응용통계연구 (2009)
22(3), 553–568

인터넷웹상의 숫자들과 벤포드법칙

장대흥[1]

[1]부경대학교 수리과학부 통계학전공

(2009년 2월 접수, 2009년 3월 채택)

요 약

인터넷 상의 웹페이지에 나타나는 숫자들의 빈도수를 조사한 후 이러한 숫자들이 이루는 집합체의 성질을 알아보고 이러한 자료들이 각 종 법칙들(거듭곱 법칙, 지프 법칙, 벤포드 법칙)이 성립하는 지를 살펴보았다.

주요용어: WWW, 거듭곱 법칙, 지프 법칙, 벤포드 법칙.

1. 서론

인터넷에서 웹(WWW)의 구조와 특징에 대하여서는 많은 학자들이 10여년 전부터 연구를 하여 오고 있다 (Huberman 등, 1998; Albert 등, 1999; Barabàsi와 Albert, 1999; Huberman과 Adamic, 1999). 검색엔진을 통하여 우리가 만나는 웹페이지에는 글자, 숫자 뿐만이 아니라 그림이나 동영상도 포함하는 멀티미디어 정보를 포함하고 있다. 웹을 이해하기 위한 독특한 접근 방법으로서 Dorogovtsev 등 (2006)은 웹페이지에 나타나는 숫자의 빈도수를 분석함으로써 웹의 구조를 이해하고자 하는 시도가 있었다. 우리는 웹을 거대한 숫자들의 집합체로 볼 수 있다. 우리들이 사회생활을 영위하며 만들어내고 사용하는 숫자들을 웹페이지를 통하여 우리들은 서로에게 전달하고 있는 것이다. 이러한 숫자들의 집합체가 어떤 통계적 성질을 가지며 이러한 숫자들의 집합체에서 어떤 법칙들(예로 거듭곱 법칙, 지프법칙, 벤포드 법칙)을 발견할 수 있는 지를 밝히는 작업은 의미가 있는 작업이 될 수 있을 것이다. 이러한 작업을 통하여 서로 다른 사회공동체가 서로 다른 문화생활을 영위하며 만들어내는 숫자들의 동적구조가 서로 비슷한 구조를 갖는 지, 다르다면 왜 다른 지를 밝힐 수 있는 단서를 얻게 된다. Dorogovtsev 등 (2006)은 2004년 12월 검색엔진 google을 이용하여 웹페이지에 나타나는 숫자들의 빈도수를 분석함으로써 인간들이 만들어내는 이 거대한 숫자들의 집합체가 어떤 시스템을 형성하고 있는 지를 귀납법적으로 밝혔다. 본 논문을 통하여 Dorogovtsev 등 (2006)이 주장하는 웹 시스템의 특징이 우리나라 웹 DB에도 적용되는 지를 살펴봄으로써 이러한 웹 시스템의 특징이 전세계에서 통용될 수 있는 보편적인 특징인 지를 알아보고자 한다. 우리나라 대표적인 포털사이트 naver를 대상으로 하였다. 2009년 2월 6일 조사에 착수하여 1998년 1월 1일부터 2008년 12월 31일까지의 naver 웹페이지를 대상으로 웹페이지에 나타나는 숫자들의 빈도수를 측정하였다. 측정 시 naver 뉴스 검색 기능을 이용하였다. Dorogovtsev 등 (2006)이 한 것처럼 연속적인 자연수 수열을 선정하여 이 자연수에 대응하여 이 자연수가 나타나는 웹페이지수를 측정하는 방법을 사용하였다. 자연수로 1에서 2200까지를 선정하여 조사하였

이 논문은 2007학년도 부경대학교 연구년 교수 지원사업에 의하여 연구되었음(PS-2007-009).
[1](608-737) 부산광역시 남구 대연3동 599-1 부경대학교 수리과학부 통계학전공, 교수.
E-mail: dhjang@pknu.ac.kr

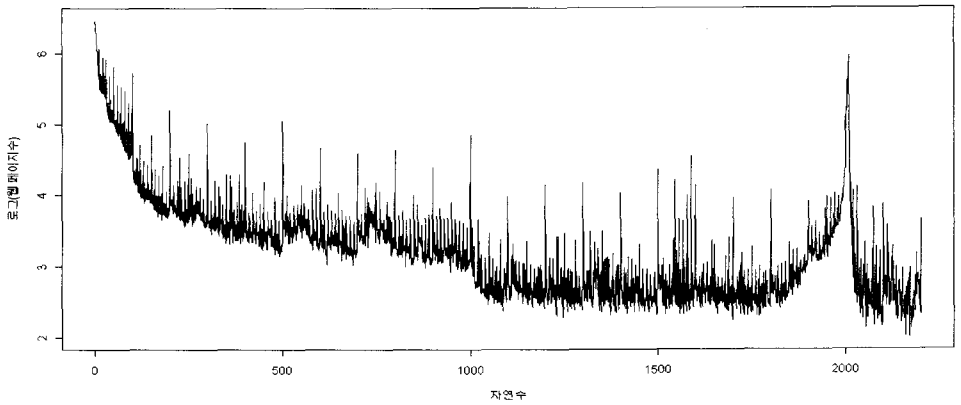

그림 **2.1.** 자연수 1–2200에 대응하는 로그(웹페이지수)를 나타내는 꺾은선그래프

다. 저자의 생각에 자연수 1–1000을 조사하면 대략 인터넷웹상의 숫자들의 특징을 알 수 있다고 보았고 캘린더의 연도와 관련한 자연수는 1000에서 2200이면 대략적인 특징을 알 수 있다고 생각하여 1에서 2200까지 고려하였다. 포털사이트의 웹 DB에서는 일정기간이 지난 오래된 웹페이지는 사라지고 새로운 웹페이지가 계속 축적됨으로 이 웹 DB는 항상 가변적인 시스템이 된다. 그래서 며칠만 지나도 측정 데이터가 변하는 이러한 속성 때문에 조사를 2009년 2월 6일 하루로 한정하여 조사하였다. 2절에서는 이렇게 구한 자료를 이용하여 자료분석을 시도하였고 3절에서 결론 및 추후의 과제에 대하여 언급하였다.

2. 인터넷 웹상의 숫자들에 대한 자료분석

검색엔진을 이용하여 자연수 1에서 2200을 대상으로 해당 자연수가 나타나는 웹페이지수를 측정하여 얻는 자료에서 웹페이지수에 로그를 취한 후 자연수에 대한 로그(웹페이지수) 꺾은선그래프를 그리니 다음 그림 2.1과 같았다. 자연수 100, 1000을 경계로 구조적인 차이가 보이고 자연수 1000을 넘어서면 자연수 2008에서 비대칭적인 임계값(critical value)이 형성됨을 알 수 있다. 이는 자연수가 1000을 넘으면 웹페이지에 나타나는 자연수들이 캘린더의 연도와 밀접한 관계를 갖는다는 것을 암시한다.

1000보다 작은 자연수 부분에 대하여 더 자세히 알기 위하여 자연수에 로그를 취하여 로그(자연수)에 대한 로그(웹페이지수) 로그-로그 그래프를 그리니 다음 그림 2.2와 같았다. 자연수 10을 경계로는 뚜렷한 구조적인 차이가 보이지 않으나 자연수 100, 1000을 경계로 구조적인 차이가 있음을 알 수 있다.

