인용 네트워크 분석에 근거한 문헌 인용 지수 연구*

A Study on Document Citation Indicators Based on Citation Network Analysis

이 재 윤(Jae Yun Lee)**

목 차

초 록

이 연구는 최근 발표된 단일 문헌에 대한 인용 영향력을 측정하는 여러 인용 지수에 대해서 각 지수의 특성과 지수 간 관계를 살펴보는 것을 목적으로 한다. 분석 대상 인용 지수로는 페이지랭크, SCEAS Rank, CCI, f-값, 단일 논문 h-지수의 다섯 가지와 h-지수를 변형한 세 가지 지수를 더하여 8가지를 포함하였다. 우선 단일 문헌에 대한 인용 영향력을 측정하는 다섯 가지 인용지수에 대해서 살펴보고 단일문헌 h-지수를 변형한 단일문헌 h_S-지수, h1-지수, h_S1-지수의 세 가지를 추가로 제안하였다. 각 인용 지수의 특성을 파악하기 위해서 국내 인용 데이터베이스인 KSCI 데이터베이스를 대상으로 실제 네트워크 인용 지수를 측정해보았다. 상관관계 분석과 군집분석을 수행하여 8가지 인용 지수 사이의 관계를 분석한 결과, 유사한 측정 행태를 보이는 인용 지수 군을 파악할 수 있었다. 또한 인용빈도 요인과 각 인용 지수 간의 상관관계 분석을 통해서 각 지수의 특성을 설명하였다. 마지막으로 인용 지수의 적용을 위한 고려사항과 후속 연구 방향을 제안하였다.

ABSTRACT

This study identifies the characteristics of recent citation-based indicators for assessing a single paper in the context of their co-relationships. Five predefined indicators were examined with three variants of h-index which are convened in this study; the formers are PageRank, SCEAS Rank, CCI, f-value, and single paper h-index and the latters are h_S-index, h1-index, and h_S1-index. The correlation analysis and cluster analysis were performed to group the indicators by common characteristics, after which the indicators were calculated with the dataset from KSCI DB. The results show statistical evidence that distinguishes h-index type indicators from others. The characteristics of the indicators were verified with citation frequency factors using correlation analysis. Finally, the implications for applications and further studies are discussed.

키워드: 인용분석, 문헌 인용 네트워크, 영향력 지수, 페이지랭크, h-지수, 네트워크 분석, 단일 문헌 평가
Citation Analysis, Document Citation Network, Impact Factor, PageRank, h-Index, Network Analysis, Single Paper Assessment

---

* 본 연구는 2008학년도 경기대학교 학술연구비(일반연구과제) 지원에 의하여 수행되었음. 또한 본 연구는 행정안전부가 지원하는 한국정보화진흥원의 국가DB사업의 일환으로 한국과학기술정보연구원이 구축하여 제공하는 과학기술 참고문헌 인용색인 DB정보를 이용하였음.

** 경기대학교 문헌정보학과 부교수(memexlee@kgu.ac.kr)

논문접수일자: 2011년 4월 19일 최초심사일자: 2011년 4월 19일 게재확정일자: 2011년 5월 11일
한국문헌정보학회지, 45(2): 119-143, 2011. [DOI:10.4275/KSLIS.2011.45.2.119]

1. 서 론

SCI를 비롯한 Thomson사의 인용색인 DB가 학술활동과 그 평가에 미치는 영향력에서 알 수 있듯이, 인용빈도가 학술활동에 대한 계량적인 척도처럼 인식되고 있는 것이 현실이다. 심지어 최근 국내 일부 대학에서는 논문이 게재된 학술지의 ISI 영향력 지수 수준에 따라서 교수의 연구업적 평가에 가산점을 차등적으로 부여하는 정책을 채택하는 것으로 나타났다(윤희윤, 김신영 2005). 이와 같이 ISI 영향력 지수를 연구자 개인의 성과를 평가하는 용도로 사용하는 것에 대해서는 SCI DB를 제공하는 Thomson사 측에서도 무리한 적용이므로 주의할 것을 당부한 바 있다(Adam 2002). 무엇보다도 학술지의 영향력 지수가 개별 논문의 인용 영향력을 대변하는 것은 아니기 때문이다(Seglen 1993; 1997). Neylon과 Wu(2009)도 학술지를 대상으로 하는 어떤 인용 지표도 개별 논문의 중요성을 평가하기 위한 용도로 쓰여서는 안 된다고 하였다. 그럼에도 불구하고 한국연구재단과 교육과학기술부에서는 특정 연구개발 사업을 평가할 때 사업 성과 논문이 게재된 학술지의 영향력 지수(Impact Factor)를 해당 사업의 질적 측정 지표로 활용하고 있는 실정이다.

나아가서 인용 분석 자체에 대한 전통적인 비판으로서, 인용빈도는 문맥을 반영하지 않으므로 비판을 포함한 다양한 이유에서의 인용을 동일하게 취급한다는 문제가 있다(Neylon and Wu 2009). 정경희(1999)도 평가의 목적으로 인용분석을 사용할 경우 인용이 가치 인정을 의미한다는 가정은 많은 논의를 불러일으키는 문제라고 지적하였다. 학술지의 경우에는 평균적으로 긍정적인 이유에서의 인용 비율이 높을 것이므로 그 중의 일부인 비판적인 인용이 큰 문제가 되지 않는다. 그러나 문헌 단위에서는 비판적인 인용이 많은 논문도 영향력이 높은 것으로 오인될 여지가 있다. 이에 따라서 일부에서는 논문에 대한 평가를 위해서 이용 통계를 보완적인 지표로 사용하려는 시도가 나타나고 있다. 이용 통계에는 원문제공 사이트의 클릭 빈도와 다운로드 빈도를 사용하는 경우(Bollen and Sompell 2008)와, Zotero나 CiteULike와 같은 웹 2.0 서지정보 관리 사이트의 북마크 빈도를 분석한 경우(Priem and Hemminger 2010)가 있다.

아직까지 이용 통계가 논문의 질적 수준이나 영향력에 대한 학술적인 지표가 되기에는 데이터베이스의 수록 범위 및 이용 가능성 등을 비롯하여 해결해야할 문제가 많다. 따라서 현실적으로 논문 단위의 연구성과에 대한 계량적인 측정 수단으로 인용을 대신하여 이용 통계를 사용하기는 어렵다. Moed(2005)는 인용분석 자체가 평가라고 할 수는 없지만, 계량서지적 지표가 학술활동을 평가할 때 질적인 수준에 대한 통찰을 얻고 판단을 내리는 데 도움을 받는 하나의 연구 평가 도구는 될 수 있다고 언급하였다.

결국 인용을 수단으로 하여 단일 논문의 영향력을 평가하려는 목적으로는 게재된 학술지에 대한 인용이 아닌 해당 논문에 대한 인용을 근거로 하는 것이 바람직하다. 물론 논문에 대한 인용을 파악하기 위해서는 논문이 발표된 후 일정 기간이 지나야만 한다는 한계가 있다. 그러나 게재된 학술지의 영향력 지수를 논문에 대한 평가 근거로 사용하는 오류를 방지하기

위해서라도 논문에 대한 인용 영향력 분석 방안이 개발되어야 한다.

학술정보서비스에 있어서도 개별 논문의 인용 영향력 측정 결과는 연구자에게 유용한 정보가 될 수 있다. 최근 많은 학술지들이 웹사이트에서 "Most cited article" 등의 이름으로 가장 인용을 많이 받은 논문에 대한 정보를 제공하고 있다(정희경, 이춘실 2009). 비록 단순 인용 빈도에 근거하고 있긴 하지만 이와 같은 인용 분석 서비스는 이용자에게 개별 논문의 가치에 대한 판단 근거를 제공하므로 논문 조사 과정에서 유용한 자료를 파악하는데 도움이 된다.

개별 논문의 영향력 측정은 최근 활발히 이루어지고 있는 지적구조 분석 연구와 지식지도 관련 연구에도 도움이 될 수 있다. 특히 저자나 단어를 분석 단위로 하는 지적구조 분석 연구(김판준, 이재윤 2007; 김희전, 조현양 2010; 유종덕, 최은주 2011; 이재윤 2008)에서는 주제적인 구조만 제시될 뿐, 지적구조나 연구동향 내에서 개별 연구의 영향력이나 가치에 대한 접근은 다루지 못하고 있다. 따라서 기존의 지적구조 연구에 논문의 영향력 측정 결과를 결합한다면 더 유용한 분석이 가능할 것이다.

인터넷 검색엔진 구글의 개발자들이 제시한 페이지랭크(PageRank) 알고리즘(Page et al. 1999)은 인용과 유사한 하이퍼링크 네트워크에서 각 노드의 구조적 중요도를 측정하기 위한 방식으로서 단순 빈도 이상의 정보를 제공하는 것으로 인정되고 있다. 이를 서지데이터베이스에 적용하려는 시도는 특히 학술지 단위에서 큰 성과를 얻어서 Scopus 데이터에 대한 SCImago Journal Rankings나 Web of Science 데이터에 대한 Eigenfactor score와 같은 실제적인 응용도 개발되었다. 그러나 학술지가 아닌 단일 논문을 대상으로 페이지랭크 방식을 적용하는 것에 대해서는 비판적인 의견이 다수 제기되었다(Sidiropoulos and Manolopoulos 2006). 이와 함께 논문 인용 네트워크에 대한 구조적 인용 지수로서 페이지랭크를 개선하거나 대체하고자 하는 시도가 최근들어 계속해서 이루어지고 있다.

이 연구에서는 이와 같은 페이지랭크 알고리즘과 그 대안 지수들을 검토하여 특징을 살펴본 후, 국내 인용데이터베이스에 대한 적용 실험을 통해서 단일 논문의 영향력을 측정할 때 고려할 사항과 각 지수 간 관계를 파악하였다. 이를 통해서 국내 학술 논문 데이터베이스에서 논문의 인용 영향력을 측정하기 위한 방안을 제시하였다.

2. 논문 인용 영향력 측정 지수

이 장에서는 페이지랭크를 비롯하여 논문의 인용 영향력을 측정하는 여러 지수에 대해서 살펴보고, 일부 지수는 영향력 측정 결과의 변별력을 높이기 위해서 변형하는 방안도 제시하였다. 이 논문에서 다룬 인용 지수 이외에 정준민(2010)의 연구에서 제시한 방식도 대상이 될 수 있으나, 그 경우는 인용한 논문의 수명까지 고려하는 방식이어서 기존 지수와 직접 비교하기 어려운 것으로 판단하여 제외하였다. 각 절에서는 우선 각 인용 지수의 개념과 공식 및 산출 방식을 살펴보았다. 인용 지수를 수식으로 표현할 때에는 〈표 1〉에 제시한 항을 공통으로 사용하여 공식의 비교가 용이하도록 하였다.

