정보관리연구, vol.39, no.3 2008, pp.163-177

NDSL 검색 질의어와 기술용어간의 관계에 대한 분석적 연구

A Relation Analysis between NDSL User Queries and Technical Terms

강 남 규* · 조 민 희** · 권 오 석***
Nam-Gyu Kang · Min-Hee Cho · Oh-Seok Kwon

차 례

초 록

본 논문에서는 NDSL을 검색하기 위해 이용자가 입력하는 검색 질의어를 대상으로 질의어에 사용되는 키워드와 학술지에서 추출한 기술용어와의 관계를 분석하고자 한다. 관계 분석을 위해 사용된 키워드는 17개월 동안의 NDSL 검색 질의어에서 추출한 약 83만3,000개, 기술용어는 NDSL, INSPEC, FSTA 3개 영문 학술지 데이터베이스 약 4,100만건에서 추출한 약 97만5,000개이다. 그리고 분석에 사용된 키워드와 기술용어는 2어절 이상의 영어 단어이며, 이들 간의 관계 분석은 키워드와 기술용어 간의 일치성, 연관성, 기술용어에 대한 빈도 분석 등이다.

키 워 드

검색 키워드 추출, 기술용어 추출, 관계 분석

* 한국과학기술정보연구원 정보시스템개발팀 선임연구원
(Senior Researcher, Information System Development Team, KISTI, ngkang@kisti.re.kr)
** 한국과학기술정보연구원 정보시스템개발팀 연구원
(Researcher, Information System Development Team, KISTI, mini@kisti.re.kr)
*** 충남대학교 컴퓨터공학과 교수
(Professor, Computer Science & Engineering, Chungnam National University, oskwon@cnu.ac.kr)
• 논문접수일자 : 2008년 8월 18일
• 게재확정일자 : 2008년 9월 9일

ABSTRACT

In this paper, we analyzed the relationship between user query keywords that is used to search NDSL and technical terms extracted from NDSL journals. For the analysis, we extracted about 833,000 query keywords from NDSL search logs during nearly 17 months and approximately 41,000,000 technical terms from NDSL, INSPEC, FSTA journals. And we used only the English noun phrase in extracted those and then we did an experiment on analysis of equality, relationship analysis and frequency analysis.

KEYWORDS

Query Extraction, Tech Terminology Extraction, Relation Analysis

1. 서론

검색 질의어로 빈번하게 사용된 키워드가 최근 또는 그 이전부터 이슈화되어 활발히 진행되고 있는 연구 분야 또는 특정 기술 등을 찾는데 도움이 될 수 있을까? 또한 검색 질의어로 사용된 수많은 키워드 중에서 특정 기술을 정확히 표현하는 것은 얼마나 될까? 본 논문은 위와 같은 질문을 해결하기 위하여 검색 질의어에 사용되는 키워드와 학술지에서 추출한 기술용어를 대상으로 이들 간의 연관성, 기술용어에 대한 빈도수 분석 등을 실험한다.

기술용어는 전문용어와 동일한 의미를 갖고 있으며, 특정 기술의 개념을 표현하기 위한 언어적 기호라고 정의할 수 있다. 즉, 한 특정 분야의 개념적 정보와 표현의 총체를 용어, 코드, 그래픽 또는 비언어적 기호 및 정의 혹은 다른 서술적 표현을 통하여 나타낸 것이다. 전문용어 필요한 이유는 전문용어가 없이는 전문적 의사소통이 불가능하고, 전문적 의사소통이 없으면 지식이전이 불가능하다. 또한 지식이전이 없으면 지적, 물질적 발전이 불가능하고 교육, 훈련 및 전문적 연구가 불가능하여 장기적인 기술발전이 불가능하기 때문이다. 또한 기술의 발전으로 인해 기술의 개념과 이를 지칭하는 용어가 지속적으로 새롭게 만들어지고 있다. 이러한 이유로 문서에서 특정 분야의 기술용어를 자동으로 추출하는 연구 개발 활동이 활발히 이루어지고 있다.

본 논문의 실험을 위하여 영어로 구성된 데이터베이스와 검색 질의어가 필요하며, 이를 위해 NDSL, INSPEC, FSTA 3개 영문 데이터베이스와, NDSL 로그 데이터베이스를 사용하였고, INSEPC과 FSTA 로그 데이터베이스는 수집 불가로 인하여 본 실험에서는 제외되었다. 검색 키워드는 NDSL 로그 데이터베이

스에 저장된 검색식에서 제목과 초록 항목에 입력된 키워드 중 2어절 이상의 영문 키워드만을 추출하였다. 기술용어는 NDSL, INSPEC, FSTA 3개 영문 데이터베이스로부터 2어절 이상의 영문 단어만을 추출하였으며, 추출 대상 항목은 제목과 초록 2개로 한정하였다.

본 논문에서는 1장은 서론, 2장은 NDSL 검색 질의어로부터 키워드 추출 과정 및 결과, 3장은 학술지 데이터베이스로부터 기술용어 추출 과정 및 결과, 4장은 추출된 키워드와 기술용어간의 분석한 결과를 설명한다. 그리고 4장에서는 결론과 향후 연구 과제를 말한다.

