학술커뮤니케이션 네트웍을 통한 정보검색 시스팀의 개발*

鄭 遵 民**

〈 目 次 〉

Ⅰ. 서 론

지난 반세기에 걸쳐서 학술문헌의 기하급수적인 성장은 문헌정보학 연구자 뿐만 아니라 일반 이용자들에게 까지도 심각한 문제로 대두되어지고 있다. 특히, 학술정보 데이터베이스를 관리하는 도서관이나 정보센터로서는 기하급수적으로 늘어나는 학술정보를 어떻게 하면 효율적으로 입력 유지할 수 있을까 하는 것이 그들의 큰 관심분야라 할 수 있다. 그러나 관리자의 입장과는 반대로 정보검색 시스팀을 이용하고 있는 이용자들에게 있어서 가장 당면한 문제는 대규모의 데이터베이스 내에서 자신의 정보 욕구를 충족시켜줄 수 있는 소수의 정보만을 어떻게 하면 효율적으로 검색할 수 있을까 하는 일이다. 실제로 정보검색 시스팀을 사용하는 이용자들의 대부분은 그들이 이용하고 있는 데이터베이스의 구조와 시스팀 자체에 대해선 거의 지식을 갖지 못하고 있는 실정이다.

* 본 논문은 1989 년도 문교부 학술연구 조성비에 의하여 이루어진 것임.
** 전남대학교 문헌정보학과 부교수

정보검색 시스팀의 대부분은 그것이 전산화가 되었든 아니든 간에 데이터베이스를 구성하고 있는 정보와 그것을 특징짓고 구분지을 수 있는 경로를 그 정보에 부여함으로써 이용자들로 하여금 데이터베이스 내의 정보를 탐색할 수 있도록 설계되어져 있다. 데이터베이스 구성에 사용되어지는 주된 탐색 경로는 분류번호, 저자명, 서명, 주제명 (색인어, 디스크립터 등 포함) 등을 들 수 있으며 전산화된 정보검색 시스팀에서는 대부분 이들 탐색 경로에 부울 (Boole) 대수의 논리 형식을 적용함으로써 이용자들의 정보 욕구를 충족시켜 주고 있다.

부울검색기법이란 부울 논리연산자 (NOT, OR, AND)를 이용하여 둘 이상의 탐색어에 관계를 설정, 검색의 깊이와 정도를 조절하는 검색기법을 말한다. 그러나 논리연산자가 갖는 이진법적 사고는 복잡한 주제와 많은 탐색어를 사용할 경우 그 효과를 기대하기 어려우며 이와 같은 문제를 해결하기 위한 발전된 형태의 기법 (가중치) 조차도 만족할 만한 성과를 얻지 못하고 있다. 한편, 실제 시스팀에 적용된 예는 많지 않으나 실험실적으로 개발되어진 것에는 확률이론과 클러스터링 기법을 응용한 다차원적인 검색 시스팀이 있다. 이 시스팀은 문헌에 주어진 속성을 이진법적 사고가 아니라 집합적 논리에 의해 두 문헌의 관계를 확률적 스케일로 변환하여 정의하고 있다. 확률시스팀은 그 이론적 배경에 있어 문헌정보학에서 의미하는 검색논리로서의 적합성 (Relevance) 개념을 충족시켜줄 수 있는 장점이 있는 반면 현실적으로 시스팀을 설계, 유지하기에는 막대한 비용과 노력이 요구되어지는 문제를 안고 있다.

정보검색 시스팀에 있어서 검색 논리와 아울러 중요한 사실은 탐색 경로의 선택이다. 앞에서도 언급하였듯이 대부분의 시스팀은 주된 탐색 경로로 주제명 또는 색인어를 채용하고 있다. 즉, 이용자의 정보 욕구를 다수의 색인어 또는 주제명으로 변환하여 시스팀이 채용하고 있는 검색 논리에 따라 적합한 문헌군으로 접근시키고 있

다. 그러나 정보 이용자들의 정보 접근 패턴을 조사해 보면 데이터베이스 내의 관련 문헌 전부를 탐색하여 보기 보다는 가장 적합하다고 생각되어지는 또는 쉽게 접근할 수 있는 문헌 중 가장 최근의 문헌을 채택하여 그 문헌이 갖고 있는 인용문헌을 통한 소급 탐색을 선호하고 있다. 결국 이것은 원 문헌과 그것이 인용한 문헌간엔 상당한 주제적 상관관계가 있다는 것을 입증하는 것이며 이러한 가설을 통한 인용문헌간의 나아가 그 문헌의 저작자들간의 학술정보 커뮤니케이션 네트웍을 구축할 수 만 있다면 주제명 또는 색인어에 의한 탐색 보다는 좀 더 효율적인 검색 시스팀을 설계할 수 있으리라 본다.

본 연구는 이와 같은 가설 위에서 서지기술 사항 중의 하나인 인용문헌을 데이터베이스의 주된 탐색 경로로 하는 새로운 학술 커뮤니케이션 네트웍을 구축하여 보았으며 실험을 통하여 그 검색 효율을 측정하여 보았다.

II. 이론적 배경

1. 그래프 이론 (Graph Theory)

그래프를 수학적으로 정의하면 다음과 같다.

그래프 G 는 두 집합 V 와 E 로 이루어진다. V 는 절점 (vertex, node, point) 의 집합으로서 $V \neq 0$ 이고 유한 집합이다. E 는 절점의 쌍 (pair) 으로 된 절선 (arc, edge, line) 의 집합이다.