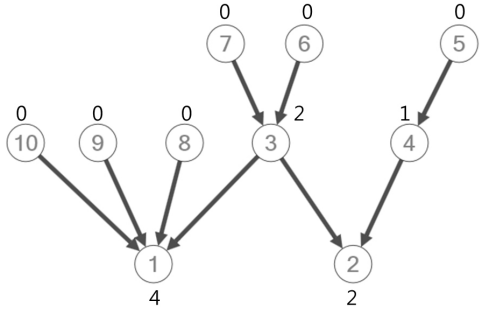
〈표 1〉 인용 지수 산출에 사용되는 항

항	의미	비고
n	전체 문헌 수	
d_i	i번째 문헌	
d_j	특정 문헌을 인용하는 문헌들 중에서 j번째 문헌	
$W(d_i)$	문헌 d_i의 영향력 가중치	각 인용 지수의 측정 결과값임.
$CO(d_j)$	문헌 d_j의 참고문헌 수	
$CI(d_j)$	문헌 d_j의 인용빈도	
$CR(d_j)$	특정 문헌을 인용하는 문헌들 중에서 문헌 d_j의 인용빈도 순위	h-지수와 그 변형지수에서만 사용됨.
b	(반복 계산에서) 인용빈도가 0인 논문으로부터의 인용 영향력이 0이 되지 않도록 하는 보정 상수	SCEASRank에서 사용되며 f-값이나 h1-지수, hs1-지수에서 1을 더하는 것과 비슷한 의미임. 페이지랭크의(1-d)/n도 유사한 역할을 함.
a, d, RF, β	0~1 사이의 상수(영향력 감쇄용)	d는 페이지랭크와 SCEASRank에서, a는 SCEASRank에서, β는 CCI에서, RF는 f-값 공식에서 사용함.

〈그림 1〉은 이하 각 절에서 다룰 네트워크 인용 지수를 설명할 때 사용할 가상의 인용 네트워크로서 화살표는 인용이 이루어지는 방향을 나타낸다. 인용 빈도를 기준으로 하면 이 네트워크에서 가장 인용빈도가 높은 것은 1번 논문(4회)이며 2번 논문과 3번 논문이 각각 2회씩의 인용 빈도로 그 다음 순위를 차지하고 있다.

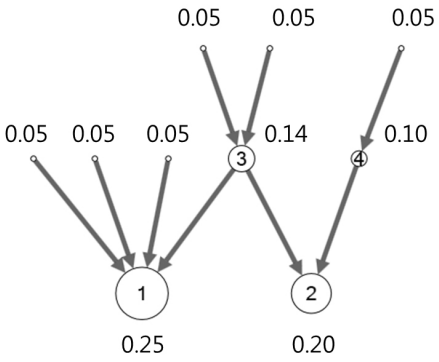
2.1 페이지랭크

페이지랭크(PageRank) 공식은 웹 사이트를 대상으로 중요도를 측정하기 위해서 개발한 것이다(Page et al. 1999). 이를 서지 인용에 적용하여 학술지나 저자의 중요도를 측정하기 위한 다양한 시도가 이루어지고 있다(Chen et al.



(원 안의 숫자는 논문 번호, 원 밖의 숫자는 인용빈도)

〈그림 1〉 가상의 인용 네트워크 사례



(원 안의 숫자는 논문 번호, 원 밖의 숫자는 지수 값 - 이하 동일)

〈그림 2〉 그림 1의 인용 네트워크에서 논문별 페이지랭크값을 측정한 결과

2007; Dellavalle et al. 2007; Ding et al. 2009). 페이지랭크 공식의 계산식(Langville and Meyer 2006)을 〈표 1〉의 항을 사용하여 나타내면 다음과 같다.

$$W(d_i) = \frac{1-d}{n} + d \times \sum_j \frac{W(d_j)}{CO(d_j)}$$

이 공식에서 d는 보통 0.85로 설정된다. 각 문헌의 페이지랭크 값 $W(d_i)$를 실제로 산출하기 위해서는 반복 계산을 통해서 각 단계마다 이전 단계의 각 문헌의 가중치를 현 단계에서 인용하는 문헌으로 보내주는 작업을 수행해야 한다. 반복 계산 시 초기 가중치는 모든 문헌에 동일하게 1/n로 설정한다. 내보내는 인용이 없는 문헌의 가중치는 전체 n개의 문헌에게 고르게 분산되도록 설정한다(Langville and Meyer 2006). 반복 계산을 통해서 가중치의 변화가 일정한 값 이하로 수렴되면 계산을 종료한다. 네트워크 내에 순환 고리가 없이 한 방향으로만 인용 링크가 연결되어 있을 경우에는 비교적 빠르게 수렴이 이루어진다.

〈그림 1〉의 가상 인용 네트워크 사례에 대해서 페이지랭크 값을 산출해보면 〈그림 2〉와 같다. 인용빈도가 4회로 가장 높은 1번 논문이 영향력 1위로 나타나며, 2번과 3번 논문은 인용빈도가 2회씩으로 동일하지만 3번 논문을 인용하는 두 논문보다 2번 논문을 인용하는 두 논문의 영향력이 더 크므로 2번 논문이 더 높은 영향력을 가진 것으로 나타난다.

2.2 SCEAS_B1

Sidiropoulos와 Manolopoulos(2005; 2006)는 서지 데이터에 적용할 때 페이지랭크의 3가지 문제점을 다음과 같이 지적하였다. 첫째, 인용이 돌고 도는 순환 고리에 포함된 개체에 유리하다. 둘째, 직접 인용빈도의 가치를 너무 낮게 간주한다. 셋째, 한 개체의 중요도가 바뀌는 것이 여러 인용단계를 거쳐서 멀리 떨어져 있는 개체에게까지 전파된다. 이중에서 특히 두 번째 문제점에 주목해서 직접 인용빈도의 가치를 더 반영하도록 보완한 평가 지수가 SCEAS Rank이다(Sidiropoulos and Manolopoulos

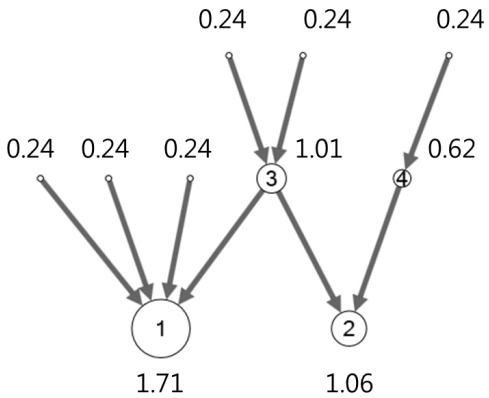
2005; 2006). 이들은 개별 문헌뿐만 아니라 문헌의 집합인 저자나 학술대회를 대상으로 평가하기 위한 용도로 SCEAS Rank 방식을 개발하였다. 여기서 SCEAS는 이들이 구축한 시스템인 Scientific Collection Evaluator with Advanced Scoring의 약자이다. SCEAS Rank 공식은 특정 파라미터의 포함 여부에 따라서 여러 형태가 있는데, 가장 복잡한 형태를 〈표 1〉에 제시한 항으로 표현하면 다음과 같다.

$$W(d_i) = (1-d) + d \times \sum_j \frac{W(d_j)+b}{CO(d_j)} \times a^{-1}$$

여기서 b는 '직접 인용 강화 요소(direct citation enforcement factor)'이며, a는 간접 인용의 영향력을 반복 계산하는 과정의 수렴 속도를 조절하는 요소이다. Sidiropoulos와 Manolopoulos (2005)는 이 공식에서 d는 페이지랭크에서와 같은 0.85로 설정하였고 다양한 실험 결과 b는 1 또는 0으로 설정하고 a는 e로 설정하는 것이 적절하다고 보고하였다. 특히 b를 1로 설정한 경우를 이들은 SCEAS_B1 또는 SCEAS1이라고 이름지었다. 파라미터가 b=1, d=0.85, a=e로 결정된 SCEAS_B1 공식을 다시 제시하면 아래와 같고, 이를 〈그림 1〉의 인용 네트워크에 적용하여 각 논문의 영향력을 측정한 결과는 〈그림 3〉과 같다.

$$W(d_i) = 0.15 + 0.85 \times \sum_j \frac{W(d_j)+1}{CO(d_j)} \times e^{-1}$$

〈그림 3〉의 SCEAS_B1 적용 결과를 앞 절의 페이지랭크 적용 결과와 비교해보면 각 논문의 순위는 같다. 다만 직접 인용빈도가 4회인 1번 논문과 2회인 2번 논문 사이의 영향력 격차가 더 벌어졌고, 직접 인용빈도가 2회로 동일하며 간접 인용만 차이가 나는 2번 논문과 3번 논문 사이의 영향력 격차는 크게 감소한 것으로 나타난다. 이런 결과는 SCEAS_B1이 페이지랭크에 비해서 직접 인용을 비교적 더 중요하게 고려한다는 특성을 보여준다.



〈그림 3〉 그림 1의 인용 네트워크에서 논문별 SCEAS_B1을 측정한 결과

2.3 CCI

CCI(Comprehensive Citation Index)는 페이지랭크에서 각 논문이 직접 인용되는 빈도가 반영되지 않는다는 점을 비판하면서 이를 개선하는데 주력하여 제안된 인용 지수이다(Bi, Wang, and Lin 2011). CCI의 계산식을 〈표 1〉의 항을 사용하여 나타내면 다음과 같다.

