Dol: 10.3743/OPEM.2007.24.4.255

검색엔진에서 일간 질의어 분포의 정상성에 관한 연구*

On Regularity of Daily Distribution of Queries in Search Engine

박상규(Sang-Gue Park),이찬규(Chan-Kyu Lee)***,윤경현(Kyung-Hyun Yoon)****,
김성희(Seong-Hee Kim)*****,이준호(Jun-Ho Lee)******

초 록

하루에도 수백만 건 이상의 방대한 질의어가 입력되는 검색엔진을 통해 인터넷 검색 패턴을 분석하는데 활용될 수 있는 질의어 분포의 정상성을, Pareto 분포와 Zipf 모형을 이용해 검토하였다. 실제 2주일간의 일간 검색어 분포를 Pareto 모형으로 적합 시킨 결과, 그 기울기가 1.33에서 1.34 정도로 일관되게 측정되고 있어 Pareto 분포를 이용해 일간 질의어 분포의 정상성을 평가할 수 있음을 제시하였다. 이런 연구결과는 사회적 관심사와 그 변동 흐름을 해석하는데 보다 객관적이며, 효율적인 자료를 제공해 줄 수 있을 것이다.

ABSTRACTS

In this paper we analyzed regularity of daily patterns of distribution of Queries coming from internet search engine. And then, we proposed a Pareto distribution and Zipf law for identifying the query distribution and applied them to daily queries on the search engine during 2 week. We found that there is some evidence that Pareto and Zipf laws can be applied to evaluate the regularity of daily patterns of distribution of queries in search engine. Those results can be used to provide a better understanding of the social interests and trends using the query distribution patterns.

키워드 : 질의어, 파레토 법칙, 지프 법칙, 검색엔진, 정상성
daily query patterns, pareto model, regularity, search engine, zipf's law.

* 본 논문은 정통부 및 정보통신 연구진흥원의 정보통신 선도기반기술 개발사업 연구결과로수행되었음.
** 중앙대학교 수학통계학부 교수(spark@cau.ac.kr)
*** 중앙대학교 국문학과 교수(leeck@cau.ac.kr)
**** 중앙대학교 컴퓨터공학과 교수(khyoon@cau.ac.kr)
***** 중앙대학교 문헌정보학과 부교수(seonghee@cau.ac.kr)
****** 숭실대학교 컴퓨터학부 교수(joonho@naver.com)
■ 논문접수일자 : 2007년 11월 29일
■ 게재확정일자 : 2007년 12월 15일

1. 서 론

인터넷이 사회 전반에 걸쳐 일상생활의 한 부분으로 자리 잡으면서, 정보에 대한 요구가 있을 때 정보 검색을 위해 웹에 의존하는 정도가 점점 커져가고 있다. 최근 정보통신부와 한국인터넷진흥원이 발표한 정보화 실태조사 결과에 따르면, 2007년 6월 우리나라 만 6세 이상 국민의 인터넷 이용률은 75.5%로 이용자 수는 3,443만 명으로 전년대비 2.0%가 증가한것으로나타났고, 인터넷이용자의88.7%가 검색엔진 등을 통해 자료나 정보를 획득하는 것으로 나타나 이를 입증해주고 있다.

검색엔진(search engine)이란 인터넷상에 산재해 있는 제반 정보를 미리 수집하고 이를 체계적으로 저장한 후, 사용자가 원하는 정보를 수시로 찾을 수 있도록 해주는 일종의 데이터베이스 관리시스템을 뜻한다. 인터넷 이용자들에게 검색엔진은 의사결정을 위한 정보 수집 도구로서 매우 중요한 의미를 가지고 있기 때문에 인터넷 포털 사이트는 사용자가 원하는 다양한 정보를 쉽게 제공할 수 있는 강력한 검색엔진을 제공하기 위해 많은 노력을 하고 있다. 이러한 이유로 포털 사이트는 인터넷 사용자들이 검색엔진을 통해 어떤 정보에 관심 있고, 어떤 정보를 찾고자 하는지 그 행태를 지속적으로 파악하고자 하고 있다.

일부 포털 사이트의 경우 검색엔진을 통해 입력되는 질의어의 수가 하루 2,000만개를 넘고 있다. 매일 검색엔진을 통해 입력되는 질의어의 크기도 크지만 그 내용도 워낙 다양하기 때문에 검색어의 내용을 기반으로 질의어의 특성을 분석하는 것은 쉽지 않다. 따라서 대부분의 포털 사이트들은 매일 혹은 실시간으로 검색엔진을 통해 입력되는 질의어의 빈도와 순위 등을 제공하면서 빈도가 높은 질의어 위주로 내용이나 특징을 분석하고 있는 상황이다.

