인문학 및 사회과학 분야 국내 학술논문의 저자키워드 출현빈도와 피인용횟수의 상관관계 연구*

A Study on the Correlation between the Appearance Frequency of Author Keyword and the Number of Citation in the Humanities and Social Science Journal Articles of the Korea Citation Index (KCI)

고영만 (Young Man Ko)**, 송민선 (Min-Sun Song)***,
김비연 (Bee-Yeon Kim)****, 민혜령 (Hye-Ryoung Min)*****

초 록

본 연구의 목적은 저자키워드의 출현빈도와 해당 키워드가 속한 논문들의 총피인용횟수 간 상관관계 여부를 확인하고자 하는 것이다. 연구의 배경은 인문사회과학 분야 학술용어사전을 구축하는데 있어서 실제 연구에서의 활용도가 높고 다른 키워드와의 의미적 연관관계가 많은 학술용어를 추출하기 위한 방법론을 개발해 보고자 하는 것이다. 본 연구의 목적을 이루기 위해 한국연구재단 한국학술지인용색인(KCI)에 수록된 2007년에서 2011년까지의 인문학 및 사회과학 분야 학술지 논문의 저자키워드와 피인용횟수를 분석하였다. 분석 결과 저자키워드의 출현빈도와 해당 키워드가 속한 논문들의 총피인용횟수는 통계적으로 상관관계가 있으며, 저자키워드의 출현빈도가 늘어날수록 논문의 총피인용횟수도 많아지는 것으로 나타났다.

ABSTRACT

The purpose of this study is to verify the correlation between the appearance frequency of author keyword and the number of citation in journal articles. In this study, we were trying to develop a methodology that can select the term having semantic relation with other terms and higher utilization to build a structured scientific glossary. In order to achieve this purpose, we analyzed the number of citation and the author keyword of the humanities and social science journal articles of the Korea Citation Index (KCI) from 2007 to 2011. This study found a correlation between appearance frequency of author keyword and the number of citation of the journal articles, with higher appearance frequency of author keyword of the journal articles being more cited.

키워드: 학술용어사전, 저자키워드, 출현빈도, 피인용횟수, 한국학술지인용색인
scientific glossary, author keyword, appearance frequency, number of citation, Korea Citation Index

---

* 본 연구는 2012년 한국연구재단의 토대연구 지원을 받아 수행 중인 「인문사회 분야 연구성과물의 지식지도 형성을 위한 구조적 용어사전 지식베이스 구축」 연구 과제의 일환으로 진행된 연구임.
** 성균관대학교 문과대학 문헌정보학과 교수(ymko@skku.edu) (제1저자)
*** 성균관대학교 정보관리연구소 연구원(songsery@skku.edu) (공동저자)
**** 성균관대학교 정보관리연구소 연구원(korkby@skku.edu) (공동저자)
***** 성균관대학교 정보관리연구소 연구원(minipalm@skku.edu) (공동저자)
■ 논문접수일자 : 2013년 5월 21일 ■ 최초심사일자 : 2013년 6월 4일 ■ 게재확정일자 : 2013년 6월 11일
■ 정보관리학회지, 30(2), 227-243, 2013. 〔http://dx.doi.org/10.3743/KOSIM.2013.30.2.227〕

1. 서 론

1.1 연구의 목적

본 연구의 배경은 인문학 및 사회과학 분야의 구조적 학술용어사전을 구축하는데 있어서 연구에서의 활용도가 높으며 다른 용어와 풍부한 의미적 연관관계를 맺고 있는 학술용어를 추출하기 위한 방법을 모색하고자 하는 것이다. 본 연구에서 말하는 구조적 학술용어사전이란 유사한 속성을 가진 학술 용어들을 동일 개념으로 범주화하여 이들을 범주별로 분류하고, 각각의 개념들이 가지는 속성을 체계화한 다음, 학술 용어 하나하나의 의미를 용어가 속한 개념의 속성에 따라 정의하는 사전을 말한다. 이때 학술용어 하나하나는 해당되는 개념 범주의 속성 항목에 따라 구조화된 목록 즉 구조화된 메타데이터 형식으로 정의되며, 여러 개념의 속성 항목 간에 형성되는 연관 관계를 설정해줄 경우 연관 관계를 네트워크 형식으로 조직하는 것이 가능하게 된다. 그리고 이렇게 규정된 여러 연관 관계로부터 추론 규칙을 생성시킬 경우 의미적 연관 검색이 가능한 지식베이스가 생성될 수 있다.

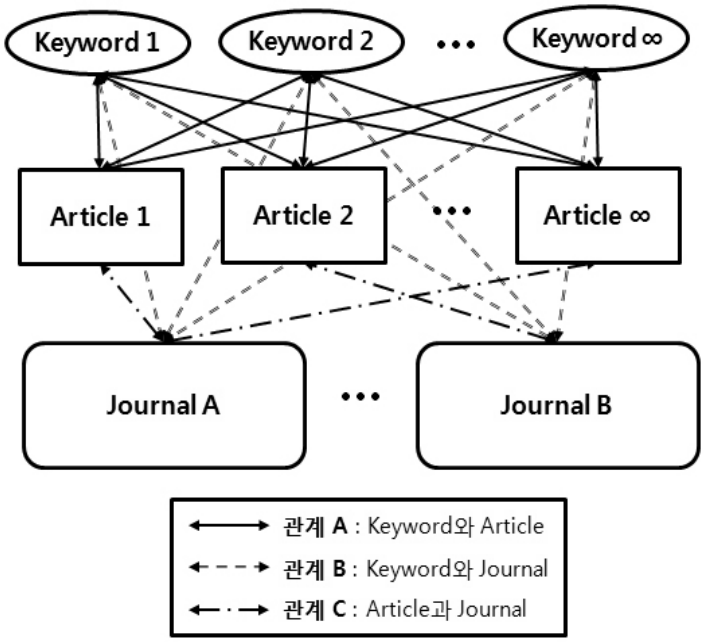
구조적 학술용어사전의 구축에 있어서 추출 대상 용어군으로 가장 많이 사용되는 것 중의 하나가 학술 논문의 저자키워드이다. 학술 논문의 저자키워드는 의미적인 면에서 해당 논문의 주제와 관련이 되며, 또한 저자가 해당 논문의 내용에서 가장 핵심적이고 중요하다고 판단하여 추출한 용어로 구성되어 있다(이춘실, 문혜원, 2000; Šauperl, 2004). 저자키워드를 대상으로 구조적 학술용어사전을 구축할 경우 연구주제와 키워드의 의미적 연관성에 따라 생성되는 연구논문의 지식 지도를 그려볼 수 있다. 그리고 해당 지식지도에서 나타나는 관계들에 대한 의미 추론 규칙을 생성할 수 있다면, 후속 연구자들이 또 다른 연구를 수행할 때 선행 연구 분석 과정에서 기존 연구자들이 중요하게 생각했던 학술적 의미와 연관 정보를 보다 효율적으로 추적하고 탐색할 수 있는 검색 시스템을 개발할 수 있게 된다. 연구자들은 이러한 시스템의 도움으로 관련 연구의 내용들을 보다 쉽게 분석하고 정리할 수 있을 것이다.