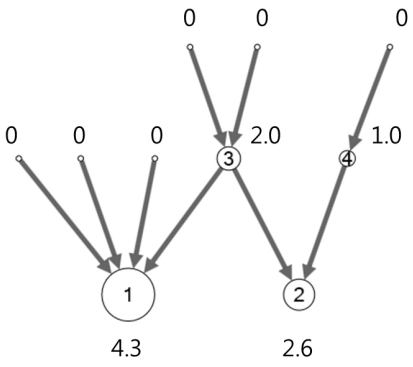
$$W(d_i) = CI(d_i) + \beta \sum_j \frac{W(d_j)}{CO(d_j)}$$
$$= \sum_j \left\{ 1 + \beta \frac{W(d_j)}{CO(d_j)} \right\}$$

CCI 공식은 페이지랭크와 유사하게 인용하는 논문의 영향력을 참고문헌 수로 나누어주는 항을 가지고 있으며, 이에다가 해당 논문의 인용빈도를 직접 더해주는 방식을 취하고 있다. SCEAS_B1도 동일한 문제점의 개선에 주력하였지만 CCI가 직접 인용빈도를 더 적극적으로 수용하는 방식을 취하고 있다. 공식에서 β가 0이면 CCI는 직접 인용빈도를 그대로 영향력으로 삼게 된다. CCI를 제안한 Bi 등(2011)은 β값을 다양하게 설정하여 산출 결과를 비교해본 결과 0.3에서 0.9 사이에서는 별다른 차이가 없다는 결과를 얻었다. 따라서 이들은 β값을 0.3으로 설정하여 측정한 값을 다른 인용 지수와 비교하였으므로 이 논문에서도 β는 0.3으로 설정하였다.

〈그림 1〉의 인용 네트워크에 대해서 각 논문의 CCI를 산출해본 결과는 〈그림 4〉와 같다. 직접 인용빈도를 그대로 반영하므로 인용빈도가 4회인 1번 논문의 영향력이 가장 높게 측정된다. 직접 인용빈도가 2회로 동일한 2번과 3번 논문 중에서는 간접 영향력을 추가로 획득하는 2번 논문의 영향력이 더 크게 측정된다. 간접 인용이 없는 3번과 4번 논문은 직접 인용빈도가 그대로 영향력이 된다.

2.4 f-값

f-값(f-value)은 해당 논문을 인용하는 논문들의 가중치를 모두 합한 값에 일정한 감쇄상수 RF(reducing factor)를 곱해서 인용받는 논문의 가중치로 삼는 방법을 사용한다(Fragkiadaki et al. 2011). 감쇄상수 값으로 f-값을 제안한 논문에서는 1/2.2을 적용하였다. 감쇄상수의 분



〈그림 4〉 그림 1의 인용 네트워크에서 논문별 CCI를 측정한 결과

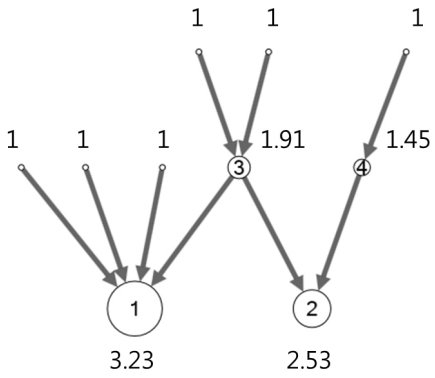
모를 2.2로 설정한 이유는 이들이 분석한 데이터에서 1세대 인용빈도와 그 인용논문들에 대한 인용인 2세대 인용빈도의 비율이 약 2.2였기 때문이다. 그리고 영향력 계산에서 인용하는 논문의 인용빈도가 0이면 전달받을 가중치가 연쇄적으로 없어지므로 가중치에 1을 항상 더하도록 하였다. f-값 공식을 〈표 1〉에 제시된 항으로 나타내면 다음과 같다.

$$W(d_i) = 1 + RF\sum_j W(d_j)$$

이 공식을 보면 페이지랭크나 SCEAS_B1처럼 인용하는 논문의 영향력을 받아서 합하는 형태를 취하고 있다. 다만 인용을 내보내는 논문의 영향력이 각 인용논문으로 나누어 분산되지 않고 모두에게 그대로 전달된다는 것이 큰 차이점이다. 직접 인용을 통해서 감쇄상수를 곱한 만큼 영향력이 전달되므로 t단계 떨어진 간접 인용을 통해서는 감쇄상수의 t제곱을 곱한 만큼의 영향력이 전달된다.

f-값 계산에서는 인용한 논문의 가중치가 감쇄상수를 곱한 만큼 인용된 논문으로 전해지므로 가장 긴 인용 사슬(citation chain 또는 citation thread)의 길이 만큼 반복해서 계산이 수행된다. 만약 문헌 간 상호 인용이나 순환 인용이 있어서 인용 링크를 통해서 한 논문의 영향력이 자신에게 되돌아는 경우가 발생하면 계산이 종료되지 않으며 순환 고리 상에 위치한 논문은 인용 영향력이 비정상적으로 크게 측정되는 문제가 있다. f-값을 제안한 Fragkiadaki(2011) 등은 이런 문제를 해결하기 위해서 인용한 논문과 인용된 논문의 출판년도가 같은 인용은 모두 제외하고 지수값을 측정하였다. 그러나 이런 방식은 제외되는 인용빈도가 각 논문의 영향력 순위에 영향을 끼칠 가능성이 있으므로 특히 규모가 작아서 인용빈도가 크지 않은 데이터베이스에 적용하기에는 적절하지 않다. 또한 최근에 발표된 논문의 영향력을 제대로 측정하기 어렵다는 문제도 발생한다.

〈그림 1〉의 인용 네트워크에 대해서 논문별 f-값을 측정한 결과는 〈그림 5〉와 같다. 인용을 받지 않은 논문들은 기본 값인 1을 가지게 되고, 이들로부터 1회 인용된 4번 논문의 영향력은 1 + 1/2.2 = 1.45, 2회 인용된 3번 논문의 영향력



〈그림 5〉 그림 1의 인용 네트워크에서 논문별 f-값을 측정한 결과

은 1 +(1/2.2) × 2 = 1.91로 산출된다. 같은 방식으로 1번 논문의 영향력은 1 +(1/2.2) × (1 + 1 + 1 + 1.91) = 3.23이 되고 2번 논문은 그보다 적은 2.53의 영향력을 가진 것으로 측정된다. 각 논문의 순위는 앞의 세 인용 지수 측정 결과와 같다.

2.5 단일 문헌 h-지수와 변형 지수

Schubert(2009)는 인용을 통한 영향력이 직접 인용 이외에도 간접 인용을 통해서도 전달되므로 이에 대한 측정이 필요하다고 주장하면서 단일 문헌 h-지수(single paper h-index)를 제안하였다. 단일 문헌 h-지수는 연구자의 인용 영향력을 측정하기 위해서 개발된 h-지수(Hirsch 2005)를 단일 문헌의 영향력 측정에 응용한 것이다. Schubert(2009)가 단일 문헌 h-지수를 정의한 문장은 다음과 같다.

> "한 문헌의 h-지수 h는 그 문헌을 인용하는 논문 집합의 인용빈도 h-지수로 정의할 수 있다. 이는 해당 문헌을 인용하는 논문 중에서 최대 h개가 h회 이상의 인용빈도를 가지고 있다는 뜻이다."

이 정의를 〈표 1〉의 항을 사용하여 공식으로 표현하면 다음과 같이 논문 i를 인용하는 논문 중에서 순위보다 인용빈도가 더 크거나 같은 문헌의 수를 구하는 형태가 된다.

$$W(d_i)=\sum_j f(d_j)\ \ , f(d_j)=\begin{cases}1 & if\quad CI(d_j)\geq CR(d_j)\\ 0 & else\end{cases}$$

이와 같은 단일 문헌 h-지수는 계산 방식이 원래의 h-지수와 마찬가지로 간단하므로 Web of Science나 Scopus와 같은 인용 데이터베이스에서 특정 논문을 찾았을 때 추가 검색 없이 그 논문을 인용한 논문의 인용빈도 목록만으로 산출할 수 있다는 장점이 있다. Egghe(2009)는 단일 문헌 h-지수를 "h-지수의 주목할 만한 새로운 응용"이라고 호평하였으며 Google Scholar를 이용해서 특정 논문을 검색하면 단일 문헌 h-지수를 산출해주는 서비스도 등장했다(Thor and Bornmann 2011a; 2011b).

단일 문헌 h-지수는 원래의 h-지수와 마찬가지로 정수값으로만 측정되며 h위 이내 논문의 인용빈도가 h보다 큰 정도를 전혀 반영하지 않으므로 정밀도와 변별력이 떨어진다는 단점이 있다. h-지수에 대해서도 이를 보완할 수 있도록 여러 변형 지수가 개발된 바 있으며, 이를 이용한 학술지의 영향력 측정이 시도된 바 있다(김판준, 이재윤 2010). 이 논문에서는 h_S-지수(이재윤 2006)를 적용하여 단일 문헌 h_S-지수를 측정해보았다. 단일 문헌 h_S-지수는 일단 단일 문헌 h-지수를 측정한 후 h위 이내에 속한 문헌의 인용빈도마다 제곱근을 취하여 합산하면 된다. 이를 공식으로 표현하면 다음과 같다.

$$W(d_i)=\sum_j f(d_j)$$
$$, f(d_j)=\begin{cases}\sqrt{CI(d_j)} & if\quad CI(d_j)\geq CR(d_j)\\ 0 & else\end{cases}$$

또한 단일 문헌 h-지수는 한 논문을 인용한 문헌들이 모두 인용빈도가 0일 경우에는 아무리 많이 인용된 논문이라도 값이 0에 머무는 단점이 있다. 이를 해소하기 위해서 지수 계산에서 각 논문의 인용빈도에 항상 1을 더하는 방식

을 고려해볼 수 있다. 이 방식은 f-값 공식에서 각 논문의 인용빈도에 1을 더하는 경우나, 페이지랭크 공식에서 받은 영향력에 (1-d)를 더하는 것과 유사한 경우이다. 인용 네트워크 분석의 효시라고 할 수 있는 Price(1976)도 인용 네트워크의 성장 모델을 제안하면서 발표시점의 인용빈도를 0이 아닌 1로 설정하는 모델을 발표한 바 있다.