그래프 G 는 공집합이 아닌 유한 집합으로서 p 개의 요소로 이루어진 집합 V 의 요소들의 q 개의 쌍으로 이루어진 집합 (E) 을 의미하며 집합 E 의 요소 x 의 두 절점을 u, v 라 할 때 x 는 두 절점 u, v 를 잇는 절선을 의미한다. 즉, 그래프 G 는 p 개의 절점과 q 개의 절선으로 이루어진 집합을 말하며 이것을 (p, q) 그래프라 부른다. 특히, 집합 E 가 순서쌍의 집합일때 우리는 이것을 방향성 그래프 (Directed Graph, Digraph) 라 부른다.

일반적으로 그래프를 도식적으로 나타낼 때, 절점은 점 (.) 으로 표시하고, 절선은 선(-)으로 표시한다. 또 방향성 그래프에서의 절점은 화살표 (→) 로 나타낸다. 그래프를 도식화하여 보면 그림 2-1 같다.

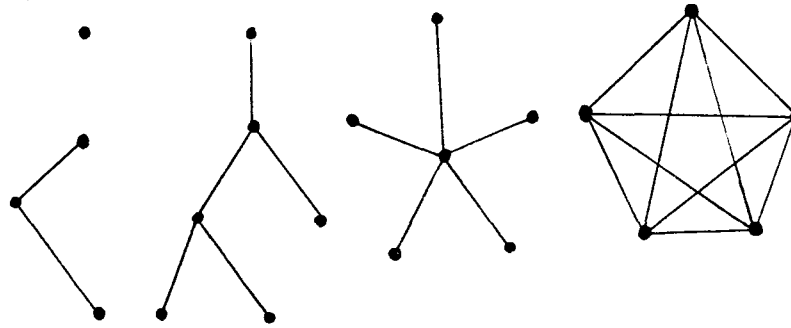

그림 2-1 그래프의 여러가지 형태

그래프를 이용한 문제 해결의 최초 기록은 1736년의 것이다. 우리에게 오일러의 정리 (Euler's Theorem, Leonhard Euler 1707-1783)로 알려진 쾨니히스베르크 (K nigsberg) 시의 〈다리건너기문제〉 이다 (그림 2-2a : 프레겔강의 7 개의 다리를 단 한번씩만건너 모두 돌아오는 방법).

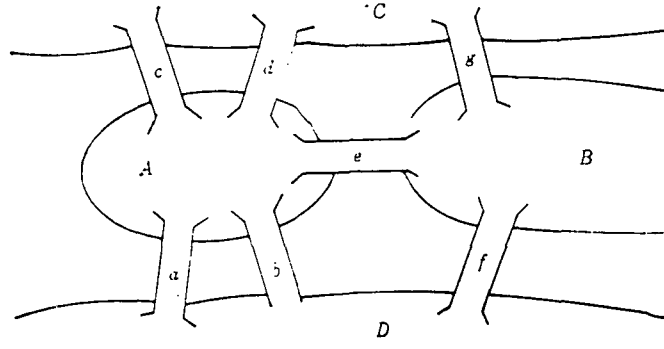

그림 2-2 Euler의 다리건너기문제

Eule 는 다음과 같은 그래프로써 그림 2-2a 를 표현하였다.

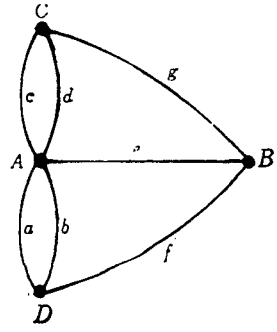

그림 2-2b Euler의 그래프식 표현

1736 년 이후 그래프 이론은 많은 분야에 이용되었다. 전자 회로의 분석, 최단 거리 찾기, 언어학, 사회학 등 그 응용 분야는 광범위하다. 특히, 게임이론으로 알려진 OR (Operations Research) 의 한 분야로서 그 역할은 매우 크다고 볼 수 있겠다.

문헌정보학에서는 정보검색 이론에 주로 이용되었으며 색인어 선정 또는 비공식 커뮤니케이션의 경로 분석에 활용 되어지기도 한다. 정보검색에의 활용은 집락분석 (Clustering Technique) 을 통한 문헌 (절점) 간의 관계규정 (절선) 의 형태로 사용되었으며 학자 (절점) 들 사이에서의 커뮤니케이션 경로 (절선) 를 도식화하여 특정 집단 내에서의 핵심 저자의 파악 및 사상의 전달 경로를 해석하는데에도 그래프 이론을 적용할 수 있다. 한 예로 전산화 된 도서관에서 도서관의 장서를 인용한 문헌과 인용된 문헌의 관계로 규정 지었다고 했을 때 어떤 특정한 규칙 (커어널 기법) 에 따른 최소한의 문헌군을 설정하여 직접 검색에 활용하고 나머지 문헌군은 일차 검색된 문헌을 통한 간접 검색을 시도함으로써 검색의 효율성을 높일 수 있다.

