Journal of the Korean Data & Information Science Society
2013, **24**(3), 437–444

http://dx.doi.org/10.7465/jkdi.2013.24.3.437
한국데이터정보과학회지

주요 포털사이트의 응답패킷분포에 관한 연구

류귀열[1]

[1]서경대학교 컴퓨터과학과

접수 2013년 3월 5일, 수정 2013년 3월 31일, 게재확정 2013년 4월 15일

요 약

연구의 목적은 대한민국 주요 3개 포털사이트의 응답패킷 분포를 연구하는 것이다. 대상은 Naver, Daum, Nate 등 포털사이트의 대표페이지이며, 실험기간은 2010년 5월 19일에서 2012년 11월 7일까지이며, 실험횟수는 4,642회 실시하였다. 분포 형태는 응답패킷의 양이 많은 Naver, Nate는 양봉분포를 이루고 있으며, Daum은 오른쪽 꼬리가 긴 분포를 이루고 있다. 분포가 동일한가에 대한 검정은 카이제곱 검정, 이표본 콜모고로프-스미르노프 검정을 통하여 3개 포털의 분포가 유의수준 1%하에서 다르다는 사실을 밝혔다. 상대도수와 백분위 수의 비교를 통해 응답패킷의 분포를 비교하였다. 그 결과 응답패킷의 분포가 Naver가 가장 큰 값을 가졌으며, 다음은 Nate이었으며, Daum은 가장 작은 값을 가졌다. 응답속도를 높이기 위해서는 네트워크 속도를 높이는 것과 병행하여 페이지를 가볍게 만들어야 한다. 본 논문은 포털들이 페이지를 가볍게 만들게 하는 경쟁을 활성화하는데 기여하기를 기대한다.

주요용어: 응답패킷, 이표본 콜모고로프-스미르노프 검정, 카이제곱 검정, 포털사이트

1. 서론

우리나라에서 유선인터넷은 1982년 서울대와 한국전자기술연구소와의 시범망을 구축함으로써 시작되었다. 이 때 학술용도로 인터넷 네트워크를 이용하기 시작하여, 1986년에는 국가도메인인.kr을 도입하고 IP (internet Protocol) 배정을 시작하였다. 1994년에는 마우스만으로 정보를 접속할 수 있으면서 멀티미디어 기능을 가지는 브라우저인 네스케이프가 나오면서 일반 이용자들이 사용하기 시작하여 이 때 부터 인터넷이 급속도로 확산되기 시작하였다. 우리나라에서도 1994년에는 상용 인터넷 서비스 제공자 (internet service provider)인 한국통신, 데이콤, 아이네트 등이 텍스트 기반의 인터넷서비스를 제공하기 시작하였다. 1999년에는 하나로통신이 전화선에 주파수 대역을 나누어, 높은 주파수 대역에 데이터 서비스를 하고, 다운로드는 빠르게, 업로드는 늦게 제공하는 ADSL (asymmetric digital subscriber line) 서비스를 시작하였다. 이 때 부터 우리나라는 인터넷에 집중 투자하여 전송속도도 빠르게 높이고 인터넷 네트워크도 빠르게 확장시켜 나갔다. 한국인터넷진흥원 (2012)에 의하면 2001년 다운로드 속도가 초당 1 Mbps (Mbyte per second)를 넘는 초고속망 네트워크에서 세계 1위를 차지하였고, 인터넷 이용자도 2,000만명을 돌파하였다. 이 후 진보된 기술인 VDSL (very high-bit rate digital subscriber line), FTTH (fiber to the home), LAN (local area network) 등으로 발전하여 우리나라는 인터넷분야에서 세계에서도 유래를 찾아 볼 수 없을 정도로 빠르게 성장해 왔다.

