논문 인용의 영향요인 분석

Analysis of Factors Influencing Journal Articles' Citations

유재복_한국원자력연구원 책임연구원
김재호_(주)링크소프트 개발이사

초 록

최근 논문의 가치평가가 크게 강조되고 있으며, 그 평가의 수단으로 논문의 피인용횟수가 매우 유용한 척도 중의 하나로 받아들여지고 있다. 그에 따라 이 연구에서는 논문의 피인용횟수와 이에 영향을 미칠만한 형태적 및 개념적 요인의 11개 변수들 간의 상관관계를 문헌정보학분야 11종의 학술지 논문을 대상으로 분석하였다. 분석결과, 논문의 피인용횟수와 일정 수준 이상의 상관관계 즉 5% 이상의 설명력을 갖는 변수는 문헌간유사도 뿐인 것으로 나타났으며, 문헌간유사도가 높아질수록 논문 상호 간에 피인용횟수에 대한 상관관계도 증가하는 것으로 나타났다

ABSTRACT

Recently, the valuation of research papers has been greatly emphasized, and their citation has been accepted as a very useful indicator. In this study, we performed correlation analyses between the paper citation counts and 11 explanatory variables of morphological and conceptual factors with a test dataset of the papers of 11 journals in library and information science. The analysis results of the correlations show that only the document similarity has 5% or more standardized variances(r2) with paper citation counts and the document similarity with citation counts get higher as the variable value increases.

키워드: 논문인용, 인용분석, 논문 피인용횟수, 논문인용 영향요인
research paper citations, citation analysis, paper citation counts, factors influencing paper citation

1. 서론

1.1 연구의 목적

학술논문은 연구자 자신은 물론 소속기관 및 국가의 경쟁력이자 성과평가에 있어서도 매우 중요한 요소 중의 하나이다. 그에 따라 연구역량을 측정하는 수단으로 지금까지는 주로 논문건수를 기반으로 한 정량적인 척도가 사용되어 왔는데, 최근 들어 논문의 가치가 크게 강조되고 있으며 논문의 피인용횟수가 그의 매우 유용한 척도 중의 하나로 받아들여지고 있다(Fu and Aliferis, 2010). 즉 자주 인용된 논문은 그 만큼 질적 측면에서 보다 우수한 가치를 지닌 것으로 나타난 것이다(Levitt and Thelwall, 2007).

그러나 논문의 인용은 발표된 후 일정 기간이 경과되어야만 활발하게 이루어지는 특성으로 인해 최근에 발표된 논문의 피인용횟수를 제대로 파악할 수 없다는 문제점이 있다(유재복, 정영미 2010). 즉 논문이 다른 논문으로부터 인용을 받기 위해서는 일정 기간이 경과되어야 하는데, Yi 등(Yi, Ao, and Ho 2008)에 따르면 논문의 인용빈도가 최고점에 이르는 시기는 논문이 학술지에 발표된 후 4년째가 되는 시점이라고 발표하였다. 결국 논문의 가치를 측정하는 중요한 척도인 피인용횟수는 이처럼 시간상의 제약으로 인해 즉각적인 활용이 어렵고, 특히 최근에 발표된 논문의 경우 사실상 피인용횟수를 정확하게 파악하는 것은 거의 불가능하다고 할 수 있을 것이다.

그에 따라 최근에 발표된 새로운 논문에 대한 피인용횟수를 예측할 수 있는 예측모형에 관한 연구가 활발하게 진행되고 있다. 아울러 예측모형 개발의 일환으로 논문의 인용에 영향을 미치는 여러 가지 요인들에 대한 연구 또한 활발하게 이루어지고 있는데, 최적의 인용 예측모형을 개발하기 위해서는 보다 다양한 변수들에 대한 분석이 여전히 필요한 실정이다.

이에 이 연구에서는 최적의 논문인용 예측모형을 개발하기 위한 기초적 자료를 제공하는 데 의의를 두고, 세계적인 인용색인 웹 데이터베이스인 SCOPUS에 등재된 문헌정보학분야 학술지의 논문을 토대로 논문의 인용에 영향을 미칠만한 제반 변수들을 크게 형태적 요인과 개념적 요인으로 나누어 종합적으로 분석하였다. 그 결과 어떠한 변수들이 논문의 피인용횟수와 어느 정도의 상관관계가 있고 얼마만큼의 설명력, 즉 영향력이 있는지를 분석하였다. 이 연구에서의 분석결과는 보다 정교한 논문인용 예측모형을 설계하는데 있어서 직접 활용할 수 있을 것으로 기대된다.

