 하고, $Y_{(1)},\cdots,Y_{(n)}$은 $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$을 만족하는 순서통계량이라 하자. 여기서

$$W_i = (n-i+1)(Y_{(i)} - Y_{(i-1)}),\ 1 < i \le n,$$

$$D_k = \sum_{i=1}^{k} W_i, \qquad 1 \le i \le n,$$

이라고 하면, TTT 타점법은

$$\left(\frac{D_i}{D_n}, \frac{i}{n}\right), \qquad 1 \le i \le n$$

에 대하여 타점하는 것이다. 만일 타점 그래프가 직선이 되고, 기울기가 1이면, 지수분포로 판정한다.

3.2 후미성 판정법

TTT 타점법의 특징은 지수분포를 따르는 확률표본인 $(Y_1,\cdots,Y_n)$로 부터 변환된 $(W_1,\cdots,W_n)$도 지수분포의 확률표본이 된다는 성질을 이용한 것으로, 이는 지수분포의 독특한 특성이다. 이 방법을 이용하여 분포의 후미성 여부는 다음과 같이 판정할 수 있다.

$X_1,\cdots,X_n$을 파일의 크기를 나타내는 확률변수라고 하고, $X_{(1)},\cdots,X_{(n)}$은 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$을 만족하는 순서통계량이라고 하자. 만일 $X_1,\cdots,X_n$이 $X_{(k)}$이후 파레토 분포를 갖는 경우 식 (3)에 의하여 $\ln X_{k+1},\cdots,\ln X_n$는 지수분포에서 추출된 확률 표본과 같게 된다. 이때 $Y_i = \ln X_i$, $k+1 \le i \le n$로 하여 TTT 타점을 하면 직선을 얻을 수 있다. 이와 같이 타점 그래프가 직선으로 나타나면 확률 표본이 파레토 분포를 따른다는 것을 의미하게 되므로, 이는 결국 후미성을 갖는 분포라는 것을 알 수 있게 된다. 즉, 파레토 분포의 경우에는 상위 꼬리 부분에 위치하고 있는 데이터에 Log를 취한 다음 TTT 타점법을 적용하면 그래프가 직선으로 표현된다. 일반적으로 꼬리가 두꺼운 경우에는 Hill 추정법에서와 같이 $\ln\frac{X_{(i)}}{X_{(k+1)}}$을 데이터로 이용하면 된다. 즉, (1)에서 L(x)가 상수가 아닌 경우에는 Resnick[14]의 결과에 따라

$$Y_i = \ln\frac{X_i}{X_k},\ k+1 \le i \le n$$

으로 치환하면 같은 결과를 얻을 수 있다.

4. 제안하는 판정법의 유효성 검증 실험

본 절에서는 분포의 후미성 판정하기 위하여 TTT 타점법을 이용한 결과를 앞에서 제시한 여러 가지 방법과 비교하여 이의 유효성을 검증하고자 한다.

4.1 모의 데이터를 이용한 비교실험

후미성 판정 방법의 유효성을 비교하기 위하여 파레토 분포(Pareto distribution), 지수분포(exponential distribution), 정규분포(normal distribution), 와이블 분포(Weibull distribution), 대수정규분포(lognormal distribution) 등 5가지 분포를 대상으로 실험하였다. 그리고 각 분포를 따르는 난수들을 발생시킨 후 여러 가지 방법의 판정 결과를 비교하였다. 실험을 위하여 사용된 각 분포에서의 모수와 확률밀도함수는 <표 1>과 같다. 여기서 모수는 Log(x)들의 평균이 1이 되도록 설정한 것이다.

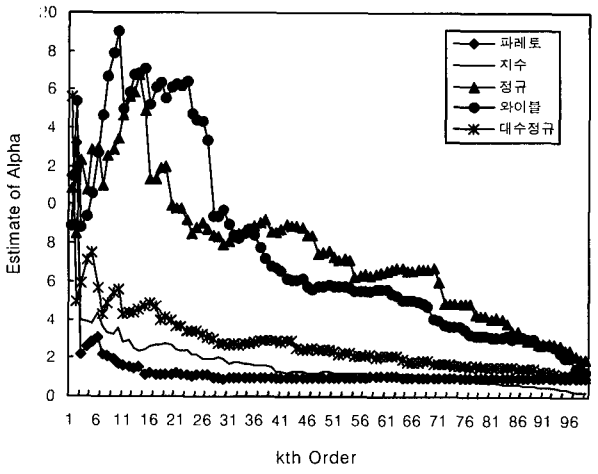
(그림 1)은 <표 1>의 분포들로 부터 난수를 100개씩 발생시켜 Hill 추정법으로 타점한 결과이다.