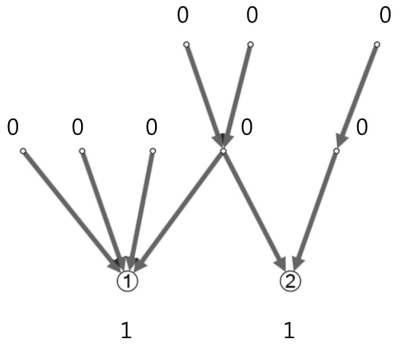
이처럼 인용빈도에 1을 더해서 h-지수 및 h_S-지수와 유사하게 적용하는 경우를 각각 h1-지수와 h_S1-지수라고 부르기로 한다. h1-지수를 구하는 공식은 다음과 같다.

$$W(d_i)=\sum_j f(d_j)$$
$$, f(d_j)=\begin{cases}1 & if \quad CI(d_j)+1 \geq CR(d_j)\\ 0 & else\end{cases}$$

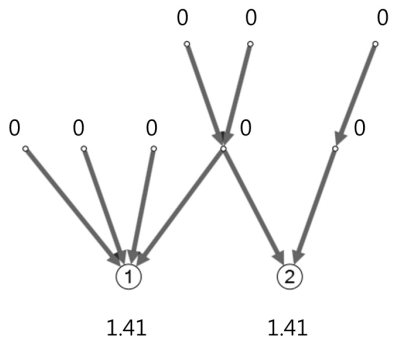
h_S1-지수를 구하는 공식은 다음과 같다.

$$W(d_i)=\sum_j f(d_j)$$
$$, f(d_j)=\begin{cases}\sqrt{CI(d_j)} & if \quad CI(d_j)+1 \geq CR(d_j)\\ 0 & else\end{cases}$$

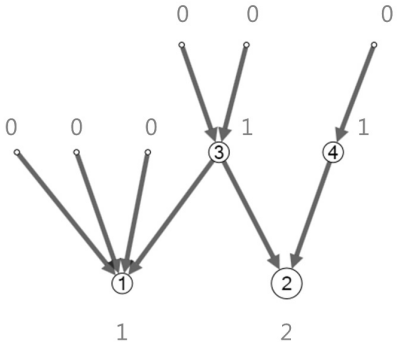
〈그림 1〉의 인용 네트워크 사례에 대해서 h-지수 계열 4가지를 각각 측정해본 결과는 〈그림 6〉 ~ 〈그림 9〉와 같다.



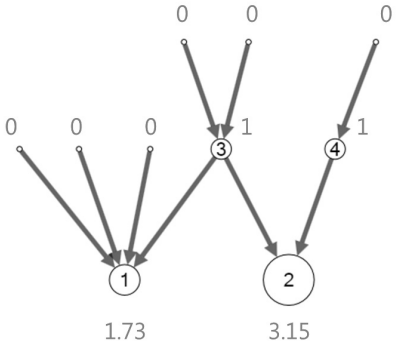
〈그림 6〉 그림 1의 인용 네트워크에서 논문별 h-지수를 측정한 결과



〈그림 7〉 그림 1의 인용 네트워크에서 논문별 h_S-지수를 측정한 결과



〈그림 8〉 그림 1의 인용 네트워크에서 논문별 h1-지수를 측정한 결과



〈그림 9〉 그림 1의 인용 네트워크에서 논문별 h_S1-지수를 측정한 결과

〈표 2〉 사례 인용 네트워크에 대한 인용 지수 산출 결과

번호	CFRQ (1gen)	CFRQ (2gen)	Page Rank	SCEAS _B1	CCI	f-value	h-index	h_S-index	h1-index	h_S1-index
1	4	2	0.25	1.71	4.30	3.23	1.00	1.41	1.00	1.73
2	2	3	0.20	1.06	2.60	2.53	1.00	1.41	2.00	3.15
3	2	0	0.14	1.01	2.00	1.91	0.00	0.00	1.00	1.00
4	1	0	0.10	0.62	1.00	1.45	0.00	0.00	1.00	1.00
5...10*	0	0	0.05	0.24	0.00	1.00	0.00	0.00	0.00	0.00

* 5번부터 10번 논문까지는 모두 같은 측정값임.

〈그림 1〉의 인용 네트워크 사례에 대해서 지금까지 측정한 8가지 인용 지수를 한꺼번에 비교해보면 〈표 2〉와 같다. 이 표에서 CFRQ(1gen)은 각 논문이 직접 받은 1세대 인용빈도이며, CFRQ(2gen)은 각 논문을 인용한 논문이 받은 2세대 인용빈도의 합계이다. 1세대 인용빈도가 가장 높은 1번 논문의 영향력을 가장 크게 평가한 지수는 페이지랭크, SCEAS_B1, CCI, f-값이다. 이와 달리 h1-지수와 h_S1-지수는 2세대 인용빈도가 가장 큰 2번 논문의 영향력을 1위로 측정하였다. h-지수와 h_S-지수는 1번 논문과 2번 논문의 영향력을 동등하게 평가하였다. 이 측정 결과로만 보면 h-지수 계열에서는 다른 지수에 비해서 1세대 인용빈도보다 2세대 인용빈도를 상당히 고려하고 있는 것으로 나타난다.

3. 지수 측정을 위한 인용 데이터 가공

앞 절에서는 가상의 소규모 인용 네트워크를 대상으로 8가지 인용 지수를 측정해보았다. 각 인용 지수의 특징을 실제 인용 데이터베이스를 통해서 살펴보기 위하여 이 논문에서는 한국과학기술정보연구원의 KSCI인용 데이터베이스를 대상으로 하는 분석 실험을 준비하였다. KSCI 데이터베이스는 상대적으로 인용이 활발한 과학기술 분야를 대상으로 하고 있으며, 수록 학술지 종수가 400여종 이상으로 대규모이고 해외 논문에 대한 인용 정보도 누락하지 않고 포함하고 있어서 이 연구의 분석 대상으로 적절하다고 판단하였다.

실험을 준비하는 시점에서 인용 데이터와 참고문헌 데이터를 고유 식별기호로 연결하여 분석할 수준으로 가공된 것은 2005년부터 2007년까지 3년간의 인용 데이터였으므로 이를 대상으로 실험을 진행하였다. 엑셀 파일 형태로 입수한 데이터에서 인용된 논문의 식별기호를 정리하여 데이터베이스로 구축하고 직접 처리 프로그램을 작성하여 인용 지수를 산출하였다. 부분적인 인용 네트워크의 시각화를 위해서는 Microsoft NodeXL(Hansen, Shneiderman, and Smith 2010)을 사용하였다.

3년간 데이터에서 1회 이상 인용이 파악된 KSCI 등재 학술지는 409종이었으며 전체적으로 파악된 논문은 82,106건, 파악된 인용은 80,353건이었다. 참고문헌이 있거나 1회 이상 인용된 논문 82,106건 중에는 KSCI에 등재된 학술지의 논문과 등재되지 않은 학술지의 논문이 섞여 있다. 등재되지 않은 논문은 모두가 KSCI에 등재된 논문의 참고문헌으로서 파악

된 것이므로 1회 이상의 인용정보를 가지며 해당 논문으로부터 다시 인용된 논문은 파악되지 않았다. 이들은 대부분 해외 학술지의 논문이 국내 학술지에 인용된 경우이다.

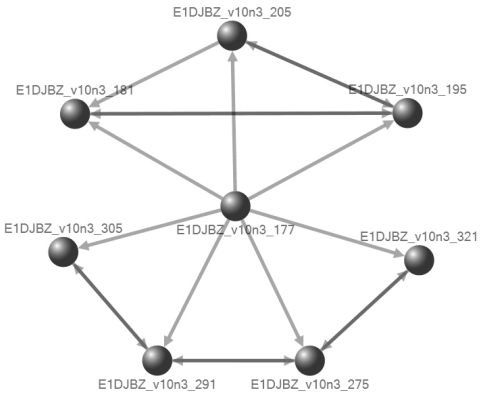
앞 장에서 살펴본 인용 지수를 모두 측정하기 위해서는 인용 데이터에 오류나 순환이 없도록 정련할 필요가 있다. 원칙적으로 논문 인용 네트워크에서는 먼저 출간된 논문에 대해서만 인용이 이루어지므로 자기 인용이나 순환 인용 고리가 없는 순수한 무환(acyclic) 네트워크이어야 한다(Newman 2010). 그러나 출간 전 사본의 형태로 인용하거나 데이터베이스 구축 과정의 오류 등으로 인하여 순환 인용이나 자기 인용 사례가 예외적으로 발생할 가능성이 있다.

KSCI 데이터에 대한 분석 결과 인용한 논문과 인용된 논문의 ID가 같아서 자기 인용으로 의심되는 경우가 30건 발견되었다. 논문의 자기 인용은 실제로 있을 수 없기 때문에 인용 데이터베이스 구축이나 데이터 변환 과정에서의 오류라고 판단되어 모두 삭제하였다.

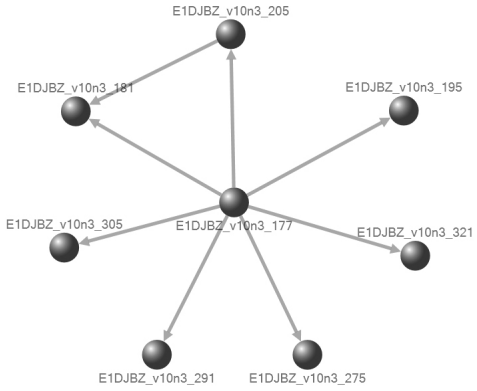
논문 간의 상호 인용은 논문의 자기 인용처럼 원칙적으로 불가능한 것은 아니지만, 인용 네트워크에서의 영향력 계산에서는 문제를 발생시킬 수 있다. Sidiropoulos와 Manolopoulos (2006)는 순환되는 인용 고리에 포함된 논문이 피드백을 통해서 영향력이 너무 높게 측정된다는 것을 페이지랭크 공식의 첫번째 문제점으로 지적하고 있으며, 그 인용 고리의 크기가 클수록 정도는 더 심해진다고 지적하였다. 논문 인용 네트워크에서는 상호 인용이 드물긴 하지만 발생하므로 페이지랭크 뿐만 아니라 반복적으로 계산하여 영향력을 산출하는 방식에서는 대부분 비정상적인 영향력 인플레이션이 발생하게 된다. 상호 인용하고 있는 두 논문이 서로 영향력을 반복하여 주고 받으면서 쌓아가기 때문이다. 앞에서 살펴본 방법 중에서는 h-지수 계열을 제외한 모든 공식에서 이런 현상이 생길 수 있다.

