한국지역지리학회지 제13권 제3호(2007) 290-300

인터넷 문서빈도를 통해 본 도시순위규모에 관한 연구

– 미국 10만 이상의 인구를 갖는 도시들을 사례로 –

홍 일 영*

Rank-Size Distribution with Web Document Frequency of City Name: Case study with U.S incorporated places of 100,000 or more population

Ilyoung Hong*

요약 : 본 연구는 인터넷 문서상에 나타나는 도시 지명의 문서 빈도를 통계량으로 도시규모에 대한 분석을 실시하였다. 검색어가 갖는 의미상의 차이에 따른 조건과 검색의 범위를 제약하면서 나타나는 유의적인 차이점들에 대해 분석하였고, 도시규모분포의 상관계수에 대한 분석을 통해 인구와 문서빈도와의 차이점을 분석하였다. 각 도시의 인구와 문서빈도와 상관관계 분석에서는 검색어의 종류를 보다 공간적의 의미로 제약할수록 더 높은 상관관계가 나타났고, 문서의 종류는 상용, 네트워크, 기관의 경우에 있어서 높은 상관관계가 나타났다. 그리고 인구와 문서빈도의 통계량을 이용한 군집분석을 통해서, 인구에 비해 더 많은 혹은 낮은 문서빈도를 보이는 도시들을 파악하였다. 이와 같은 분석은 웹 문서라는 정보통신사회 속에서 반영되는 각 도시의 특성을 분석하는 새로운 방안을 제시한다는 점에서 큰 의미를 갖는다고 할 수 있다.

주요어 : 문서빈도, 웹 마이닝, 도시규모분포, 지프의 법칙

Abstract : In this study, web document frequency of city place name is analyzed and it is used as the dataset for rank-size analysis. The search keywords are compared in the context of spatial meaning and the different domain corpus is applied. The acquired search results are applied for the further analysis. Firstly, the rank-size analysis is applied to compare the result between population and document frequency. Secondly, in case of correlation analysis, the significant changes are revealed when the spatial criteria for search keywords are increased. In case of corpus, COM, NET, and ORG shows the higher coefficient values. Lastly, the cluster analysis is applied to classify the list of cities that shows the similarity and difference. These analyses have a significant role in representing the rank-size distribution of city names that are reflected on the web documents in the information society.

Key Words : Document Frequency, Web Mining, Rank-Size Rule, Zipf's Law

1. 서 론

지프의 순위규모에 대한 법칙은 단어 사용에 관한 경험적 자료에 기초하여 도출된 것으로서 비록 이론적 기초가 미약하다는 점에서 많은 비판을 받지만 특정 현상에 대한 규모와 위계에 대한 일정한 규칙을 제시하는 분석방식이라는 점에서 많이 활용되는 방법들 중의 하나이다(Zipf, 1949). 이를 도시와 인구의 관계에 적용한 도시순위규모 법칙은 한 국가 내에 대도시와 중소도시의 차이가 존재한다는 것을 바탕으로 도시간 인구 규모의 정도에 따라 나타나는 일정한 규칙을 제시하고 그 기준에 근거한 도시간의 규모를 분석하는 방안을 제공한다(강지은 등, 2003). 우리나라의 경우 도시순위와 규모에 대한 특징은 서울을 비롯한 대도시에 과밀화를 특징으로 하는 종주도시에 대한 논의, 그리고 대도시와 중도도시와의 극심한 격차 등을 가장 큰 특징으로 제시하고 있다(권용우, 1999).

인터넷의 대중적인 보급 이후, 인터넷 상의 폭발적인 문서 수의 증가로 누적된 웹 문서로 구성된 데이터베이스 속에서 사용자가 원하는 웹 문서를 얼마나 빨리 찾아내고 또한 검색에 있어서 어느 정도의 정확성까지 갖출 수 있는가는 웹 검색 엔진에서 갖추어야 할 필수 요소가 되었다. 이와 같은 웹 상의 다양한 정보에 대한 분석은 최근 데이터 마이닝의 연구분야 중에서 웹 마이닝 연구로 발전하고 있다(김정자・이도헌, 1998; 윤종필 등, 1998). 웹 마이닝의 한 분야인 웹 내용분석에 있어

* 뉴욕주립대 버팔로대학 지리학 박사(Ph.D, Department of Geography, University at Buffalo, The State University of New York)(ilyoung.hong@gmail.com)

서 특정 용어가 웹 문서 상에서 나타나는 문서빈도는 각 검색 용어가 담고 있는 중요도를 표현한다고 할 수 있다. 따라서 웹 문서상에서 나타나는 각 지명의 빈도는 최근 급속하게 발달하는 정보화 시대 속에 누적된 웹 문서상에서 각 지명들이 갖는 중요성을 나타내는 지표라고 할 수 있다(Jicheng et al., 1999; Cooley et al., 1997).

Dodge(2000)는 인터넷이라는 사이버공간 속에 나타나는 여러 통계적 지표에 대한 분석을 시도한 가장 대표적인 지리학 관련 연구자라 할 수 있다. 국내의 경우, 허우긍(2004)의 하이퍼링크를 통한 도시네트워크 분석과 이희연·이용균(2004)의 인터넷의 확산의 따른 공간구조의 변화에 관한 연구 등이 대표적이다. 이러한 연구들은 인터넷이라는 공간상에서 나타나는 통계적 지표에 지리학의 분석방법을 적용하여 공간적인 특성을 분석한 연구라는 점에서 공통적인 특징을 보여주고 있다.

한편, 도시경제를 연구하는 지리학자들에게 있어서 도시간 위계 혹은 체계적 특성을 분석하는데 가장 중요하게 사용된 지표는 인구라는 변수이다. 인터넷이라는 사이버 공간 속에서 생성된 웹 문서는 사용자들의 다양한 활동을 반영하는 방대한 데이터베이스라고 할 수 있다. 이 같은 데이터베이스 속에 반영된 각 도시 지명의 문서빈도라는 지표가 도시의 규모를 나타내는 통계적 변수인 인구와 과연 어느 정도의 상관관계를 갖고 있을 것인지, 혹은 문서를 포함하는 도메인의 성격에 따라 도시들은 어떠한 그룹으로 유사하게 혹은 상이하게 구분지어질 수 있는 지 등과 같은 질문들은 정보화사회 속에서 도시가 갖는 특성을 표현하는 주요한 연구주제라고 볼 수 있을 것이다.