인터넷에서 사용되는 HTTP (hypertext transfer protocol) 프로토콜은 FTP (file transfer protocol)에서의 파일전송의 비효율성을 줄이기 위해 제안된 것으로 지금은 널리 사용되고 있는 프로토콜이

[1] (136-704) 서울시 성북구 정릉동 산16-1, 서경대학교 컴퓨터과학과, 부교수. E-mail: gyryu@skuniv.ac.kr

다. Touch 등 (1996)에 의하면 HTTP에서의 데이터 수신은 먼저 클라이언트와 서버의 채널을 열기 위해 RTT (round-trip time) 시간이 소요된다. 이는 클라이언트가 서버에 채널을 요청 (TCP syn)하는데 소요되는 시간 0.5RTT와, 서버가 클라이언트에게 채널을 열어 (TCP syn+ack)주는데 소요되는 시간 0.5RTT이다. 채널이 열리면 클라이언트가 서버에 데이터를 요청하는데 소요되는 시간은 0.5RTT가 된다. 그 이 후 요청한 데이터를 수신하게 된다. HTTP에서 데이터를 전송하는 프로토콜은 Figure 1.1에 나와 있다.

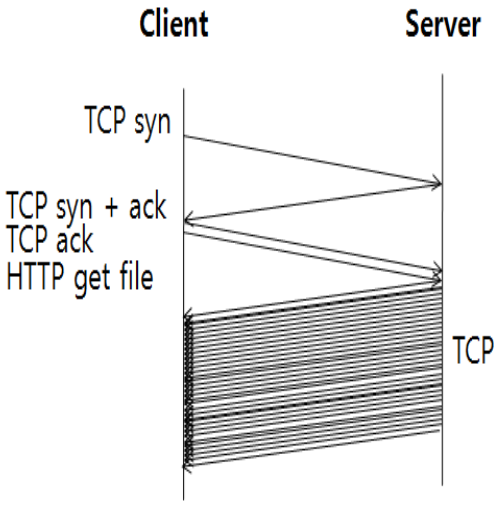

Figure 1.1 Protocol of data transmission in HTTP

포털 서비스는 이용자들이 포털 서버에 서비스를 요청하면 서버는 HTTP 프로토콜을 이용하여 정보를 패킷화하여 이용자들에게 전달하는 방식으로 이루어지고 있다. 따라서 이용자들은 포탈 사이트로부터 어느 정도 패킷을 받고 있는가가 기본적인 정보이지만 확인하기 매우 어렵다. 또한 다른 요인들이 같다는 가정 하에서 응답속도는 응답패킷에 반비례하므로, 이용자들은 이러한 정보들을 이용하여 어느 사이트가 상대적으로 빠른 응답을 할 것이라는 전망을 할 수 있다. 따라서 포탈 사이트들은 이용자들의 빠른 응답에 대한 욕구를 충족시키는 방법 중의 하나로 응답 패킷을 줄이려는 기술개발에 노력할 것이다. 이는 인터넷 서비스관련 기술개발로 이어질 것으로 예상된다. 연구 대상 사이트는 텍스트 기반의 단순 소개 페이지를 제공하는 Google을 제외하고, 현재 서비스 제공 중인 포털 사이트 중 방문자 수 기준으로 상위 3대 포탈 사이트인 Naver와 Daum, Nate를 선정하였다. 따라서 본 논문은 Naver와 Daum, Nate의 응답 패킷 분포를 비교 분석하는 것이 목적이다.

2. 연구설계

연구에 사용된 컴퓨터는 데스크탑 PC (personal computer)와 노트북 컴퓨터이다. 데스크탑 PC의 사양으로 CPU (central processing unit)는 인텔 2.33GHz, 메인메모리는 1.99Gbyte, OS (operating system)는 Windows XP이며, 노트북 PC의 사양으로 CPU는 셀러론 1.60GHz, 메인메모리는 1.2Gbyte, OS는 Windows XP이다. 본 실험에서 사용하는 브라우저는 브라우저가 제공하는 측정도구를 사용하기 위해 네스케이프 사에서 개발한 firefox를 사용하였으며, 패킷량 측정은 firebug에서 제공하는 네트워크 관리 툴을 사용하였다. Figure 2.1은 firebug가 패킷량을 측정하는 화면이다. 이 방법은

Ryu (2012, 2013)에서도 사용하였다. 실험은 2010년 5월 19일에서 2012년 11월 7일까지 2년 6개월 동안 실시되었으며, 실험횟수는 4,642회 실시하였다.