1.2 연구의 방법 및 범위

이 연구에서는 논문의 인용에 어떠한 요인들이 얼마만큼의 영향을 미치는가를 종합적으로 살펴보고자 세계적인 인용색인 웹 데이터베이스인 SCOPUS에 등재된 문헌정보학분야 학술지 중 1990년 이후의 논문을 제공하는 11종의 학술지의 논문을 대상으로 분석하였다.

이 연구에서는 논문의 피인용횟수에 영향을 미칠만한 형태적 요인 및 개념적 요인의 11개의 변수를 대상으로 연구가설을 설정하였다. 연구가설을 검증하기 위해 이 논문에서는 SPSS for windows 12.0 version 프로그램을 이용하여 종속변수인 논문의 피인용횟수와 11개의 독립변수들을 대상으로 상관관계분석을 실시하였다.

1.3 선행연구

논문의 질적 가치가 그의 피인용횟수와 상당한 상관관계가 있다는 것이 많은 연구를 통해 밝혀짐에 따라 논문의 인용에 영향을 미치는 요인에 대한 분석과 이를 토대로 한 논문인용 예측모형 개발에 대한 연구가 활발하게 진행되고 있다. 주요 선행연구를 살펴보면 다음과 같다.

Walkers(2006)는 2003년도에 발행된 범죄심리학분야 12개 저널을 대상으로 15개의 변수를 토대로 논문의 피인용횟수와의 상관관계를 분석한 결과 저자특성(성별), 소속기관(대학-기타), 국적(미국-기타), 최근 2년간의 주저자의 논문 피인용횟수, 최근 2년간 주저자의 공저여부, 페이지 수, 리뷰논문 여부, 주제분야, 저널 영향력지수 등 9개의 변수가 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있음을 밝혀냈다. 아울러 NB(Negative binominal) 회귀분석을 실시한 결과, 이들 9개의 변수들 중에서 최근 2년간 주저자의 논문 피인용횟수, 국적, 리뷰논문 여부 등 3개의 변수가 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있음을 밝혀냈으며, 저자의 영향력이 저널의 영향력보다 중요함을 밝혀냈다.

Castillo 등(Castillo, Donato, and Gionis)는 특정 논문의 저자에 의해 발표된 직전의 최근 논문의 저자정보를 사용하여 그 논문의 피인용횟수를 예측할 수 있는 연구를 수행한 결과, 실제값과 예측값 간에는 일정 수준의 상관관계(r=0.57)가 있는 것으로 나타났다. 아울러 예측율을 높이기 위해 저자정보와 함께 그 저자의 직전의 최근 논문에 대한 몇 가지 특징들을 추가시켜 분석한 결과 실제값과 예측값 간에는 상당히 높은 수준의 상관관계(r=0.81)가 있는 것으로 나타났다.

Lokker 등(2008)은 온라인 논문 심사평가를 위해 관련분야 전문가들에게 공개된 지 3주 이내의 임상학분야 저널의 논문 데이터를 이용하여 향후 2년간의 피인용횟수를 예측할 수 있는지를 실험하였다. 20개의 독립변수에 대한 상관관계분석을 실시한 결과, 20개 변수 중 저자 수, 개요 저널(synoptic journal) 내 초록 유무, 임상학적 적합성 점수, 페이지 수, 구조화된 초록 유무, 참고문헌 수, 원저논문(original article) 여부, 학제성, 치료방법 여부, 색인 유무, 초록화된 논문비율 등 11개의 변수가 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다. 아울러 이들 11개의 변수들을 대상으로 다중회귀분석을 실시한 결과 논문의 피인용횟수를 설명할 수 있는 변량(r^2)이 0.60으로 약 60% 정도의 설명력, 즉 예측력을 갖는 것으로 나타났다.

