

수자원문헌의 주제탐색과 인용탐색의 검색효율 비교 연구*

이 명 희**

목 차

1. 문제의 제기 2. 연구문제 및 가설 3. 연구방법 및 설계 3.1 예비연구의 결과 3.2 변인의 설정 3.3 CD-ROM 데이터베이스의 선정 3.4 표본집단의 선정 및 탐색질문의 작성 3.5 탐색과정 3.6 적합성의 판단	3.7 통계적 분석 4. 자료 분석 4.1 검색된 문헌의 총수 4.2 검색된 적합문헌의 총수 4.3 두 탐색방법에 중복되어 검색된 문헌의 수 와 한 방법에만 독특하게 검색된 문헌의 수 4.4 정도율 5. 선행연구와의 비교 6. 연구결과의 의미, 제언 및 결론
--	--

1. 문제의 제기

서지데이터베이스를 탐색하기 위한 주요 검색방법은 주제색인에 의한 탐색(이하 “주제탐색”)과 인용색인에 의한 탐색(이하 “인용탐색”)의 두가지 방법이 있다. 주제탐색은 정보검색의 대표적인 탐색방법이지만, 정보의 과잉(Information overload)과 탐색 실패(Search failure)에 따른 외부적인 문제점과 단어의 어휘사용에 관계되는 내부적인 문제점을 가지고 있다 (Larson, 1991). 외부적인 문제점의 정보과잉은 주로 레코드수의 증가에 따른 데이터베이스 크기의 확장과 빈번

* 이 논문은 1993년 미국 University of Wisconsin at Madison의 School of Library and Information Studies에 박사학위논문으로 제출된 것임.

** 성균관대학교 사서교육원 강사.

한 용어 절단 등에 의해 야기되어지며, 탐색실패는 논리연산자 “AND”의 빈번한 사용과 여러 데이터베이스의 다양한 명령어에 대한 무지 등에서 기인한다. 주제 탐색에 의한 정보검색의 내부적 문제는(비록 그것이 자연언어 탐색방법이든 통제언어 탐색방법이든간에) 탐색을 위해 사용된 어휘에 의해 발생한다. 통제언어 탐색방법에서는 언어의 특정성 결여, 잘못된 어휘의 조합, 부정확한 용어간의 관계, 색인작성자간 일관성의 결여등으로 낮은 검색효율이 초래된다. 자연언어 탐색방법에서도 동의어, 동음이의어에 따른 어의적 애매성, 그리고 잘못된 조합(False drop)에 의한 문맥의 애매성등에 의해 문제가 제기된다 (Saracevic and Kantor, 1988).

인용탐색에 의한 정보검색은 주제탐색에 비해 용어간의 애매성이 배제되어지고 주관적인 인간의 해석이 필요없으며 색인자의 주제지식이 불필요한 등과 같은 장점을 가지고 있다. 그러나 과다한 자기인용, 인용의 과용 및 오용, 인용색인에 나타난 타이핑에러 및 내용의 가치에 관계없이 최신 논문이 적게 인용되어지는 등의 문제점은 인용탐색에서 검색효율의 저하를 초래한다 (Garfield, 1979 ; Smith, 1981).

정보탐색자들은 이러한 문제점을 고려하지 않고 데이터베이스를 사용하고 있으나, 색인작성자들이 주제의 개념을 잘 이해하고 있는지, 또는 주제색인에서 할당된 용어들이 개념과 밀접하게 관계되어 있는지는 명확하지 않다(Maron, 1977 ; Weinberg, 1988). 마찬가지로 인용된 문헌과 그 문헌을 인용한 문헌과의 관계는 잘 알려져 있지 않다(Pao, 1993 ; Cronin, 1981).

최근 정보검색의 여러 연구자들은 주제탐색과 인용탐색이 모두 적합한 자료를 검색함에도 불구하고 각기 다른 자료를 검색하고 있으며, 같은 일련의 질문에 대해서도 이 두 방법에 의한 검색효율은 매우 다르다는 것을 지적하고 있다 (Hurt, 1982 ; White et al., 1984 ; Hurt, 1985). 즉 같은 질문에 대해 소수의 자료들만이 두 방법 모두에 의해 중복되어 검색되고 있으며 많은 자료가 한 방법에 의해 독특하게 검색되므로 검색효율의 향상을 위해서 두 탐색방법을 동시에 사용하기를 제안하고 있다 (McCain et al., 1987 ; White and Griffith, 1987 ; McCain, 1989 ; Pao and Worthen, 1989 ; Pao, 1993).

그럼에도 불구하고 이 두 탐색방법이 어떠한 특성때문에 다른 검색결과를 초래하는지는 밝혀져 있지 않다. 따라서 본 연구는 탐색방법 이외에 두 특성인 질문타입(Query type)과 필드(Field)가 검색결과에 어떻게 영향을 미치는지를 파악하

기 위해 시도되었다. 이 연구는 수자원의 하위주제인 자연과학, 공학, 그리고 사회과학의 박사과정에 재학중인 연구자들에 의해 제기된 개념적 질문(Conceptual query)과 방법론적 질문(Methodological query)에 대해 주제탐색과 인용탐색이 어떤 문헌을 검색하는가를 파악하기 위해 수행되었다.

2. 연구문제 및 가설

이 연구의 연구문제는 주제탐색과 인용탐색이 수자원분야의 다학문간(Multidisciplinary) 구조내에서 개념적 질문과 방법론적인 질문에 대해 어떠한 문헌들을 검색하는가 하는 것으로서 세부적인 연구문제는 아래와 같다.

