http://dx.doi.org/10.3745/KIPSTD.2012.19D.2.133

용어 자동분류를 사용한 검색어 범주화의 분석적 고찰

이 태 석[†] · 정 도 헌[††] · 문 영 수[†] · 박 민 수[†] · 현 미 환[†]

요 약

검색 창을 통해 입력된 검색어는 정보이용자가 의미 있는 자료를 찾아내는 적극적인 활동의 산물이다. 따라서 검색로그는 정보이용자의 관심사항을 알 수 있는 중요한 분석 데이터이다. 본 연구의 목적은 입력한 검색어의 범주화 결과와 엑세스한 문서의 범주가 어느 정도 유사한 상관관계를 가지는지 분석적으로 고찰해보는 것이다. KISTI(한국과학기술정보연구원)의 NDSL(과학기술정보센터) 사이트의 2009년 검색로그의 검색세션을 식별하고 검색세션단위로 검색어와 이용 자료를 추출한 후, 검색어에 대해 어떤 주제 분류에 속하는 용어인지 자동분류기로 식별한 결과가 실제 이용한 자료의 주제 분야와 잘 맞는지 비교하였다. 그 결과 상위 100개 검색어 분류에 대한 유사도 평균이 58.8%로 파악되었다. 결국 전체적인 유사도는 58.8%이하이며, 관련 연구에서 수행한 자료의 자동분류 검색성능 전문가 평가 결과인 76.8%에 비해 낮다. 이것은 검색어로 쓰인 용어가 다른 연구 분야의 관심 용어로 새롭게 주목 받고 있기 때문이라는 사실을 알 수 있었다.

키워드 : 용어 자동분류, 검색로그, 질의어 분석, 유사도

An Analytic Study on the Categorization of Query through Automatic Term Classification

Tae-seok Lee[†] · Do-heon Jeong[††] · Young-su Moon[†] · Minsoo Park[†] · Mi-hwan Hyun[†]

ABSTRACT

Queries entered in a search box are the results of users' activities to actively seek information. Therefore, search logs are important data which represent users' information needs. The purpose of this study is to examine if there is a relationship between the results of queries automatically classified and the categories of documents accessed. Search sessions were identified in 2009 NDSL(National Discovery for Science Leaders) log dataset of KISTI (Korea Institute of Science and Technology Information). Queries and items used were extracted by session. The queries were processed using an automatic classifier. The identified queries were then compared with the subject categories of items used. As a result, it was found that the average similarity was 58.8% for the automatic classification of the top 100 queries. Interestingly, this result is a numerical value lower than 76.8%, the result of search evaluated by experts. The reason for this difference explains that the terms used as queries are newly emerging as those of concern in other fields of research.

Keywords : Automatic Term Classification, Search Log, Analysis of Query, Similarity

1. 서 론

KISTI(한국과학기술정보연구원)의 NDSL(국가과학기술정보센터)은 국내 최대의 과학기술정보 포털 사이트로 제공자 입장에서는 국내 과학기술 연구자들의 연구 분야는 항상 중요한 이슈로 받아들이고 있다. 이러한 이유는 연구 이슈가 많은 주제 분야에 대해 최신의 논문, 특허, 보고서 등에 대한 종합적 정보를 제공해야 하기 때문이다. NDSL 논문 정보에는 STEAK를 사용한 용어 자동분류(automatic term classification)기를 통하여 18개의 분야로 미리 분류하여 서비스하고 있다[4].

본 논문의 목적은 검색로그에 기록된 키워드 분석을 통하여 연구자들이 어떤 주제 분야에 대해 관심이 있는지 파악하고, 정보이용자가 선택하는 문서의 자동분류정보와 용어의 범주 사이의 유사도를 살펴보는 것이다. 즉, 논문을 기초로 한 용어 자동분류와 검색로그의 키워드 측면에서 바라본 분류체계에 대한 검증을 함으로써 검색결과에 대한 전문가의 평가결과와 비교하여 어느 정도 의미 있는지 검증해 보는 것이 필요하기 때문이다.

[8]에서 비정형 잡음요소들을 추출하고 범주자질을 전거어로 활용하여 유사어 사전을 이용한 모델을 제안하고 실험한 결과 분석에서 정확도는 87.05%, 재현율은 76.8% 조사되

† 정 회 원 : 한국과학기술정보연구원 NDSL서비스실 선임연구원
†† 정 회 원 : 한국과학기술정보연구원 소프트웨어연구실 선임연구원(교신저자)
논문접수 : 2011년 6월 24일
수 정 일 : 1차 2011년 11월 14일, 2차 2012년 1월 9일, 3차 2012년 2월 22일
심사완료 : 2012년 2월 28일

〈표 1〉 과학기술표준분류표

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
수학	물리학	화학	생명과학	지구과학	기계	재료	화학공정	전기전자	정보	통신	농림수산	보건의료	환경	에너지자원	원자력	건설교통	우주항공천문해양

었다. 재현율과 정확도에 대한 편차가 커 불안정한 효율을 보였다. 또한 [2]에서 용어 자동분류기 STEAK를 사용한 검색결과 평가는 전문가에 의해 수행하였다. 전문가 평가결과 76.8%로 나타났다.