구조적 학술용어사전의 구축에 있어서 선행되어야 할 가장 중요한 작업 중의 하나는 사전 구축의 목적에 적합한 용어를 선정하는 일이다. 또한 학술논문의 저자키워드를 대상으로 용어를 선정할 경우 의미적 연관성이 풍부하고 연구에서 많이 활용되는 키워드를 선정하는 것이 사전의 질적 향상을 위해 무엇보다 중요하다(김판준, 이재윤, 2012; 양창진, 2010; 이혜영, 곽승진, 2011; 조민희, 정도헌, 2010; Gil-Leiva & Alonso-Arroyo, 2007; Hurt, 2010). 이와 관련하여 본 연구는 연구 논문이 포함하고 있는 저자키워드의 출현빈도와 키워드가 포함된 논문의 총피인용횟수의 상관 관계 및 저자키워드 출현빈도가 논문의 총피인용횟수에 영향을 미치는지를 확인하고자 하는 목적을 가진다. 저자키워드의 출현빈도가 늘어날수록 해당 키워드가 속한 논문의 총피인용횟수가 많아지는 관계를 보일 경우, 출현빈도가 높은 저자키워드를 구조적 학술용어사전의 구축 대상 용어로 선정하는 것이 적절한 방법의 하나가 될 수 있기 때문이다.

1.2 연구의 방법 및 범위

본 연구에서는 연구의 목적을 이루기 위하여 두 가지의 가설을 검정하고자 하였다. 하나의 가설은 인문학 및 사회과학 분야에 있어서 저자키워드의 출현빈도와 해당 키워드가 속한 논문의 피인용횟수 간의 관계를 확인하기 위한 것으로 '저자키워드의 출현빈도와 해당 키워드가 속한 논문들의 총피인용횟수 간에는 상관관계가 있을 것이다'이며, 검정을 위해 Pearson 상관분석을 실시하였다. 다른 하나의 가설은 키워드가 출현한 후 피인용이 된다는 시간적 인과 관계에 근거하여 저자키워드 출현빈도가 해당 키워드가 속한 논문의 총피인용횟수에 영향을 미치는지의 여부를 확인하기 위한 것으로 '저자키워드의 출현빈도는 해당 키워드가 속한 논문들의 총피인용횟수에 영향을 미칠 것이다'이며, 검정을 위해 단순회귀분석을 실시하였다. 가설 검정을 위한 분석에 사용된 데이터는 한국연구재단의 한국학술지인용색인(Korea Citation Index, 이하 KCI)에 수록된 2007년에서 2011년까지의 인문학 및 사회과학 분야 학술지 논문의 저자키워드와 피인용횟수이다.

저자키워드는 해당 내용이 포함되어 있는 군집에 따라 크게 2가지의 관계를 맺는다(〈그림 1〉 참조). 하나는 저자키워드가 삽입된 논문 단위에서의 관계(관계 A)이며 다른 하나는 저자키워드가 속한 논문이 발간된 학술지 단위와의 관계(관계 B)이다. '관계 A'에서는 보통 출현한 논문수를 의미하는 키워드의 출현빈도와 해당 키워드가 속한 논문의 실제 활용 정도의 척도가 되는 피인용횟수, 주제분야 분포 등



〈그림 1〉 키워드-논문-학술지의 관계

을 볼 수 있으며, '관계 B'에서는 각 저자키워드가 출현한 학술지에 대한 영향력지수(Impact Factor, 이하 IF)와 주제분야 분포 등을 분석할 수 있다.

학술 논문에 들어있는 저자키워드를 이용해 구조적 학술용어사전을 구축하기 위해서는 해당 키워드가 들어있는 학술 논문 및 학술지와 관계된 모든 변수들을 고려해 '관계 A'와 '관계 B' 간의 공통분모들에 대한 속성들을 분석하는 것이 필요하다. IF 값은 평가년도 기준 2년 전에 발간된 학술지에 속한 전체 논문에 대해 피인용된 횟수를 가지고 계산된 값이다. 키워드-논문-학술지의 매개 관계를 염두에 두고 연구를 진행할 경우 논문의 피인용횟수 산출 기준년도와 IF 값 산출 기준년도를 동일하게 설정해 집단을 구성할 때 명확한 기준이 되는 공통된 지점을 설정하기 어렵다는 문제가 있다. 따라서 본 연구에서의 분석은 학술 논문과 저자키워드와의 '관계 A'에 대한 것으로 한정하여 이루어졌다.

2. 인문학 및 사회과학 분야 학술지 논문의 저자키워드 출현 및 피인용횟수 특성

2.1 저자키워드 출현빈도

본 연구는 한국연구재단의 KCI DB에 구축된 인문학 및 사회과학 분야에 속한 2007년부터 2011년까지의 학술지 논문에 수록된 저자키워드를 대상으로 이루어졌다. 수집된 저자키워드 데이터는 학술지 논문에 기술된 서식을 그대로 유지한 키워드 나열형의 문자열로 구축되어 있었다. 따라서 키워드열을 키워드 단위로 분리하고 괄호, 특수기호, 공백을 처리하는 정제작업을 거쳐 자체 저자키워드 데이터베이스를 구축하였다.