2. 인용문헌 분석 (Citation Analysis)

문헌정보학은 그 학문적 특성을 문헌 또는 문헌과 관계 지어진 제반 현상을 계량학적 또는 사회학적으로 분석하는 데에 두고 있다. 특히, 문헌이 갖는 서지적 요소를 수량학적으로 분석하여 봄으로써 대상 학문의 특성을 파악하거나 이용자들의 학술정보에 대한 욕구를 충족시켜주고자 하는 연구를 계량서지학이라 부른다. 계량서지학은 1960 년대 문헌정보학의 영역이 전통적인 도서관학에서 정보학으로 발전되는 시점에서 구체화 되기 시작하였으며 문헌의 서지적 요소를 그 주된 연구의 대상으로 하고 있다. 계량서지학의 연구 대상인 문헌의 서지적 요소란 문헌이 갖고 있는 형태적 요소 (저자, 서명, 출판사항 등) 뿐 아니라 문헌과 문헌이 갖는 관계, 학술 커뮤니케이션 집단 내에서의 사상, 정보의 흐름 등을 내포하고 있으며 이와 같은 서지적 요소들의 인과 관계를 규명하여 그 현상을 이론적으로 정립시키는데 학문의 목적이 있다. 계량서지학의 연구 분야 중에서도 특히 활발한 분야는 인용문헌 분석 분야라 할 수 있겠다.

인용문헌이란 학술논문에 있어서 저자가 표현하고자 하는 사상이나 이론을 학술적으로 뒷받침받을 수 있도록 저자에 의해 선택되어져 학술논문의 말미에 수록하고 있는 문헌 또는 문헌군을 의미한다. 즉, 저자가 전달하고자 하는 내용과 주제적으로 일치하는 논문의 집합이라 할 수 있다. 위의 사실에서 우리는 다음 두가지 가설을 설정할 수 있다. 첫째, 서로 비슷한 인용문헌 목록을 갖는 두 문헌은 주제적으로 매우 유사할 것이다 (서지결합; Bibliographic Coupling). 둘째, 한 문헌에 인용되어진 인용문헌들은 서로 서로 주제적 상관관계가 클것이다 (동시인용; Co-Citation). 이상과 같은 두가지 가설에 대해 문헌정보학에서는 인용문헌의 상관관계 알고리즘을 개발하여 그것의 타당성을 증명하고 학술 커뮤니케이션 시스템에 적용시키려는 연구가 매우 활발히 진행되고 있다. 인용문헌에

대한 분석은 주로 색인어를 통한 검색시스템의 문제점을 해결하고 보완하기 위한 수단으로 연구되어지고 있으며 저자동시인용 또는 공저지도 등을 통하여 학술 커뮤니케이션 집단의 사상의 흐름과 그 연구전선 (Research Front) 규명에 활용되어지고 있다.

위의 두 가설을 좀 더 자세히 설명하기 위하여 다음과 같은 예를 들어 설명하여 보기로 한다.

다섯 개의 문헌 A, B, C, D, P 가 있고, 다섯 논문 모두 각자의 참고문헌 리스트가 있으며 다음 그림 2-3 과 같다고 가정하여 보자.

(가상문헌)

	문헌 A	문헌 B	문헌 C	문헌 D	문헌 P
참	a	a	h	d	a
	b	d	i	g	b
고	c	e	j		c
	d	f			e
문		g			f
					h
헌					

문헌 P 에 수록된 참고문헌 리스트는 이용자가 제시한 문헌 리스트로 이용자가 일차 검색을 통하여 확보한 데이터임.

그림 2-3 문헌 A, B, C, D, P 와 참고문헌 리스트

2.1. 서지결합 (Bibliographic Coupling)

그림 2-3 에서 두가지 경우로 나누어 생각해 보자. 첫째, 이용자의 정보요구 (가상문헌 P) 를 만족 시킬 수 있는 문헌을 찾는 예,

둘째, 이용자의 정보요구와 무관하게 데이터베이스 내에서의 문헌간의 관계를 설정하는 예.

인용문헌 리스트에 나타난 문헌들을 그것을 인용한 문헌들의 속성이라고 전제하여 본다면 우리는 문헌들간에 유사한 속성을 갖고 있음을 그림을 통하여 알 수 있다. 즉, 문헌 A 와 B 동일한 참고문헌 a, d 를 갖고 있음을 알 수 있으며 같은 방법으로 문헌 B 와 D (참고문헌 d, g), 문헌 A 와 D (참고문헌 d) 도 같은 참고문헌을 공유하고 있음을 알 수 있다. 한편, 문헌 C 는 문헌 A, B, D 와 무관함을 알 수 있다.

그러나 이용자가 특수한 목적으로 설정한 질문 (문헌군) 과 데이터베이스 내의 문헌과 비교했을 때는 다른 결과를 얻을 수 있다. 그림 2-3 에 따르면 앞의 결과와는 달리 가상문헌 P 와 관계있다고 보이는 문헌은 A, B, D 가 아닌 A, B, C 로 나타남을 알 수 있다. 그림 2-4 는 이것을 그래프로 도식화한 것이다.