2. NDSL 검색 질의어 분석

2.1. NDSL

NDSL(National Digital Science Library)은 국내 학계, 연구계, 산업계의 모든 연구자들을 위한 해외 학술저널 및 프로시딩 포털로써 2008년 6월 현재 6만3,000여종의 학술저널과 19만 8,000여종의 프로시딩을 서비스하고 있다. 또한 NDSL이 보유한 전자원문링크 정보는 2만 3,000여종의 학술저널과 8,000여종의 프로시딩이다. 전체 NDSL 데이터베이스 중에서 응용과학 41%, 자연과학 16%를 차지하고 있어 과학기술분야에 해당하는 정보는 약 57%를 포함하며, 인문과학 14%, 사회과학 27%를 차지한다. 〈표 1〉은 2008년 6월까지의 NDSL 데이터베이스 구축현황을 나타낸 것이다.

또한 NDSL은 대학, 연구소, 기업체, 병원 등 학술연구기능을 수행하는 각 기관의 학술저널 콘텐트를 대폭 확충하기 위한 전자저널 공동구매컨소시엄(KESLI : Korean Electronic Site License Initiative)을 운영하고 있으며, KESLI에 참가하고 있는 기관에서 보유중인 인쇄저널의 공동이용을 위한 도서관 협력망을 운영하고 있고, 2008년 6월까지 96개의 컨소시엄과 271개의 참가기관수를 갖고 있다.

2.2. 검색 질의어로부터 다어절 키워드 추출

2007년 1월부터 2008년 5월까지 약 17개월 동안의 NDSL 검색 질의어 약 427만여건

〈표 1〉 NDSL 데이터베이스 구축 현황

구분		저널	프로시딩	합계
서지정보	전체	63,227	198,195	261,422 (종)
	전자원문링크정보	23,653	8,176	31,829 (종)
논문정보	전체	42,465,444	6,843,977	49,309,421 (건)
	전자원문링크정보	22,902,649	929,311	23,831,960 (건)

을 대상으로 약 83만 3,000건의 다어절 키워드를 추출하였다. 본 실험에서의 추출 대상 검색 질의어는 NDSL 전체 분야 즉, 과학기술분야를 포함한 인문과학, 사회과학 등을 대상으로 검색한 질의어이며, 추출 방법은 〈그림 1〉과 같은 단계로 진행되고 각 단계별 추출 건수를 표기하였다.

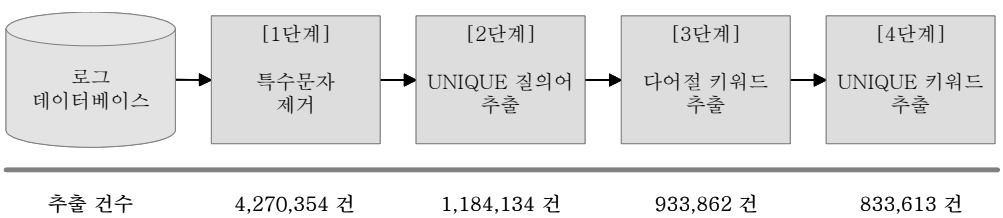
1단계는 로그 데이터베이스로부터 NDSL 검색 질의어를 추출하며, 연산자를 제외한 특수문자 제거 및 다국어에 대한 캐릭터 셋을 변경한다. 2단계는 1단계의 결과에서 유니크 질의어를 추출한다. 〈표 2〉는 2단계에서 추출된 검색 질의어의 빈도별 건수를 나타낸 것이다. 〈그림 1〉에서 전체 검색 질의어 427만여건에서 유니크 질의어는 118만4,000여건으로 약 73%의 질의어가 반복됨을 알 수 있다. 〈표 2〉 분포도에서 알 수 있듯이 질의어 빈도수가 1회인 경우 약 59%, 2회에서 10회인 경우 35%를 차지하고 있어 10회 이내의 빈도수를 갖는 질의어는 전체의 약 94.6%로 나타났다. 그러나 2007년 1월부터 2008년 5월까지의 질의어 중에서 11회 이상 반복된 검색의 경우, 질의어는 약 6만 2,000개, 검색 횟수로는 약 206만1,000건이나 된다는 것을 알 수 있다. 이것은 전체 검색 중에서 5.3%의 질의어가 48.2%의 검색 횟수를 차지한다고 말할 수 있다.

〈표 3〉은 빈도수 상위 15개의 검색 질의어를 나열한 것이다. 상위 15개 검색 질의어 중에서 분석 대상이 되는 2어절 이상의 키워드는 5개로 나타났으며, 대부분 이용자들이 1개 단어를 검색으로 활용하는 것을 예측할 수 있다.

NDSL 검색 시스템에서는 이용자에게 제목, 초록, 년도, 저자 등 다양한 검색 항목을

〈표 2〉 검색 질의어 빈도별 건수

빈도수		1	2 ~ 10	11 ~ 100	101 ~ 200	201 ~ 300	301 이상
질의어	건수	701,544	419,549	61,005	1,457	262	217
	분포율	59.25	35.43	5.15	0.12	0.02	0.02
검색	건수	701,544	1,507,786	1,413,913	195,955	63,459	387,697
	분포율	16.43	35.31	33.11	4.59	1.49	9.08



〈그림 1〉 다어절 키워드 추출 단계

제공하는데, 3단계에서는 제목과 초록 항목에 입력된 단어를 대상으로 다어절 영문 키워드를 추출한다. 그리고 질의어 연산자(AND, OR, WITHIN 등)는 제거하고, 추출된 결과는 분석 대상으로 선정하여 기술용어 매칭에 사용한다. 본 논문에서는 영문 다어절 키워드만을 대상으로 하는데, 한글로 구성된 키워드를 제외시키는 이유는 다음과 같다. 일반적인 기술용어에 있어서, 단일명사로 이루어진 것보다 복합명사로 이루어진 용어 또는 여러 개의 단어가 하나의 기술용어를 구성하는 경우가 많은데, 다어절 한글 용어의 경우에는 띄어쓰기, 영어식 발음표기 등의 문제로 인한 정확한 용어 매칭이 어려울 수 있기 때문이다. 그리고 1어절 단어의 경우는 문장 또는 문맥의 이해 없이 일반어휘와 기술용어의 정확한 구분이 어렵기 때문에 1어절 단어에 대해서도 분석 대상에서 제외시킨다. 〈표 4〉는 3단계의 처리 과정인 검색 질의어에 대한 키워드 추출 및 선정 결과를 예로 나타낸 것이다.