하지만 현재와 같이 질의어 빈도를 통한 순위 정보는 많은 경우 사회적인 관심사의 큰 변화를 실시간으로 보여준다는 장점은 있지만, 날짜나 연도가 달라질 때 그 패턴을 비교할 수 있는 방법이 없다. 또한 순위 정보를 비교할 때 비교하려는 순위 정보에 속해있는 빈도 분포가 서로 다를 경우 그 의미를 분석하는 것이 제한적이고 또 비교하는 것 자체가 어려울 수 있다는 문제점이 있다.

또한 빈도가 높은 일부 질의어를 통해 사용자 관심사에 관한 다양한 패턴을 비교할 때 비교하려는 대상이 서로 다른 분포를 가지고 있다면[1], 분석할 때 그 결과를 왜곡시킬 수 있다. 그래서 매일같이 검색엔진을 통해 입력되는 질의어가 그 내용은 다르더라도 그 전체적인 빈도나 순위의 곱이 일정하게 유지된다는 Zipf의 법칙을 활용하여 일정한 패턴을 찾아낸다면 그 패턴을 질의어 분포의 정상성과 특이성을 구분할 수 있는 기준으로 삼을 수 있다고 본다.

이에 따라 여러 학문분야에서 웹 검색에 관한 연구들이 수행되어왔다. Silverstein et al.(1999)은 6주동안알타비스타이용자들이9억 9천만개 이상의 질의를 트랜잭션 로그를 이용하여 분석하였다. Spink et al.(2001)은

1) 예를 들어 순위는 같지만 빈도가 전혀 다른 두 분포를 비교할 때 나타나는 결과는 그 의미가많이 다를 수 있다.

1997년부터 2001년까지 2년에 한번씩 하루를 선정하고 그날 익사이트에 입력된 질의들 중 무작위로 추출된 약 2,500개의 주제를 분류하였다. 그 결과 검색질의어가 엔터테인먼트와 성관련으로부터 전자 상거래 관련으로 변화하였으나 이용자들의 전반적인 검색행태는 변하지 않았음을 보고하였다. Jansen, Spink, Pedersen(2005)은 2002년 알타비스타에서 생성된 질의들로부터 2,603개를 무작위로 추출하여 이를 주제별로 분류하고 Silverstein et al. (1999)의 연구결과와 비교 하였다.그 결과 질의어의 주제가 좀 다양해졌으며 성과 관련된 주제가 감소된 것으로 나타났다. 국내의 연구로는 박소연, 이준호(2002, 2005)는 1년동안 네이버에 입력된 검색 질의를 기반으로 이용자의 검색 행태 추이를 분석한 결과 계절별, 주중과 주말, 요일별 질의 형태의 분포에 있어서 유의한 차이가 있다고 하였다. 이상에서와 선행연구를 보면 그 동안의 연구결과는 검색어를 주제별로 적절하게 분류되어 보다 포탈 사용자들이 효율적으로 구축된 콘텐츠를 잘 사용할 수 있는 환경을 제공할 수 있으나 질의어들의 전반적인 빈도분포를 분석하거나 특정 패턴에 관한 연구를 하여 지속적으로 변하는 사회적 현상이나 관심을 반영할 수 있는 검색어 패턴에 관한 연구는 거의 없다. 즉, 현재 대부분의 포털 사이트에서는 매일같이 검색엔진을 통해 입력되는 질의어 빈도와 순위 정보를 제공하고 있지만 분포 자체에 대한 연구에 기초하여 분석되고 있지 못하다.