구분된 저자키워드 중 본 연구의 분석 대상은 출현빈도 2회 이상인 한글 키워드로 제한하였다. 출현빈도 1회 키워드를 제외한 이유는 해당 키워드가 속한 논문이 피인용 되었더라도 인용을 한 연구자가 자신의 저술 논문에 해당 키워드를 저자키워드로 선정할 만큼의 중요성을 인정하지 않아 1회 출현으로 그친 것이라는 의미로 해석할 수 있기 때문이다(〈표 1〉 참조). 한글 키워드만을 분석 대상으로 삼은 것은 연구의 정확성을 기하기 위한 것이다. 외국어 키워드의 대부분이 한글 키워드에 대응되는 것이어서 의미가 중복될 뿐만 아니라 정제작업에도 불구하고 수집된 원본 데이터 자체에서 외국어 문자의 깨짐 현상과 같은 오류가 교정되지 않았다.

인문학 및 사회과학 분야에서 2007년부터 2011년까지 발간된 학술지에 포함된 저자키워드 출현수는 인문학 분야에서 192,186회, 사회과학 분야에서 282,459회로 나타났다. 이를 중분류별로 살펴볼 경우 저자키워드 출현수가 상위에 위치하는 분야는 인문학의 경우 한국어와문학(52,198회), 역사학(31,683회), 기타인문학(21,478회) 순으로, 사회과학에서는 법학(63,542회), 교육학(52,163회), 경영학(19,342회) 순인 것으로 나타났다(〈표 2〉 참조).

〈표 1〉 고유저자키워드의 출현빈도별 순위표

연번	저자키워드	출현빈도	순위	연번	저자키워드	출현빈도	순위
1	중국	797	1	41	주체	255	39
2	직무만족	613	2	42	유럽연합	254	40
3	정체성	582	3	43	타자	253	41
4	조직몰입	554	4	44	리더십	252	42
5	청소년	512	5	45	의사소통	252	42
6	우울	506	6	46	기억	252	42
7	신뢰	461	7	47	다문화사회	251	43
8	자아존중감	447	8	48	스토리텔링	251	43
9	자기효능감	427	9	49	기업지배구조	249	44
10	문화	422	10	50	효율성	243	45
11	민족주의	402	11				
12	민주주의	395	12				
13	세계화	394	13				
14	다문화교육	383	14				
15	번역	377	15				
16	고객만족	375	16	75255	힘의남용	2	254
17	교육과정	371	17	75256	힙합	2	254
18	질적연구	357	18	75257	POP광고유형	2	254
19	사회적지지	356	19	75258	광고메시지속성	2	254
20	한국	355	20	75259	교통광고	2	254
21	일본	352	21	75260	내직	2	254
22	욕망	351	22	75261	단발령	2	254
23	다문화주의	335	23	75262	디지털사이니지	2	254
24	인권	332	24	75263	매장요인	2	254
25	이미지	323	25	75264	발명의공개	2	254
26	근대성	316	26	75265	상호연동성	2	254
27	창의성	315	27	75266	수용어휘력	2	254
28	만족도	313	28	75267	시각적인표현	2	254
29	서비스품질	301	29	75268	시모노세키條約	2	254
30	자유	292	30	75269	신규성의제	2	254
31	통합교육	283	31	75270	신용가산금리	2	254
32	북한	279	32	75271	신제품평가	2	254
33	죽음	278	33	75272	옥외광고법	2	254
34	자연	271	34	75273	우민관	2	254
35	스트레스	265	35	75274	이정귀	2	254
36	교육	265	35	75275	인류세	2	254
37	여성	263	36	75276	정지승	2	254
38	저작권	259	37	75277	지방사족	2	254
39	인터넷	258	38	75278	쿠폰이용의도	2	254
40	거버넌스	255	39	75279	홍한주	2	254

〈표 2〉 한국연구재단 학술연구분야 중분류 기준 학술논문 수록 저자키워드 출현수

인문학	저자키워드 출현수	사회과학	저자키워드 출현수
일반	8,966	일반	13,010
가톨릭신학	1,042	경영학	19,342
기독교신학	8,542	경제학	12,582
기타동양어문학	663	관광학	8,310
기타인문학	21,478	교육학	52,163
독일어와문학	5,567	군사학	161
러시아어와문학	3,195	기타사회과학	7,009
문학	145	농업경제학	338
불교학	3,025	무역학	4,390
사전학	111	법학	63,542
서양고전어와문학	508	사회과학일반	14,439
스페인어와문학	147	사회복지학	10,893
언어학	5,260	사회학	5,628
역사학	31,683	신문방송학	10,194
영어와문학	9,856	심리과학	8,106
유교학	1,268	정책학	4,806
일본어와문학	4,085	정치외교학	12,732
종교학	4,090	지리학	5,563
중국어와문학	2,683	지역개발	4,009
철학	20,108	지역학	8,299
통역번역학	601	행정학	12,907
프랑스어와문학	6,965	회계학	4,036
한국어와문학	52,198	-	
인문학 소계	192,186	사회과학 소계	282,459
총 합계 474,645			

2.2 고유저자키워드가 속한 논문의 총피인용횟수

논문과 키워드의 관계는 동일 키워드가 여러 논문에 출현하는 1대 다수의 관계가 된다. 해당 키워드가 수록된 논문 편수를 의미하는 고유저자키워드별 출현빈도와 해당 키워드가 속해 있는 논문의 피인용횟수를 1대1로 비교해 보기 위해 키워드와 해당 키워드가 속한 논문들의 피인용횟수를 모두 합한 총피인용횟수를 구해 데이터를 정렬하였다. 이를 위해 가장 먼저 각 논문들을 기준으로 해당 논문에 출현한 고유저자키워드와 피인용횟수를 정리하였고, 그 다음 고유저자키워드를 기준으로 해당 키워드가 속한 전체 논문들의 피인용횟수를 모두 합해 총피인용횟수를 구하여 분석에 사용하였다.

한국연구재단 학술연구분야 분류표 기준 인문학 및 사회과학 분야에서 2007년부터 2011년까지 발간된 학술지는 인문학 500종, 사회과학 674종으로 총 1,174종이며 여기에 수록된 논문은 인문학 46,914편, 사회과학 83,263편으로 총 130,177편이다. 각 논문에 출현한 저자키워드는

총 474,645개이며 동일 형태 용어의 중복을 제거한 고유저자키워드 수는 총 75,279개이다. 해당 저자키워드들이 속한 논문의 피인용횟수 누적 합계는 자기피인용을 구분하지 않을 경우 총 791,847회로 집계되어 고유저자키워드 1개 당 평균 출현빈도는 약 6.3회, 평균 총피인용횟수는 약 10.5회인 것으로 나타났다(〈표 3〉 참조).