우리는 자연수 1–2200과 이에 대응하는 웹페이지수 자료가 거듭곱 법칙을 따르는 지를 조사할 필요가 있다. 거듭곱 법칙(power law)은 다음과 같이 정의된다.

$$F = \alpha N^{-\beta},$$

여기서, N은 자연수, F는 자연수에 대응되는 웹페이지수이고 α와 β는 상수이다. 우리는 우리의 일상생활이나 과학세계에서 거듭곱 법칙의 활용 예들을 자주 접하게 된다. 예로 스테판-볼츠만 법칙, 곰페르츠 사망법칙, 램버그-오스굿 스트레스-긴장 관계, 프락탈, 파레토법칙 등이 있다. 거듭곱법칙 분포로서는 파레토분포, 제타분포, 율-사이먼분포 등이 있다.

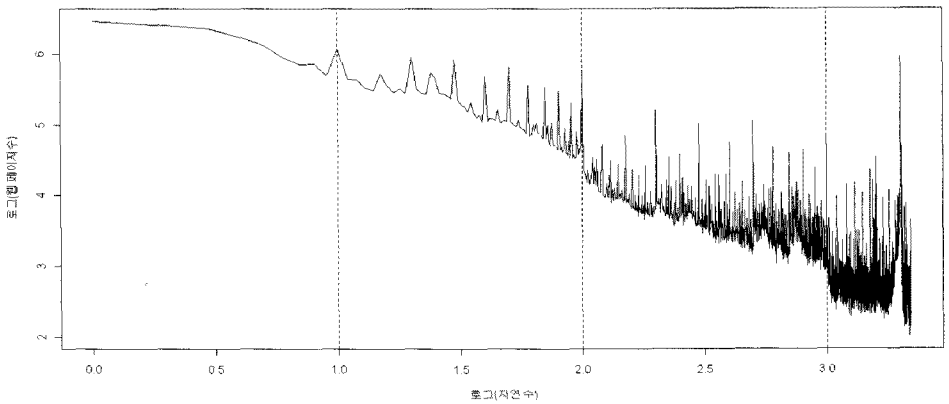

그림 **2.2.** 자연수 1-2200를 대상으로 로그(자연수)에 따른 로그(웹페이지수)를 나타내는 로그-로그 그래프

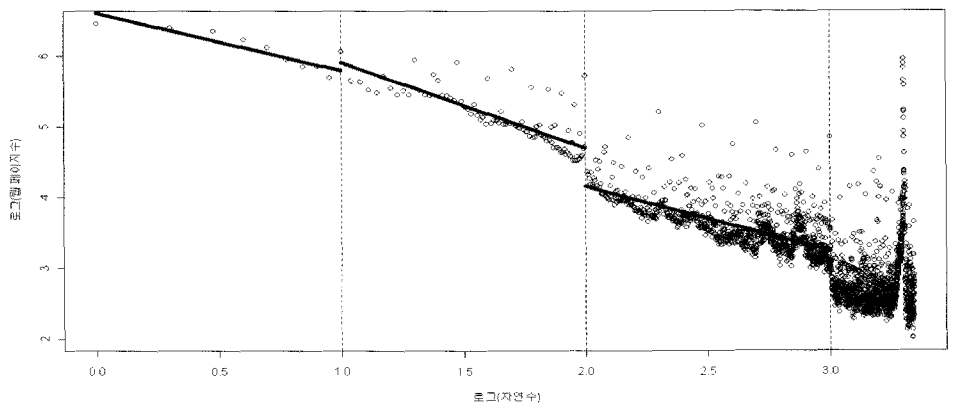

그림 **2.3.** 자연수 1-2200에 대응하는 로그(웹페이지수)를 나타내는 산점도와 추정회귀직선(1)

자연수 1-2200과 이에 대응하는 웹페이지수 자료가 거듭곱 법칙을 따르는 지를 조사하기 위하여 자연수 10, 100, 1000을 경계로 직선회귀식들을 차례로 구해보니 첫째로 자연수 1에서 9까지는 $\log_{10} F = 6.611 - 0.822 \log_{10} N$이 나왔다($R^2 = 0.85$, p-값 < 0.001). 정리하면 $F = 4083194/N^{0.822}$이 된다. 둘째로 자연수 10에서 99까지는 $\log_{10} F = 7.125 - 1.215 \log_{10} N$이 나왔다($R^2 = 0.68$, p-값 < 0.001). 정리하면 $F = 13335214/N^{1.215}$이 된다. 셋째로 자연수 100에서 999까지는 $\log_{10} F = 5.978 - 0.914 \log_{10} N$이 나왔다($R^2 = 0.47$, p-값 < 0.001). 정리하면 $F = 950605/N^{0.914}$이 된다. 다음 그림 2.3은 앞에서 구한 세 개의 직선회귀식들을 자료 위에 두꺼운 직선으로 표시한 그림이다.

이 차이를 무시하고 자연수 1에서 999까지 하나의 직선회귀식을 찾으니 $\log_{10} F = 6.967 - 1.274 \log_{10} N$이 나왔다($R^2 = 0.80$, p-값 < 0.001). 정리하면 $F = 9268298/N^{1.274}$이 된다. 다음 그림 2.4는 앞에서 구한 직선회귀식을 자료 위에 두꺼운 직선으로 표시한 그림이다. 자연수 1에서 1000까지 하나의 직선회귀식을 구하여도 앞에서 구한 직선식과 큰 차이가 없었다.

만일 자연수 1에서 2200 전체에 대하여 구조적인 차이를 무시하고 하나의 직선회귀식을 구하니 $\log_{10} F = 7.028 - 1.317 \log_{10} N$이 나왔다($R^2 = 0.66$, p-값 < 0.001). 정리하면 $F = 10665961/N^{1.317}$

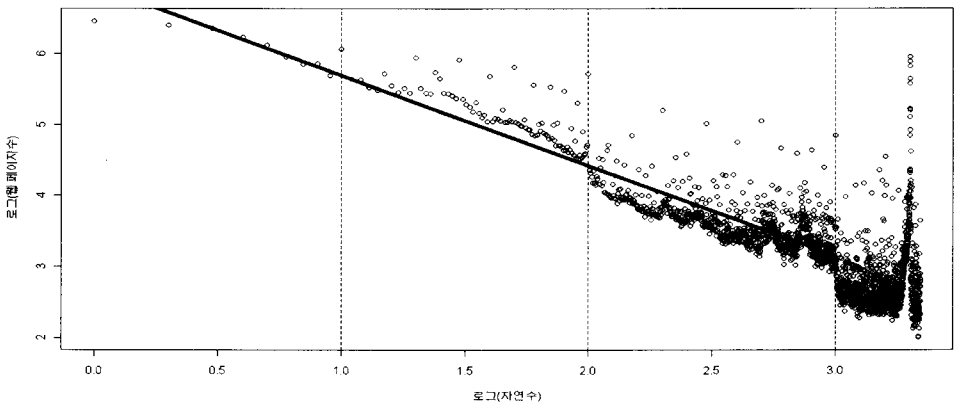

그림 2.4. 자연수 1–2200에 대응하는 로그(웹페이지수)를 나타내는 산점도와 추정회귀직선(2)

