단일 문헌의 인용 영향력 측정 방식의 개선

Improved methods for assessing single paper's citation impact

이재윤, 경기대학교, memexlee@kgu.ac.kr

Jae Yun Lee, Kyonggi University

최근 인용 네트워크 분석을 통해서 단일 문헌의 인용 영향력을 측정하려는 시도가 다양하게 전개되고 있다. 널리 알려진 PageRank를 보완하려는 일부 시도와 함께 h-index를 단일 문헌의 인용 영향력 측정에 적용한 단일문헌 h-index도 제안되었다. 이 연구에서는 Web of Science에서 검색한 계량정보학 분야 문헌집합을 대상으로 여러 측정 방식의 특징을 비교해본 후, 새로운 인용지수 2종을 제안하였다. 제안한 인용지수는 PageRank처럼 전역 네트워크 분석 방식인 지수 1종과 h-index처럼 지역 네트워크 분석 방식인 지수 1종으로서 상황에 따라 선택하여 사용할 수 있다.

1. 서론

인용빈도를 단서로 하여 학술적인 영향력을 평가하는 인용지수는 주로 학술지나 연구자/연구집단과 같이 여러 문헌의 집합체가 되는 대상에 대해서 적용되었다. 단일 논문에 대해서는 단순 인용빈도 이외의 평가 지표가 없었으나 최근 수년 동안 PageRank(Page et al. 1999) 공식을 문헌 인용 네트워크에 적용하거나(Chen et al. 2007; Radev et al. 2009) h-index(Hrisch 2005) 방식을 응용하여 단일 문헌의 인용 영향력을 측정하려는 시도(Schubett 2009)를 비롯한 여러 연구가 발표되었다.

이재윤(2011)에서는 이런 기존 시도들 중에서 선정한 5가지 지수와 수정 제안한 3가지 지수를 포함한 8가지 지수들을 KISTI의 KSCD 데이터베이스를 대상으로 측정하여 비교분석한 바 있다. 분석된 지수는 PageRank 이외에 Schubert(2009)의 단일문헌 h-index, Sidiropoulos와 Manolopoulos (2005)의 SCEAS_B1, Bi 등(2011)의 CCI, Fragkiadaki 등(2011)의 f-value, 그리고 단일문헌 h-index를 이재윤(2011)이 개량한 h_S-index, h1-index, h_S1-index이다.

선행연구에서 분석된 지수 중에서 h-index와 그 변형 지수는 인용 네트워크 전체를 보지 않고 평가 대상 문헌과 그 문헌을 인용한 문헌만 파악하면 되므로 국지적 네트워크(local network)만 분석하면 된다. 반면에 PageRank를 비롯한 나머지 지수들은 전체 네트워크의 인용 관계를 모두 파악해야 하므로 전역 네트워크(global network)를 분석해야 한다. 이는 마치 사회 네트워크 분석에서 중심성 지수가 지역 중심성 지수와 전역 중심성 지수로 나뉘는 것과 유사하다. 전역 네트워크 분석은 지역 네트워크 분석에 비해서 훨씬 많은 자원이 필요하며, 경우에 따라서는 전체 인용 네트워크를 획득하기 어려운 경우도 많다.

이 논문에서는 WoS에서 검색한 계량정보학 분야 문헌 집합(Lee and Choi 2011)을 대상으로 8가지 인용지수를 측정하여 특징을 비교해본다. 또한 비교적 단순하지만 분석 결과에서 단점이 뚜렷하게 드러난 f-value 공식을 개량하여 전역 네트워크와 지역 네트워크에 각각 적용할 수 있는 직관적이면서 단순한 공식 2종을 제안한다.