분석 결과, 총 75,279개의 고유저자키워드 중 가장 많은 총피인용횟수를 기록한 키워드는 '직무만족'으로, 총 613편의 논문에서 2,185회 피인용되었다. 가장 많은 출현빈도를 보인 '중국'의 경우 총피인용횟수 순위는 21위로 797편의 논문에서 785회 피인용된 것으로 나타났다(〈표 4〉 참조). 한편 2편 이상 13편 이하의 논문에 출현은 하였지만 한 번도 인용되지 않은 고유저자키워드는 10,849개로 전체 고유저자키워드 중 약 14.4%를 차지하였으며, 2편 이상 22편 이하 논문에 출현하였지만 피인용이 1회에 그친 고유저자키워드는 10,052개로 13.4%를 차지하여, 고유저자키워드가 출현하는 논문의 총피인용횟수가 1회 이하인 비율이 전체 고유저자키워드 수의 20%가 넘는 것으로 나타났다.

〈표 3〉 2007년-2011년 발간 학술논문수, 저자키워드 출현수, 총피인용횟수

구분 \ 년도	2007	2008	2009	2010	2011	합계
학술논문 수	20,351	24,587	27,570	29,678	27,991	130,177
저자키워드 출현수 (고유저자키워드 수)	73,792	89,962	101,154	108,547	101,190	474,645 (75,279)
총피인용횟수	226,264	235,349	194,345	113,426	22,463	791,847

〈표 4〉 고유저자키워드가 출현한 논문들의 총피인용횟수 기준 순위표

연번	저자키워드	출현빈도	총피인용횟수	순위	연번	저자키워드	출현빈도	총피인용횟수	순위
1	**직무만족**	**613**	**2,185**	**1**	41	이익조정	215	564	40
2	다문화교육	383	2,166	2	42	구매의도	175	557	41
3	우울	506	2,133	3	43	몰입	155	556	42
4	조직몰입	554	2,024	4	44	세계화	394	552	43
5	다문화주의	335	1,692	5	45	학업성취도	219	543	44
6	신뢰	461	1,446	6	46	민족주의	402	527	45
7	자아존중감	447	1,431	7	47	민주주의	395	526	46
8	사회적지지	356	1,387	8	48	내용분석	234	526	46
9	고객만족	375	1,374	9	49	인권	332	510	47
10	청소년	512	1,292	10	50	리더십	252	510	47
11	자기효능감	427	1,253	11					
12	다문화사회	251	1,134	12			·		
13	서비스품질	301	1,084	13			·		
14	질적연구	357	1,005	14			·		
15	매개효과	226	932	15					
16	노인	232	911	16	75255	흔암리유형	2	0	400

연번	저자키워드	출현빈도	총피인용횟수	순위	연번	저자키워드	출현빈도	총피인용횟수	순위
17	스트레스	265	874	17	75256	흡연반감	2	0	400
18	만족도	313	840	18	75257	흡연의사	2	0	400
19	다문화	212	823	19	75258	흡입력	2	0	400
20	정체성	582	818	20	75259	흥분	2	0	400
21	**중국**	**797**	**785**	**21**	75260	흥성문화	2	0	400
22	이직의도	204	785	21	75261	흥학	2	0	400
23	만족	218	762	22	75262	흥행통계	2	0	400
24	학업성취	178	761	23	75263	희	2	0	400
25	다문화가정	175	759	24	75264	희귀식물	2	0	400
26	삶의질	223	741	25	75265	희망버스	2	0	400
27	교육과정	371	739	26	75266	희생적사랑	2	0	400
28	조직시민행동	197	738	27	75267	희화	2	0	400
29	통합교육	283	686	28	75268	흰두교	2	0	400
30	행동의도	134	659	29	75269	히구치이치요	2	0	400
31	직무스트레스	204	656	30	75270	히라가나	2	0	400
32	사회자본	204	617	31	75271	히메유리학도대	2	0	400
33	문화	422	615	32	75272	히브리서	2	0	400
34	자살생각	85	604	33	75273	힌두민족주의	2	0	400
35	스토리텔링	251	595	34	75274	힌두이즘	2	0	400
36	인터넷	258	586	35	75275	힌두철학	2	0	400
37	교사교육	172	583	36	75276	힐리스밀러	2	0	400
38	창의성	315	570	37	75277	힘의균형	2	0	400
39	거버넌스	255	566	38	75278	힘의남용	2	0	400
40	기업지배구조	249	565	39	75279	힙합	2	0	400

2.3 저자키워드가 출현한 학술논문의 총피인용횟수에 따른 구간별 분포

저자키워드가 속해 있는 논문의 총피인용횟수에 따라 저자키워드 출현 분포에 차이가 있는지를 살펴보기 위해 논문의 총피인용횟수를 기준으로 4개의 집단으로 나누어 집단별 특성을 살펴보았다. 집단을 나누기 위해 먼저 각각의 키워드별로 해당 키워드가 속한 논문들의 총피인용횟수를 상위에서 하위까지 내림차순으로 정렬하여, 각 키워드별로 상위부터 하위까지 차례대로 총피인용횟수의 누적값을 구하였다(〈표 5〉 참조).

총피인용횟수의 누적값을 토대로 전체 키워드의 총피인용횟수 누적합계 대비 각 키워드별 총피인용횟수 누적 값의 백분율을 구해 25%씩 네 구간으로 나누었다. 이때 동일한 수치의 총피인용횟수가 많이 나타나 구간별로 정확하게 구분되지 않는 문제점을 해결하기 위하여 구간 구별의 기준 값인 25%, 50%, 75%에 가장 근접해서 나누어지는 값을 구간의 구별점으로 삼았다(〈표 6〉 참조).