Figure 2.1 Response time using firebug

3. 6개 포털사이트들의 응답패킷 분포 비교

3.1. 이론적 배경

확률분포에서 주 관심분야는 Hong (2000), Song (2002), Hur (2011) 등에서 설명되어 있는 바와 같이 이론적인 분야이다. 본 논문은 확률분포를 응용하기 때문에 분포를 표현하는 것이 중요한 문제이다. 분포를 표현하기 위해서 일반적으로 Jiang 등 (2010)과 같이 확률밀도함수를 이용하는 방법과 Kim (2011)과 같이 누적 확률을 이용하는 방법이 있다. 우리는 확률밀도함수의 형태인 히스토그램 형태로 분포를 표현하고, 분포를 비교하기 위해서 누적확률을 이용하는 방법인 백분위수를 이용할 것이다. 두 모집단을 비교하는 방법은 Han 등 (2012), Lee 등 (2012), Han과 Kang (2012) 등에서 설명된 바와 같이 주로 평균을 비교하는 방법이다. 그러나 우리는 분포를 비교하는 연구로서 여기에는 모수적 방법과 비모수적 방법이 있다. 모수적 방법은 Press 등 (2002)에 설명되어 있는 것과 같이, 카이제곱 검정을 실시할 수 있다. 알려지지 않은 모집단의 분포 F와 G에서 독립적으로 얻은 랜덤 표본 $X_1, X_2, \cdots X_n$과 $Y_1, Y_2, \cdots Y_n$에 대하여 두 분포의 동일성을 검정하기 위한 카이제곱 검정통계량은

$$\chi^2 = \sum_{i=1}^{B} \frac{(R_i - S_i)^2}{R_i + S_i}.$$

여기서 B는 구간의 수이며, R_i와 S_i는 각각 $X_1, X_2, \cdots X_n$와 $Y_1, Y_2, \cdots Y_n$중에서 i번째 구간에 속하는 자료의 개수이다. 이는 두 모집단의 분포가 동일하다면 근사적으로 카이제곱 분포를 한다는 사실에 기초하고 있다. 비모수적 검정은 Kim과 Oh (2003), Hong과 Kim (2003) 등에서 설명되어 있는 것과 같이 이표본 콜모고로프-스미르노프 (Kolmogov-Smirnov) 검정을 사용할 수 있다. 이표본 콜모고로프 스미르노프 검정통계량은

$$D_{n,m} = \sup_x |F_{1,n}(x) - G_{2,m}(x)|.$$

여기서 $F_{1,n}(x)$와 $G_{2,m}(x)$는 표본의 수가 n개와 m개인 두 모집단 분포의 경험적 누적 분포함수이다. $Z = D_{n,m}\sqrt{\frac{nm}{n+m}} > K_\alpha$이면 두 분포가 동일하다는 귀무가설은 기각된다. 여기서 K_α는 유의수준 α하에서 귀무가설을 기각할 수 있는 기각값이다.

3.2. 연구결과

응답패킷 분포의 수평축에는 0Kbyte에서 1Mbyte까지 사용하였으며, 1Mbyte이상은 한 범주로 표현하였다. 왜냐하면 이상의 범주는 대체로 비율은 낮지만 범위가 넓어 분포를 비교하기 어렵기 때문이다. 수직축에는 범주에 해당되는 비율을 사용하였다. 3개 포털사이트의 응답패킷분포는 Figure 3.1에서 Figure 3.3에 나와 있다.