- ① 주제탐색과 인용탐색은 검색문헌의 총수, 검색된 적합문헌의 총수, 두 방법 모두에 의해 중복되어 검색된 문헌의 수, 오직 한 방법에 의해 독특하게 검색된 문헌의 수, 그리고 정도율에 대해 각각 다른 결과를 초래하나?
- ② 개념적인 질문에 대해, 주제탐색과 인용탐색은 검색문헌의 총수, 검색된 적합문헌의 총수, 두 방법 모두에 의해 중복되어 검색된 문헌의 수, 오직 한 방법에 독특하게 검색된 문헌의 수, 그리고 정도율에 대해 각각 다른 결과를 초래하나 ?
- ③ 방법론적인 질문에 대해 주제탐색과 인용탐색은 검색문헌의 총수, 검색된 적합문헌의 총수, 두 방법 모두에 의해 중복되어 검색된 문헌의 수, 오직 한 방법에 독특하게 검색된 문헌의 수, 그리고 정도율에 대해 각각 다른 결과를 초래하나 ?

개념적 질문과 방법론적인 질문에 대한 아이디어는 화이트(M. White)의 논문으로부터 얻어졌다(1975). 화이트에 의하면 개념적 질문은 연구자가 아이디어를 인식하거나 개념의 정의 또는 연구문제의 추출을 고려할 때 가지는 질문이다. 그러나 방법론적인 질문은 연구디자인을 개발하거나 데이터의 수집, 또는 통계방

법을 개발할 때 가지는 질문이다.

첫번째의 종속변인인 검색문헌의 총수를 평가하기 위해 세개의 가설이 설정되었다.

1. 주제탐색은 인용탐색에 비해 많은 수의 검색문헌을 초래한다.
2. 개념적질문에 대해 주제탐색은 인용탐색에 비해 많은 수의 검색문헌을 초래한다.
3. 방법론적 질문에 대해 인용탐색은 주제탐색에 비해 많은 수의 검색문헌을 초래한다.

다른 4개의 종속변인에 대해서도 같은 일련의 가설이 되풀이되었다.

3. 연구방법 및 설계

이 연구에서는 3가지의 독립변인(탐색방법, 질문의 타입, 필드)이 5가지의 종속변인(검색문헌의 총수, 검색된 적합문헌의 수, 두 방법에 의해 중복되어 검색된 문헌의 수, 한 방법에 의해 독특하게 검색된 문헌의 수, 정도율)에 어떻게 영향을 미치는가를 발견하기 위하여 실험연구(Experimental design) 방법으로 수행되었다.

3.1 예비연구(Pilot study)의 결과

실험연구에 유용한 정보를 얻기위해 예비연구가 행해졌는데, 6명의 수자원분야 연구자들이 자연과학분야와 응용과학 분야에서 선택되었으며 필자에 의해 만들어진 탐색요구서가 그들에게 제공되었다. 예비연구의 결과에 따라 실험연구에 다음 두가지 사실을 보완하였다.

첫째, 기존의 연구자들(Advanced researchers)은 연구과제수행 중 방법론에 대한 정보를 따로 필요로 하지 않고 초록과 색인서비스를 거의 이용하지 않는데 반해, 신입연구자들(Young researchers)은 연구과제중의 개념적 질문과 방법론적인 질문에 대해 명확히 구분할 뿐 아니라 기존의 색인 및 초록서비스를 잘 이용

하였으므로 이 연구의 목적을 위해 오직 신입 연구자만을 대상으로 표본추출하고, 특히 수자원은 다학문간 분야인 점을 고려 자연과학, 공학, 사회과학의 세 하위 주제로 구분하여 표본추출하였다.

둘째, SCISEARCH에서 사용되는 “Seed Article”의 선정기준은 “잡지에 수록된 논문으로서 참고 문헌을 가지고 있고, 또한 검색될 인용의 누적을 위해 최소한 5년정도 이전에 발표 되어야 한다”는 것이다.

3.2 변인의 설정

A. 통제된 변인(Controlled variables)

① 탐색자의 일관성 ② CD-ROM 탐색이므로 탐색시간에 따른 비용에 제한을 받지 않음 ③ 탐색결과에 있어서 동일한 양의 정보만 출력하기 위하여 저자명, 제목, 소스잡지명, 출판연도만 출력 ④ SCISEARCH와 SWRA에 동시에 수록된 잡지명만 소스로서 채택함으로써 내적타당도(Internal validity)를 증가시켰다.

B. 조작된 변인(Manipulated variables)

1) 독립변인(Independent variables)

- a. 탐색방법 (주제탐색과 인용탐색)
- b. 분야(Field : 자연과학, 공학, 사회과학)
- c. 질문타입(Query type : 개념적 질문과 방법론적 질문)

2) 종속변인(Dependent variables)

- a. 검색된 문헌의 총수
- b. 검색된 적합 문헌의 총수
- c. 두 방법에 의해 중복되어 검색된 문헌의 수 (Overlap)
- d. 한 방법에만 독특하게 검색된 문헌의 수 (Uniqueness)
- e. 정도율