집단별 분석 결과 전체 고유저자키워드 75,279개 중 하위 약 80%(79.9%)에 속하는 '집단 4'의

〈표 5〉 총피인용횟수 누적 비율에 따른 집단 구분 산정 방식

키워드	총피인용횟수	누적 총피인용횟수	누적 비율
A	500	500	5% [=(500/10,000)*100]
B	400	900	9% [=(900/10,000)*100]
C	300	1200	12% [=(1,200/10,000)*100]
		⋮	
X	2	9,999	99.9% [=(9,999/10,000)*100]
Y	1	10,000	100%
Z	0	10,000	100%

〈표 6〉 총피인용횟수 기준 구간별 고유저자키워드 수 및 출현빈도 분포

	집단 구분 기준		고유저자키워드 수(%)	고유저자키워드별 출현빈도 분포
	누적 피인용 비율	피인용횟수		
집단 1	상위 25.09%	104~2,185회	869(1.2%)	8회 ~ 797회
집단 2	25.09% ~ 49.95%	32~103회	3,758(5.0%)	2회 ~ 155회
집단 3	49.95% ~ 74.14%	12~31회	10,489(13.9%)	2회 ~ 80회
집단 4	74.14% ~ 100%	0~11회	60,163(79.9%)	2회 ~ 44회
총계			75,279(100%)	

60,163개 고유키워드가 피인용횟수 누적 비율 기준 하위 약 26%(25.86%, 피인용횟수 0~11회)에 분포하는 것으로 나타났다. 그리고 상위 20%에 속하는 '집단 1, 2, 3'의 고유저자키워드 15,116개가 피인용횟수 누적 비율 기준 상위 약 74%(74.14%, 피인용횟수 12회 이상)에 분포하는 것으로 나타났다.

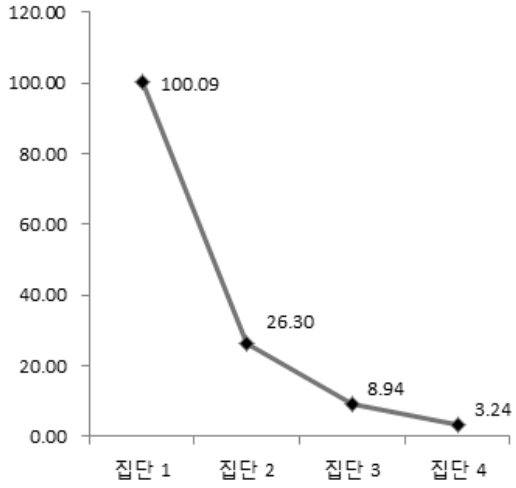
2.4 집단별 저자키워드 출현빈도

고유저자키워드가 출현한 학술논문의 총피인용횟수 순위에 따라 구분한 4개 집단의 2007년부터 2011년까지의 연도별 저자키워드 출현수는 〈표 7〉과 같다.

각 해당 집단의 전체 고유저자키워드 출현수 소계를 고유저자키워드 수로 나눈 '집단별 고

〈표 7〉 총피인용횟수 순위 기준 집단의 연도별 저자키워드 출현수

	2007	2008	2009	2010	2011	소계	고유키워드수
집단 1	13,224	16,181	18,292	20,169	19,112	86,978	869
집단 2	15,844	19,255	21,295	22,140	20,288	98,822	3,758
집단 3	15,758	18,808	20,595	20,599	18,027	93,787	10,489
집단 4	28,966	35,718	40,972	45,639	43,763	195,058	60,163
합계						474,645	75,279



〈그림 2〉 총피인용횟수 순위 기준 집단별 고유키워드의 평균 출현빈도 비교

유키워드의 평균 출현빈도'는 '집단 1'이 약 100회, '집단 2'는 약 26.3회, '집단 3'은 약 8.9회, '집단 4'는 약 3.2회로 나타나 집단에 따라 고유저자키워드 당 평균 출현수는 상당한 차이가 있었다(〈그림 2〉 참조).

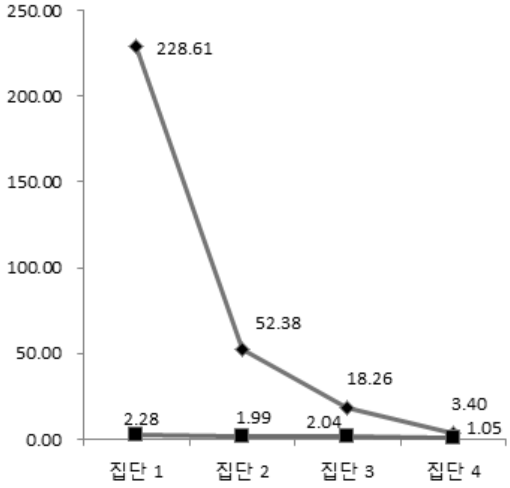
2.5 집단별 논문 총피인용횟수

2007년부터 2011년까지 4개 집단에 속한 키워드가 출현하는 논문의 총피인용횟수는 〈표 8〉과 같다.

해당 집단의 총피인용횟수 소계를 고유저자키워드 수로 나눈 '집단별 고유저자키워드 평균 피인용횟수'는 '집단 1'이 약 228.6회, '집단 2'는 약 52.4회, '집단 3'은 약 18.3회, '집단 4'는 약 3.4회로 나타났다. 해당 집단의 총피인용횟수 소계를 해당 집단의 키워드 출현수 소계로 나눈 '집단별 저자키워드 출현 1회당 평균 피인용횟수'는 '집단 1'이 약 2.3회, '집단 2'는 약 2.0회, '집단 3'은 약 2.0회, '집단 4'는 약 1.0회로 나타났다(〈그림 3〉 참조).

〈표 8〉 총피인용횟수 순위 기준 집단의 연도별 총피인용횟수

	2007	2008	2009	2010	2011	소계	고유키워드수
집단 1	59,576	59,511	46,966	27,166	5,443	198,662	869
집단 2	59,929	59,816	46,745	25,534	4,812	196,836	3,758
집단 3	57,112	58,325	47,342	24,476	4,290	191,545	10,489
집단 4	49,647	57,697	53,292	36,250	7,918	204,804	60,163
합계						791,847	75,279



〈그림 3〉 집단별 고유저자키워드 당 평균 피인용횟수 및 저자키워드 출현 1회당 평균 피인용횟수

3. 학술논문의 저자키워드 출현빈도와 논문의 총피인용횟수

3.1 저자키워드의 출현빈도와 해당 키워드가 속한 논문의 총피인용횟수 간 상관관계

저자키워드 출현빈도 및 해당 키워드가 속한 논문들의 총피인용횟수 간 상관관계를 보기 위해 '저자키워드의 출현빈도와 해당 키워드가 속한 논문들의 총피인용횟수 간에는 상관관계가 있을 것이다'는 가설에 대하여 Pearson 상관 분석을 실시하였다. 분석 결과 〈표 9〉와 같이 저자키워드 출현빈도와 총피인용횟수에는 유의수준 0.01 수준에서 높은 상관관계(=.858)가 있는 것으로 나타나 가설은 채택되었다.