(그림 1)을 살펴보면 우선 파레토 분포의 경우에는 Hill

추정치가 수렴하고 있음을 알 수 있다. 또한 상대적으로 얇은 꼬리를 갖는 와이블 분포나 정규분포에서는 수렴성이 보이지 않는다. 그러나 후미성이 없는 것으로 알려진 분포인 지수분포나 대수정규분포의 경우에도 Hill 추정치는 수렴하고 있음을 볼 수 있다. 이러한 결과는 Hill 추정법이 이용하는 추정치가 확률표본의 평균만을 고려하고 있기 때문에 초래된 것으로 판단된다. 따라서 이 방법은 파레토 분포와 같이 특정한 분포에서만 추정치가 수렴된다고 할 수 없으며 분포의 후미성 판정도 부정확할 수 있다.

〈표 1〉 실험 대상 분포의 모수 및 확률밀도함수

분 포	모 수	확률밀도함수
파레토 분포	$\alpha = 1$	$kx^{-(\alpha+1)}$
지수 분포	$\theta = 4.95$	$\frac{1}{\theta} e^{-\frac{x}{\theta}}$
정규 분포	$\mu = 2.72,\ \sigma = 0.5$	$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
와이블 분포	$\alpha = 3,\ \beta = 5$	$\frac{\beta}{\alpha} x^{\beta-1} e^{-\frac{x^\beta}{\alpha}}$
대수정규 분포	$\mu = 1,\ \sigma = 0.5$	$\frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$

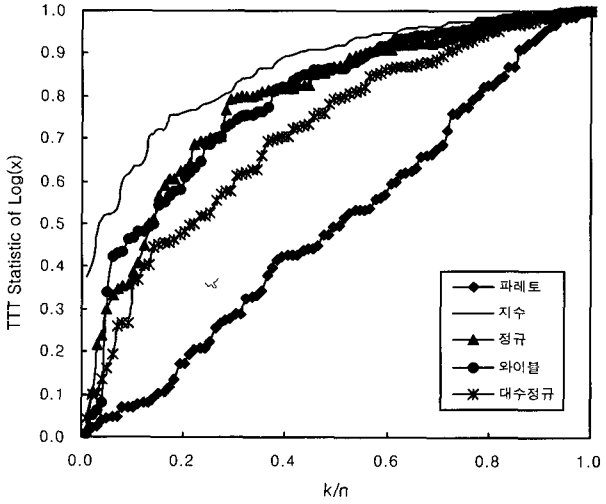


(그림 1) 모의 데이터 분포의 Hill 추정치 타점 결과

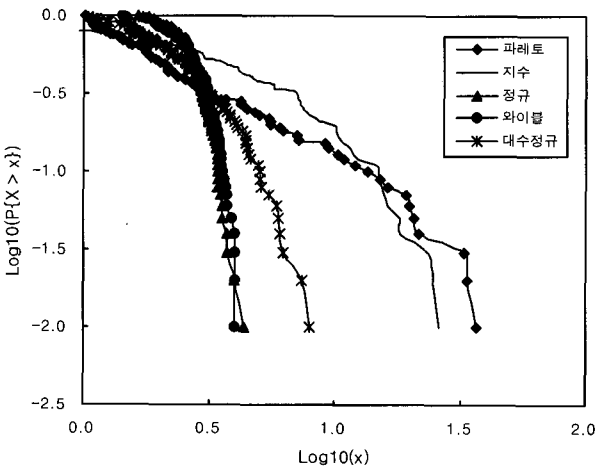
한편, TTT 타점법과 LLCD 타점법은 모두 그래프의 직선성으로 후미성을 판단하는 방법이다. (그림 2)와 (그림 3)은 Hill 추정치 실험시 사용하였던 표본들에 대하여 TTT 타점과 LLCD 타점을 실시한 결과이다.

(그림 2)와 (그림 3)을 살펴보면 두 방법 모두 파레토 분포가 아닌 경우에는 직선이 아님을 알 수 있다. 이는 두 방법 모두 Hill 추정법보다는 판정의 정확도면에서 우수하다는 것을 의미한다.

파레토 분포에서는 TTT 타점법과 LLCD 타점법은 그래프 상으로는 큰 차이를 보이지 않는다. 다만 LLCD 타점법의 경우에는 오른쪽으로 갈수록, 즉 파일의 크기가 커질수록 직선에서 멀어지는 현상을 보이고 있다.



