수자원분야 색인시스템의 검색효율 비교와 질적분석

Retrieval Effectiveness of the Two Indexing Systems in the Water Resources : A Qualitative Analysis

이 명 희(Myeong-Hee Lee)*

목 차

1. 서 론
2. 실험연구
 2.1 연구설계
 2.2 자료의 분석 및 결과
3. 질적연구
 3.1 특정성
 3.2 복잡성
 3.3 최신성
 3.4 자연과학과 공학질문의 분석
 3.5 사회과학질문의 분석
4. 결론 및 제언
 4.1 결 론
 4.2 제 언

초 록

수자원 분야의 데이타베이스를 대상으로 한 실험연구에서 검색효율의 차이는 질문특성의 차이에서 기인한다는 것이 밝혀졌기 때문에 특정성, 복잡성, 최신성의 세요소를 가지고 질문에 대한 질적분석이 행해졌다. 그 결과, 특정적인 질문보다 일반적인 질문을 위해서 보다 많은 적합문헌이 주제탐색을 통하여 검색되었으며, 박사학위논문의 연구질문은 특정성을 지닌 질문이었기 때문에 주제탐색보다 인용탐색을 통해 적합문헌이 더욱 많이 검색되었다. 또한 자연과학분야, 공학분야 및 사회과학 분야의 질문에 대한 개별적인 분석이 이루어졌다.

ABSTRACT

The previous study showed a large variation in performance within the queries and suggested that characteristics of queries cotribute to retrieval performance. Three attributes, specificity, complexity and recency were used to analyze the different results within queries. The result showed that subject searching retrieve more relevant documents for a query with low specificity than a query with high specificity and that queries from the doctoral students' dissertations were specific queries with high specificity.

* 성균관대학교 사서교육원 강사
접수일자 : 95. 7. 13

1. 서 론

서지데이타베이스를 탐색하기 위해 널리 사용되는 탐색방법은 주제탐색으로서 정보검색 연구자들에게 문헌의 내용을 파악하여 검색하기 위한 유용한 탐색방법으로 인식되었으나 정보과다(information overload)와 탐색실패(search failure)에 따른 문제점을 가지고 있다(Larson, 1991). 인용한 문헌과 인용된 문헌사이의 관계를 탐색하는 인용탐색 역시 과다한 자기인용, 인용의 오용과 남용, 인용색인에 나타난 타이핑 에러, 최근 논문에 대한 바이어스 등의 문제점을 가지고 있다(Garfield, 1979; Smith, 1981).

최근에 동일한 질문에 대한 주제탐색과 인용탐색의 검색결과를 비교한 많은 연구가 여러 연구자에 의해 행해졌는데 이 두 방법에 의해 매우 적은 양의 공통되는 문헌이 검색되었다(Chapman and Subramanyam, 1981; McCain, 1989; Pao and Worthen, 1989; Pao, 1993). 그러나 그 두 탐색방법이 어떻게 다르게 작용한다거나 또는 어떤 다른 요소때문에 다른 검색결과가 초래되는지를 설명하지는 못하였다. 따라서 이 논문에서는 이들 두 탐색방법에서 어떤 요소들이 다른 탐색결과를 낳게 하는지 이해하기 위하여 두가지의 다른 요소인 질문타입(query type)과 필드(field)가 조사되었다. 이 연구는 1차와 2차에 걸쳐서 진행되었는데, 먼저 1차연구인 실험연구는 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. 실험연구의 결과에서, 질문의 성격에 따라 탐색결과가 크게 달라진다는 것이 밝혀졌으므로 2차연구인 질적연구에서는 어떤 요소들이 이러한 결과의 차이를 초래하는지 규명하기 위하여 개념적인 질문에 대하여 데이타의 질적분석이 이루어졌다.

2. 실험연구

2.1 연구 설계

연구질문은 주제탐색과 인용탐색이 다학문간분야인 수자원 분야 문헌내에서 개념적/방법론적 질문을 위해 다른 내용의 문헌을 검색하는가이다.

먼저 용어의 정의를 보면, 개념적 질문(conceptual query)이란 연구자가 아이디어를 창출하거나 개념을 정의한다든지 또는 연구문제의 추출을 고려할 때 가지는 질문이다. 방법론적인 질문(methodological query)은 연구디자인을 개발하거나 데이타를 수집·분석할 때 또는 통계방법을 고려할 때 가지는 질문이다.

이 연구에서는 세가지 독립변인(탐색방법, 질문타입, 필드)이 다섯가지 종속변인(전체 검색문헌양, 검색된 적합문헌의 양, 유니크하게 검색된 문헌의 양, 유니크하게 검색된 적합문헌의 양, 정확률)에 어떻게 영향을 미치는가를 알아보기 위하여 실험연구(experimental design) 방법이 수행되었다.

첫번째의 종속변인인 전체 검색문헌의 양을

평가하기 위하여 세개의 가설이 설정되었다.

1. 주제탐색은 인용탐색에 비해 많은 양의 검색문헌을 가진다.
2. 개념적 질문에 대해 주제탐색은 인용탐색에 비해 많은 양의 검색문헌을 가진다.
3. 방법론적 질문에 대해 인용탐색은 주제탐색에 비해 많은 양의 검색문헌을 가진다.

다른 4개의 종속변인에 대해서도 같은 일련의 가설이 되풀이되었다.

검색효율을 비교하기 위하여 SWRA와 SCISEARCH(SOCIAL SCISEARCH 포함)의 CD-ROM 데이타베이스에 공통으로 포함되어 있는 243종의 잡지만을 비교대상으로 선정하였다. 미국 University of Wisconsin-Madison의 수자원분야 박사과정에 등록한 학생들을 자연과학, 공학, 사회과학의 세그룹으로 나눈 후 각각 7명씩 무작위로 추출하여 전체 21명의 학생을 선발하였다. 각자에게 이 연구의 목적과 진행과정을 설명하는 편지가 전달되었는데, 그 편지에는 두가지 타입의 질문 탐색요구서(개념적 질문을 위한 요구서와 방법론적 질문을 위한 요구서), 두 종류의 연구질문에 대한 설명과 예비연구에서 한 연구자에 의해 제공된 탐색요구질문의 샘플이 포함되었다. 따라서 21 연구자의 박사학위 논문주제에 근거한 42 질문이 추출되었고, 각 질문당 2개씩의 인용문헌(seed article)이 제공되어 84번의 데이타베이스 탐색(42 주제탐색과 42 인용탐색)이 이루어지게 되었다.