본 연구는 웹마이닝의 분석방법들 중에서 웹 내용을 분석하는 방법을 통해, 웹 페이지에 나타나는 각 도시지명의 문서빈도라는 통계적 자료를 기초로 하여, 지리학의 도시순위규모를 분석하는 방법론을 적용한 결과와의 상관관계를 비교함으로써 인터넷 웹 문서 속에 반영된 도시들의 특성을 분석하였다. 분석을 위한 자료는 가장 대중적인 검색엔진의 데이터베이스를 대상으로 하였고 이들이 제공하는 API(Application Programming Interface)[1] 및 검색 방식을 이용하여 미국 내 10만 이상의 인구를 갖는 254개의 도시들[2]에 대한 검색결과를 도시순위분석에 적용하여 인구를 기초로 하는 도시규모분석과 어떠한 상관관계를 보여주는가를 분석하였다. 이와 함께, 빈도분석을 위한 문서의 집합인 코퍼스 (corpus)[3]를 상이한 성격을 갖는 도메인으로 구별하고 검색이 발생할 수 있는 단어의 의미에 따른 특성을 구분하여 각각의 상이한 경우에 따라서 나타나는 도시인구규모의 특징과 인구와의 상관관계 그리고 도시들간의 유사성과 차이점을 군집분석을 통해 검토하였다.

2. 지프의 법칙과 인터넷 문서빈도

1) 지프법칙과 문서빈도

지프의 법칙(Zipf's law)은 단어의 집합으로 나타나는 문서의 집합인 코퍼스에 있어서 나타나는 특정 단어의 빈도는 전체적인 단어에 대한 빈도 테이블의 순위와 대체로 역비례의 관계로 나타난다는 것을 말한다. 다시 말해, 단어의 순위가 내려갈수록 사용 빈도수가 기하급수적으로 떨어진다는 특징을 밝힌 것이다. 지프의 법칙은 경험적인 측정에 의해 도출된 법칙이라 할 수 있으며, 단어를 모아 만든 코퍼스의 경우가 아닌 일반적인 다양한 현상 속에서도 관찰이 가능하다.

지프법칙의 가장 단순한 경우는 가장 일반적인 것에 대해서 다음의 것이 반의 빈도를 보이고 세 번째는 1/3, n번째의 것을 1/n의 빈도를 보인다는 것이다. 예를 들어 한 책에 수록된 단어를 조사해 본다면, 가장 많이 사용된 단어가 모두 1,000번 등장했다면, 두번째로 많이 사용된 단어는 1순위 단어 빈도수의 1/2인 약 500번, 세번째로 많이 나온 단어는 1순위 단어 빈도수의 1/3, 네번째로 1순위 단어 빈도수의 1/4만큼 등장한다는 것이다. 이렇게 점점 줄어들어서 나머지 대부분의 단어들은 극히 제한적인 횟수만큼 사용된다는 것이다. 다시 말해, 소수의 몇몇 단어가 지극히 많은 빈도로 사용되고, 나머지의 단어들은 적은 빈도로 활용된다고 할 수 있다. 이러한 측정 결과가 항상 정확하게 맞는 것은 아니지만, 상당히 많은 분야와 현상 속에서 적절한 근사치를 보이는 현상들을 발견할 수 있다. 지프의 법칙을 만족하는 특성은 도처에서 발견되고 이를 도시의 인구를 변수로 도시순위와 규모에

대한 규칙으로 제시한 것이 도시순위규모법칙(rank-size rule)이다. 도시규모분포는 다음의 수식으로 설명된다.

$$\log y = \log A - a \log x$$

여기서 x는 특정 인구의 크기, y는 x보다 큰 인구를 갖는 도시의 수, 그리고, A와 α는 상수로서 (A, α > 0)이다. 일반적으로, 도시의 분포가 α = 1인 경우를 순위규모로, α의 값이 1보다 크면 종주분포의 패턴을 나타나나는 것으로 설명하는 것으로 요약된다.

2) 지명에 관한 문서빈도

인터넷의 다양한 활용은 대중의 커뮤니케이션 방식에 영향을 주고, 정보를 수집하거나 물건을 구입하는 방식 등에 있어서도 많은 변화를 가져왔다. 인터넷 공간에서 많은 사용자들의 웹 브라우징 혹은 이메일의 사용 및 서버에 대한 접속현황에 대한 통계자료를 분석하는 과정에서 우리는 지프의 분포를 쉽게 찾아볼 수 있다(Adamic and Huberman, 2002). 초기에 인터넷의 특징을 사용자, 웹 사이트, 도메인 수 등의 폭발적인 증가와 같이 규모에 있어서 보여지는 양적인 면에서 급속한 성장의 특징들을 보여주었다면, 최근에 보여지는 인터넷 상에서의 변화는 사용자들이 인터넷을 사용하는 행태와 및 유형과 관련하여 새로운 특징들이 발견되고 있다. 이것은 적은 요소들이 거대한 부분들을 설명한다는 지프의 규칙을 반영하는 경우가 많다. 예를 들어, 몇몇의 사이트들이 수백만의 페이지들로 구성되어있다는 점, 혹은 몇몇 소수의 포탈사이트들이 수백만의 링크를 포함한다는 점 등은 가장 대표적인 사례라고 할 수 있다. 이러한 사실들은 역으로 해석하면 수많은 사이트들은 몇몇의 링크만을 갖고 있다고 말할 수 있으며, 수백만의 사용자들이 몇몇 소수의 사이트에 상당히 많은 접속률을 보여주고, 나머지 수백만의 사이트에는 별 접속이 없다는 사실 등은 모두 지프의 법칙을 인터넷의 활용 속에서 쉽게 발견할 수 있는 특징들이다.