사실 검색엔진 이용자들의 검색사용 패턴을 빨리 예측하여 이에 대비하는 것은 포탈 사이트로서는 매우 중요한 문제이다. 현재 대다수 포탈 사이트에서는 검색엔진을 통해 입력된 질의어들의 일간 혹은 주간 순위를 제공하고 있다. 이러한 순위 정보는 포탈 사이트의 콘텐츠를 구성하는데 도움을 주고 또한 사용자에게 최근 사회의 관심 동향이 어디로 흘러가는지 보여주면서 인터넷 소비자 행동 패턴을 예측할 수 있는 정보를 준다는 점에서 매우 중요하게 사용되고 있을 뿐만 아니라 최근에는 이러한 순위 정보가 인터넷 광고에까지 연결되어 사용되고 있어 그 중요성이 갈수록 강조되고 있다. 또한 질의어는 정보 대한 요구를 직접 표출하고 있는 표현이기 때문에 사회적, 문화적 흐름을 파악하는데 매우 중요하다고 할 수 있다. 예를 들어 미국의 이라크 침공에 대한 뉴스 발표가 나면 곧바로 네티즌들은 상세정보획득을 위해 포탈 사이트 검색창에‘ 이라크 침공 ‘‘ 이라크' ‘ 부시' 와 같은 키워드들이 입력되고 인기 키워드 순위에서 해당 키워드 들이 급상승하게 된다. 물론 이와 같은 사회적인 현상은 질의어 분석을 통하지 않더라도 동향 파악에 무리가 없겠지만‘ 얼짱'과 같은 문화적 키워드에 대한 동향은 네티즌 사회에서 자생적으로 발생해 증폭되는 문화현상 이므로 사회 문화 전반에 걸친 변화를 감지하기란 쉽지 않다. 이런 동향 변화를 질의어 분석을 통해 빠르게 인식하거나 예측할 수 있다. 또한 이러한 인기 질의어를 광고와 결합시켜 최근 포탈 사이트들은 막대한 수입을 벌어들이고 있다. 따라서 최근 포탈 사이트 들은 차세대 서비스 방향을 결정하기 위해 이런 분석에 많은 시간과 노력을 투자하고 있다.

본 연구에서는 검색 엔진을 통해 들어오는 수많은질의어가어떤패턴(pattern)을이루고

있는지를 분석하고자 한다. 먼저 이를 수행하기 위해서 빈도가 높은 질의어를 1년을 기준으로 표본조사 해 본 결과 질의어가 일정한 패턴을 띠고 있다는 가설을 갖게 되었다. 입력어가 일정한 패턴을 갖는다는 것은 사람들의 관심사에 어떤 흐름이 존재한다는 것을 의미한다. 그래서 이러한 패턴을 파악하게 되면 사람들의 사회적 관심사에 대한 전체적인 흐름을 분석해 낼 수 있다는 큰 의의를 지니게 되는 것이다.

본 연구에서는 질의어가 일정한 패턴을 유지하는경우를검색질의어가정상성(reguhrity)을 유지하고 있다고 규정하고자 하는 것이다. 정상성을 유지하고 있는 날들과 그렇지 않은 날들이 존재할 경우 비정상 패턴을 보다 심도 있게 분석할 수 있을 것이다. 만약 어떤 날의 패턴이 이 정상성으로 부터 벗어나 있다면 그 날은 뭔가 특별한 사회적 이슈가 등장한 날로 간주하고 분석할 수 있게 되어, 사회적 변동에 대한 대처와 예측이 가능해 지는 것이다. 즉, 검색 질의어의 정상성이란 검색질의어의 변동을 분석할 수 있는 기준 척도로 삼을 수 있을 것이다.

본 연구는 검색엔진을 통해 입력된 질의어의 일간 분포를 Zipf의 법칙을 이용하여 분석하고 그것을 통해 분포의 정상성을 정의하는 것을 목적으로 한다. 즉, 매일 검색엔진을 통해 입력된 질의어의 빈도와 순위로 모형을 정한 후 정상성의 의미를 부여할 수 있는 모형을 통해 경험적으로 분석한다. 또한 분석된 모형에서 정상성의 의미를 평가하고자 한다. 이런 연구를 통해 얻을 수 있는 기대효과로는 질의어 분석을 통해 검색어들이 사회의 각종 이벤트나 트렌드, 정상성을 발견할 수 있을 것으로 기대된다. 예를 들면 질의어 로그 파일의 주기성이나 질의어가 언제 급상승하는지, 특정시점에서 관심 있는 뉴스패턴을 찾을 수 있을 것으로 예상된다. 또한 질의어를 특정패턴을 기준으로 분리해서 저장함으로써 검색의 효율성을 제공하는데 활용될 수 있을 것이다.

