논문 인용 영향력 측정 지수의 편향성에 대한 연구: KCI 수록 논문을 대상으로*

Discipline Bias of Document Citation Impact Indicators: Analyzing Articles in Korean Citation Index

이재윤 (Jae Yun Lee)**
최상희 (Sanghee Choi)***

초 록

학술지의 인용빈도를 특정하여 산출된 지수로 단일 논문의 영향력을 평가하는 것에 대한 비판으로 인해 단일 논문의 인용 영향력을 측정하는 인용지수에 대한 연구가 다양하게 시도되었다. 이 연구에서는 8개의 단일 논문 인용영향력 평가 지수를 살펴보고 KCI 논문 데이터베이스를 대상으로 각 인용지수의 분야별 편향성을 조사하여 보았다. 대상 지수는 단순 인용빈도, 페이지랭크, f-값, CCI, c-지수, 단일문헌 h-지수, 단일문헌 hs-지수, cl-지수였다. 분석결과 페이지랭크가 학문 분야별 균등성, 학문 분야 내에서 학술지별 균등성 영역에서 가장 편향성이 없는 것으로 나타났다. 반면에 단순 인용빈도는 특정 학문분야나 특정 학술지에 편향된 결과를 산출할 가능성이 높은 것으로 나타났다. KCI 데이터베이스에서는 논문의 단순 인용빈도만 제공하고 있는데, 분야별 균등성을 가장 잘 유지하는 지수인 논문 페이지랭크를 함께 제공할 필요가 있다. 아울러 인용한 문헌의 인용빈도만으로 산출이 가능해서 이용자의 검색 결과로부터 바로 산출할 수 있는 지역 네크워크 지수 중에서는 cl-지수가 가장 균등성을 잘 유지하므로 계산 과정과 서비스가 손쉬운 지수로 함께 제공하는 것도 검토해야 한다.

ABSTRACT

The impact of a journal is commonly used as the impact of an individual paper within that journal. It is problematic to interpret a journal's impact as a single paper's impact of the journal, so there are several researches to measure a single paper's impact with its own citation counts. This study applied 8 impact indicators to Korean Citation Index database and examined discipline bias of each indicator. Analyzed indicators are simple citation counts, PageRank, f-value, CCI, c-index, single publication h-index, single publication hs-index, and cl-index. PageRank has the least discipline bias at highly ranked papers and journal bias in a discipline. On the contrary, simple citation counts showed strongly biased results toward a certain discipline or a journal. KCI database provides only simple citation counts. It needs to show PageRank (global indicator) to discover influential papers in diverse areas. Furthermore it needs to consider to provide the best of local indicators. Local indicators can be calculated only with papers in users' search results because they uses citation counts of citing papers and the number of references. They are more efficient than global indicators which explore the whole database. KCI should also consider to provide Cl-index (local indicator).

키워드: 연구 평가, 인용 지수, 인용 네트워크, 분야 편향성, 페이지랭크
research evaluation, citation indicator, citation network, discipline bias, PageRank

---

* 본 논문에는 한국연구재단에서 구축하여 제공하는 한국학술지인용색인(KCI) DB 정보를 이용하였음.
** 명지대학교 문헌정보학과 부교수(memexlee@mju.ac.kr) (제1저자)
*** 대구가톨릭대학교 도서관학과 부교수(shchoi@cu.ac.kr) (공동저자)
■ 논문접수일자 : 2015년 11월 29일 ■ 최초심사일자 : 2015년 12월 7일 ■ 게재확정일자 : 2015년 12월 8일
■ 정보관리학회지, 32(4), 205-221, 2015. 〔http://dx.doi.org/10.3743/KOSIM.2015.32.4.205〕

1. 서 론

연구비 지원이나 연구자 평가를 위한 국가 사업이나 연구기관 정책에서 논문의 내용이 아닌 논문이 게재된 학술지에 대한 위상을 기준으로 평가하는 것은 흔한 일이 되었다. 특히 톰슨 로이터의 영향력지수(Impact Factor)를 연구성과 평가의 핵심 지표로 활용하는 경향이 우리나라뿐만 아니라 해외에서도 증가하는 추세이다. 심지어는 Science Citation Index나 한국학술지인용색인 KCI와 같은 인용색인 데이터베이스에 학술지가 색인되고 있는지 여부가 그 학술지에 게재된 논문의 위상을 좌우하는 핵심 요인이 되었다. 국내의 경우 한국연구재단에서 2015년 7월 발표한 학술지평가에서 상당수의 학술지가 등재지에서 탈락함에 따라 여러 학술단체에 큰 충격을 주었고 한국연구재단은 쏟아지는 민원 재기에 시달렸다. 2015년 7월에는 신청자격 및 체계평가 부문에 대해서만 재심 신청을 받았으나, 결국 2015년 9월에 정성평가에 대한 재심사를 받는다는 공지가 발표되었다. 연구 평가가 논문이 아닌 학술지를 기준으로 이루어지는 상황에서는 이와 유사한 혼란이 언제든지 다시 발생할 수 있다.

2012년 연말에 미국 샌프란시스코에서 개최된 Annual Meeting of The American Society for Cell Biology에 모인 학술지 편집자들과 출판사들은, 이와 같이 왜곡된 연구성과물 평가의 현실이 개선되어야 한다는 인식을 공유하고, 올바른 연구 평가 방법에 대해 권고하는 DORA 선언(San Francisco Declaration on Research Assessment)을 발표하였다. DORA 선언의 핵심 내용에는 연구비 지원기관과 연구기관이 연구 평가를 할 때 학술지가 아닌 논문 자체의 내용을 기준으로 해야 한다는 것과, 출판사가 다양한 논문 단위의 계량 지표를 제공해야 한다는 항목이 포함되어 있다.

최근 들어 상당수의 국제 학술지 출판사가 논문 단위의 계량 지표로 단순 인용빈도나 altmetrics 지표를 제공하기 시작하였다. 그러나 단순 인용빈도는 학문 분야별로 전반적인 수준의 차이가 크며, altmetrics는 순수한 연구 영향력으로 인정하기 어렵고 활성화되지 않은 분야가 많은 것이 한계이다. 개별 논문에 대한 계량적 평가에서는 해당 논문을 직접 인용한 경우뿐만 아니라 그 이후의 후속 연구 전반에 어느 정도로 영향을 끼쳤는지를 평가하는 것이 바람직하다. 이런 점에 착안하여 단일 논문 단위의 인용 영향력 평가 지표를 다룬 연구가 최근 다수 발표되었다. 이재윤(2011a)은 페이지랭크(PageRank) 알고리즘(Page, Brin, Motwani, & Winograd, 1999)을 비롯한 기존 시도들 중에서 선정한 5가지 지수와 수정 제안한 3가지 지수를 포함한 8가지 지수들을 KISTI의 KSCD 데이터베이스를 대상으로 측정하여 비교분석한 바 있다. 또한 이재윤(2011b)은 인용 데이터베이스 전체를 분석하는 경우에 해당하는 새로운 논문 영향력 지수로 c-index를 제안하고, 인용 DB 검색만으로 간단히 측정할 수 있는 변형 지수로 cl-index를 함께 제시하였다.