한편, 초기 인터넷을 무한하게 연결된 하이퍼링크로 구성된 문서의 집합으로 표현한다면, 지금의 인터넷은 구글, 야후 혹은 국내의 경우 다음, 네이버와 같은 포탈 사이트를 통해서 원하는 정보에 도달하는 이른바 정보 검색의 시대라 할 수 있을 것이다. 링크를 통한 사이트 이동은 대부분 이와 같은 검색엔진에서 제공하는 검색결과를 통해 접속한 이후의 이동이며, 현재 웹 문서의 네비게이션에 있어서 가장 절대절인 가이드는 이 같은 검색엔진이 제공한다. 검색엔진은 검색을 위한 데이터베이스를 구성하기 위해 검색을 위한 모든 웹 페이지로 이동하여 그것을 읽고, 다시 각 페이지 상의 하이퍼텍스트 링크를 사용하여 그 사이트에 연결된 다른 페이지들을 읽어서 정보를 수집한다. 이러한 정보 수집은 스파이더(Spider), 크롤러(crawler) 또는 봇(bot) 이라고도 불리는 프로그램을 사용하게 된다. 이와 같은 프로그램들은 인터넷의 각 페이지에서 읽어 들인 정보에 대해 거대한 색인 또는 카달로그를 만들어서 사용자들의 다양한 검색요구에 대해서 자신들의 데이터베이스속의 색인내에 있는 내용과 비교한 후, 검색 결과를 돌려 주는 것이 일반적이다. 이와 같은 인터넷 자료의 검색을 위해서 주기적으로 각 검색엔진들은 인터넷 상의 문서들에 대하여 크롤러를 통해 이들을 수집하여 자신에 데이터베이스로서 저장하여 관리하게 된다(주길홍 등, 2004) 즉, 각 검색엔진들은 인터넷의 모든 문서들을 요약한 거대한 자료집을 자신의 코퍼스로서 갖고 있다고 할 수 있고, 이러한 거대한 데이터베이스는 인터넷 공간을 활용하는 정보화시대의 대중활동을 반영한 문서집이라고 할 수 있을 것이다.

검색엔진의 다양한 검색 방식 속에서도 검색엔진의 가장 기초적인 검색의 시작은 *TFxIDF*의 방식이라 할 수 있다(Salton and Buckley, 1988). *TFxIDF*는 단어빈도(Term Frequency)와 역문서빈도(Inverse Document Frequency)로 설명되는 정보검색에서의 검색결과에 대한 순위를 결정하는 수식이다. 단어빈도는 특정 문서 내에서 검색어가 몇 번이나 나타났는가를 계산하여 문서의 중요도의 수치로 사용한다. 즉 검색어 "Buffalo"가 *a*라는 문서에서 3번, *b*라는 문서에서 7번 나왔다면 문서 *b*가 *a* 보다 높은 순위로 평가하게 된다. 다음으로 역문서빈도란 검색어 "Buffalo"가 총 몇 개의 문서에서 나오는가 즉 문서빈도(DF: Document Frequency)

의 값의 역수를 적용하는 것으로 너무 많이 널려 있는 단어는 상대적으로 중요도를 낮추는 것을 의미한다. 이 수식은 검색 엔진의 가장 기본적인 방법으로 실제에서는 방식을 확장 및 변영하여 사용하게 된다. 요약하자면, 인터넷 검색엔진의 검색순위를 결정짓는 가장 중요시 되는 요인은 문서속에 포함된 검색어의 빈도라고 할 수 있고, 문서빈도라는 통계치는 역으로 검색어의 중요성 혹은 인지도를 반영한다고 할 수 있다.

Tezuka and Tanaka(2005)는 이 같은 문서검색 방식을 지리를 표현하는 단어들인 지명에 적용하는 연구를 수행하였고 문서 내 지명의 중요도와 그의 검색방식에 대해 다음의 <표 1>과 같은 5가지 방식을 제안하였다. 첫째는, 가장 단순한 방식으로 검색어의 문서빈도(DF)를 측정하는 것으로서 이 수치는 지명이 갖는 인지도에 비례하여 나타나게 되고 측정치는 텍스트마이닝(text mining)에서 아주 일반적으로 사용되는 방식이다. 이와 같은 문서빈도는 단어 그 자체가 갖는 비중을 파악하게 하는 지표로서 의미를 갖지만, 지명이 공간적 맥락인지 아닌지를 반영하지 못하는 문제점을 남긴다. 이러한 문제점을 고려하기 위해 공간적 의미를 갖는 단어를 추가하는 방식들이 제안된 것이 두번째 방식으로 검색어와 함께 자주 사용 가능한 단어와의 동시발생합계를 사용하는 근린지역용어와의 동시발생빈도 (RF)이다. 문서 속의 지명은 그 지역을 잘 설명하거나 혹은 밀접한 관계를 갖는 단어와 자주 언급되게 되고 이것은 그 단어의 사용을 공간적인 맥락으로 제한하고, 이러한 단어의 사용은 장소명의 애매함을 줄여준다고 할 수 있다. 셋째는 지역적 동시발생의 변동을 고려하기 위해, 지명이 공간적으로 인접한 주변의 단어와 함께 사용되는 경우를 계산하는 것이다. 예를 들어 검색에 있어서, 검색어에 검색지명을 포함하는 상위에 지명을 검색어에 포함시킨다거나 혹은 인접한 지명을 부가하는 방식이다. 넷째는, 공간적 문장 빈도(Spatial Sentence Frequency)를 계산하는 것으로서 중요한 장소명은 공간에 관한 문장 속에서 더욱 자주 인용된다는 점을 고려하는 것이다. 이것은 공간에 관한 주제의 문서들만을 추려낸 다음 이들 속에서 사용된 장소의 빈도를 계산한 경우를 의미한다. 다섯 번째는, 경우 빈도(Case Frequency)로서, 장소명은 문서의 구조상의 위치에 따라 중요성이 달라진다는 것으로, 문장 속에서 검색어가 어떠한 품사를 갖는가에 비중을 두는 방식이다. 이와 같은 예로서, 지명이 문장 속에서 어떠한 역할을 하느냐, 주어냐, 목적어냐, 장소명이나 등에 따라 상이한 비중을 주는 방식이다. 이와 같은 5가지 방식은 각 검색어를 포함하는 문장의 맥락을 고려함으로서 검색어에 대한 사용자의 공간적인 의도를 반영하는 특징을 갖는다고 할 수 있다.