(그림 2). 모의 데이터 분포의 TTT 타점 결과



(그림 3) 모의 데이터 분포의 LLCD 타점 결과

TTT 타점법과 LLCD 타점법은 모두 타점된 점들의 선형성을 바탕으로 하는 것이다. 특히, 타점된 형태가 직선과 가까울수록 파레토 분포, 즉 분포가 후미성을 갖는다고 판단하는 방법이다. 본 연구에서는 두 방법간의 차이성을 비교하기 위해, 각 방법에서 생성되는 결정 계수(coefficient of determination) R^2를 분석하였다. R^2은 선형회귀분석에서 선형 모형을 판단하는 기준으로 사용된다. 일반적으로 R^2는 1에 가까울수록 강한 직선성을 나타내는 지표로서 0.9 이상이면 직선으로 간주한다. <표 2>는 앞에서 설명한 방법과 같은 실험을 100회 실시하여 구한 R^2의 평균과 표준편차를 나타낸 것이다. <표 2>에서와 같이 TTT 타점법이나 LLCD 타점법 모두 파레토 분포의 경우에만 R^2의 값이 1에 가까운 값으로 나타나고 있다. 그러나 파레토 분포의 경우에서도 TTT 타점법이 LLCD 타점법에 비하여 평균값은 크고, 표준편차는 작게 나타나고 있다. LLCD 타점법의 경우 앞

에서도 설명한 바와 같이 타점 결과의 오른쪽 부분이 비선형으로 휘고 있어 R^2의 평균 값은 작아지고 표준편차는 증가된 것으로 판단된다. 이러한 경우 직선 여부를 판단하는데 어려움을 줄 수 있다.

〈표 2〉 결정계수(R^2)의 비교(n=100)

구 분	평 균		표준편차	
	TTT	LLCD	TTT	LLCD
파레토	0.9923	0.9788	0.0054	0.0167
지 수	0.7405	0.6955	0.0674	0.0590
정 규	0.7897	0.7771	0.0542	0.0431
와이블	0.7360	0.6908	0.0605	0.0514
대수정규	0.8553	0.8531	0.0392	0.0393

본 연구에서는 같은 방법으로 1,000회의 실험을 반복 수행하였는데 유사한 결과를 얻을 수 있었다. 이상과 같은 결과를 미루어 볼 때 그래프의 직선성 판단 방법으로 TTT 타점법이 LLCD 타점법 보다 우수하다고 할 수 있다.

4.2 실측 데이터를 이용한 비교실험

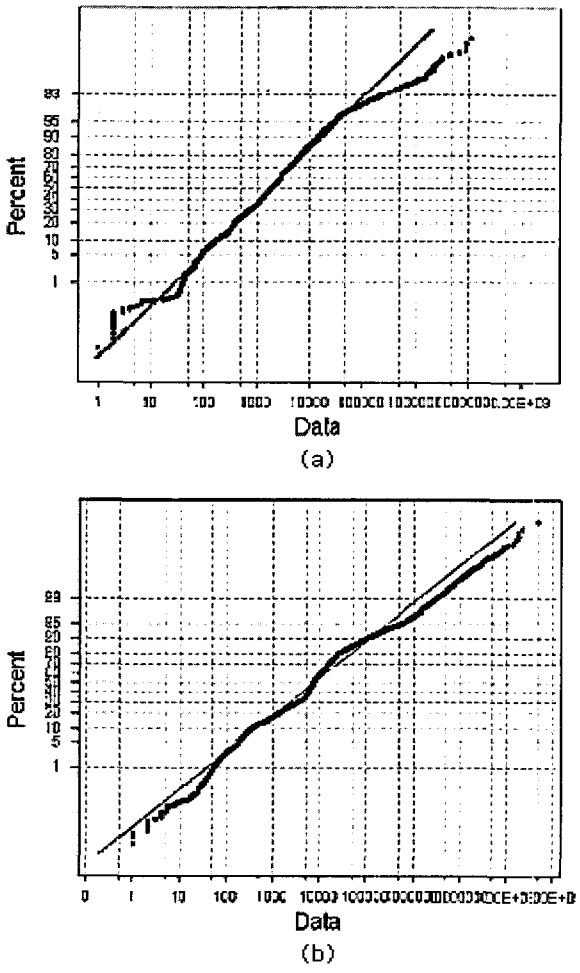
앞서 제시된 방법을 적용하기 위하여 본 연구에서는 한국교육학술정보원이 운영하고 있는 에듀넷의 사용자 인증서버(Edu01)와 평생교육서버(Hp01)로 부터 2000년 7월 22일부터 14일간의 로그 데이터와, NASA가 운영하고 있는 웹서버로부터 1996년 7월 1일부터 35일간 수집된 로그를 이용하여 분석하였다[16]. 단, NASA의 경우 웹로그에 대한 정보만 수집되어 서버에 대한 분석은 불가능하였음을 밝혀둔다. 또한 본 연구에서는 요청되는 파일 크기들에 대하여 관심이 있기 때문에 서버에서 실행되는 프로그램인 CGI 파일들은 분석에서 제외하였다. <표 3>은 에듀넷 서버에 탑재되어 있는 파일들 중 이용 빈도가 비교적 높은 파일들의 통계 정보이다.