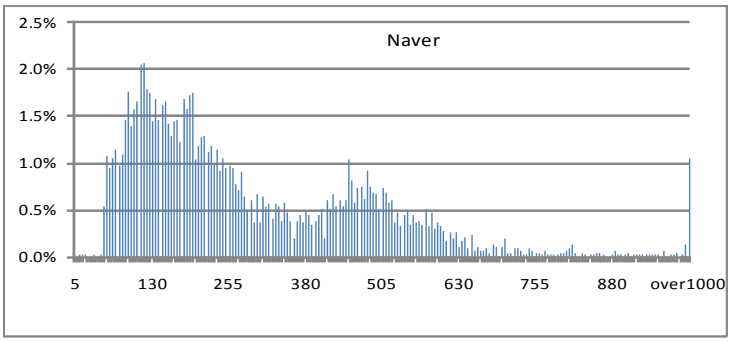

Figure 3.1 Distribution of response packets for Naver

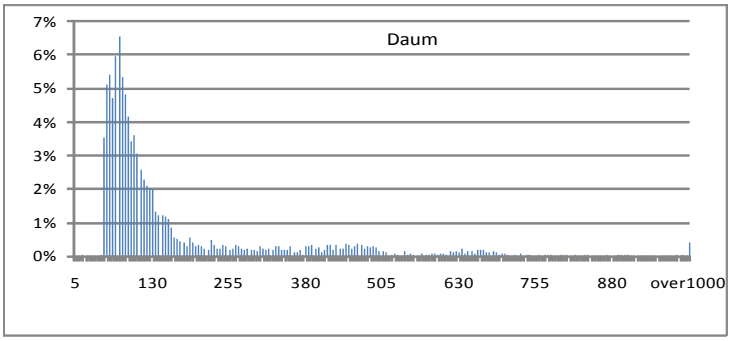

Figure 3.2 Distribution of response packets for Daum

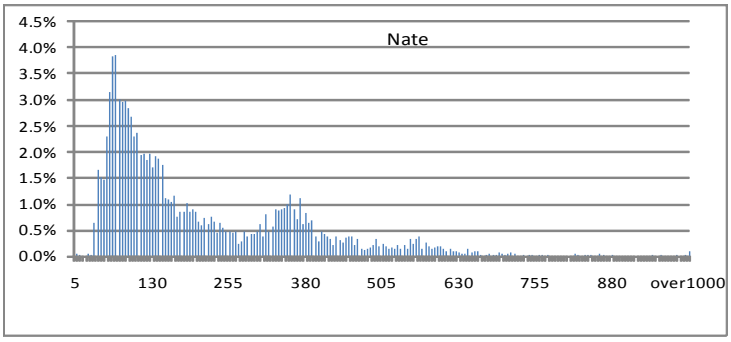

Figure 3.3 Distribution of response packets for Nate

Naver의 응답패킷의 분포는 Figure 3.1에서 보듯이 양봉분포이며 1Mbyte가 넘는 경우도 1.06%로 높은 비율을 보이고 있다. 양봉분포는 이질적인 두 분포가 합쳐진 형태로, Naver의 응답 패킷은 패킷량이 작은 분포와 많은 분포로 나뉘어질 수 있다는 것을 알 수 있다. 평균은 304.8Kbyte, 표준편차는 328.8Kbyte, 최소값은 7.9Kbyte, 최대값은 5.846Mbyte로 응답패킷의 분포가 넓게 분포되어 있다. Daum의 응답패킷의 분포는 Figure 3.2에서 보듯이 오른쪽 꼬리가 긴 분포를 하고 있다. 1Mbyte가 넘는 비율은 0.41%를 보이고 있다. 평균은 174.5Kbyte, 표준편차는 226.6Kbyte, 최소값은 1.6Kbyte, 최대값은 6.229Mbyte로 응답패킷의 분포가 넓게 분포되어 있다. Nate의 응답패킷의 분포는 Figure 3.3에서 보듯이 양봉분포로 Naver와 유사하며, 1Mbyte가 넘는 경우도 0.108%을 보이고 있다. 평균은 200.6Kbyte, 표준편차는 165.5Kbyte, 최소값은 1.9Kbyte, 최대값은 4.369Mbyte로 응답패킷의 분포가 넓게 분포되어 있다. Yahoo의 응답패킷의 분포는 Figure 3.6에서 보듯이 오른쪽 꼬리가 긴 분포를 하고 있다. 1Mbyte가 넘는 비율은 0.09%로 낮은 비율을 보이고 있다. 평균은 203.1byte, 표준편차는 226.6Kbyte, 최소값은 1.9Kbyte, 최대값은 4.369Mbyte로 넓게 분포되어 있다.