분석대상 데이터를 조사해본 결과 상호 인용에 해당되는 경우가 84건(42쌍)이 발견되어서 양쪽의 인용을 모두 삭제하였다. 상호 인용이 가장 복잡했던 경우의 예를 들면 〈그림 10〉과 같다. 이는 *Geosciences Journal*의 2006년 특집호에 실린 7편의 논문들 사이에 얽힌 복잡한 인용망의 예이다. 가운데에서 각 논문들을 인용하고 있는 논문이 서두 논문에 해당하며 이로부터 인용된 6편의 논문들 사이에 5쌍의 상호 링크가 발견되었다. 각 논문의 인용 지수를 산출할 때에는 모든 상호 링크를 제거하여 〈그림 11〉과 같은 상태로 변환하여 처리하였다.

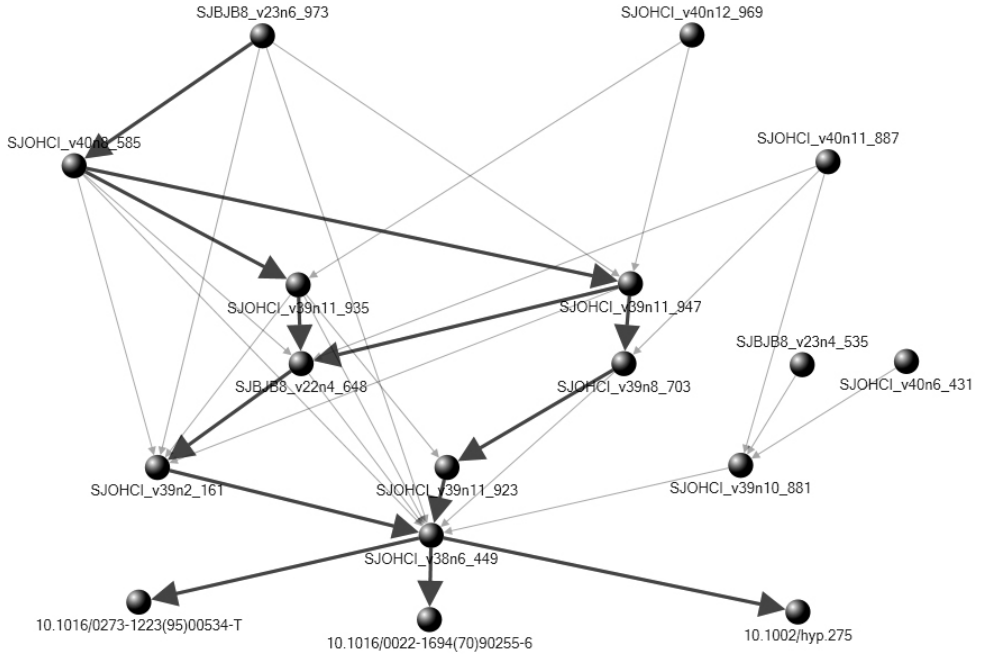
분석 대상 논문들 사이에서 가장 긴 인용 사슬은 〈그림 12〉와 같이 6단계 인용이 이어진 경우였다. 최초 인용이 2005년에 발생한 이후 2007년까지 3년 이내에 6단계까지 인용이 진행되었으므로 1년에 두 단계씩 인용이 이어진 셈이다. 〈그림 12〉에서는 6단계 인용이 이루어진 경로를 굵은 선으로 표시하였다. 논문 사이의 인용 관계가 얽혀 있으므로 마지막 논문에서 최초 논문까지 이어지는 2단계나 3단계의 더 짧은 인용 경로도 존재하였다. *Water Science and Technology*와 *Journal of Hydrology*, 그리고 *Hydrological Processes*에 게재된 논문을 『한국수자원학회논문집』 38권 6호에 게재된 논문(SJOHCI_v38n6_449)이 인용하면서 인용 사슬이 시작되며, 『한국수자원학회논문집』과 『수질보전』(SJBJB8)에 게재된 다른 논문들이 뒤를 잇고 있다.



〈그림 10〉 상호 인용이 포함된 인용망



〈그림 11〉 상호 인용을 제거한 인용망



〈그림 12〉 분석 대상 인용망에서 가장 긴 6단계 인용 사슬 사례

4. 인용 지수 측정 결과 분석

4.1 인용빈도 상위 논문의 인용 지수 측정 결과

인용빈도가 높은 논문은 영향력도 큰 것으로 간주되는 경우가 많기 때문에 학계의 주된 관심 대상이 된다. 논문 인용 지수 측정의 목적도 영향력이 낮은 논문들끼리 비교하기 위한 것이기 보다는 영향력이 큰 논문을 파악하고 영향력의 대소를 비교하기 위한 경우가 많다. 예를

들어 Sidiropoulos와 Manolopoulos(2005)는 우수 논문에 대한 시상을 위한 목적으로 인용 지수 분석을 응용할 수 있는 가능성을 제시하였으며, 정준민(2010)은 인용될 가능성이 높거나 많이 인용되는 논문을 모아 별도의 데이터베이스로 유통시키는 방안을 제시한 바 있다. 따라서 논문 인용 지수에 대한 분석의 첫 단계에서는 인용빈도 최상위 논문에 대한 지수 측정 결과의 비교를 통해서 각 인용 지수의 특성을 살펴보기로 한다.

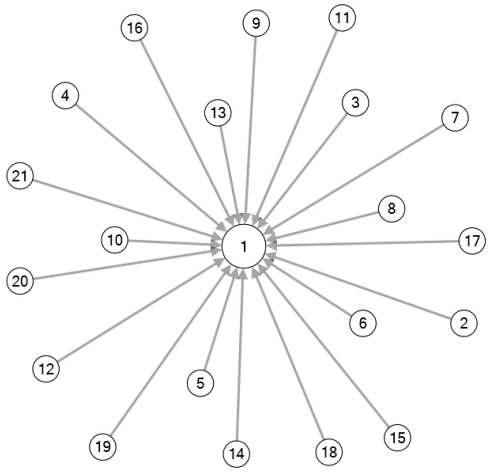
3년간 KSCI 데이터베이스에서 인용하거나 인용된 논문 중에서 인용빈도가 가장 높은 논문은 39회 인용된 논문이며, 상위 10위 이내에 포함된 논문은 인용빈도 16회로 공동 9위인 논문 3편을 포함하여 11편이다. 상위 11편의 논문에 대해서 1세대 인용빈도(CFRQ(1gen)) 및 2세대 인용빈도(CFRQ(2gen))와 함께 8가지 인용 지수를 측정하고 각 논문의 전체 순위를 파악한 결과는 〈표 3〉과 같다.

〈표 3〉에서 11개 최상위 논문의 지수별 평균 순위를 보면 CCI를 기준으로 측정한 평균 순위가 1세대 인용빈도 기준의 평균 순위인 5.6위와 비슷한 6.9위, f-값을 기준으로 측정한 평균 순위가 16.5위로 나타나서 1세대 인용빈도가 높은 논문의 영향력을 크게 평가하는 지수임을 알 수 있다.

〈표 3〉에서 38077번 논문은 1세대 인용빈도가 20회로 전체 4위이지만 2세대 인용빈도는 전혀 없는 것으로 나타났다. 〈그림 13〉과 같이 중심의 한 논문을 많은 논문들이 인용하는 형태이지만 2세대 인용은 전혀 없는 상황이다. 이에 대한 각 지수별 순위를 보면 CCI로는 7위, f-값으로는 24위에 해당하므로 1세대 인용빈도에 걸맞는 높은 순위로 평가되었지만, 나머지 지수로는 모두 1,000위 이하로 평가되었다. 특히 h-지수 계열로는 h1-지수를 제외한 나머지 3개 지수에서 모두 10,000위 이하로 극히 낮은 순위로 판정되었다. 실제로 38077번 논문의 서

〈표 3〉 인용빈도 상위 10위 이내 논문의 인용 지수별 순위

문헌 번호	인용빈도		순위									
	CFRQ (1gen)	CFRQ (2gen)	CFRQ (1gen)	CFRQ (2gen)	Page Rank	SCEAS _B1	CCI	f-value	h-index	h_S-index	h1-index	h_S1-index
3624	39	26	1	5	1	1	1	1	35	125	16	70
3552	24	30	2	3	5	4	2	2	3	15	1	9
30828	21	6	3	325	573	385	4	11	433	4127	101	1117
38077	20	0	4	12251	1672	1183	7	24	12319	12319	1644	12319
38277	19	0	5	12251	3128	2285	10	28	12319	12319	1644	12319
38902	18	24	6	9	2	2	3	4	3	18	1	10
2485	18	4	6	747	3325	2688	11	25	433	1762	101	710
63309	17	28	8	4	31	13	5	5	1	3	1	5
80233	16	20	9	15	8	7	6	8	35	42	16	25
36817	16	2	9	2220	362	204	14	41	433	4127	101	1117
26110	16	4	9	747	1899	2116	13	33	433	1762	101	710
평균 순위			5.6	2597.9	1000.5	808.0	6.9	16.5	2404.3	3329.0	338.8	2582.8



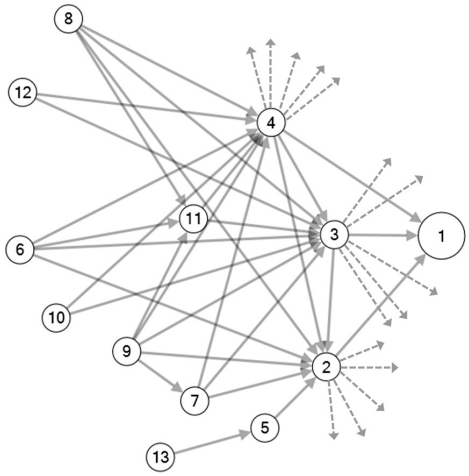
〈그림 13〉 가운데 논문에 대한 1세대 인용만 다수 존재하는 경우

지사항을 확인해보면 다음과 같다.

> Guttman, Antonin. 1984. "R-trees: A dynamic index structure for spatial searching." *Proceedings of the 1984 ACM SIGMOD*, 47-57.

이 논문은 데이터베이스나 파일 저장 구조에서 널리 활용되고 있는 자료구조인 R-trees를 제안한 논문으로서 이 자료구조를 사용하는 연구에서는 기본적으로 인용하는 논문이다. 따라서 38077번 논문에 대한 인용은 주제적인 영향을 받았다기보다는 도구적인 차원에서 다룬 경우가 대부분이다. 그러다 보니 이를 인용한 논문들 사이에도 인용 관계가 드문 편이다. 이와 같이 도구로 사용한 인용이 많은 논문에 대해서 CCI와 f-값은 1세대 인용빈도와 비슷하게 높이 평가한 반면에, 페이지랭크와 SCEAS_B1, 그리고 h1-지수는 보통 정도로 평가하고, h-지수, h_S-지수, h_S1-지수는 매우 영향력이 낮다고 평가하였다. 이처럼 동일한 논문에 대해서 CCI 및 f-값은 높은 영향력이라고 인정하는 반면에 h-지수 계열에서는 2세대 인용빈도가 0이라는 점에서 매우 낮게 평가하는 결과가 나타났다.