또한 저자키워드를 총피인용횟수 순위 기준으로 누적 구간을 나눠 구분한 4개의 집단별 분석에 있어서도 〈표 10〉과 같이 유의수준 0.01에서 통계적으로 유의한 상관관계가 있는 것으로

〈표 9〉 저자키워드 출현빈도와 총피인용횟수 간의 상관 분석

		출현빈도	총피인용횟수
출현빈도	Pearson 상관계수 (샘플수)	1 (75,279)	.858** (75,254)
총피인용횟수	Pearson 상관계수 (샘플수)	.858** (75,254)	1 (75,254)

* p<.05, ** p<.01, *** p<.001

※ 전체 75,279개의 고유저자키워드 중 25개 키워드는 피인용횟수가 표기되지 않은 채 데이터가 반입되어, 데이터 분석 시 결측값으로 처리함

〈표 10〉 각 집단별 고유저자키워드 출현빈도와 총피인용횟수 간의 상관 분석

			출현빈도	총피인용횟수
집단 1	출현빈도	Pearson상관계수 (샘플수)	1 (869)	.757** (869)
	총피인용횟수	Pearson 상관계수 (샘플수)	.757** (869)	1 (869)
집단 2	출현빈도	Pearson 상관계수 (샘플수)	1 (3,758)	.525** (3,758)
	총피인용횟수	Pearson 상관계수 (샘플수)	.525** (3,758)	1 (3,758)
집단 3	출현빈도	Pearson 상관계수 (샘플수)	1 (10,489)	.358** (10,489)
	총피인용횟수	Pearson 상관계수 (샘플수)	.358** (10,489)	1 (10,489)
집단 4	출현빈도	Pearson 상관계수 (샘플수)	1 (60,163)	.405** (60,138)
	총피인용횟수	Pearson 상관계수 (샘플수)	.405** (60,138)	1 (60,138)

* p<.05, ** p<.01, *** p<.001

나타났다. 다만 '집단 1'의 경우는 높은 상관관계(=.757), '집단 2'와 '집단 4'는 다소 높은 상관관계(각각 =.525, =.405), '집단 3'은 낮은 상관관계(=.358)가 있는 것으로 나타나 집단별로 관련성의 정도에는 차이가 있었다.

3.2 저자키워드의 출현빈도가 해당 키워드가 속한 논문의 총피인용횟수에 미치는 영향

저자키워드의 출현빈도가 해당 키워드가 속한 논문의 총피인용횟수에 미치는 영향을 알아보기 위하여 '저자키워드의 출현빈도는 해당 키워드가 속한 논문들의 총피인용횟수에 영향을 미칠 것이다'는 가설에 대하여 단순 회귀분석을 실시한 결과 〈표 11〉과 같이 저자키워드 출현빈도가 총피인용횟수에 유의한 영향을 주는 것(p=.000<.05)으로 나타나 가설은 채택되었다. 회귀식에 의하면 저자키워드 출현빈도가 1회 증가함에 따라 해당 키워드가 속한 논문들의 총피인용횟수는 약 1.982회 증가하는 것으로 나타났다. 저자키워드 출현빈도가 총피인용횟수에 미치는 영향력에 대한 설명력(R^2)은 73.5%였다.

또한 총피인용횟수 누적 구간에 따라 구분한 4개의 집단별 분석에 있어서도 〈표 12〉와 같이 저자키워드 출현빈도가 총피인용횟수에 유의한 영향을 주는 것(p=.000<.05)으로 나타났다. 하지만 각 집단별로 회귀식에 따른 저자키워드 출현빈도 별 총피인용횟수 증가 정도와 영향력에 대한 설명력(R^2)에는 차이가 있는 것으로 나타났다. '집단 1'의 경우는 회귀식에 의해 저자키워드 출현빈도가 1회 증가함에 따라 해당 키워드가 속한 논문들의 총피인용횟수가 약 1.992회 증가하는 것으로 나타났고, 저자키워드 출현빈도가 총피인용횟수에 미치는 영향력에 대한 설

〈표 11〉 저자키워드 출현빈도가 총피인용횟수에 미치는 영향

모형	비표준화계수		표준화 계수	T	p
	B	표준오차 (SE)	β		
상수	-1.977	.072		-27.400	.000
저자키워드 출현빈도	1.982	.004	.858	457.382	.000
R^2 = .735 F = 209197.910 p = .000***					

* p<.05, ** p<.01, *** p<.001

〈표 12〉 집단별 저자키워드 출현빈도가 총피인용횟수에 미치는 영향

집단 구분	모형	비표준화계수		표준화 계수	T	p
		B	표준오차 (SE)	β		
집단 1	상수	29.277	7.604		3.850	.000
	저자키워드 출현빈도	1.992	.058	.757	34.106	.000
	R^2 = .573 F = 1163.245 p = .000***					
집단 2	상수	38.703	.442		87.620	.000
	저자키워드 출현빈도	.502	.014	.525	37.802	.000
	R^2 = .276 F = 1428.954 p = .000***					
집단 3	상수	15.762	.080		196.740	.000
	저자키워드 출현빈도	.280	.007	.358	39.297	.000
	R^2 = .128 F = 1544.246 p = .000***					
집단 4	상수	1.585	.020		78.476	.000
	저자키워드 출현빈도	.561	.005	.405	108.729	.000
	R^2 = .164 F = 11821.944 p = .000***					

* p<.05, ** p<.01, *** p<.001

명력(R^2)은 57.3%였다. '집단 2'의 경우에는 저자키워드 출현빈도 1회 증가에 따라 해당 키워드가 속한 논문들의 총피인용횟수는 약 0.502회 증가하는 것으로 나타났고 설명력(R^2)은 27.6%였다. '집단 3'은 저자키워드 출현빈도 1회 증가에 따라 해당 키워드가 속한 논문들의 총피인용횟수는 약 0.280회 증가하며 설명력은 12.8%였다. '집단 4'는 저자키워드 출현빈도 1회 증가에 따라 해당 키워드가 속한 논문들의 총피인용횟수가 약 0.561회 증가하고 설명력(R^2)은 16.4%였다.