〈표 3〉 Edu01 서버와 Hp01 서버 파일의 종류별 분포

구 분	파일종류	개 수	평균크기	표준편차	왜 도	첨 도
Edu01 서버	gif	6,255	3,212	7,214	13.86	389.91
	html/htm	5,163	4,292	8,925	7.51	86.18
	jpg	882	11,982	10,488	4.84	50.30
	txt	254	11,422	56,749	11.59	157.87
	total	19,261	41,119	2,008,439	136.78	18887.21
Hp01 서버	gif	18,061	7,994	21,873	11.44	201.28
	html/htm	31,914	5,090	12,714	9.26	169.12
	jpg	33,012	25,755	29,960	6.94	153.66
	pdf	40,777	97,128	679,985	31.60	2161.25
	wav	6,019	496,534	965,652	5.78	62.04
	hwp	854	148,015	600,116	9.52	103.31
	total	171,362	172,903	2,529,464	122.01	22522.64

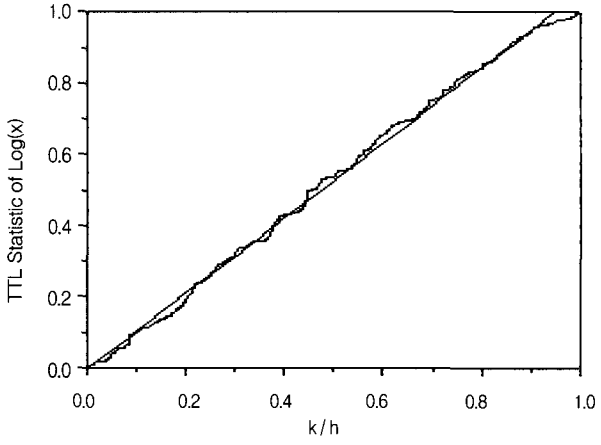
<표 3>을 보면 우선 각 서버별로 파일의 개수나 크기 면에서 많은 차이를 보이고 있음을 알 수 있다. 또한 평균이나 표준편차 등도 많은 차이를 보이고 있다. 두 개의 서버 모두 평균에 비하여 표준편차가 대단히 크다는 것을 알 수 있다. 이는 파일들의 크기가 넓은 범위에 걸쳐 분포하고 있음을 의미한다. 대부분의 왜도는 0 보다 매우 큰데, 이것은 파일의 분포가 오른쪽으로 기울어진(right skewed) 분포 임을 나타내며, 첨도가 3 보다 큰 것은 정규분포보다 뾰족한 분포를 갖고 있음을 의미한다. 이러한 통계 데이터로부터 서버에 탑재되어 있는 자료들이 후미성을 지니고 있을 것이라는 것을 유추할 수 있다.

다음은 서버 탑재 파일 분포의 TTT 타점법을 이용한 후미성 판정 방법을 적용한 결과이다. 먼저 파일크기 분포 중에서 후미성을 발생하고 있는 부분을 찾기 위해 확률지를 이용하였다. (그림 4)에서 보는 바와 같이 분포의 몸체 부분은 대수정규분포 확률지에 가장 적합하다. 이는 서버에 탑재된 대부분의 파일들은 대수 정규 분포를 따르고 있다는 것을 나타낸다.

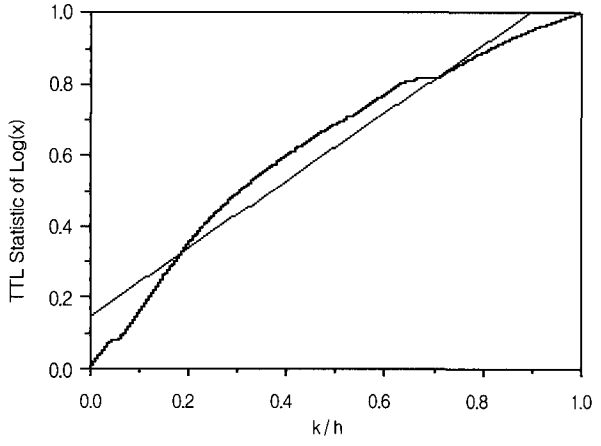


(그림 4) 서버 파일의 대수정규확률지 타점 결과 - (a) Edu01, (b) Hp

그러나 Edu01의 97%, Hp01의 90% 이상의 파일들은 대수정규분포를 벗어나고 있음을 보여주고 있다. 이 부분이 바로 후미성을 갖는 분포의 경계가 된다. 이 경계점 이후의 해당 데이터를 추출하고, 이들 분포에 대해 TTT 타점법을 이용한 후미성 검정 결과는 (그림 5)와 같다.