질문을 제공한 다수의 학생들은 논문의 선행연구 개관을 위하여 적절한 정확률과 높은 재현율을 요구하였으므로 주제탐색에서는 한명의 정보전문가가 "building block strategy"를 사용하여 탐색하였다. 인용탐색을 위해서는 각 질문당 2개의 seed article이 제공되었는데 필자가 그들을 가지고 cycling이 없이 "citation strategy"만 사용하여 수행하였다. Seed article의 제공기준은 참고문헌을 가진 잡지논문으로서 최소한 5년전에 출판되어야 한다는 것이다. 검색된 문헌중 오직 243종의 공통잡지에 수록된 문헌에 한해서만 적합성의 판단이 요구되었다.

질문을 제공한 학생 각자에게 적합성 판정을 위한 문서화된 판단기준이 주어지고, 같은 양의 정보인 저자명, 제목, 소스잡지명, 출판년에 한해 적합성의 판정이 요구되었다. 1개의 블럭간 실험처치(필드)와 2개의 블럭내 실험처치(탐색방법과 질문타입)를 규명하여야 하므로 Split-Plot Factorial design 통계기법이 사용되었으며, 어떠한 평균값이 변인들 사이에서 특별히 어떻게 다른가 하는 것을 밝히기 위하여 Scheffe Post Hoc 비교를 수행하였다.

2.2 자료의 분석 및 결과.

두 탐색방법에 의해 검색된 문헌의 평균양은 모두 25건으로 대체로 유사했으나, 개념적 질문에 대해 검색된 문헌의 양이 방법론적인 질문에 대해 검색된 문헌의 양보다 많았다. 필드가운데서는 자연과학분야에서 가장 많은 문헌(40건)을 검색했으며 그 다음이 공학(30건), 그리고 사회과학(6건)의 순이었다. 주제탐색은 자연과학 질문(37건)보다 공학질문(58건)에 더욱 효과적이었으며, 인용탐색은 공학(29건)

보다 자연과학(64건)에서 많은 문헌을 탐색해 내었으나 사회과학 질문을 위해서는 두 탐색 방법이 다 효과적이지 못했다 (주제탐색:인용탐색= 6건:7건).

일반적인 결과는 평균 오버랩이 아주 적다는 것을 보여주고 있으며 개념적 질문에 대한 평균 오버랩은 방법론적인 질문에 대한 오버랩보다 높았고 평균은 각각 1.9건과 0.4건이었다. 또한 그들은 아주 적합한(highly relevant) 문헌인 것으로 나타났다. 검색된 문헌 중 유니크한 문헌의 평균치는 24건이며 개념적 질문에는 두 탐색방법에서 거의 동일한 수의 유니크한 문헌이 검색되었으나 방법론적인 질문에는 주제탐색보다 인용탐색에서 더욱 유니크한 문헌이 많이 검색되었다.

개념적 질문을 위해서는, 인용탐색에 의해 검색된 문헌(0.50)이 주제탐색에 의해 검색된 문헌(0.37)보다 높은 정확률을 가졌으나, 방법론적인 질문을 위해서는 주제탐색에 의해 검색된 문헌이 0.43의 정확률을 가진데 반해 인용탐색에 의한 것은 0.39의 정확률을 가졌다. 주제별로는 공학분야 문헌(0.52)이 최고의 정확률을 가지고 자연과학분야 문헌(0.44), 그리고 사회과학분야 문헌(0.30)의 순이었다.

선행연구의 대부분은 주제탐색이 인용탐색보다 많은 양의 문헌을 검색하고 인용탐색은 보조적인 탐색방법으로 사용되었다고 보고하고 있으나(Hurt, 1982: Snow and Ifshin, 1984: Vidal-Arbona, 1986: Pao, 1993) 이 연구에서는 보다 많은 적합한 문헌이 인용탐색에서 검색되었다.

3. 질적연구

이 연구에서는 박사학위 논문의 제목만이 탐색질문으로 사용되어 질문의 수준이 통제되었음에도 불구하고, 질문의 내용에 따라 검색결과에 큰 차이가 있다는 것이 연구결과에서 밝혀졌다〈표 1〉. 따라서 어떤 요소들이 이러한 결과의 차이를 초래하는지 규명하기 위하여 개념적인 질문에 대한 데이타의 질적 분석이 이루어졌다. 특히 자연과학분야 질문(번호 1-7)과 공학질문(번호 8-14)에서 질문의 성격에 따라 검색결과가 크게 달라졌기 때문에 그 결과를 설명하기 위하여 특정성(specificity), 복잡성 (complexity), 최신성(recency)의 세가지 속성을 가지고 질문을 분석하였다. 이중 특히 특정성과 복잡성의 측정방법에 대한 아이디어는 Saracevic과 Kantor(1988)의 논문에서 얻었다. 또한 주제탐색과 인용탐색의 두 탐색방법이 어떠한 경우에 더욱 효과적인 결과를 낳는지 알아보고 그들의 특성을 좀더 자세히 살펴보기 위하여 각각의 방법에 대한 개별적 분석이 행해졌다. 그러나 사회과학 질문을 위해서는 SWRA와 SCISEARCH(SOCIAL SCISEARCH 포함) 모두에서 극히 소수의 논문만 검색되었으므로, 질문의 세가지 속성 파악은 오직 자연과학과 공학질문에만 적용되었고 사회과학 분야의 질문(번호 15-21)은 달리 취급하여 개별로 분석하였다.