선행 연구에서 소개된 지수 중에서 단순 인용빈도나 단일 문헌 h-지수(Schubert, 2009), 그리고 cl-index(이재윤, 2011b)는 인용 네트워크 전체를 분석하지 않고 평가 대상 문헌과 그 문헌을 인용한 문헌의 인용빈도만 파악하면 되므로 국지적 네트워크(local network)만 분석하는 지수이다. 이는 특정 논문을 인용한 논

문과 그 인용빈도만 검색하면 산출이 가능하므로 인용 데이터베이스의 이용자라면 누구나 간단하게 산출할 수 있는 지수가 된다. 반면에 페이지랭크를 비롯한 나머지 지수들은 전체 네트워크의 인용 관계를 모두 파악해야 하므로 전역 네트워크(global network)를 분석해야 한다. 결국 인용 데이터베이스의 모든 데이터를 수집하여 분석해야 하므로 인용 데이터베이스를 서비스하는 기관 이외에는 산출이 어려운 지표이다. 지수를 계산하는 과정도 전역 네트워크를 분석하는 페이지랭크 류의 지수는 국지적 네트워크를 대상으로 산출되는 지수보다 오랜 시간과 복잡한 알고리즘이 사용된다.

이 글에서는 기존에 검토된 논문 영향력 지수를 2장에서 간단히 살펴본 다음, KCI의 논문을 대상으로 단순 인용빈도를 비롯한 8종의 인용지수를 적용하여 개별 논문의 영향력을 분석하고 인용지수의 분야 편향성을 살펴보았다. 적용한 인용지수는 유사한 성격의 지수 군에서 대표적인 지수들이다.

2. 논문 인용 영향력 측정 지수

2.1 전역 네트워크 분석 지수

네트워크 전체 인용 관계 정보를 분석하는 전역 네트워크 분석 지수는 계산 시간이 상당히 소요되지만, 여러 단계에 걸친 인용 사슬이 모두 분석되므로 장기간에 걸쳐 후속 연구에 영향을 끼친 문헌을 식별하기에 유리한 방법이다. 기존에 제안된 여러 전역 네트워크 분석 지수 중에서 이 논문에서는 페이지랭크, 복합 인용지수 CCI, f-값, c-지수를 살펴보기로 한다.

페이지랭크 공식은 웹 사이트를 대상으로 중요도를 측정하기 위해서 개발한 것이다(Page et al., 1999). 전체 논문 수가 n개일 때 참고문헌 수가 REF(d_i)개인 논문 i의 페이지랭크 PR(d_i)를 계산하는 공식은 다음과 같다.

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

복합인용지수 CCI(Comprehensive Citation Index)는 페이지랭크와 달리 각 논문이 직접 인용되는 빈도를 크게 반영하도록 고안된 지수이다(Bi, Wang, & Lin, 2011). CCI 공식은 인용하는 논문의 영향력을 참고문헌 수로 나누어 주는 항을 가지고 있으며, 해당 논문의 인용빈도를 더해주므로 직접 인용빈도를 적극적으로 고려하는 방식이다. 아래 CCI 공식에서 β가 0이면 CCI는 직접 인용빈도를 그대로 영향력으로 삼게 되는데 일반적으로 0.3으로 설정한다.

$$CCI(d_i) = CIT(d_i) + \beta \sum_j \frac{CCI(d_j)}{REF(d_j)}$$
$$= \sum_j \left\{ 1 + \beta \frac{CCI(d_j)}{REF(d_j)} \right\}$$

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, Evangelidis, Samaras, & Dervos, 2011). f-값은 인용을 내보내는 논문의 영향력을 참고문헌 수로 나누지 않고 전체를 각 인용논문에게 동일하게 전달하는 것이 가장 큰 특징이다. 이 때문에 f-값 기준으로 분석하면 상위권에 유사한 주제의

논문이 집중될 가능성이 높다. f-값 공식은 아래와 같으며 감쇄상수 값은 1/2.2로 제안되었다.

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

c-지수(c-index)는 앞의 지수들을 고려하여 다음과 같은 원칙을 지키도록 새롭게 제안된 것이다(이재윤, 2011b). 첫째, 각 인용의 중요도는 인용하는 문헌의 중요도에 따라서 차별되어야 한다. 둘째, 인용하는 문헌의 중요도는 인용받는 문헌 각각에게 나누어 전달되어야 한다. 셋째, 인용빈도가 0인 논문으로부터의 인용도 최소 수준의 중요도를 전달해야 한다. 넷째, 인용하는 문헌의 중요도는 일정한 비율로 감쇠되면서 전달되어야 한다. 다섯째, 인용하는 문헌의 중요도를 인용받는 각 문헌에 배분할 때, 배분되는 몫이 참고문헌 수에 따라 지나치게 좌우되지 않도록 한다. 이 다섯 가지 요건을 반영하여 개발된 c-지수 공식은 다음과 같다.

$$c(d_i) = 1 + d \times \sum_j \frac{c(d_j)}{\sqrt{REF(d_j)}}$$

이 공식에서 1은 인용빈도가 0인 논문도 인용하는 논문으로 영향력을 전달하기 위해서 설정한 기본값이다. d는 인용 단계를 건너 전파되는 값을 일정 비율로 감소시키는 감쇠지수로서 페이지랭크와 같이 0.85로 설정해볼 수 있다. 여기까지는 f-값 공식과 유사하나, 전달되는 영향력을 참고문헌의 수를 고려하여 나누는 점이 다르다. 다만 페이지랭크 등과 같이 참고문헌의 수로 나누지 않고 제곱근을 취한 값으로 나누는 이유는, 인용영향력을 분산하여 전달하는 문제를 일종의 문헌길이정규화(document length normalization) (Singhal, Salton, Mitra, & Buckley, 1996) 문제로 간주했기 때문이다(이재윤, 2011b). 정보검색에서 각 문헌에 출현한 단어의 가중치를 산출할 때 문헌 길이의 편차가 심할 경우에는 단어의 출현빈도를 그대로 사용하지 않고 문헌의 길이를 고려하여 정규화한다. 이와 유사하게 인용하는 문헌으로부터 인용되는 문헌으로 전달되는 인용 영향력을 결정할 때 인용하는 문헌의 길이(참고문헌 수)를 고려하여 전달되는 인용영향력을 정규화하였다. 이때 인용하는 문헌의 길이는 참고문헌의 수에 따라 결정되도록 하되, 문헌길이정규화 공식 중에서 전통적으로 오래 사용되어온 코사인정규화 공식을 적용하였다. 단어가중치 계산에서 코사인정규화는 단어 출현빈도의 제곱을 합한 후 제곱근을 취한 값(벡터 norm)으로 나눠주는 것인데, 각 문헌으로의 인용은 빈도가 모두 1이므로 제곱은 무의미하고 내보내는 인용 건수의 제곱근으로 인용영향력을 나누면 된다. 이는 자동분류 실험에서 성능이 좋은 것으로 보고된 루트정규화와 같은 방식이기도 하다(정영미, 이재윤, 2001). 루트 정규화는 단어 출현빈도를 전체 단어빈도합계로 나눈 후 제곱근을 취한 것인데, 한 문헌에서 다른 문헌을 인용한 빈도는 구분하지 않고 1로 처리하므로 분자는 그대로이고 분모의 참고문헌 수만 제곱근을 취한 결과가 되기 때문이다. 이와 같이 내보내는 인용건수가 아닌 인용건수의 제곱근으로 인용영향력을 나누게 되면 문헌 길이, 즉 참고문헌 수에 따라 전달되는 영향력이 지나치게 좌우되는 문제를 해소할 수 있다(이재윤, 2011b).