관련 연구의 용어 자동분류에 대한 선행연구들이 실험결과의 평가에서 정확한 분류의 기준을 전문가가 미리 정해놓고 비교하거나 범주화의 결과에 대한 전문가 평가를 통하여 이루어 졌다[2]. 범주화에 대한 전문가 평가는 주관적인 견해와 지식의 부족으로 보다 일반적이고 객관적인 평가를 내리는 데 한계를 지니고 있는 문제가 있다. 이를 보완하고 검증할 수 있는 방법은 보다 많은 사람의 다양한 선택을 분석하는 것이다. 국내에서 가장 많은 과학기술정보를 제공하고 많은 연구자들이 사용하는 NDSL 사이트의 검색로그를 분석하여 범주화와 검색결과를 평가하는 것은 정보 이용자의 입장에서 적합한 범주화에 더 가깝게 다가갈 수 있을 것이다. 비록, 특이한 행동을 하는 일부 잘못된 이용 패턴이 있겠지만, 많은 사람들이 일반적으로 자주 이용하는 용어와 자료를 대상으로 제한하여 분석함으로써 객관적이고 일반적인 검색결과 평가를 얻을 수 있다.

본 연구를 통해 과학기술에 대한 이용자의 트렌드를 읽을 수 있으며, 무엇보다도 용어 자동분류가 새로운 용어 또는 기존 용어가 새롭게 다른 분야에서 주목 받을 때 해당 범주에 대응할 수 있어야 한다는 것을 확인하고자 한다.

2. 용어 자동분류를 사용한 검색어 범주 분석

검색로그 분석은 NDSL 사이트의 2009년 12개월 서비스 로그 4천 7백만건중 검색로그 2천 3백만건을 대상으로 하였다. 실험의 처리 과정은 검색로그의 방문 세션처리, 검색어 분리, 용어 자동분류, 용어-검색어 매칭, 검색어 주제 분류 검증 과정으로 나누었다. 주제 분류는 KISTEP(한국과학기술기획평가원) 2005년 <표 1> 과학기술표준분류 18개 대분류를 사용하였다.

2.1 검색로그 분석

일반적인 로그 분석은 웹서버에 사용자가 들어오는 순간부터 하나의 데이터에 접속(hit), 실제 이용자가 하나의 완성된 페이지를 보는 행위(view), 특정 사용자가 일정시간 내에 계속적으로 웹서버를 검색(search)하는 등 웹서버의 방문(visit) 데이터를 기반으로 어떤 목적에 맞도록 분석을 수행하는 계량적 방법을 말한다. 이와 같은 다양한 방문 데이터들이 통계분석의 대상이 될 수 있으며, 이를 바탕으로 해당 기관의 웹서버에 대하여 얼마나 많은 사람들이, 언제 방문하는지, 가장 오래 보는 자료와 가장 많이 보는 자료는 어떤 것인지 등 다양하고 의미 있는 정보들을 파악해 낼 수 있다[10].

검색로그 분석은 크게 두 가지 유형으로 구분할 수 있다. 첫 번째는 이용자가 검색을 위해 입력한 질의인 검색어만을 대상으로 분석하는 질의로그 분석 또는 검색어 로그 분석이다. 두 번째는 이용자가 입력한 검색어뿐만 아니라 검색 결과 중에서 이용자가 실제로 사용하기 위해 자료를 선택한 행위를 보여주는 클릭로그 데이터를 분석하는 클릭로그 분석 또는 트랜잭션 로그 분석이다[1].

질의로그 분석은 이용자가 검색을 위해 검색창에 입력한 검색어만을 대상으로 분석하는 방법이다. 주로 포털사이트를 대상으로 많은 연구가 이루어지고 있으며, 장기간에 걸친 방대한 자료를 바탕으로 이용자의 대략적인 검색 행태를 파악할 수 있다. 클릭로그 분석은 이용자가 실제로 관심을 가지고 찾고자 하는 정보는 단순히 검색창에 입력하는 검색어가 아니라 검색 결과에서 실제로 선택하고 이용한 자료를 분석 대상으로 삼으며, 이용자의 관심 주제를 파악할 수 있는 분석방법이다[5][6].