3.3 CD-ROM 데이터베이스의 선정

검색효율의 비교를 위해 SWRA(Selected Water Resources Abstracts)와 SCISEARCH의 2가지 데이터베이스를 선정하였다. SCISEARCH는 Science Citation Index(SCI)의 온라인 버전으로 ISI(Institute for Scientific Information)에 의해 누적되는 CD-ROM 데이터베이스로서 서지적 인용 색인에 의해 탐색되고 있다. SWRA는 미국 Dept. of Interior의 Geological Survey에 의해 제공된 수자원분야 데이터베이스의 온라인 버전으로 디스크럽터와 키워드에 의해 탐색된다. 주제탐색과 인용탐색의 검색효율을 비교 측정하기 위하여 단일 데이터베이스를 사용하는 것이 이상적이지만 현재 이 두가지 탐색방법을 동시에 제공하는 데이터베이스는 수자원분야에 없으므로 두 데이터베이스를 다음과 같이 수정·사용하였다. 즉 SCISEARCH는 잡지의 내용만 수록하고 있으므로 SWRA에서도 잡지만 소스로 선정하여 이들 두 데이터베이스에 동시에 수록되어 있는 243종의 잡지만을 비교 대상으로 선정하였다. 그럼에도 불구하고 두 데이터베이스의 수록범위나 색인정책등의 차이가 검색결과에 영향을 미칠것으로 생각되어졌으나 실제로 행해진 한 연구에서 이들의 차이는 검색결과에 크게 영향을 미치지 않는 것으로 드러났다. SCISEARCH와 MEDLINE이 같은 소스 잡지의 수록범위에도 불구하고 색인정책의 차이때문에 각기 다른 잡지 기사를 수록하고 있는지를 조사한 한 연구에서 Pao(1993)는 오직 하나의 데이터베이스에 의해 색인된 잡지 기사는 극히 소수인 것을 발견했다.

3.4 표본집단의 선정 및 탐색질문의 작성

먼저 University of Wisconsin-Madison의 수자원분야 박사과정에 등록한 학생들의 명단을 입수한후, 필자는 그들의 논문주제가 자연과학, 공학, 사회과학의 3카테고리중 어디에 속하는지 파악하기 위하여 전화확인을 했다. 확인된 세 그룹에서 각각 7명씩을 무작위 추출하여 전체 21명의 학생들을 선발하였다. 필자는 이들 21명의 학생들 모두로부터 이 연구에 실험대상으로 참여하겠다는 동의를 받은 후 연구의 목적과 진행과정을 설명하는 편지를 전달하였는데 그 편지는 2가지 타입의 질문탐색 요구서(한가지는 개념적 질문을 위한 요구서와 다른 한가지는 방법론적 질문을 위한 요구서), 두 타입의 연구 질문에 대한 설명과 예비연구에서 한 연구자에 의해 제공되어진 탐색요구질문의 샘플을 포함했다. 따라서 21명 연구자의 학위논문 주제에 근거한 42질문과 각 질문당 2개씩의 인용문헌이

제공됨으로써 84번의 데이터베이스 탐색(주제탐색 42회, 인용탐색 42회)이 이루어졌다. 한명의 주제 전문가와 또 다른 한명의 정보전문가가 학생들에 의해 제공된 두 질문이 같은 연구목적에 위한 한쌍의 질문인지 검토하였으며, 또한 이들 질문이 연구 계획에서 요구되어진, 확실히 개념적이고 방법론적인 질문인지 검토하였다. 또한 3명의 주제전문가는 21명의 학생들이 자신의 주제가 수자원 하위분야의 어느 한가지에 속한다고 결정한 카테고리를 검토하였으며, 그 결과 각 질문은 자연과학(6), 사회과학(7), 공학(8)으로 구분되었다.

3.5 탐색과정

연구질문을 제공한 다수의 학생들은 그들 학위논문의 선행연구개관(Literature review)을 위해 적절한 정도율과 높은 재현율을 요구했으므로, 주제탐색은 “Building Block Strategy”를 사용해서 한명의 정보전문가에 의해 진행되었다. 먼저 수행된 결과로부터 얻어진 “Learning bias”를 최소화하기 위해 탐색전략은 데이터베이스를 사용하지 않고 2주에 걸쳐 수립되었으며, 최적의 탐색결과를 얻기 위해 첫번째 정보전문가에 의해 수립된 탐색전략은 두번째의 다른 정보전문가에 의해 검토되었으며 필요시에 새용어가 추가되거나 기존의 용어가 삭제되었다. 탐색자와 이용자의 상호작용을 배제하기 위해, 수립된 탐색 전략은 이용자와의 상호 교류나 결과에 대한 수정없이 필자에 의해 수행되었다.

인용탐색에 있어서, SCISEARCH는 제목의 키워드 탐색을 제공하고 있으나 두 색인이론의 공정한 비교를 위해 이 연구에서는 오직 “Citation Strategy”만 사용하기로 하였다. 최적의 탐색결과를 얻기위해 각 질문당 2개의 “Seed Article”이 제공되었으며 이들의 제공 기준은 참고문헌을 가진 잡지 논문으로서 최소한 5년 전에 출판되어야 한다는 것이다.

출판연도에 대한 필자의 임의적 선정이유는 Citation strategy가 그 Seed Article을 인용한 문헌만을 검색하므로, 인용의 누적, 다시말해서 최적수준의 검색효과를 얻기 위해 필요한 것이다. 이것은 주제탐색에 있어서 최적의 검색결과를 얻기위해 색인에 동의어와 유사어등 모든 가능한 단어를 키워드로 사용한 것과 유사한 것이다. 42질문에 대한 일관된 탐색 결과를 얻기 위해 인용탐색은 필자에 의해 Cycling이 없이 행해졌다. 검색된 논문들은 체크되고, 오직 243종의

공통 잡지에 수록된 논문에 한해서만 적합성의 판단을 요구하였다.