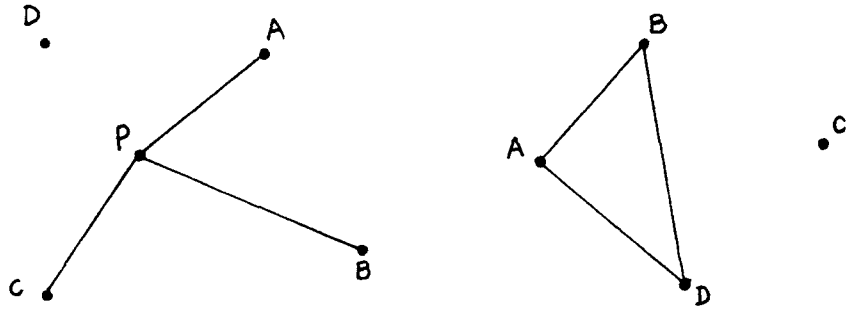

그림 2-4 서지결합의 두가지 예

이상과 같이 서지결합은 대상 문헌이 갖는 인용문헌 (참고문헌) 을 일반 검색 시스템에서의 분류 속성인 색인어와 유사하게 처리하

고 있음을 알 수 있다. 즉, 문헌의 속성을 색인어 대신 인용문헌으로 대치하고 검색을 시도하고자 하는 것이다. 특히, 문헌과 문헌의 관계 정도를 공유하는 인용문헌의 숫적 비례로 평가함으로써 재래의 부울 대수에 의한 검색이기 보다는 확률 검색 기법에 가깝다고 볼 수 있다. 그러나 논문 저자들의 참고문헌의 중요성에 대한 인식 부족과 잘못된 인용 패턴으로 인하여 아직은 색인어 만큼 관심을 끌지는 못하나 앞으로 참고문헌의 질적 평가와 더블어 검색의 중요한 요소가 될 수 있으리라 본다.

2.2. 동시인용 (Co-Citation)

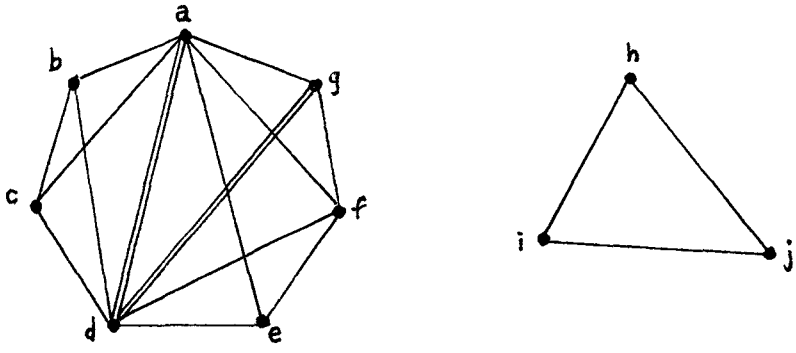

그림 2-5 동시인용 그래프

동시인용 기법은 서지결합과 반대로 인용한 문헌과 인용된 문헌에 의한 관계가 아니라 동시에 인용되어진 문헌간의 관계를 규명지으려는 시도라 볼 수 있다. 그림 2-3 에서 보듯이 문헌 A 와 다른 문헌의 관계 규정이 아니라 문헌 A 가 인용한 문헌 a, b, c, d 의 상호 관련성을 측정하려는 시도이다. 즉, 인용되어진 문헌들을 두개 씩 짝 (쌍) 을 이루었을 때 각 쌍은 서로 주제적 상관성이 있다고 보는 것이며 이러한 쌍이 둘 이상의 문헌에 나타나질 때 그 쌍은 주제 관

련성이 다른 쌍들에 비해 더 강함을 의미한다. 그림 2-3 과 그림 2-5 에서 보듯이 인용된 문헌 중 a 와 d 는 두개의 문헌 A 와 B 에서 동시에 나타났으며 d, g 역시 문헌 B, D 에 동시에 나타나고 있음을 알 수 있다.

동시인용 분석은 서지결합에서 보인 잘못된 인용 패턴의 영향을 무수히 많은 데이터를 처리함으로써 희석시킬 수 있는 장점이 있으나 일단 대상 문헌이 다른 문헌과 함께 인용되어야 비로소 데이터베이스에 등록이 될 수 있다는 시간의 한계를 갖는다. 뿐만 아니라 서지결합이 일반적인 데이터베이스 관리방법으로 처리가 가능한 방면 동시인용은 많은 데이터 용량과 함께 데이터 처리에 장시간을 요구함으로써 실질적 응용이 불가능한 상태이다.

Ⅲ. 연구 방법

1. 가설의 설정

모든 정보검색 시스팀은 이용자들로 하여금 데이터베이스로 부터 이용자가 원하는 정보만을 검색할 수 있도록 설계되어진 탐색 경로를 갖고 있다. 검색 시스팀의 주된 탐색 경로는 저자, 서명과 아울러 일반적으로 색인어가 많이 사용되고 있으며 탐색 논리로는 부울 대수의 연산자를 사용하고 있다. 그러나 이와 같은 시스팀은 이용자에게 데이터베이스와 시스팀 자체에 대한 전문 지식을 요구할 뿐만 아니라 자신이 원하는 정보에 대한 완전하고 정확한 지식을 갖고 있어야 한다. 그렇지 못했을 경우, 시스팀은 이용자를 대신할 수 있는 전문 탐색 요원을 확보하여야 하며 이는 검색에 많은 경비를 투자하여야 함을 의미한다. 더구나 시스팀이 관리하는 데이터베이스의 주제가 너무 전문적이거나 최신 분야일 경우 주제전문 사서 (전문 탐색자) 의 확보는 용이하지 않다.