〈표 3〉 상위 15개 검색 질의어

빈도수	질의어	빈도수	질의어	빈도수	질의어
1,798	computer	1,221	gcms	907	zno
1,547	oled	1,032	xanthomonas oryzae	871	cerebral palsy
1,516	propolis	1,025	gc ms	870	random
1,508	work environment	962	creativity	810	cnt
1,452	rfid	918	solar cell	790	led

〈표 4〉 NDSL 검색 질의어에서 키워드 추출 결과

#	검색 질의어	
	추출 결과	선정
1	(bacillus〈in〉title 〈and〉 thuringiensis〈in〉title)	
	≫ bacillus thuringiensis	○
2	(bacillus〈in〉title 〈and〉 thuringiensis〈in〉title) 〈and〉 (Pubdate〉20040000)	
	≫ bacillus thuringiensis	○
4	(bacillus〈in〉title 〈and〉 thuringiensis〈in〉title) 〈and〉 (sourthern〈in〉abstract 〈and〉 blot〈in〉abstract)	
	≫ bacillus thuringiensis	○
	≫ sourthern blot	○
5	(bacillus〈in〉title 〈and〉 thuringiensis〈in〉title) 〈and〉 (spectrometer〈in〉abstract)	
	≫ bacillus thuringiensis	○
	≫ spectrometer	×

4단계는 3단계의 결과에서 유니크 키워드를 추출한다. 3단계에서는 〈표 4〉 #5의 예와 같이 제목과 초록 항목을 구분하여 키워드를 추출하지만, 4단계의 유니크 키워드 추출시에는 제목과 초록 항목을 구분하지 않는다. 〈표 5〉는 4단계의 결과인 키워드 빈도별 건수를 나타낸 것이다. 〈표 6〉은 빈도수 상위 20개 키워드를 나열하였으며, 추출된 각 키워드가 의미 있는 용어인지 알아보기 위해 구글(Google)과 위키피디아(Wikipedia)를 검색한 결과를 표시하였다.

3. 기술용어 추출

3.1 데이터베이스

본 논문에서는 영문 학술지 데이터베이스를 대상으로 다어절 기술용어를 추출하기 위하여 FSTA(Food Science and Technology Abstracts), INSPEC(Information Services for the Physics and Engineering Communities), NDSL 3개의 대용량 데이터베이스를 사용한다. FSTA 데이터베이스는 전 세계 식품과학, 식품공학, 식품기술 및 음식과 관련된 식품 영

〈표 5〉 키워드 빈도별 건수

빈도수		1	2 ~ 10	11 ~ 100	101 ~ 200	201 ~ 300	301 이상
키워드	건수	472,360	315,493	44,421	1,050	168	121
	비율	56.66	37.85	5.33	0.13	0.02	0.01
출현	건수	472,360	1,115,101	1,034,390	141,243	40,784	64,886
	분포율	16.47	38.87	36.06	4.92	1.42	2.26

〈표 6〉 빈도수 상위 20개 키워드

빈도수	키워드	G	W	빈도수	키워드	G	W
2,091	bacillus thuringiensis	○	○	910	social support	○	○
1,926	work environment	○	×	892	microbial fuel cell	○	○
1,746	solar cell	○	○	870	liquid crystal	○	○
1,694	cerebral palsy	○	○	839	breast cancer	○	○
1,599	carbon nanotube	○	○	760	internet addiction	○	○
1,598	fuel cell	○	○	741	sensory integration	○	○
1,588	gc ms	○	○	728	business model	○	○
1,034	xanthomonas oryzae	○	○	722	bipolar plates	○	×
938	information literacy	○	○	693	six sigma	○	○
934	bipolar plate	○	×	655	lc ms	○	○

G : Google W : Wikipedia

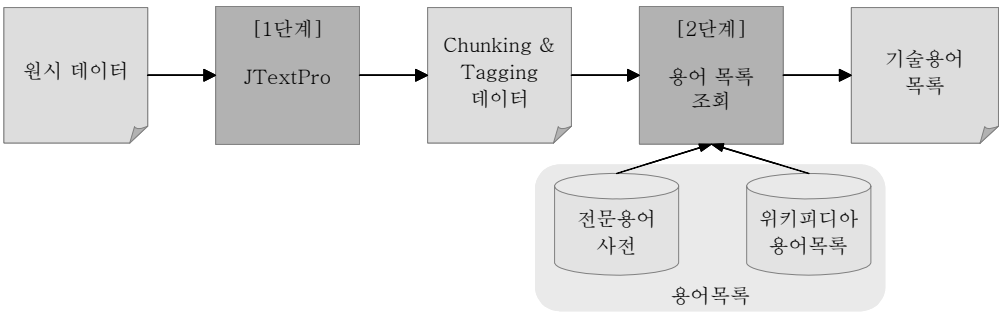
양학에 관한 광범위한 초록정보를 제공하는 데이터베이스로 기술용어 추출을 위해 76만 3,300건의 레코드를 사용하였다. INSPEC 데이터베이스는 컴퓨터, 제어공학, 전기, 전자, 물리 등에 관한 학술 정보를 제공하는 데이터베이스로 본 논문에서는 955만 6,847건을 사용하였다. NLDS 데이터베이스는 전체 데이터 중에서 3,109만4,899건을 사용하여, 전체 대상 데이터 4,141만5,046건을 사용하였다.