이상의 전역 네트워크 분석 지수 중에서 페

이지랭크와 복합인용지수 CCI, 그리고 c-지수는 한 논문이 내보내는 영향력을 논문의 참고문헌 수로 나누어줌으로써 영향력을 평균적으로 고르게 배분하도록 고안되어 있는 반면에, f-값은 이와 같은 영향력 배분 장치가 없고 한 논문의 영향력 중에서 감쇄 상수에 의해 일정한 비율로 감쇄된 몫 전체를 선행 연구 각각에게로 전달하도록 되어 있다. 이와 같은 영향력 배분 장치는 지수 산출 결과의 편향성에 영향을 끼칠 가능성이 있다.

2.2 지역 네트워크 분석 지수

지역 네트워크 분석 지수는 평가 대상 문헌과 그 문헌을 인용한 문헌의 인용빈도로 산출될 수 있는데 가장 단순한 방식으로는 논문의 인용빈도를 들 수 있다. 단일 문헌 h-지수(single paper h-index)는 이런 단순 인용빈도를 넘어서서 한 논문을 인용한 논문들의 인용빈도까지 고려하도록 개발된 지수이다. 이는 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."

이와 같은 단일 문헌 h-지수는 계산 방식이 원래의 h-지수와 마찬가지로 간단하므로 Thor와 Bornmann(2011)은 Google Scholar를 이용해서 단일 문헌 h-지수를 산출하는 어플리케이션을 구현하기도 하였다. 이에 대해서 Egghe(2010)는 'h-지수의 특출난 응용'이라는 찬사를 보냈고, Bornmann, Schier, Marx, Daniel(2011)은 Angewandte Chemie International Edition의 논문 심사 과정 데이터를 분석하여 심사자의 판단과 단일 문헌 h-지수가 상관성이 높다고 보고한 바도 있다.

단일 문헌 hs-지수는 일단 단일 문헌 h-지수를 측정한 후 h위 이내에 속한 문헌의 인용빈도마다 제곱근을 취하여 합산한 것이다(이재윤, 2011a). 이는 단일 문헌 h-지수가 원래의 h-지수처럼 정수값으로 산출되므로 변별력이 떨어지고, h위 이내에 포함된 논문의 인용빈도는 아무리 높아져도 h-지수 산출에 영향을 주지 않는다는 문제점을 보완하기 위한 것이다. 연구자의 h-지수를 응용하여 단일 문헌 h-지수가 제안되었듯이, 연구자의 hs-지수(이재윤, 2006)를 응용한 것이 단일 문헌 hs-지수이다.

전역네트워크 지수인 c-지수를 지역 네트워크 분석에 적용하기 위해서 변형한 것이 cl-지수(이재윤, 2011b)이다. c-지수를 설계할 때 고려한 다섯 가지 원칙 중에서 지역 네트워크에 적용할 수가 없는 넷째 원칙인 인용하는 문헌의 중요도가 일정한 비율로 감쇠되면서 전달되어야 한다는 원칙만 제외하고 나머지 네 가지는 cl-지수 설계에도 적용되어 다음과 같이 인용빈도(CIT)와 참고문헌 수(REF)로 구성된 공식이 제안되었다.

$$cl(d_i) = \sum_j \sqrt{\frac{CIT(d_j)+1}{REF(d_j)}}$$

인용을 보내는 논문의 영향력은 인용빈도로 산출하되 인용빈도에 1을 더함으로써 인용빈도

가 0인 논문도 영향력을 전달할 수 있게 설정하였다. 이와 달리 단일 문헌 h-지수는 인용빈도가 0인 논문으로부터의 인용은 전혀 영향력을 향상시키지 못한다. cl-지수에서 인용빈도에 제곱근을 씌운 이유는 인용이 빈익빈 부익부 현상을 보이기 때문에 인용빈도의 차이를 그대로 영향력의 차이로 반영하지 않고 차이를 완화시키기 위해서이다(이재윤, 2011b).

이상의 지역 네트워크 분석 지수 중에서 cl-지수는 한 논문이 내보내는 영향력을 논문의 참고문헌 수로 나누어줌으로써 영향력을 평균적으로 고르게 배분하도록 고안되어 있는 반면에, 단일문헌 h-지수와 hs-지수는 이와 같은 영향력 배분 장치가 없다. 따라서 지역 네트워크 분석 지수의 경우에도 이와 같은 영향력 배분 장치가 지수 산출 결과의 편향성에 영향을 끼칠 가능성이 있다.

3. KCI 논문의 인용 영향력 측정

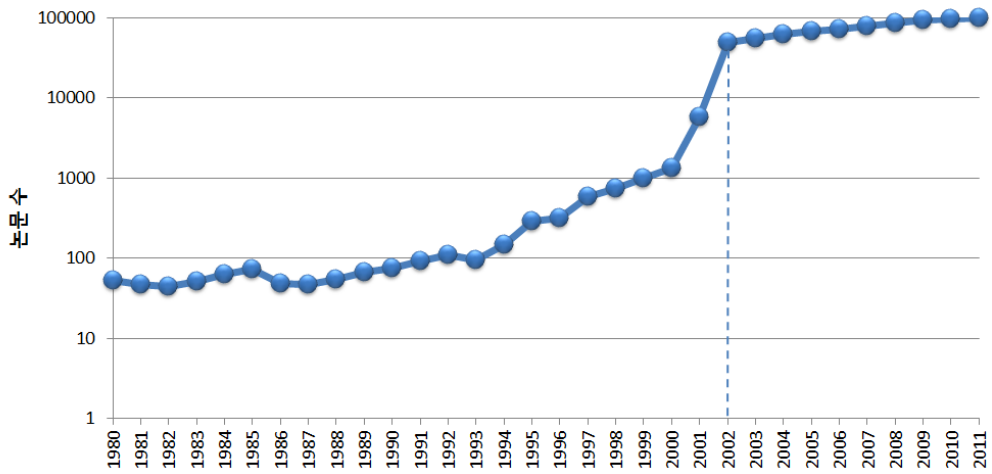
3.1 측정 대상 KCI 논문 데이터

인용 영향력 측정을 위해서 2013년 중반까지 발표된 전체 논문과 인용 정보를 한국연구재단으로부터 입수할 수 있었다. 이중에서 2012년까지 KCI에 등록된 914,114건의 논문들 사이에 이루어진 인용 관계 중에서 오류를 제외하고 정리한 결과 987,943건의 인용 관계가 추출되었다. 논문 1건 당 약 1건의 링크가 존재하는 셈이지만 입수 가능한 데이터의 분석 시점에서는 2012년 논문의 참고문헌 정보가 구축되지 않았으므로 실제로는 2011년까지 발행된 논문 768,353편이 분석 대상이 되었다. 발행연도별 논문 수는 〈표 1〉과 같다.

