인용문헌에 의한 정보검색 효과에 관한 고찰*

이 란 주**

목 차

1. 서 론

정보검색에 있어서 인용문헌을 통한 검색은 인용하는 문헌과 인용되는 문헌 사이에 주제의 관련성이 내재한다는 전제하에 이루어지고 있다. 인용문헌 색인은 주제색인과는 달리 주제에 대한 지식 및 경험과 훈련을 크게 요구하지 않는다. Brooks(1986)에 의하면 "학자들이 그들의 문헌에 참고문헌을 기입할 때 문헌을 색인하고 있는 셈이고, 확실히 그들만큼 더 지식을 가지고 색인을 맡고 있는 사람이 있다고 볼 수 없다" (p. 34)고 언급했다. 그러므로 인용검색은 이렇게 자동적으로 인용하는 사람들에 의하

* 이 논문은 Indiana University, Graduate School, Doctoral Dissertation Grant-in-Aid (1991)와 School of Library and Information Science, Indiana University, Research & Development Fund (1990, 1991a, 1991b)에 의하여 연구된 연구자의 박사학위논문 (1992)에 근거하였음.

** 동덕여자대학교 문헌정보학과 조교수

접수일자 : 1994. 9. 30.

여 인용된 문헌들을 검색하므로 역시 색인할 때와 마찬가지로 검색하는 과정에서도 주제에 대한 지식 및 검색경험과 훈련을 요구하지 않는다는 큰 이점을 갖고 있다.

주제검색과 인용검색의 장단점에 대해서는 Pao(1984)와 Vidal-Arbona(1986)가 각각 그들의 논문에서 자세히 다루고 있다. 인용검색은 앞서 언급했드시 이용자가 검색하려는 주제에 대한 지식이 없이도 검색할 수 있기 때문에 가장 간단한 검색방법으로 인정되어 왔다. 한편 주제검색은 단어(words)와 어휘(phrases)의 개념을 표현하는데 의미론의 문제를 지니고있다(Pao, 1984). 인용검색은 단지 질문 (참고질문/검색주제/정보요구)에 관련이 있는 문헌을 사용하여 그 문헌을 인용하고 있는 모든 문헌들을 검색해 낼 수 있다. 인용문헌 검색의 장점들에도 불구하고, 이 검색방법은 현재 온라인 정보 시스템에서 주된 검색전략으로 사용되기 보다는 주제검색의 보완적인 역할에 그치고 있다. Harter는 "현재 인용문헌 검색은 아마 온라인 정보검색의 도구로써 충분히 이용되고 있지 않다"(1986, p. 59)고 지적하였다.

같은 주제에 대하여 인용검색과 주제검색을 실행한 후 그 검색효과를 비교한 몇몇 연구들이 행하여졌다(McCain, 1989; Goffman & Pao, 1980; Pao, 1986, 1988; Pao & Worthen, 1989; Vidal-Arbona, 1986). 이 연구들에서 두 검색방법은 각기 많은 다른 문헌들을 검색하였다. 즉 두 검색방법은 매우 적은 수의 공통된 문헌을 검색하였다. 또한 Vidal-Arbona(1986)의 연구를 제외하고 인용검색이 주제검색보다 적은 수의 적합(relevant)문헌들을 검색하는 가운데 인용검색은 주제검색보다 적은 수의 부적합(nonrelevant)문헌들을 검색하였다.

두 검색방법을 비교한 이전의 대부분의 연구들은 신뢰성 있는 비교를 위하여 가능한 많은 변수들을 통제하려고 노력하였다. 하지만 그들은 특히 인용검색에 사용된 인용문헌 수에는 관심을 기울이지 않았다. 검색과정이 검색 결과에 큰 영향을 준다는 연구 결과들이 있다(Fenichel, 1980-81; Harter, 1990). Harter의 사례연구(1990)에서도 적은 수의 검색 용어만을 사용하면 많은 적합한 문헌들을 검색하기 어렵다는 것을 잘 보여주고 있다. 직관적으로 생각할 때, 만일 인용검색에 있어 인용문헌의 수를 늘렸을 때 적합한 문헌의 수는 더 증가될 것으로 여겨진다(high recall). 반면 정확률(precision)은 떨어질 것이다.

인용검색에 있어 사용된 인용문헌(혹은 인용저자)의 수는 검색 결과에 중요한 요소일 수 있다. 그러나 아직까지 인용검색의 이러한 면을 깊이 있게 다룬 연구는 없었다.

본 논문은 세 사례연구를 통하여 온라인 인용검색에 있어 인용문헌의 수를 증가시킬 때 정보검색 효과에 미치는 영향에 대하여 조사하고 있다. 사용된 인용문헌의 수와 검색된 문헌들과의 관계, 검색된 문헌들의 중복에 대해서도 조사하며, 인용문헌의 형태가 미치는 검색 효과도 분석하고자 한다.

2. 연구 방법

2.1 연구 설계

기존의 온라인 검색을 다루고 있는 연구에서는, 실제 온라인 검색에서와 같이, 실제 이용자, 이용자가 제공하는 질문, 그리고 시스템의 삼대 요건으로 되어 있다. 인용검색을 위해서는 주로 실제 이용자로 부터 그들의 질문에 적합한 하나 혹은 몇개의 문헌들을 제공 받기도 한다. 본 연구에서는 기존의 연구 방법과는 달리 문헌 정보학 잡지에서 세 논문들을 채택하였다. 여기서 각각의 논문을 이용자가 제시하는 하나의 질문처럼 쓰고 있다. 그래서 질문은 채택된 논문에서 중점적으로 다루고 있는 주제를 의미한다. 따라서 본 연구에서는 명시된 질문은 없고, 채택된 논문을 정보요구(information need)로 가정한다. 채택된 논문과 논문에 실린 참고문헌들을 인용검색을 위한 검색 키(search key)로 이용한다. 인용검색을 위하여 DIALOG의 Scisearch와 Social Scisearch 데이터베이스를 이용하였다. 이 두 파일은 Institute for Scientific Information(ISI)이 만든 Science Citation Index와 Social Sciences Citation Index의 기계가독형 버젼(machine-readable versions)이다. 채택된 논문의 저자가 검색된 문헌들의 적합성 판정을 하였다. 본 연구에서는 다음과 같은 연구질문들이 구체적으로 조사되고 있다.

1) 인용검색에서 하나의 인용문헌이 사용되었을 때 정확률(precision)과 재현율(recall)의 의미에서 검색 효과는 무엇인가?
2) 인용검색에서 두개, 세개, 혹은 그이상의 인용문헌들이 사용되었을 때 정확률과 재현율의 의미에서 검색 효과는 무엇인가?
3) 인용검색에서 둘 이상의 인용문헌들이 사용되었을 때 검색 결과에서 어느정도의 중복을 예상하는가?