2.2 검색세션 처리

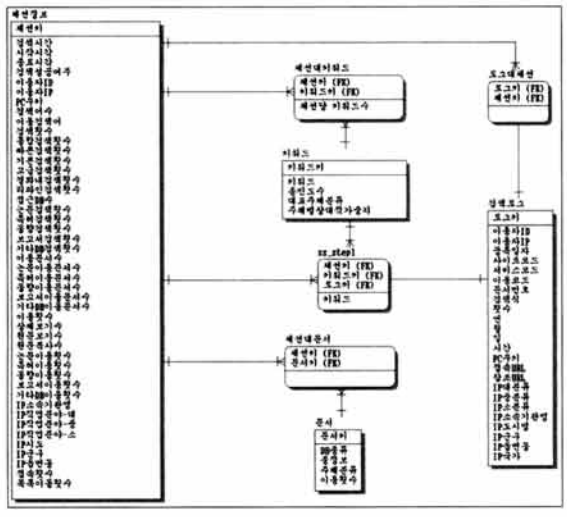
검색로그 분석에서 사용된 방법은 일반적인 로그 분석 방법과 동일하게 처리했다. 검색결과 목록 이동, 문서 상세 페

〈표 2〉 검색세션 처리 형태

항목명	설명
PCID	정보검색 이용자에 대한 영구 쿠키 ID, PC를 구분해줌.
검색세션	동일 이용자의 검색 스트림으로 로그 간격이 30분 이내로 이어지는 형태 (간격이 30분 이상이 되면 다음 검색세션으로 간주함.)
검색시간	검색세션 종료시각 - 시작시각
검색어수	검색세션에 나타난 중복 제거한 키워드 개수
이용검색어	검색세션에서 이용한 검색어 집합
검색횟수	통합검색횟수+빠른검색횟수+기본검색횟수+고급검색횟수+결과내검색횟수
접근DB수	논문, 특허, 동향, 보고서등 이용 DB종류 개수
이용문서수	검색세션에서 이용된 콘텐트 중 중복 제거한 콘텐트 숫자
이용횟수	상세보기수 + 원문보기수 + 원문복사수
접속횟수	검색세션당 검색로그 총 수
목록이동횟수	검색결과 페이지 이동횟수

이지 보기, 원문 다운로드, 원문복사 신청 등의 행위가 30분 시간 내에 계속적으로 일어나는 것을 방문으로 보고 하나의 세션으로 처리하였다. 사이트 특성에 맞게 검색세션은 <표 2>의 형식으로 검색세션 데이터를 집계하였다. 이용자를 구분하는 방법은 IP를 사용할 수 있으나, IP 공유로 인한 식별성이 떨어지는 문제를 고려하여 PC 식별 쿠키를 사용하였다.

검색로그, 검색세션, 검색에 이용된 검색어 집합, 이용한 (access) 문서 집합에 대한 (그림 1)과 같은 데이터 모델을 설계하여 분석이 용이하도록 DB로 구축하였다.



(그림 1) 데이터 모델

(그림 2)와 같은 방문세션 처리로직으로 자바 프로그램을 작성하여 검색로그 분석을 수행하였다.

step 1. 검색로그를 기록 일자/시간 순으로 정렬한다.
step 2. 오래된 순서로 검색로그를 가져온다.
  step 2-1 "검색로그 pcid와 같은 세션 최종시간 > 검색로그 시간-30분" 조건을 만족하는 기존 검색세션 레코드가 있는지 검사
  step 2-2 있으면, 기존 세션 연장 처리 작업 수행 (기존 세션 레코드 갱신)
  step 2-3 없으면, 신규 세션 추가 작업 수행 (신규 세션 레코드 생성)
step 3. 모든 데이터에 대해 step 2. 처리 반복

(그림 2) 검색로그 세션분석 처리 로직

검색 키워드 식별 및 처리방법은 검색세션에서 추출된 검색로그에서 불용어 사전에 있는 용어를 제거하고, FAST 검색엔진[3]에서 제공하는 형태소 분석기를 통해 확장된 검색식으로부터 유효 키워드를 추출하였다. 그 방법은 (1)과 같은 정규표현식을 사용하였다.