3.6 적합성의 판단

이 연구의 대상은 실제 박사 학위 연구논문에 근거한 질문을 제공한 학생들로써 그들은 실지 그 정보의 요구자이므로 적합성 판단의 가장 적절한 평가자로 인정되었다. 각 평가자에게 주어진 검색결과는 모두 같은 양의 정보인 저자명, 제목, 소스잡지명, 출판년을 포함하고 있었다. 적합성 판정을 위한 문서화된 판단 기준이 각자에게 주어지고 그들은 일반적 연구과제에 관계없이 오직 주어진 2가지의 연구질문에만 근거하여 검색된 논문의 적합성을 판단할 것이 요구되었다.

이 연구에서는 선행연구에서와 달리 적합성 판단을 두가지로 측정하였는데, 첫째, 적합한 논문의 수를 세기 위해서는 2 (① “상당히 유용한”과 “좀 유용한” ② 유용하지 않은)방법을 사용하고 둘째, 정도율의 계산을 위해서는 3 방법(① 상당히 유용한 ② 좀 유용한 ③ 유용하지 않은)을 사용하였다.

3.7 통계적 분석

일반적으로 3개의 독립변인 (탐색방법, 질문타입과 필드)을 테스트 하기 위해서는 삼원분산분석(3-Way ANOVA)이 사용되어지나 이 연구에서는 연구질문과 탐색방법이 독립적으로 관찰되어지지 않고 반복되어 측정·사용되므로 Split-Plot Factorial(SPF) 디자인을 사용하였다. SPF디자인은 Randomized-Block Factorial Design 대한 대안으로 제시된 것으로 연구자의 관심이 블럭내 실험처치(Within-Block Treatment)와 그들사이의 상호작용효과(Interaction Effect)에 있으면 아주 적합한 디자인이다.

특히 이 연구에서는 1개의 블럭간 실험처치(필드)와 2개의 블럭내 실험처치(탐색방법과 질문타입)를 규명하여야 하므로 SPF-P.qr디자인을 사용하였으며, 어떠한 평균값이 변인들 사이에서 특별히 어떻게 다른가 하는것을 밝히기 위하여 Scheffe Post Hoc 비교를 수행하였다.

4. 자료의 분석

이 연구의 실험디자인은 3개의 독립변인이 5개의 종속변인에 어떻게 영향을 미치는가하는 SPF-P,q,r 디자인이므로 BMDP통계 패키지를 사용하여 수행되었다. 그리고 이 연구에서 시도된 모든 테스트에는 단측검증(one-tail test)을 사용했고 유의수준 5%에서 신뢰도를 측정하였다.

4.1 검색된 문헌의 총수

< 표 1 >에서 보여주는 바와같이 대체로 두 탐색방법에 의해 검색된 문헌의 총수는 매우 비슷했으나(25) 개념적 질문에 대해 검색된 문헌의 수는 방법론적 질문에 대해 검색된 문헌의 수보다 많았다. 검색된 문헌의 총수는 세개의 필드에서 달랐다. 즉 자연과학 분야에서 가장 많은 논문(40)을 검색했으며, 그 다음이 공학(30), 그리고 다음이 사회과학(6)의 순이었다. 검색의 한 패턴이 Subset 사이에 존재하고 있었는데 주제탐색은 자연과학 질문(37)보다 공학질문(58)에 더욱 효과적이었으며 인용탐색은 공학(29)보다 자연과학(64)에서 많은 문헌을 탐색해 내었다. 그러나 사회과학질문을 위해서는 두 탐색방법이 다 효과적이지 못했다(주제색인:인용색인 = 6:7). 질문과 필드가 전체검색 문헌의 총수에 영향을 미치는 요소들이었으나 모든 가설은 부정되었다.

< 표 1 > 검색된 문헌의 총수의 평균

탐색타입	자연과학	공 학	사회과학	한계평균	한계평균	
					탐색	질문
개념적x주제	37	58	6	34	24	33
개념적x인용	64	29	7	32	(주제) (개념적)	
방법론적x주제	16	15	8	13	26	17
방법론적x인용	45	18	2	21	(인용) (방법론적)	
한계평균	40	30	6	25		

4.2 검색된 적합 문헌의 총수

질문당 검색된 적합 문헌수의 평균은 15였으며, 선행연구와는 달리 인용탐색이 주제 탐색보다 적합한 문헌을 더 검색해 내었다(표2 참조). 개념적 질문에는 두 탐색방법이 비슷한 수의 적합한 논문을 검색했으나 방법론적인 질문에는 인용탐색에 의해 검색된 적합 논문의 총수(13)가 주제탐색에 의해 검색된 적합 논문의 수(7)보다 많았다. 자연과학분야 적합논문수의 평균치는 21이고 공학은 20이었으나 사회과학은 4였다. 질문과 필드효과는 적합문헌수의 전체에 영향을 미치는 요소였으나 모든 가설은 부정되었다.