2. 분석 대상 지수와 문헌 집합

이 연구에서 분석한 8가지 기존 인용 지수

<표 1> 문헌 d_i의 중요도 $W(d_i)$를 계산하는 공식

명 칭	공 식	
PageRank	$W(d_i) = \frac{1-d}{n} + d \times \sum_j \frac{W(d_j)}{CO(d_j)}$	, d = 0.85
SCEAS_B1	$W(d_i) = (1-d) + d \times \sum_j \frac{W(d_j)+b}{CO(d_j)} \times a^{-1}$	, d = 0.85; b = 1; $a = e$
CCI	$W(d_i) = CI(d_i) + \beta \times \sum_j \frac{W(d_j)}{CO(d_j)}$	, β = 0.3
f-value	$W(d_i) = 1 + RF \times \sum_j W(d_j)$	, RF = 0.45
h-index	$W(d_i) = \sum_j f(d_j)$, $f(d_j) = \begin{cases} 1 & if \quad CI(d_j) \geq CR(d_j) \\ 0 & else \end{cases}$	
h_S-index	$W(d_i) = \sum_j f(d_j)$, $f(d_j) = \begin{cases} \sqrt{CI(d_j)} & if \quad CI(d_j) \geq CR(d_j) \\ 0 & else \end{cases}$	
h1-index	$W(d_i) = \sum_j f(d_j)$, $f(d_j) = \begin{cases} 1 & if \quad CI(d_j)+1 \geq CR(d_j) \\ 0 & else \end{cases}$	
h_S1-index	$W(d_i) = \sum_j f(d_j)$, $f(d_j) = \begin{cases} \sqrt{CI(d_j)} & if \quad CI(d_j)+1 \geq CR(d_j) \\ 0 & else \end{cases}$	

공식은 <표 1>과 같다. 각 공식에 사용된 항의 의미는 <표 2>에 설명을 제시하였다(각 공식에 대한 자세한 설명은 이재윤(2011)을 참고).

인용 지수 측정 대상이 되는 문헌집합은 WoS에서 검색한 계량정보학 분야 문헌 1,715개와 이 문헌들로부터 인용된 39,795개를 합친 41,510개이다. 문헌들 사이의 인용 링크는 총 65,806개로 나타났다. 인용 정보는 초기 검색된 1,715개 문헌에서 인용된 경우만 파악하였으므로 초기 문헌집합으로부터 인용을 받지 못한 문헌은 설사 검색되지 않은 다른 문헌으로부터 인용되었을지라도 인용빈도가 0으로 간주된다. 또한 초기 검색문헌에 속하지 않았으나 이들로부터 인용된 39,795개는 받은 인용만 파악되어 있고 이들로부터 나간 인용 정보는 배제되어 있다. 문헌 검색 범위를 2001년 이후 WoS에 포함된 문헌으로 제한하였으므로 이 인용 데이터는 2001년 이후 계량정보학 연구에서 활용된 연구 주제를 반영하는 집합이라고 할 수 있다(검색식과 함께 검색된 1,715건에 대한 자세한 설명은 Lee and Choi(2011)를 참고).

<표 2> 문헌 인용 영향력 측정에 활용되는 항

항	의미
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
$CO(d_j)$	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

3. 기존 인용지수 적용 결과 분석

3.1 인용빈도 상위 문헌 비교

41,510개 문헌에 사이의 인용 네트워크에서 각 문헌의 인용빈도(CFRQ)와 2세대 인용빈도(CFRQ-2gen)를 산출한 후, 8가지 인용지수를 적용하여 인용 지수를 측정해보았다. 측정 결과 파악된 인용빈도 상위 10건에 대해서 각 지수별 순위를 산출한 결과를 <표 3>에 제시하였다.