다음으로 검색세션에서 이용한 문서를 추출하였다. 검색세션과 문서키를 연결하고 DB 종류와 이용횟수, 주제 분류를 가지는 데이터를 만들었다.

```
((^|\\s)[0-9\\p{Punct}]+[X]?(\\s|$))|
((^|\\s)(NJOUINPRO)[0-9]*(\\s|$))|
((^|\\s)[A-Z][0-9]+[A-Z][-]?[0-9]+[/][0-9]+[*]?(\\s|$))|
((^|\\s)[A-Z][0-9]+[A-Z][-]?[0-9]+[*]?(\\s|$))|
((^|\\s)[A-Z][0-9]+[A-Z][*]?(\\s|$))|
((^|\\s)[A-Z][0-9]+[*](\\s|$))|
((^|\\s)[/][NW][0-9]*(\\s|$))
```

(1)

검색성공률은 검색세션에서 원문보기와 원문복사신청 유도된 경우를 검색 이용자가 찾는 자료를 획득한 경우 성공하였다고 가정하여 측정하였다. 성공률 계산 결과 <표 3>와 같이 100.5만회의 검색 방문 중 성공한 경우는 18.9만회로 18.82%로 나타났다.

〈표 3〉 검색 성공 비율

구분	성공	실패	합계
횟수	189,262	816,485	1,005,747
비율	18.82%	81.18%	100%

2.3 용어 자동분류

용어 자동분류는 용어의 문맥적인 성질을 근거로 하여 용어의 의미를 결정, 동의어 및 관계어군을 자동으로 만드는 방법이다[12]. 전문가가 수작업으로 용어를 분류하여 주제명 표목을 구축하는 경우에는 시간과 노력이 많이 들고, 객관적으로 일관성 있는 분류결과를 얻기 어렵다. 또한 수작업에 의한 주제명 표목은 일반적이고 전역적인 수준의 개념체계로서, 일반적인 개념간의 관련성을 표현하는데 효과적인 반면 최신의 학문분야를 표현하는 개념이나 구체적인 개념을 즉각적으로 반영하기가 어렵다. 이러한 문제점을 해결하기 위해 많은 연구자들이 자동 용어 분류에 관심을 갖기 시작했다. 특히 1960년대 후반에 들어 지식의 자동분류를 위해 클러스터링의 개념이 도입되고 컴퓨팅 기술의 발전으로 인해 그 처리 속도가 향상되면서, 1980년대부터 용어뿐만 아니라 문헌을 대상으로 한 자동분류 연구가 증가했다[7]. 현재 용어 자동분류는 탐색용 시소러스를 이용한 질의 확장뿐만 아니라 정보조직 및 접근 도구와 데이터 마이닝 등 그 적용범위가 넓다.

2.4 자동분류 용어와 검색어 매칭

STEAK 시스템은 비통제 어휘의 의미망을 이용한 다국어 질의확장 및 색인 분류를 위한 기반 시스템으로 현재 KISTI에서 개발 중인 언어자원 생성 및 분석도구이다. STEAK 시스템은 크게 두 영역으로 구분되는데 첫 번째는 다국어 어휘 간의 관련 네트워크를 자동생성하고 동적으로 해석하여 제공하는 기능이며 두 번째는 구축된 자원으로부터 언어자원의 학습 환경을 구축하고 이를 이용해 학술정보를 자동분류하는 기능이다[2][4]. 검색어의 범주분석을 위해 STEAK 시스템의 용어의 범주를 동적으로 해석하는 기능을 활용하였다.

과학기술 주제 분류에 따라 논문등 문서 50만 건에서 추출/정제한 용어를 과학기술 주제 분류별 적합성에 따라 231,156개의 자동분류 용어 DB를 생성하였다. 용어-범주 유사도값 계산은 분류별 용어 출현 문서수를 기준으로 <표 4>와 같은 2x2 연관행렬을 작성한 다음 연관성 척도 공식으로 (2)오치아이 코사인 유사도 계수를 사용하였다. 오치아이 유사도계수는 이진모델로 고빈도 용어에 대한 유사도값을 선호하는 방식이다[9][11].

〈표 4〉 용어와 범주사이 2x2 분할표

	범주 c_j 소속	c_j 이외 범주 소속
용어 f_i 출현	a	b
용어 f_i 미출현	c	d

$$similarity_{ab} = \frac{a}{\sqrt{(a+b)*(a+c)}} \quad (2)$$

검색어 매칭은 2009년 동안 검색어 이용 순위 1만등까지 추출하여 자동분류 용어 DB와 비교하여 일치하는 용어와 검색어를 추출했다. 그리고 매칭된 키워드-용어에 대해 주제 분류별 상대적 가중치를 계산한 결과는 "LED/I/I::1.43517::43.62||B::1.06088::32.24||G::0.60679::18.44||K::0.18747::5.7"의 형태로 출력했다. "LED"는 키워드, "I" 대표 주제 분류(전기전자), "I::1.43517::43.62~" 각 주제 분류별 상대적 자질값 1.43517 과 백분율 43.62 이다.