〈표 1〉에 제시된 논문 중에서 KCI 데이터베

〈표 1〉 발행년도별 논문 수

발행년도	논문 수	누계	발행년도	논문 수	누계
1980	53	53	1997	583	2,250
1981	47	100	1998	729	2,979
1982	44	144	1999	982	3,961
1983	51	195	2000	1,315	5,276
1984	63	258	2001	5,717	10,993
1985	72	330	2002	49,357	60,350
1986	48	378	2003	55,069	115,419
1987	47	425	2004	62,367	177,786
1988	54	479	2005	66,818	244,604
1989	67	546	2006	71,778	316,382
1990	75	621	2007	78,370	394,752
1991	91	712	2008	84,939	479,691
1992	110	822	2009	92,173	571,864
1993	94	916	2010	96,547	668,411
1994	149	1,065	2011	99,942	768,353
1995	287	1,352	2012	101,217	869,570
1996	315	1,667	2013	44,544	914,114



〈그림 1〉 발행년도별 논문 수 (로그 척도)

이스 구축 이전 시기의 논문은 KCI 등재지 제도가 시작된 이후에 발표된 논문으로부터 인용된 정보가 추가된 것이다. 등재지 제도 시작 이전에 발표된 논문은 참고문헌 정보가 구축되지 않았으므로 선행 인용 관계를 분석할 수는 없다. 따라서 인용 관계 네트워크를 분석하는 논문 인용지수는 주로 2002년 이후의 논문에 유리한 지표가 산출된다.

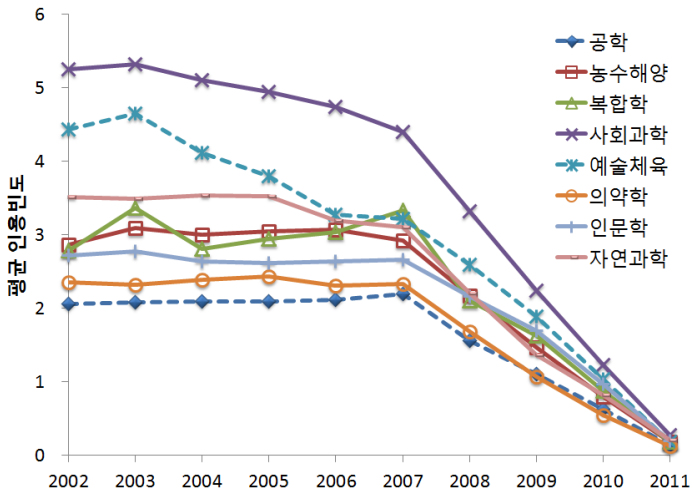
KCI에서 구분된 8개 대분류 분야 논문의 평균 인용빈도는 〈표 2〉와 같이 사회과학 분야 논문이 3.040으로 가장 높고 예술체육 분야 논문과 자연과학 분야 논문이 2.479와 2.133으로 그 다음 순이었다. 〈표 2〉 및 〈그림 2〉에서 논문의 발행년도별 인용빈도 평균을 살펴보면 8개 대분류 분야 중 인문학, 사회과학, 예술체육, 농수해양, 복합학 분야에서 2003년에 발행된 논문의 평균 인용빈도가 가장 높은 것으로 나타난 반면에 공학 분야는 2007년, 의약학 분야는 2005년, 자연과학 분야는 2004년에 발행된 논문의 평균 인용빈도가 가장 높아서 최근 논문에 대한 인용이 많은 분야임을 알 수 있다.

〈그림 2〉에 제시된 발행연도별 평균 인용빈도를 살펴보면 모든 분야에서 2007년까지는 거슬러가면서 인용빈도가 급속하게 증가하지만 2006년 이전 인용빈도의 증가추세가 누그러지거나 정체 상태에 머무는 것으로 보인다. 이는 출판 후 4년을 경계로 인용빈도의 증가 추세가 완만해지기 때문이라고 해석할 수도 있으나, KCI의 인용 데이터베이스 구축 작업이 2007년부터 본 궤도에 올랐기 때문에 2006년 이전 논문에 대한 인용 파악이 상대적으로 불완전한 것에 원인이 있기도 하다. KCI 데이터베이스가 가진 이런 특징은 국내에서 논문 출판 후 경과 시간에 따른 인용패턴을 파악하기에는 아직 더 많은 시간이 필요함을 시사한다. 이는 다음에 살펴볼 논문 인용지수와 출판년도의 상관관계 분석에도 제한점으로 작용하게 된다.

〈표 2〉 8개 대분류 분야 논문의 출판년도별 평균 인용빈도

출판년도	인문학	사회과학	예술체육	복합학	자연과학	공학	농수해양	의약학	전체
2002	2.715	5.254	4.438	2.788	3.510	2.064	2.873	2.353	3.476
2003	2.781	5.322	4.656	3.367	3.488	2.085	3.097	2.318	3.587
2004	2.643	5.110	4.116	2.808	3.535	2.096	3.007	2.394	3.487
2005	2.620	4.950	3.805	2.945	3.531	2.098	3.053	2.433	3.414
2006	2.645	4.741	3.282	3.040	3.201	2.123	3.077	2.307	3.268
2007	2.666	4.407	3.219	3.337	3.104	2.197	2.928	2.335	3.172
2008	2.161	3.327	2.595	2.102	2.205	1.565	2.170	1.681	2.344
2009	1.696	2.248	1.893	1.624	1.374	1.106	1.469	1.067	1.625
2010	0.966	1.232	1.041	0.878	0.827	0.628	0.804	0.554	0.911
2011	0.203	0.276	0.185	0.191	0.211	0.143	0.185	0.132	0.203
전체	1.864	3.040	2.479	1.780	2.133	1.317	1.967	1.498	2.132