Ibanez 등(Ibanez, Larranaga, and Bielza 2009)은 Informatics 저널의 초록정보를 토대로 확률기반 분류기인 나이브베이즈 분류기와 로지스틱 회귀분석을 사용하여 출판 후 4년 내의 논문에 대한 피인용횟수를 예측할 수 있는 모형을 설계하였다. 이 연구에서는 예측의 정확율을 높이기 위해 저널 섹션(9개 분야)별로 예측모형을 설계하였는데, 각 섹션별 평균 예측율은 89.4%에서 91.5%로 매우 높게 나타났다.

Fu와 Aliferis(2008, 2009, 2010)는 생의학분야의 6개 저널을 대상으로 SVM 분류기를 이용한 기계학습방법을 사용하여 내용기반 요소와 서지적인 특징들을 혼합 사용하여 논문 출판 후 10년간의 피인용횟수를 예측할 수 있는 장기적인 예측모형을 개발하였다. 이 연구에서는 제목, 초록, MeSH 용어, 주저자 논문 수, 주저자 논문 피인용횟수, 끝저자 논문 수, 끝저자 논문 피인용횟수, 출판유형, 저자 수, 기관 수, 저널 영향력지수, 주저자 기관의 질 등 12개의 변수를 사용하였으며, 개발된 예측모형의 논문 피인용횟수에 대한 예측율은 77~82%(AUC=0.86~0.92)로 비교적 높게 나타났다.

한편, 상기 선행연구 중에서 특히 Ibanez 등과 Fu와 Aliferis의 연구는 기계학습방법 중의 하나인 자동분류기를 사용하여 예측모형을 설계하였는데, 심경(2005)에 따르면 통제된 실험환경에서는 자동분류기의 분류정확

도가 80~90%에 이른다고 할지라도 실제환경에서는 평균 30% 정도로 매우 낮은 수치를 보인다고 주장하였다. 이로써 보면, 실험환경에서 설계된 예측모형을 실제환경에 그대로 적용시키기에는 아직 상당한 무리가 있을 것이라 여겨진다.

2. 연구설계

2.1 데이터 수집 및 전처리

이 연구에서는 세계적인 인용색인 웹 데이터베이스인 SCOPUS에 등재된 문헌정보학분야 학술지 가운데 1990년 이후의 논문을 제공하는 11종의 학술지에서 초록정보를 제공하는 논문만을 대상으로 분석하였다.

분석대상 기간은 분석대상인 11종의 문헌정보학분야 학술지에 수록된 1990~2009년까지의 논문의 인용정보를 토대로 특허 피인용반감기를 산출한 결과 피인용반감기가 8년(7.6년)으로 나타남에 따라 이를 적용시켜 1990~2001년으로 설정하였다. 여기에서의 피인용반감기는 JCR(Journal Citation Reports)에서 사용하는 아래의 피인용반감기 산출방법을 토대로 산출하였다.

$$\text{피인용반감기} = \text{누적인용률 50\% 이전 연도 수} + \frac{\text{누적인용률 50\%에서 50\% 직전 연도 누적인용율을 뺀 값}}{\text{50\% 직후 연도 누적인용률에서 50\% 직전 연도 누적인용률을 뺀 값}}$$

〈표 1〉은 문헌정보학분야 학술지의 피인용반감기를 적용한 분석대상 논문건수로 총 5,444건이다.

〈표 1〉 분석대상 학술지 및 건수

저널명	간기	전체 수록건수('90~'08)	분석대상 건수('90-'01)
Electronic Library	Bimonthly	932	308
Information Processing and Management	Bimonthly	1,269	584
Information Sciences	Semi-monthly	3,152	1,288
J. of Academic Librarianship	Bimonthly	1,068	284
J. of Classification	Semi-Annual	256	121
J. of Documentation	Bimonthly	543	237
J. of Information Science	Bimonthly	922	480
J. of Libraianship and Information Science	Quarterly	350	196
J of the American Society for Information Science and Technology	monthly	1,960	621
Library Trends	Quarterly	829	444
Scientometrics	monthly	2,108	881
계		13,389	5,444

2.2 변수 및 데이터 산출

이 연구에서 일부 변수의 데이터는 SCOPUS 사이트를 통해 다운로드한 값을 그대로 사용한 반면, 대부분 변수의 데이터는 수작업 또는 별도로 설계한 프로그램을 통해 산출하였다. 〈표 2〉는 이 연구에서 필요한 각종 변수 및 데이터에 대한 산출방식을 정리한 것이다.