<표 3> 인용빈도 상위 10건의 지수별 순위

빈도		문헌	인용지수별 순위							
CFRQ	CFRQ-2gen		Page Rank	SCEAS_B1	CCI	f-value	h-index	hs-index	h1-index	hs1-index
172	240	HIRSCH JE, 2005, An index to quantify an individual's scientific research output	1	1	1	174	59	116	49	95
106	270	GARFIELD E, 1972, Citation analysis as a tool in journal evaluation	2	2	2	141	37	61	30	50
92	302	GARFIELD E, 1955, Citation indexes for science	3	3	3	29	28	38	22	35
85	219	SEGLEN PO, 1997, Why the impact factor of journals should not be used for evaluating research	4	4	4	218	37	109	49	116
82	389	WHITE HD, 1998, Visualizing a discipline: An author co-citation analysis of information science	5	6	5	28	19	21	13	18
82	315	SMALL H, 1973, Co-citation in the scientific literature: A new measure of the relationship between two documents	10	10	6	118	10	22	13	25
76	280	WHITE HD, 1981, Author cocitation: A literature measure of intellectual structure	14	14	7	125	37	47	30	42
74	157	MOED HF, 2005, Citation Analysis in Research Evaluation	6	5	9	340	59	124	49	106
74	349	MACROBERTS MH, 1989, Problems of citation analysis: A critical review	15	15	8	14	19	24	22	26
69	684	INGWERSEN P, 1998, The calculation of web impact factors	17	19	10	1	1	1	1	1

측정결과 PageRank와 SCEAS_B1, CCI는 1세대 인용빈도가 1위인 Hirsch의 2005년 논문(h-index를 제안한 논문)을 1위로 판정하였다. 반면에 f-value와 h-index 계열 4개 지수는 모두 2세대 인용빈도가 1위인 Ingwerson의 1999년 논문(웹 IF를 제안한 논문)을 1위로 판정하였다.

또한 PageRank와 SCEAS_B1은 인용빈도 20위 이내 문헌들이 지수별 상위 10위 이내에 자리잡았고, CCI 기준 상위 10위 문헌은 인용빈도 10위 이내 문헌과 순위만 약간 다를 뿐 거의 일치하였다. 이 세 지수는 모두 인용하는 논문의 참고문헌 수 $CO(d_j)$가 공식의 분모에 포함되어있다. 이와 대조적으로 h-index를 비롯한 나머지 5개 지수별 상위 10위 문헌 중에서는 인용빈도 10위 이내 문헌을 찾기가 어려웠다.

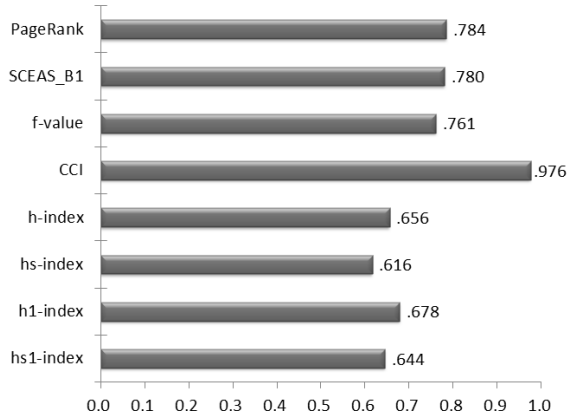
이와 같은 결과는 어느 지수를 선택하느냐에 따라서 최상위권에 포함되는 문헌들이 크게 달라질 수 있음을 의미한다. f-value를 제외하면 대체로 전역 네트워크 분석에 해당하는 인용지수(PageRank, SCEAS_B1, CCI)와 지역 네트워크 분석에 해당하는 지수(h-index 계열)로 크게 구분된다고 볼 수도 있다.