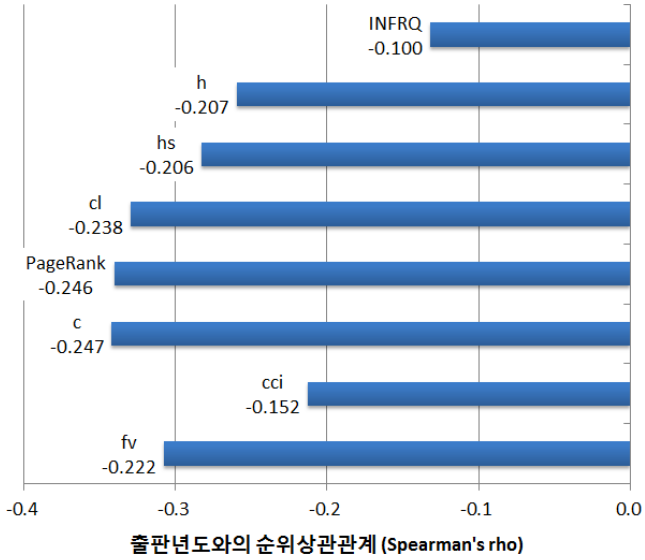
(진한 상자는 해당 대분류 영역에서 최댓값 표시)



〈그림 2〉 8개 대분류 분야 논문의 출판년도별 평균 인용빈도

3.2 KCI 논문 발표 연도와 인용 영향력의 상관관계 분석

2011년까지 발간되어 KCI 데이터베이스에 수록된 768,353건의 논문들 사이의 987,943건의 인용 관계로부터 7종의 논문 영향력 지수를 측정해보았다. 7종에는 단일문헌 h-지수(h로 약칭), 단일문헌 hs-지수(hs로 약칭), cl-지수(cl로 약칭), c-지수(c로 약칭), 페이지랭크, 포괄적 인용 지수 CCI(cci로 약칭), f-값(fv로 약칭)이 포함된다. 이중에서 단일문헌 h-지수, 단일문헌 hs-지수, cl-지수의 세 종은 단순 검색 결과만으로 측정할 수 있는 지역 네트워크 분석 방식 지수표이고, 페이지랭크, c-지수, 복합 인용지수 CCI, f-값의 네 종은 데이터베이스내 전체 인용 관계를 분석해야 측정할 수 있는 전



〈그림 3〉 논문 영향력 지수와 출판년도와의 순위 상관관계

역 네트워크 분석 방식 지수이다.

인용 지수가 측정된 논문 중에서 인용빈도 3회 이상인 논문들에 대해서 논문의 발표 연도와 각 인용 지수와의 스피어맨 순위상관관계를 측정해보았다. 일반적으로 오래된 논문일수록 인용될 기회가 많아지고 후속 논문에 대한 인용도 연쇄적으로 발생하여 영향력이 더 커질 수 있다. 〈그림 3〉의 측정 결과를 보면 단순 인용빈도(INFRQ)가 출판년도와의 상관성이 가장 낮고 c-지수, 페이지랭크, cl-지수가 가장 상관성이 높은 것으로 나타났다. 전역 네트워크 분석 방식의 지수 중에서 c-지수와 페이지랭크가 CCI나 f-값에 비해서 출판년도와의 상관성이 높은 이유는, 영향력을 연쇄적으로 전달하면서 다음 인용단계로 전달되는 영향력 비율이 CCI는 0.3, f-값의 경우는 약 0.5 정도인 반면에 c-지수와 페이지랭크는 0.85였기 때문인 것으로 짐작된다. c-지수나 페이지랭크를 측정하면서 오래된 논문에 유리한 정도를 낮추려면 c-지수 공식과 페이지랭크 공식에서 사용하는 d값을 0.85보다 낮은 0.5 내외로 줄이는 것도 고려해볼만 하다.

3.3 논문 영향력 지수의 학문 분야별 균등성

일반적으로 논문의 피인용은 주제 분야에 따라 매우 상이한 형태를 보인다. 분야에 따라 연구자 수, 발표 논문 수, 평균 참고문헌 수, 출판 소요시간, 피인용 횟수 등이 달라 주제 분야 간에 직접적인 비교를 하는 것은 타당하지 않기 때문이다(신은자, 2013, p. 132). 논문 영향력 지수가 가져야 할 중요한 특성 중 하나는 이처럼 특정 학문 분야 논문이 타 분야 논문에 비해서 높은 수치를 획득하는 경향이 적어야 한다는 점이다. 이를 학문 분야별 균등성이라고 표

현할 수 있다. 각 지수에 대해서 이런 학문 분야별 균등성이 어느 정도인가를 알아보기 위해서 각 분야별 1위 논문이 전체 순위에서는 몇 위에 해당하는가를 파악해보았다. 각 분야별 1위 논문의 전체 순위의 평균을 인용 영향력 지수별로 산출한 결과를 〈표 3〉에 제시하였다. 〈표 3〉

〈표 3〉 각 분야별 인용빈도 1위 논문의 전체 순위

학문분야	순위	학문분야	순위	학문분야	순위	학문분야	순위
경영학	1	영어와문학	410	기술정책	2259	약리학	7191
역사학	2	통계학	447	토목공학	2259	영화	7191
교육학	4	인문학	551	피부과학	2259	예술일반	7191
회계학	5	자연과학	608	기타자연과학	2576	자원공학	7191
체육	7	공학일반	652	대기과학	2576	기계공학	8564
사회과학	9	유교학	697	독일어와문학	2576	기타공학	8564
사회복지학	11	해상운송학	697	디자인	2576	사전학	8564
행정학	14	간호학	752	미용	2576	의상	8564
사회학	16	농업경제학	814	신경과학	2576	임상병리학	8564
식품과학	16	재활의학	886	원자력공학	2942	전기공학	8564
정책학	20	조경학	886	산부인과학	3361	조선공학	8564
예방의학	23	환경공학	886	수의학	3361	병리학	10200
정치외교학	26	무역학	973	연극	3361	자동차공학	10200
학제간연구	28	일반외과학	973	음악학	3361	해양학	10200
심리과학	29	의학일반	1079	일본어와문학	3361	러시아어와문학	12251
의공학	36	한의학	1079	종교학	3361	서양고전어와문학	12251
관광학	38	가정의학	1191	지질학	3361	이비인후과학	12251
문학	40	재료공학	1191	무용	3880	기타동양어문학	15028
사회과학일반	48	축산학	1191	전자/정보통신공학	3880	생화학	15028
신문방송학	54	산업공학	1324	통역번역학	3880	정형외과학	15028
생활과학	83	예술체육	1471	프랑스어와문학	3880	제어계측공학	15028
철학	86	약학	1633	물리치료학	4469	항공우주공학	15028
소아과학	90	치의학	1633	미술	4469	기독교신학	18855
경제학	102	문헌정보학	1823	비뇨기과학	4469	기타예술체육	18855
기타사회과학	103	물리학	1823	생물학	4469	섬유공학	18855
농수해양	103	방사선과학	1823	응급의학	4469	미생물학	23926
여성학	124	불교학	1823	해양공학	4469	신경외과학	23926
한국어와문학	135	생물공학	1823	고분자공학	5227	흉부외과학	23926
농학	162	컴퓨터학	1823	교통공학	5227	가톨릭신학	30985
임학	162	공학	2027	안전공학	5227	금속공학	30985
정신과학	228	기타인문학	2027	자연과학일반	5227	스페인어와문학	30985
의약학	243	내과학	2027	작업치료학	5227	군사학	40994
지역개발	313	수산학	2027	감성과학	6119	성형외과학	40994
기타의약학	379	수학	2027	인지과학	6119	해부학	40994
법학	379	지구과학	2027	중국어와문학	6119	면역학	79157
지리학	379	지역학	2027	마취과학	7191	천문학	116995
화학	379	화학공학	2027	복합학	7191	과학기술학	185290
언어학	410	건축공학	2259	안과학	7191	뇌과학	326432