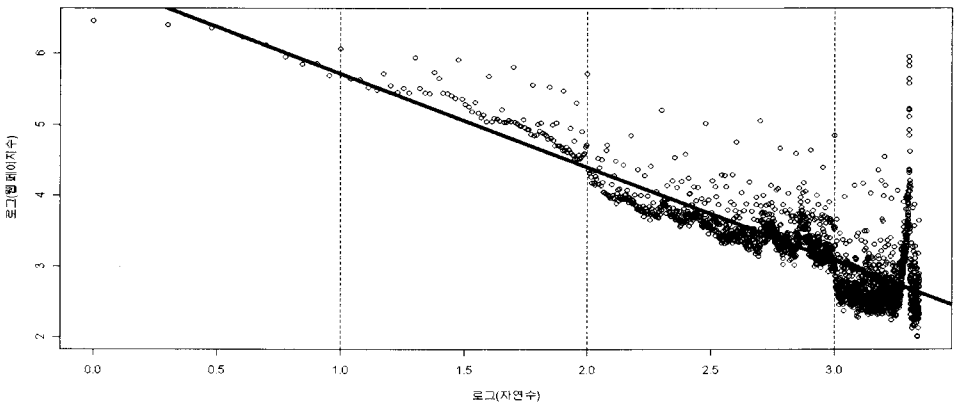

그림 2.5. 자연수 1–2200에 대응하는 로그(웹페이지수)를 나타내는 산점도와 추정회귀직선(3)

이 된다. 자연수 1에서 999까지 하나의 직선회귀식을 구한 결과와 아주 큰 차이는 없음을 알 수 있다. 다음 그림 2.5는 앞에서 구한 직선회귀식을 자료 위에 두꺼운 직선으로 표시한 그림이다.

이상을 정리하면 자연수 1–2200과 이에 대응하는 웹페이지수 자료는 거듭곱 법칙을 따름을 알 수 있고, 또한 상수 β가 한 개가 아니고 여러 개가 존재하는 복잡한 시스템인 것을 알 수 있다.

앞의 논의에서 자연수 1000을 넘어서면 자연수 2008에서 비대칭적인 임계값이 형성되는 데 웹페이지에 나타나는 자연수들이 캘린더의 연도와 밀접한 관계를 갖는다는 것을 암시한다고 언급하였다. 다음 그림 2.6은 자연수 1801–2199에 따른 로그(웹페이지수)를 그린 꺾은선그래프이다. 2008년에서 최대 웹페이지수가 나타남을 알 수 있다. 2008년(현재)를 중심으로 2008년 전(과거)과 2008년 후(미래)의 패턴이 비대칭적일 뿐만이 아니라 2008년 직전(2007년)에 웹페이지수가 증가하는 속도보다 2008년 직후(2009년)에 웹페이지수가 감소하는 속도가 더 빠르다. 또한 2008년 전(과거) 웹페이지수가 2008년 후(미래) 웹페이지수보다 평균적으로 많다. 포털사이트에 쌓이는 웹페이지 자료가 2008년 기준으로 과거 자료가 미래 자료보다 더 많다는 이야기가 된다.

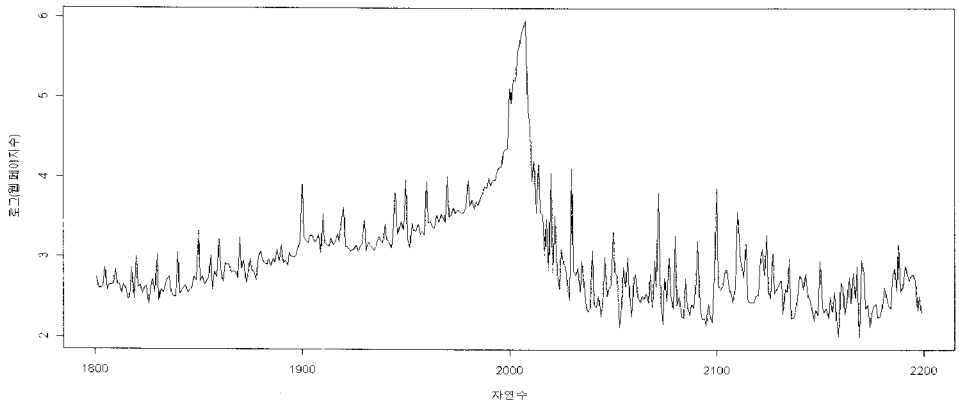

그림 **2.6.** 자연수 1801–2199에 따른 로그(웹페이지수)를 나타내는 꺾은선그래프

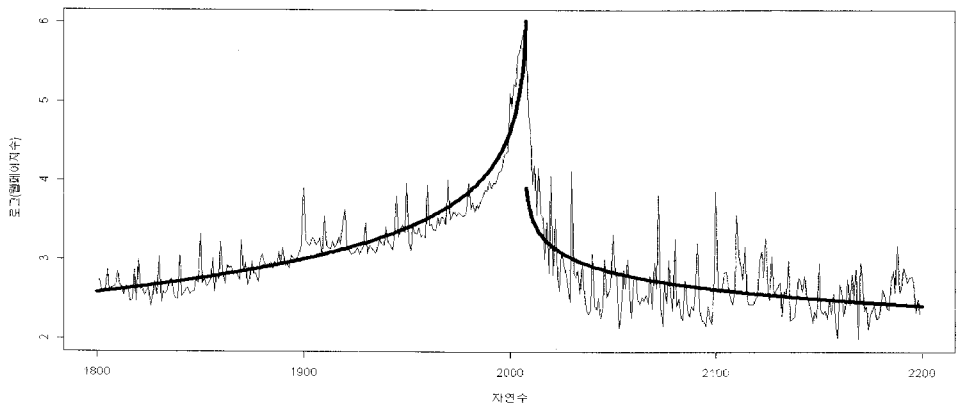

그림 **2.7.** 자연수 1801–2199에 따른 로그(웹페이지수)를 나타내는 꺾은선그래프와 추정회귀선

자연수 2008을 중심으로 비대칭적인 패턴이 나타나므로 2008년 이전과 2008년 이후로 나누어 거듭곱 법칙이 성립하는 지를 조사하기 위하여 2008년 이전에는 함수식을 $F = \alpha(2009 - N)^{-\beta}$로, 2008년 이후에는 함수식을 $F = \alpha(N - 2007)^{-\beta}$로 잡고 자료에 적합시켜 보았다. 2008년 이전에는 $F = 1073989(2009 - N)^{-1.484}$이 되고($R^2 = 0.90$, p-값 < 0.001), 2008년 이후에는 $F = 8091(N - 2007)^{-0.651}$이 되었다($R^2 = 0.30$, p-값 < 0.001). 다음 그림 2.7은 꺾은선그래프 위에 두 개의 적합 회귀식을 나타낸 그림이다.

우리는 앞에서 자연수 1–2200과 이에 대응하는 웹페이지수 자료는 거듭곱 법칙을 따름을 알 수 있고, 또한 상수 β가 한 개가 아니고 여러 개가 존재하는 복잡한 시스템인 것을 알 수 있다고 하였다. 1000과 2200 사이의 자연수에 대해서도 거듭곱 법칙을 따름을 알 수 있었고, 하나의 임계값을 중심으로 좌우가 거듭곱 법칙이 다르게 적용되는 것을 확인하였다. 한 마디로 정리하면 자연수 1–2200과 이에 대응하는 웹페이지수 자료는 거듭곱 법칙을 만족하는 시스템으로 크게 보면 상수 β가 한 개인 시스템으로 볼 수도 있으나 좀 더 들여다보면 상수 β가 여러 개(5개 이상)가 존재하는 복잡한 시스템인 것을 알 수 있다.