〈표 2〉 각종 변수 및 데이터 산출방법

구분		변수	데이터 산출방법
종속변수		피인용횟수	다운로드 데이터를 토대로, 피인용반감기를 적용하여 산출
독립변수	형태적 요인	저자 수*	다운로드 데이터를 토대로 산출
		저자 소속	다운로드 데이터를 토대로 산출 [1=대학, 0=기타]
		저자 국적	다운로드 데이터를 토대로 대륙별로 산출 [1=북미, 2=남미, 3=유럽, 4=아시아, 5=아프리카, 6=오세아니아]
		페이지 수	다운로드 데이터를 토대로 산출
		문헌 유형	다운로드 데이터를 토대로 산출 [1=article, 2=conference paper, 3=review, 4=short survey, 5=editorial]
		저널논문 수	저널별 수록논문 수를 산출
		저널 발행주기	저널의 발행주기를 그대로 사용 [1=semi-monthly, 2=monthly, 3=bimonthly, 4=quarterly, 5=semi-annual]
		SJR 지수	저널의 SJR Indicator를 그대로 사용
	개념적 요인	참고문헌 수	각 논문별로 수작업을 통해 일일이 확인하여 산출
		참고문헌 평균 피인용횟수	각 논문별로 수작업을 통해 해당 참고문헌의 피인용횟수를 계산한 후, 이들 합을 참고문헌 수로 나누어 산출
		문헌간유사도	별도로 설계한 프로그램을 통해, 논문의 제목과 초록에 출현하는 용어를 대상으로 코사인 유사계수를 이용하여 산출

주) SJR(SCImago Journal Rank Indicator) 지수는 SCImago 연구그룹(스페인 소재)에서 제공하는 자료로, 가중치를 부여한 논문의 인용값을 기반으로 한 저널의 영향력(prestige 지표임

참고로, 여기에서 개념적 요인 중 문헌간유사도의 경우 〈표 2〉의 데이터 산출방법에 따라 개념적으로 서로 관계가 있는 2개 특허 간의 결합쌍 단위로 산출된 문헌간유사도를 토대로, 분석대상 특허와 쌍결합된 대응특허의 피인용횟수를 독립변수의 변수값으로 사용하였다.

2.3 연구가설

이 연구의 목적은 최적의 논문인용 예측모형을 개발하기 위한 기초적 자료를 제공하고자 논문의 피인용횟수에 어떠한 변수들이 얼마만큼의 상관관계를 갖고 있는지를 종합적으로 분석하는 것이다. 여기에서는 논문의 피인용횟수를 종속변수로 하고, 종속변수에 영향을 미칠 수 있는 제반 독립변수를 형태적 요인과 개념적 요인의 2가지 측면으로 나누어 분석하였다.

형태적 요인은 저자 수, 저자 소속, 저자 국적, 페이지 수, 문헌유형, 저널논문 수, 저널 발행주기, SJR 지수 등 8가지 변수를 선정하였고, 개념적 요인은 참고문헌 수, 참고문헌 평균 피인용횟수, 문헌간유사도 등 3가지

변수를 선정하였다.

이들 각각의 독립변수가 종속변수인 논문의 피인용횟수에 얼마간의 상관관계가 있는지를 검증하기 위해서 〈표 3〉과 같이 11가지의 연구가설을 설정하였다. 참고로, 독립변수 중 더미변수인 저자 소속, 저자 국적, 저널 발행주기, 및 문헌유형을 제외한 나머지 모든 독립변수와 종속변수인 피인용횟수는 사회과학분야에서 가장 보편적으로 사용되는 순위척도인 리커트 7점 척도를 사용하여 실제값을 7개 구간으로 나누어 분석하였다. 각 구간의 설정은 각 변수별로 해당 변수값의 분포비율과 변수의 특성을 충분히 고려하여 임의로 조정하였다.