3. 문서빈도 지표로 본 도시순위규모

1) 연구 자료와 연구방법

본 연구에서는 미국의 도시들 중 10만 이상의 인구를 갖는 254개의 도시를 분석의 대상으로 선

표 1. 지명의 인지도에 대한 5가지 측정방식

측정방식	약어	특징	예제
단순문서빈도	DF	검색어가 나타나는 문서의 수	Buffalo
지리개념용어와 동시발생 빈도	GF	검색어가 지리개념용어와 함께 사용되는 문서의 수	Buffalo City
근린지역용어와의 동시발생빈도	RF	검색어가 지역적으로 관련된 용어와 함께 사용되는 문서의 수	Buffalo New York
공간맥락을 의미하는 문장 빈도	SF	검색어가 공간적 맥락을 의미하는 문서 속에서 사용되는 경우만을 고려	The east of the Buffalo
주제어 경우 빈도	CF	검색어가 문장 속에서 주어, 목적어와 같은 핵심적 역할을 하는 문서만을 고려하는 경우	Buffalo is the oldest city

출처: Tezuka and Tanaka(2005)에서 인용

정하였다. 각 도시들에 대한 인구자료는 2000년 인구센서스 자료를 사용하였고, 각 도시 지명의 도시 빈도는 가장 많이 활용되는 2개의 검색사이트인 야후와 구글에서 제공하는 검색엔진을 사용하였다. 각 도시 지명의 빈도추출을 위해서는 검색엔진에서 제공하는 API를 사용하여, 입력된 도시명에 대한 빈도를 자동 추출할 수 있는 프로그램을 펄(Perl)언어[4]로 작성하여 추출하였다(Calishain and Dornfest, 2003). 엄격한 의미에서 각 도시의 인구와 문서빈도 간의 상관관계는 동일한 시점에서 얻어낸 자료를 사용해야하지만 미국통계청에서 일반인에게 공개하는 자료는 10년간의 주기로 제공하기에 인구는 2000년 센서스 자료를 활용하고 문서빈도의 경우 2007년 4월 12일의 결과를 이용하였다. 다음의 <그림 1>은 전체적인 자료의 흐름과 분석의 과정을 보여주는 것으로서, 각 도시에 대해서 상이한 코퍼스 별로 그리고 검색어에 대한 정의에 따라 반복해서 검색결과를 외부파일로 저장하고 이를 다시 통계 프로그램을 통해 분석하는 과정을 보여주고 있다.

각 지명에 대한 코퍼스는 국제적으로 많이 활용되는 5가지의 도메인으로 분류하여 254개의 도시들에 대한 검색결과를 취득하였다. 국제 도메인으로서 가장 대표적인 도메인은 COM, NET, ORG, EDU, GOV 등의 5가지 종류이다. COM은 상업 도메인, NET은 네트워크 관련, ORG의 경우 비영리기관 및 단체에 허가된 것으로 일반 도메인 등록을 통해 자유로운 등록이 가능하기에 비교적 많은 문서들이 존재하고, GOV와 EDU는 정부기관 및 학교 교육기관에 제한된 도메인이기에 상대적으로 위의 것들에 비해서는 적은 문서 수를 갖는다. 그러나 이들은 모두 인터넷 문서에서 가장 많이 활용하는 대표적인 도메인들이다. 각 도메인에 대한 검색은 Tezuka and Tanaka(2005)가 제시한 5가지 검색방식 중 3가지의 검색방식을 채택하여 각각 실행하였다. 첫째는 도시명 그 자체를 이용하는 방식으로서 도시명이 갖은 광범위한 의미를 모두 포괄하는 분석을 실행하였다. 둘째는 도시명에 “city”라는 검색어를 추가하여 검색하려는 범위를 “city”라는 도시의 의미를 포함할 수 있는 검색어로 제한하는 방식이다. 셋째는 각 도시명과 그 도시가 포함된 주(State)를 합성한 검색어로서 지리적으로 인접한 지명을 추가함으로써 공간적인 의미적 맥락을 추가하였다.

위에서 서술한 것과 같이 각각의 검색을 위해 코퍼스 별, 검색어의 의미 별 그리고 두가지의 검색엔진별로 구분된 검색어를 통해 획득한 자료를 이용하여 분석을 시행하였다. 분석은 우선 지프의 도시순위규모 분석을 위한 계수 a를 추출함으로써 각 항목별 계수치가 어떻게 나타나는 지를 분석하였다. 다음으로는 각 도시의 인구변수와 각 변수 별로 추출한 변수와의 상관관계를 추출하였고, 인구