4) 인용검색에서 인용문헌 수가 증가할 때 어느정도의 체감효과가 나타나는지?

2.2 사례문헌 선정

본 연구에서 세개의 선택된 문헌들은 정보검색에서 이용자의 정보요구를 의미하는 질문으로 간주한다. 세개의 논문은 1980에서 1983년 사이에 Journal of the American Society for Information Science(JASIS)에 게제된 논문들 중에서 채택하였다. 선택된 논문들의 출판 년도는 그 문헌이 인용될 수 있을 만큼 햇 수가 지나야 하며 또한 논문 저자가 적합성 판정을 하는데 어려움을 격지 않도록 너무 오래되지 않아야 한다는 전제 조건에 맞추어 이 시기가 적절하다고 연구자가 결정하였다. 본 연구에서는 채택된 논문과 그것에 포함된 참고문헌들이 인용검색에 사용되므로 채택된 논문은 반드시 참고문헌을 포함하고 있어야 한다.

논문을 선택할 때 참고문헌 수도 고려해야 될 중요한 사항이었다. 참고문헌 수가 너무 적으면 연구 목적상 효과가 적기 때문이다. 한편 참고문헌 수가 너무 많으면 콤퓨터 프로그램을 돌리는 데 걸리는 시간을 고려할 때 비실용적이다. 따라서 참고문헌 수는 10개에서 25개까지를 적당한 수로 잡았다. 시간과 비용을 고려하여 미국내에 거주하는 저자의 논문만을 채택하기로 하였다. 게다가 채택된 논문은 하나의 검색 질문으로 이용되고 논문의 저자가 적합성 판정을 하므로 2명 이상의 저자들이 쓴 논문은 이 연구목적에 적합하지 않으므로 제외하기로 했다.

2.3 적합성 판정을 위한 콤퓨터 프린트

지나치게 많이 인용된 문헌(over-cited)이나 유명한 문헌(classic)들은 자료 분석에 있어 한쪽으로 너무 치우치게 하므로 제외 시켰다. 한 질문을 위하여 200개 이상의 문헌이 검색되었을 때는 판정자의 무리한 작업을 피하기 위하여 판정을 위한 문헌 수를 200 미만으로 정하였다. 이 결정은 이전의 연구에서 "정보의 과잉에 따르는 포용력" ("tolerances for information overload", Trivison, Saracevic, & Kantor, 1986, p. 343)에 따르고 있다. 이를 위하여 한 문헌(a cited reference)이 20개 이상의 문헌을 검색해 낼 때 단지 처음부터 20번째 검색되는 문헌까지만 포함 시켰다.

적합성 판정을 위하여 검색된 문헌들의 서지 사항을 나타내는 콤퓨터 프린트를 논

문 저자에게 보냈다. 서지 사항은 저자명, 서명, 잡지명, 출판 년도, 언어, 문헌 형태와 초록을 포함하고 있다. 1991년 이전의 Social Scisearch와 Scisearch 파일들 (files)은 초록을 제공하지 않기에 연구자가 직접 가능한 여러 출처들에서 초록을 구하려고 노력하였다.

2.4 적합성 판정

채택된 논문의 저자가 이용자로 간주되는 가운데, 그 저자는 검색된 문헌들이 그의 논문에서 중점적으로 다루는 주제(topics or problems)에 적합한지 판정한다. 본 연구에서는 Harter (1990)가 이용한 적합성(relevance)의 정의를 그대로 쓰고 있다. 즉 판정자가 '검색된 문헌의 서지 사항를 보고 만일 그것이 그의 논문 주제와 관련이 있어 그 문헌을 찾아 읽고 싶은 마음이 들면 적합한 문헌으로 보고 그렇지 않으면 부적합 문헌으로 판정'한다. 이와같은 적합성의 정의에 따라 판정자는 적합한 문헌에는 R (Relevant) 부적합 문헌에는 N(Nonrelevant) 그리고 판정이 어려운것에는 I (Impossible)를 주어진 서지 사항 옆에 표시한다.

정보검색 효과를 측정하기 위하여 재현율(recall)과 정확률(precision)을 사용하였다. 재현율은 시스템이 주어진 주제에 대하여 어느정도의 적합문헌을 검색할 수 있는지를 측정하는 것이다. 한 질문에 대한 적합문헌이 검색하려는 데이타베이스 안에 어느정도 있는지를 측정하는 간단한 방법은 없다(Pao, 1988). 그래서 본 연구에서는 한 질문을 위하여 사용한 모든 검색 키들에 의하여 검색된 적합문헌들을 총 적합문헌의 양으로 정하고 상대적 재현율(relative recall)을 사용하고 있다. 그래서 주어진 단일 질문을 위하여 재현율은 다음과 같이 계산되고 있다:

$$\text{재현율} = \frac{\text{주어진 하나의 인용문헌에 의하여 검색된 적합문헌수}}{\text{주어진 질문에 대한 적합문헌 총수}}$$

정확률은 시스템이 주어진 주제에 대하여 어느정도의 부적합문헌을 검색하지 않는 능력을 나타내는 것이다. 여기서 주어진 단일 질문을 위하여 정확률은 다음과 같이 계산되고 있다:

$$\text{정확률} = \frac{\text{주어진 하나의 인용문헌에 의하여 검색된 적합문헌수}}{\text{주어진 하나의 인용문헌에 의하여 검색된 문헌 총수}}$$

3. 사례문헌 성격

본 연구 방법에서 제시한 기준에 근거하여 Fenichel(1981)의 논문이 첫번째 사례연구로 선택되었다. Fenichel 논문의 연구목적은 온라인 검색의 성공과 관련된 검색의 습성(searching behavior)을 조사하는 것이다. 이 연구는 실험 방법을 이용한 비교적 특정적이고 좁은 주제를 다루고 있다.

Cooper(1983)의 논문이 두번째 사례연구로 선택되었다. Cooper 논문의 연구목적은 전통적인 정보검색 시스템의 결점을 극복할 수 있는 방법을 설계하는 것이다. 그는 질문들이 주제 용어 정확도 추정으로 가중치를 주거나 주지 않는 용어의 집합으로 구성된 검색 방법과 "극대화 엔트로피 원리(maximum entropy principle)"를 이용한 확율에 의하여 검색 결과를 등급하는 방법을 설계하였다. 이 연구는 다른 분야에서 빌려온

〈표 1〉 각각의 사례연구를 위한 세 문헌들

사례연구 1	Fenichel, Carol Hansen. "Online searching: measures that discriminate among users with different types of experiences." Journal of the American Society for Information Science, 32, 1 (1981): 23-32.
사례연구 2	Cooper, William S. "Exploiting the maximum entropy principle to increase retrieval effectiveness." Journal of the American Society for Information Science, 34, 1 (1983): 31-39.
사례연구 3	Yerkey, A. Neil. "A cluster analysis of retrieval patterns among bibliographic databases." Journal of the American Society for Information Science, 34, 5 (1983): 350-355.