3.2 기술용어 추출

3개 영문 학술지 데이터베이스의 데이터 4,141만5,046건을 대상으로 기술용어를 추출하였으며 그 과정은 〈그림 2〉와 같다.

원시 데이터는 학술지 서지 정보로 제목, 저자, 발행일자, 키워드, 초록 등 많은 항목을 갖고 있으나, 3개 영문 데이터베이스에 공통으로 포함된 항목 중에서 기술용어가 포함될 수 있는 항목인 제목과 초록 정보만을 활용하였다. 그리고 키워드 정보는 아래에 설명했듯이 기술용어 판단을 위해 사용하는 용어목록으로 활용하기 때문에 기술용어 추출에서 제외시켰다.

1단계에서는 원시 데이터의 제목과 초록 정보를 JTextPro를 활용해 청킹과 품사 태깅을 수행한다. JTextPro는 Java 기반의 텍스트 처리를 위한 도구로써 자연어 처리, 텍스트 또는 웹 데이터 마이닝, 정보 추출 등에 활용할 수 있도록 개발되었다. 2단계에서는 청킹과 품사 태깅이 완료된 데이터에 대해서 기술용어 여부를 판단하는데, 추출된 다어절의 데이터가 〈그림 2〉의 용어목록에 있는지를 조사하여 해당 용어가 존재하는 경우에 기술용어로 판단하고, 기술용어로 판단된 것에 대해서는 별도의 태깅(B:시작, E;끝, TT:기술용어, KW:키워드) 작업을 수행한다. 여기서 사용하는 용어목록은 3개의 용어집으로 전체 용어 수는 285만274개이며, 각 용어집은 원시 데이터 중 키워드 항목에서 2어절 이상의 단어를 추출하여 만든 15만8,621개 용어와 전문용어언어공학연구센터(KORTERM : Korea Termi-



〈그림 2〉 기술용어 추출 단계

nology Research Center for Language and Knowledge Engineering)의 전문용어사전에서 추출한 25만3,603개 전문용어, 그리고 위키피디아 용어집 중에서 2어절 이상의 243만8,050개 용어를 사용하였다. 〈표 7〉은 기술용어 추출 과정을 예로써 나타낸 것이다.

추출이 완료된 기술용어집합은 〈그림 3〉과 같이 문헌 종류(TP), 기술용어(TT)와 그 용어가 나타난 논문의 발행년도(YE), 년도별 빈도수(CNT), 논문의 관리번호(ANS) 5개 항목으로 구성되어 전체 1억9,315만1,686건이 추출되었으며, 유니크 기술용어만으로는 97만5,158건이 추출되었다.

기술용어별 빈도수를 알아보기 위하여 1998년부터 2007년까지 10년간 각 년도별 기술용어 빈도수가 상위 1,000개인 것들을 합하였더니 1,430개의 기술용어들이 나타났다. 〈표 8〉은 그 중 일부를 보이고 있다. 출력순서는 순위와는 상관없으며, 빈도수에 '◇'로 표기되어 있는 부분은 기술용어가 해당 년도에 상위 1,000개에 포함되지 않은 경우를 의미한다. 대부분의 기술용어에서 2007년에 빈도수가 하락함을 보이는데, 〈표 8〉의 데이터 구축 건수에서 예측할 수 있듯이 2007년 데이터의 확

〈표 7〉 기술용어 추출 결과

단계	데이터
원본	Simultaneous Measurement of Bacillus thuringiensis Cry1Ab and Cry3B Proteins in Corn Extracts
1단계	Simultaneous/JJ/B-NP Measurement/NN/I-NP of/IN/B-PP Bacillus/NN/B-NP thuringiensis/NN/I-NP Cry1Ab/NN/I-NP and/CC/O Cry3B/NN/B-NP Proteins/NNS/I-NP in/IN/B-PP Corn/NN/B-NP Extracts/NNS/I-NP
2단계	Simultaneous/JJ/B-NP/B-KW Measurement/NN/I-NP/E-KW of/IN/B-PP Bacillus/NN/B-NP/B-TT thuringiensis/NN/I-NP/E-TT Cry1Ab/NN/I-NP and/CC/O Cry3B/NN/B-NP Proteins/NNS/I-NP/S-TT in/IN/B-PP Corn/NN/B-NP/S-TT Extracts/NNS/I-NP/S-TT
결과	≫ simultaneous measurement ≫ bacillus thuringiensis

```
@NEW_DOCUMENT
#TP=j
#TT=bacillus thuringiensis
#YE=1993
#CNT=190
#ANS=FSTA199300013468|FSTA199300016795||JAFO000025744482...
```