응답패킷의 양이 많은 Naver, Nate는 접속날짜에 따라 응답패킷이 많은 날과 적은 날로 구분되고, 이는 이질적인 두 개의 분포가 붙어 있는 형태를 보이는 양봉분포의 형태로 나타났다. Daum은 오른쪽으로 꼬리가 긴 분포를 이루고 있는데, 이는 응답패킷이 클 확률이 급속도로 낮아진다는 사실을 보여 준다.

3개 포털들의 기본통계량들은 Table 3.1에 나와 있다. 평균 응답패킷을 기준으로 Naver가 가장 무겁고 다음으로 Nate, Daum 순이다. Naver의 응답패킷은 Daum의 응답패킷의 평균 1.75배, Nate의 응답패킷의 평균 1.5배임을 알 수 있다. 따라서 다른 요인이 동일하다면 응답속도는 Daum이 가장 빠르고 그 다음이 Nate이며 Naver가 가장 느릴 것이라 추정할 수 있다.

Table 3.1 Descriptive statistics for 3 portals (unit: Kbyte)

	Iteration	Average	Standard deviation	Min	Max
Naver	4,642	307.440	359.700	7.9	5,846.0
Daum	4,642	153.151	168.174	1.6	4,000.0
Nate	4,642	200.593	165.525	1.9	4,368.6

포털사이트의 분포를 비교하기 위해, 먼저 모수적 방법인 카이제곱 검정을 실시하였다. 카이제곱 검정을 위해 구간을 31개로 분할하였다. 검정결과는 Table 3.2에 나와 있다. 카이제곱 검정에 따르면 6개 포털들의 응답 패킷량의 분포는 유의수준 1%하에서 모두 다르다는 사실을 알 수 있다. 본 검정에서 기대빈도가 5이하인 범주가 10%이하이다.

Table 3.2 Chi-square test

	Daum	Nate
Naver	2116.6**	1115.7**
Daum		265.8**

* means we can reject H0 under α =0.05.

** means we can reject H0 under α =0.01.

비모수적 검정을 위해 이표본 콜모고로프-스미르노프 검정을 실시하였다. 검정결과는 Table 3.3에 나와 있다. 이표본 콜모고로프-스미르노프 검정에 따르면 3개 포털들의 응답 패킷량의 분포는 유의수준 1%하에서 모두 다르다는 사실을 알 수 있다. 따라서 모수적 방법과 비모수적 방법 모두 3개 포털들의 응답패킷 분포가 다르다는 사실을 알 수 있다.

Table 3.3 Two sample Komogorov-Smirnov test statistics

	Daum	Nate
Naver	20.8**	11.6**
Daum		9.59**

* means we can reject H0 under α =0.05.
** means we can reject H0 under α =0.01.

다음으로 응답패킷 양의 많고 적음을 평가하기 위해, 첫 번째 방법으로 동 시간 대 수신한 패킷의 양을 직접비교하여 패킷 양이 많을 비율로 추정할 수 있다. 결과는 Table 3.4에 나와 있다. 결과를 보면 Naver가 Daum보다 응답패킷이 많을 비율은 79.77%, Nate에 70.10%이므로 Naver의 응답패킷이 Daum과 Nate보다 많음을 알 수 있다. Daum이 Nate보다 응답패킷이 많을 비율은 42.01%이므로 Nate가 Daum보다 많음을 알 수 있다. 따라서 응답이 많을 비율 기준으로 보면 Naver, Nate, Daum 순으로 응답패킷 양이 많음을 알 수 있다.