〈표 3〉 연구가설

가 설		내 용
형태적 요인	H1	저자 수는 논문의 피인용횟수와 관련이 있을 것이다.
	H2	저자 소속은 논문의 피인용횟수와 관련이 있을 것이다.
	H3	저자 국적은 논문의 피인용횟수와 관련이 있을 것이다.
	H4	페이지 수는 논문의 피인용횟수와 관련이 있을 것이다.
	H5	문헌유형은 논문의 피인용횟수와 관련이 있을 것이다.
	H6	저널논문 수는 논문의 피인용횟수와 관련이 있을 것이다.
	H7	저널발행주기는 논문의 피인용횟수와 관련이 있을 것이다.
	H8	SJR 지수는 논문의 피인용횟수와 관련이 있을 것이다.
개념적 요인	H15	참고문헌 수는 논문의 피인용횟수와 관련이 있을 것이다.
	H16	참고문헌 평균 피인용횟수는 논문의 피인용횟수와 관련이 있을 것이다.
	H17	문헌간유사도는 논문의 피인용횟수와 관련이 있을 것이다.

3. 가설검증 및 분석

논문을 인용함에 있어서 어떠한 독립변수들이 종속변수에 얼마만큼의 상관관계가 있는지를 종합적으로 살펴보기 위해서 문헌정보학분야 11종의 학술지에 수록된 총 5,444건의 논문을 분석대상으로 분석하였다. 이 연구에서는 연구가설을 검증하기 위해 SPSS for window 12.0을 사용하여 상관관계분석을 실시하였다.

가설검증에 앞서 이 논문에서의 유일한 종속변수인 논문의 피인용횟수에 대한 분포현황을 살펴보면 〈표 4〉와 같다. 여기에서의 피인용횟수는 문헌정보학분야 학술지 논문의 피인용반감기를 적용시켜 산출한 것이며, 실제값을 7개 구간으로 나눈 리커트 7점 척도를 사용하였다. 〈표 4〉를 토대로 살펴보면, 논문의 피인용횟수는 0회, 1~2회, 3~5회가 각각 20% 이상의 높은 점유율을 보이는 반면 21~30회, 31회 이상은 각각 3% 대의 낮은 점유율을 보인 것으로 나타났다.

〈표 4〉 피인용횟수 분포

	구간	0	1~2	3~5	6~10	11~20	21~30	31이상	계
피인용 횟수	빈도(건)	1,233	1,438	1,114	783	529	168	179	5,444
	비율(%)	22.6	26.5	20.4	14.4	9.7	3.1	3.3	100

3.1 형태적 요인

논문의 피인용횟수와 형태적 요인의 8가지 변수들 간의 상관관계를 분석하기에 앞서 각 변수들의 분포현황을 살펴보면 〈표 5〉와 같다.

〈표 5〉를 통해 살펴보면, 저자 수는 1명 또는 2명인 경우가 압도적으로 많고, 저자 소속은 대학이 매우 높으

〈표 5〉 형태적 요인의 제반변수 분포

	구간	1	2	3	4	5	6	7이상	계
저자 수	빈도(건)	2,708	1,681	710	242	70	16	17	5,444
	비율(%)	49.7	30.9	13.1	4.4	1.3	0.3	0.3	100
저자 소속	구간	0	1						계
	빈도(건)	1,341	4,103						5,444
	비율(%)	24.6	75.4						100
저자 국적	구간	1	2	3	4	5	6	기타	계
	빈도(건)	2,344	92	1,887	762	124	146	89	5,444
	비율(%)	43.05	1.69	34.67	14.00	2.28	2.68	1.63	100
페이지 수	구간	1~5	6~10	11~15	16~20	21~25	26~30	31이상	계
	빈도(건)	446	1,332	1,498	1,038	621	278	231	5,444
	비율(%)	8.2	24.5	27.5	19.1	11.4	5.1	4.2	100
문헌유형	구간	1	2	3	4	5			계
	빈도(건)	5,187	59	196	1	1			5,444
	비율(%)	95.3	1.1	3.6	0.0	0.0			100
저널논문 수	구간	200미만	200-299	300~399	400~499	500~599	600~699	700이상	계
	빈도(건)	317	521	308	924	584	621	2,169	5,444
	비율(%)	5.82	9.57	5.66	16.97	10.73	11.41	39.84	100
저널 발행주기	구간	1	2	3	4	5			계
	빈도(건)	1,288	1,502	1,893	640	121			5,444
	비율(%)	23.7	27.6	34.7	11.8	2.2			100
SJR 지수	구간	0.036	0.042~0.043	0.051	0.070~0.071	0.083	0.087	0.099	계
	빈도(건)	444	504	521	1,064	121	1,909	881	5,444
	비율(%)	8.2	9.3	9.6	19.4	2.2	35.1	16.2	100