〈그림 3〉 기술용어집합 샘플데이터

〈표 8〉 최근 10년간 년도별 기술용어 빈도수

기술용어	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	합계
데이터 구축 건수 (× 1,000)	1,667	1,121	1,780	1,761	1,816	1,866	1,904	1,933	1,952	1,360	18,160
thin films	11,829	11,935	12,496	13,335	14,420	14,485	14,265	12,505	17,573	10,224	133,067
	0.71%	1.06%	0.70%	0.76%	0.79%	0.78%	0.75%	0.65%	0.90%	0.75%	
room temperature	10,515	11,480	11,005	11,137	12,241	12,392	12,274	10,782	16,113	9,831	117,770
	0.63%	1.02%	0.62%	0.63%	0.67%	0.66%	0.64%	0.56%	0.83%	0.72%	
magnetic field	11,939	12,064	12,601	12,010	12,458	11,964	11,072	8,893	14,717	6,869	114,587
	0.72%	1.08%	0.71%	0.68%	0.69%	0.64%	0.58%	0.46%	0.75%	0.51%	
x-ray diffraction	6,589	7,014	7,301	7,460	8,708	8,909	9,472	8,399	12,813	8,568	85,233
	0.40%	0.63%	0.41%	0.42%	0.48%	0.48%	0.50%	0.43%	0.66%	0.63%	
carbon nanotubes	〈577〉	839	1,131	1,867	2,345	2,913	3,394	3,256	4,994	3,099	24,415
	0.03%	0.07%	0.06%	0.11%	0.13%	0.16%	0.18%	0.17%	0.26%	0.23%	
fuel cell	〈452〉	〈427〉	〈681〉	759	1,067	1,303	1,743	1,691	2,967	1,516	12,606
	0.03%	0.04%	0.04%	0.04%	0.06%	0.07%	0.09%	0.09%	0.15%	0.11%	
hepatitis c virus	867	1,264	1,206	1,087	1,034	934	817	786	756	〈661〉	9,412
	0.05%	0.11%	0.07%	0.06%	0.06%	0.05%	0.04%	0.04%	0.04%	0.05%	

보가 완료되지 않은 것으로 판단된다.

추출된 최근 10년간 년도별 기술용어 빈도수로부터 기술 추이 예측도 가능한데, 'carbon nanotubes'과 'fuel cell'의 경우 시간이 경과할수록 연구 활동이 활발해짐과 'hepatitis c virus'의 경우 1999년을 정점으로 지속적인 하향 및 현상 유지 상태임 예측할 수 있다.

4. 검색 키워드와 기술용어 분석

4.1 키워드와 기술용어간의 일치성 및 연관성 분석

키워드와 기술용어 간의 일치성 분석을 위한 실험 방법은 다음과 같다. 추출한 기술용어는 데이터베이스에 저장시키고, 키워드를 변수로 활용하여 두 용어간의 일치여부를 판단하였으며, SQL의 equal 연산자로 매치되는 경우 exact match, exact match가 아니면서 like 연산자로 매치되는 경우 related match, 앞의 두 경우가 모두 아닐 경우 mismatch로 판단한다. Related match는 exact match는 아니지만, 연관된 기술용어가 존재한다고 판단할 수 있는데, 예를 들어 키워드가 'signal to noise'일 때, 기술용어가 'signal to noise ratio', 'signal to noise statistic'등의 경우라면 related match로 판단한다.

〈표 9〉는 검색 질의어에서 추출한 키워드 중 빈도수 10회 이상의 키워드 5만1,749건을 대상으로 1980년부터 2007년까지의 기술용어와 일치 여부를 실험한 결과이다. 실험 결과로써 exact match와 related match는 키워드와 기술용어간의 관계를 지을 수 있다고 판

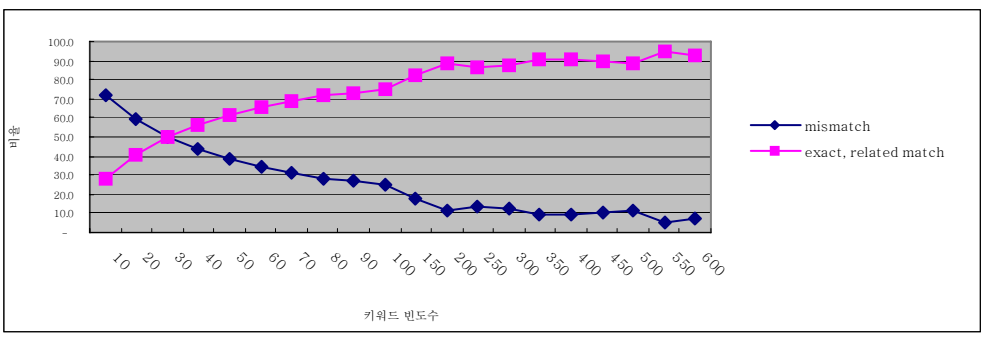
단하므로, 약 29%가 키워드와 기술용어가 일치한다고 할 수 있다. 그러나 mismatch의 경우 약 71%로 높은 비율을 차지하므로, 이것에 대한 분석이 필요할 것이다.

Mismatch가 발생하는 원인은 정확하지 않은 검색 질의어를 입력하거나, 기술용어를 추출할 때 사용한 사전에 해당 키워드가 포함되지 않는 경우 즉, 기술용어로써 사전에 등록이 되지 않는 경우로 예상된다. 부정확한 검색 질의어 입력의 경우에 대한 예로써, 검색 질의어로 'toughen epoxy', 'emotion intelligence'가 입력되었더라도 검색시스템은 스테밍(stem-ming)을 통해 검색 질의어를 처리하거나 색인어를 구축하므로 'toughened epoxy'와 'emotional intelligence'과 같은 정확한 정보를 제공할 수 있다. 하지만 본 논문에서는 키워드나 기술용어에 대한 추출 및 분석 과정에서 스테밍이 고려되지 않았기 때문에 'toughen epoxy'나 'emotion intelligence'는 mismatch로 분류되어 위와 같은 상황이 발생한 것이다. 다른 mismatch 원인으로는 검색 질의어로부터 키워드 추출시 오류가 있을 것이다. 이러한 오류들은 AND, OR 등과 같은 검색 연산자를 무시함으로써 발생할 수도 있고, 'epoxy toughen'과 같이 여러 개의 키워드를 조합하는 과정에서 순서가 바뀜으로 인해 발생할 수도 있을 것이다.