(a)



(b)

(그림 5) TTT 타점 결과 - (a) Edu01, (b) Hp01

(그림 5)와 같이 파일의 크기에 대한 TTT 타점 결과는 직선성을 나타내고 있으며, 결정계수를 구해본 결과 Edu01은 98.09%, Hp01은 96.76%로 직선에 가까움을 보이고 있다. 따라서 서버 파일 분포에는 후미성이 존재하고 있다고 할 수 있다.

서버가 가지고 있는 파일들은 모두가 서비스되는 것이 아니다. 서버에 탑재는 되어 있으나 전송되지 않는 파일들이 다수 존재하며, 분포의 후미 부분에도 전송되지 않는 파일들이 다수 존재한다. 따라서 서버의 파일들 중 실제 서비스된 파일들, 즉 전송파일들에 대한 요소들도 작업 부하를 생성시 반영할 필요가 있다. <표 4>는 앞에서 제시한 각 서버들에서의 전송파일을 분석한 결과이다.

<표 4>는 전송된 파일 중 요청 횟수가 최상위인 파일들에 대한 통계량을 구해 본 것으로 서버에 탑재되어 있는 파일들과 같이 표준편차가 매우 크며, 대부분의 파일들의 왜도와 첨도가 모두 크게 나타나고 있다. 여기에서 파일들의 빈도는 웹 서버의 로그에 기록된 전송 파일 정보를 이용하여 유일한 파일을 구하고, 이들 각각에 대하여 파일들의 종류별 통계량을 분석하여 얻은 데이터이다. <표 4>에서 보는 바와 같이 전송파일의 분포 역시 서버 전체 탑재 파일과 마찬가지로 후미성을 갖는다고 할 수 있다. (그림 6)은 전송파일들을 대수정규확률지에 타점한 결과이다.

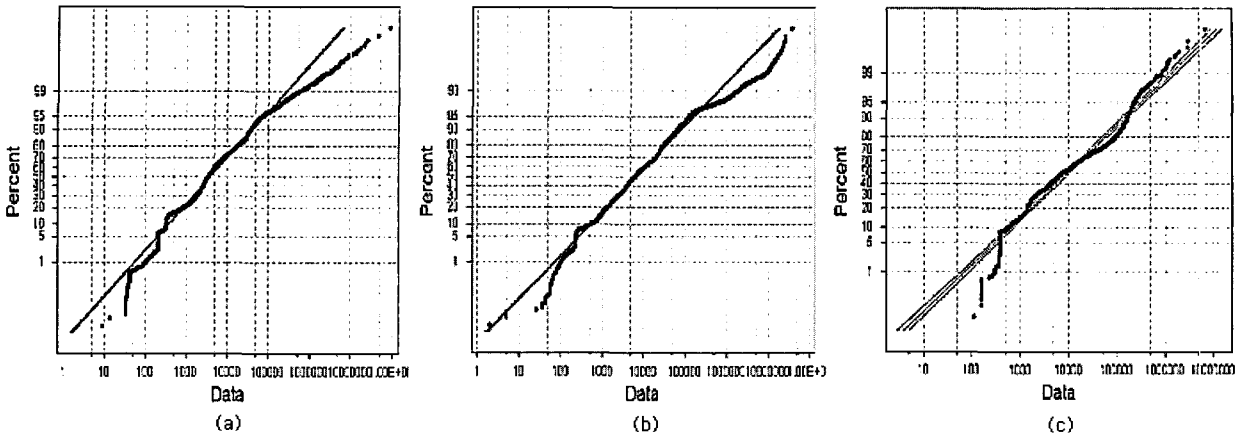
〈표 4〉 에듀넷, NASA 전송파일들의 통계정보

구분	파일 종류	개수	평균크기 (Byte)	표준편차	왜도	첨도
Edu01 Unique File	gif	11,196	10,450	24,005.24	9.34	155.39
	html/htm	9,376	5,021	11,679.15	17.00	492.40
	jpg	4,520	22,542	37,211.12	8.21	118.76
	txt	109	3,416	4,872.70	4.31	25.81
	total	28,023	44,119	722,229.28	81.47	8891.59
Hp01 Unique File	gif	8,994	9,216	19,014.95	8.74	155.09
	html/htm	5,604	6,325	15,743.96	10.33	187.25
	jpg	7,975	25,533	30,579.95	2.32	6.12
	pdf	1,185	194,357	561,763.00	7.05	62.58
	wav	192	1335,710	1,428,032.00	6.08	56.37
	hwp	765	78,751	187,766.20	7.01	59.10
	total	26,449	91,928	787,183.13	19.53	505.93
NASA Unique File	gif	554	60,137	69,893.97	1.56	3.45
	html/htm	438	45,828	245,714.50	9.59	100.33
	jpg	161	120,462	101,904.80	1.27	1.63
	txt	201	37,576	90,819.35	4.93	35.33
	mpg	26	586,493	353,776.70	0.36	-1.11
	wav	9	253,730	263,998.00	1.25	0.07
	total	1,675	65,224	237,236.50	17.19	424.62