정보검색 시스팀이 제안하는 탐색 논리 외에도 우리는 쉽게 자신이 원하는 정보에 접근할 수 있는 방법을 찾을 수 있다. 대부분의

이공계 연구자들에 의해 이루어지고 있는 참고문헌을 통한 접근이 그것이다. 그들은 자신이 필요로 하는 문헌이나 정보를 자신이 획득한 문헌의 참고문헌을 순차적으로 접근하여 봄으로써 그들의 정보욕구를 만족시키고 있다. 이것은 학술논문에 수록된 인용문헌 리스트가 저자가 자신의 견해와 사상을 피력하기 위하여 주제적으로 관련된 기존의 논문을 인용한 것으로서 자신이 발표하는 논문에 학술적 권위와 아울러 자신이 발표한 논문의 이론적 타당성을 입증하기 위한 방편으로 사용되었기 때문이다. 즉, 원 문헌과 그에 수록된 인용문헌 사이에는 주제적으로 깊은 관계가 있음을 간접적으로 시인하는 것이며 함께 인용되어진 문헌 간에도 높은 주제적 상관관계가 있음을 암시하고 있는 것이다.

그러나 이와 같은 접근은 이용자가 원하지 않은 많은 문헌도 탐색해야하는 문제점을 안고 있을 뿐 아니라 그들의 탐색을 어디에서 멈출 것인지에 대한 객관적 기준도 갖고 있지 못하다. 만일 원 문헌과 아울러 인용문헌에 대한 정보도 함께 제공하며 동시에 인용 되어진 문헌의 주제적 상관관계에 대한 이론적 척도를 제시할 수만 있다면 색인 시스팀이 갖는 표현의 문제점을 쉽게 해결할 수 있을 뿐 아니라 참고문헌을 통한 순차적 접근이 주는 비관련 문헌의 탐색을 제거할 수 있을 것이다.

위에서 제시한 가정을 요약, 정리하여 보면 다음과 같은 가설을 설정할 수 있다. "동시에 인용된 문헌이 갖는 주제적 상관성을 이용하여 인용문헌을 탐색 요소로 하는 데이터베이스를 설계하여 검색을 실시할 경우 색인어를 통한 검색과 유사한 결과를 얻을 수 있다."

2. 모형 설계

주어진 가설을 검증하기 위해서는 먼저 인용문헌 간의 의미있는 상관계수를 설정해야 한다. 두 문헌이 동시에 인용되었다 함은 두 문헌 간에 주제적 상관계수가 높음을 의미하며 이와 같은 패턴이 여

러 문헌에서 나타나면 그 두 문헌의 주제 관련성은 더 높을 것이라는 것이 주어진 가설의 기본 가정이므로 두 문헌의 주제 상관성을 측정하는 것이 가장 중요한 문제다.

한 문헌에 인용되어진 인용문헌들을 동시인용 패턴에 따라 도식화해 보면 그래프 이론에서 말하는 완전연결 그래프를 형성한다. 동시인용 패턴 자체가 완전연결 그래프의 형식을 취하고 있으므로 본 논문에서의 주제 상관성의 척도를 주어진 기준값 (Thereshold)에 따라 인용문헌 지도를 만들고 완전연결 그래프를 형성하는 집단을 하나의 집합으로 설정하여 한 집합에 소속된 문헌들의 관계만을 인정하고 그렇지 않은 경우는 2 차 접근을 위한 경로로 간주한다.

그림 2-6a 는 기준값을 1 로 했을 때의 집합과 경로를 나타내고 있으며 그림 2-6b 는 같은 데이터의 기준값을 2 로 했을 때의 결과이다. 그림에서도 알 수 있듯이 기준값 1 인 경우는 원 문헌의 참고문헌 리스트와 일치함을 알 수 있다. 그러나 일반 참고문헌 리스트와는 달리 2 차 접근을 위한 방향과 확장성을 제시하고 있다.

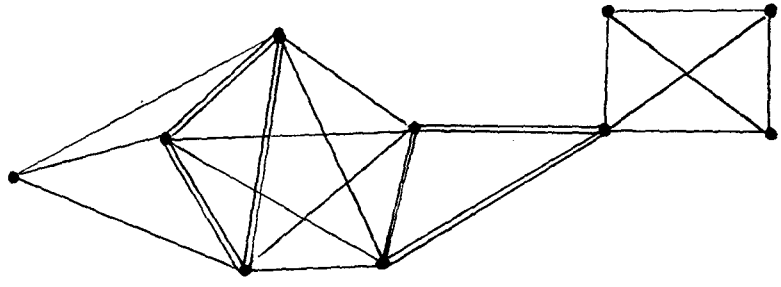

그림 2-6a 동시인용 지도 (기준값 = 1)

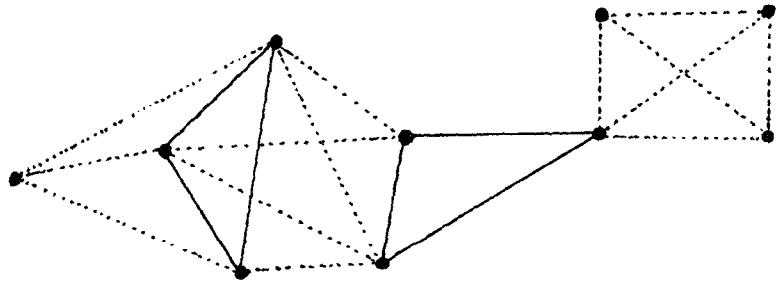

그림 2-6b 동시인용 지도 (기준값 = 2)

동시인용 지도를 살펴보면 기준값을 높였을 때 완전연결 그래프가 줄어드는 현상을 파악할 수 있으며 나머지 연결선 (경로) 를 통하여 검색의 폭을 확장하면 기준값을 낮췄을 때와 같은 결론을 도출할 수 있음을 알 수 있다. 이것은 검색에서 의미하는 적합성에 따른 재현율과 정확율을 이용자의 요구에 따라 쉽게 조절할 수 있음을 의미하며 보통 검색 시스템에서 주어지는 검색 문헌의 주제 상관성에 따른 순서 리스트와 같은 효과를 창출할 수 있다.