〈그림 4〉는 기술용어와 키워드간의 연관성을 살펴보기 위하여 키워드 빈도수를 10회 이상부터 600회 이상까지 조절하면서 exact match와 related match, 그리고 mismatch 비율을 실험한 결과이다. X축은 키워드 빈도수, Y축은 비율을 나타낸 것으로, 이 실험으

〈표 9〉 키워드와 기술용어 간의 일치율

구분	exact match	mismatch	related match	합계
건수	12,599	36,979	2,171	51,749
비율	24.3%	71.5%	4.2%	100

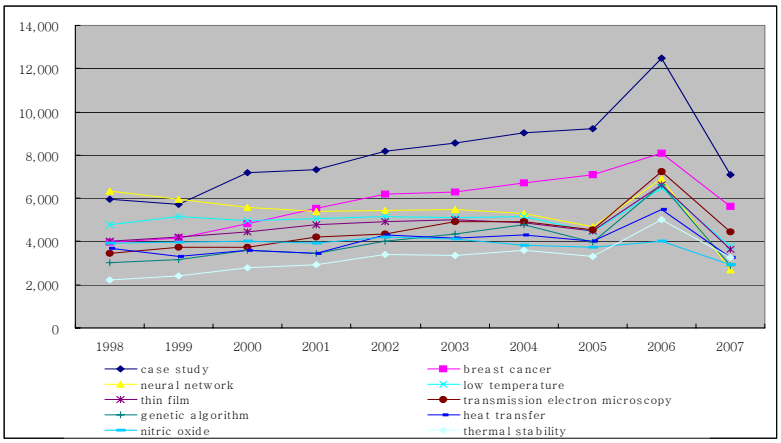


〈그림 4〉 키워드 빈도별 mismatch 비율

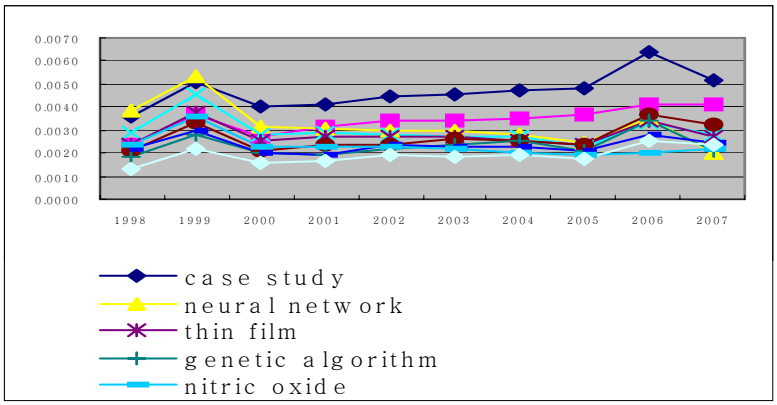
로 알 수 있듯이 빈번하게 질의어로 사용되는 키워드는 기술용어일 확률이 높다는 것을 알 수 있다. 그리고 키워드 빈도수 150회 이상을 기준으로 기술용어와의 연관성이 85.2% 이상 나타남을 알 수 있다.

4.2 년도별 빈도수 분석

키워드 빈도수 100회 이상의 데이터를 대상으로 1998년부터 2007년까지 최근 10년간의 기술용어와의 일치성 검토한 결과를 이용하여 년도별 빈도수 분석을 수행하였다. 〈그림 5〉는 기술용어 빈도수 상위 10개를 출력한 것이며 〈그림 6〉는 각 빈도수를 년도별 데이터베이스 구축건수로 나누어 빈도수에 대한 분포율을 나타낸 것이다. 〈그림 6〉에서 볼 수 있듯이 대부분의 기술용어에서 분포율이 크게 변동 없이 일정함을 보이거나 다소 감소하지만, 'breast



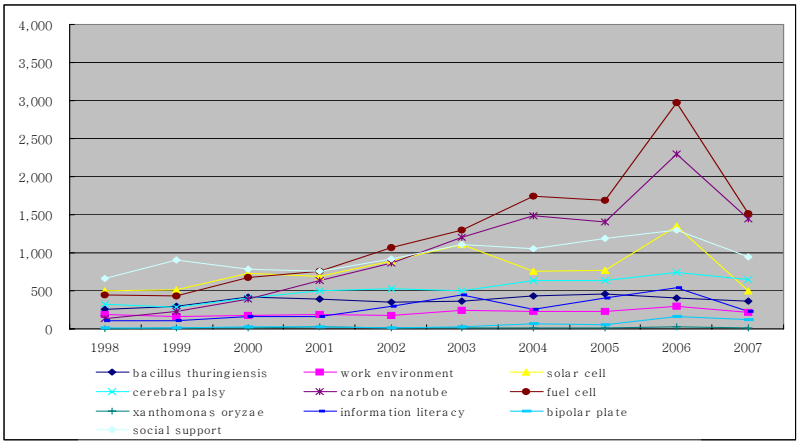
〈그림 5〉 기술용어 빈도수 상위 10개 (exact match)