3.1 특정성

특정성은 질문용어의 상세한 정도에 근거하

〈표 1〉 개념적 질문에 대한 두 탐색방법과 검색된 적합문헌의 양

질문번호	검색된 적합문헌의 양	
	주제 탐색	인용 탐색
1	0	38
2	12	3
3	13	66
4	144	49
5	3	6
6	61	0
7	13	21
8	19	2
9	6	12
10	39	49
11	12	10
12	2	3
13	3	2
14	41	18
15	5	18
16	10	0
17	1	3
18	4	7
19	0	2
20	1	1
21	1	4
평균	19	15

여 질문을 등급화한 것으로서, 특정성의 범위는 특정적인 것과 일반적인 것으로 양분되었다. 특정성과 검색된 적합문헌의 양의 사이에 어떤 관계가 있는가를 알아보기 위하여 특정성의 정도(degree)가 조사되었다. 처음에 연구질문을 제공했던 자연과학과 공학분야의 각 연구자에게 그들의 질문이 아직도 여전히 학위논문의 제목으로서 유효한 것인가 하는, 다시 말해 여전히 충분히 특정적인가 하는 물음이 주어졌다〈표 2〉. 높은 특정성을 가진 질문(특정적인 질문을 가진)을 위해 더욱 많은 적합한 문헌이 검색되리라고 예측되었다. 이 연구의 데이타는 질문들 사이에서 비대칭적인 분포를 보이고 있었기 때문에 정상분포를 가지고 있지 않았다. 따라서 비모수통계의 일종인 Wilcoxon 2 sample test를 사용해서 특정성과 검색된 적합문헌양과의 관계에 대한 가설을 테스트하였다. 낮은 특정성을 가진 일반적인 질문에 대한 적합문헌양과 높은 특정성을 가진 특정한 질문에 대한 적합문헌 양의 평균치는 각각 58건:9건이었다. 따라서 첫번째 가설 "주제탐색은 높은 특정성을 지닌 질문을 위해서보다 낮은 특정성을 지닌 질문을 위해 더욱 많은 적합문헌을 검색해낸다"는 $p<0.05(p=0.038)$에서 채택되었다. 이 연구에서 일반적 질문을 위한 적합문헌의 평균은 선행연구에서의 적합문헌의 평균과 근사했는데, 이 사실은 본 연구의 결과가 선행연구의 결과와 다른 결과를 초래한 이유를 설명하는것 같다. 다시 말해서 이 연구에서 주제탐색이 선행연구에 비해 적은 양의 적합문헌을 검색해 내었기 때문에 전체적으로 선행연구의 결과와 다른 결과를 초래하게 되었는데 그 이유는 이 연구에서의 질문이 선행연구에서의 질문에 비해 높은 특정성을 가지고 있는데 기인하는 것으로 생각된다. 여기서 알 수있는 것은 특정성과 적합문헌의 양은 반비례한다는 것이다. Lancaster(1979)에 의해 관찰된 바와 같이 높은 특정성을 가진 색인언어는 주제를 더욱 자

세히 정의하고 정확률을 증가시키나 재현율과 적합문헌의 양을 감소시킨다. 이 연구에서도 특정성이 낮은 질문에 비해 특정성이 높은 질문을 위해 약간의 문헌만이 검색되었다.

〈표 2〉 특정성의 정도와 검색된 적합문헌의 양

질문번호	질문번호	검색된 적합문헌의 양
일반적	4	144
	6	61
	10	39
	14	41
	평균	58
특정적	1	0
	2	12
	3	13
	5	3
	7	13
	8	19
	9	6
	11	12
	12	2
	13	3
	평균	9

3.2 복잡성(Complexity)

복잡성은 탐색전략과 관련된 탐색개념의 수에 근거하여 질문을 등급화한 것으로 표시된다. 탐색개념은 탐색용어(디스크립터)나 탐색파셋 (facet)의 수로 나타내어진다. 따라서 복잡성의 두가지 측정법은 탐색파셋의 수와 탐색용어의 수에 의해 이루어진다. 탐색파셋은 탐색자에 의해 파악된 개념(concept)을 나타내는데, 탐색파셋의 수는 불리안 연산자 AND로 구분되었다. 탐색식에서 탐색파셋이 많으면 많을수록 적합문헌의 양은 적어진다. 많은 탐색파셋을 가지면서 높은 복잡성을 가진 질문보다 더욱 적은 탐색파셋을 가지면서 낮은 복잡성을 가진 질문을 위해 보다 많은 적합문헌이 검색될 것이라고 기대되었다.

〈표 3〉 복잡성에 따른 탐색파셋수와 검색된 적합문헌의 양

복잡성	질문번호	탐색파셋의 수	검색된 적합문헌의 양
낮음	1	2	0
	4	2	144
	5	3	3
	6	3	61
	7	2	13
	8	3	19
	10	3	39
	평균	2.6	40.0
높음	2	4	12
	3	4	13
	9	4	6
	11	4	12
	12	4	2
	13	4	3
	14	4	41
	평균	4.0	14.1
전체평균		3.2	27.0

〈표 3〉에서 보는 바와 같이 질문의 복잡성

의 척도로서 먼저 탐색전략에 사용된 탐색파셋의 수와 검색된 적합문헌의 양을 가지고 비교하였다. 14질문에 대한 탐색파셋의 평균은 3.2로서 이 숫자는 복잡성의 비교를 위한 절단기준(cutoff point)으로 선택되었다. 탐색전략에서 4이상의 탐색파셋수를 사용한 질문은 높은 복잡성을 지닌 질문으로, 탐색전략에서 3이하의 탐색파셋수를 사용한 질문은 낮은 복잡성을 가진 질문으로 정의되었다. 두번째 가설 "3 이하의 탐색파셋수를 가지면서 낮은 복잡성을 가진 질문은 4 이상의 탐색파셋수를 가지면서 높은 복잡성을 가진 질문보다 많은 적합문헌을 검색해 낸다"는 Wilcoxon test를 사용해서 테스트되었으나 결과는 기각되었다. 따라서 탐색 파셋의 수와 적합문헌양 사이에는 아무런 상관관계가 없는 것으로 판명되었다.