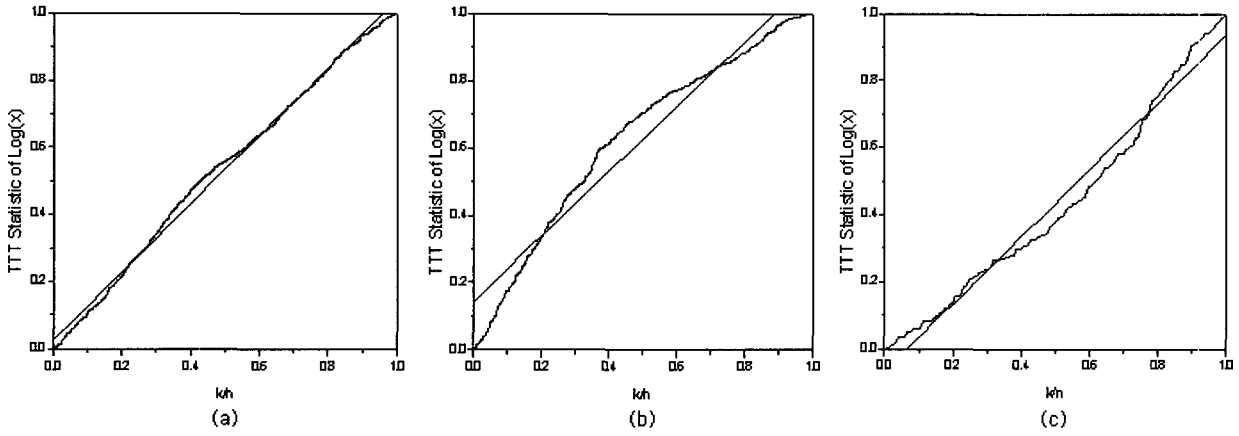
(그림 6)에서와 같이 Edu01은 상위 95%이상, Hp01은 상위 97% 이상, NASA는 90% 이상에서는 적합하지 않고 있으며 이 부분이 후미성을 일으키는 부분이라고 할 수 있다. (그림 7)은 해당 데이터를 추출하여 TTT 타점법을 적용한 결과이다.

이상의 그림들은 직선에서 크게 벗어나지 않고 있음을 보이며, 결정계수 R^2값을 보면 Edu01 99.75%, Hp01 95.12%, NASA 98.26%로 전송파일 분포 역시 후미성이 존재함을 알 수 있다.

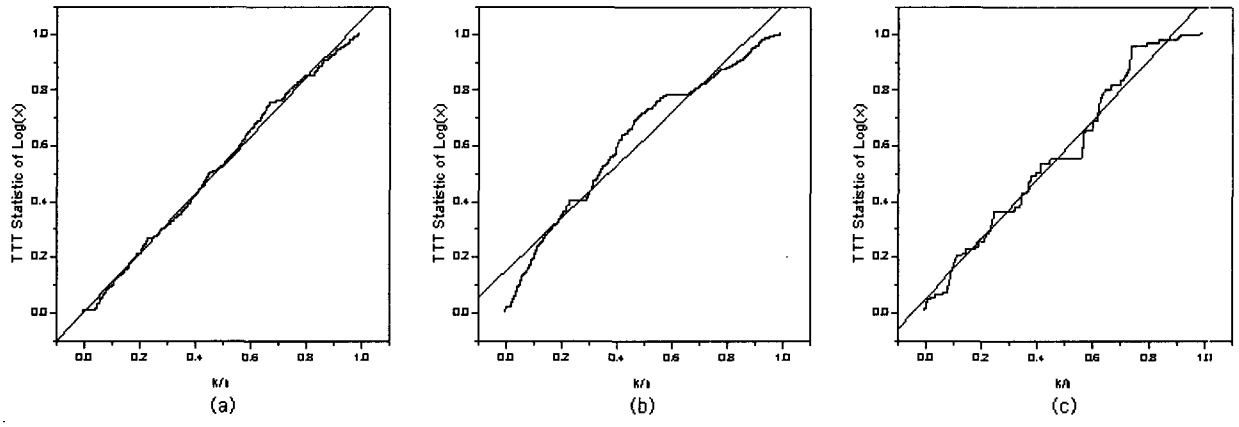
앞에서는 서버에 탑재되어 있는 파일들과 전송파일들에 대한 후미성을 분석하였다. 그러나 작업부하는 실제로 서비스된 파일들을 나타내는 로그, 즉 요청 크기(request size)에 의존하게 된다. 그런데 요청 크기의 분포는 파레토 분포로서 잘 설명되어진다고 알려져 있다. 본 연구에서는 3개 서버의 로그파일에 대해 TTT 타점법을 적용한 결과는 (그