통계학적 이론인 극대화 엔트로피 원리을 이용한 이론적인 연구이다.

Yerkey(1983)의 논문이 세번째 사례연구로 선택되었다. Yerkey의 연구는 더 효과적인 검색전략의 개발을 제공하기 위하여 데이타베이스들의 유사점과 상이점을 조사하려고 시도하였다. 실험 방법을 통하여 많은 다른 분야에서도 적용되고 있는 군집 분석(cluster analysis)기술을 이용하였다.

4. 자료분석과 결과

판정자의 적합성 판정을 받은 후, 파스칼(Pascal) 프로그램으로 짠 콤퓨터 프로그램으로 한 파일(a file)을 만들었다. 이 파일은 인용검색으로 검출된 인용들의 ISI 액서스 넘버(ISI accession number - ISI가 각각의 논문에 붙힌 고유번호), 문헌의 서명 그리고 적합성 판정을 포함하고 있다. 또 다른 파스칼 컴퓨터 프로그램으로 기술적(descriptive)요약 통계를 계산하였다. 두 변수들 사이의 상관관계는 SPSS를 사용하여 검증하였다.

표 2는 세 사례연구의 전반적인 통계를 보여 주고 있다. 첫번째 사례연구는 두번째와 세번째의 사례연구들 보다 검색된 문헌들에 있어 훨씬 적은 수의 잡지 종류와 ISI의 주제부류를 나타내고 있다. 첫번째 사례연구에 사용된 인용문헌들은 대부분 문헌정보학에서 인용되었다(91%). 한편 두번째와 세번째 사례연구들에서 사용된 인용문헌들은 단지 57%와 56%만이 각각 문헌정보학에서 인용되었다.

자료분석은 네 부분으로 구성되어 있으며 이 분석은 각각의 세 사례연구들에 적용되었다: (1) 각각의 인용문헌에 의한 재현율과 정확률; (2) 둘 이상의 인용문헌의 결합에 의한 검색 중복도; (3) 재현율 극대화를 위한 노선; 그리고 (4) 두 인용문헌에 의한 검색 중복도. 본 논문에서는 단지 첫번째 사례연구의 표만 보여 주고 있다. 좀 더 상세한 연구 결과에 하여 알고자 하는 독자들은 연구자의 박사학위 논문(Yoon, 1992)을 참조하기 바란다.

4.1 각각의 인용문헌에 의한 재현율과 정확률

첫번째 사례연구에 사용된 23개의 인용문헌들 각각에 의하여 검색된 문헌들의 집합

〈표 2〉 세 사례연구들의 전반적인 통계

	사례연구 1	사례연구 2	사례연구 3
원 논문의 출판년도	1981	1983	1983
원 논문의 참고문헌수	25	22	12
검색키로서 인용문헌수	23	22	12
판정된 총문헌수	178	252	144
Social Scisearch로 부터 검색된 문헌수	175	238	124
Scisearch로 부터 검색된 문헌수	3	14	20
고유한 잡지명	40	88	72
ISI의 잡지 주제부류	15	38	40
ISI 문헌정보학 주제에 속하는 부류의 %	91%	57%	56%
평균 정확률	0.52	0.21	0.24

을 '이니셜 포스팅 세트'(an initial postings set)라고 하며 S1 - Sn(표 3 참조)으로 표기한다. 각각의 인용문헌에 의한 재현율과 정확률은 판정자의 판정에 기준을 두고 계산한다. 이 계산에서 각각의 인용문헌의 검색 결과는 다른 인용문헌의 검색 결과와 관계됨이 없이 따로 계산된다.

첫번째 사례연구에서는 검색된 178개의 문헌 중에서 84개가 적합한 것으로 판정을 받았다. 23개의 인용문헌의 평균 정확률은 0.52이다. 표 3은 23개의 이니셜 포스팅 세트(S1 - S23)의 검색 결과를 보고하고 있다. 즉 검색된 문헌수, 검색된 적합 문헌수, 정확률 그리고 재현율을 나타내고 있다. 23개의 인용문헌 중 13개가 정확률에 있어 0.50 보다 높은 값을 보였다.

23개의 인용문헌들의 검색 결과가 인용문헌의 형태와 인용문헌이 원 논문(source article)에서 인용된 위치에 의하여 분석되었다. 가장 검색 효과가 높은 인용문헌의 형태는 잡지 논문(journal articles) 2개와 박사학위 논문(dissertations) 2개이었다. 이 인용문헌들은 원 논문의 서문에서 인용되고 있다. 중간 정도의 검색 효과를 보인 인용문헌의 형태는 학회 회의록(proceedings)이 2개, 잡지 논문이 2개, 마지막 보고서(a final report)가 1개, 메뉴얼(a manual) 1개 그리고 리뷰 논문(a review article) 1개로 되어 있다. 인용된 위치에서 살펴 보면 서문, 방법론 그리고 결과를 다루는 부분에서 인용

되고 있다. 가장 효과가 낮은 인용문헌의 형태는 가이드(guides) 2개, 학회 회의록 2개, 연보(annual reports) 3개, 마지막 보고서 2개 그리고 잡지 논문이 3개이다.

두번째 사례연구에서는 검색된 200개의 문헌들 중에서 적합한 문헌 수는 19개이었다. 22개의 인용문헌의 평균 정확률은 0.21이다. 최고의 재현율은 0.47이다. 22개의 인용문헌에서 2개만이 0.50 이상의 정확률을 보이고 있다.