복잡성의 또 하나의 척도는 탐색전략에 사용된 탐색용어의 수인데, 탐색용어의 수는 각 파셋내에서 불리안 연산자 OR로 연결된 것이다〈표 4〉. 하나의 탐색파셋은 동의어나 유사동의어 등으로 구성된 하나 이상의 탐색용어를 가질 수 있다. 여러 탐색용어들은 단어, 구, 디스크립터와 identifier 등으로부터 선택된다. 한 파셋에서 탐색용어가 많으면 많을수록 검색된 적합문헌의 양은 많아진다. 보다 많은 적합문헌들이 적은 탐색용어를 가지면서 낮은 복잡성을 가진 질문보다 많은 탐색용어를 가지면서 높은 복잡성을 가진 질문을 위해 검색되어질 것으로 기대되었다. 14질문에 대한 탐색용어의 평균은 9.1로서 이것 역시 복잡성의 비교를 위한 절단기준(cutoff point)으로 사용되었다. 질문이 복잡할 때 검색된 적합문헌의 양은 증가될 것이라고 기대되었다. 세번째 가설 "10 개 이상의 탐색용어를 가지고 높은 복잡성을 가진 질문은 9개 이하의 탐색용어를 가지고 낮은 복잡성을 가진 질문보다 많은 적합문헌을 검색해 낸다"는 Wilcoxon test에 의해 검증되었으나 결과는 기각되었다. 따라서 적합문헌의 양과 탐색용어수 사이에는 아무런 상관관계가 없음이 밝혀졌다. 검색된 적합문헌의 양의 차이를 이해하기 위하여 SWRA의 제작자에 의해 사용된 용어의 특성과 그들의 색인정책을 고려하는 것이 필요할 것으로 생각된다.

〈표 4〉 복잡성에 따른 탐색파셋수와 검색된 적합문헌의 양

복잡성	질문번호	탐색파셋의 수	검색된 적합문헌의 양
낮음	1	6	0
	4	6	144
	5	8	3
	6	7	61
	7	6	13
	8	9	19
	10	7	39
	14	9	41
	평균	9.0	40.0
높음	2	13	12
	3	13	13
	9	17	6
	11	10	12
	12	10	2
	13	10	3
	평균	12.2	8.0
전체평균		9.1	24.0

3.3 최신성 (Recency)

최신성은 seed article의 출판년도에 근거하여 결정되는 것으로, 효과적인 인용탐색을 위해서는 seed article이 출판된 후에 인용이 축적되어지도록 몇년이 지난 seed article을 선택하는 것이 좋다. 따라서 최신성과 검색된 적합문헌의 양은 반비례 관계에 있는 것으로 기대되었다. 데이타 결과에서 최근에 출판된 인용문헌(예를 들면 1989년과 1991년 사이에 출판된 논문의 인용)보다 오래된 인용문헌(예를 들면 80년대 초반에 출판된 논문의 것)을 사용하면 많은 문헌이 검색된다는 것이 밝혀졌다. 최근에 출판된 문헌의 인용을 위해 적합한 문헌양의 범위는 0에서 3이었다. 최근에 출판된 인용문헌을 사용해서 검색된 문헌의 양이 너무 적었기 때문에 두 집단 사이의 차이를 결정하기 위한 테스트는 실시되지 않았다. 따라서 Seed article의 성격이 검색결과에 크게 영향을 미친다는 사실이 관찰되었으나 seed article의 적절성 여부는 평가되지 않았다.

3.4 자연과학과 공학질문의 분석

질문의 세가지 특성인 특정성, 복잡성, 최신성 이외에, 주제탐색과 인용탐색에 의해 각각 잘 검색되는 질문과, 검색에 효과적이지 못한 질문이 있었다. 따라서 자연과학과 공학분야의 14개의 질문과 두 탐색방법에 의해 검색된 적합문헌에 대한 개별적인 분석이 이루어졌다.

3.4.1 주제탐색의 특징.

질문 4, 6, 10, 14를 위해 주제탐색에서 상당한 양의 적합문헌이 검색되었다. 이 4개의 질문은 그들을 제공한 각 연구자들에 의해 일반적인 질문으로 확인되었다. 특히 SWRA를 가지고 질문 4를 검색했을 때 144개의 적합문헌이 검색되었다. 이 사실은 SWRA를 사용한 주제탐색은 일반적인 질문에 적합하다는 것을 말하고 있다. 이 4 주제를 위해 색인된 문헌들은 이미 잘 발달된 주제들(well established subjects)에 관한 것으로서 이들은 SWRA에서 표준용어에 의해 색인되었다. 이 질문들을 위해 검색된 문헌의 결과는 전형적인 탐색질문에서 보여진 검색결과와 아주 유사했다.

질문 1, 2, 3, 5, 12, 13에 대해 주제탐색은 많은 문헌을 검색해 내지 못했는데, 질문 13을 제외한 모든 질문은 특정한 분야에 촛점을 맞추고 있는 구체적인 박사학위 논문제목들이었다. 주제탐색에 의해 적은 양의 문헌이 검색되었다는 것은 주제 디스크립터가 비교적 최근에 도입되었거나 탐색기간 동안에 그 용어로 색인된 문헌이 비교적 적다는 것을 반영할 지도 모른다. 예를 들면 "membrane technology"는 비교적 최근에 발달된 개념이고 SWRA에서는 오직 1문헌만이 이 디스크립터를 사용해서 색인되었다. 유사하게, 질문 12의 "atrazine" 용어를 사용해서 약간의 문헌만이 검색되었는데 그 이유는 "groundwater" 중에서 "atrazine"에 관한 연구가 이루어진 지는 약 10년도 채 못되었기 때문이다. 비록 "groundwater"는 오래된 주제이고 이 주제하에 많은 양의 문헌이 축적되었지만 SWRA는 새로 발달된 용어를