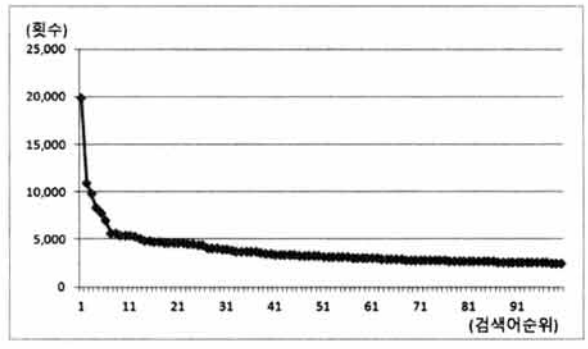
3. 검색어 주제 분야 식별 결과 분석 및 검증

3.1 검증 모형 설계

앞 장에서 분석된 용어 자동분류를 사용한 검색어 범주 분석결과와 클릭로그 분석을 하여 얻은 문서 범주가 상호 유사한 상관관계가 있는지 검증하기 위한 모형을 만들었다. 검증 방식은 18개 범주별로 나타난 상대적인 비율을 유사도 공식으로 측정하는 것이다. 즉, 검색세션에서 열어본 문서의 주제 분류를 추출하여 분류별 문서수를 세고 상대적인 문서 비율을 계산하였다.

검증용 데이터는 검색로그 분석 데이터 모델에서 검색세션에서 사용된 상위 100위 까지 검색어와 검색세션에서 단순 초록 열람이 아닌 원문 다운로드 또는 원문 복사 신청에 해당하는 문서와 과학기술표준 분류코드를 추출하였다. (그림 3)과 같이 상위 100위의 검색어 사용 횟수는 10위 이하에서 점진적으로 감소하는 모습을 보이고 100위 이하의 검색어의 사용횟수는 2,500회 이하이다. 상위 검색어만 분석대상으로 한 이유는 검색로그에서 빈도수가 낮은 검색어는 개인적인 성향에 따른 특이값으로 분석 의미가 없고, 다양한 사람이 많이 선호하는 검색어와 클릭한 자료의 일반적인 경향을 분석하는 목적을 가지고 있기 때문이다.

그리고 검색어가 단순하게 하나의 분류 범주에만 쓰이는 것이 아니다. 그래서 최대 빈도만으로 적합성을 따지지 않



(그림 3) 검색어 사용 횟수(상위 100위)

았다. 왜냐하면 연구주제의 다양한 이슈가 바뀌고 연구자의 관심사항이 다양하게 표현되기 때문이다. 주제별 상대적 비율에 대한 유사성 확인은 빈도에 민감한 코사인 유사도 공식이 적합하다. 검색로그에 이러한 다양한 측면이 반영되어 있기 때문에 연구주제의 변화를 용어 자동분류와 검색어-이용 문서 주제 분류 차이를 통해 살펴보는 것은 의미가 있다고 가정하였다. 용어 자동분류 vs 검색어-이용 문서 주제 분류 유사도 계산은 코사인 유사도 계수 (3) 수식을 이용하였다[13].

$$similarity_{ab} = \frac{\sum_i (w_{ai} * w_{bi})}{\sqrt{\sum_i w_{ai}^2} * \sqrt{\sum_i w_{bi}^2}} \quad (3)$$

(단, 과학기술표준분류 $i = 1 \sim 18$, w_{ai} 분류별 이용 문서수 비율, w_{bi} 용어 자동분류 가중치, 유사도 값 0 ~ 1)

이용 문서수에 대한 분류 비율과 자동분류 용어와 검색어 매칭결과 검색어 분류 비율을 유사도 계산 공식(3)으로 계산하여 검색어의 범주화가 어느 정도 일치하는지 분석하였다.

3.2 유사도 분석 결과

검색세션에서 추출한 실제 이용 문서 집합의 주제 분야가 검색어-용어 매칭을 통해 식별된 주제 분야와 어느 정도 일치하는지 유사도 검증을 통해 살펴보았다. <표 5>은 이용 문서수 비율 A 와 검색어 자동분류 비율 B에 대하여 유사도 분석을 통한 검증 결과 100개 중 상위 10개이다.