〈표 3〉 23개의 이니쉘 포스팅 세트들의 검색 결과 (사례연구 1)

이니쉘 포스팅 세트*	검색된 문헌수	적합 문헌수	정확률	재현율
S1	20	16	0.800	0.190
S2	19	15	0.789	0.178
S3	16	15	0.937	0.178
S4	7	4	0.571	0.048
S5	6	5	0.833	0.060
S6	2	1	0.500	0.012
S7	3	2	0.667	0.024
S8	9	4	0.444	0.048
S9	15	6	0.400	0.071
S10	19	15	0.789	0.179
S11	6	2	0.333	0.024
S12	19	10	0.526	0.119
S13	3	1	0.333	0.012
S14	3	2	0.667	0.024
S15	19	9	0.474	0.107
S16	2	2	1.000	0.024
S17	20	4	0.200	0.048
S18	5	4	0.800	0.048
S19	2	0	0.000	0.000
S20	3	0	0.000	0.000
S21	20	11	0.550	0.131
S22	1	0	0.000	0.000
S23	20	8	0.400	0.095

* 이니쉘 포스팅 세트(an initial postings set)란 각각의 인용문헌에 의하여 검색된 문헌들의 집합을 말함

두번째 사례연구에서는 가장 검색 효과가 높은 2개의 인용문헌의 형태는 잡지 논문이다. 중간 정도의 검색 효과를 보인 인용문헌의 형태는 기술 보고서(a technical report)가 1개와 8개의 잡지 논문들이다. 가장 검색 효과가 낮은 인용문헌의 형태는 한 개의 박사학위 논문과 5개의 도서(books)들과 5개의 잡지 논문들이다. 인용문헌들 중에서 가장 검색 효과가 낮은 인용문헌들은 원 논문에서 방법론을 다루는 부분에서 인용되었다.

세번째 사례연구에서는 검색된 144개의 문헌들 중에서 적합한 문헌 수는 18개 이었다. 12개의 인용문헌의 평균 정확률은 0.24이다. 최고의 재현율은 0.28이다. 단지 한 인용문헌만이 0.50 이상의 정확률을 보이고 있다.

세번째 사례연구에서는 가장 검색 효과가 높은 4개의 인용문헌의 형태는 잡지 논문이다. 중간 정도의 검색 효과를 보인 인용문헌의 형태는 3개의 잡지 논문들과 1개의 학회 회의록이다. 가장 효과가 낮은 인용문헌의 형태는 2개의 잡지 논문들과 1개의 시소러스(a thesaurus)와 1개의 콤퓨터 프로그램(a computer program)이다. 가장 검색 효과가 많은 인용문헌들은 원 논문에서 서문이나 방법론을 다루는 부분들에서 인용되었다. 한편 가장 검색 효과가 낮은 인용문헌들은 방법론을 다루는 부분에서 인용되었다.

4.2 두개 이상의 인용문헌들의 결합에 의한 검색 중복도

인용검색에 있어 두개 이상의 인용문헌들을 이용했을 때 검색된 문헌들의 중복도를 조사하기 위하여 한 질문에 사용된 인용문헌들의 결합에 의한 검색 중복도를 조사하고 있다. 만일 주어진 질문을 위하여 N개의 수 만큼의 인용문헌들을 이용한다면, 가능한 결합의 수는 $2^N - 1$(Gluck, 1990; Harter, 1990). 예를 들면, 주어진 질문에 세개의 인용문헌들이 이용되었다면, 일곱개($2^3 - 1 = 7$)의 가능한 결합이 형성된다: 1; 2; 3; 1, 2; 1, 3; 2, 3; 그리고 1, 2, 3.

첫번째 사례연구를 위한 표 4는 다섯 컬럼으로 되어 있다: 인용문헌 결합의 수; 평균 재현율; 증가된 재현율; 평균 정확률; 증가된 정확율. 표 4의 두번째와 네번째 컬럼에서 보여 주듯이 하나의 인용문헌에 의한 재현율과 정확률만 제외하면 검색에 사용된 인용문헌의 결합 수가 늘어 날수록, 재현율은 증가하고 정확률은 줄어든다. 평균 재현율은 평균 정확률과 높은 음적 상관관계(strong negative correlation)를 보여준다

($r = -.98$, $p < .001$). 인용문헌중 거의 반이 결합되었을 때 평균 재현율이 0.56이고 평균 정확률은 0.50이다. 따라서 80%의 재현율을 얻기 위해서는 23개의 인용문헌들 중 18개가 필요함을 보여 주고 있다.

두번째와 세번째 사례연구들에서도 재현율과 정확률의 값의 차이는 있지만 첫번째 사례연구와 비슷한 결과를 보이고 있다. 여기서도, 인용문헌의 결합 수가 늘어 날수록, 재현율은 증가하고 정확률은 줄어든다. 두번째와 세번째의 사례연구들에서 재현율과

〈표 4〉 23개의 인용문헌들의 결합에 의한 평균 재현율과 평균 정확률

(사례연구 1)

인용문헌 결합수	평균 재현율	증가된 재현율	평균 정확률	증가된 정확률
1	0.070	0.000	0.522	0.000
2	0.132	0.061	0.539	-0.017
3	0.188	0.056	0.536	0.003
4	0.241	0.053	0.531	0.005
5	0.291	0.050	0.526	0.005
6	0.340	0.049	0.522	0.005
7	0.387	0.047	0.518	0.004
8	0.432	0.045	0.514	0.004
9	0.476	0.044	0.510	0.004
10	0.518	0.043	0.507	0.003
11	0.560	0.042	0.503	0.003
12	0.600	0.041	0.500	0.003
3	0.188	0.056	0.536	0.003
13	0.640	0.040	0.498	0.003
14	0.679	0.039	0.495	0.003
15	0.717	0.038	0.493	0.002
16	0.754	0.037	0.490	0.002
17	0.791	0.037	0.488	0.002
18	0.827	0.036	0.486	0.002
19	0.863	0.036	0.484	0.002
20	0.898	0.035	0.482	0.002
21	0.932	0.035	0.480	0.002
22	0.966	0.034	0.479	0.002
23	1.000	0.034	0.477	0.002

정확률 사이의 피어슨 상관계수(r)는 각각 -.96 ($p < .001$)과 -.95 ($p < .001$)이다. 따라서 이 두 변수들은 세 사례연구들 모두에서 높은 음적 상관관계를 보이고 있다. 두번째 사례연구에서는 80%의 재현율을 얻기 위해서 22개의 인용문헌들 중에서 13개가 필요하고 세번째 사례연구에서는 12개 중에서 9개가 필요하다.

표 4의 세번째와 다섯번째 컬럼들은 인용문헌에 의하여 검색된 적합한 문헌들 중 중복된 문헌들 때문에 나타나는 체감효과를 반영하고 있다. 인용문헌의 결합의 수가 늘어남에 따라, 이전에 검색된 문헌들과 중복되어 검색되는 문헌들이 있기 때문에, 점점 더 적은 수의 적합문헌들이 검색된다. 그 결과로 재현율과 정확률의 증가되는 값은 점점 줄어든다. 이 결과는 두번째와 세번째 사례연구들에서도 같은 현상을 보여주고 있다.