을 보면 경영학, 역사학, 교육학, 회계학, 체육 등과 같이 연구자의 수가 많은 학문분야의 경우 분야별 1위가 전체 순위 10위 이내에 포진하고 있는 것으로 나타난다. 반면에 러시아어와문학, 기타동양어문학과 같이 연구자가 적은 학문분야의 인용빈도 1위 논문은 전체 10,000위 밖으로 밀린다. 병리학, 이비인후과학, 생화학 등과 같이 해외 문헌을 주로 인용하는 분야의 경우에도 분야 내 인용빈도 1위 논문이 전체 순위 10,000위보다 아래로 나타났다.

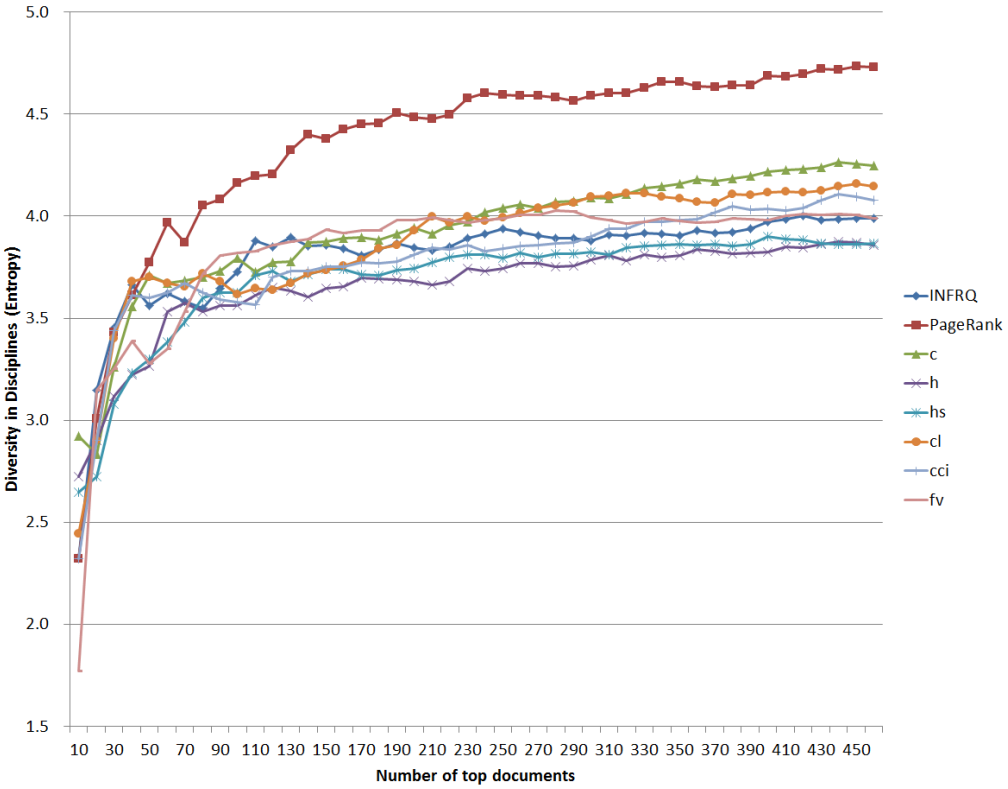
각 분야별 1위 논문의 전체 순위의 평균을 인용 영향력 지수별로 산출한 결과를 〈표 4〉에 제시하였다. 분야별 최고 인용빈도가 극단적으로 낮은 학문분야도 있어서 면역학은 인용빈도 3회, 천문학은 인용빈도 2회, 과학기술학은 인용빈도 1회 논문이 최고에 불과했으며 뇌과학은 모든 논문의 인용빈도가 0회였다. 〈표 4〉에서는 전체 평균 순위 이외에 이 네 분야를 제외한 평균 순위도 제시하였다.

〈표 4〉를 보면 분야별 1위 논문의 전체 순위 평균이 가장 낮은 지수는 페이지랭크인 것으로 나타났다. 이는 페이지랭크를 기준으로 하였을 때 상대적으로 학문 분야별로 고르게 상위 논문이 파악된다는 의미이다. 직접 인용빈도 기준 평균 순위보다 낮은 평균 순위를 보여주는 지수는 페이지랭크 이외에 c-지수, cl-지수, 단일문헌 h-지수가 있었으며, CCI, 단일문헌 hs-지수, f-값 등은 인용빈도 기준 평균 순위보다도 평균 순위가 더 커서 여러 분야의 논문이 고르게 상위 순위를 차지하지 못하는 것을 알 수 있다.

분야별 1위 논문을 기준으로 하지 않고 전체 KCI 논문 전체 중에서 각 지수별 상위권에 다양한 학문 분야의 논문이 고르게 포함되는가도 살펴보았다. 〈그림 4〉는 10위부터 500위까지 전체 순위를 10씩 증가시키면서 각 순위까지의 논문 중에서 학문 분야별 논문의 분포가 얼마나 고르게 되어 있는지를 엔트로피로 측정한 결과이다. 이처럼 각 지수의 학문 분야 다양성 반영 정도를 엔트로피로 측정하면 수치가 높을수록 다양한 학문 분야의 논문이 상위권에 고르게 포함된다는 의미이다. 〈그림 4〉를 보면 페이지랭크 기준 순위가 타 지수 순위에 비해서 최상위에 다양한 학문분야의 문헌이 포함되고 있는 것으로 뚜렷하게 나타난다. 다른 지수들의 엔트로피는 크게 차이나지는 않지만 250위 이후부터는 전역 네트워크 분석 지수인 c-지수와 지역 네트워크 분석 지수인 cl-지수가 지속적으로 높은 엔트로피를 보였다. 다른 지역 네트워크 분석 지수인 단일 문헌 h-지수와 hs-지수는 가장 낮은 엔트로피를 보여서 일부 학문분야에 편향된 순위를 산출하는 경향이 있음을 알 수 있다.