3. 시스템 평가

주어진 가설과 모형에 따라 설계되어지고 구축되어진 시스템은 이용자들의 검색 만족도에 따른 피이드 백 과정이 요구된다. 검색 시스템에 있어서의 피이드 백은 검색된 문헌집단을 가지고 이용자의 만족도를 즉각 측정할 수도 있으나 이용자의 문헌 검색의 의미 자체가 검색된 문헌들을 통하여 새로운 문헌을 창출하는 데에 있으므로 검색된 문헌을 이용하여 이용자가 발표한 문헌에 수록된 참고문헌을 간접적으로 조사해 봄으로써 그 결과를 측정할 수 있다. 연구자가

학술논문을 발표하기 까지에는 많은 연구와 문헌 탐색이 요구되어지며 그 과정에서 발생하는 정보검색 시스팀을 통한 문헌의 획득에는 많은 시행착오를 가질 수 있다. 그러나 일단 논문이 발표되어지고 그 논문에 참고문헌을 실었을 경우는 연구자가 행한 그간의 연구 업적과 참고문헌이 갖는 주제가 일치함을 의미한다고 볼 수 있다. 그러므로 시스팀을 통한 탐색 때마다 만들어진 문헌군에 대한 평가는 객관성을 상실할 수 있으며 궁극적으로 정확한 판단이었다고 볼 수 없다.

검색 시스팀에 대한 평가는 주어진 질문 (문헌)에 대한 결과와 그에 의해 창출된 논문의 참고문헌을 비교함으로써 가능하며 그 일치된 정도에 따라 시스팀의 효율 또는 발표된 논문의 가치를 객관적으로 평가하여 볼 수 있다.

Ⅳ. 실 험

1. 데이터

본 연구의 가설을 검증하기 위하여 두 개의 주제를 선정 실험 대상으로 삼았다. 실험 결과에 대한 해석에 당위성을 기하기 위하여 인문사회과학 분야 중 주제 특성이 강한 문헌정보학 (서지학 제외)을 선택하였으며 자연과학 분야에서는 두 개 이상의 주제가 합쳐져서 이루어진 신분야로 유전공학을 선택하였다.

본 연구가 실험적 성격을 띤 연구이므로 두 주제와 관련된 문헌을 총망라하기에는 불가능하여 1990 년대에 학회지에 발표된 문헌만을 선택하였으며 유전공학 분야는 한국과학기술연구원 부설 유전공학센터의 자문을 구해 관련 문헌을 수집하였으며 문헌정보학은 전국적 규모의 학회지 두 개를 선정하였다. 표 4-1 은 질문문헌군을 형성할 데이터의 분포 상황을 보여주고 있다.

유 전 공 학		문 헌 정 보 학	
학 술 지 명	논문수	학 술 지 명	논문수
유전공학 (A)	6	정보관리학회지(F)	13
미생물학회지 (B)	11	도서관학 (G)	13
산업미술학회지(C)	11	——	
동물학회지 (D)	6	——	
분자생물학뉴스(E)	3	——	
계	37	계	26

번역 및 단신 등은 제외시켰음

표 4-1 실험 질문문헌 분포

유전공학과 관련된 문헌 37 개와 문헌정보학 분야의 26 개 문헌을 질문문헌군으로 설정하여 본격적인 실험을 시도하기 위한 데이터는 이들 문헌이 갖고 있는 참고문헌을 마스터 파일로 하여 작성하였다. 마스터 파일이 되는 두 개의 문헌군으로 부터 수집한 참고문헌의 수는 유전공학 974 개, 문헌정보학 953 개로 문헌당 수록하고 있는 참고문헌의 수는 유전공학 26.32 개 (SD = 16.7478), 문헌정보학 36.65 개 (SD = 24.5105) 로 나타났다. 표 4-2 는 두 주제집단으로 부터 추출된 참고문헌의 연도별 분포를 보여준다.

주제 / 년도	#90	#89	#88	#87	#86	#85	#84	#83	#82	#81	#80	#79	#78	#77
유 전 공 학	20	61	81	86	89	77	66	53	45	41	48	24	21	26
문 헌 정 보 학	15	38	50	57	71	52	53	82	50	45	40	38	24	35

주제 / 년도	#76	#75	#74	#73	#72	#71	#70	#69-#65	#64-#60	#50년대	ETC
유 전 공 학	21	17	9	17	14	12	11	47	24	34	30
문 헌 정 보 학	24	27	28	16	22	22	11	76	28	30	19

표 4-2 참고문헌의 출판연도별 분포

2. 분 석

수집된 데이터의 특성을 살펴보면 거의 대부분의 논문이 한번 정도 인용되어지고 있으며 (표 4-3) 두번 이상 인용되어진 문헌을 분석해 본 결과 유전공학 분야는 실험 결과 및 방법론에 대한 문헌이 많은 반면 문헌정보학은 대개가 개론적 성격을 띤 문헌들로 되어 있음을 보여준다.