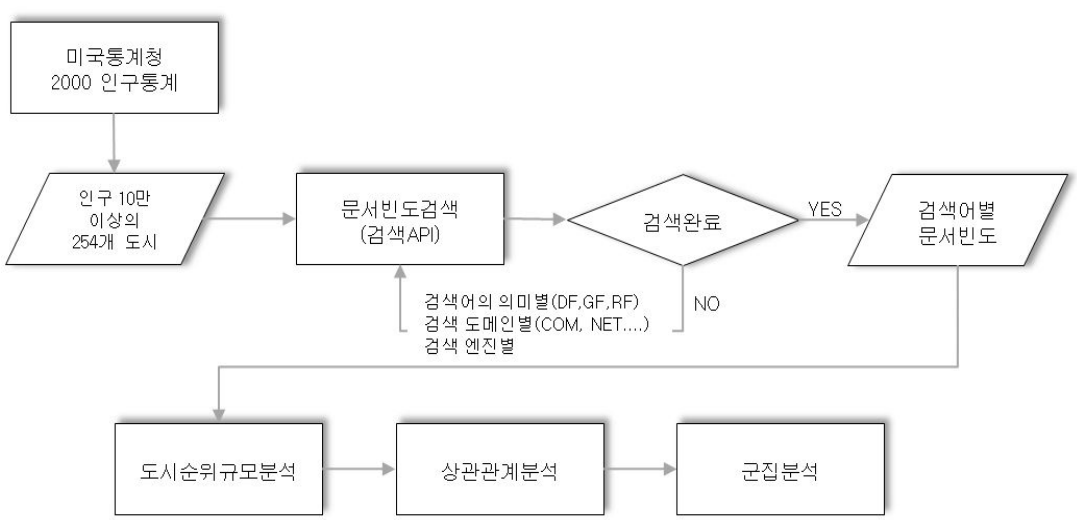


그림 1. 자료의 흐름과 분석의 과정

표 2. 도시순위규모 계수 α의 변화

	전체	COM	NET	ORG	EDU	GOV
인구	0.740	-	-	-	-	-
야후의 경우						
DF	1.242	1.230	1.351	1.313	1.361	1.062
GF	1.193	1.201	1.381	1.283	1.301	0.960
RF	0.998	0.984	1.316	1.077	1.175	0.933
구글의 경우						
DF	1.277	1.557	0.783	1.424	0.932	0.551
GF	1.610	1.315	0.814	0.863	0.974	0.629
RF	1.459	1.186	1.406	1.334	1.711	1.670

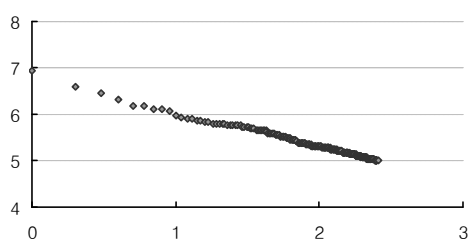
주: 2007.4.12일자 검색결과를 이용함

와 문서간에 상대적 비율에 따라 달라지는 도시들의 그룹을 비교하기 위해 군집분석을 실행하였다.

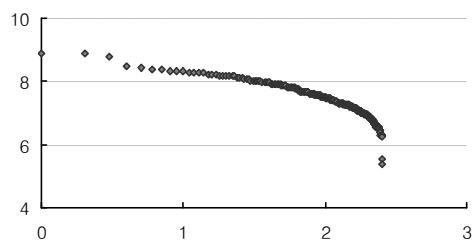
2) 연구결과

(1) 순위상관계수의 변화

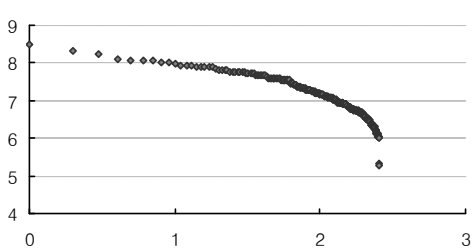
순위상관계수는 도시의 순위규모분포 현상을 구체적으로 설명해준다고 할 수 있다. 다음의 <표 2>는 각 검색어의 변화에 따르는 순위상관계수를 정리 및 요약한 것이다. 미국 도시들의 인구규모에 따른 순위상관분포는 계수 값이 0.740으로 해당하는 비교적 안정적인 단계에 있음을 보여주고 있다. 이에 비해, 문서빈도를 통해 얻어진 수치를 통해 본 도시간의 분포의 경우는 1 보다 큰 값을 보여줌으로써 비교적 상위에 위치한 도시명으로의 집중도가 높게 나타나고 있음을 알 수 있다. 또한 상관계수의 변화는 공간적 검색어를 추가할 수록 낮아지는 경향을 보이고 있고, 상업, 네트워크, 비영



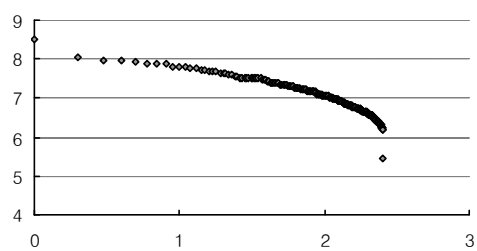
(a) 미국 254개 도시를 대상으로 한 도시인구규모분포



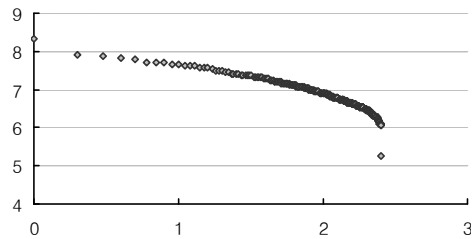
(b) 전체문서를 대상으로 한 문서빈도(DF)



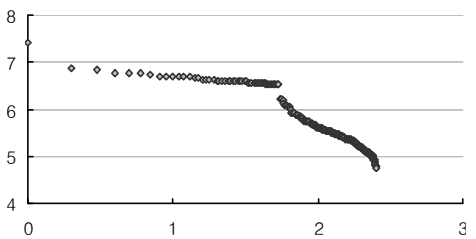
(c) 전체문서를 대상으로 한 지리용어동시발생빈도(GF)



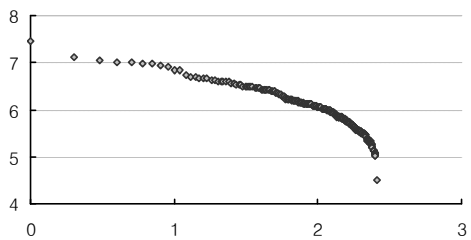
(d) 전체문서를 대상으로 한 지역용어동시발생빈도(RF)



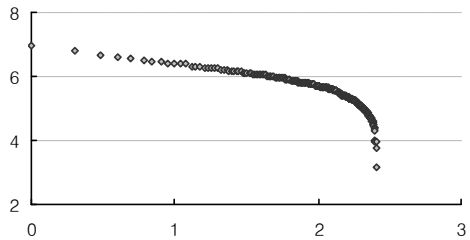
(e) 상업도메인(COM)을 대상으로 한 지역용어동시발생빈도(RF)