4. 구조적 학술용어사전 구축을 위한 용어 선정 기준으로서의 저자키워드 출현빈도와 피인용횟수

4.1 저자키워드 출현빈도와 피인용횟수의 중요성과 한계

저자키워드 출현빈도와 해당 키워드가 속한 논문들의 총피인용횟수 간의 상관관계 및 영향력에 대한 가설 검정 결과를 보면, 4개로 구분한

〈표 13〉 저자키워드 출현빈도와 논문의 총피인용횟수를 고려한 용어 선정 시 주의사항

		저자키워드 출현빈도	
		낮음	높음
총피인용횟수	적음	신규 출현 용어이거나 해당 주제분야 희소성 등의 고려가 필요함 ※ 예: 빠움짜리우, 스와힐리 정체성 등	후속 연구자가 키워드로 선정하기에는 중요성이 떨어지지만 해당 논문에서는 필요한 용어였을 가능성이 높음 ※ 예: 등장인물, 화두 등
	많음	연구 주제의 독창성 등에 의해 의미적인 면에서 주제의 특수성을 반영한 용어일 가능성이 높음 ※ 예: 다문화적 효능감, 노인자살 등	선정 용어로 적합하지만, 전 주제 분야에 걸쳐 출현하는 경우 지나치게 일반적인 용어가 아닌지 고려가 필요함 ※ 예: 중국, 자아존중감, 만족도 등

집단 간 차이는 조금씩 있으나 두 변인 간에는 유의한 상관관계가 있는 것으로 나타났다. 그리고 출현빈도가 높을수록 총피인용횟수 또한 많아지는 것으로 나타났다. 이 결과는 특정한 키워드가 다수의 논문에 출현할수록 피인용 될 수 있는 기회가 많아지기 때문이라는 논리적 해석이 가능하다. 따라서 연구자들이 많이 이용할 수 있는 학술용어사전을 구축하고자 하는 경우에는 전체적으로 출현빈도가 높고, 해당 키워드가 속한 논문의 총피인용횟수도 많은 키워드 군에서 용어를 선정하는 것이 사전의 활용성을 높이는 방법 중의 하나가 될 수 있을 것이다.

이와 관련하여 키워드의 출현빈도와 해당 키워드가 속한 논문의 피인용횟수 간의 관계만을 가지고 용어 선정 기준을 삼는다고 할 경우, 저자키워드 출현빈도 및 키워드가 속한 논문의 총피인용횟수의 정도에 따라 용어 선정시 고려 사항을 〈표 13〉과 같이 정리해 볼 수 있다. 출현빈도가 낮고 총피인용횟수도 적은 키워드의 경우, 주제의 희소성이나 특수성에 따라 발간된 논문의 편수 자체도 얼마 안 되고 인용횟수도 적을 수 있기 때문에 의미적인 중요성이 결코 떨어진다고 할 수는 없다. 특히 전체 고유저자키워드의 20% 가량이 총피인용횟수가 1회 이하인 점을 고려해보면, 단지 총피인용횟수가 적고 출현빈도가 낮은 집단에 해당하는 저자키워드라는 이유만으로 용어 선정 대상에서 제외할 수는 없을 것이다.

따라서 출현빈도와 총피인용횟수가 용어 선정의 유일한 기준이 될 수는 없을 것이며, 해당 용어가 속해 있는 논문의 주제 분야 특성이나 용어의 출현 시기 등에 대한 고려가 필요하다(〈표 13〉 참조). 새로 출현한 키워드의 경우 출현빈도가 낮고 총피인용횟수도 적은 경향이 있으므로 키워드의 생명주기와 관련된 용어의 선정 기준에 대해서는 다른 방식의 방법론을 적용해 접근해야 할 것이다.

4.2 저자키워드의 일반성과 출현 분야에 따른 특수성

전체 모집단 및 4개로 구분한 집단에 따라 주제 분야별 저자키워드의 출현 분포를 분석한 결과, 전체 모집단과 4개 집단 각각에서의 주제분야별 저자키워드 출현 분포는 중분류를 기준으로 출현수 순위에 다소 차이가 있었으나 대체

적으로 집단 간에 큰 차이가 없이 비슷한 분포 양상을 보였다. 그렇지만 중분류별 분포에서 저자키워드 출현수가 가장 많은 분야와 가장 적은 분야의 차이는 매우 심한 것으로 나타났다(인문학: 한국어와문학 52,198 ↔ 사전학 111, 사회과학: 법학 63,542 ↔ 군사학 161). 따라서 학술용어사전 구축을 위한 용어 선정에 있어서는 해당 키워드가 출현한 주제의 내용적 측면과 함께 각 주제 분야별로 키워드가 출현한 비율을 함께 고려해 선정할 필요가 있을 것이다.

이와 관련하여 주제의 성격이 상이한 분야에 동시에 다수 출현하는 키워드는 키워드의 출현빈도가 높고 총피인용횟수도 많은 경우에도 지나치게 일반적으로 사용되는 용어가 아닌지에 대한 고려가 필요하다. 그렇지만 해당 용어가 일반적인 용어라 할지라도 출현하는 주제에 따라 독자적인 성격으로 사용될 수 있는 여지가 있으므로 용어 선정 후 정의를 하는 작업이나 다른 용어와의 관계 설정 시 주의가 필요할 것이다.

5. 결론 및 제언

본 연구의 목적은 학술논문이 포함하고 있는 저자키워드의 출현빈도와 키워드가 포함된 논문의 총피인용횟수의 상관 관계 및 저자키워드 출현빈도가 논문의 총피인용횟수에 영향을 미치는지를 확인하는 것이다. 연구 목적을 이루기 위하여 KCI에 등록된 2007년에서 2011년까지의 인문학 및 사회과학 분야 학술지 수록 논문에 2회 이상 출현한 한글 저자키워드와 학술지 논문의 피인용횟수를 분석하였다. 본 연구의 분석 결과는 다음과 같다.