반면에 h-지수 계열에서 다른 지수에 비해 영향력을 월등하게 높게 평가하는 상황도 존재한다. 7346번 논문(〈그림 14〉에서는 최우측의 1번)은 1세대 인용빈도가 3회에 불과하므로 〈표 3〉에 제시되었던 상위 논문과는 상당한 격차가 있는 논문이다. 그런데 이를 인용한 3개 논문(〈그림 14〉의 2, 3, 4번 논문)이 모두 인용빈도가 높아서 2세대 인용빈도는 22회에 달하고 〈표 4〉와 같이 h-지수 계열로는 모두 20위 이내로 평가된다. 반면에 페이지랭크와 SCEAS_B1으로는 7346번 논문이 3,000위 이하로 평가된다. 그 이유는 이 논문을 인용하는 3개 논문이 모두 인용빈도가 높아서 영향력이 높긴 하지만 모두 참고문헌이 여러 편이어서 정작 7346번 논문으로 전달되는 영향력은 아주 일부분에 불과하기 때문이다. 〈그림 14〉에서는 2번, 3번, 4번 논문



〈그림 14〉 다단계 인용이 존재하는 경우

〈표 4〉 다단계 인용이 존재하는 사례 네트워크(그림 14)의 지수별 순위

번호	PageRank	SCEAS_B1	CCI	f-value	h-index	h_S-index	h1-index	h_S1-index
1	5303	12635	1427	18	3	4	16	15
2	1725	2163	171	31	35	35	101	80
3	3153	2516	122	71	35	59	101	117
4	3613	3412	239	302	433	993	101	566

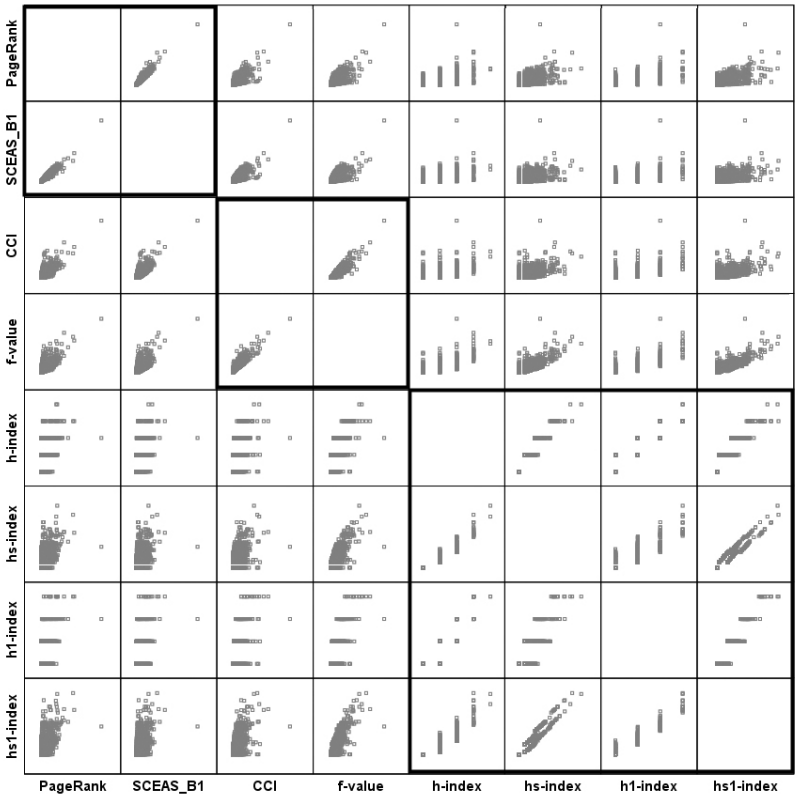
의 영향력이 점선으로 표현된 다른 논문들로의 인용으로 분산되고 있음을 보여주고 있다.

이상과 같이 각 인용 지수가 평가하는 논문의 영향력이 인용 네트워크의 연결 구조에 따라서 매우 달라진다는 점을 실제 인용 데이터를 통해서도 확인할 수 있었다. 인용빈도 상위 논문에 대해서 h-지수 계열의 인용 지수를 적용할 경우에는 여타 인용 지수를 적용하는 경우에 비해서 상당히 다른 결과를 얻게 되는 것으로 나타났다.

4.2 인용 지수 간 관계 분석

앞 절에서 인용빈도 상위 논문을 대상으로 각 인용 지수가 산출하는 순위를 비교해본 결과 인용 지수 사이의 순위에 상당한 차이가 있었다. 한편으로는 페이지랭크와 SCEAS_B1과 같이 유사한 순위를 산출하는 인용 지수군도 있었다. 여기서는 전체 논문을 대상으로 인용 지수 간 상관관계를 분석하여 각 지수 간 관계를 구체적으로 파악해보고자 한다.

인용 지수의 값 분포를 관계와 함께 살펴보기 위해서 8개 인용 지수를 교차하여 비교하는 산포도를 〈그림 15〉와 같이 그려보았다. 이때 인용 빈도가 0회인 논문의 지수는 모두 공통적으로 최저값에 해당하므로 비교의 의미가 없어서 제외하고 인용 빈도가 1회 이상인 논문 61,154건에 대해서만 지수값의 비교를 수행하였다.



〈그림 15〉 8개 인용 지수 사이의 관계를 보여주는 산포도

〈그림 15〉의 지수 간 교차 산포도를 보면 선형적인 관계를 보이는 인용 지수 쌍이 드러난다. 일단 왼쪽 상단에서 페이지랭크와 SCEAS_B1, 그리고 CCI와 f-값의 두 쌍이 뚜렷한 선형 관계를 보여주고 있다. 그리고 배치 순서상 여섯 번째인 h_S-지수와 여덟 번째인 h_S1-지수 사이의 선형 관계도 두드러지게 나타나고 있다. 이밖에 오른쪽 하단에 자리잡은 네 가지 h-지수 계열 지수 사이의 관계도 다소 선형에 가깝게 나타나고는 있으나, 정수값으로 산출되어 변별력이 떨어지는 h-지수와 h1-지수 때문에 계단식으로 관계가 나타난다. 한편 h-지수 계열의 4개 지수와 나머지 4개 지수의 관계는 오른쪽 상단과 왼쪽 하단에서 보듯이 뚜렷하지 않았다.

산포도에서 보듯이 인용 지수의 값 분포는 정규분포보다는 낮은 값 쪽으로 심하게 치우친 비정규분포에 가깝다. 따라서 인용 지수 간 상관관계 분석을 위해서는 피어슨 상관계수보다는 비모수통계인 스피어맨 상관계수를 산출하는 것이 바람직하다. 8개 지수 간 스피어맨 상관계수를 구하면 〈표 5〉와 같다.

산포도에서도 확인하였듯이 〈표 5〉에서는 뚜렷한 상관관계를 보이는 지수 쌍이 드러난다. 특히 0.99 이상의 매우 높은 상관관계를 보이는 지수들끼리는 산출하는 영향력 순위가 거의 동일하다는 것을 의미한다. 이와 같이 예외적으

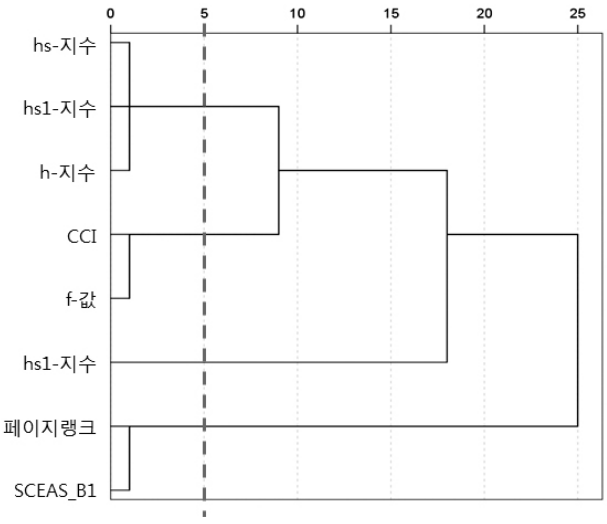
로 높은 상관관계가 나타난 이유는 측정 대상 자료에 동률 순위가 많았기 때문이다. 그럼에도 불구하고 인용 지수 사이의 관계는 상관이 큰 경우와 그렇지 못한 경우로 뚜렷하게 구분됨을 알 수 있다.

상관관계가 높은 지수끼리 뭉치는 현상을 군집분석을 통해서 확인해보면 〈그림 16〉과 같다. 이 덴드로그램은 인용 지수 간 스피어맨 상관계수를 연관성 행렬로 입력하여 집단내평균연결 기법으로 군집을 생성한 결과이다. 이에 따르면 밀접하게 관련되어 하나의 군집을 이루는 인용 지수 군은 페이지랭크와 SCEAS_B1, f-값과 CCI, h-지수와 h_S-지수 및 h_S1-지수이다. 이들이 각각 모여 3개의 군집을 이루고 h1-지수는 여타 지수와는 상당히 다른 특성을 보여서 단독 군집을 구성하였다.

〈표 5〉 8개 인용 지수 간 스피어맨 상관계수

	PageRank	SCEAS_B1	CCI	f-value	h-index	h_S-index	h1-index
SCEAS_B1	.995						
CCI	.457	.421					
f-value	.441	.399	.995				
h-index	.271	.209	.714	.734			
h_S-index	.276	.211	.714	.739	.995		
h1-index	.206	.196	.327	.335	.350	.363	
h_S1-index	.280	.216	.717	.742	.995	.998	.400

* 음영은 0.9 이상으로 매우 높은 상관값임.
* 상관행렬의 오른쪽 윗부분은 표시하지 않았음.



〈그림 16〉 8개 지수 간 관계의 군집 분석 덴드로그램

4.3 빈도 변수와 인용 지수 간 관계 분석

앞에서 살펴본 바와 같이 단일 문헌의 영향력을 평가하는 8개 인용 지수는 4가지 유형으로 나눌 수 있었다. 8개 인용 지수 간의 차이와 유사성이 나타나는 원인을 파악하기 위해서 인용 지수 공식을 구성하는 주요 빈도 변수를 추출해서 이들과 인용 지수 사이의 상관관계를 측정해보았다.