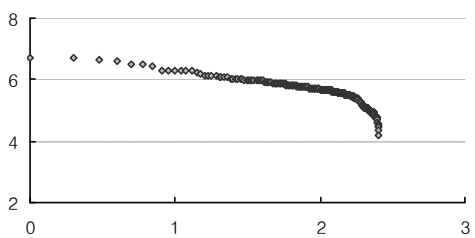
(f) 네트워크도메인(NET)를 대상으로 한 지역용어동시발생빈도(RF)



(g) 비영리기관(ORG)을 대상으로 한 지역용어동시발생빈도(RF)



(h) 교육기관(EDU)을 대상으로 한 지역용어동시발생빈도(RF)



(i) 정부기관(GOV)을 대상으로 한 지역용어동시발생빈도(RF)

(주: 그림 (a)의 경우 X축은 도시의 순위, Y축은 인구수를 로그값으로 변환한 값을 말하고, 그림 (b)-(i)의 경우 X축은 도시의 순위 Y축은 문서빈도값을 로그값으로 변환한 값을 의미한다. 상관계수의 수치상으로는 양 검색엔진간에 미묘한 차이를 보이지만 검색어 혹은 도메인에 변화에 따른 변동에는 큰 차이가 없기에 야후의 검색결과만을 제시하였다.)

그림 2. 인구와 문서빈도의 순위상관분포

리기관과 같은 도메인이 교육, 정부기관의 계수보다 높은 경향을 보여주고 있다.

다음의 <그림 2>는 인구와 검색어의 변화에 따른 순위상관분포의 변화를 그래프로 표현한 것이다. <그림 2>의 (a)에서와 같이, 미국의 도시인구규모분포는 비교적 직선에 가까운 안정적인 상태에 있음을 보여주고 있다. 문서빈도에 있어서는 공간적 제약을 추가할 수록 좀더 도시인구규모분포의 모습과 유사해지고 있음을 발견할 수 있다. 문서빈도의 검색 코퍼스의 변화에 있어서는 상업, 네트워크, 비영리 기관이 비교적 교육, 정부기관의 경우 보다 직선에 유사한 모습을 보여주고 있고, 교육과 정부기관의 경우 빈도수가 낮은 도시들의 경우 급격하게 수치가 낮아지는 현상을 발견할 수 있다.

(2) 도시명의 빈도와 도시인구와의 상관관계

다음의 <표 3>는 도시명의 빈도와 도시인구와의 상관관계를 분석한 결과를 요약해서 보여주고 있다. 우선, 가장 특징적인 것으로는 전반적으로

표 3. 도시명의 빈도와 도시인구와의 상관관계

	Total	COM	NET	ORG	EDU	GOV
야후의 경우						
DF	.639**	.676**	.716**	.668**	.617**	.382**
GF	.703**	.703**	.767**	.623**	.716**	.468**
RF	.828**	.816**	.833**	.773**	.752**	.525**
구글의 경우						
DF	.585**	.525**	.544**	.615**	.624**	.140*
GF	.650**	.730**	.712**	.727**	.419**	.286**
RF	.772**	.815**	.832**	.851**	.471**	.352**

(주: 표 안의 수치는 상관계수 (r)이다. **는 유의수준 1%에서의 유의함을 의미하고, *은 유의수준 5%에서 유의함을 의미한다.)

인구와 문서빈도간에 높은 상관관계를 보인다는 점을 들 수 있고, 둘째로, 검색어에 공간적 맥락을 주어질 수록 상관관계가 높아진다는 점, 마지막으로 코퍼스의 변화에 있어서는 상업, 네트워크, 비영리기관 등의 도메인에서 비교적 높은 상관관계를 보여주고, 교육과 정부기관의 경우 낮은 상관관계를 보이고 있다는 점을 들 수 있다. 전제적으로 도시명의 빈도와 인구와는 비교적 높은 상관관계를 보여주고 있고, 상관계수는 검색어에 대해 공간적인 특성을 부가할 수록 증가하고 있는 것이 특징적이다. 특히, 각 도시지명에 지역적으로 관계된 지명(각 도시가 속해있는 주의 이름)을 부가해서 검색했을 때는 아주 높은 상관관계를 보여주고 있음을 확인할 수 있었다. 다음으로는 검색도메인에 대한 제약에 있어서는 상업, 네트워크, 기관 등과 같은 도메인에 있어서 비교적 높은 상관관계를 보여주었고, 교육, 정부기관 등의 도메인 문서는 상대적으로 낮은 상관관계를 보여주었다. 검색엔진간

표 4. 군집별 판별분석 정확도와 오류율

		Predicted Group Membership			Total
		군집1	군집2	군집3	
Count	군집1	5	0	0	5
	군집2	3	232	4	239
	군집3	0	0	6	6
%	군집1	100	0	0	100
	군집2	1.255	97.071	1.674	100
	군집3	0	0	100	100

		Predicted Group Membership				Total
		군집1	군집2	군집3	군집4	
Count	군집1	5	0	0	0	5
	군집2	3	230	6	0	239
	군집3	0	0	4	0	4
	군집4	0	0	0	2	2
%	군집1	100	0	0	0	100
	군집2	1.255	96.234	2.510	0	100
	군집3	0	0	100	0	100
	군집4	0	0	0	100	100

에 차이에 있어서는 야후의 결과가 구글의 결과보다 상대적으로 높은 상관관계를 보이고 있었다.

분석의 과정에서 나타나는 문제점으로는 장소명만으로 분석을 했을 경우, 공간적 의미 외에 다양한 여러 가지 의미를 갖는 장소명의 경우(예를 들어, New York, Mobile) 등의 도시명에 있어서 인구수에 비해 높은 빈도를 보이는 경우도 나타났다. 또한, 지리개념용어와의 동시발생빈도(GF)의 분석에 있어서는 주(State)는 다른데 도시명이 같아서 인구수에 상관없이 결과가 같아지는 문제점(콜롬부스:오하이오, 콜롬부스:조지아)을 발견할 수 있었다. 마지막으로, 인지도가 낮은 도시들에 있어서 검색어가 두 단어 이상으로 길게 이루어진 지명의 경우 상대적으로 빈도가 낮아지는 문제점 등이 발생하였다.