첫째, 저자키워드의 출현빈도와 해당 키워드가 속한 학술 논문의 총피인용횟수는 상관분석을 실시한 결과 통계적으로 유의한 상관관계를 보였다. 또한 저자키워드 출현 후 피인용이 일어나는 인과관계에 따라 출현빈도와 총피인용횟수 간에 단순 회귀 분석을 실시한 결과, 출현빈도가 높아질수록 총피인용횟수도 증가하는 것으로 나타났다. 이러한 결과는 전체 모집단뿐만 아니라, 총피인용횟수를 기준으로 나눈 4개의 집단에서도 정도의 차이는 다소 있으나 모두 유의한 것으로 확인되었다. 따라서 연구자들이 많이 이용할 수 있는 학술용어 사전을 구축하고자 하는 경우에는 전체적으로 출현빈도도 높고, 해당 키워드가 속한 총피인용횟수도 많은 키워드 군에서 용어를 선정하는 것이 활용성 측면에서 적절한 것으로 해석할 수 있다.

둘째, 총피인용횟수를 기준으로 나눈 집단 내에서도 각 키워드가 속한 논문들의 총피인용횟수가 동일하더라도 키워드별 출현빈도의 분포에 차이가 크게 나타나는 경향을 보였다. 출현빈도 2 이상과 한글 키워드라는 제한 기준을 두어 전체 군집을 통제하였으나 모집단의 규모가 매우 크기 때문에 표준 범위를 벗어난 변칙적인 사례일 가능성을 배제하기는 어렵다. 그렇지만 이는 용어 선정에 있어 각 키워드가 출현한 논문의 주제분야가 가지는 특수성과 용어의 일반성을 함께 고려하는 것이 필요함을 보여주는 것이라 할 수 있다.

셋째, 한국연구재단의 학술연구분야 분류표 주제 구분에 따라 대분류 및 중분류에 분포된 키워드의 동시 출현수를 분석한 결과, 전체 모집단의 분포 비율과 총피인용횟수로 나눈 4개 집단의 분포 비율은 대체적으로 비슷한 양상을

띠었다. 그렇지만 중분류별 분포에 있어서는 주제 분야의 특성에 따라 저자키워드 출현수의 차이가 매우 심한 것으로 나타났다. 따라서 해당 키워드가 출현한 주제의 내용적 측면과 함께 각 주제 분야별로 키워드가 출현한 비율을 함께 고려해 선정하는 것이 필요하다.

본 연구는 저자키워드와 학술 논문 간의 관계만을 염두에 두고 분석을 진행하였다는 제한점을 가지고 있다. 이는 검색 시 저자키워드의 이용 효율성을 측정하거나, 저자키워드 출현 분야별 특성에 대한 클러스터링 분석 결과를 제시한 기존 선행연구들과는 달리, 구조적 용어사전 구축에 실제 활용할 수 있는 용어의 추출 방법론을 개발해 보고자 한 것에서 기인한다. 따라서 저자키워드와 학술지 단위 간의 관계에서 고려할 수 있는 내용들에 대해 다루지 않았으며, 용어별 생명주기나 용어 간 클러스터링 분석 등에 대한 내용이 고려되지 않은 본 연구의 한계는 후속 연구를 통해 보완될 필요가 있다.

본 연구는 인문학 및 사회과학 분야의 구조적 학술용어사전을 구축하는데 있어서 연구에서의 활용도가 높으며 다른 용어와 풍부한 의미적 연관관계를 맺고 있는 학술용어를 추출하기 위한 방법을 모색하고자 하는 것을 연구의 배경으로 한다. 구조적 학술용어사전을 통해 학술지 논문의 의미적 지식 지도를 형성하기 위해서는 용어 선정뿐만 아니라 각각의 용어에 대한 풍부한 의미 정보 및 효율적 검색을 위한 추론 규칙의 생성도 중요하다. 따라서 본 연구의 범위에는 포함되지 않았던 의미적 연관관계 형성에 관한 내용들에 대해서도 후속 연구와 심화 연구가 이루어져야 할 것이다.

참 고 문 헌

김판준, 이재윤 (2012). 디스크립터 자동 할당을 위한 저자키워드의 재분류에 관한 실험적 연구. 정보관리학회지, 29(2), 225-246. http://dx.doi.org/10.3743/KOSIM.2012.29.2.225

양창진 (2010). 학술 논문의 주제어 표기 및 활용 방안 연구: DB 구축 및 정보연계의 관점에서. 인문콘텐츠, 19, 395-416.

이춘실, 문혜원 (2000). 한국의학학술 논문의 저자선정 주제어와 MeSH 용어의 비교 분석. 정보관리학회지, 17(3), 109-204.

이혜영, 곽승진 (2011). 국내 학술지 논문의 주제어를 통한 학술연구분야 관계분석. 한국비블리아학회지, 22(3), 354-371.

조민희, 정도헌 (2010). 학술정보데이터의 키워드 연관성 분석. 2010년도 한국인터넷정보학회 추계학술발표대회 논문집, 11(2), 149-150.

한국학술지인용색인 홈페이지. Retrieved from https://www.kci.go.kr/

Hurt, C. D. (2010). Automatically generated keywords: A comparison to author-generated keywords in the sciences. Journal of Information and Organizational Sciences, 34(1), 81-88.

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. Journal of the American Society for Information Science and Technology, 58(8), 1175-1187. http://dx.doi.org/10.1002/asi.20595

Šauperl, A. (2004). Catalogers' common ground and shared knowledge. Journal of the American Society for Information Science and Technology, 55(1), 55-63. http://dx.doi.org/10.1002/asi.10351

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

Cho, Min-Hee, & Jeong, Do-Heon (2010). The analysis of the keyword relevance in academic information data. Proceedings of 2010 Korean Society for Internet Information Fall Conference, 11(2), 149-150.

Kim, Pan Jun, & Lee, Jae Yun (2012). A study on the reclassification of author keywords for automatic assignment of descriptors. Journal of the Korean Society for Information Management, 29(2), 225-246. http://dx.doi.org/10.3743/KOSIM.2012.29.2.225

Korea Citation Index. Retrieved from https://www.kci.go.kr/

Lee, Choon-Shil, & Moon, Hye-Woon (2000). A comparison study of subject words of Korean medical journal papers: Author keywords vs MeSH terms assigned by MEDLINE. Journal of the Korean Society for Information Management, 17(3), 109-204.

Lee, Hye-Young, & Kwak, Seung-Jin (2011). Relation analysis among academic research areas using subject terms of domestic journal papers. Journal of the Korea Biblia Society for Library and Information Science, 22(3), 354-371.

Yang, Chang-Jin (2010). Study on keywords and their use of academic theses: Focused on database development and information link. Human Contents, 19, 395-416.