유사도 검증 결과 과학기술표준분류 범주의 상대적인 출현 비율을 확인할 수 있다. 2009년 가장 많이 사용되었던 검색어 "LED"의 경우 I(전기전자)에서 검색 문서 범주와 용어 범주가 일치하게 높은 비율을 보이고 있다. 검색 이용문서에서 F(기계) 분야의 문서가 LED 검색 후에 이용한 문서로 나타나는 반면 용어-범주에서는 나타나지 않았다. 용어 자동분류와는 다른 현상을 보인다. 이것은 실제 검색 이용자가 I(전기전자) 분야의 문서를 주로 보지만 F(기계) 분야의 문서에도 관심을 가지고 있다고 할 수 있다. "태양전지"의 경우 용어-범주에서는 O(에너지자원)가 37%의 비율로 가장 높았지만, 실제로 검색에 이용된 문서 분야는 I(전기전

〈표 5〉 이용 문서 분류 vs 검색어 범주 비교표
(ratio A: 이용 문서 분류 비율, ratio B: 용어 자동분류를 사용한 검색어 범주 비율)

rank	검색어	대표 범주	유사도	문서	과학기술표준분류 (* 최대값)																	Totals
					A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	led	I	0.8357	counts	0.00	159.00	77.00	56.00	7.00	178.00	52.00	68.00	*548.00	27.00	3.00	114.00	29.00	8.00	7.00	1.00	19.00	11.00	1,364
				ratio A	0.00	11.66	5.65	4.11	0.51	13.05	3.81	4.99	*40.18	1.98	0.22	8.36	2.13	0.59	0.51	0.07	1.39	0.81	100%
				ratio B	0.00	32.24	0.00	0.00	0.00	0.00	18.44	0.00	*43.62	0.00	5.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100%
2	태양 전지, solarcell	O	0.5713	counts	1.00	13.00	96.00	4.00	4.00	172.00	50.00	135.00	*417.00	12.00	0.00	5.00	4.00	3.00	38.00	0.00	25.00	0.00	979
				ratio A	0.10	1.33	9.81	0.41	0.41	17.57	5.11	13.79	*42.59	1.23	0.00	0.51	0.41	0.31	3.88	0.00	2.55	0.00	100%
				ratio B	0.00	7.91	0.00	0.00	0.00	0.00	16.05	11.42	23.16	0.00	0.00	0.00	0.00	4.45	*37.00	0.00	0.00	0.00	100%
3	연료 전지, Fuel cell	O	0.3436	counts	0.00	8.00	130.00	13.00	2.00	*372.00	21.00	118.00	208.00	8.00	1.00	7.00	2.00	14.00	57.00	1.00	38.00	2.00	1,002
				ratio A	0.00	0.80	12.97	1.30	0.20	*37.13	2.10	11.78	20.76	0.80	0.10	0.70	0.20	1.40	5.69	0.10	3.79	0.20	100%
				ratio B	0.00	0.00	0.00	0.00	0.00	0.00	10.37	16.89	20.91	0.00	0.00	0.00	0.00	0.00	*51.83	0.00	0.00	0.00	100%
4	나노, nano	G	0.6233	counts	0.00	2.00	72.00	19.00	2.00	81.00	39.00	*162.00	61.00	6.00	2.00	25.00	25.00	22.00	8.00	0.00	6.00	2.00	534
				ratio A	0.00	0.37	13.48	3.56	0.37	15.17	7.30	*30.34	11.42	1.12	0.37	4.68	4.68	4.12	1.50	0.00	1.12	0.37	100%
				ratio B	0.00	23.69	0.00	0.00	0.00	0.00	*35.57	32.95	7.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100%
5	인삼, ginseng	B	0.3821	counts	0.00	0.00	14.00	*402.00	0.00	16.00	0.00	14.00	5.00	1.00	0.00	325.00	138.00	2.00	0.00	1.00	0.00	0.00	918
				ratio A	0.00	0.00	1.53	*43.79	0.00	1.74	0.00	1.53	0.54	0.11	0.00	35.40	15.03	0.22	0.00	0.11	0.00	0.00	100%
				ratio B	0.00	*62.41	0.00	36.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.25	0.00	0.00	0.00	0.00	0.00	100%
6	풍력, wind power	O	0.3098	counts	0.00	3.00	0.00	1.00	11.00	*194.00	11.00	2.00	148.00	7.00	0.00	4.00	1.00	25.00	11.00	0.00	35.00	2.00	455
				ratio A	0.00	0.66	0.00	0.22	2.42	*42.64	2.42	0.44	32.53	1.54	0.00	0.88	0.22	5.49	2.42	0.00	7.69	0.44	100%
				ratio B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	23.95	0.00	0.00	0.00	0.00	16.38	*49.72	9.95	0.00	0.00	100%
7	solarcell	O	0.6878	counts	1.00	72.00	70.00	1.00	0.00	143.00	13.00	50.00	*183.00	2.00	0.00	1.00	4.00	4.00	144.00	0.00	4.00	0.00	692
				ratio A	0.14	10.40	10.12	0.14	0.00	20.66	1.88	7.23	*26.45	0.29	0.00	0.14	0.58	0.58	20.81	0.00	0.58	0.00	100%
				ratio B	0.00	5.97	0.00	0.00	0.00	0.00	17.58	15.29	15.92	0.00	0.00	0.00	0.00	8.67	*36.57	0.00	0.00	0.00	100%
8	water	C	0.5949	counts	0.00	19.00	73.00	36.00	10.00	56.00	15.00	*110.00	25.00	1.00	0.00	41.00	44.00	76.00	6.00	0.00	17.00	5.00	534
				ratio A	0.00	3.56	13.67	6.74	1.87	10.49	2.81	*20.60	4.68	0.19	0.00	7.68	8.24	14.23	1.12	0.00	3.18	0.94	100%
				ratio B	0.00	21.24	*21.69	0.00	7.94	0.00	0.00	2.97	0.00	0.00	0.00	9.67	0.00	20.17	13.03	0.00	0.00	3.29	100%
9	oled	I	0.7922	counts	0.00	53.00	75.00	2.00	0.00	73.00	7.00	67.00	*251.00	39.00	0.00	4.00	2.00	0.00	0.00	1.00	2.00	0.00	576
				ratio A	0.00	9.20	13.02	0.35	0.00	12.67	1.22	11.63	*43.58	6.77	0.00	0.69	0.35	0.00	0.00	0.17	0.35	0.00	100%
				ratio B	0.00	20.84	0.00	0.00	0.00	0.00	24.27	15.92	*38.19	0.00	0.00	0.00	0.00	0.00	0.78	0.00	0.00	0.00	100%
10	nano	H	0.5605	counts	1.00	42.00	79.00	13.00	0.00	*174.00	64.00	112.00	34.00	8.00	2.00	13.00	34.00	14.00	12.00	0.00	1.00	0.00	603
				ratio A	0.17	6.97	13.10	2.16	0.00	*28.86	10.61	18.57	5.64	1.33	0.33	2.16	5.64	2.32	1.99	0.00	0.17	0.00	100%
				ratio B	0.00	23.20	0.00	0.00	0.00	0.00	31.25	*32.82	12.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100%