포함한 문헌을 별로 수록하지 않았기 때문에 적합한 문헌을 별로 검색하지 못한 것으로 보인다. 마찬가지로, 질문 3과 질문 5의 주제는 최근에 발달된 주제들이다. 비록 "water"와 "air"는 오랜 연구전통을 가진 용어들이지만 "water-air interface"는 새롭게 발달된 주제이다. 또한, "water"와 "sand"는 오래된 전통의 용어들이지만 "water-sediment interface"는 새로운 주제이므로 약간의 문헌만이 이 용어하에 색인되었다. 이 두 질문을 위한 탐색전략에서 이들 용어를 사용하여 탐색하였을 때 검색결과가 만족스럽지 않았다. 질문 2와 13은 그 질문을 제공한 연구자에게 상당히 많은 문헌이 검색되리라 기대되어진, 잘 발달된 연구분야로 인식되었다. 질문 13은 〈그림 1〉에서 보는 바와 같이 상당히 일반적인 질문으로서 각 파셋은 많은 양의 문헌을 검색했으나 탐색전략에서 불리안 연산자 3개의 AND로 결합했을 때 약간의 문헌만이 검색되었다. 만약 "characteristics"를 탐색전략에서 배제했더라면 보다 많은 문헌이 검색되었을 것이다.이 분석결과에서 특정적이고 좁은 범위의 탐색전략이 좋은 결과를 낳지 않을 때에는 일반적인 탐색전략이 때때로 필요하다는 것을 알 수 있다. 일반적인 탐색은 적절한 정확률을 유지하면서 때때로 재현율을 상승시키기도 하는 것이다. 질문 2를 위해서는 이 주제가 비록 공학에서 잘 발달된 주제에 속하나 적은 양의 문헌만을 검색해 내었는데, 그 이유는 질문을 제공한 연구자에 의해 확인된 바와 같이 이 분야에서 비교적 적은 연구가 이루어졌기 때문인 것으로 보인다.

〈그림 1〉 질문13을 위한 탐색전략

lake* and ((nutrient or carbon or nitrogen or phosphorus or silica or nitrate or phosphate) and sedimentation) and characteristics

3.4.2 인용탐색의 특징

질문 1, 3, 4, 10과 14는 인용탐색으로 좋은 결과를 가진 예들이다. 질문 1과 질문 3은 좁은 범위의 특정적인 질문이고 질문 4, 10, 14는 일반적인 질문이다. 인용탐색을 위해 제공된 대부분의 seed article은 1980년대 초에 발표된 논문들이었다.

질문 14는 인용탐색에서 가장 좋은 결과를 보여주는데, 그 질문은 그 분야의 연구자들에 의해 잘 발달된 전형적인 주제에서 나온 것이다. 그 연구주제는 상당한 기간동안 잘 발달된 분야이고 seed article의 저자는 그 분야에 탁월한 연구자들이었다. 질문 1과 3은 좁은 범위의 특정적인 질문이고, seed article을 사용하여 인용탐색을 수행하였을 때 상당한 양의 적합문헌이 검색되었다. "membrane technology"는 비교적 새롭게 발달된 일반적인 주제임에도 불구하고 주제탐색에서는 별로 효과적이지 못했으나 인용탐색의 검색결과는 좋은 것으로 나타났다. 또한 질문 3은 새로운 주제임에도 불구하고 인용탐색을 통해 적합한 문헌이 상당히 많이 검색되었다. 이 사실에서 좋은 seed article만 주어진다면 좁은 범위의 특정적인 질문이든 일반적인 질문이든간에 상관없이 인용탐색에 의해 비교적 많은 문헌이 검색된다는

것을 알 수 있다. 이점은 주제탐색에 대한 인용탐색의 장점인데, 주제탐색은 일반적인 질문에는 효과적이지만 인용탐색은 일반적인 질문뿐 아니라 특정적인 질문에도 효과적임을 보여 준다.

질문 2, 6, 8과 12를 위해서 인용탐색은 효과적이지 못했는데, 실제로 이들 질문을 위해서는 인용탐색에서 적합문헌이 전혀 검색되지 않거나 소수의 문헌만이 검색되었다. 질문 2는 좁은 주제분야의 질문이고 제공된 seed article들은 1982년과 1986년에 출판된 것들이지만 인용탐색에서 많은 문헌이 검색되지 못했다. 이러한 결과에 대해서 두가지 설명이 가능한 것으로 보인다. 첫째, 그들은 적절하지 못한 seed article일 지도 모르며 따라서 잘못된 seed article의 선정때문에 검색 결과가 만족스럽지 못할 지도 모른다. 둘째, 이 질문을 제공한 연구자가 예측했던 것처럼 이 분야에서는 비교적 적은 양의 연구가 행해졌을 지도 모른다는 추측이 가능하다.

질문 6, 8, 12를 위한 seed article은 1989년에서 1991년 사이에 출판된 것들로서, 이들은 문헌에서 인용되는 기회가 적었기 때문에 관련 연구를 연결하는 인용 네트워크가 잘 형성되지 않았던 것으로 생각된다. 질문 6과 질문 8의 주제들은 비록 잘 발달된 연구주제이지만 충분한 선행연구가 이루어지지 않았기 때문에 연구자들은 오래된 인용을 제공할 수 없었다고 진술했다. 특히 질문 12를 위해서는 주제탐색과 인용탐색 모두가 다 효과적이지 못했다. 이점은 학위논문 토픽의 일반적인 특징을 설명하는데 도움을 주고 있다. 즉, 학생들은 최신 연구분야에서 학위논문 토픽을 구하기를 좋아하며 이들 분야는 때때로 선행연구가 거의 이루어지지 않은 분야들이다. 사실상 "atrazine in groundwater"는 선행연구가 거의 누적되어 있지 않은 새롭고 좁은 범위의 주제였다.

3.5 사회과학 질문의 분석

질문 15에서 21까지는 사회과학 질문들로서, 불행히도 SWRA나 SCISEARCH (SOCIAL SCISEARCH 포함) 모두 사회과학 질문에 대해 효과적이지 못했다. 따라서 위에서 사용된 3가지 속성인 특정성, 복잡성, 최신성은 사회과학 분야 질문의 분석에는 적합하지 않았기 때문에 각각의 질문에 대한 간단한 분석이 행해졌다.

질문 15는 농업경영에서의 핫이슈가 되는 연구분야이다. 이 분야 연구를 위해서는 SWRA보다는 AGRICOLA가 농업분야의 중요한 데이타베이스이기 때문에 AGRICOLA를 사용하는 것이 더욱 적당한 것으로 보인다. 그 외에도 특정적인 용어인 "midwest"는 탐색전략을 제한시켰기 때문에 소수의 논문들이 SWRA로부터 검색되었다. 비슷한 상황이 질문 17에서도 발생하였는데, 특정한 용어인 "Amazon"이 탐색전략에서 제외되었다면 더욱 많은 적합한 문헌들이 SWRA로부터 검색되었을 것이다.