자연수 1–2200과 이에 대응하는 웹페이지수 자료가 지프 법칙(Zipf's law)을 따르는 지를 보기 위하여

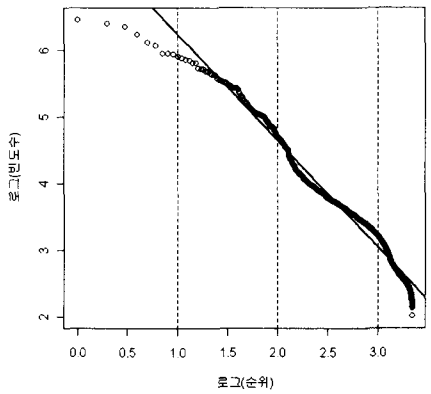

그림 **2.8.** 로그(순위)와 로그(빈도수)를 나타내는 산점도와 추정회귀직선

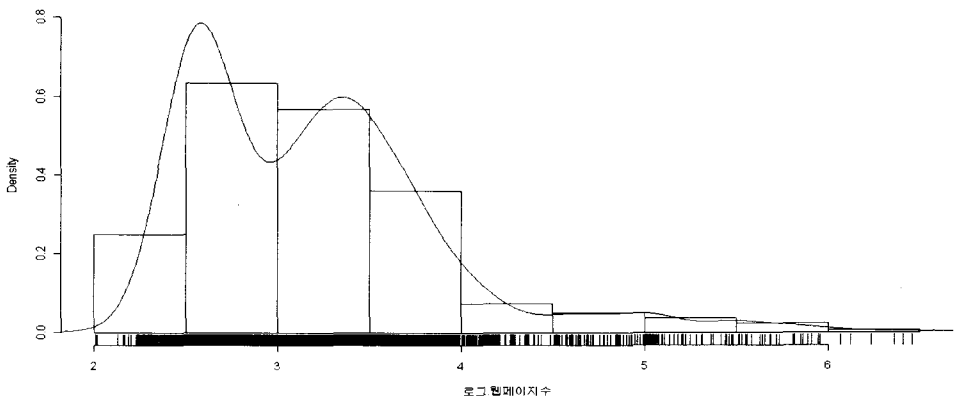

그림 **2.9.** 로그(웹페이지수)에 대한 히스토그램과 커널밀도추정량

다음과 같은 지프 법칙에 관한 수식을 정의하자.

$$F = \frac{K}{R^{\beta}},$$

여기서, R은 내림차순 순위, F는 빈도수이고 K는 $N = 2{,}200$(2,200개의 자연수 숫자), $M = 40{,}733{,}062$(자연수 2,200개에 대응하는 총웹페이지수)가 주어졌을 때의 상수이다. 그리고 $\beta \geq 1$이다. 지프 법칙은 임의의 문서에서 사용된 언어의 빈도수와 이 언어의 빈도수에 대응되는 순위와의 관계를 나타내는 법칙이다.

자연수 1–2200과 이에 대응하는 웹페이지수를 이용하여 로그(순위)에 대한 로그(빈도수)를 구한 후 산점도를 그리면 다음 그림 2.8과 같다. 지프의 법칙이 성립하는 지를 보기 위하여 직선회귀식을 구하니 $\log_{10} F = 7.850 - 1.599 \log_{10} R$이 나왔다($R^2 = 0.97$, p-값 < 0.001). 추정된 직선회귀선 양쪽 끝에서 적합이 잘 맞지 않는 경향이 있으나 전체적으로는 추론에 큰 문제가 없다. 추정된 직선회귀식을 정리하면 $F = 70794578/R^{1.599}$이 되어 자연수 1–2200과 이에 대응하는 웹페이지수 자료는 지프 법칙을 따른다고 볼 수 있다.

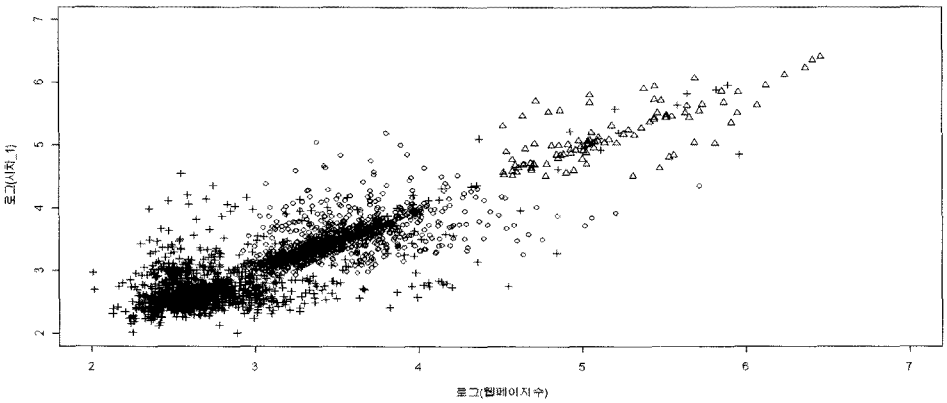

그림 2.10. 시차도(1)

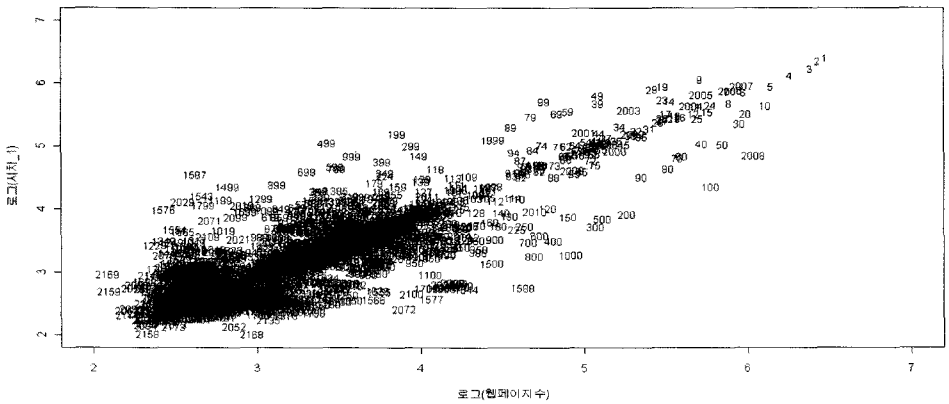

그림 2.11. 시차도(2)

자연수 1–2200에 대응하는 로그(웹페이지수)에 대한 히스토그램과 커널밀도추정량을 그려 보면 다음 그림 2.9와 같다. 두 개의 그룹이 현저하게 나타나는 데, 많은 자연수들이 웹페이지수가 100에서 1,000 사이인 그룹과 웹페이지수가 1,000에서 10,000 사이인 그룹으로 나뉜다. 또한 값이 큰 특이값들이 꽤 많아 오른쪽으로 매우 치우친 분포를 이룸을 알 수 있다(웹페이지수에 로그함수를 적용했는데도 대칭이 되지 않을 정도로).