〈 표 2 〉 검색된 적합문헌의 총수의 평균

탐 색 타 입	자연과학	공 학	사회과학	한계평균	한계평균	
					탐색	질문
개념적x주제	17	33	3	19	13	19
개념적x인용	34	23	0	20	(주제) (개념적)	
방법론적x주제	8	8	4	7	17	10
방법론적x인용	24	15	2	13	(인용) (방법론적)	
한 계 평 균	21	20	4	15		

〈 표 3 〉 두 탐색방법에 중복되어 검색된 문헌의 수의 평균

질 문	자연과학	공 학	사회과학	한계평균
개 념 적	5.0	1.1	0.1	1.9
방법론적	1.3	0.0	0.0	0.4
한계평균	3.2	0.6	0.1	1.2

〈 표 4 〉 한 방법에만 독특하게 검색된 문헌의 수의 평균

탐색 타입	자연과학	공학	사회과학	한계평균	한계평균	
					탐색	질문
개념적x주제	32	56	6	32	22	31
개념적x인용	59	28	6	30	(주제) (개념적)	
방법론적x주제	14	15	8	12	25	16
방법론적x인용	44	18	2	20	(인용) (방법론적)	
한계평균	37	29	6	24		

4.3 두 탐색방법에 중복되어 검색된 문헌의 수(Overlap)와 한 방법에만 독특하게 검색된 문헌의 수(Uniqueness)

〈표 3〉에서 개념적 질문에 대한 평균 오버랩은 방법론적인 질문의 오버랩보다 높았고 평균은 각각 1.9와 0.4였다. 일반적인 결과는 평균오버랩이 아주 적다는 것을 보여주고 있으며 그들은 아주 적합한(highly relevant) 논문인 것으로 나타났다(표 4 참조). 검색된 논문중 유니크한 논문의 평균치는 24이며 개념적인 질문에는 두 탐색방법이 거의 동일한 수의 유니크한 논문을 검색했으나 방법론적인 질문에는 인용탐색이 주제탐색보다 더욱 유니크한 논문을 검색했다. 유니크한 적합 문헌의 수는 질문과 필드에 의해 영향을 받았다. 그럼에도 모든 가설은 부정되었다.

4.4 정 도 율

〈 표 5 〉에서 평균 정도율은 42%였으며, 일반적으로 개념적 질문의 탐색에는 인용탐색(50%)이 주제탐색(37%)보다 높은 정도율을 가졌으나, 방법론적인 질문에는 주제탐색이 43%의 정도율을 가진데 비해 인용탐색은 39%의 정도율을 가졌다. 그리고 주제별로는 공학(52%)이 최고의 정도율을 가지고 다음은 자연과학(44%) 그리고 사회과학(30%)의 순이었다. ANOVA 결과로부터, 개념적인 질문을 위해 인용탐색은 주제탐색보다 높은 정도율을 나타내고 방법론적인 질문

을 위해 주제탐색은 인용탐색보다 높은 정도율을 나타낸다는 가설은 인정되었다.

〈 표 5 〉 정도율의 평균

탐색타입	자연과학	공 학	사회과학	한계평균	한계평균	
					탐색	질문
개념적x주제	30	42	37	37	40	43
개념적x인용	44	59	43	50	(주제)	(개념적)
방법론적x주제	55	54	20	43	44	41
방법론적x인용	45	52	19	39	(인용)	(방법론적)
한 계 평 균	44	52	30	42		

5. 선행연구와의 비교

선행연구의 대부분은 주제탐색이 인용탐색보다 많은수의 문헌을 검색한다고 보고하고 있으나 이 연구에서는 보다 많은 적합한 문헌이 인용탐색에서 검색되었다. 이 연구와 선행연구 결과의 차이는 질문타입의 차이에 기인하는 것 같다. 선행연구의 탐색질문은 일반적인데 반해 이 연구의 탐색질문은 높은 특정성을 가진 박사과정 학생들의 학위논문 제목으로서 소수의 논문들이 검색되었다. 검색 논문수에서의 큰 차이가 질문사이에 존재하고 있었으므로 질문의 특성이 검색 문헌수에 크게 영향을 미치는 것으로 나타났다. 질문의 특성이 기존 선행연구의 질문의 성격과 다른가를 비교하기 위해 질문에 대한 질적분석(Qualitative analysis)이 특정성(Specificity), 복잡성(Complexity), 최신성(Recency)의 3요소를 가지고 이루어졌다.

특히, 개념적질문에 대해 특정성이 분석되었는데 그 결과 현저히 다른 두 그룹이 대두되었다. 첫째 그룹은 그들의 질문이 여전히 학위논문의 제목으로서 특정적인 그룹과 둘째 그룹은 학위논문의 제목으로서 어느정도 일반적이고 넓은 그룹이었다. 이들 두 그룹은 테스트되고, 주제탐색에서 일반적인 질문은 많은 양의 적합한 논문을 검색하는데 비해 특정적인 질문은 적은 양의 논문을 검색해 내는 것이 밝혀졌다. 특히 이 논문에서 주로 사용된 질문은 아주 특정적인 질문

이었으므로 소수의 논문만 검색했다. 여기에 반해 인용탐색은 질문의 특정성의 정도에 영향을 받지 않고 비교적 효과적으로 검색했다. 따라서 특정적인 질문에는 인용탐색이 효과적이고 일반적인 질문에는 주제탐색도 효과적이었다.