4.3 재현율 극대화를 위한 노선 (Potential Maximizing Recall Route)

재현율 극대화를 위한 노선은 가능한 최고의 높은 정확률을 가진 인용문헌 검색을 시작으로 최대의 높은 재현율에 도달할 수 있는 알고리즘 (algorithm)을 말한다. 이 알고리즘은 검색자들이 검색 전에 이미 인용문헌들의 검색 결과를 안다는 가정하에 이루어진다. 만일 검색자들이 검색 전에 미리 검색 목적을 세운다면, 이 노선을 이용하여 그 목적을 이룰 수 있다.

재현율 극대화를 위한 노선의 알고리즘은 원래 Harter의 연구 (1990)에서 처음 구상되었는데 이는 세단계 과정으로 이루어져 있다. 다음에서 첫번째 사례를 이용하여 그 과정을 소개하기로 한다.

(1) 표 5에서 보듯이 처음에 S16 - 최고의 높은 정확률을 얻은 이니셜 포스팅 세트가 선택된다: 만일 최고의 높은 정확률이 둘 이상일 때는, 그들 중에서 최고의 재현율을 얻은 세트를 선택해야 한다.

(2) 두번째로 높은 정확률로 S3이 선택되었다. 다음 단계는 S3과 S16의 합을 내고, 여기서 중복된 문헌은 제외된다.

(3) 이제 그다음으로 높은 정확률을 가진 세트를 선택하여 계속 합을 낸다. (2) 단계에서 보여준 대로 계속 같은 방법으로 재현율 극대화를 위한 노선을 만들기 위하여 마지막 세트인 S20이 합해 질 때까지 계속 한다.

〈표 5〉 재현율 극대화를 위한 노선 (사례연구 1)

가산된 이니셜포스팅	축적검색된 문헌수	축적검색된 적합 문헌수	축적된 재현율	축적된 정확률
S16	2	2	0.024	1.000
S3	18	17	0.202	0.944
S5	22	20	0.238	0.909
S1	41	35	0.417	0.854
S18	45	38	0.452	0.844
S2	57	47	0.560	0.825
S10	68	55	0.655	0.809
S7	70	56	0.667	0.800
S14	72	57	0.679	0.792
S4	75	57	0.679	0.760
S21	87	62	0.738	0.713
S12	102	68	0.810	0.667
S6	103	68	0.810	0.660
S15	117	72	0.857	0.615
S8	123	73	0.869	0.594
S23	141	80	0.952	0.567
S9	150	82	0.976	0.547
S11	152	82	0.976	0.539
S13	153	82	0.976	0.536
S17	170	84	1.000	0.494
S19	172	84	1.000	0.488
S22	173	84	1.000	0.486
S20	176	84	1.000	0.477

표 5가 보여 주듯이 재현율은 꾸준히 증가하고, S17이 합해졌을 때 최고의 재현율 1을 보이고 있다. 정확률은 최고의 값인 1에서 시작하여 서서히 줄어 들어 마지막 S20이 합해졌을 때 0.48을 나타낸다. 재현율 극대화의 노선이 보여 주듯이 S12가 합하여졌을 때 80%를 조금 넘는 재현율을 보이고 S17이 합해졌을 때 100% 의 재현율을 기록했다. 여기서 S19, S22, S20은 재현율 증가에 전혀 영향을 미치지 않았기 때문에 비생산적인 검색결과를 낳았다. 이러한 결과는 두번째와 세번째의 사례연구들에서도 비슷했다.

본 연구에서 정의된 재현율 극대화를 위한 노선은 검색자들의 검색 목적을 이루도록 도와 주는 한 수단이 될 수 있다. 본 연구에서 기술된 알고리즘 외에도 다양한 검색 목적들을 위한 다른 알고리즘들이 개발될 수 있다. 검색자들이 미리 인용문헌들의 검색 결과를 알기는 어렵기 때문에 현 시점에서는 이 재현율 극대화를 위한 노선이 실제적인 도구(a practical tool)이기 보다는 이론적인 설계(a theoretical model)이다.

4.4 두 인용문헌들에 의한 검색 중복도

한 질문에서 사용된 두 인용문헌들에 의하여 검색된 적합한 문헌들의 중복도 조사는 두 인용문헌들의 관계를 살펴 보기 위하여 중요한 자료를 제공 한다. N개 만큼의 인용문헌들의 모든 페어와이즈 비교(pairwise comparisons)를 계산한 결과, 두 인용문헌들, i, j의 중복도는 다음과 같이 계산되었다.

$$\text{중복도}_{(i,j)} = \frac{\text{i와 j에 의하여 공통적으로 검색된 적합문헌수}}{\text{i에 의하여 검색된 적합문헌수}}$$

이 측정을 "에시메트릭-오버랩"("asymmetric-overlap")이라고 부르며, 다른 연구들에서도 사용되었다(Harter, 1990; Katzer et al., 1982; Saracevic & Kantor, 1988).

세 사례연구 모두에서 두 인용문헌들에 의하여 검색된 적합문헌들의 중복도는 낮은 경향을 보이고 있다. 이 분석에서 비교적 높은 중복도를 보이는 두 인용문헌들은 비슷한 문제를 다루고 같은 방법론을 사용하고 있었다. 이 결과가 보여 주는 것은 중복도의 정도가 두 인용문헌들이 다루고 있는 주제의 유사성을 의미한다. 세번째 사례연구에서 세개의 인용문헌들이 서문에서 인용되었고, 네 개의 인용문헌들이 방법론 부분에서 인용이 되었는데, 이 두 구룹의 문헌들에 의하여 검색된 문헌들 사이에서는 전혀 중복된 문헌들이 없었다.

5. 정보검색에 사용된 인용문헌 수와 인용문헌 형태가 미치는 효과

세 사례들 모두가 그들 사이에 재현율과 정확률의 양적인 차이는 있지만 일관된 결

과들을 보여 주었다. 본 연구의 결과들로 부터 몇가지 결론들을 끌어 낼 수 있고 미래의 연구들을 위하여 몇가지 가설들이 제언되고 있다.

본 연구의 발견들은 정보검색에 있어 더 많은 인용문헌들이 사용될수록 평균 재현율은 증가하고, 평균 정확률은 감소한다는 것이다. 이 두 변수들은 높은 음적 상관관계을 보이고 있다. 한편 인용문헌들에 의하여 검색된 문헌들 사이의 중복으로 체감효과가 나타났다. 여기서의 체감효과란 더 많은 인용문헌들이 사용될수록 점점 더 적은 적합문헌들을 얻는 것이다. 따라서 더 많은 인용문헌들이 사용될수록 더 많은 비용이 들고 검색 효과는 점점 줄어드는 비용 효과적인 측면을 고려하게 된다. 그러나 인용문헌들이 하나 하나 첨가 되어 검색에 이용될 때 이전의 검색된 문헌들과의 중복도는 아주 낮다는 것을 지적하고자 한다.