이러한 두 개의 그룹을 더 확연하게 보기 위하여 자연수 1–2199에 따른 로그(웹페이지수)와, 자연수 1–2199 각각에 시차(lag)를 1로 갖는 자연수 2–2200에 따른 로그(웹페이지수) 사이의 시차도(lag plot)를 그려 보면 다음 그림 2.10과 2.11과 같다. 확연하게 웹페이지수가 100에서 1,000 사이인 그룹(주로 1000에서 2200 사이의 자연수, 그림에서 십자표시 점)과 웹페이지수가 1,000에서 10,000 사이인 그룹(주로 100에서 1000 사이의 자연수, 그림에서 원표시 점)을 확인할 수 있다. 1에서 100 사이의 자연수(그림에서 삼각형표시 점)와 1999–2009 사이의 자연수(그림에서 십자표시 점)가 또 다른 작은 그룹(주로 특이값들)을 형성한다. 특이한 것은 그림에서 보는 것처럼 1000에서 2200 사이의 자연수는 웹페이지수가 100에서 1,000 사이인 그룹을 주로 형성하나 웹페이지수가 1,000에서 10,000 사이인 그룹

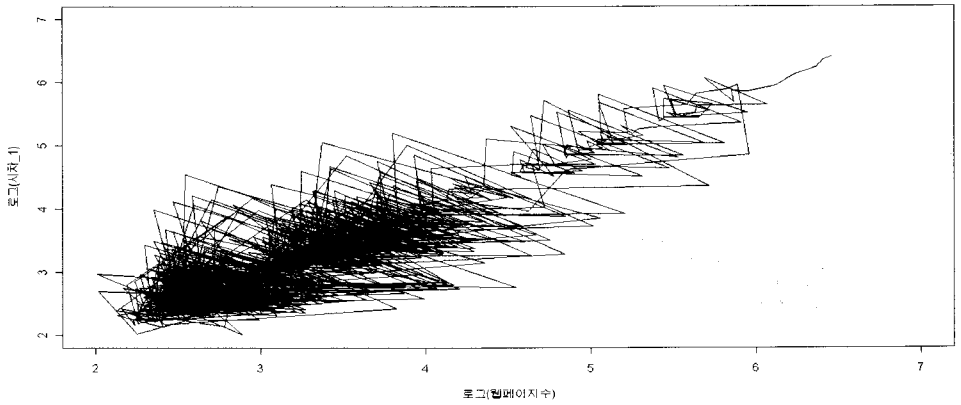

그림 **2.12.** 시차도(3)

표 **2.1.** 웹페이지수에 대한 수치적인 자료요약

최소값	Q_1	중앙값	Q_3	최대값	산술평균	MAD	IQR	표준편차
102	423.8	1342	3497	2865275	18520.0	1497.43	3073.5	126794.7

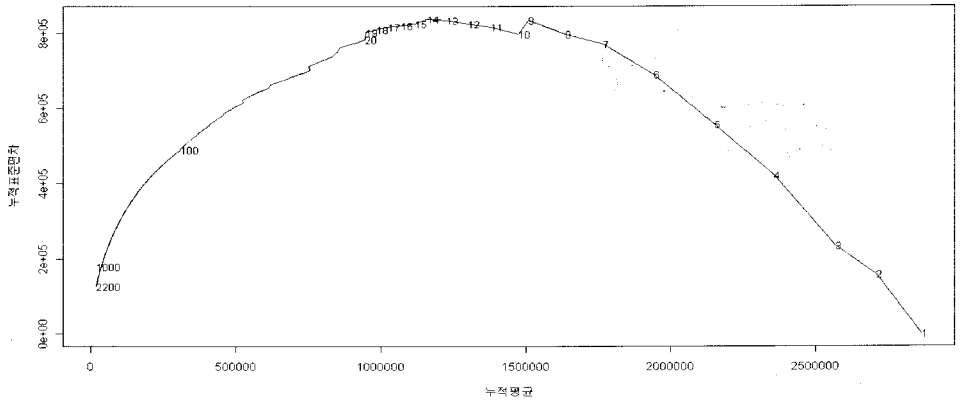

그림 **2.13.** 누적산술평균과 누적표준편차를 나타내는 꺾은선그래프

이나 특이값 그룹에도 다수 포함된다는 것이다. 그림 2.10과 2.11에서 1000에서 2200 사이의 자연수가 대각선을 따라 길게 세 개의 그룹으로 흩어져 있음을 알 수 있다.

다음 그림 2.12는 자연수 1–2199에 따른 로그(웹페이지수)와, 자연수 1–2199 각각에 시차를 1로 갖는 자연수 2–2200에 따른 로그(웹페이지수) 사이의 시차도를 1에서 시작하여 2199까지 연결하여 그린 그림이다. 자연수 1에서 2199로 진행하면서 웹페이지수의 변화가 아주 심함을 알 수 있다.

자연수 1–2200에 대응하는 웹페이지수에 대한 수치적인 자료요약은 표 2.1과 같다. 값이 큰 특이값이 많은 관계로 산술평균과 중앙값, 표준편차와 IQR(MAD)를 서로 비교하면 많은 차이가 있음을 알 수 있다.

표 **2.2.** 첫 숫자에 대한 웹페이지 비율과 벤포드 법칙

첫숫자	웹페이지수	웹페이지 비율	벤포드 법칙 비율
1	9,020,525	0.253	0.301
2	7,192,904	0.202	0.176
3	5,124,660	0.144	0.125
4	3,628,790	0.102	0.097
5	3,316,552	0.093	0.079
6	2,255,135	0.063	0.067
7	2,043,851	0.057	0.058
8	1,707,660	0.048	0.051
9	1,295,873	0.036	0.046
합계	35,585,950	1	1

다음 그림 2.13은 자연수 1에서 2200으로 진행하면서 구한 누적산술평균과 누적표준편차를 나타내는 꺾은선그래프이다. 자연수 1에서 2200으로 진행하면서 누적산술평균은 급격히 줄지만 누적표준편차는 크게 줄지 않음을 알 수 있다. 자연수 14에서 누적표준편차가 최대가 된다.

자연수 1–999에 대응하는 웹페이지수를 이용하여 벤포드 법칙(Benford's law)이 성립하는가를 알아보기 위하여 자연수 1–999를 첫 숫자(first digits, leading digits)에 따라 9개의 그룹으로 나누었더니 다음 표 2.2와 같았다. 첫 숫자란 각각의 숫자에서 유의한 첫 숫자를 가리키는 데, 예로 '351'이나 '0.0351'에서의 첫 숫자는 3이 된다. 인간이 생활을 하면서 발생시키는 숫자들의 첫 숫자가 각각 1, 2, 3, 4, 5, 6, 7, 8, 9가 될 확률은 균등분포처럼 되지 못하고 다음 표 2.2 마지막 열에서 보는 것처럼 제일 큰 것은 1일 때 0.301, 제일 작은 것은 9일 때 0.046이 되고 1에서 9 순으로 발생 확률이 작아진다. 벤포드 법칙에 대해서는 통계논문에서도 나타난다 (예로, Hill (1996), Irmay (1997), Schatte (1998), Leemis 등 (2000), Engel과 Leuenberger (2003), Geyer와 Williamson (2004), Hill과 Schürger (2005), Diekmann (2007), Göb (2007), Schürger (2008) 등이 있고 Bradley와 Farnsworth (2009)는 간단한 예와 함께 벤포드법칙에 대하여 소개하고 있다.). 벤포드법칙의 응용으로서 허위로 작성하여 보고한 세무자료를 밝혀내는 데 사용되거나 (Drake와 Nigrini, 2000; Durtschi 등, 2004) 자료조작, 화상인식, 생존분포, 차분방정식, 인터넷 옥션에서의 상품가격, 블랙-숄츠 모형 등 다양한 분야에서 활용되고 있다 (Jolion, 2001; Sehity 등, 2005; Lagarias와 Soundararajan, 2006; Berger와 Siegmund, 2007; Giles, 2007; Costas 등, 2008; Hales 등, 2008; Lipovetsky, 2008). 웹페이지 비율과 벤포드 법칙에서의 비율을 서로 비교하여 보면 첫 숫자 1에서 3까지 비율에 차이가 나나 첫 숫자 4에서 9까지는 비율에 차이가 많이 나지 않는다. 첫 숫자 1에서 웹페이지 비율이 벤포드 법칙에서의 비율보다 0.048 작은 반면 첫 숫자 2와 3을 합쳤을 때 웹페이지 비율이 벤포드 법칙에서의 비율보다 0.045 크다. 웹페이지에서 첫 숫자 1이 벤포드 법칙보다는 약 5퍼센트 적게 나타나고 첫 숫자 2와 3을 합쳤을 때 벤포드 법칙보다는 약 5퍼센트 많이 나타난다는 뜻이다. 그림 2.14는 표 2.2의 결과를 나타내는 막대그래프이다.