주제/인용 빈도	5 번	4 번	3 번	2 번	1 번	계
유 전 공 학	1	0	2	17	929	949
문 헌 정 보 학	0	0	0	7	939	946

표 4-3 인용빈도별 문헌 수

동시인용분석 그래프의 일반적인 형태는 완전연결 그래프의 형태를 취하고 있으므로 문헌 (절점) 의 수와 동시인용 (절선) 의 관계는 다음의 관계를 갖는다.

$$\text{동시인용 관계의 수} = \frac{\text{문헌의 수} * (\text{문헌의 수} - 1)}{2}$$

위의 식에 따른 두 개 문헌 집단의 문헌 총 수와 동시인용 관계 (절선)의 수는 표 4-4와 같다. 표 4-4 에서 보듯이 유전공학 분야가 37 개 문헌 그룹에 949 개의 인용문헌으로 17,520 개의 동시인용 관계를 만든 반면 문헌정보학은 25 개그룹 (946 개의 인용문헌) 에서 24,799 개의 동시인용 관계를 만들었다.

주제	문헌수	동시인용
그룹I	949	17,520
그룹II	946	24,799

표 4-4 동시인용 관계

이것은 문헌정보학이 유전공학에 비해 평균 인용문헌 수가 높음을

의미하기도 한다. 그러나 본 실험에 사용된 데이터의 동시인용 관계의 값을 증가 시키면 거의 모든 문헌이 독립되어 버린다는 사실이다. 즉, 기준값 (Threshold Value, τ) 를 $\tau = 1$ 에서 $\tau = 2$ 로 올렸을 때, 거의 모든 문헌이 동시인용 관계에서 단독의 형태로 존재하게 된다.

$\tau = 2$ 인 상태란 두 개 이상의 문헌에서 공통으로 동시 인용된 경우를 의미하며 표 4-3 과 표 4-5 에 나타났듯이 유전공학에서는 모두 20 개의 문헌이, 문헌정보학에서는 7 개의 문헌이 두 번 이상 인용 되었으나 이 중에 두 번 이상 동시 인용 된 경우는 그림 4 - 6 에 나타났듯이 극히 소수에 불과 하였다.

주제	인용문헌	원문헌 (인용한문헌)	주제	인용문헌	문헌 (인용한문헌)
	1949W	C01,C05		71리재철	F01,F07
	1951L	B01,B02,C08,D04,D05	문	72정필모	F01,F07
유	1959M	B05,C02,C11		1976GAT	F07,G05
	1970L	B03,D04,D06		1980BUS	G04,G12
	1972F	C01,C05		1980CON	F01,G05
	1972M	B11,C03	헌	87정영미	F05,G03
	1975P	C01,C05		88신성철	F05,F08
전	1977W	B06,B07			
	1981G	B04,C08			
	1981N	B04,B05	정		
	1982A	B04,B05			
	1982F	B06,B07			
공	1982G	C01,C05			
	1983M	B04,B05	보		
	1984H	B04,B05			
	1984K	C01,B10			
	1984Y	B04,B05			
학	1987C	B04,B05	학		
	1987K	B04,B05			
	1988A	B06,B07			

표4-5 두번 이상 인용된 문헌표 (실험에 사용한 코드로 표기하였음)