질문 18과 19를 위해서 많은 문헌이 SWRA에서 검색되지 못했는데, 그 이유는 그 데이타베이스는 이들 주제에 관계되는 색인어를 가

지고 있지 못했기 때문이다. SWRA에서는 질문 18의 "conflict", 질문 19의 "hydropower"와 "benefit", 질문 20의 "institution"하에는 많은 논문이 색인되지 않았다. 이 사실에서 수자원 분야의 사회과학 주제를 위한 검색에는 SWRA의 수록범위가 너무 제한되어 있다는 것을 알 수 있다.

질문 16과 질문 21을 위해서 SWRA의 각 탐색 파셋을 통해 상당한 양의 문헌이 검색되었다. 그러나 4개의 AND로 연결된 5개의 파셋을 가지고 탐색했을 때 소수의 논문만이 검색되었는데, 사실상 다른 질문과 비교했을 때 이들은 많은 수의 AND로 연결된 복잡한 질문들이었다. 따라서 만약 이 질문들을 위해 여기서 사용된 탐색전략보다 적은 AND를 사용한 일반적인 탐색전략이 사용되었다면 적어도 "groundwater management"와 "property rights"를 위해서 지금보다는 많은 적합문헌이 검색되었을 것이다.

질문 16을 위한 seed article을 제외한 모든 seed article들은 검색되기에 충분히 오래된 것들이었다. 그러나 인용탐색에서 어떤 seed article도 많은 문헌을 검색해 내지 못했다. 이것은 SOCIAL SCISEARCH의 수록범위가 수자원 분야의 사회과학 주제를 광범위하게 포함하고 있지 못하다는 을 의미한다. 질문 18은 조금 다른 경우인데, 비록 SOCIAL SCISEARCH는 상당한 양의 문헌을 검색했으나 SWRA는 이들 문헌을 전혀 포함하고 있지 않았기 때문에 비교자체가 이루어지지 않았다. 일반적으로 SWRA와 SOCIAL SCISEARCH는 수자원의 사회과학 분야 주제를 위한 충분한 문헌을 수록하고 있지 못했다. 이 사실로 미루어 사서들은 사회과학 분야 연구자들에게 현재와는 다른 서비스를 제공해야 하며, 데이타베이스 제작자들은 자체 데이타베이스 제작시에 수자원 분야의 사회과학적 영역의 넓은 범위를 수록하도록 해야 하겠다.

4. 결론 및 제언

4.1 결론

두 탐색방법으로부터 낮은 오버랩과 많은 유니크한 논문이 검색된다는 것은 각각의 탐색방법이 다른 세트의 적합문헌을 검색했다는 것을 의미한다. 한 문헌과 할당된 색인용어 사이의 관계는 인용하는 문헌과 인용된 문헌사이의 관계와는 다르다. 주제탐색에 의한 적합성은 서지레코드에서의 디스크립터와 키워드를 포함하고 있는 용어들로 구성된다. 인용탐색에 의한 적합성은 한 문헌을 인용하는 문헌들과 한 문헌에 의해 인용되는 문헌들로 구성된다. 그럼에도 불구하고 적합성의 다른 두 기준에 근거한 두 타입의 탐색은 주제적 적합성의 두 타입을 보여준다. 첫째는 디스크립터의 할당에 의해 개념적으로는 잘 정의되었으나 연구자들에게는 공통되는 기본 백그라운드 문헌을 결여한 것이고, 둘째 서지적 인용에 의해 결합되어 있고 연구자들에게는 잘 알려져 있으나 주제적 문제와 쉽게 인식되는 terminology를 결여하고 있는 것이다.

이 연구에서 검색결과의 차이는 탐색방법이

나 데이타베이스의 특징보다 질문타입의 특징에 크게 좌우된다는 것을 보여주고 있다. 특히 주제탐색은 높은 특정성을 가진 특정한 질문보다 낮은 특정성을 가진 일반적인 질문에 대해 더욱 많은 적합문헌을 검색해 낸다는 사실과, 학위논문의 주제는 높은 특정성을 가지고 있으며 이를 위해 인용탐색이 주제탐색보다 더욱 효과적이라는 사실의 발견은 추후의 탐색전략 수립에 시사하는 바가 크다고 하겠다. 그것은 인용탐색이 더 이상 주제탐색의 보조적인 탐색방법이 아니라 질문의 유형에 따라서 독립적으로 사용되어질 수 있는 탐색방법이라고 하는 사실이다. 최근에 와서 인용탐색에 대한 심층적인 연구가 Harter(1992)와 Yoon(1994)에 의해 시도되었다.

4.2 제언

이 연구의 결과는 수자원 데이타베이스의 탐색을 위한 실제적인 의미를 가지고 있으며 아래와 같은 여러가지의 제언을 할 수 있다.

4.2.1 탐색방법

동일한 질문을 위해서도 주제탐색과 인용탐색이 다른 set의 문헌을 검색했기 때문에 검색결과를 위하여 두 탐색방법은 상호 보완적이라는 과거연구의 결론은 이 연구에 의해서도 확인되었다. 이러한 차이는 두 데이타베이스의 색인원칙의 차이에서 기인할 지도 모른다: SCISEARCH의 내용은 그들의 연구에 도움을 준다고 판단하는 저자들의 인용에 의해 제공된다. 그러나 SWRA의 범위는 그 문헌 내용의 "aboutness"에 근거해서 전문적인 색인자에 의해 결정되는데 만약 색인언어와 이용자의 탐색용어가 일치하면 그들은 적합한 문헌으로 검색된다.

두 탐색방법이 적은 오버랩과 많은 유니크한 논문을 검색하고 특히 오버랩된 논문이 상당히 적합한 문헌이라는 사실은 탐색전략의 구축에 영향을 줄 수 있다. 즉, 적은 양의 아주 적합한(highly relevant) 문헌의 탐색을 요구하는 이용자를 위해서는 여러개의 탐색방법을 독립적으로 탐색한 후 그중 오버랩된 문헌만을 제공하는 것이다.