선행연구에서와 마찬가지로 이 연구에서도 주제탐색과 인용탐색은 중복된 문헌이 거의 없이 유니크한 논문을 검색한다는 것을 보여준다. 또한 선행연구에서는 주제탐색이 대부분의 논문을 검색해 내었으며 인용탐색은 단지 주제탐색을 보조하는데 불과하였으나 이 연구에서는 두 탐색방법이 거의 같은 수의 유니크한 문헌을 검색하였고 방법론적인 질문에서는 오히려 인용탐색이 주제탐색보다 많은 양의 문헌을 검색해 내었다 (20:12)

선행연구는 또한 인용탐색이 비록 적은 양의 논문을 검색했으나 주제탐색보다 높은 정도율을 가진다고 보고하고 있다. 그러나 이 논문에서는 개념적 질문을 위해서 인용탐색이 주제탐색에 의해 검색된 논문보다 높은 정도율을 가지고 방법론적인 질문을 위해서는 주제탐색에 의해 검색된 논문이 높은 정도율을 가졌다. 일반적으로 이 연구에서의 정도율은 선행연구에서의 정도율보다 낮는데 그 이유는 선행연구에서의 이분법적인 적합성의 척도와 달리 이 연구에는 3스코어(Scores)의 변형된 적합성의 측정척도가 사용되었기 때문이다. 3스코어를 사용한 이 연구에서 정도율이 50%라면 이분법을 사용한 선행연구에서는 거의 65%의 정도율을 가졌을 것이다.

6. 연구결과의 의미, 제언 및 결론

이 연구의 결과는 수자원데이터베이스의 탐색에 있어서 중요한 여러가지 실질적인 함축적인 의미를 내포하고 있다.

6.1 연구의 의미 (Implication)

A. 탐색방법

만약 주제탐색과 인용탐색에 의해 검색된 논문들이 오버랩이 없이 일관성있게

적합한 것으로 판단된다면 데이터베이스 디자이너에게 어떤 암시를 줄 수 있을 것이다. 즉 미래의 데이터베이스는 키워드, 디스크립터, 초록, 인용된 참고문헌을 포함하고 있는 데이터베이스가 될 것이다.

또한 두 탐색방법이 적은 오버랩과 많은 유니크한 논문을 검색하고 특히 오버랩된 논문이 상당히 적합한 문헌이라는 사실은 탐색전략의 구축에 영향을 줄 것이다. 즉 적은 수의 아주 적합한(highly relevant) 문헌을 검색하기 위해서는 여러개의 탐색방법을 사용하는 것보다 오버랩된 문헌을 탐색하게 하는 것이다.

SWRA는 여러가지의 접근점(access point)을 가지고 있기 때문에 SCISEARCH보다 적합한 문헌을 검색해낼 가능성이 높았음에도 불구하고 결과가 거의 비슷하거나 오히려 인용탐색이 방법론적인 질문에서는 더욱 효과적이었다. 주제탐색은 잘 확립된(well established) 주제분야와 표준화된 terminology를 가진 용어에 의해 색인되고, 일반적이고 broad한 유형의 질문에는 효과적이었다. 그러나 최근에 발달된 주제나 특정성을 지닌 질문(예를 들면 박사학위논문의 토픽)에는 인용 탐색이 더욱 효과적이었다. 특히 인용탐색에서는 잘 선택된 "Seed Article"만 주어진다면 특징적인 주제이거나, 넓은 범위의 주제이거나, 새로운 분야의 주제에 상관없이 잘 검색했으나 "Best Seed Article"의 선택을 위하여 신중을 기해야 한다는 것이 밝혀졌다.

B. 필드

수자원분야는 여러개의 학문분야가 합쳐진 다학문간 분야이므로 연구자들의 탐색 질문과 검색결과는 필드에 따라 다를 것으로 가정되었다. 이러한 가정은 필드에 따라 각기 다른양의 문헌들이 검색됨으로써 증명되었다. 공학분야 질문은 SWRA에 의해 가장 잘 검색되었으며 자연과학분야 질문은 SCISEARCH에 의해 가장 잘 검색되었다. 그러나 사회과학분야 질문을 위해서는 SWRA와 SCISEARCH는 모두 효과적이지 못했다. 이것은 사회과학연구자들의 정보요구를 만족시키기 위해서 사서는 다른 종류의 정보자료를 제공해야 하며, 또한 데이터베이스의 생산자들은 사회과학연구자들을 위해 법률, 정책, 의사결정, 경제학 등을 포함한 사회적 제 양상을 데이터베이스 수록 범위에 추가시켜야 한다는 것을 의미한다.

C. 질문타입

사용된 두 타입의 질문은 개념적 질문과 방법론적 질문이다. 개념적 질문은 내용의 “aboutness”에 근거하고 있기 때문에 주제탐색에 의해 잘 검색되리라고 생각되어졌고, 저자들은 그들이 사용하는 논문의 방법론을 위해 다른 논문을 잘 인용하고 있기 때문에 방법론적인 질문을 위해 인용 탐색이 더욱 효과적일 것이라고 예측되었다. 연구결과 SWRA와 SCISEARCH는 모두 개념적인 질문에 잘 응답했으므로 SCISEARCH도 SWRA와 마찬가지로 주제 데이터베이스의 한 형태임이 밝혀졌고, 방법론적인 질문에는 인용탐색이 더욱 효과적이라는 가설은 증명되었다.