또 다른 발견은 두 인용문헌들에 의하여 검색된 적합문헌들의 중복이 매우 적다는 것이다. 어떤 쌍은 그들 사이에 전혀 중복이 없었다. 이러한 발견들은 다음과 같은 임시적인 결론을 얻을 수 있다:

> 적합문헌을 검색해 낼수 있는 가능성을 가진 많은 수의 인용문헌 중에서 단지 적은 수의 인용문헌들만 검색에 사용한다면 높은 재현율은 기대하기 어렵다.

인용문헌의 검색 결과가 재현율과 정확률에 의하여 크게 차이를 보였다. 다시 말하면, 어떤 인용문헌들은 다른 문헌들 보다 훨씬 검색 효과가 좋았다. 그래서 인용문헌들의 성격을 분석한 결과 인용문헌의 형태와 인용문헌이 원 논문에서 인용된 위치에 따라서 검색 효과가 달라지며 이 두 변수들이 인용검색의 효과를 예견하는 좋은 지표가 될 수 있음을 보여 주었다. 잡지 논문 형태가 연보와 같은 다른 형태들 보다 효과적이었다. 두번째 사례연구에서는 도서 형태가 잡지 논문 보다 덜 효과적이었다. 이를 뒷받침하는 이론적 근거는 잡지 논문 형태가 다른 형태들 보다 주어진 질문의 주제나 문제들을 좀 더 깊이 기술할지도 모른다는 것이다. 이 발견은 인용문헌의 형태가 인용검색에 있어 효과를 예견하는 좋은지표가 될 수 있음을 말해 준다.

인용문헌의 검색 효과를 예견하는 또 다른 지표는 원 논문에서 인용문헌이 인용된 위치이다. 서문, 요약, 혹은 결론을 다룬 부분에서 인용된 인용문헌들이 방법론을 다룬 부분에서 인용된 인용문헌들 보다 검색 효과가 더 높다. 이것은 서문, 요약, 혹은 결론 부분에서 인용된 인용문헌들이 방법론의 부분에서 인용된 인용문헌들 보다 중심 주제

라는 의미에서 그들이 원 논문과 더 가깝게 관련되어 있다는 것이다. 이 발견에 근거하여 다음과 같은 가설이 제안된다.

> 검색자들이 인용검색에서 검색 키로써 원 논문의 참고문헌들을 사용하기로 한다면, 원 논문에서 서문, 요약, 혹은 결론 부분에서 인용된 참고문헌들의 사용이 방법론 부분에서 인용된 참고문헌들의 사용보다 정보검색을 향상시킨다는 점에서 더 효과적이다.

또한 Voos와 Dagaev(1976)도 많이 인용된 네개의 잡지 논문들을 이용해 잡지 논문들에서 인용된 위치를 다루었다. 그들의 발견도 많이 인용되는 문헌들은 다른 위치보다도 주로 원 논문의 서문에서 인용된 인용문헌들이었다. 그들은 인용된 위치가 "인용의 가치"로써 인식될 수 있다고 결론지었다.

평균 정확률에 있어 첫번째 사례연구는 세번째 사례연구보다 더 효과적이었다. 이것의 원인으로써 가능한 한 요소는 첫번째 사례연구의 판정자가 세번째 사례연구의 판정자보다 좀 더 관대하였다고 볼 수있다. 다른 가능성은 첫번째 사례연구에서는 인용문헌들이 세번째 사례연구에서 보다도 많은 수가 서문에서 인용되었다. 인용문헌들의 원문의 위치 분석은 다음과 같은 가설로 이끈다.

> 만일 검색자들이 인용검색에 있어 원 논문에 포함된 참고문헌들을 검색 키로 사용하기로 한다면 방법론의 부분에서 보다는 서문에 더 많은 참고문헌들이 포함되어 있는 원 논문이, 서문보다는 방법론 부분에 더 많은 참고문헌을 포함하고 있는 원 논문보다, 재현율과 정확률을 향상시키는 의미에서 더 효과적이다.

인용문헌들의 출판년도와 그들의 검색 능력의 관계에 대하여 본 연구는 세가지 방법들로 조사하였다: (1) 출판년도와 정확률; (2) 출판 년도와 재현율; (3) 중간 나이(median age/median year)와 세가지 검색능력.

표 6과 7이 지적하듯이 인용문헌의 나이는 좋은 지표가 되지 못 하였다. 다시 말하면, 반드시 오래된 인용문헌이 최신문헌보다 더 많이 적합 문헌을 검색하는 것은 아니다.

〈표 6〉 세 사례들에서 출판년도와 정확률/재현율과의 상관관계

	사례수	출판년도	확 률
사례연구 1			
정확률	23	.1469	.503
재현율	23	.4908	.017*
사례연구 2			
정확률	22	-.5300	.011*
재현율	22	.3699	.090
사례연구 3			
정확률	12	.0523	.872
재현율	12	-.3607	.125

양측검증: *P 〈 .05.

본 연구에서 세개의 잡지 논문들이 주어진 질문을 대신하고 있다. 이 세 논문들은 주어진 질문에 대하여 주제 혹은 문제를 가장 잘 대표하는 인용문헌으로 가정되고 있다. 이러한 문맥에서 첫번째와 두번째의 사례연구들에서 질문으로 선택된 잡지의 논문들과 이들에 포함된 참고문헌들을 비교할 때 원 논문들이 참고문헌들보다 더 검색 효과가 좋았다. 이 발견은 주어진 질문의 주제를 잘 대표하는 인용문헌들이 문제를 덜 대표하는 인용문헌들보다 효과적인 것을 나타낸다. 또한 이 발견은 다음과 같은 일반적인 가설을 제시한다.

> 인용검색에 있어 원 논문과 그의 참고문헌이 검색 키로 사용되었을 때, 정보검색의 효과면에서 원 논문이 참고문헌보다 검색 키로써 더 낫다.

그러나 세번째 사례연구에서는, 원 논문이 그의 참고문헌들보다 검색 능력에 있어 월등한 점을 보이지 않았다. 이 결과는 세번째 사례연구에 사용된 원 논문이 많은 다른 문헌들에 의하여 인용되지 않았다는 사실에 기인한다고 볼 수 있다. 이 발견은 적은 수의 적합문헌의 검색은 인용문헌이 적게 인용된 것이 주요 원인이라고 한 Vidal-Arbona의 연구 결과(1986)와 일치한다.