2. Pareto 분포와 Zipf의 법칙

기업의 규모분포를 최초로 연구한 Pareto(1897)는 일정 규모를 넘는 기업들의 소득분포가 두터운 꼬리를 갖게 되는 분포, 즉 거듭제곱 법칙을 따른다는 것을 보여주었다. 거듭제곱 법칙이라는 것은 한 확률변수 가 보다 크거나 같을 확률이 의 거듭제곱에 비례한다는 것 $\Pr(X \geq x) = x^{-\beta}$ ($\beta > 0$) 로 나타낼 수 있다는 것이다. 이러한 거듭제곱 법칙은 흔히 80-20법칙으로도 인용되며 적은 비율이지만 전체에서 상당히 큰 비중을 차지하고 있는 현상을 설명할 때 많이 사용되어 왔다. 이후 경제학 분야를 중심으로 소득과 관련한 다양한 경제적 지표들의 특성이 Pareto 분포를 이용해서 연구되었고 최근 Pareto Hart & Oulton(1997)는 미국과 영국 등의 국가에서 제조기업 규모에 대한 β가 1.1~1.2의 값을 실증적으로 보이면서 그 의미를 해석하였다.

하버드 대학 언어학 교수 지프(Goerge Kingsley Zipf, 1902-1950)는 파레토의 법칙을 적용하여 특정 코퍼스[2]의 어휘를 극도로 많이 쓰이는 소수의 단어와 제한된 횟수로만 사용되는 대다수의 단어로 구분할 수 있음을

발견하였고이것을‘ Zipf의법칙 이라고하였다.

문서상에서 단어의 출현 빈도와 순위간에는 β=1 인 Zipf의 법칙이 성립하는 것으로 밝혀졌는데, 단어의 출현 빈도와 순위에 Zipf 분포를 적용하면 두 번째 빈도가 높은 단어는 가장 높은 빈도의 단어의 절반 수준이며, 세 번째 빈도가 높은 단어는 가장 큰 빈도의 단어의 1/3 수준으로, 빈도와 순위간의 관계를 예측하는데 불확실성이 존재하지 않는 예측 가능성의 의미를 부여할 수 있고 이런 현상들이 많이 관측된다면 정규성의 의미 역시 부여할 수 있다. 이 Zipf의법칙은기업의규모(Hart & Oulton, 1997), 도시의 규모(Gabaix & Ioannides, 2003) 등 다양한 분야에 적용되어 왔고, 최근에는 Barasi & Albert(1999)는 WEB Page가 이 법칙을 따른다는 연구를 발표하였다.

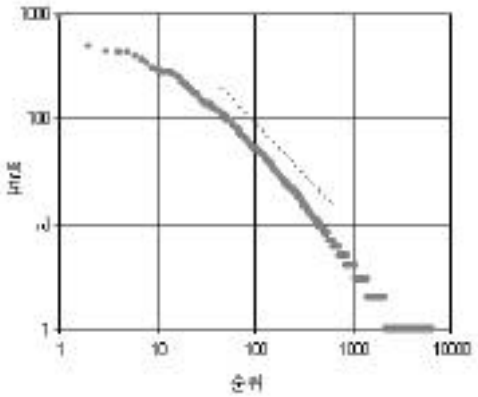
검색엔진의 일간(日間) 질의어 분포를 살펴보기 위해 보다 일반적인 Pareto 모형을 고려한다.

$$F_i = \frac{\alpha}{R_i^{\beta}}, \quad i = 1, \cdots, n, \quad \alpha, \beta > 0. \qquad (1)$$

모형 (1)에서 R_i은 일간 질의어중 i-번째 순위, F_i는 i-번째 순서의 질의어의 빈도, α는 상수, β는 Pareto 지수를 의미한다. 모형 (1)은 양변에 로그를 취해 서 다음과 같은 선형회귀모형으로 표현할 수 있다.

$$\ln(F_i) = \ln(\alpha) - \beta \ln(R_i). \qquad (2)$$

식(2)의 선형회귀모형의 Pareto 지수 β의 수치가 1보다 작을수록 질의어들의 빈도가 비슷한 크기로 분포되어있는 것을 의미하며, β의 수치가 1보다 클수록 불균등한 빈도 분포를 보이는 것을 의미한다. β가 1인 경우 일반적으로 Zipf 모형이라고 하다. Zipf 모형이 일반적인 Pareto 모형보다 선호되는 이유는 빈도와 순위의 관계가 일정한 기울기의 반비례 관계로 예측 타당성이나 설명이 명확하기 때문이다.