3.2 인용지수의 특성 분석

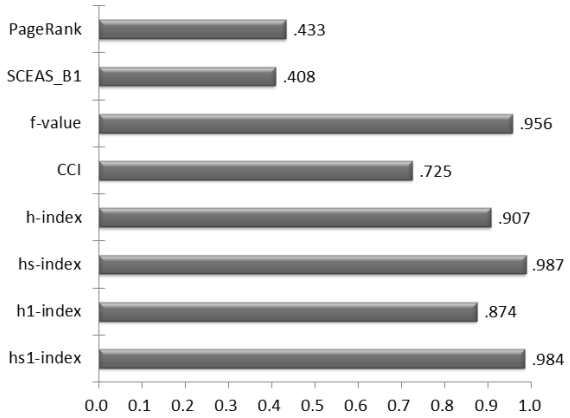
각 인용지수의 특성을 살펴보기 위해서 직접 인용빈도, 2세대 인용빈도, 출판년도, 8개 인용지수 측정값의 11개 변수 사이의 스피어맨 순위상관계수를 측정하였다. 상관관계 분석은 41,510개 문헌 중에서 인용빈도 3회 이상이면서 2001년 이후에 발표된 문헌 3,798건을 대상으로 하였다.

직접 인용빈도와 각 지수 사이의 상관관계는 <그림 1>과 같이 나타났다. 대체로 h-index 계열 지수가 다소 낮게 나타났지만 CCI를 제외하면 큰 차이는 없었다. CCI는 직접 인용빈도와의 상관도가 매우 큰 것으로 나타났다.

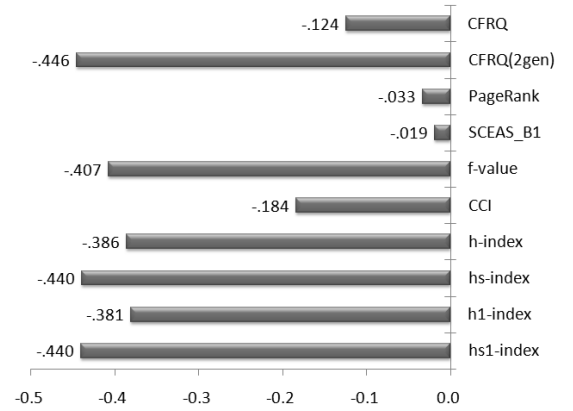
2세대 인용빈도와 각 지수 사이의 상관관계는 <그림 2>와 같이 나타났다. 여기서는 PageRank와 SCEAS_B1이 상대적으로 낮게 나타났으며 h-index 계열 지수들과 f-value



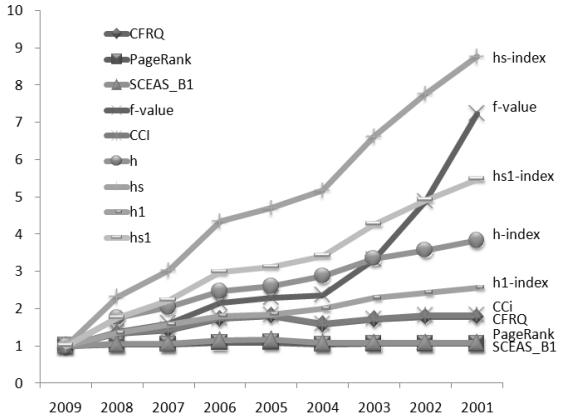
<그림 1> 직접 인용빈도와의 순위상관관계
(인용빈도 3회 이상인 2001년 이후 논문 1,790건)



<그림 2> 2세대 인용빈도와의 순위상관관계
(인용빈도 3회 이상인 2001년 이후 논문 1,790건)



<그림 3> 출판년도와의 순위상관관계 (인용빈도 3회 이상인 2001년 이후 논문 1,790건)



<그림 4> 2009년 발표논문의 평균값 대비 각 연도별 발표논문의 지수값 평균 비교

는 매우 높게 나타났다.

출판년도와 각 변수 사이의 상관관계는 <그림 3>과 같이 나타났다. 모두가 음의 상관관계를 보인 것은 출판년도가 최근일수록 낮은 값을 가지는 것을 의미한다. 특히 2세대 인용빈도가 가장 강한 음의 상관관계를 보였는데, 이는 특정 문헌을 인용한 문헌이 다시 인용되기 위해서는 어느 정도 이상의 시간이 경과되어야 하기 때문이다. 인용지수 중에서 PageRank와 SCEAS_B1은 출판년도와의 순위 상관 정도가 0에 가깝게 나타났다. 반면에 f-value와 h-index 계열 인용지수는 -0.4 내외로 다소 강한 음의 상관관계를 보이므로 오래된 문헌일수록 지수값이 크게 측정될 가능성이 높음을 알 수 있다.