첫번째 사례연구와 다른 두 사례연구들을 비교할 때 전자가 후자들 보다 더 높은 정확률을 보였다. 이 발견은 세 질문으로 선택된 세 잡지 논문들의 성격에 관련된 문

〈표 7〉 세 사례들에서 인용문헌들의 중간 나이와 그들의 검색 능력과의 관계

검색능력 부류	중간 나이		
	사례연구 1	사례연구 2	사례연구 3
가장 검색능력이 높은 구룹의 인용문헌	3	0.5	3
중간 정도의 검색능력을 갖은 구룹의 인용문헌	5	7	2.5
가장 검색능력이 낮은 구룹의 인용문헌	7.5	6	3.5

제들과 주제들에서 설명되어질 수 있다. 왜 첫번째 사례연구가 비교적 높은 평균 정확률을 얻었는가에 대한 이유들은 잡지 논문이 다루고 있는 전문성과 세분화된 주제의 특성과 검색된 많은 문헌들이 문헌정보학에 속하기 때문으로 볼 수 있다. 두번째 사례연구에 선택된 잡지 논문은 많은 다른 분야에서 이용되고 있는 극대화 엔트로피 원리의 개념을 이용한 이론적인 연구이다. 첫번째 사례연구와 달리, 인용문헌들도 문헌정보학외의 많은 다른 분야에서 인용되었다.

세번째 사례연구에서도 다른 분야에서 많이 사용되는 군집 분석을 사용하는 가운데 두번째 사례연구에서와 비슷한 양식을 보이고 있다. 두번째와 세번째의 사례연구들에서 비교적 낮은 평균 정확률은 두 사례연구들에서 사용된 인용문헌들이 많은 다른 분야에서 인용되었고, 그들은 많은 수의 부적합한 문헌들을 검색하였다.

나중의 두 사례연구들에서 많은 수의 적합하지 않은 문헌들을 검색해 낸 주요 원인들 중 하나는 검색된 문헌들이 다른 분야에서도 많이 쓰이는 방법론에 관련되어 있기 때문일 수도 있다. 이 발견을 현 시스템에 적용하자면, 여러 분야에서 쓰는 방법론을 이용한 인용문헌은 인용검색에서 사용됨이 비효과적이라고 추측할 수 있다.

정확률의 효과에 대하여, 첫번째 사례연구에서는 만일 인용문헌이 주어진 질문의 주요 문제를 극히 적게 다루고 있다면 그것은 많은 수의 부적합 문헌들을 검색함을

보여 주었다. 두번째 사례연구에서는 만일 인용문헌의 주제 범위가 주어진 질문에서 다루고 있는 것보다 더 넓거나 더 일반적이라면 그 인용문헌은 또한 많은 수의 부적합 문헌들을 검색함을 보여 주었다. 이 발견은 다음과 같은 일반적인 가설을 제안한다.

> 인용검색의 향상을 위한 주요 요소는 인용문헌이 얼마나 특정성이 높게, 혹은 구체적으로 주어진 질문의 주제를 다루고 있는가에 달려 있다.

재현율 극대화를 위한 노선의 분석으로 부터 얻은 또 다른 발견은 어느 정도의 정확률을 가지고 재현율의 극대화를 이루기 위하여 어떻게 인용문헌들의 순서를 정하는가를 보여 주었다. 그러나 이 이론적인 설계는 만일 검색자들이 주어진 질문에 사용되는 각각의 인용문헌의 검색 능력을 미리 알 수 있다면 실제적인 온라인 시스템에 적용될 수 있다. 본 연구는 인용문헌의 성격을 조사하였고, 주어진 질문의 적합성을 예시하는 몇몇 변수들을 좋은 지표로써 제안하였다. 만일 이 변수들이 전문가 시스템들에 의하여 현 온라인 정보검색 시스템으로 합병된다면 이 이론적인 설계는 실제 환경 시스템에 이행될 수 있다.

6. 결 론

6.1 요약

본 연구는 온라인 인용검색의 분야에서 비교적 연구되지 않은 현상을 탐구하고 있다. 연구의 주 목적은 인용검색에 있어 사용된 인용문헌들의 수를 늘렸을 때 검색의 효과에 대하여 조사하는 것이었다.

본 연구의 발견들은 인용문헌 수가 증가함에 따라 재현율은 증가하고 정확률은 감소한다는 것이다. 이 발견은 Harter의 발견(1990)인 검색 용어가 추가됨에 따라 재현율이 증가한다는 것과 병행하여 해석할 수 있다. 즉 적절한 많은 수의 검색 용어들 중에서 적은 수의 용어들만을 사용한 주제검색으로는 망라적인 검색 목적을 이룰 수 없듯이, 적절한 많은 수의 인용문헌들 중에서 적은 수의 인용문헌들만 사용한 인용검색

으로는 높은 재현율을 얻을 수 없다. 따라서 검색자의 목적이 높은 재현율에 있다면 이 검색 목적을 이루기 위하여 이 변수는 중요한 것으로 고려되어져야 한다.

인용문헌들에 의한 재현율과 정확률의 명확한 차이를 설명하기 위하여 인용문헌들의 성격이 재현율과 정확률과의 관계에서 조사되었다. 여기서 얻은 발견은 어떤 인용문헌들이 정보검색의 효율성 증진을 위하여 유용한지 예견하는 지표를 제공한다.

6.2 연구 한계 및 제언

본 연구의 발견들은 이 세 사례연구들을 넘어서 일반화될 수 없다. 이 세 표본 잡지 논문들은 무작위표본 추출법을 이용하지 않고 단지 JASIS로 부터 선택되었기 때문에 이 세 잡지 논문들이 문헌정보학의 문헌들을 대표한다고 말할 수 없다.

각각의 사례연구에서 각각의 인용문헌에 의하여 검색된 문헌들의 수를 20개로 제한했다. 이 방법론적인 제약은 가장 최근 출판물에 유리한 편견으로 기울 수 있다. 왜냐하면 최신 출판물들을 먼저 검색되게 하는 DIALOG의 파일 구성 때문이다. 본 연구는 이 방법론적인 제한점을 가지고 얻어진 자료의 한계를 갖고 있다. 미래의 연구는 표본이 되는 잡지 논문의 선택과 검색된 문헌의 수를 일반화할 수 있는 방법들을 고려해야만 한다.