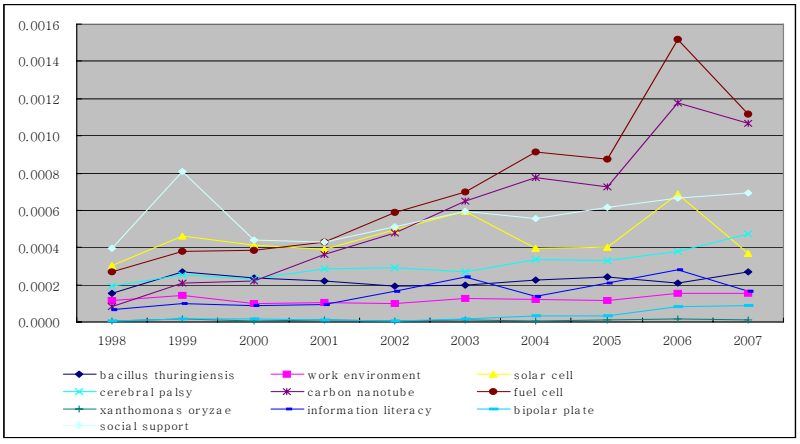
〈그림 6〉 기술용어 빈도수 분포율 (exact match)

cancer'와 'case study' 용어는 지속적으로 증가함을 알 수 있다. 이 그래프에서 특이한 점은 1999년과 2006년의 분포율이 갑자기 변하는 모양을 띄고 있는데, 1999년의 경우는 다른 년도보다 데이터 구축건수가 매우 적음으로 인해 이러한 현상이 나타남을 추측할 수 있다. 하지만 2006년의 경우에는 데이터 구축건수도 다른 년도와 비슷하여 1999년과 동일한 현상으로 판단할 수 없고, 기술용어의 출현이 급작스럽게 증가했다고 예상할 수 있을 것이다.

〈그림 7〉은 키워드 빈도수 상위 10개에 대한 것이며 〈그림 8〉은 빈도수에 대한 분포율을 나타낸 것이다. 〈그림 8〉에서 대부분의 기술용어에서 빈도수가 증가하는 모습을 보이고 있다. 게다가 'fuel cell', 'carbon nanotube'는 급격히 증가하는 모습을 보이고 있다. 〈그



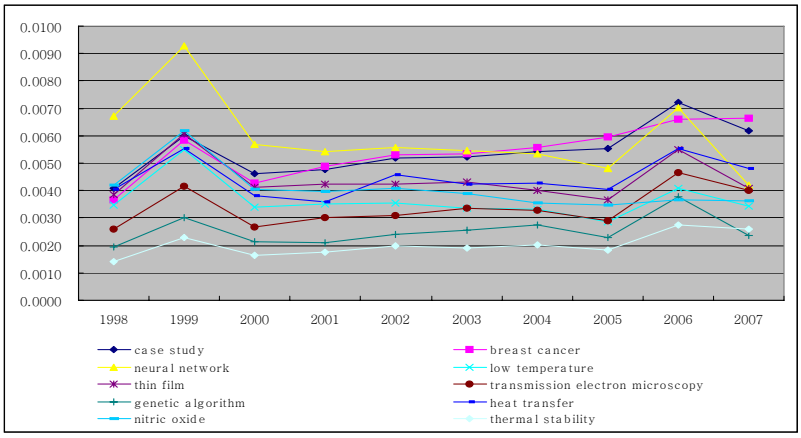
〈그림 7〉 키워드 빈도수 상위 10개 (exact match)



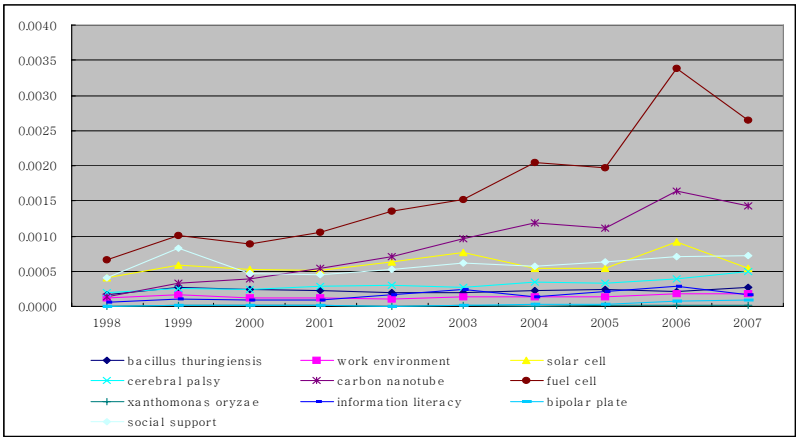
〈그림 8〉 키워드 빈도수 분포율 (exact match)

림 6〉과 〈그림 8〉의 그래프를 비교해 볼 때, 키워드와 연관된 기술용어의 출현 빈도가 더 증가하는 모습을 볼 수 있는데, 검색시스템 이용자가 최근에 많이 검색하는 키워드는 해당 분야의 연구 개발 활동이 활발하다는 것을 예측할 수 있을 것이다.