Table 3.4 Proportions that A is greater than B using response packets

$A \diagdown B$	Daum	Nate
Naver	79.77	70.1%
Daum		42.01%

두 번째 방법으로 포털들의 응답패킷의 백분위수를 비교하였다. 결과는 Table 3.5에 나와 있다. 백분위 수 기준으로 Naver의 백분위 수는 최댓값을 제외한 나머지 1분위 수에서 99분위 수에서 Daum의 백분위 수 보다 크며, Nate에 대해서는 모든 백분위 수에서 크다. Daum의 백분위 수는 Nate의 백분위 수의 19% 만이 크기 때문에 Nate가 Daum보다 백분위 수 기준으로 클 비율은 81%로 Nate가 Daum보다 크다는 사실을 알 수 있다. 따라서 직접비교에 의한 크기 비교에서나 백분위 수 비교나 두 가지 방법으로 본다면 응답패킷의 양은 Naver, Nate, Daum 순으로 유의하게 크다는 사실을 알 수 있다.

Table 3.5 Proportions that A is greater than B using percentiles

$A \diagdown B$	Daum	Nate
Naver	99%	100%
Daum		19%

4. 결론

본 연구의 목적은 대한민국 주요 3개 포털사이트의 응답패킷분포를 연구하는 것이다. 실험 대상은 포털사이트의 대표페이지이며, 실험기간은 사이트의 패킷이 사계절을 포함하고 많은 자료의 수집을 위해 2010년 5월 19일에서 2012년 11월 7일까지 2년 6개월 동안 실시되었으며, 실험횟수는 4,642회 실시하였다.

분포 형태는 응답패킷의 양이 많은 Naver, Nate는 양봉분포를 이루고 있으며, 이는 접속날짜에 따라 응답패킷이 많은 날과 적은 날로 구분되어 이질적인 두 개의 분포가 붙어 있는 형태를 보였다. Daum은 오른쪽으로 꼬리가 긴 분포를 이루고 있으며 이는 응답패킷이 큰 경우의 확률이 급격히 낮아지는 형태를 보이고 있다.

응답패킷 양의 분포비교는 모수적 방법으로 카이제곱 검정, 비모수적 방법으로 이표본 콜모고로프-스미르노프 검정을 실시하였다. 두 방법 모두 유의수준 1% 하에서 포털들의 응답패킷의 분포가 다르다

는 사실을 알 수 있었다. 또한 분포의 비교를 위해 직접비교와 백분위 수의 비교를 실시하였다. 그 결과 Naver의 응답패킷의 분포가 가장 컸으며 다음으로 Nate, Daum 순이었다.

본 논문은 Naver와 Daum, Nate의 대표페이지의 패킷의 분포를 밝히고 비교하였다. 응답속도를 높이기 위해서는 대표 페이지를 최대한 가볍게 만들어야 한다. Naver의 응답패킷은 Daum의 응답패킷의 평균 1.75배, nate의 응답패킷의 평균 1.5배로 나타나 다른 조건이 동일하다면 Naver의 응답속도가 가장 느리고 다음으로 Nate이고 Daum의 응답속도가 가장 빠를 것으로 예상된다. 이는 포털 사이트의 특성 상 대표 페이지에는 광고 등 이해당사자들의 관계를 고려하여 패킷 양을 조절하는데 한계가 있음을 반영하는 현상이라고 생각된다.

그러나 연구를 통해 같은 페이지라 할지라도 응답패킷의 양이 클 경우 반복 실험 시 패킷의 변화량이 크고 양이 작을 경우 패킷의 변화량이 적다는 사실을 알 수 있었다. 후 속 연구를 통해 이러한 현상을 규명할 필요가 있다. 또한 어떤 콘텐츠가 데이터 량에 많은 영향을 주는지를 밝힘으로써 동일한 콘텐츠라도 가볍게 만들 수 있는 기술을 확보할 수 있으며 응답속도를 높일 수 있다. 또한 포털사이트들의 응답속도가 이용자들의 만족도에 많은 영향을 주기 때문에 응답속도에 관한 후속 연구도 필요하다. 무선인터넷의 요금은 주로 종량제이므로 무선인터넷 포탈의 응답패킷 연구는 이용자들의 효율적인 이용에 도움을 줄 것으로 기대된다. 이러한 후속 연구들은 포탈들의 기술경쟁을 촉진시켜 더욱 효율적인 페이지를 구축하도록 유도할 것으로 기대된다.