(3) 군집분석

각 도시들이 지니고 있는 다양한 특성의 유사성을 바탕으로 동질적인 집단으로 구분하기 위해 군집분석을 실행하였다. 유사성의 측정방법을 위해서는 가장 일반적으로 사용되는 유클리디안 거리(Euclidean distance)를 사용하였으며, 거리 계산시 변수 값을 표준화여 분석하였다. 각 도시 별 인구와 야후 검색엔진의 결과 중 근린지역용어와의 동시발생빈도(RF) 검색의 대한 전체 문서빈도를 표준화한 Z-Score점수를 사용하였다. New York City의 경우 유사성의 정도가 월등하게 떨어져 나타나기에 제외하고 분석하였다. 군집화는 가장 보편적으로 사용되는 agglomerative clustering에 의한 계층적 군집 방법을 사용하였고 가장 적절한 군집수의 적합성을 파악하기 위해서는 판별분석을 실행하였다. 판별 분석의 결과 다음의 <표 4>과 같이 군집 수가 3개 일 때 정확성은 97%로 오류율은 3%이고 군집수가 4개일때는 각각 96.2%와 3.8%로 나타났다. <표 5>은 각 군집에 따라 나타나는 Wilks Lamda 값과 정준상관계수를 보여주고 있는데, 여기서는 군집이 4개인 경우가 3개인 경우보 더 높은 상관관계와 보다 더 강한 군집간에 응집도를 보여주고 있다.

판별분석의 결과를 반영하여 4개의 군집유형을 따라서 도시들을 유형화하여 분류하였다. 각 군집에 해당하는 도시들과 인구 및 문서빈도의 평균값은 다음의 <표 6>을 통해 확인할 수 있다. 군집1에 해당하는 도시들은 비교적 많은 인구수와 문서빈도를 보여주지만 인구의 평균이 더욱 크게 나타나는 도시들로 Philadelphia, Phoenix, San Diego 등이 사례로 나타났다. 군집 2의 경우, 인구와 문서빈도가 다 비교적 적게 나타나는 도시들로서 254개의 도시 중 230개 즉 96%정도가 해당된다. 군집 3의 경우 인구와 빈도가 높은 중에서도 문서빈도

표 5. 판별분석에 따른 각 군집별 Eigenvalue와 Wilk's Lambda

	Eigenvalue				Wilks' Lambda			
군집	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation	Wilks' Lambda	Chi-square	df	Sig.
3개 군집	1.112	53.927	53.927	0.726	0.243	348.986	4	0.000
4개 군집	1.390	56.287	56.287	0.763	0.201	394.503	6	0.000

표 6. 각 군집별 도시명과 인구 및 지역동시발생빈도(RF)의 Zscore 평균

	도시명	인구평균(Zscore)	RF평균(Zscore)
군집1	Philadelphia, Phoenix, San Diego, San Antonio, Dallas	4.797	1.840
군집2	San Jose, Detroit, Indianapolis, Jacksonville, Columbus, Memphis, Baltimore, Fort Worth, Charlotte, Nashville, El Paso, Boston, Washington, Milwaukee, Denver, Louisville, Las Vegas 외 228개	0.931	-0.146
군집3	San Francisco, Austin, Seattle, Oklahoma	1.702	3.672
군집4	Jersey, Kansas	-0.273	5.445

가 크게 나타나는 도시들로서 San Francisco, Austin, Seattle, Oklahoma 등의 도시가 이에 해당한다. 마지막으로 군집 4의 경우 역시 빈도가 더욱 높게 나타난 도시들로서 Jersey, Kansas 등이 이에 해당한다. 군집 3의 경우, San Francisco의 경우 구글의 본사가 위치한 도시이고, Seattle의 경우 마이크로소프트사의 본사가 위치한 도시란 점은 정보통신 관련 기업을 유치한 도시에서 인터넷 상의 문서빈도가 인구에 비해 높게 나타나게 하는 요인으로 작용할 수 있음을 보여준다고 할 수 있을 것이다.

4. 결 론

본 연구는 인터넷 문서상에서 나타나는 도시지명의 빈도라는 통계치를 데이터베이스로 변환하여 인구를 이용한 도시순위규모와의 상관관계를 분석하였다. 첫째로, 순위상관계수의 분석에 있어서 인구의 경우 1 보다 낮은 값을 보인 반면, 상업, 네트워크, 비영리기관 등의 도메인들에서 1 보다 큰 값이 나타났고, 교육, 관공서 등의 도메인에서 1에 가까운 수치를 보여주었다. 또한, 공간적 의미를 갖는 검색어를 사용할 수록 1의 값에 가까워지는 변화를 볼 수 있었다. 둘째로 상관관계 분석에 있어서는 전체적으로 도시명의 빈도와 인구와는 비교적 높은 상관관계를 보여주었고, 상관계수는 검색어에 대해 공간적인 제약을 부가할 수록 증가하고 있는 것이 특징적이었다. 특히, 각 도시지명에 지역적으로 관계된 지명(각 도시가 속해있는 주의 이름)을 부가해서 검색했을 경우에는 높은 상관관계가 나타나고 있음을 확인할 수 있었다. 다음으로는 검색도메인에 대한 제약에 있어서 상업, 네트워크, 기관 등과 같은 도메인에 있어서 비교적 높은 상관관계가 나타났고, 교육, 정부기관 등의 도메인 문서는 상대적으로 낮은 상관관계를 보여주었다. 셋째로, 군집분석에 있어서는 크게 4가지의 군집으로 분류되었는데 San Francisco, Austin, Seattle, Oklahoma 등의 도시들이 인구에 비해 문서빈도가 크게 나타나는 특징적인 도시들로 나타났다.