각각의 사례연구에 중점을 둔 포괄적인 분석들이 잠정적인 결론들과 가설들을 산출하는 발견들을 제공하였다. 현 환경에 적용할 수 있는 확실한 제안들을 만들기 위하여, 본 연구는 문헌정보학의 다른 주제들에서 뿐만 아니라, 물리학 같은 잘 정의되고 발달된 학문 분야에서도 되풀이 되는 것이 필요하다. 이 연구는 인용검색이 주제검색보다 효과적인, 새로이 나타나는 학제간 연구 문제의 영역(an emerging interdisciplinary problem area)에서도 이루어지기를 기대한다.

미래의 연구로 또 다른 가능성은 본 논문이 사용한 같은 질문들을 이용한 주제검색을 행하여 본 연구와 비교한 후 두 검색방법의 효과를 고찰하는 것이다. 또한 주제검색과 본 연구로 부터 얻은 결과는 이전의 두 검색방법들 - 주제검색과 인용검색 - 의 비교 연구(e.g., Pao, 1986)들과 비교할 기회를 줄 수 있을 것이다. 이때 주제검색은 원 논문의 서명과 초록에 포함된 용어들을 이용하여 행하면 될 것이다.

마지막 미래을 위한 연구로 검색자들이 어떻게 적절한 정확률을 가지고 재현율을 극대화 할 수 있는지를 그린 이론적 설계인 "재현율 극대화 노선"의 확장이다. 본 연

구에서 이 설계를 소개하고 이것을 어떻게 현 시스템에 적용할 수 있을지 논하였다. 그러나 이 이론적 설계를 현 온라인 정보검색 시스템에 실행할 수 있게 발전시키는 연구는 결실이 풍부한 연구로써 높게 제언된다.

참고문헌

Brooks, T.A. (1986). Evidence of complex citer motivations. Journal of the American Society for Information Science, 34, 34-36.

Chubin, D.E., Alan, L.P. and Frederick, A.R. (1984). Citation Classics analysis: An approach to characterizing interdisciplinary research. Jouranal of the American Society for Information Science, 35, 360-368.

Cleverdon, C.W. (1960). ASLIB Cranfield research project on the comparative efficiency of indexing systems. ASLIB Proceedings, XII, 421-431.

Fenichel, C.H. (1980/1981). The process of searching online bibliographic databases: A review of research. Library Research 2, 107-127.

Garfield, E. (1965). Can citation indexing be automated? In M.E. Stevens et al.(Eds.), Statistical Association Methods for Mechanized Documentation (NBS Misc. Pub. 269). Washington, DC: National Bureau of Standards.

Gluck, M. (1990). A review of journal coverage overlap with an extension to the definition of overlap. Jouranl of the American Society for Information Science, 41, 43-60.

Goffman, W. and M.L. Pao. (1980). Retrieval of biomedical information for emerging interdisciplinary problems. Proceedings 4th International Congress on Medical Librarianship, 39-50.

Harter, S.P. (1990). Search term combinations and retrieval overlap: A proposed methodology and case study. Journal of the American Society for Information Science, 41, 132-146.

Harter, S.P. (1986). Online information retrieval, concepts, principles, and techniques. San Diego, CA: Academic Press.

Katzer, J., McGill, M.J., Tessier, J.A., Frakes, W. and DasGupta, P. (1982). A study of the overlap among document representations. Information Technology: Research and Development, 2, 261-274.

McCain, K.W. (1989). Descriptor and citation retrieval in the medical behavioral sciences

literature: Retrieval overlaps and novelty distribution. Jouranl of the American Society for Inforamtion Science, 40, 110-114.

Pao, M.L. (1988). Term and citation searching: A preliminary report. Proceedings of the Annual Meeting of the American Society for Information Science, 25, 177-179.

Pao, M.L. (1986). Comparing retrievals by keywords and citations. Proceedings of the National Online Meeting, 341-346.

Pao, M.L. (1984). Semantic and pragmatic retrieval. Proceedings of the Annual Meeting of the American Society for Information Science, 21, 134-136.

Pao, M.L. and Worthen, D.B. (1989). Retrieval effectiveness by semantic and citation searching. Journal of the American Society for Inforamtion Science, 40, 226-235.

Saracevic, T., Kantor, P., Chamis, A.Y. and Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. II. Users, questions, and effectiveness. III. Searchers, searches, and overlap. Journal of the American Society for Information Science, 39, 161-176, 177-196, 197-216.

Swanson, D.R. (1965). The evidence underlying the Cranfield results. Library Quarterly, 35, 1-20.

Trivison, D., Saracevic, T. and Kantor, P. (1986). Effectiveness and efficiency of searchers in online searching: Preliminary results from a study of information seeking and retrieving. Proceedings of the Annual Meeting of the American Society for Information Science, 341-349.

Vidal-Arbona, C. (1986). Comparing the retrieval effectiveness of free-text and citation search strategies in the subject of technology planning. Ph.D. dissertation, Case Western Reserve University, Cleveland, OH.

Voos, H., & Dagaev, K.S. (1976). Are all citations equal? or, did we op. cit. your idem? Journal of Academic Librarianship, 1, 20-21.

Yoon, L.L. (1992). The Performance of cited references as an approach to information retrieval. Ph.D. dissertation, Indiana University, Bloomington, IN.

ABSTRACT

A Study on Information Retrieval Effectiveness by Cited References

Lanju Lee Yoon *

Databases publicly available for online searching permit both citation and subject searching, however, subject searching has dominated the online search environment. Despite the power of citation searching, it may be underutilized. This study explored the relationship between the number of cited references used in a citation search and information retrieval effectiveness, a relatively unstudied phenomenon.

Three articles in the library and information science literature were chosen to represent sample questions. Cited reference searches were conducted for each article and each of its references. All searches were conducted in Social Scisearch and Scisearch on DIALOG. Relevance judgments on the retrieved citations were obtained from the authors of the original articles.

This research focused on analyzing, in terms of information retrieval effectiveness, the overlap among postings sets retrieved by various combinations of cited references. The findings from the three case studies clearly showed that the more cited references used for the citation search, the better the performance, in terms of retrieving more relevant documents, up to a point of diminishing returns. In addition, generally the overall level of overlap among relevant documents sets was found to be low. Therefore, if only some of the cited references among many candidates are used for a citation search, a significant proportion of relevant documents may be missed. The analysis of the characteristics of cited references provided the ways to predict which cited refereces would be useful to improve informtion retrieval.

* Assistant Professor, Dept. of Library and Information Science, Dongduck Women's University.

The findings of this comprehensive exploratory study are of interest for both theoretical and practical reasons. They contribute to the development of a theoretical model for the effective use of the citation search. This model might also be implemented in operational online systems. In addition, the findings potentially will help online searchers improve their search strategies using the citation search so that they can better achieve their information retrieval goals: the retrieval of items relevant to a given question and the suppression of nonrelevant items.