〈그림 5〉부터 〈그림 8〉까지는 키워드와 기술용어간의 정확한 매치 즉 exact match에 대한 결과이고, 〈그림 9〉와 〈그림 10〉은 related match에 대한 결과를 그래프로 나타낸 것으로 〈그림 9〉는 기술용어 상위 10개에 대한 빈도수 분포율, 〈그림 10〉은 키워드 상위 10개에 대한 빈도수 분포율을 나타낸 것이다. 두 그래프를 비교해보면 위에서 설명한 바와 같이 키워드와 연관된 기술용어의 출현 빈도가 더 증가함을 볼 수 있다.



〈그림 9〉 기술용어 빈도수 분포율 (related match)



〈그림 10〉 키워드 빈도수 분포율 (related match)

5. 결론 및 향후 연구 과제

기술용어는 개념간의 관계를 기반으로 지식을 표현하는 것으로써 과학기술 정보의 정리, 기술이전의 기초, 언어의 인덱싱, 시소러스, 분류체계 등에 기초가 된다. 본 논문에서는 이러한 기술용어와 검색 질의어와의 관계를 분석하였다.

NDSL 로그 데이터베이스에 저장된 검색식에서 제목과 초록항목에 입력된 키워드와 NDSL, FSTA, INSPEC 영문 학술지 데이터베이스에서 추출한 기술용어간의 관계 분석을 수행하였다. 추출과 분석에 사용된 키워드와 기술용어는 2어절 이상의 영문을 사용하였고, 이들 간의 관계 분석은 키워드와 기술용어간의 연관성, 일치성, 빈도수 분석 등이다.

NDSL 로그 데이터베이스에서 최근 17개월간의 NDSL 검색 질의어 약 420만건으로 부터 2어절 이상의 영문 유티크 키워드 약 83만 3,000건과 3개 영문 데이터베이스에서 약 410만건의 학술지 서지 정보로부터 제목과 초록 항목을 대상으로 2어절 이상의 영문 유니크 기술용어 약 97만5,000건을 추출하였다.

본 논문에서는 검색 질의어로부터 추출한 키워드의 빈도수를 조절하면서 기술용어와의 일치성 및 연관성을 조사하였다. 일치성 및 연관성은 exact match, related match, mis-match로 구분하여 실험하였으며, 150회 이상 반복된 검색 키워드는 기술용어일 확률이 약 85% 이상 된다는 결과를 얻어냈다.

그리고 키워드 빈도수 100회 이상의 데이터를 대상으로 최근 10년간의 학술지에서 추출한 기술용어와 비교를 통해 빈도수 분석을 수행하였다. 그 결과 빈번하게 검색되는 키워드의 빈도수 분포율이 그렇지 않은 경우와 비교할 때 급격한 증가를 보이고 있음을 알 수 있다. 이는 최근에 많이 검색하는 키워드는 해당 분야의 연구 개발 활동이 활발하다는 것을 예측할 수 있다.

일반적인 연구 동향 파악은 특정 기술 또는 분야에 한정하는 경우가 대부분이지만, 본 논문을 통해 분석된 결과는 다양한 기술 또는 다양한 분야에 대한 기술 추이 분석 등 연구 동향을 파악하는데 도움이 될 수 있을 것이다. 또한 높은 빈도수를 갖는 키워드를 분석한다면 최근 중요하게 이슈화되는 기술들에 대한 동향 파악도 가능할 것이다.

검색 키워드와 기술용어간의 관계를 좀 더 세밀하게 파악하기 위하여 1어절의 영어 단어에 대한 추출 및 분석이 필요할 것이다. 하지만 1어절 단어가 기술용어인지 일상용어인지에 대한 판단은 단어 자체로만 판단할 수 있는 것이 아니라 단어가 포함된 문헌의 분야, 문장 또는 문맥 등의 다양 주변 요소들을 함께 분석해야하는 과정이 필요하다. 이러한 분야, 문장 또는 문맥 등 분석을 통한 1어절 영어 단어와의 연관성 파악은 향후 연구과제로 남긴다.

참고문헌

강정미. 1999. 전문용어사전 표제어 기술형식에 대한 연구.『한국정보관리학회 학술대회 논문집』, 6: 39-42.

박소연, 이준호. 2007. 웹 검색 분야에서의 로그 분석 방법론의 활용도.『한국문헌정보학회지』, 41(1): 231-242.

오종훈, 이경순, 최기선. 2002. 분야간 유사도와 통계기법을 이용한 전문용어의 자동 추출.『정보과학회논문지: 소프트웨어 및 응용』, 29(3/4): 258-269.

오종훈. 2000.『전문분야 사전과 코퍼스 및 외래어 인식에 기반한 전문용어 추출』. 석사학위논문, 한국과학기술원, 전산학과.

Alan L.Porter, Scott W. Cunningham. 2004. Tech Mining : Exploiting New Technologies for Competitive Advantage. New York: John Wiley & Sons, Inc.

Michael J. Cafarella, Christopher Re, Dan Suciu, Oren Etzioni, Michele Banko. 2007. "Structured Query of Web Text." 3rd Biennial Conference on Innovative Data Systems Research.

Michael J. Cafarella, Oren Etzioni. 2005. "A Search Engine for Natural Language Applications." International World Wide Web Conference Committee.

Doug Downey, Stefan Schoenmackers, Oren Etzioni. 2007. "Sparse Information Extraction : Unsupervised Language Models to the Rescue." ACL2007.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, Oren Etzioni. 2007. "Open Information Extraction from the Web." IJCAI-07 Proceedings, 2670-2676.

JTextPro. 〈http://jtextpro.sourceforge.net〉.

KORTERM. 〈http://www.korterm.or.kr〉.

KnowItAll. 〈http://www.cs.washington.edu/research/knowitall〉.