주) '저자 국적'의 구간에서 '기타'는 국적 미상으로 분석대상에서 제외함

며, 저자 국적은 북미와 유럽의 점유율이 매우 높은 것으로 나타났다. 페이지 수의 경우 6~10, 11~15, 16~20 페이지가 각각 20% 내외로 높게 나타났고, 문헌유형은 논문(article)이 압도적으로 높게 나타났으며, 저널 논문 수는 700편 이상이 상대적으로 높게 나타났다. 또한 저널 발행주기는 격월간이 상대적으로 높게 나타났으며, SJR 지수는 0.087인 경우가 가장 높은 것으로 나타났다.

논문의 피인용횟수와 형태적 요인의 8가지 변수들 간의 상관관계를 분석한 결과는 〈표 6〉과 같다.

〈표 6〉 피인용횟수와 형태적 요인 변수 간의 상관분석 결과

	상관계수(r)	유의확률(p)	변량(r2)
저자 수	.103	.000	.011
저자 소속	.062	.000	.004
저자 국적	-.068	.000	.005
페이지 수	.135	.000	.018
문헌유형	.046	.001	.001
저널논문 수	-.048	.000	.001
저널 발행주기	-.020	.137	.001
SJR 지수	.157	.000	.025

첫째, 저자 수는 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.103, p<.05). 따라서 가설-1은 채택되었으며, 저자 수가 논문의 피인용횟수를 설명할 수 있는 변량(r2)은 .011로 약 1.1% 정도의 설명력을 갖는 것으로 나타났다.

둘째, 저자 소속은 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.062, p<.05). 따라서 가설-2는 채택되었으며, 저자 소속이 논문의 피인용횟수를 설명할 수 있는 변량은 .004로 약 0.4% 정도의 설명력을 갖는 것으로 나타났다.

셋째, 저자 국적은 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=-.068, p<.05). 따라서 가설-3은 채택되었으며, 저자 국적이 논문의 피인용횟수를 설명할 수 있는 변량은 .005로 약 0.5% 정도의 설명력을 갖는 것으로 나타났다.

넷째, 페이지 수는 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.135, p<.05). 따라서 가설-4는 채택되었으며, 페이지 수가 논문의 피인용횟수를 설명할 수 있는 변량은 .018로 약 1.8% 정도의 설명력을 갖는 것으로 나타났다.

다섯째, 문헌유형은 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.046, p<.05). 따라서 가설-5는 채택되었으며, 저자 수가 논문의 피인용횟수를 설명할 수 있는 변량은 .001로 약 0.1% 정도의 설명력을 갖는 것으로 나타났다.

여섯째, 저널논문 수는 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=-.048, p<.05). 따라서 가설-6은 채택되었으며, 저널논문 수가 논문의 피인용횟수를 설명할 수 있는 변량은 .001로 약 0.1% 정도의 설명력을 갖는 것으로 나타났다.

일곱째, 저널 발행주기는 논문의 피인용횟수와 통계적으로 유의미하지 않은 상관관계가 있는 것으로 나타났다(r=-.020, p<.05). 따라서 가설-7은 기각되었다.

여덟째, SJR 지수는 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.157, p<.05). 따라서 가설-8은 채택되었으며, SJR 지수가 논문의 피인용횟수를 설명할 수 있는 변량은 .025로 약 2.5% 정도의 설명력을 갖는 것으로 나타났다.

3.2 개념적 요인

논문의 피인용횟수와 개념적 요인의 3가지 변수들 간의 상관관계를 분석하기에 앞서 각 변수들의 분포현황을 살펴보면 〈표 7〉과 같다. 〈표 7〉을 통해 살펴보면, 참고문헌 수는 11~20, 참고문헌 평균 피인용횟수는 100~500, 그리고 문헌간유사도는 0.4 이상인 경우가 상대적으로 높은 것으로 나타났다.