자)가 가장 높고 O(에너지자원)은 매우 낮은 편이다. 이것은 검색 이용자가 I(전기전자)의 문서를 O(에너지자원)쪽 문서보다 많이 본다고 할 수 있으며 태양전지와 관련한 연구에서 전기전자관련 관심 이슈가 많다고 볼 수 있다.

<표 5>와 같이 일부 키워드 범주가 문서이용 분류와 80%이상으로 일치하나 일부는 30%정도의 낮은 유사도로 나타났다. 평균적인 수준은 58.8%로 관련연구 전문가 평가 수준인 76.8%보다 낮다[2]. 전체 건수를 대상으로 한 전문가 평가 수준은 범주화에 대한 적합성 평가의 기준점이 된다. 상위 최다 빈도의 키워드 범주를 대상으로 나타난 결과 이므로 최소한 전문가 평가 수준보다 같거나 높아야 하지만, 정보이용자 측면에서 전문가평가와 같이 동일한 분류기준을 가지고 있는 것이 아니기 때문에 일부 키워드의 경우 낮게 나왔다. 이것은 일반적인 이용자 측면에서 나타난 정보이용의 경향의 변화로 볼 수 있다.

4. 결 론

과학기술정보 이용자의 연구 주제와 이슈사항은 새로운 용어와 학문이 접목됨에 따라 바뀌게 되므로 정해진 용어와 분류 범주는 이러한 연구 흐름을 충분히 반영하지 못한다. 이용자의 검색로그는 다수의 연구자들이 각기 다른 목적으로 방문하여 검색한 행위에 대한 기록이며 검색 키워드는 보다 직접적인 관심사항의 표현이다. 검색로그 분석과 키워드에 대한 분석을 통해 많은 정보를 얻을 수 있지만, 처리 과정이 복잡하고 시간이 많이 소요되는 일이다.

용어 자동분류를 이용하여 검색어의 주제 분류를 식별한 결과가 클릭로그 분석으로 식별한 이용 문서의 주제 분류와 얼마나 일치하는 유사도 검증을 수행하였다. 검색에 많이 이용된 키워드 100개까지의 유사도 평균은 58.8%의 분류 비율의 유사성이 있는 것으로 나타났다. 이것은 검색결과 전문가 평가 수준인 76.8%보다 낮았으며, 이것은 용어의 범주가 새로운 연구영역으로 확대되면서 실제 정보이용자가 해당 키워드로 선택하는 관심 논문이 변하기 때문이라고 파악되었다.