<그림 4>는 2009년 발표논문의 지수별 평균값 대비 매년 발표된 논문의 평균값을 비교한 것이다. 전역 네트워크 분석 방식의 지수 중에서 PageRank와 SCEAS_B1은 연도별 평균값이 거의 변화가 없고 CCI도 약 2배 이내까지만 증가하였다. 반면에 f-value와 지역네트워크 지수인 h-index 계열 지수들은 출판년도가 오래될수록 평균값도 크게 높은 것으로 나타났다. 특히 f-value는 2003년 이전으로 거슬러 올라가면 평균값이 급속하게 증가하였다. 전역 네트워크 분석 방식의 지수 중

에서 f-value만 오래된 논문에 지나치게 유리한 결과가 나타나는 이유는 전역 네트워크 분석 방식이어서 인용 영향력이 여러 단계를 건너서까지 전달됨에도 불구하고 한 문헌의 영향력을 참고문헌 수로 나누지 않고 통째로 인용하는 논문으로 전달하므로 누적 효과가 크게 반영되기 때문이다.

이와 같이 오래된 문헌에 유리한 인용지수는 출판후 경과된 햇수로 나누는 방식을 검토해볼 수 있다. 연구자의 성과를 평가하기 위해서 h-index를 제안한 Hirsch(2005)도 h-index 값을 경과한 햇수로 나눈 m 지수를 제안한 바 있다. 이처럼 단일문헌 h-index도 경과한 햇수로 나눈 m 지수를 산출해볼 수 있을 것이다. 전산언어학 분야 논문집합을 대상으로 분석한 Radev 등(2009)의 연구에서는 PageRank도 오래된 논문에 다소 유리한 결과가 나타난다고 판단하고 PageRank 값을 출판후 경과된 햇수로 나눈 PPY(PageRank Per Year) 값을 산출하였다. 그 결과 PPY는 대체로 최근 논문에 유리한 것으로 보고되었는데, 이는 PageRank가 경과된 햇수로 나눌 만큼 충분히 오래된 논문에 유리한 값이 산출되는 지수는 아님을 의미한다.

4. 새로운 인용지수 제안

앞에서 살펴본 각 인용지수 공식과 특성을 감안하여 문헌의 인용영향력을 계산하는 공식에 필요한 요건을 정리해보면 다음과 같다.

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, 전역 네트워크 정보를 이용할 경우에는 인용하는 문헌의 중요도가 일정한 비율로 감쇠되면서 전달되어야 한다.

이런 점을 고려하여 전역 네트워크 분석 방식에 해당하는 새로운 지수인 c-index 공식을 다음과 같이 제안한다.

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

이 공식에서 1은 인용빈도가 0인 논문도 인용하는 논문으로 영향력을 전달하기 위해서 설정한 기본값이다. d는 인용 단계를 건너 전파되는 값을 일정 비율로 감소시키는 감쇠지수로서 PageRank와 같이 0.85로 설정해볼 수 있다. 여기까지는 f-value 공식과 유사하나, 전달되는 영향력을 참고문헌의 수를 고려하여 나누는 점이 다르다. 다만 PageRank 등과 같이 참고문헌의 수로 나누지 않고 제곱근을 취한 값으로 나누는 이유는, 인용영향력을 분산하여 전달하는 문제를 일종의 문헌길이정규화 문제로 간주했기 때문이다. 정보검색에서 각 문헌에 출현한 단어의 가중치를 산출할 때 단어의 출현빈도를 그대로 사용하지 않고 문헌의 길이를 고려하여 정규화한다. 이와 유사하게 인용하는 문헌으로부터 인용되는 문헌으로 전달되는 인용 영향력을 결정할 때 인용하는 문헌의 길이를 고려하여 전달되는 인용영향력을 정규화하였다. 이때 인용하는 문헌의 길이는 참고문헌의 수에 따라 결정되도록 하되, 문헌길이정규화 공식 중에서 가장 단순하면서 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. 이와 같이 인용건수가 아닌 인용건수의 제곱근으로 인용영향력을 나누게 되면 문헌 길이, 즉 참고문헌 수에 따라 전달되는 영향력이 지나치게 좌우되는 문제를 해소할 수 있다.