(그림 6) 전송 파일의 대수정규확률지 타점 결과 - (a) Edu01, (b) Hp01, (c) NASA



(그림 7) 유일한 전송 파일의 TTT 타점 결과 - (a) Edu01, (b) Hp01, (c) NASA



(그림 8) 요청크기 TTT 타점 결과 - (a) Edu01, (b) Hp01, (c) NASA

림 8)과 같다.

(그림 8)에서와 같이 타점결과는 직선에서 크게 벗어나지 않고 있음을 보이고 있으며, 결정계수 R^2값을 구해보면 Edu01은 99.62%, Hp01은 95.45%, NASA는 98.04%의 값을 나타내고 있으므로 파일들이 요청크기 분포에서도 후미성을 일으키고 있음을 알 수 있다.

5. 결 론

웹서버의 성능 평가, 웹서버의 분산화 방법, 웹서버의 스케일링 등 다양한 웹서버 관련 연구시, 서버의 부하를 적절하게 생성하고 적용하는 것은 제안하는 방법에 대한 타당성 검증시 중요한 의미를 갖는다. 특히, 서버가 제공하는

파일들의 크기를 적절하게 모델링하는 것은 서버의 부하 생성에 매우 중요한 요소가 되며, 일반적인 웹 서버에서 파일들의 크기 분포는 후미성을 가진다는 것도 잘 알려진 사실이다. 이러한 배경으로 웹서버의 파일 분포가 지니는 후미성을 평가하기 위하여 여러 가지 방법이 제안되고 있다.

본 연구에서는 웹 서버가 보유하고 있는 파일들의 크기 분포의 후미성을 판단하기 위하여 TTT 타점법을 이용하는 방법을 제안하였다. 본 연구에서 새로이 제안하는 방법은 모의실험을 통하여 그 유효성 검증하였으며, 동시에 실제로 운영중인 웹 서버가 제공하는 데이터를 이용한 실험 결과도 제시하였다. 그 결과 본 연구에서 제안하는 방법인 TTT 타점법을 이용하는 방법이 기존의 방법인 Hill 추정법과 LLCD 타점법에 비하여 후미성을 정확하게 판단하고, 판단의 효율성 면에서도 그들보다 우수하다는 것을 확인하였다. 특히 제안하는 방법은 웹 서버의 파일 분포 판정이나 통계학에서의 파레토 분포 판정시 나타날 수 있는 판정의 오류 가능성을 개선할 수 있음도 보여주고 있다.

한편 본 연구에서 제안하는 방법을 포함한 여러 가지 연구에서 보여주는 후미성 분포 판정시 주관적 판단이 아닌 수치적인 방법들을 응용할 필요가 있으며, 특히, 각종 모수들을 정확하게 생성하는 방법을 통하여 웹 서버의 부하를 보다 정확히 모델링하기 위한 연구가 계속되어야 할 것이다. 또한 본 연구는 웹 서버의 부하 모델링시 사용하는 여러 가지 요소들 중에서 파일들의 크기 분포만을 판단하는 방법을 지시한 것이다. 따라서, 본 연구 결과를 이용하여 파일들의 선호도, 임시지역성 등 여러 가지 종류의 웹 서버 부하 결정 요소들에 대한 통계적 분포를 분석하고 모델링하는 방법에 대한 연구도 계속되어야 하는 과제임을 밝혀둔다.

참 고 문 헌

[1] A. Bestavros, "Discovering Spatial Locality in WWW Access Patterns using Data Mining of Document Clusters in Server Logs," *Technical Report*, TR-97-016, Computer Sci. Dept. Boston Univ., August, 1997.

[2] Binzhang Liu, Ghaleb Abdulla, Tommy Johnson and Edward A. Fox, "Web Response Time and Proxy Caching," *Technical Report*, TR-98-07, Computer Sci. Dept. Virginia Polytechnic Inst. and State University, March, 1998.

[3] Daniel Andresen and Tao Yang, "Adaptive Scheduling with Client Resources to Improve WWW Server Scalability," *Technical Report*, TRCS-96-27, Department of Computer Science UC Santa Barbara, November, 1996.

[4] E. P. Markatos, "Main Memory Caching of web documents," *Proceedings of the 5th International World Wide Web Conference*, Paris, May, 1996.

[5] Hill B. M., "A Simple General Approach to Inference about the Tail of a Distribution," *Ann. Statist.*, Vol.3, No.5, pp. 1163-1174, 1975.