6.2 미래연구를 위한 제언

이 연구에서 사용된 방법론이 다른 주제분야에서도 같은 검색결과를 낳는지 알아보기 위하여 다른 주제분야에서도 시행될 필요가 있다. 이 연구는 수자원이라는 여러 학문분야가 합친 다학문분야에서 테스트 되었는데, 다른하나의 가능성은 이 연구가 물리학, 화학, 지학등 잘 정의되고 발달된 학문분야에서도 같은 연구결과를 가지는지 시행해 보는 것이다. 또한 다른 사회과학분야나 자연과학분야에서의 적용은 학문분야의 특징과 기존 데이터베이스의 수정, 새 데이터베이스의 디자인등 탐색전략의 개발에 중요한 영향을 미칠 것이다.

실험디자인의 한 변형은 탐색전문가와 최종이용자가 시스템과 상호작용하게 함으로써 최적의 탐색결과를 갖게 하는 것이다. 선행연구의 결과는 이용자와 시스템이 상호 작용할때 최적의 검색효과를 가진다고 보고하고 있다(Swanson, 1977 ; Harter, 1992). 이 디자인에서는 learning bias를 방지하기 위하여, 탐색전략이 미리 수립되고, 수정이나 데이터베이스와의 상호작용이 없이 어느정도 인위적인 탐색전략이 사용됨으로써 주제탐색으로부터 최적의 결과를 얻지 못하게 되었다.

실험연구의 다른 한 변형은 최종이용자로 하여금 직접 CD-ROM의 탐색전략을 수립하게 하고 수행하게 하는 것이다. 사실 CD-ROM 탐색은 최종이용자에 의해 사용 되어지도록 고안된 것이다. 최종 이용자가 자신에게 특수한 상황에서 특수한 목적을 위해 검색함으로써 검색의 효율성을 제고할 수 있을 것이다.

다른 또 하나의 이슈는 질문의 타입이다. 비록 이 연구가 박사학위 논문 제목

만을 질문으로 선택함으로써 질문의 수준을 통제했으나 검색된 문헌의 수는 질문에 따라 큰 차이가 있었다. 이것은 질문타입의 특성이 검색결과에 크게 영향을 미친다는 것을 의미한다. 이것은 더욱 연구가 되어져야 할 항목이며 여기에 테스트된 특성 이외에 다른 특성들이 밝혀져야 할 것이다.

6.3 결 론

이 논문은 주제탐색과 인용탐색이 적은 오버랩과 함께 많은 유니크한 논문을 검색했다는 것을 보여준다. 비록 두 다른 방법에 의해 문헌이 검색되었지만 다른 기준에 근거한 두 타입의 적합성은 주제검색의 분명한 두 타입을 보여준다. 첫째는 디스크립터의 할당에 의해 잘 정의되었으나 연구자들에게는 공통되는 백그라운드 문헌을 결여하고 있는 것이고 둘째는 서지적 인용에 의해 결합되어 있고 연구자들에게는 잘 알려져 있으나 주제문제와 Terminology를 결여하고 있는 것이다.

이 연구의 결과는 같은 질문에 대해 주제탐색과 인용탐색이 다른 SET의 문헌을 검색하고 각 탐색방법은 전체 적합한 문헌의 검색에 기여한다는 것을 보여준다. 이것은 탐색키(key)의 수나 접근점의 수가 각각 색인의 질이나 데이터베이스의 범위보다 더욱 중요하다는 것을 제시하고 있다. 이것은 또한 다른 여러 탐색방법을 통한 다양한 탐색은 높은 재현율의 구현에 매우 중요하며 그러므로 적합한 문헌의 전체 양(body)을 창조하기 위해서는 종합적인 탐색이 필요하다는 것을 말해준다.

검색결과는 탐색방법이나 데이터베이스의 특징보다 질문타입의 특징때문에 크게 달라진다는 것을 보여주고 있다. 이것은 문헌의 주제나 데이터베이스의 수록 범위보다 질문타입의 특성이 검색결과에 더욱 영향을 미친다는 것을 의미하며 더욱 많은 연구가 질문타입에서 수행되어야 할 것이다. 탐색요구시에 이용자의 질문을 처리하는 과정에 있어서 다양한 테크닉이 더욱 개발되어야겠다.

참고문헌

- Cooper, W.S.(1973). On Selecting a Measure of Retrieval Effectiveness. *Journal of the American Society for Information Science*, 24, 87-100.
- Cronin, B.(1981). The Need for a Theory of Citing. *Journal of Documentation*, 37(1), 16-24
- Garfield, E.(1979). *Citation Indexing-Its Theory and Applications in Science, Technology and Humanities*. New York: John Wiley.
- Goffman, W.(1985). A Pragmatic Approach to Literature Selection. ed. by K. S. Warren, *Selectivity in Information Systems*, New York : Praeger Pub., pp. 117-143.
- Goffman, W. and Pao, M.L.(1980). Retrieval of Biomedical Information for Emerging Interdisciplinary Problems. *Proceedings of 4th Congress on Medical Librarianship*, Belgrade, Yugoslavia, 39-50.
- Harter, S.P.(1986). *Online Information Retrieval: Concepts, Principles and Techniques*. New York: Academic Press.
- Harter, S.P.(1992). Psychological Relevance and Information Science. *Journal of the American Society for Information Science*, 43, 602-615.
- Harter, S.P. and Peters, A.R.(1985). Heuristics for Online Information Retrieval; a Typology and Preliminary Listing. *Online Review*, 9, 407-424.
- Horry, A.(1981). *A Comparison of Citation Similarities and Index Term Similarities for Linking Subject Related Documents*. Ph.D. Dissertation, Cleveland, OH: Case Western Reserve University.
- Hurt, C.D.(1982). Important Items in the Database; an Investigation of Two Methods of Identification. *Online Review*, 6(3), 227-233.
- Hurt, C.D.(1985). A Comparison of Bibliometric Approach and Historical Approach to the Identification of Important Literature. *Information Processing and Management*, 19(3), 151-157.
- Hutchins, W.J.(1978). The Concept of "Aboutness" in Subject Indexing. *Aslib Proceedings*, 30, 172-181.