<그림 1> 빈도와 순위 관계 Zipf 모형

3. 일간 질의어 분포 실증 분석

일간 질의어 분포의 정규성을 실증적으로 분석하기 위해 2003년 1월 1일부터 14일까지 2주일 동안 포털 사이트 *N*의 검색엔진에 입력된 모든 종류의 질의어를 분석하였다. *N*의 검색자료를 사용한 이유는 *N*이 일정 기간의 전체 질의어를 구할 수 있는 포탈사이트 중 분석

2) 코퍼스는 언어 자료 구축의 범위나 시기를정하고 그안에 있는 모든 어휘와문장에 태그를 붙여 언어 분석에 용이하도록 구축한 일련의 말뭉치를 말한다.

허가를 해주었고, 현재 가장 선호되고 있는 포탈사이트라 그 대표성에 의미를 두었다. 2주일을 분석한 이유는 문헌에 의하면 질의어 사용 빈도와 패턴이 요일별로 차이가 나기 때문에 반복의 의미를 두었다. 웹 검색엔진의 사용도가 급증하는 시점에서 자료가 비교적 오래되었지만 질어어 분포를 패턴화하는데 Pareto 분포와 Zipf 모형을 적용 방법론을 적용할 수 있는지 그 타당성 여부를 확인하는 데는 큰 무리가 없다고 생각하였다. 검색 분석에 사용된 자료의 기초통계량은 <표 1>에 나타나 있다.

1월 1일이라는 특수한 상황을 감안하면 주말인 토요일, 일요일에 500만 개 이하로 검색어 수가 떨어지고 주중에는 모두 500만 개 이

<표 1> 검색 질의어의 일간 기초통계량

	일간 질의어수	개별 질의어수	최대 빈도 질의어수 (비율,%)
2003.1.1(수)	4,081,960	203,550	110,657 (2.71)
2003.1.2(목)	5,471,997	237,259	139,081 (2.54)
2003.1.3(금)	5,406,911	236,736	136,248 (2.52)
2003.1.4(토)	4,638,271	221,052	119,654 (2.58)
2003.1.5(일)	4,075,251	211,048	105,887 (2.60)
2003.1.6(월)	5,551,299	243,639	138,215 (2.49)
2003.1.7(화)	5,805,736	243,895	278,919 (4.80)
2003.1.8(수)	5,724,836	244,701	184,606 (3.22)
2003.1.9(목)	5,567,749	244,313	146,964 (2.64)
2003.1.10(금)	5,541,086	242,444	138,071 (2.49)
2003.1.11(토)	4,833,728	223,557	121,095 (2.51)
2003.1.12(일)	4,198,654	215,682	108,231 (2.58)
2003.1.13(월)	5,987,183	250,013	142,990 (2.39)
2003.1.14(화)	5,836,792	250,135	138,941 (2.38)

상 검색어수가 나타나고 있다. 최대빈도와 일간 질의어수의 비율은 특정 날에는 4.8%정도까지 크게 높아지고 있지만 거의 2.5% 정도로 유지되고 있음을 볼 수 있어 어느 정도 패턴을 유지하고 있음을 알 수 있다. 일간 질의어의 총수가 500만 개 정도 되도 실제 서로 다른 개별 질의어의 수는 20만~25만개 정도 수준임을 알 수 있다. 최대 빈도수를 감안하면 상당히 많은 수의 질의어는 한 두 개정도의 아주 적은 개별 질의어임을 알 수 있다.

전체 질의어 분포에 기초한 분석이 보다 더 일관된 분석 결과를 줄 수 있어 모형 (2)를 적합하여 와 의 추정치를 <표 2>와 같이 계산하였다.

<표 2>를 참고해보면 의 값이 1과는 상당한 차이를 보여 검색어의 빈도가 검색어에 따라 그 빈도가 상당히 불균등하게 분포되어 있음을 알 수 있고 또한 기울기가 1인 Zipf 분포를 따르고 있지 않음을 알 수 있다. 하지만 모형의 기울기 추정값들이 상당히 비슷하게 움직이고 있음을 볼 수 있고, 적합도를 의미하는 가 매우 높아서 Pareto 분포에 잘 적합 되고 있음을 볼 수 있다.