이 연구와 선행연구의 중요한 차이는 인용탐색이 주제탐색의 보조적인 탐색방법이 아니라 실제적인 중요한 탐색방법이라는 것이다. 분명히 SWRA로부터의 주제탐색은 잘 확립된 주제안에서 표준 terminology를 가진 용어에 의해 색인된, 일반적이고 넓은 주제범위의 질문에는 효과적이다. 그러나 최근에 발달된 주제를 위해서 또는 특정성을 지닌 질문(예를 들면 박사학위논문의 제목)에는 인용탐색이 더욱 효과적이었다. 인용탐색이 새로운 주제의 탐색에 효과적이라고 하는 사실은 인용탐색이 근거하고 있는 전제(즉 어느정도의 문헌의 축적이 되어야 한다는)를 생각할 때 매우 흥미있는 것으로 평가된다. 잘 선택된 seed article만 주어진다면 특정적인 주제거나, 넓은 범위의 주제이거나, 새로운 분야의 주제에 상관없이 인용탐색에 의해 잘 검색되었다. 따라서 seed article 선택의 여지가 탐색결과의 중요한 관건이 된다.

4.2.2 필드

수자원분야는 여러개의 학문분야가 합쳐진 다학문간 분야이므로 연구자들의 탐색질문은 각 주제영역에 따라 다를 것으로 가정되었다. 선행연구의 결과는 이들 주제영역의 연구자들은 각각 다른 정보요구와 정보이용 패턴 및 탐색질문을 가지고 있다고 보고하고 있기 때문에 연구자들의 탐색질문에 따라서 검색결과가 다를 것으로 가정되었는데, 자연과학, 공학, 사회과학의 3 주제분야에서 상당히 다른 양의 문헌이 검색됨으로써 이 가정은 사실로 증명되었다. 두 데이타베이스에서 가장 많은 양의 문헌이 자연과학분야 질문을 위해 검색되었으며 그 다음이 공학, 사회과학의 순이었다. 데이타베이스별로 좀더 자세히 말하면, 공학분야 질문은 SWRA에 의해 가장 잘 검색되었으며 자연과학분야 질문은 SCISEARCH에 의해 가장 잘 검색되었다. 그러나 사회과학분야 질문을 위해서는 SWRA와 SCISEARCH 모두 다 효과적이지 못했다. 특히 SWRA는 법률, 정책, 의사결정, 경제학 등을 포함한 수자원의 사회적 제 양상에 관한 충분한 정보를 포함하지 못하였으므로 사회과학 연구자들의 정보요구를 만족시키지 못하였다. 사회과학자들의 정보요구를 만족시키기 위해서 사서는 다른 종류의 자료를 제공해야 하며, 또한 데이타베이스의 생산자들은 사회과학적 제 양상을 데이타베이스 수록범위에 추가시켜야 할 것으로 생각된다.

자연과학분야 주제를 위해서는 왜 SCISEARCH가 더욱 많은 문헌을 검색했으며, 공학분야 주제를 위해서는 SWRA가 더욱 많은 문헌을 검색했는지는 분명하지 않다. 하나의 분명한 사실은 자연과학이 이론적 이슈에 촛점을 맞추는데 반해 공학은 연구의 응용에 더욱 관심을 갖고 있다는 것이다. 그러나 이 두 필드의 다른 검색결과가 데이타베이스의 수록범위의 차이에 의한 결과인지, 다른 색인정책에 의한 결과인지, 또는 다른 어떤 요소에 의한 것인지는 분명치 않다.

4.2.3 연구 질문타입

두타입의 연구질문인 개념적 질문과 방법론적 질문이 사용되었는데, 개념적 질문은 내용이 "aboutness"에 근거하고 있기 때문에 주제탐색에 의해 보다 잘 검색되어질 것이라 생각되었고, 저자들이 흔히 사용하는 방법론적 질문을 위해서는 인용탐색이 더욱 효과적일 것으로 예측되었다. 연구결과 SWRA와 SCISEARCH는 모두 개념적 질문에 잘 응답했으므로 SCISEARCH도 SWRA와 마찬가지로 주제 데이타베이스의 하나임이 밝혀졌다. 방법론적인 질문에는 인용탐색이 주제탐색보다 더욱 효과적이라는 가설도 사실로 증명되었다. 따라서 연구방법론을 고려하는 단계에서는 인용탐색의 사용이 좋은 결과를 가져올 수 있는 것으로 보인다.

또 하나의 이슈는 질문의 타입으로서, 비록 이 연구가 연구자들의 박사학위논문 제목만을 질문으로 선택함으로써 질문의 수준을 통제했으나 검색된 문헌의 양은 질문의 내용에 따라 큰 차이가 있었다. 이것은 질문 타입의 특징이 검색결과에 영향을 미친다는 것을 의미한다. 그러므로 효과적인 검색을 위해서는 이 연구

에서 사용된 질문의 세가지 특성 이외에 다른 특성들도 조사 연구되어야 할 것이다.

인용문헌

Ahn, Myeonghee Lee. 1993. Retrieval Effectiveness of Subject Descriptor and Citation Searching in the Water Resources Literature. Doctoral Dissertation. Madison, WI: University of Wisconsin-Madison.

Chapman, J. and Subramanyam, K. 1981. "Co-citation Search Strategy." Proceedings of the 2nd National Online Meeting. New York, 97-102.

Garfield, E. 1979. Citation Indexing-Its Theory and Applications in Science, Technology and Humanities. New York : John Wiley.

Harter, S.P. 1992. "Psycological Relevance and Information Science." Journal of the American Society for Information Science. 43:602-615.

Hurt, C.D. 1982. "Important Items in the Database: an Investigation of Two Methods of Identification." Online Review. 6(3):227-233.

Lancaster, F.W. 1979. Information Retrieval Systems: Characteristics, Testing and Evaluation. 2nd ed. New York: John Wiley & Sons.

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.

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.

Pao, M.L. 1993. "Term and Citation Searching : a Preliminary Report." Proceedings of the 9th National Online Meeting. New York, 177-179.

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.

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.

Smith, L. 1981. "Citation Analysis." Library Trend. 30(1):85-105.

Tenopir, C. 1984. Retrieval Performance in a Full Text Journal Article Database. Doctoral Dissertation. Urbana, IL: University of Illinois.