인용 지수 공식의 주요 구성 요소로서 각 지수의 측정에 주로 사용되는 빈도 변수는 논문 d_i에 대한 1세대 인용빈도(CFRQ(1gen)), 논문 d_i를 인용하는 논문인 d_j들의 인용빈도 합계인 2세대 인용빈도(CFRQ(2gen)), 그리고 논문 d_i가 받는 인용의 비중(Outfrq(2gen)_portion)이다. 논문 d_i가 받는 인용의 비중이란 인용하는 논문인 d_j들의 영향력 중에서 논문 d_i로 전달되는 영향력의 비율을 뜻한다. 이는 인용하는 논문의 참고문헌 수를 분모로 하는 값이다. 이들 외에 h-지수 계열에만 사용되는 인용하는 논문들의 빈도 순위는 2세대 인용빈도와 같은 의미이므로 제외하고 기타 상수항도 제외할 수 있다. 세 가지 주요 빈도 변수를 〈표 1〉의 항들로 나타내면 다음과 같다.

$$CFRQ(1gen) = CI(d_i)$$
$$CFRQ(2gen) = \sum_j CI(d_j)$$
$$Outfrq(2gen)\_portion = \sum_j \frac{1}{CO(d_j)}$$

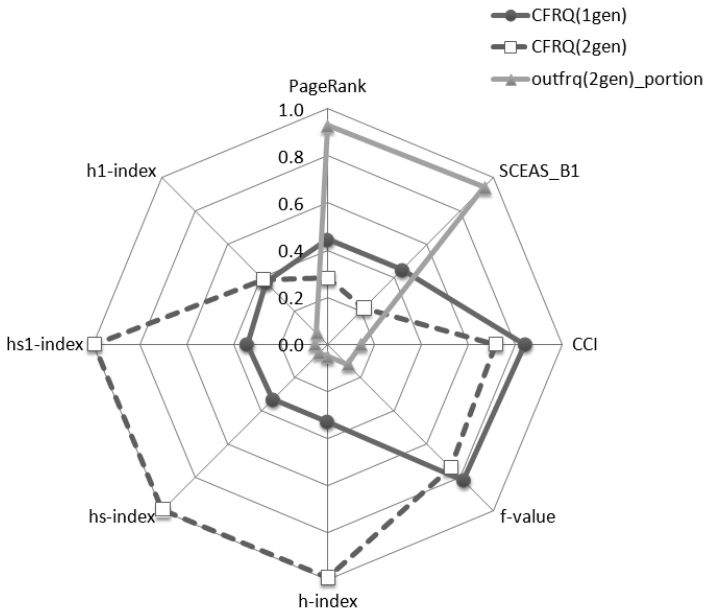
인용빈도가 1회 이상인 61,154건을 대상으로 각 논문의 세 가지 빈도변수와 8가지 인용 지수 사이의 스피어맨 상관계수를 산출한 결과를 〈표 6〉과 〈그림 17〉에 제시하였다. 결과를 살펴보면 1세대 인용빈도와 상관관계가 매우 높은 지수는 CCI와 f-값이다. 이 두 지수는 2세대 인용빈도와도 상관관계가 높고 받는 인용의 비중은 매우 낮았다. 이 두 지수가 유사한 영향력 측정 결과를 보이는 것은 다른 지수에 비해서 1세대 인용빈도를 반영하는 비중이 뚜렷하게 높고 2세대 인용빈도를 적절히 반영하며 받는 인용의 비중은 거의 반영하지 않기 때문이라고 판단된다.

2세대 인용빈도와의 상관관계만 높게 나타난 지수는 h-지수 계열 중에서 h-지수, h_S-지수, h_S1-지수의 세 가지이다. 이들이 유사한 영향력 측정 결과를 보이는 것은 1세대 인용빈도와의 상관관계에 비해서 2세대 인용빈도와의 상관관계가 뚜렷하게 높기 때문이라고 할 수 있다. 이 세 지수는 받는 인용의 비중 변수와의 상관관계는 거의 미미한 것으로 나타났다.

〈표 6〉 빈도변수와 인용 지수 간 스피어맨 상관계수

인용 지수 / 빈도변수	PageRank	SCEAS_B1	CCI	f-value	h-index	hs-index	h1-index	hs1-index
CFRQ(1gen)	.445	.448	.840	.819	.329	.331	.373	.342
CFRQ(2gen)	.280	.216	.716	.742	.991	.996	.387	.997
Outfrq(2gen)_portion	.929	.943	.142	.120	.052	.051	.068	.053

* 음영은 0.7 이상의 높은 상관값임.



〈그림 17〉 인용 지수와 빈도변수 간 스피어맨 상관계수

다른 지수들과 달리 받는 인용의 비중 변수와 관련이 매우 깊은 것은 페이지랭크와 SCEAS_B1이다. 이 두 지수는 1세대 인용빈도 변수나 2세대 인용빈도 변수와의 상관계수는 높지 않은 반면에 받는 인용의 비중 변수와의 상관계수가 0.9대로 매우 높았다. 이 두 지수가 앞 절에서 살펴본 바와 같이 다른 지수들과 매우 다른 측정결과를 보이는 이유는 받는 인용의 비중을 크게 반영하기 때문임을 알 수 있다. 특이한 점은 받는 인용의 비중을 반영하는 항이 공식에 포함되어 있는 또 하나의 지수인 CCI는 이 두 지수와 매우 다른 순위를 산출한다는 점이다. CCI 공식에서 1세대 인용빈도를 직접 지수값에 더함으로써 다른 변수의 반영 정도를 약화시킨 것이 이런 결과를 가져온 것으로 이해된다.

8가지 인용 지수 중에서 7가지 인용 지수가 세 빈도변수 중 하나 이상과 높은 상관관계를 보이는 반면에, h1-지수는 어느 지수와도 뚜렷한 상관관계를 보이지 않았다. h1-지수가 정수값으로 측정되므로 낮은 변별력을 가지고 있음에도 불구하고 인용하는 논문들의 인용빈도에 일괄적으로 1을 더해준 것이 영향을 끼친 결과라고 추정된다.

정리해보면 CCI와 f-값은 1세대 인용빈도와 2세대 인용빈도를 고르게 반영하고 있으며, 페이지랭크와 SCEAS_B1은 받는 인용의 비중과 관련이 높고, h1-지수를 제외한 h-지수 계열은 2세대 인용빈도의 비중과 관련이 높음이 확인되었다. 이와 같이 관련이 높은 빈도변수 측면에서 살펴본 각 지수의 특성은 앞 절에서 8가지 지수를 서로 간의 상관관계에 근거해서 군집분석으로 나누어본 결과와 정확히 일치한다. 따라서 8가지 지수의 특성은 이 세 가지 빈도변수로 설명된다고 할 수 있다.

5. 결 론

페이지랭크를 비롯한 단일 논문에 대한 평가 지수 8가지를 대상으로 국내 인용데이터베이스에 대한 적용 실험을 통해서 각 지수 간 관계와 특성을 파악하였다. 주요 확인 결과는 다음과 같다.

첫째, 기존에 제안된 단일 논문의 인용 영향력 측정 지수 5가지와 3가지 변형 지수를 일관된 공식으로 표현하여 비교할 수 있었다. 기존 연구에서는 각각 다른 표기 방식을 사용하였으므로 지수 간 비교가 어려웠다.

둘째, 가상의 데이터와 실제 인용 데이터에 대한 측정을 통해서 각 인용 지수가 평가하는 논문의 영향력이 인용 네트워크의 연결 구조에 따라서 매우 달라진다는 점이 확인되었다.

셋째, 인용 지수 간 상관관계 분석과 군집분석을 통해서 8가지 인용 지수를 4가지 유형(군집)으로 나눌 수 있었다. 이중에서 3가지 지수 유형은 페이지랭크와 SCEAS_B1, f-값과 CCI, h-지수와 h_S-지수 및 h_S1-지수가 각각 모인 것이며, h1-지수는 여타 지수와 구별되어 독립 유형을 구성하였다.

넷째, 빈도변수와 인용 지수 사이의 상관관계 분석을 통해서 각 인용 지수 유형에 대한 해석이 가능하였다. CCI와 f-값은 1세대 인용빈도와 2세대 인용빈도를 고르게 반영하고 있으며, 페이지랭크와 SCEAS_B1은 받는 인용의 비중과 관련이 높고, h1-지수를 제외한 h-지수 계열은 2세대 인용빈도의 비중과 관련이 높음이 확인되었다.

이상의 실험 과정 및 분석 결과에 근거하여 국내 학술 논문 데이터베이스에서 논문의 인용 영향력을 측정할 때 고려할 사항을 제시하면 다음과 같다.

첫째, 인용 데이터베이스에 대한 분석에서는 데이터의 오류 검증이나 순환 인용 파악 등의 정련 과정이 매우 중요하다. 데이터 정련이 일부에 대해서만 미흡할 경우에도 네트워크의 구조적 분석에 기반한 인용 지수는 크게 잘못된 결과를 산출할 수도 있다.

둘째, 단일문헌 h-지수 계열의 네 가지 지수 중에서는 변별력과 상관 정도를 고려할 때 이 논문에서 제안한 h_S1-지수를 사용하는 것이 특히 국내 인용 데이터베이스에 적절할 것으로 판단된다. 나머지 인용 지수 중에서 h1-지수는 타 지수와의 상관관계가 매우 낮으며 인용 빈도 요소와의 관련성이 뚜렷하지 않아서 측정 결과에 대한 신뢰성이 떨어진다. h-지수는 정수로만 측정되므로 변별력이 낮으며, h_S-지수는 인용빈도가 0인 논문의 영향력을 고려하지 못한다는 단점이 있다. 이와 비교해보면 h_S1-지수는 변별력이 높고 여타 h-지수 계열과 상관성이 크며 인용빈도가 0인 논문으로부터의 인용 영향력도 고려한다는 장점이 있다. 또한 페이지랭크를 비롯하여 순환 계산을 통해서 산출하는 다른 지수들과는 달리 2단계 인용까지만을 고려하므로 계산이 신속히 이루어지고 순환 인용이 큰 문제가 안 된다는 장점도 있다.