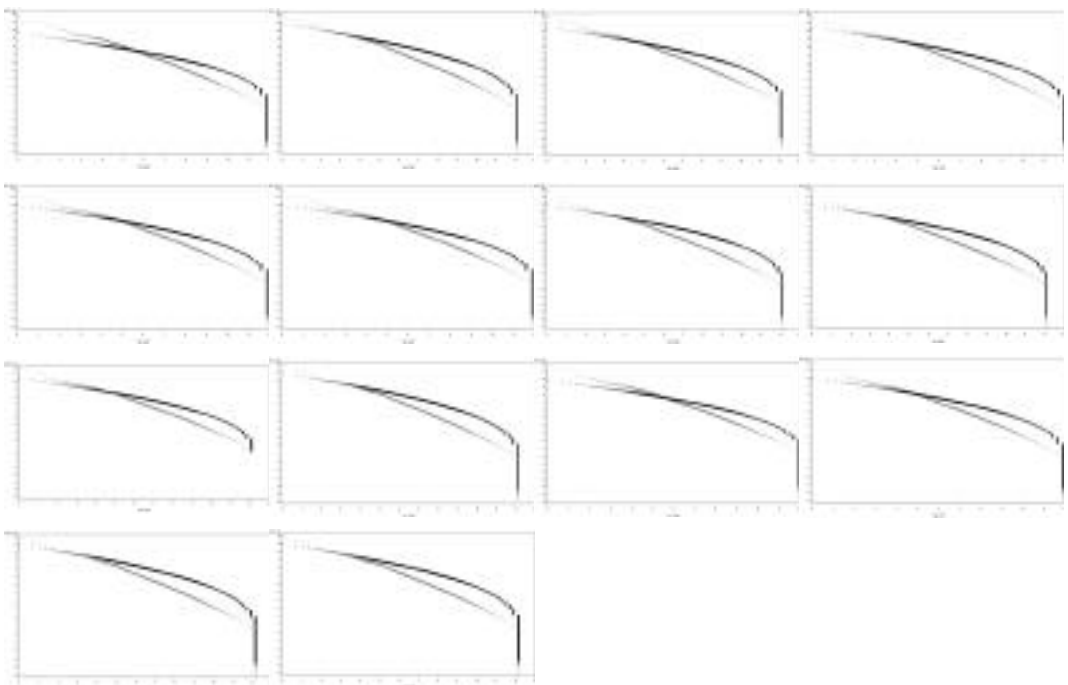
500만개가 넘는 질의어를 분석하다 보면 상당히 많은 수의 질의어가 한 두 개의 정도의 소수 빈도만을 가지고 있다. 이러한 질의어는 크게 의미는 없지만 그 양이 많아 전체 분포를 왜곡 시킬 수 있어 이러한 자료를 적당한 기준으로 제외하거나 영역별로 묶고 Pareto 분포에 적합시켜 Zipf 법칙을 따르는지 여부를 확인하는 경우가 많다. 본 분석자료에서도 일간 입력되는 질의어 중 약 80%의 자료가 하루

<표 2> 파레토 모형 모수 추정치와 적합도

	$\ln(\hat{\alpha})$	$\hat{\beta}$	R^2
2003.1.1(수)	7.01538	1.34438	0.9738
2003.1.2(목)	7.20212	1.34854	0.9659
2003.1.3(금)	7.18124	1.34552	0.9665
2003.1.4(토)	7.09016	1.34235	0.9700
2003.1.5(일)	6.99815	1.33527	0.9727
2003.1.6(월)	7.19986	1.34344	0.9650
2003.1.7(화)	7.20158	1.34311	0.9657
2003.1.8(수)	7.20597	1.34322	0.9650
2003.1.9(목)	7.18029	1.33945	0.9655
2003.1.10(금)	7.19954	1.34468	0.9667
2003.1.11(토)	7.11344	1.34568	0.9707
2003.1.12(일)	7.01347	1.33487	0.9723
2003.1.13(월)	7.25226	1.34726	0.9636
2003.1.14(화)	7.21915	1.34181	0.9641

10개 이하의 질의어로 나타나고 있고, 이들은 어떤 트렌드를 보여주는 값이라기 보다는 랜덤하게 나타날 수 있는 여지가 많다. 만약 이렇게 작은 빈도의 질의어를 분석에 포함시킬 경우 일부 트렌드를 보여줄 수 있는 질의어에 영향을 주어 그 의미를 제대로 파악하기 위하여는 적당한 기준으로 이들을 제외하는 것이 타당하다고생각하였다. 최희갑, 박원석(2006)은 기업의 규모 분포를 분석하는데 실제 자료로부터 얻은 Zipf 분포와 이론적 Zipf 분포를 동시에고려하면서두터운꼬리(heavy tail)모형을 보다 명확하게 설명하기 위해 실제 빈도-순위 그림이 이론적 Zipf 그림을 상회하는 부분만을 고려하면서 실증분석을 수행하였다. 본 연구에서도 검색엔진을 통해 분석하는 대상이 일단 빈도가 높은 질의어가 중요한 의미를 가지고 있고, 다수의 소수 빈도를 가진 질의어의 왜곡을 막을 수 있다는 의미에서 그 방법을 적용할 수 있다. <그림 2>는 실제 빈도-순위 분포와이론적 Zipf 분포를함께그려본그림이다.