〈표 4〉 분야별 최상위 문헌의 전체 순위 평균

	INFRQ	h	hs	cl	PageRank	c	cci	fv
분야별 최상위 문헌의 전체 순위 평균	9843.3	7984.9	11371.0	8745.6	5459.6	8335.8	10836.4	13672.5
최다 인용이 3회 이하인 4개 분야를 제외한 분야별 최상위 문헌의 전체 순위 평균	5326.4	5114.3	8310.3	4472.7	2011.8	4248.5	6033.4	9013.9



〈그림 4〉 전체 순위 기준 상위 문헌의 학문분야별 분포 다양성

3.4 논문 영향력 지수의 학술지별 균등성

논문 영향력 지수가 학문 분야별로 균등한 순위를 산출하더라도 한 학문 분야 내에서 특정 학술지에 유리한 편향성을 가질 가능성도 있다. 각 지수에 대해서 이런 학술지별 균등성이 어느 정도인가를 알아보기 위해서 시험적으로 언어학 분야에 대해서 인용빈도, 페이지랭크, 단일 문헌 h-지수 기준으로 각각 상위 논문이 게재된 학술지를 파악하여 〈표 5〉, 〈표 6〉, 〈표 7〉에 제시하였다. 각 표에서 해당 학술지의 논문의 순위 내에 포함되지 못해서 값이 0인 경우는 칸을 비워 놓았다. 단순 인용빈도와 단일 문헌 h-지수로는 50위 이내에 11종의 학술지 논문만 포함되었으나, 페이지랭크 기준으로는 훨씬 더 다양한 16종의 학술지 논문이 포함된 것으로 나타났다. 또한 50위 이내에 가장 많은 논문이 포함된 학술지의 비중을 살펴보아도, 단순 인용빈도와 단일 문헌 h-지수로는 〈이중언어학〉의 논문이 21개로 42%를 차지하였지만 페이지랭크 기준으로는 그 절반 정도인 11개로 50위 이내 상위 논문의 22%만 차지한 것으로 나타났다. 결국 페이지랭크를 기준으로 논문의 영향력을 측정하면 상위권에 더 다양한 학술지의 논문이 포함될 뿐만 아니라 특정 학술지의 상위권 과점 현상도 적음을 알 수 있다.

〈표 5〉 인용빈도 기준 상위 순위에 포함된 언어학 분야 학술지별 논문 수

언어학 분야 학술지	언어학 분야 상위 순위에 포함된 논문 수				
	10위	20위	30위	40위	50위
이중언어학	8	13	15	19	21
Foreign Languages Education	1	4	7	9	11
텍스트언어학	1	1	2	2	2
언어학		1	2	4	5
응용언어학		1	1	1	3
담화와 인지			1	2	2
언어과학연구			1	1	2
형태론			1	1	1
언어와 정보				1	1
사회언어학					1
생성문법연구					1

〈표 6〉 페이지랭크 기준 상위 순위에 포함된 언어학 분야 학술지별 논문 수

언어학 분야 학술지	언어학 분야 상위 순위에 포함된 논문 수				
	10위	20위	30위	40위	50위
Foreign Languages Education	6	7	8	8	10
이중언어학	3	5	5	8	11
응용언어학	1	1	1	1	1
생성문법연구		2	2	2	3
언어학		1	2	4	5
언어과학연구		1	2	2	2
어학연구		1	1	2	3
형태론		1	1	1	2
사회언어학		1	1	1	1
담화와 인지			2	3	3
텍스트언어학			2	2	2
음성과학			1	3	3
언어			1	1	1
언어와언어학			1	1	1
기호학 연구				1	1
언어연구					1

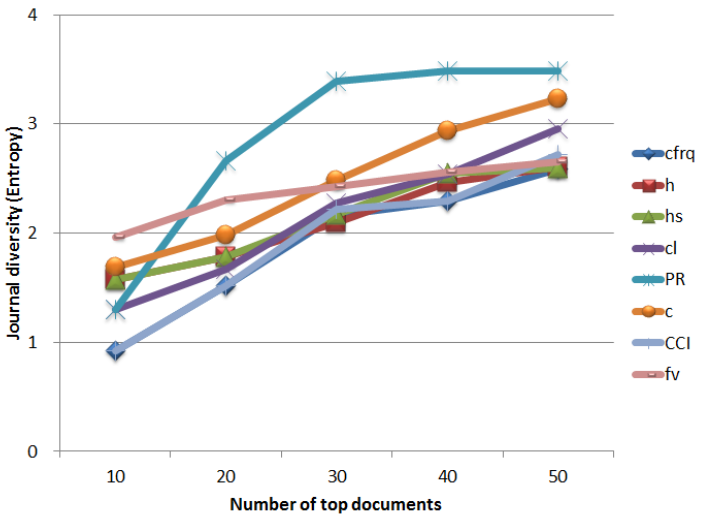
8개 인용 지수별로 언어학 분야 10위부터 50위까지의 논문이 게재된 학술지가 어느 정도로 다양하게 분포되었는가를 엔트로피로 측정해 보면 〈표 8〉 및 〈그림 5〉와 같다. 50위권까지 범위를 넓히게 되면 페이지랭크가 1위, c-지수가 2위, cl-지수가 3위로 나타난다.

〈표 7〉 단일 문헌 h-지수 기준 상위 순위에 포함된 언어학분야 학술지별 논문 수

언어학 분야 학술지	언어학 분야 상위 순위에 포함된 논문 수				
	10위	20위	30위	40위	50위
이중언어학	6	11	16	18	21
언어학	2	3	3	3	4
텍스트언어학	1	1	1	2	2
언어	1	1	1	1	2
Foreign Languages Education		4	6	9	11
담화와 인지			1	2	3
언어과학연구			1	2	3
형태론			1	1	1
응용언어학				1	1
생성문법연구				1	1
기호학 연구					1

〈표 8〉 상위 문헌의 게재 학술지 다양성(엔트로피) - 언어학 분야

	10위 이내	20위 이내	30위 이내	40위 이내	50위 이내
INFRQ	0.922	1.517	2.165	2.291	2.591
h	1.571	1.781	2.098	2.463	2.608
hs	1.571	1.781	2.170	2.547	2.594
cl	1.295	1.671	2.275	2.541	2.952
PageRank	1.295	2.659	3.387	3.484	3.486
c	1.685	1.980	2.485	2.937	3.226
cci	0.922	1.517	2.219	2.296	2.722
fv	1.961	2.304	2.423	2.556	2.648



〈그림 5〉 언어학 분야 상위 순위에 포함된 논문의 게재학술지 분포 다양성(엔트로피)

4. 결 론

연구 성과에 대한 평가 기준으로 널리 사용되고 있는 게재 학술지의 인용 수준을 대체, 또는 보완할 수 있는 지표로 단일 논문의 인용 영향력을 측정할 필요가 있다. 이 글에서는 단순 인용빈도를 비롯해서 기존에 제안된 8개 지수를 살펴보고 KCI 논문 데이터베이스를 대상으로 측정해보았다.

논문의 발표 연도와 각 인용 지수와의 스피어맨 순위상관관계를 측정해본 결과 c-지수, 페이지랭크, cl-지수의 순서대로 가장 상관성이 높은 것으로 나타났다. 이 세 지수가 단계적 인용을 통해 누적되는 영향력을 상대적으로 더 크게 반영하는 것으로 해석된다. 반면 단순 인용빈도는 논문 발표연도와 상관성이 가장 낮았다.