〈표 7〉 개념적 요인의 제반변수 분포

참고문헌 수	구간	1	2	3	4	5	6	70|상	계
	빈도(건)	2,708	1,681	710	242	70	16	17	5,444
	비율(%)	49.7	30.9	13.1	4.4	1.3	0.3	0.3	100
참고문헌 평균 피인용횟수	구간	0	1						계
	빈도(건)	1,341	4,103						5,444
	비율(%)	24.6	75.4						100
문헌간유사도	구간	1	2	3	4	5	6	기타	계
	빈도(건)	2,344	92	1,887	762	124	146	89	5,444
	비율(%)	43.05	1.69	34.67	14.00	2.28	2.68	1.63	100

논문의 피인용횟수와 개념적 요인의 3가지 변수들 간의 상관관계를 분석한 결과는 〈표 8〉과 같다.

〈표 8〉 피인용횟수와 개념적 요인 변수 간의 상관분석 결과

		상관계수(r)	유의확률(p)	변량(r2)
참고문헌 수		.173	.000	.030
참고문헌 평균 피인용횟수		.213	.000	.045
문헌간유사도 (쌍)	0.3 이상	.264	.000	.070
	0.4 이상	.372	.000	.139
	0.5 이상	.503	.000	.253
	0.6 이상	.617	.000	.381
	0.7 이상	.719	.000	.517
	0.8 이상	.837	.000	.700
	0.9 이상	.914	.000	.836

첫째, 참고문헌 수는 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.173, p<.05). 따라서 가설-9는 채택되었으며, 참고문헌 수가 논문의 피인용횟수를 설명할 수 있는 변량(r2)은 .030으로 약 3.0% 정도의 설명력을 갖는 것으로 나타났다.

둘째, 참고문헌 평균 피인용횟수는 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.213, p<.05). 따라서 가설-10는 채택되었으며, 참고문헌 평균 피인용횟수가 논문의 피인용횟수를 설명할 수 있는 변량은 .045으로 약 4.5% 정도의 설명력을 갖는 것으로 나타났다.

셋째, 문헌간유사도는 각 논문쌍 간의 문헌간유사도 값에 따라 상관관계를 분석하였는데, 모든 구간에 걸쳐 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 나타났다(r=.264~.914, p<.05). 따라서 가설-11는 채택되었으며, 문헌간유사도가 논문의 피인용횟수를 설명할 수 있는 변량은 .070~.836으로 약 7.0~83.6% 정도의 설명력을 갖는 것으로 나타났다.

이상에서의 상관관계분석 결과를 종합 정리하면, 분석대상인 11개의 변수 가운데 저널 발행주기를 제외한 11개의 변수가 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있는 것으로 밝혀졌다. 그러나 논문의 피인용횟수를 설명할 수 있는 변량을 살펴보면, 피인용횟수와 일정 수준 이상의 상관관계, 즉 변량이 .05 이상으로 5% 이상의 설명력이 있는 변수는 문헌간유사도 밖에 없는 것으로 나타났다.

4. 결론

이 연구에서는 논문의 인용에 영향을 미칠 수 있는 제반 요인들을 종합적으로 살펴보고자 문헌정보학분야 11종의 학술지 논문을 대상으로 각종 변수들을 형태적 요인과 개념적 요인의 두 가지 측면으로 나누어 분석하였다. 그 결과 어떠한 변수들이 논문의 피인용횟수와 상관관계가 있는지를 검토하였다. 논문의 피인용횟수에 어떠한 변수들이 얼마만큼의 관련이 있는지를 확인하고자 실시한 상관관계분석 결과는 다음과 같다.

첫째, 형태적 요인의 경우 논문 발행주기를 제외한 7개의 변수들이 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있지만, 설명변량은 매우 낮은 것으로 밝혀졌다.

둘째, 개념적 요인의 경우 3개의 변수 모두 논문의 피인용횟수와 통계적으로 유의미한 상관관계가 있으며, 설명변량은 참고문헌 수와 참고문헌 평균 피인용횟수는 매우 낮은 것으로 나타난 반면, 문헌간유사도는 논문쌍 간의 문헌간유사도값이 높아질수록 점차 높아지는 것으로 밝혀졌다.

이 연구는 논문의 인용에 영향을 미칠만한 형태적 및 개념적 측면의 다양한 변수들을 종합적으로 분석하였다는 데 그 의미가 있다. 이 연구에서의 분석결과를 토대로 향후 논문인용 예측모형을 설계함에 있어서 시사하는 바를 정리하면 다음과 같다.

