과학기술 논문의 저자 키워드 분석

Analysis on the author keywords in the scientific articles

김 태 중, 이 석 형, 김 광 영, 김 환 민
한국과학기술정보연구원

Kim Tae-Jung, Lee Seok-Hyoung,
Kim Kwang-Young, Kim Hwanmin
Korea Institute of Science and Technology Information

요약

대부분 국내에서 발행되는 과학기술 분야의 논문에는 저자 키워드가 포함되어 있다. 이 키워드는 논문을 이해를 돕고 온라인 검색에 유용하게 활용되고 있다. 특히 많은 논문에서 키워드를 영문과 국문을 동시에 부여하도록 하고 있어 과학기술 용어로서의 가치도 있다. 일정 기간 국내에서 발행되는 논문으로부터 저자 키워드들을 추출하여 다양한 각도에서 부여 키워드의 현황을 분석하였다. 결론으로 바람직한 키워드 부여의 방향을 제시한다.

Ⅰ. 서론

1. 연구 목적

국내에서 발간되는 대부분 학회의 논문지와 학회지에 논문을 제출할 때에 국문과 영문으로 저자 키워드를 함께 제출하도록 하고 있다. 이러한 키워드는 논문 데이터베이스 구축시에 검색어로 활용되며 논문의 대강을 이해하는데 참고가 된다. 또한 논문의 저자가 생각하는 중요한 키워드 즉 용어이므로 해당 분야의 학술 용어로서의 가치도 있을 것으로 판단된다. 더욱이 국문과 영문 키워드를 동시에 부여하고 있어 영-한 또는 한-영 대역어로 활용될 수 있을 것이다. 특히 학회 논문지 또는 학회지는 최근의 연구결과를 발표하고 있어 새로운 학술 용어를 수집하는 도구로 활용될 수 있다.

이에 학회 논문지 등에 수록된 저자 키워드의 실태를 분석하여 이러한 활용 가능성을 조사하고자 한다.

2. 연구 방법

한국과학기술정보연구원(KISTI)이 수집한 2008-2011년에 국내에서 발간된 366종의 학회지 및 논문지에 수록된 논문 90,214건[1]을 대상으로 논문을 제출할 당시에 국문과 영문으로 저자가 부여한 키워드를 조사·분석하였다. 주요 조사내용은 평균 키워드의 수(국·영문), 학문 분야별 키워드 부여의 특성 여부, 영문 키워드 수와 국문 키워드 수의 일치 여부, 국문과 영문 키워드의 수가 같은 경우에 국영 대역의 적합성 여부 등을 조사하였다.

국영 대역 적합성 여부는 국문과 영문 키워드의 수가 같은 경우(27,652건)에 대하여 논문의 제목을 한글 자모순으로 배열한 후에 1%(276건)를 무작위 추출하여 이로부터 1,226건의 키워드 쌍을 대상으로 조사하였다. 부여된 키워드를 기계적으로 추출한 후 활용 가능성을 조사하기 위해 여러 개의 키워드를 부여했을 때 한영의 순서가 서로 다르면 적합하지 않은 것으로 평가하였다. 평가는 '적합', '이해가능', '부적합' 등의 3가지로 분류하였다.

Ⅱ. 본론

1. 일반 현황

논문당 평균 키워드의 수를 보면 전체 90,214건의 논문에서 국문 키워드수가 총 149,054개이어서 1개의 논문당 평균 1.65개가 그리고 영문 키워드는 총 350,682개로서 논문당 3.89개가 부여되어 있다. 각각 키워드가 없는(부여되지 않은 경우)를 제외하면 국문 키워드는 4.12개, 영문 키워드는 4.34개이다. 90,214건의 논문 가운데 국문과 영문 모두 키워드가 없는 논문이 8,814건, 국문 키워드가 없고 영문 키워드만 있는 경우 44,954건, 국문 키워드만 있는 논문은 591건이다.

이상의 결과에서 보면 10%에 가까운 논문에 국문과 영문 모두 저자 키워드가 없으며, 대다수의 논문에 영문 키워드는 있으나 국문 키워드는 그렇지 않다. 영문과 국문 키워드가 모두 있는 논문은 39.74%인 35,855건에 불과하다. 즉, 영어가 매우 중요시되며 40%미만이 국문과 영문 키워드의 비중을 같게 보고 있다고 할 수 있다. 이 가운데 국문과 영문 키워드 수가 같은 경우가 27,652건

1) 2012년에 중국 연변데이터베이스센터에서 국가표준과학기술분류체계(교육과학기술부, 2010년)에 따라 주제 분류하였다.

이며 8,203건은 다르다. 표 1은 키워드 현황을 그리고 표 2는 평균 키워드 수이다. 표 3은 학문 분야별 논문의 분포를 보여준다. 366종의 발간물로 보면 국영문 모두 없는 경우가 40종, 국문 키워드가 전혀 없는 경우가 129종, 영문 키워드가 없는 경우는 43종이다.