카이제곱검정이나 Kuiper 검정을 시행하여 보면 양 쪽 검정 모두 p-값이 0.001보다도 작아 모두 기각된다. 이런 결과가 나온 것은 우리가 조사란 것이 자연수 1에서 999를 대상으로 해당 자연수의 빈도수를 모두 측정한 것이 아니라 해당 자연수가 나타나는 웹페이지수를 측정한 것이어서 그렇다고 생각된다. 웹페이지 내에 해당 자연수는 한 개가 존재할 수도 있고 복수개가 나타날 수도 있다. 자연수 1에서 999를 대상으로 해당 자연수의 빈도수를 모두 측정한 후 검정을 행한다면 검정결과는 달라질 수도 있을 것이다. 검정에서는 기각되나 웹페이지 비율은 벤포드 법칙을 따라가려고 노력하고 있다는 것을 알 수 있다.

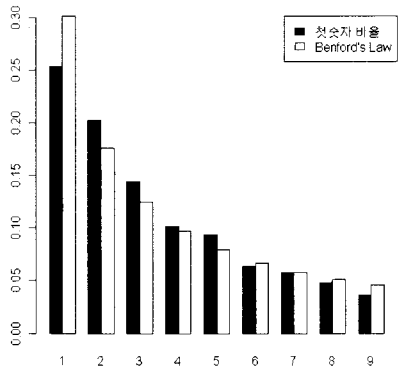

그림 2.14. 첫 숫자에 대한 웹페이지 비율과 벤포드 법칙을 나타내는 막대그래프

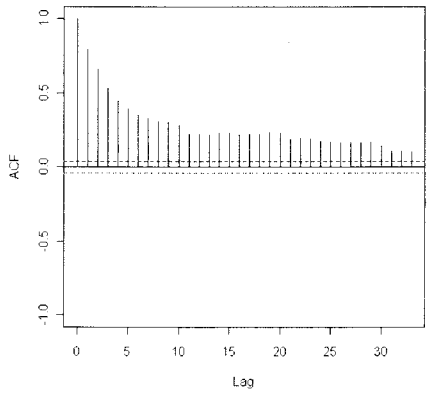

그림 2.15. 자기상관도

그림 2.15는 자연수 1–2200이 나타나는 웹페이지수에 대하여 구한 자기상관도(autocorrelation function plot)이다. 자기상관관계가 매우 강함을 알 수 있다.

사전편찬학(lexicography)에서 두 단어 사이의 연관성을 알기 위하여 우리는 다음과 같은 상호정보(mutual information)를 주로 이용한다 (Church와 Hanks, 1990).

$$\mathrm{MI} = \log_{10} \frac{p(x,y)}{p_1(x)p_2(y)} = \log_{10} \frac{\frac{n(x,y)}{NN}}{\frac{n_1(x)}{NN} \times \frac{n_2(y)}{NN}} = \log_{10} \frac{NN \times n(x,y)}{n_1(x)n_2(y)},$$

여기서, NN은 총단어의 총빈도수이고 $p(x,y)$와 $n(x,y)$는 각각 단어 x와 단어 y가 동시에 나타나는 비율과 빈도수, $p_1(x)$와 $n_1(x)$는 각각 단어 x가 나타나는 비율과 빈도수, $p_2(y)$와 $n_2(y)$는 각각 단어 y가 나타나는 비율과 빈도수이다.

자연수 1–2200이 나타나는 웹페이지수 사이에 그림 2.15에서 본 것처럼 자기상관관계가 강함을 알 수 있었다. 이를 더 구체적으로 확인하기 위하여 해당 웹페이지수가 많은 자연수 1–20 그룹과 캘린더 연도

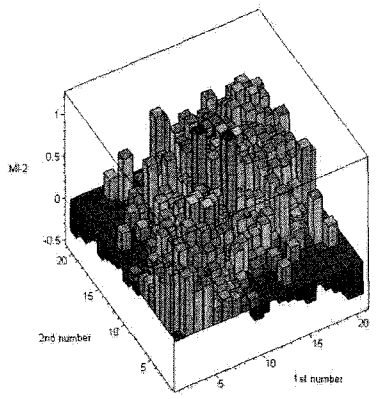

그림 **2.16.** 자연수 1–20에서의 상호정보행렬그림

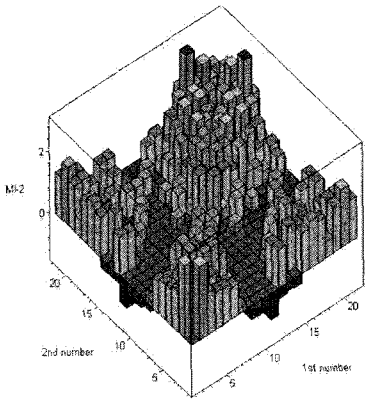

그림 **2.17.** 자연수 1998–2018에서의 상호정보행렬그림

와 관련이 있는 1998–2018 그룹 두 개의 그룹을 표본으로 하여 상호정보를 계산하여 웹페이지수 사이의 연관성을 알아보았다. 웹페이지에 나타나는 모든 자연수를 알 수는 없으므로 편의상 $NN = 10^{10}$으로 고정하였다.

자연수 1–20에서의 상호정보행렬을 그림으로 나타내면 다음 그림 2.16과 같았다. 상호정보량을 자세히 비교하기 위하여 z축은 상호정보에서 2를 뺀 값을 표시하였다.

그림 2.16에서 대각선 원소들은 의미가 없다. 일반적으로 인접한 두 자연수 사이의 연관성이 강하고 두 자연수가 멀리 떨어질수록, 즉 두 자연수의 차이가 클수록 연관성이 약해짐을 알 수 있다. 특이한 것은 바로 인접한 자연수가 아닌 데도 불구하고 자연수 9와 11 사이의 연관성이 인접한 자연수들보다 상대적으로 연관성이 강하다는 사실이다.

자연수 1998–2018에서의 상호정보행렬을 그림으로 나타내면 다음 그림 2.17과 같았다. 그림 2.17에서 x축과 y축에서의 숫자는 1998을 1로, 2018을 21로 나타내었다. 상호정보량을 자세히 비교하기 위하여 z축은 상호정보에서 2를 뺀 값을 표시하였다.