도시의 규모를 연구하는 데 있어서 인구라는 변수가 일정기간을 갖고 많은 비용을 들여 조사원을 통해 수집하는 정보라고 한다면, 웹 문서에서 나타나는 각 도시지명의 문서빈도라는 정보는 온라인상에서 자발적으로 누적되는 정보로서 이들 자료에 대한 분석은 상대적으로 적은 비용과 자동화된 방식으로 각 시점에서 빠른 정보의 수집이 용이하다. 따라서, 점진적으로 변화하는 시계열별 도시규모의 변화를 예측하는데 유리한 장점을 제공한다. 그러나, 그 분석의 결과에 대한 신뢰도에 있어서는 몇 가지의 문제점들이 나타났다. 우선, 영문 문서에 대한 미국 도시명에 대한 분석이기는 했으나, 상업, 네트워크, 기관 등의 경우 미국이 아닌 다른 나라의 문서를 포함할 가능성을 배제하지 못한 점 등의 문제점을 갖는다. 다음으로, Tezuka and Tanaka (2005)가 제안한 5가지 방식 중 공간적 문장빈도(SF)와 주제어빈도(CF)에 대한 분석이 이루어지지 못한 점, 각 도시가 갖는 경제, 교육, 정부기관에 따른 외부적인 변수들과 각 도메인에 따르는 변이에 대한 추가적인 연관관계에 대한 분석 등은 본 논문에서 미처 논의하지 못한 것으로, 향후의 연구과제라 할 수 있다. 끝으로, 본 연구는 인터넷이라는 사이버 공간에서 공간에서 투영된 지리적 특징에 대해 지리학의 분석기법을 적용하여 분석을 했다는 점 그리고 자동화된 방식을 통해 빠르게 접근할 수 있는 방안을 제시한다는 점 등에서 연구의 의미를 갖는다.

註

1) API(Application Programming Interface), 응용 프로그래밍 인터페이스는 외부에서 각 응용 프로그램이 제공하는 기능을 접근하고 제어할 수 있도록해주는 인터페이스를 말한다. 이와 같은 인터페이스는 외부 소프트웨어와의 통합 및 기능의 확장을 가능하게 한다. 본 연구에서는 구글(Google)과 야후(Yahoo)에서 제공하는 API를 이용하여 지명의 문서빈도를 취득하였다.

2) 254개 도시들에 대한 인구통계자료는 미국 통계청 홈페이지에서 제공하는 2000년 통계자료를 사용하였다. 자료원: http://www.census.gov/population/www/cen2000/phc-t5.html

3) 코퍼스(corpus)란 전산언어학에 있어서 문서빈도와 같은 통계적 분석의 대상이 되는 문서의 집합을 의미한다. 본 연구에서는 각 검색엔진이 제공하는 문서집합을 코퍼스로 상정하고 분석을 실시하였다.

4) 펄(Perl)언어는 래리 월이 만든 인터프리터 방식의 프로그래밍 언어로서, 단어나 문자열 처리에 있어서 뛰어난 처리능력을 갖기에 전산언어학 분야에서 많이 활용된다. 최근 인터넷의 발달로 인해 HTML의

문자열 처리를 위해 많이 사용되는 CGI프로그램으로서 더욱더 알려져 있다.

文 獻

강지은·윤철현·허성곤, 2003, 도시산업구조가 도시순위에 미치는 영향에 관한 연구, 대한국토도시계획학회지(국토계획), 38(7), 187-200.

권용우, 1998, 한국도시의 순위규모법칙 1789 - 1995, 지리학연구, 32(1), 57-70.

김정자·이도헌, 1998, 데이터 마이닝 기술 및 연구동향, 정보과학회지, 16(9), 6-14.

윤종필·김희숙·최옥주, 1998, 데이터 마이닝의 유용성, 정보과학회지, 16(9), 15-23.

이희연·이용균, 2004, 인터넷의 확산에 따른 디지털 격차와 공간구조의 변화, 한국지역지리학회지, 10(2), 407-427.

주길홍·이준휘·이원석, 2004, 스타일 기반 키워드 추출 및 키워드 마이닝 프로파일 기반 웹 검색 방법, 정보처리학회논문지, 11(5), 1049-1062.

허우긍, 2003, 인터넷 하이퍼링크로 본 도시 네트워크, 대한지리학회지, 38(4), 518-534.

Adamic, L.A. and Huberman, B.A., 2002, Zipf's law and the Internet, *Glottometrics*, 3, 143-150.

Dornfest, R., Bausch, P., and Calishain, T., 2003, *Google Hacks: 100 Industrial-Strength Tips & Tricks*, O'Reilly.

Cooley, R., Mobasher, B., and Srivastava, J., 1997, Web mining: information and pattern discovery on the World Wide Web, *Ninth IEEE International Conference*, 558-567.

Dodge, M. and Kitchin, R., 2000, *Mapping the Cyberspace*, Routledge, London.

Jicheng, W., Yuan, H., Gangshan, W., and Fuyan, F., 1999, Web mining: knowledge discovery on the Web, *IEEE SMC '99 Conference Proceedings*, 137-141.

Salton, G. and Buckley, C., 1988, Term-weighting approaches in automatic text retrieval, *Information Processing & Management*, 24 (5), 513-523.

Steven, B., Harry, G., Marrewijk, C., and Berg, M., 1999, The Return of Zipf: Towards a Further Understanding of the Rank-Size Distribution, *Journal of Regional Science*, 39 (1), 183-213.

Tezuka, T., and Tanaka, K., 2005, Landmark Extraction: a web mining approach, *Spatial Information Theory: Lecture Notes in Computer Science*, 3693, 379-396.

Zipf, G.K., 1949, *Humanbehavior and the principle of least effort*, AddisonWesley, Cambridge.

• 교신 : 홍일영(뉴욕주립대 버팔로 대학 지리학과 박사, ilyoung.hong@gmail.com) Correspondence: Hong, Ilyoung (Ph.D, Department of Geography, University at Buffalo, The State University of New York, ilyoung.hong@gmail.com)

(접수 : 2007. 5. 28, 채택 : 2007. 6. 11)