인용 지수가 학문 분야별 균등성을 유지하는지, 학문 분야 내에서 학술지별 균등성을 유지하는지를 실제로 측정해본 결과 페이지랭크가 모든 영역에서 가장 뛰어난 것으로 나타났다. 반면에 단순 인용빈도는 특정 학문분야와 학술지에 편향된 결과를 산출할 가능성이 높은 것으로 나타났다. 페이지랭크는 대표적인 전역 네트워크 지수이므로 영향력을 산출하는데 너무 많은 데이터를 필요로 하는 단점이 있다. 지역 네트워크 분석 지수인 cl-지수는 상위 250위 이내 논문에서는 페이지랭크 이외의 전역 네트워크 지수와 비슷한 분야 균등성을 보였으며, 250위 이상의 순위에서는 페이지랭크 다음으로 좋은 분야 균등성을 보여주어 상대적으로 적은 데이터로 산출가능한 지역 네트워크 인용지수로서 활용 가능성이 있는 것으로 나타났다.

페이지랭크에 비해서 cl-지수와 같은 지역 네트워크 지수가 가지는 또 하나의 장점은, 시간이 경과하면서 논문이 추가되고 인용이 늘어나면 지수가 자연 증가한다는 점이다. 페이지랭크는 전체 논문의 지수값 합계가 일정하게 유지되는 상대적인 평가 방식이므로 인용이 추가되더라도 특정 논문의 지수값은 오히려 줄어들 수도 있다. 신규 논문이 추가되면 그만큼 영향력을 나눠가지게 되므로 데이터베이스의 규모가 커질수록 절반 이상의 논문은 오히려 페이지랭크로 측정한 상대적인 영향력이 감소하는 것으로 보이게 된다. 이는 정보서비스의 측면에서 오해를 불러올 수 있으므로 바람직하지 않은 특성이다. 반면에 cl-지수와 같은 지역 네트워크 지수는 특정 논문에 대한 직접 인용이 추가되거나, 해당 논문을 인용한 논문의 인용 빈도가 증가하면 측정된 영향력 수치도 증가하게 된다. 따라서 인용 영향력이 증가하는 추세를 살펴볼 수 있다는 것이 큰 장점이다.

현재 KCI 데이터베이스에서는 논문의 단순 인용빈도만 제공하고 있는데, 전체 DB를 분석하면서 균등성을 가장 잘 유지하는 지수인 논문 페이지랭크를 정기적으로 산출하여 평가할 필요가 있다. 아울러 인용한 문헌의 인용빈도만으로 산출이 가능해서 이용자의 검색 결과로부터 실시간으로 산출할 수 있고 영향력의 증가 추세를 직관적으로 파악할 수 있으며 분야 편향성이 적은 cl-지수를 실시간 인용 지수로 제공하는 것이 바람직하다.

이 연구에서 분석한 데이터는 2013년까지의 KCI 데이터였지만, 인용 데이터는 2011년까지 인용된 데이터만 사용 가능하였다. 3.1절에서 살펴보았듯이 KCI 인용 데이터가 2006년 이전에는 참고문헌 데이터 구축이 불완전할 것이라는 부

분적인 증거가 관찰되었다. 연도에 따른 인용 빈도의 증가추세 변화가 모든 분야에서 동일하게 나타나므로 이런 현상이 본 연구의 주된 관심사인 분야 편향성 분석에 영향을 주지는 않았다. 그러나 논문 출판 후 경과 시간에 따른 인용패턴 추세, 그리고 논문 인용지수와 출판년도의 상관관계 분석을 위해서는 추후 더 오랜 기간의 데이터를 확보하여 검증할 필요가 있을 것이다.

참 고 문 헌

신은자 (2013). 한국 재료공학 논문의 피인용 영향요인에 관한 연구. 정보관리학회지, 30(1), 131-150. http://dx.doi.org/10.3743/KOSIM.2013.30.1.131

이재윤 (2006). 연구성과 측정을 위한 h-지수의 개량에 관한 연구. 정보관리학회지, 23(3), 167-186. http://dx.doi.org/10.3743/KOSIM.2006.23.3.167

이재윤 (2011a). 인용 네트워크 분석에 근거한 문헌 인용 지수 연구. 한국문헌정보학회지, 45(2), 119-143. http://dx.doi.org/10.4275/KSLIS.2011.45.2.119

이재윤 (2011b). 단일 문헌의 인용 영향력 측정 방식의 개선. 제18회 한국정보관리학회 학술대회 발표논문집, 119-143.

정영미, 이재윤 (2001). 지식 분류의 자동화를 위한 클러스터링 모형 연구. 정보관리학회지, 18(2), 203-230.

Bi, H. H., Wang, J., & Lin, D. K. J. (2011). Comprehensive citation index for research networks. IEEE Transactions on Knowledge and Data Engineering, 23(8), 1274-1278. http://doi.org/10.1109/TKDE.2010.167

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2011). Does the h index for assessing single publications really work? A case study on papers published in chemistry. Scientometrics, 89(3), 835-843. http://dx.doi.org/10.1007/s11192-011-0472-0

Egghe, L. (2010). On the relation between Schubert's h-index of a single paper and its total number of received citations. Scientometrics, 84(1), 115-117. http://dx.doi.org/10.1007/s11192-009-0062-6

Fragkiadaki, E., Evangelidis, G., Samaras, N., & Dervos, D. A. (2011). f-Value: Measuring an article's scientific impact. Scientometrics, 86(3), 671-686. http://dx.doi.org/10.1007/s11192-010-0302-9

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569-16572. http://doi.org/10.1073/pnas.0507655102

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. Technical Report. Stanford InfoLab. Retrieved from http://ilpubs.stanford.edu:8090/422/

Schubert, A. (2009). Using the h-index for assessing single publications. Scientometrics, 78(3), 559-565. http://doi.org/10.1007/s11192-008-2208-3

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. Information Processing & Management, 32(5), 619-633.

San Francisco Declaration on Research Assessment: Putting science into the assessment of research. Retrieved from http://www.ascb.org/dora/

Thor, A., & Bornmann, L. (2011). 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. http://dx.doi.org/10.1108/14684521111128050

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Chung, Young-Mee, & Lee, Jae-Yun (2001). Development of a clustering model for automatic knowledge classification. Journal of the Korean Society for Information Management, 18(2), 203-230.

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. http://dx.doi.org/10.3743/KOSIM.2006.23.3.167

Lee, Jae Yun (2011a). A study on document citation indicators based on citation network analysis. Journal of the Korean Society for Library and Information Science, 45(2), 119-143. http://dx.doi.org/10.4275/KSLIS.2011.45.2.119

Lee, Jae Yun (2011b). Improved methods for assessing single paper's citation impact. Proceedings of the 18th Annual Conference of the Korean Society for Information Management, 119-143.

Shin, Eun-Ja (2013). An analysis on the factors affecting the citation in Korean material science articles. Journal of the Korean Society for Information Management, 30(1), 131-150. http://dx.doi.org/10.3743/KOSIM.2013.30.1.131