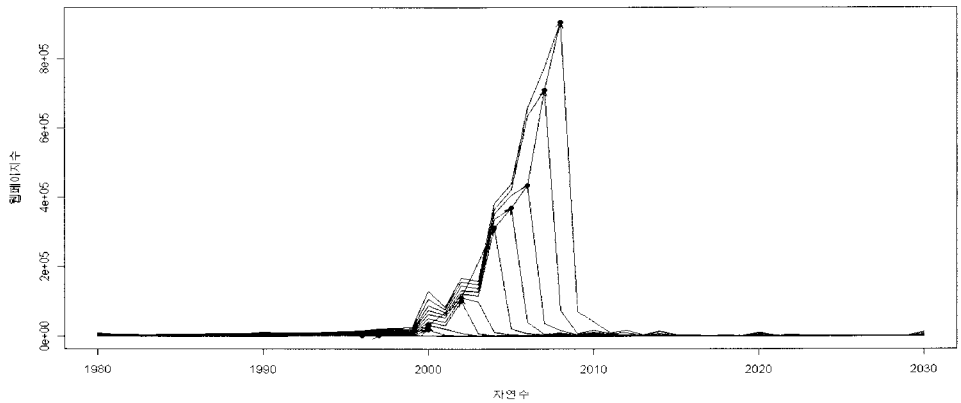

그림 2.18. 자연수 1980–2030에 따른 웹페이지수를 나타내는 꺾은선그래프(1998년→2008년: 아래에서 위로)

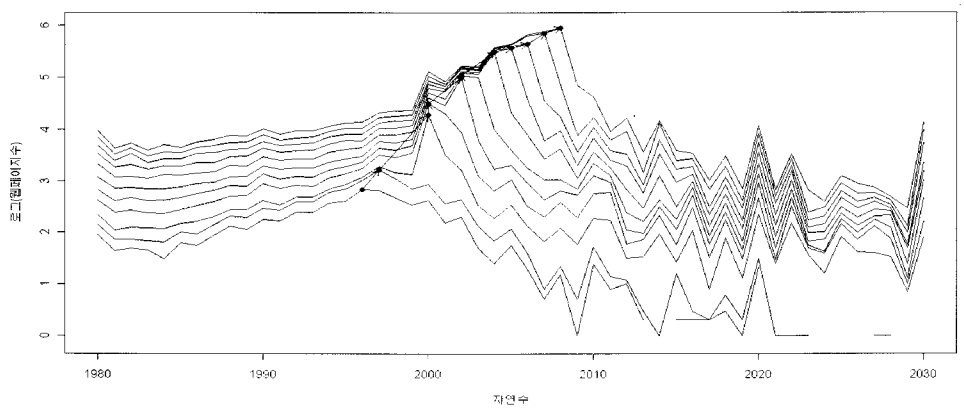

그림 2.19. 자연수 1980–2030에 따른 로그(웹페이지수)를 나타내는 꺾은선그래프(1998년→2008년: 아래에서 위로)

그림 2.17에서 대각선 원소들은 의미가 없다. 일반적으로 인접한 두 자연수 사이의 연관성이 강하고 두 자연수가 멀리 떨어질수록, 즉 두 자연수의 차이가 클수록 연관성이 약해짐을 알 수 있다. 그러나 1–20 그룹과는 다른 패턴을 보이고 있다. 연관성이 비교적 큰 영역을 살펴보면 자연수 2008(현재)을 중심으로 2008 이전(과거)과 2008 이후(미래)로 나뉜다. 즉 2008 이전(과거) 년도들은 과거년도들끼리 연관성이 깊고 2008 이전(미래) 년도들은 미래년도들끼리 연관성이 깊음을 알 수 있다. 바로 인접한 자연수가 아닌 데도 불구하고 자연수 2013과 2017 사이의 연관성이 인접한 자연수들보다 상대적으로 강하다. 또한 두 자연수 사이의 간격이 있음에도 불구하고 자연수 1998–1999은 2009, 2012–2014, 2016–2018과 연관성이 높고 자연수 2003은 2016–2018과 연관성이 높다. 전체적으로 보면 1–20 그룹보다는 좀 더 복잡한 패턴을 보이고 있다.

다음 그림 2.18은 1998년부터 2008년 기간동안 자연수 1980–2030에 따른 웹페이지수를 나타내는 꺾은선그래프이고 다음 그림 2.19는 1998년부터 2008년 기간동안 자연수 1980–2030에 따른 로그(웹페이지수)를 나타내는 꺾은선그래프이다. 최대 웹페이지수가 현재 연도(현재)에 나타나는 비평형(non-

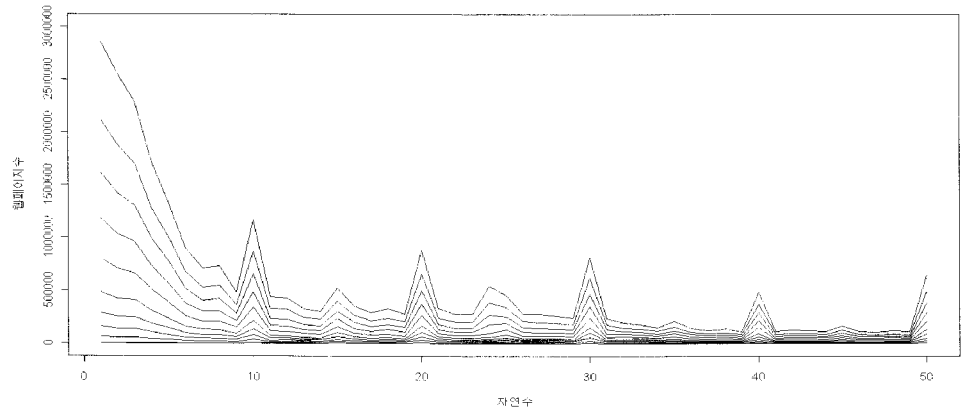

그림 2.20. 자연수 1–50에 따른 웹페이지수를 나타내는 꺾은선그래프(1998년→2008년: 아래에서 위로)

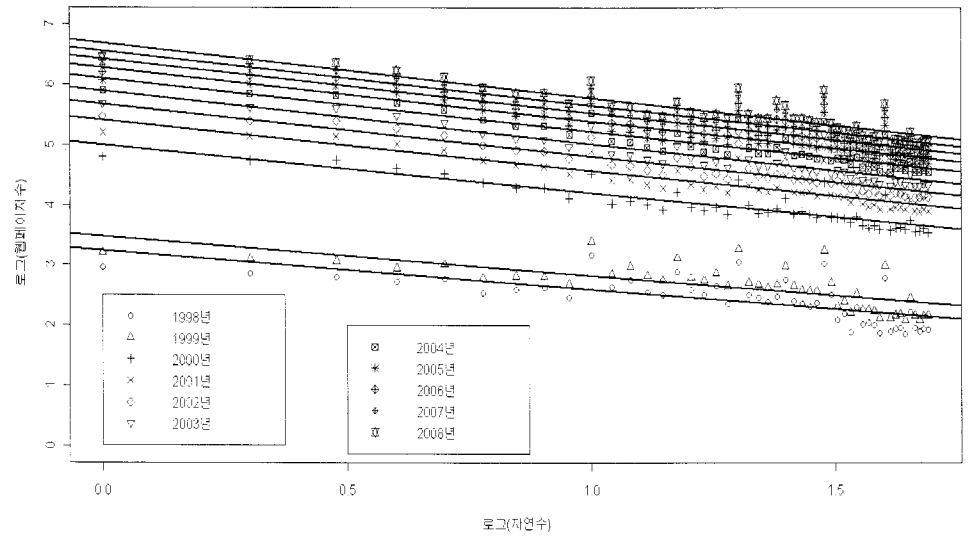

그림 2.21. 자연수 1–49를 대상으로 로그(자연수)에 따른 로그(웹페이지수)를 나타내는 산점도들과 추정회귀직선들(1998년→2008년: 아래에서 위로)

equilibrium)적이고 진화하는 시스템임을 알 수 있다. 즉 직전 연도(과거)에 나타나던 최대 웹페이지수가 해가 바뀌면 당해 연도(현재)로 계속 옮겨 가는(그림 2.18이나 2.19에서 화살표가 이동하는 모습을 보면 알 수 있다.), 안정적이지 못하고 계속 변화하는 시스템인 것이다. 또한 최대 웹페이지수도 폭발적으로 증가하고 있음을 알 수 있다.