[6] J. Dilly, "Web Server Workload Characterization," *www.hpl.hp.com*, Hewlett Packard Co., December, 1996.

[7] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic : Evidence and possible causes," *In Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, May, 1996.

[8] M. E. Crovella and M. Taqqu, "Estimating the Heavy Tail Index from Scaling Properties," *Methodology and Computing in Applied Probability*, Vol.1, No.1, 1999.

[9] M. E. Crovella, Robert Frangioso and Mor Harchol Balter, "Connection Scheduling in Web Servers," *Technical Report*, Computer Sci. Lab. M.I.T Univ., March, 1999.

[10] M. E. Crovella, M. Taqqu, A. Bestavros, "Heavy-Tailed Probability Distribution in the World wide web," citeseer.nj.nec.com/crovella98heavytailed.html.

[11] M. F. Arlitt and T. Jin, "Workload Charcterization of the 1998 World Cup Web Site," *Internet Systems and Application Laborator,* HPL-1999-35, Hewlett Packard Co., September, 1999

[12] M. Kratz and S. I. Resnick, "The QQ Estimator and Heavy Tails," *Comm. Statist.-Stoch. Models*, Vol.12., No.4, pp. 699-724, 1996.

[13] P. Barford and M. E. Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation," *Proceedings of ACM SIGMETRICS 98*, Madison, WI, pp.151-160, June, 1998.

[14] S. Resenick, "Heavy Tail Modeling and Teletraffic Data," *The Annual of Atatistics*, Vol.25. No.5, pp.1805-1869, 1997.

[15] T. Spangler, "Promising Satellite Services Emerge as Alternative to Earthbound Lines, *Internet World*, March, 1998.

[16] ftp://ita.ee.lbl.gov/traces/NASA_access_log_Aug95.gz.

[17] Yu Hayakawa, "The Total Time on Test Statistics and LI-Isotropy," *Intl. J. Reliability, Quality and Safety Eng.*, Vol.7, No.2, pp.143-151, 2000.

정 성 무

e-mail : smjung@keris.or.kr

1981년 충남대학교 전자공학과(학사)

1988년 한양대학교 전자공학과(석사)

2000년 아주대학교 컴퓨터공학과(박사)

1981년~1989년 잠실중, 경기공고 교사

1989년~1997년 한국교육개발원 부연구위원

1997년~현재 한국교육학술정보원 수석연구위원

관심분야 : e-Learning, Authoring Systems, 시스템 프로그램, 교육정보화 등

이 상 용

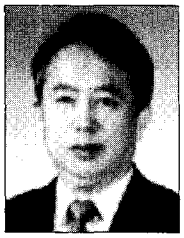
e-mail : biztech@ajou.ac.kr
1996년 아주대학교 산업공학과(학사)
2001년 아주대학교 산업공학과(석사)
2001년 현재 아주대학교 산업공학과 박사 과정
관심분야 : 정보시스템, 데이터분석, 컴퓨터 응용, 신뢰성

장 중 순

e-mail : jsjang@ajou.ac.kr
1979년 서울대학교 산업공학과(학사)
1981년 한국과학기술원 산업공학과(석사)
1986년 한국과학기술원 산업공학과(박사)
1986년~현재 아주대학교 산업정보시스템 공학부 교수
관심분야 : 정보시스템, 컴퓨터 응용, 데이터분석, 신뢰성관리

송 재 신

e-mail : song@keris.or.kr
1983년 충남대학교 전자공학과(학사)
1990년 원광대학교 전자공학과(석사)
2000년 아주대학교 컴퓨터공학과 박사 과정(수료)
1984년~1991년 중학교, 고등학교 교사
1991년~1997년 한국교육개발원 부연구위원
1997년~현재 한국교육학술정보원 연구위원
관심 분야 : e-Learning, 시스템 프로그램, 교육정보화 등



유 해 영

e-mail : yoohy@dankook.ac.kr
1979년 단국대학교 수학과(학사)
1981년 단국대학교 대학원 수학과(이학 석사)
1999년 아주대학교 컴퓨터공학과(공학 박사)
1983년~현재 단국대학교 정보 컴퓨터과학부 교수
관심분야 : 소프트웨어 공학, 시스템 프로그램 등

최 경 희

e-mail : khchoi@madang.ajou.ac.kr
1976년 서울대학교 수학교육과(학사)
1979년 프랑스 그랑데꼴 Enseeiht 대학 (석사)
1982년 프랑스 Paul Sabatier 대학 정보 공학부(박사)
1982년~현재 아주대학교 정보 및 컴퓨터공학부 교수
관심분야 : 운영 체제, 분산시스템, 실시간 및 멀티미디어 시스템 등