Vidal-Arbona, C. 1986. Comparing the

Retrieval Effectiveness of Free-text and Citation Search Strategies in the Subject of Technology Planning. Ph. D. Dissertation, Cleveland, OH: Case Western Reserve University.

Yoon, L.L. 1994. "The Performance of Cited References as an Approach to Information Retrieval." Journal of the American Society for Information Science. 45(5):287-299.

〈부록 1〉 주제탐색을 위한 개념적 질문

1. What are the advantages and disadvantages of developing and employing membrane technology for the remediation of hazardous organics in comparison to photodegradation or bioremediation?
2. What are the relationships between the flood characteristics and the characteristics of the flood plains?
3. What are the factors to control the transfer of polychlorinated biphenyls (PCBs) between the atmosphere and water bodies?
4. What are the effects of changing vegetation and soil characteristics on infiltration and surface runoff generation?
5. What charactiristics of sediments and waters control the movement of water and chemicals across a sediment-water interface in response to a density gradient?
6. How are the effects of land-use and climate on water quality predicted for a regional watershed?
7. What is the role of sediment particles in double-diffusion: could the suspended particles form the fingers in the convection?
8. What are the impacts of anthropogenic (human-induced) changes in land cover and use in a watershed on its hydrologic response: particular factors of concern are changes in land management that influence stormflow, base flow and infiltration?
9. What are the ways in which physical and chemical heterogeneities in aquifers affect contaminant removal in pump-and-treat systems?
10. What are trace metals transported in aquatic foodwebs?
11. What is the effect of sediment imputs into wetlands on wetland vegetation?
12. How does the fate of atrazine in groundwater relate to aquifer characteristics including aquifer residence times?
13. What are the important biological, chemical, and physical characteristics of a lake that influence the sedimentation of nutrient elements such as carbon, nitrogen, phosphorous and silica?
14. How do lakes respond to fish community manipulations? Does algal biomass decline following fish community manipulations?
15. What agricultural practices threaten the groundwater resources of the midwest US?
16. What are the essential components that should be included in a local(town, village, city, county) groundwater management plan?
17. In the Amazon river basin, how are the fisheries impacted by the ecology of the tropical rain forests, and how do the fisheries and local river systems affect human settlement patterns?

18. Do processes exist to apply techniques of alternative dispute resolution to problems of shared water resources?
19. What vested interests benefit from hydropower megaprojects?
20. Do relationships exist between variables associated with institutional arrangements for management of coastal zones, and institutional effectiveness as measured by management outcomes?
21. What role do property rights play in establishing policies for protecting groundwater from agricultural practices?

〈부록 2〉개념적 질문에 사용된 탐색전략

1. membrane technology and (bioremediat* or photodegrad* or photocatalysis or biodegradation or biocatalysis)
2. flood* and (peak flow or runoff volume or (peak and volume)) and (((drainage and (basin* or area)) or watershed* or flood? plain or catchment*) and (characteristic* or feature*))
3. (pcb? or polychlorinated biphenyl? or aroclor?) and ((water-air interface? or air-water interface?) or ((lake? or river? or reservoir?) and atmosphere* and (movement or transfer* or transport* or exchange*)))
4. (vegetation* or soil characteristics or land use) and (infiltration or runoff or storm-flow)
5. (SEDIMENT-WATER INTERFACE* or WATER-SEDIMENT INTERFACE*) and (DENSITY GRADIENT* or CONVECTIVE TRANSPORT or (DENSITY and (FLUX or CONVECTION* or TRANSPORT)))
6. (LAND USE or CLIMAT*) and (WATER QUALITY or WATER CHEMISTRY or WATER POLLUTION) and (WATERSHED or CATCHMENT)
7. (sediment* or suspended particles) and (convection or double diffus* or finger? or fingering)
8. (land management or land use) and (storm? flow or base? flow or infiltration or flood) and (catchment area or groundwater basin or watershed)
9. ("PUMP-AND-TREAT" or "PUMP AND TREAT METHOD" or PUMPING) and (AQUIFER* or GROUNDWATER or GROUND WATER) and ((CONTAMINANT* or POLLUT* or TOXIC CHEMICALS or HEAVY METALS or HAZARDOUS WASTERS or REMEDIAT* or BIOREDEDIAT* or CLEAN-UP or DECONTAMINANT*) and (HETEROGENEIT* or CONDUCTIVITY)
10. (trace metal or trace element or heavy metal) and (food and (pyramid or chain or web))
11. sediment? and (wetland? or marsh? or swamp? or fen? or sedge-meadow) and effect? and (plant? or vegetation or FLORA)
12. (atrazine or deethylatrazine or DEETHYLATED-ATRAZINE) and (aquifer or ground?water) and (characteristic or (time and (residence or travel or transport)))
13. lake and ((nutrient or carbon or nitrogen or phosphorous or silica or nitrate or phosphate) and sedimentation) and characteristics
14. (manipulat* or BIOMANIPULAT*) and (fish or fisheries* or fishery*) and LAKE

and (algal or algae or phytoplankton)

15. (agricultural or farm) and (contaminant? or weedkiller? or nitrogen or nitrates or fertilizer? or pesticide or pollutant or pollution or practice? or herbicide?) and (ground? water or aquifer?) and midwest*
16. (PLAN? or PLANNING) and (GROUNDWATER MANAGEMENT or ((GTROUNDWATER or AQUIFER*) and (MANAGE or MANAGING))) and MUNICIPALITIES or VILLAGE* or COUNTY or COUNTIES)
17. amazon and (fishery or fisheries) and (resettle* or settlement or rain forest? or forest? or human settlements or ecosystem? or human population or community development or social aspects)
18. ((dispute* or conflict*) and (resolution or settlement)) and water resources
19. (hydro?power or hydroelectric or electric power or water power) and benefit and (interest groups or interests or social aspects or social impact?)
20. (institution* and (coast or (coastal and (zone* or area* or region*))) and manage*) and ((constituency or constituent or citizen* or constituencies or stakeholder* or public*) and (effectiveness or effect? or result? or outcome?))
21. (property or ownership or land rights) and (aquifer? or ground?water) and ((agricult* and (activities or practice? or pollutant? or use or chemical?)) or herbicide? or pesticide? or agrichemical? or ((farm or farming or farmland or farmer?) and (policy or policies)