References

Han, D. and Kang, M. (2012). Study on application of information and communication technology in special education. *Journal of the Korean Data & Information Science Society*, **23**, 927-937.

Han, J., Lee, K. and Yang, J. (2012). The effects of the 16-weeks' combined exercise program on metabolic syndrome and autonomic nerve system of low-level physical strength group. *Journal of the Korean Data & Information Science Society*, **23**, 895-904.

Hong, J. (2000). *Statistical probability distribution*, Freeacademy, Kyeonggi.

Hong, J. and Kim, J. (2009). Nonparametric homogeneity tests of two distributions for credit rating model validation. *Journal of the Korean Data & Information Science Society*, **20**, 261-272.

Hur, M. (2011). *Mathematical statistic*, Parkyoungsa, Seoul.

Jiang, Z., Kim, J. and Yoon, S. (2010). The analysis of inequality of firm size distribution. *Journal of the Korean Data Analysis Society*, **12**, 2185-2194.

Kim, D. and Oh, K. (2003). On the equality of two distributions based on nonparametric kernel density estimator. *Journal of the Korean Data & Information Science Society*, **14**, 247-255.

Kim, T. (2011). Visualization analysis of small and medium enterprises survival distribution after IMF. *Journal of the Korean Data Analysis Society*, **13**, 441-455.

Korea Internet & Security Agency (2012). *2011 Korea internet white paper*, Korea Internet & Security Agency, Kyeonggi.

Lee, S., Cho, Y. and Yang, J. (2012). A study on the effects of a 12-week compound exercise program on obese middle school girls' leptin and insulin levels *Journal of the Korean Data & Information Science Society*, **23**, 787-796.

Press, W., Teukoisky, S., Vettering, W. and Flannery, B. (2002).*Numerical recipes in C*, Cambridge University Press, Cambridge.

Ryu, G. (2012). A study on response time of WiBro depending on signal intensity. *Journal of the Korean Data Analysis Society*, **14**, 1119-1128.

Ryu, G. (2103). A study on comparing response times between Wibro and wired internet using portals. *Journal of the Korean Data & Information Science Society*, **24**, 23-32.

Song, M. (2002). *Mathematical statistics*, Parkyoungsa, Seoul.

Touch, J., Heidemann, J. and Obraczka, K. (1996). *Analysis of HTTP performance*, USC/ISI Research Report 98-463, 1-10.

Journal of the Korean Data & Information Science Society
2013, **24**(3), 437–444

http://dx.doi.org/10.7465/jkdi.2013.24.3.437
한국데이터정보과학회지

A study on distribution comparison of response packets for major portal sites

Gui-Yeol Ryu[1]

[1]Department of Computer Science, SeoKyeong University

Received 5 March 2013, revised 31 March 2013, accepted 15 April 2013

Abstract

The object of study is to verify the distributions of response packets of 3 portal sites such as Naver, Daum, Nate. The period of experiments is from May 19th 2010 to November 7th 2012 and the number of experiments is 4,642. The distributions of Naver, Nate are biomodals. The distribution of Daum has long right tails. 3 distributions are different under 1% significance level using chi-square test and two sample Kolmogorov-Smirnov test. From proportions and percentiles, Naver has a distribution with the largest values. Nate is the second place, and Daum has a distribution with the smallest values. We must make portal pages light to increase response speed including other technologies. We expect our results to activate competition among portal sites.

Keywords: Chi-square test, portal site, response packets, two sample Kolmogov-Smirnov test.

[1] Associate professor, Department of Computer Science, SeoKyeong University, Seoul 136-704, Korea. E-mail: gyryu@skuniv.ac.kr