- Katzer, J. et al.(1982). *A Study of the Impact of Representations in Information Retrieval Systems*. Syracuse, N.Y.: Syracuse University Press.
- Lancaster, F.W.(1991). *Indexing and Abstracting in Theory and Practice*, Urbana, IL : University of Illinois Press.
- Larson, R.R.(1991). The Decline of Subject Searching; Long-term Trends and Patterns of Index Use in an Online Catalog. *Journal of the American Society for Information Science*, 42 (3), 197-215.
- Lesk, M.E. and Salton, G.(1968). Relevance Assessment and Retrieval System Evaluation. *Information Storage & Retrieval*, 4(4), 343-359
- Maron, M.E.(1977). On Indexing, Retrieval and the Meaning of About. *Journal of the American Society for Information Science*, 28, 38-43.
- McCain, K.W.(1989). Descriptor and Citation Retrieval in the Medical Behavioral Science Literature: Retrieval Overlaps and Novelty Distribution. *Journal of the American Society for Information Science*, 40(2), 110-114.
- McCain, K.W., White, H.D. and B.C. Griffith(1987). Comparing Retrieval Performance in Online Databases. *Information Processing and Management*, 23(6), 539-553.
- McCain, K.W. et al.(1983). *Test Retrievals in Five Major Databases Covering the Medical Behavioral Sciences(MBS) Literature*. Final Report. NLM-Drexel MBS Project, Study 2, Philadelphia, PA: Drexel University. N01-LM-3510
- Pao, M.L.(1993). Term and Citation Retrieval. *Information Processing & Management*, 29(1), 95-112.
- Pao, M.L. and Worthen, D.B.(1989). Retrieval Effectiveness by Semantic and Citation Searching. *Journal of the American Society for Information Science*, 40(4), 226-235.
- Robertson, S.E.(1981). The Methodology of Information Retrieval Experiment. ed by Karen Spark Jones. *Information Retrieval Experiment*, London: Butterworths, pp.9-31.
- Salton, G.(1971). Automatic Indexing Using Bibliographic Citations. *Journal of Documentation*, 27(2), 98-110.

- Saracevic, T. and Kantor, P.(1988). A Study of Information Seeking and Retrieving. 3 parts, *Journal of the American Society for Information Science*, 39, 161-216.
- Schamber, L., Eisenberg, M.B. and Nilan, M.S.(1990). A Reexamination of Relevance: Toward a Dynamic, Situational Definition. *Information Processing and Management*, 26, 755-776.
- Smith, L.(1981). Citation Analysis. *Library Trend*, 30(1), 85-105.
- Snow, B. and Ifshin, S.(1984). Online Database Coverage of Forensic Medicine. *Online*, 8(2), 37-43.
- Spark Jones, K.(1981). Retrieval Systems Tests 1958-1978. ed by Karen Spark Jones. *Information Retrieval Experiment*, London: Butterworths. pp.213-255.
- Spencer, C.(1967). Subject Searching with Science Citation Index : Preparation of Drug Bibliography Using Chemical Abstracts, Index Medicus, and Science Citation Index 1961 and 1964. *American Documentation*, 18(2), 87-96.
- Swanson, D.R.(1977). Information Retrieval as a Trial-and-Error Process. *Library Quarterly*, 47(2), 128-148.
- Swanson, D.R.(1987). Two Medical Literatures that are Logically but not Bibliographically Connected. *Journal of the American Society for Information Science*, 38, 228-233.

ABSTRACT

Retrieval Effectiveness of Subject Descriptor and Citation Searching in the Water Resources Literature

Myeong-Hee Lee*

This study measured whether subject descriptor searching and citation searching retrieve different documents for conceptual queries and methodological queries in natural science, engineering and social science. The retrieval effectiveness of two search methods was measured using as criteria, total number of documents retrieved, total number of relevant documents, overlapping and unique documents and precision ratio. The search subject was water resources and the databases used were Selected Water Resources Abstracts (SWRA) and SCISEARCH. Data were collected for 21 doctoral students working on their dissertations in the three fields of water resources. Principal findings included: 1) subject searching and citation searching each retrieved substantially equal number of documents; 2) total number of relevant documents for conceptual queries was larger than that for methodological queries, while there was a large variation among the three fields; 3) the average overlap was quite small, while citation searching yielded more unique documents than subject searching; 4) for conceptual queries, citation searching yielded a higher precision ratio than subject searching, while subject searching obtained a slightly higher precision ratio than citation searching for methodological queries ; and 5) citation searching was effective for both specific queries and broad queries if seed articles are well chosen, while subject

* Part-time lecture at Sung Kyun Kwan University.

searching only worked well for broad queries. It was further found that: 1) citation searching is not a subsidiary but a substantial retrieval method in water resources; 2) SWRA is effective for queries for engineering and SCISEARCH is appropriate for queries for natural science, while neither SWRA nor SCISEARCH work well for queries for social science; and 3) characteristics of queries affect retrieval results more than the characteristics of documents or the coverage of databases.