한편 전역 네트워크 정보를 입수하기가 불가능한 경우에 적용할 수 있는 지역 네트워크

분석 방식의 새로운 지수인 cl-index를 다음과 같이 제안한다.

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

지역 네트워크 분석에서는 여러 단계를 거치지 않으므로 감쇠지수는 필요없다. 인용을 보내는 논문의 영향력은 인용빈도로 산출한다. 인용빈도에 1을 더한 것은 인용을 받지 못한 논문으로부터의 인용도 영향력을 전달하게끔 하기 위해서이다. 또한 인용은 빈익빈부익부 현상을 보이기 때문에 인용빈도의 차이를 그대로 영향력의 차이로 반영하는 것은 바람직하지 않다. 따라서 인용빈도에 1을 더한 값의 제곱근을 한 논문으로부터 전달되는 총 영향력으로 설정하였다.

제안한 인용지수 2종을 계량정보학 분야 문헌집합에 적용해본 결과 c-index로 측정한 인용영향력 1위와 2위인 문헌은 각각 직접 인용빈도 1위인 문헌과 2세대 인용빈도 1위인 문헌으로 나타났다. 앞의 <표 3>을 보면 기존 지수는 대부분 두 문헌 중 어느 한 문헌만을 10위 이내에 포함하고 있는 것과는 차별화된 결과이다. 이는 제안한 지수가 직접 인용빈도나 2세대 인용빈도가 매우 두드러진 문헌을 모두 적절하게 고려하고 있음을 의미한다. cl-index로 측정한 경우에도 두 문헌은 각각 1위와 4위로 측정되었다.

4. 결론

제안한 두 지수는 기존 인용지수에 비해서 비교적 단순하면서도 직접 인용빈도와 2세대 인용빈도를 더 적절하게 반영하는 것으로 추정된다. 전역 네트워크 정보를 획득할 수 있는 경우에는 c-index를, 그렇지 못한 경우에는 cl-index를 적용할 수 있다. 향후 다양한 문헌집합을 대상으로 지수의 특성을 검증하는 연구가 필요하다.

참고문헌

이재윤. 2011. 인용 네트워크 분석에 근거한 문헌 인용 지수 연구. 『한국문헌정보학회지』, 45(2): 119-143.

Bi, H. H., J. Wang, and D. K. J. Lin. 2011. "Comprehensive citation index for research networks." *IEEE Transactions on Knowledge and Data Engineering*, 23(8): 1274-1278.

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

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

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.

Lee, Jae Yun and Sanghee Choi. 2011. "Intellectual structure and infrastructure of informetrics: Domain analysis from 2001 to 2010." *Journal of the Korean Society for Information Management*, 28(2): 11-36.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. "The PageRank citation ranking: Bringing order to the Web." Technical Report, Stanford InfoLab. [cited 2011.4.1]. <http://ilpubs.stanford.edu:8090/422/>.

Radev, D. R., M. T. Joseph, B. Gibson, and P. Muthukrishnan. 2009. "A bibliometric and network analysis of the field of computational linguistics." *Journal of the American Society for Information Science and Technology*, submitted. <clair.si.umich.edu/~radev/papers/133.pdf>.

Schubert, A. 2009. “Using the h-index for assessing single publications.” *Scientometrics*, 78(3): 559–565.

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