표 1. 키워드 현황

분류	논문수	구성비(%)
국문만 있음	951	0.66
영문만 있음	44,954	49.83
국·영문 모두 있음	35,855	39.74
(국·영 동수)	(27,652)	(30.65)
국·영문 모두 없음	8,814	9.77
계	90,214	100.00

표 2. 평균 키워드 수

전체 평균	국문 키워드 영문 키워드	1.65 3.89
키워드 없는 경우 제외	국문 키워드 영문 키워드	4.12 4.34

표 3. 조사 분석 대상 논문의 분야별 분포

분야	논문수	평균국문키워드수	평균영문키워드수	국영동수 논문수	샘플수	부적합 키워드쌍
건설/교통	6,387	2.46	3.39	2,931	176	7
경제/경영	1,345	1.10	3.47	252	11	2
교육	2,089	2.85	3.57	1,234	75	3
기계	10,286	3.02	4.12	6,121	315	10
농림수산 식품	8,874	0.73	4.25	1,039	37	13
물리학	1,029	1.96	3.88	371	11	
미디어 등	1,305	4.34	4.74	805	29	6
보건의료	16,905	1.07	3.89	3,874	119	9
생명과학	2,784	0.39	4.18	199		
수학	1,736	2.53	4.16	828	29	13
에너지/자원	2,499	1.76	3.81	833	27	1
원자력	897	2.89	4.03	479	43	
재료	4,720	1.39	4.23	1,245	74	2
전기/전자	6,049	0.81	3.71	865	36	1
정보/통신	12,133	1.49	3.51	2,811	84	6
지구과학	2,545	2.32	4.05	1,030	65	2
지리/지역...	714	2.52	3.83	331	8	2
화공	1,396	0.79	4.39	198		
화학	1,606	1.22	4.00	339	8	
환경	1,996	2.21	4.19	790	25	2
기타	2,919	1.86	3.37	1,077	54	5
계	90,214	1.65	3.89	27,652	1,226	84

2. 키워드 비교

국·영문 키워드의 개수가 같은 경우(27,652건)에 국한해서 1%(276건)를 무작위(랜덤)추출하여 국문과 영문 키워드를 순차적으로 하나씩 용어의 의미를 비교해 보았다. 하나의 논문에 국문과 영문 키워드가 모두 있는 경우에도 키워드의 수가 다르다면 국·영문 키워드를 순차적으로 비교할 수 없기 때문에 비교 대상에서 제외하였다. 비교는 '적합', '이해가능', '부적합' 등의 3가지로 나누어 보았으며 부적합의 경우는 이유를 찾아보았다.

276건의 논문에서 1,226쌍의 키워드를 추출하였으며 이들 키워드 쌍의 의미 적합성(일치성)을 객관적 관점에서 비교하기 위해 쉽게 이해할 수 있는 키워드에 대해서도 '네이버[2]'와 '구글[3]'에서 검색하여 판단하려고 하였다. 비교한 결과는 표 4와 같다. 276건의 논문 가운데 18.48%인 51건의 논문에서 부적합한 84쌍의 키워드가 발견되었으며 이들의 내용을 분석한 결과는 표 5와 같다. 오탈자 외에 영문 키워드에 'or'를 사용한 경우도 있었다.

표 4. 키워드 쌍의 의미 비교 결과

구분	키워드 쌍수	구성비(%)
적합	1,034	84.34
이해가능	108	8.81
부적합	84	6.85
계	1,226	100.00

표 5. 부적합 이유

구분	키워드 쌍수	구성비(%)	샘플대비(%, n/1226)
순서 불일치	38	45.24	3.10
의미 부적합	42	50.00	3.43
오탈자 등	4	4.76	0.33
계	84	100.00	6.85

Ⅲ. 결론

저자 키워드를 조사한 결과 최다 키워드를 부여한 경우는 19개이며 없는 경우도 10%에 가깝다. 표 4에서 보는 바와 같이 평균 4개 이상의 키워드를 부여하고 있다. 대부분 학회의 발행 정책에 따라 결정되므로 학문 분야별로 저자가 키워드를 부여하는데 있어 큰 차이는 없어 보인다.

서론에서도 언급한 바와 같이 논문의 키워드는 학술용어로서의 가치와 의미가 있어 신중하게 선정하고 최소한 국문과 영문을 동수로 같은 순서로 또는 '국문(영문)'과 같은 형식으로 작성할 필요가 있다. 저자 키워드를 '웹수집' 용어로 제공하고 있는 사례도 있으나 현재와 같은 상태의 저자 키워드를 직접 확인하지 않고 기계적으로 처리하여 사용하는 것은 문제가 있다. 즉, 약 7%정도의 오류가 발생할 가능성이 있다. 키워드의 수는 이번 조사 결과 평균 4개 이상인 것으로 보아 5개 내외가 가장 적당할 것이다.

2) http://www.naver.com
3) http://www.google.com