다음 그림 2.20은 1998년부터 2008년 기간동안 자연수 1–50에 따른 웹페이지수를 그린 꺾은선그래프이다. 1998년에서 2008년으로 가면서 웹페이지수가 폭발적으로 증가하고 있음을 알 수 있다. 특히 자연수 1–5에서 이런 현상이 두드러지고 어림수(round number) 10, 20, 30, 40, 50에서 웹페이지수가 특이하게 올라가는 패턴을 따르는 데 10, 20, 30, 40, 50 순으로 이러한 패턴이 두드러진다. 어림수보다는 못

표 **2.3.** 년도에 따른 추정직선회귀식들의 기울기

년도	기울기
1998	−0.641
1999	−0.653
2000	−0.802
2001	−0.842
2002	−0.865
2003	−0.884
2004	−0.884
2005	−0.895
2006	−0.900
2007	−0.907
2008	−0.914

하지만 5로 끝나는 자연수 15, 25, 35, 45가 그 뒤를 따라 이러한 패턴을 따른다.

다음 그림 2.21은 1998년부터 2008년 기간동안 자연수 1–49를 대상으로 로그(자연수)에 따른 로그(웹페이지수)를 그린 산점도들과 추정회귀직선들이다. 이 그림에서 직선회귀식들의 기울기를 구하여 보면 다음 표 2.3과 같이 1998년에서 2008년으로 가면서 기울기가 −0.641에서 시작하여 −1 근처를 향하여 변하여 가고 있음을 알 수 있다. 시간이 흘러감에 따라 포털사이트 DB에 쌓이는 웹페이지에 나타나는 숫자들이 이루는 시스템은 시간에 따라 시시각각 변하는 동적 시스템이 된다는 것을 암시한다.

3. 결론

우리는 지금까지의 자료분석으로 자연수 1–2200과 이에 대응하는 웹페이지수 자료는 거듭곱 법칙을 따름을 알 수 있었고, 또한 상수 β가 한 개가 아니고 여러 개가 존재하는 복잡한 시스템인 것을 알 수 있었다. 또한 이러한 자료에 지프 법칙이나 벤포드 법칙이 성립하는 지를 살펴보았다. 숫자가 아닌 일반 단어를 대상으로(예를 들어 통계용어) 웹페이지수 자료는 어떤 성질을 갖는 시스템인지를 밝히는 작업이 추후 과제가 될 것이다.

참고문헌

Albert, R., Jeong, H. and Barabàsi, A. L. (1999). Diameter of the World-Wide Web, *Nature*, **401**, 130–131.

Barabàsi, A. L. and Albert, R. (1999). Emergence of scaling in random networks, *Science*, **286**, 509–512.

Berger, A. and Siegmund, S. (2007). On the distribution of mantissae in nonautonomous difference equations, *Journal of Difference Equations and Applications*, **13**, 829–845.

Bradley, J. R. and Farnsworth, D. L. (2009). What is Benford's law?, *Teaching Statistics*, **31**, 2–6.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography, *Computational Linguistics*, **16**, 22–29.

Costas, E., López-Rodas, V., Toro, F. J. and Flores-Moya, A. (2008). The numbers of cells in colonies of the cyanobacterium Microcystis aeruginosa satisfies Benford's law, *Aquatic Botany*, **89**, 341–343.

Diekmann, A. (2007). Not the first digit! Using Benford's law to detect fraudulant scientific data, *Journal of Applied Statistics*, **34**, 321–329.

Dorogovtsev, S. N., Mendes, J. F. F. and Oliveira, J. G. (2006). Frequency of occurrence of numbers in the World Wide Web, *Physica A*, **360**, 548–556.

Drake, P. D. and Nigrini, M. J. (2000). Computer assisted analytical procedures using Benford's law, *Journal of Accounting Education*, **18**, 127–146.

Durtschi, C., Hillison, W. and Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data, *Journal of Forensic Accounting*, **5**, 17–34.

Engel, H. A. and Leuenberger, C. (2003). Benford's law for exponential random variables, *Statistics & Probability Letters*, **63**, 361–365.

Geyer, C. L. and Williamson, P. P. (2004). Detecting fraud in data sets using Benford's law, *Communications in Statistics-Simulation and Computation*, **33**, 229–246.

Giles, D. E. (2007). Benford's law and naturally occurring prices in certain ebaY auctions, *Applied Economics Letters*, **14**, 157–161.

Göb, R. (2007). Data conformance testing by digital analysis-A critical review and an approach to more appropriate testing, *Quality Engineering*, **19**, 281–297.

Hales, D. N., Sridharan, V., Radhakrishnan, A., Chakravorty, S. S. and Siha, S. M. (2008). Testing the accuracy of employee-reported data: An inexpensive alternative approach to traditional methods, *European Journal of Operational Research*, **189**, 583–593.

Hill, T. P. (1996). A statistical derivation of the significant-digit law, *Statistical Science*, **10**, 354–363.

Hill, T. P. and Schürger, K. (2005). Regularity of digits and significant digits of random variables, *Stochastic Processes and their Applications*, **115**, 1723–1743.

Huberman, B. A. and Adamic, L. A. (1999). Growth dynamics of the World-Wide Web, *Nature*, **401**, 131.

Huberman, B. A., Pirolli, P. L., Pitkow, J. E. and Lukose, R. M. (1998). Strong regularities in World-Wide Web surfing, *Science*, **280**, 95–97.

Irmay, S. (1997). The relationship between Zipf's law and the distribution of first digits, *Journal of Applied Statistics*, **24**, 383–393.

Jolion, J. M. (2001). Images and Benford's law, *Journal of Mathematical Imaging and Vision*, **14**, 73–81.

Lagarias, J. and Soundararajan, K. (2006). Benford's law for the $3x+1$ function, *Journal of the London Mathematical Society*, **74**, 289–303.

Leemis, L. M., Schmeiser, B. W. and Evans, D. L. (2000). Survival distributions satisfying Benford's law, *The American Statistician*, **54**, 1–6.

Lipovetsky, S. (2008). Comparison among different patterns of priority vectors estimation methods, *International Journal of Mathematical Education in Science and Technology*, **39**, 301–311.

Schatte, P. (1998). On Benford's law to variable base, *Statistics & Probability Letters*, **37**, 391–397.

Schürger, K. (2008). Extensions of Black-Scholes processes and Benford's law, *Stochastic Processes and their Applications*, **118**, 1219–1243.

Sehity, T., Hoelzl, E. and Kirchler, E. (2005). Price developments after a nominal shock: Benford's law and psychological pricing after the euro introduction, *International Journal of Research in Marketing*, **22**, 471–480.

Numbers in the Internet Web and Benford's Law

Dae-Heung Jang[1]

[1]Division of Mathematical Sciences, Pukyong National University

(Received February 2009; accepted March 2009)

Abstract

Using the information about the frequency of occurrence of numbers in WWW, we can find properties of the array of numbers and validate whether this array satisfies the several laws(Power law, Zipf's law, Benford's law).

Keywords: World-Wide Web, power law, Zipf's law, Benford's law.

This work was supported by Pukyong National University Research Abroad Fund in 2007(PS-2007-009).

[1]Professor, Division of Mathematical Sciences, Pukyong National University, 599-1 Daeyeon-dong, Nam-gu, Busan 608-737, Korea. E-mail: dhjang@pknu.ac.kr