<그림 2>를 살펴보면 <표 1>과 <표 2>에서 빈도-순위 그림을 그리고 Pareto 분포로 적합 시켰을 때 보지 못했던 흥미로운 결과를 볼 수 있다. 1월 1일과 1월 11일이 다른 그림에 비해서 실제 빈도-순위 그림과 이론적인 Zipf 분포를 동시에 그렸을 경우 실제 빈도-순위 그림이 이론적 Zipf 분포를 많이 상회하고 있음을 볼 수 있다. 이러한 점은 이론적 Zipf 분



<그림 2> 이론적 Zipf 분포와 2주일간 실제 Zipf 분포

포를 이용해서 빈도-순위 그림의 차이를 판별할 수 있음을 보여줄 수 있는 한 예일 수 있다.

이를 보다 이론적으로 증명하기 위해 이론적 Zipf 분포를 상회하는 부분만을 이용해서 모형 (2)를 다시 적합한 결과는 <표 3>과 같다. <표 2>의 서로 다른 검색어수를 <표 3>과 비교해보면 약 5% 정도가 사용되었음을 알 수 있고, 기울기 추정치가 1에 많이 가까워지면서 모형 적합도도 높아졌음을 알 수 있다. 실제 추정에 사용된 자료의 수가 여전히 매우 많아 기울기간의 통계적 동등성을 검정하기 어렵지만 어느 정도 빈도가 존재하는 일부 상위 검색 질의어를 대상으로 할 때 기울기가 1인 Zipf 분포를 따를 수 있음을 보여준다. 1월 1일이나 1월 11일의 경우 기울기 추정치가 1.1 이상으로 나타나면서 다른 날과 다른 양상을 보여주고 있어 Zipf 모형을 이용해서 어느 정도 정상성을 벗어나는 날을 판단하는 기준으로도 사용할 수 있음을 보여주고 있다.

4. 결 론

하루에도 수천만 건 이상의 방대한 질의어가 입력되는 검색엔진을 통해 검색 패턴을 분석하는데 활용될 수 있는 질의어 분포의 정상성을 Pareto 분포와 Zipf 모형을 이용해 검토하였다. 실제 2주일간의 일간 검색어 분포를 빈도-순서 그림으로 그린 후 이를 먼저 Pareto 모형으로 적합 시킨 결과, 그 기울기가 1.33에서1.34 정도로일관되게측정되고있어Pareto 분포를 이용해 일간 검색 질의어 분포의 정상

<표 3> 상방 꼬리 부분으로 적합한 모형 추정치와 적합도

	$\ln(\hat{\alpha})$	$\hat{\beta}$	R^2	서로 다른 검색어수
2003.1.1(수)	13.3666	1.1283	0.9978	9,001
2003.1.2(목)	12.4784	1.0340	0.9989	7,490
2003.1.3(금)	13.1278	1.0882	0.9980	12,384
2003.1.4(토)	12.8120	1.0709	0.9981	8,193
2003.1.5(일)	12.9540	1.0894	0.9978	8,740
2003.1.6(월)	12.1852	1.0068	0.9996	7,289
2003.1.7(화)	12.6255	1.0427	0.9991	7,194
2003.1.8(수)	12.3239	1.0175	0.9997	7,153
2003.1.9(목)	12.4915	1.0317	0.9989	7,237
2003.1.10(금)	12.2362	1.0115	0.9994	7,385
2003.1.11(토)	13.4920	1.1265	0.9976	16,270
2003.1.12(일)	12.9539	1.0870	0.9979	8,745
2003.1.13(월)	12.0710	0.9941	0.9999	6,952
2003.1.14(화)	12.2238	1.0069	0.9995	7,026