첫째, 논문의 인용에 가장 영향을 미치는 변수가 문헌간유사도인 만큼, 문헌간유사도를 토대로 논문의 피인용횟수에 대한 예측력을 높일 수 있는 보다 다양하고 심도있는 연구가 필요할 것으로 보인다. 즉 각 논문에 부여된 색인어를 활용하거나 제목과 색인어에 가중치를 부여하는 방법도 하나의 고려대상이 될 수 있을 것이며, 실험적인 선행연구에서 사용된 자동분류 결과와 문헌간유사도를 병행하여 연구하는 것도 적극 검토할 필요가 있을 것으로 보인다.

둘째, 문헌간유사도의 경우 해당 변수 값에 따라 상관관계는 물론 예측모형의 결정계수가 크게 달라질 수 있으므로 이러한 점을 고려하여 예측모형을 설계하도록 해야 할 것이다. 즉 문헌간유사도의 변수 값이 너무 낮을 경우에는 상관관계와 예측율이 크게 낮아지는 반면 너무 높을 경우에는 분석대상 논문 건수가 크게 감소될 수 있음에 유의해야 할 것이다.

셋째, 논문의 피인용횟수를 보다 정확하게 예측할 수 있는 모형을 개발하기 위해서는 이 논문에서는 물론 선행연구에서 분석한 변수 외에도 보다 다양한 변수들을 지속적으로 발굴하여 분석할 필요가 있을 것이다. 저자의 지리적 위치, 동일 저자의 논문 클러스터 형성여부, R&D 흐름도 등도 그 중 하나의 예가 될 수 있을 것이다.

넷째, 예측모형을 설계할 때 각 주제분야에 따라 각종 변수의 변수 값에 대한 평균과 상관관계의 차이는 물론 예측모형에서 유의한 변수가 다른 경우가 있을 수 있음을 고려하여, 모든 주제분야를 아우르는 통합적인 예측모형보다는 각 주제분야별로 예측모형을 설계하는 것을 고려할 필요가 있을 것이다.

▣ 참 고 문 헌 ▣

- 유재복, 정영미. 2010. 논문 인용에 영향을 미치는 요인 분석. 『정보관리학회지』, 27(1): 103-118.
- 심경. 2005. 학습문헌집합의 속성에 따른 문헌 범주화 성능 실험. 박사학위논문, 연세대학교 문헌정보학과.
- Castillo, Carlos, Debora Donato, and Aristides Gionis, 2007. "Estimatiing number of citations using author reputation." Lecture Notes in Computer Science, 4726: 107-117.
- Fu, Lawrence D. and Constantin F. Aliferis. 2008. Models for predicting and explaining citation count of biomedical articles. 『AMIA 2008 Symposium Proceedings』. 223-226.
- Fu, Lawrence D. and Constantin F. Aliferis. 2009. Method for predicting citation counts. US patent, US 2009/0157585 A1(2009.6.18).
- Fu, Lawrence D. and Constantin F. Aliferis. 2010. "Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature." Scientometrics [online], [cited 2010.7.19]. 〈www.springerlink.com/content/hg01x37463010180/fulltext.pdf〉
- Ibanez, Alfonso, Pedro Larranaga, and Concha Bielza. 2009. "Predicting citation count of Bioinformatics papers within four years of publication." Bioinformatics, 25(24): 3303-3309.
- Levitt, Jonathan M. and Mike Thelwall. 2008. "Patterns of annual citation highly cited articles and the prediction of their citation ranking: A comparison across subjects." Scientometrics, 77(1): 41-60.
- Lokker, Cynthia, K Ann McKibbon, R James McKinlay, Nancy L Wilczynski, and R Brian Haynes. 2008. "Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study." British Medical Journal, 336(7645): 655-657.
- Walters, Glenn D. 2006. "Predicting subsequent citations to article published in twelve crime-psychology journals: author impact versus journal impact." Scientometrics, 69(3): 499-510.
- Yi, Huang, Xianolan Ao, and Yuh-Shan Ho. 2008. "Use of citation per publication as an indicator to evaluate pentachlorophenol research." Scientometrics, 75(1): 67-80.