셋째, 이 연구에서 파악된 4가지 인용 지수 유형이 각각 빈도변수를 반영하는 정도가 상이하므로 어느 한 인용 지수에 의존하기 보다는 가급적 두 유형 이상의 인용 지수를 활용하는 것이 바람직하다. 예를 들면 1세대 및 2세대 인용빈도를 고르게 반영하는 f-값과 받는 인용의 비중을 적극적으로 반영하는 SCEAS_B1을 함께 적용

할 수 있다. 다만 특정한 빈도변수를 더 중요하게 고려해야 하는 상황에서는 인용 지수를 선택적으로 활용하는 것이 가능할 것이다.

향후 연구에서는 현재 실험적으로 3년 기간의 인용 데이터에 대한 분석으로 파악된 사항을, 더 장기간의 인용 데이터에 대해서 분석하여 확인할 필요가 있다. 아울러 인용 정보의 파악이 제한된 국내 데이터베이스 이외에 해외 학술지 인용 데이터를 대상으로 각 지수의 특성을 분석할 필요가 있다고 생각된다. 논문 자료 이외에 특허 자료에 대한 인용 영향력 관련 연구(유재복, 정영미 2010a; 2010b)도 기술개발 정책과 특허 관리의 수단으로 주목되고 있으므로, 이 연구에서 다룬 인용 네트워크 기반의 영향력 지수를 특허 인용 네트워크에 적용하는 연구도 필요할 것이다.

참 고 문 헌

[1] 김판준, 이재윤. 2007. 연구 영역 분석을 위한 디스크립터 프로파일링에 관한 연구.『정보관리학회지』, 24(4): 285-303.

[2] 김판준, 이재윤. 2010. 학술지 영향력 측정을 위한 h-지수의 응용에 관한 연구.『정보관리학회지』, 27(1): 269-287.

[3] 김희전, 조현양. 2010. 저자동시인용분석과 저자서지결합분석에 의한 지적 구조 분석.『정보관리학회지』, 27(3): 283-306.

[4] 유재복, 정영미. 2009a. 특허 인용에 영향을 미치는 요인 분석.『정보관리학회지』, 27(1): 103-118.

[5] 유재복, 정영미. 2009b. 특허인용 예측모형 구축에 관한 연구.『정보관리학회지』, 27(4): 239-258.

[6] 유종덕, 최은주. 2011. 저자프로파일링분석과 저자동시인용분석의 유용성 비교 검증.『정보관리학회지』, 28(1): 123-144.

[7] 윤희윤, 김신영. 2005. 학술지 영향계수와 연구업적 평가비중의 상관성 분석.『정보관리연구』, 36(3): 1-25.

[8] 이재윤. 2006 연구성과 측정을 위한 h-지수의 개량에 관한 연구.『정보관리학회지』, 23(3): 167-186.

[9] 이재윤. 2008. 서지적 저자결합분석: 연구동향 분석을 위한 새로운 접근.『정보관리학회지』, 25(1): 173-190.

[10] 정경희. 1999. 인용분석의 발전과 그에 대한 비판.『정보관리연구』, 30(2): 53-68.

[11] 정준민. 2010. 인용문헌 분석을 통한 학술 논문의 수명 및 계보에 관한 연구.『한국문헌정보학회지』, 44(2): 357-379.

[12] 정희경, 이춘실. 2009. 국내 의학 학술지의 일정 주기별 SCI 피인용 최상위 논문 선정 방법.『제16

회 한국정보관리학회 학술대회 논문집』, 127-132.

[13] Adam, D. 2002. "Citation analysis: The counting house." *Nature*, 415: 726-729.

[14] Bi, H. H., Wang, J., & Lin, D. K. J. 2011. "Comprehensive citation index for research networks." *IEEE Transactions on Knowledge and Data Engineering*, 23(x): to be published.

[15] Bollen, J., Van de Sompel, H. 2008. "Usage Impact Factor: The effects of sample characteristics on usage-based impact metrics." *Journal of the American Society for Information Science and Technology*, 59(1): 1-14.

[16] Chen, P., Xie, H., Maslov, S., & Redner, S. 2007. "Finding scientific gems with Google's PageRank algorithm." *Journal of Informetrics*, 1(1): 8-15.

[17] Dellavalle, R. P., Schilling, L. M., Rodriguez, M. A., Van de Sompel, H., & Bollen, J. 2007. "Refining dermatology journal impact factors using PageRank." *Journal of the American Academy of Dermatology*, 57(1): 116-119.

[18] Ding, Y., Yan, E., Frazho, A., & Caverlee, J. 2009. "PageRank for ranking authors in co-citation networks." *Journal of the American Society for Information Science and Technology*, 60(11): 2229-2243.

[19] Egghe, L. 2009. "On the relation between Schubert's h-index of a single paper and its total number of received citations." *Scientometrics*, 84(1): 115-117.

[20] Fragkiadaki, E., Evangelidis, G., Samaras, N., Dervos, D. A. 2011. "f-Value: Measuring an article's scientific impact." *Scientometrics*, 86(3): 671-686.

[21] Franceschini, F., Maisano, D. 2011. "Structured evaluation of the scientific output of academic research groups by recent h-based indicators." *Journal of Informetrics*, 5(1): 64-74.

[22] Glänzel, W., Moed, H. F. 2002. "Journal impact measures in bibliometric research." *Scientometrics*, 53(2): 171-193.

[23] Hansen, D., Shneiderman, B., & Smith, M. A. 2010. *Analyzing Social Media Networks with NodeXL: Insights from a Connected World*. Morgan Kaufmann.

[24] Langville, A. N., & Meyer, C. D. 2006. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.

[25] Moed, H. F. 2005. *Citation Analysis in Research Evaluation*. Springer.

[26] Newman, M. E. J. 2010. *Networks: An Introduction*. Oxford University Press.

[27] Neylon, C., & Wu, S. 2009. "Article-level metrics and the evolution of scientific impact." *PLoS Biology*, 7(11): e1000242. [online]. [cited 2011.4.1]. 〈http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1000242〉.

[28] Page, L., Brin, S., Motwani, R., & Winograd, T. 1999. "The PageRank citation ranking:

Bringing order to the Web." Technical Report. Stanford InfoLab. [online]. [cited 2011.4.1]. 〈http://ilpubs.stanford.edu:8090/422/〉.

[29] Price, D. J. de Solla. 1976. "A general theory of bibliometric and other cumulative advantage processes." *Journal of the American Society for Information Science*, 27(5): 292-306.

[30] Priem, J., & Hemminger, B. H. 2010. "Scientometrics 2.0: New metrics of scholarly impact on the social Web." *First Monday*, 15(7). [online]. [cited 2011.4.1]. 〈http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570〉.

[31] Seglen, P. O. 1993. "How representative is the journal impact factor?" *Research Evaluation*, 2: 143-149.

[32] Seglen, P. O. 1997. "Why the impact factor of journals should not be used for evaluating research." *British Medical Journal*, 314: 498-502.

[33] Sidiropoulos, A., & Manolopoulos, Y. 2005. "A citation-based system to assist prize awarding." *SIGMOD Records*, 34(4): 54-60.

[34] Sidiropoulos, A., & Manolopoulos, Y. 2006. "Generalized comparison of graph-based ranking algorithms for publications and authors." *Journal of Systems and Software*, 79(12): 1679-1700.

[35] Thor, A., Bornmann, L. 2011a. "The calculation of the single publication h index and related performance measures: A web application based on Google Scholar data." *Online Information Review*, 35(2): 291-300.

[36] Thor, A., & Bornmann, L. 2011b. *Web Application to Calculate the Single Publication h Index(and Further Metrics) based on Google Scholar.* [online]. [cited 2011.4.1]. 〈http://labs.dbs.uni-leipzig.de/gsh〉.

[37] Walker, D., Xie, H., Yan, K.-K., & Maslov, S. 2007. "Ranking scientific publications using a model of network traffic." *Journal of Statistical Mechanics*, P06010. [online]. [cited 2011.4.1]. 〈http://stacks.iop.org/JSTAT/2007/P06010〉.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Kim, Panjun, & Lee, Jae Yun. 2007. "Descriptor profiling for research domain analysis." *Journal of the Korean Society for Information Management*, 24(4): 285-303.

[2] Kim, Panjun, & Lee, Jae Yun. 2010. "A study on journal impact measurement with Hirsch-type indices." *Journal of the Korean Society for Information Management*, 27(1): 269-287.

[3] Kim, Hee-Jeon, & Cho, Hyun-Yang. 2010. "A study on intellectual structure using author co-citation analysis and author bibliographic coupling analysis in the field of social welfare science." *Journal of the Korean Society for Information Management*, 27(3): 283-306.

[4] Yoo, Jae-Bok, & Chung, Young-Mee. 2009a. "Analysis of factors influencing patent citations." *Journal of the Korean Society for Information Management*, 27(1): 103-118.

[5] Yoo, Jae-Bok, & Chung, Young-Mee. 2009b. "A study on developing a prediction model of patent citation counts." *Journal of the Korean Society for Information Management*, 27(4): 239-258.

[6] Ryoo, Jong-duk, & Choi, Eun-Ju. 2011. "A comparison test on the potential utility between Author Profiling Analysis(APA) and Author Co-Citation Analysis(ACA)." *Journal of the Korean Society for Information Management*, 28(1): 123-144.

[7] Yoon, Hee-Yoon, & Kim, Sin-Young. 2005. "An analysis on correlations between Journal Impact Factor and research performance evaluation weight." *Journal of Information Management*, 36(3): 1-25.

[8] Lee, Jae Yun. 2006. "Some improvements on h-index: Measuring research outputs by citations." *Journal of the Korean Society for Information Management*, 23(3): 167-186.

[9] Lee, Jae Yun. 2008. "Bibliographic Author Coupling Analysis: A new methodological approach for identifying research trends." *Journal of the Korean Society for Information Management*, 25(1): 173-190.

[10] Joung, Kyoung-Hee. 1999. "A review of the development and critique of citation analysis." *Journal of Information Management*, 30(2): 53-68.

[11] Chung, Jun-Min. 2010. "The study on the genealogy and impact factor of papers by citation analysis." *Journal of the Korean Society for Library and Information Science*, 44(2): 357-379.

[12] Chung, Hee-Kyung, & Lee, Choon-Shil. 2009. "A methodology to generate the ranking of the most cited articles based on SCI citations." *Proceedings of the 16th Annual Conference of Korean Society for Information Management*, 127-132.