본 연구를 통해 정보 이용자가 입력한 검색어 자동분류 결과를 분석하고 클릭로그 분석을 통해 검증을 함으로써 유사성이 높은 검색어 그룹과 그렇지 않는 검색어 그룹을 구분할 수 있었다. 후속 연구로 유사성이 높지 않은 검색어 그룹에 대해 검색로그에 기반을 두어 보정하고 성능을 높이는 방법에 대한 연구가 필요하다.

참 고 문 헌

[1] 이수상, 위성광, "디지털 도서관 이용자의 검색행태 연구", 한국도서관정보학회지, 제 40권 제 4호, pp.139-158, 2009.

[2] 정도헌, 유소영, 김환민, 김혜선, 김용광, 한희준, "웹 정보의 자동 의미연계를 통한 학술정보서비스의 확대 방안 연구", 정보관리연구, 제 40권 제 1호, pp.133-156, 2009.

[3] FAST, "FAST Enterprise Search Platform 5.3 Advanced Linguistics Guide", Document Number: ESP1036, Document Revision: A, 2009.

[4] 정도헌, 최희윤, "과학기술 전문용어의 다국어 의미망 생성과 분석", 정보관리연구, 제 37권 제 4호, pp.25-47, 2007.

[5] 박소연, 이준호, "웹 검색 분야에서의 로그 분석 방법론의 활용도", 한국문헌정보학회 학술발표논문집 제 21집, pp.81-94, 2006.

[6] 박소연, 이준호, 김지승, "클릭 로그에 근거한 네이버 검색 질의의 형태 및 주제 분석", 한국문헌정보학회지, 제 39권 제 1호, pp.265-278, 2005.

[7] 이재윤, "문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구", 정보관리연구, 제 36권 제 4호, pp.51-69, 2005.

[8] 남영준, 김규환, "유사어 사전을 이용한 웹기반 질의문의 자동 범주화에 관한 연구", 정보관리연구, 제 35권 제 4호, pp.81-105, 2004.

[9] 이재윤, "연관성 척도의 빈도수준 선호경향에 대한 연구", 정보관리학회지, 제 21권 제 4호, pp.281-294, 2004.

[10] 서진완, "로그화일(Log file)을 이용한 공공기관의 홈페이지 분석과 정책적 함의", 한국행정학회 춘계학술대회발표논문집, pp.501-517, 2001.

[11] Dunja Mladenic, Marko Grobelnik, "Feature Selection for Classification Based on Text Hierarchy, In Working notes of Learning from Text and the Web", Conference on Automated Learning and Discovery(CONALD'98), 1998.

[12] 서은경, "용어의 자동분류에 관한 연구", 석사학위논문, 연세대학교 대학원, 도서관학과, 1984.

[13] Gerard Salton, Michael J. McGill, "Introduction to Modern Information Retrieval", New York: Mc Graw Hill, 1983.

이 태 석

e-mail : tsyi@kisti.re.kr
1995년 경원대학교 전자계산학과(공학사)
1997년~2000년 산업기술정보원
2005년 고려대학교 컴퓨터학과(이학석사)
2001년~현 재 한국과학기술정보연구원 NDSL서비스실 선임연구원
관심분야 : 대용량 데이터마이닝, 데이터 웨어하우징, 정보검색 등

정 도 헌

e-mail : heon@kisti.re.kr
1997년 연세대학교 문헌정보학과(학사)
2011년 연세대학교 문헌정보학과(박사수료)
2003년~현 재 한국과학기술정보연구원 소프트웨어연구실 선임연구원
관심분야 : 정보추출, 정보분석, 기술예측 등

문 영 수

e-mail : youngsum@kisti.re.kr
1991년~2000년 산업기술정보원
2003년 서경대학교 전산정보관리학과(학사)
2005년 서경대학교 인터넷정보학과 (이학석사)
2001년~현 재 한국과학기술정보연구원 NDSL서비스실 선임연구원
관심분야 : 고객세분화, 데이터마이닝, DB 마케팅 등

박 민 수

e-mail : mspark7@gmail.com
1998년 이화여자대학교 문헌정보학과(B.A.)
2002년 Rutgers University School of Communication and Information (MLIS)
2008년 University of Pittsburgh School of Information Sciences(Ph.D.)
2009년~현 재 한국과학기술정보연구원 NDSL서비스실 선임연구원
관심분야 : Human-Information Interaction, Human-Computer Interaction

현 미 환

e-mail : mhhyun@kisti.re.kr
1999년 한양대학교 컴퓨터공학과(공학사)
2003년 숙명여자대학교 e비즈니스학과 (정보통신학 석사)
2005년~현 재 한국과학기술정보연구원 NDSL서비스실 선임연구원
관심분야 : HCI, Usability, Open API, Cloud Service 등