성을 평가할 수 있음을 제시하였다. 그러나 이것의 기울기를 보다 1에 근접시키기 위해서는 Pareto 모형으로 일간 검색어 분포의 정상성을 분류하는 시도보다는 Pareto 모형에서 기울기가 1인 Zipf 모형을 이용해서 정상성을 판단하는 것이 보다 의미가 있다고 생각하여 빈도-순서 그림에 영향을 주면서 별 의미 없는 소수 빈도를 가지는 질의어를 제외하고 검색어 분석을 시도하였다. 어느 정도 소수 빈도를 가져야 분석에서 제외할지 그 일반적인 기준을 정하기 위해, Zipf 분포를 활용하는 분석에 많이 이용되는 빈도-순서 그림과 Zipf 분포를 동시에 그린 후 빈도-순서 그림이 Zipf 분포를 상회하는 질의어만으로 다시 분석하였다. 그 결과 일반적인 Pareto 모형보다 적합도도 높아지면서 극히 일부를 제외하고는 기울기가 1 근처에서 움직이고 있음을 실증적으로 확인하였고, 향후 일간(日間) 질의어 분포를 Zipf 분포를 활용하여 그 정상성 여부를 판단할 수 있음을 제시하였다.

본 연구에서 사용한 자료가 최근 경향을 반영하지 못할 수 있는 2003년 자료라는 점, 일년 중 특정한 시점에서 2주일을 사용한 점 등 자료 분석에 어느 정도 한계가 존재하지만 질의어의 분석이 빈도에 입각한 순위에 의존해서 분석되는 현실에서 Zipf 분포를 이용해서 전체 질의어 분포의 정상성을 탐색할 수 있는 기준을 제시했다는데 의의를 두고자 한다.

향후에는 질의어 분포를 보다 과학적이고 체계적으로 분석할 수 있도록 어느 정도 순위의 자료를 Zipf 분포를 추정하는데 사용해야 하는지에 대한 연구가 필요하다. 이 연구에서 Zipf 분포를 이용해 질의어의 분포의 정상성을 판단한 결과 Zipf 분포에서 기울기가 1.1 이상인 것이 두 날(日)로 나타났고, 이 날은 특별한 관심사가 형성될 요인이 잠재되어 있었으며, 이후 이러한 비정상성은 지속적으로 나타나지 않아 변동은 하루에 그친 것으로 분석할 수 있었다. 이처럼 Zipf 분포를 통해 질의어의 정상성을 확인할 수 있다면, 검색 엔진으로 수집된 각종 인터넷 검색 결과를 통해 사회적 관심사와 그 변동 흐름을 해석하는데 보다 객관적이며, 효율적인 자료를 제공해 줄 수 있을 것이다.

참 고 문 헌

박소연, 이준호, 2005. 국내 웹 이용자의 검색 행태 추이분석,『한국문헌정보학회지』, 39(2):146-160.

박소연, 이준호, 2002. 로그분석을 통한 이용자의 웹 문서 검색 행태에 관한연구,『정보관리학회지』, 19(3):111-122.

이준호, 박소연, 권혁성. 2003. 질의 로그 분석을 통한 네이버 이용자의 검색 행태 연구『정보관리학회지』, 20(2) : 28-41.

최희갑, 박원석. 2006. Zipf 분포를 활용한 지역별 사업체 규모분포 특성,『한국지역학회』, 22(2):275-299.

한국인터넷진흥원, 2007. 2007년상반기정보화실태조사 요약보고서.

Basasi, G. and R. Albert. 1999. "Emergence of scaling in random networks". Science, 286:509-512.

Champernowne, D. 1953.“ A model for income distribution ”. Economic Journal, 63:318-351.

Gabaix, X. 1999.“ Zipf's law of city sizes : An explanation ”. Quarterly Journal of Economics, 114(3): 739-767.

Gabaix, X. 2000.“ Zipf's law and the growth of cities ”. American Economic Review, 89(2): 129-132.

Hart, P.E. and N. Oulton. 1997.“ Zipf and the size distribution of firms ”. Applied Economic Letters, 4:205-206.

Jansen, B. J., A. Spink, A., & J. Pedersen. 2005.“ A temporal comparison of Alta Vista Web searching ”. Journal of the American Society for Information Science and Technology, 56(6): 559-570.

Reed, W. 2001.“ The Pareto, Zipf and other power laws ”. Economic Letters, 49:453-457.

Silverstein, C., M. Henzinge, H. Marais and M. Moricz. 1999.“ Analysis of a very large web search engine query log ”. SIGIR. Forum 33(1):6-12.

Spink, A. et al. 2001.“ searching the web: the public and their queries. ” Journal of the American Society for information science and Technology, 52(3):. 226-34.

Zipf, G.K. 1949. Human Behavior and the Principle of Least Effort. Reading : Addision Wesley.