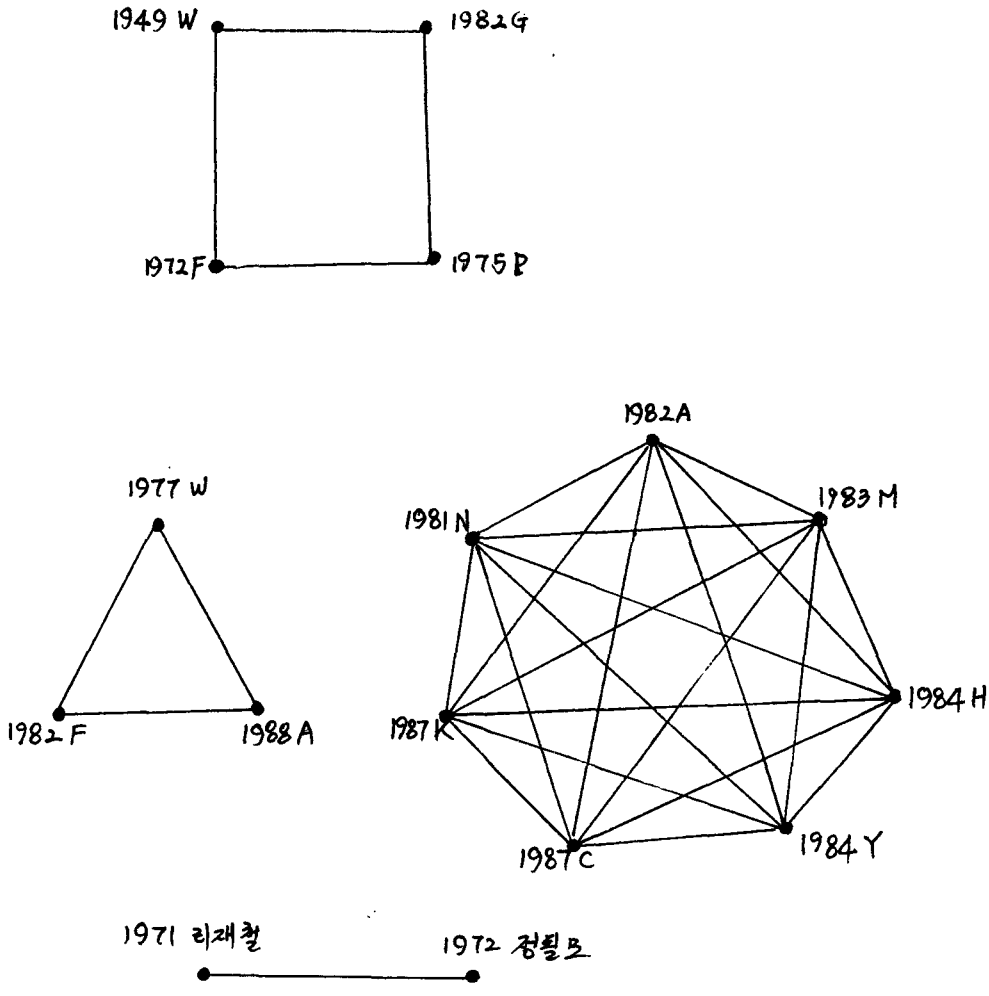

그림 4-6 유전공학, 문헌정보학 동시인용 그래프 (τ = 2)

그림 4-6에서도 보여지듯이 유전공학 분야에서는 3 개의 완전연결 그래프가 존재하고 있으며 문헌정보학 분야에서는 오직 한 개의 완전연결 그래프가 존재하고 있음을 알 수 있다. 그러나 기준값를 3으로 했을 경우는 두 분야 모두 그래프가 해체되는 현상을 보이고 있다.

3. 결 과

그림 4-6을 통하여 기준값을 2 로 주었을 때의 동시인용 그래프값을 산출해 보았다. 본 실험에서 제시하였던 '동시인용이 갖는 그래프적 해석은 완전연결 그래프' 라는 관점에서 얻을 수 있는 결론은 유전공학에서의 3 개의 문헌군과 문헌정보학의 1 개 문헌군은 연구자들의 유사 주제 연구에 동시에 인용 되어야 한다는 사실이며 만일 주제적으로 유사한 연구에 그들 문헌군의 어느 한 문헌이라도 포함되지 않았을 경우 그 논문의 질적 평가의 기준이 될 수 도 있겠다.

한편, 유전공학에 비해 문헌정보학에서 더 적은 수의 집단이 발생한 원인은 실험집단으로 선정된 두 주제의 문헌이 문헌정보학 보다는 유전공학이 훨씬 주제 밀집도가 높았다고 볼 수 있다. 그러나 유사한 주제집단을 선정하여 비교하여 보아도 (표 4-7) 동시인용 관계의 값이 낮거나 전혀 없게 나타난 것은 문헌정보학 분야의 인용패턴의 차이로 설명할 수 밖에 없다.

CODE	서 명	인용문헌수
F01	우리나라 정보학교육의 회고와 FIABID 에 기초한...	20
F07	문헌정보학의 학명에 대한 고찰	67
F10	일본의 도서관 정보학 교육	16
F11	미국의 정보학 교육	29
F12	동 서독의 사서 정보전문가 교육 비교 연구	30
G05	전문대학도서관과의 모형교육과정 수정개발에 관한.	45

표 4-7 문헌정보학 내 유사주제 문헌군

다음 표 4-8 은 기준값 $\tau = 2$ 에서 형성된 완전 연결 그래프의 문헌집단 중 문헌정보학 관련 문헌의 서명을 수록한 것이다. 표에서도 알 수 있듯이 두 문헌의 관련 주제는 학명 개칭과 아울러 우리나라 정보학 교육의 역사적 변화에 대한 개론적 의미를 내포하고 있다.

CODE	서 명	출 전
71리재철	문헌과학과 문헌사의 소임: 도서관학의 학명과 사서의 직명 개칭을 바란다	도서관보 (국립도서관) 26, 5 (1971, 5): 1
72정필모	학문명칭으로서의 '문헌과학'에 대한 재고	도협월보 13, 9 (1972, 9): 13-15.

표 4-8 $\tau = 2$ 에서의 완전연결 그래프 문헌군 (문헌정보학)

동시인용 기법과는 달리 서지결합 기법은 문헌 집단의 주제적 관련성을 제시해 주고 있다. 즉, 인용문헌을 매개로한 같은 인용문헌을 동시에 갖는 문헌간에 주제적 관계가 성립함을 기초로 한 검색지도의 설계이다. 본 실험에서 사용한 질문문헌군을 대상으로 서지결합을 실시한 결과 다음과 같은 결과를 도출하였다. 도출된 결과에 따르면 유전공학의 37 개 문헌 중 19 개 문헌이 서지결합에서 서로 아무런 관련이 없음으로 나타났으며 문헌정보학에서도 26 개 문헌 중 무려 18 개 문헌이 서로 무관함으로 나타났다. 서지결합 분석 역시 동시인용 분석과 아울러 문헌집단의 특성을 잘 말해주고 있다. 앞에 분석된 결과와 비교하여 볼 때 (표 4-7) 문헌정보학 영역 논문의 참고문헌 리스트가 완전하지 못함을 나타내고 있다.

그림 4-9 는 서지결합을 통한 유전공학 문헌 37 개와 문헌정보학 관련문헌 26 개의 결합성을 나타내고 있으며 그림에서 보듯이 문헌정보학 보다는 유전공학이 훨씬 큰 집단으로 형성되었음을 알 수 있다.

본 실험에서는 실시하지 않았으나 서지결합이나 동시인용 분석을

통한 실험에서 우리는 핵심 문헌군을 선정해 볼 수 있다. 예를 들어

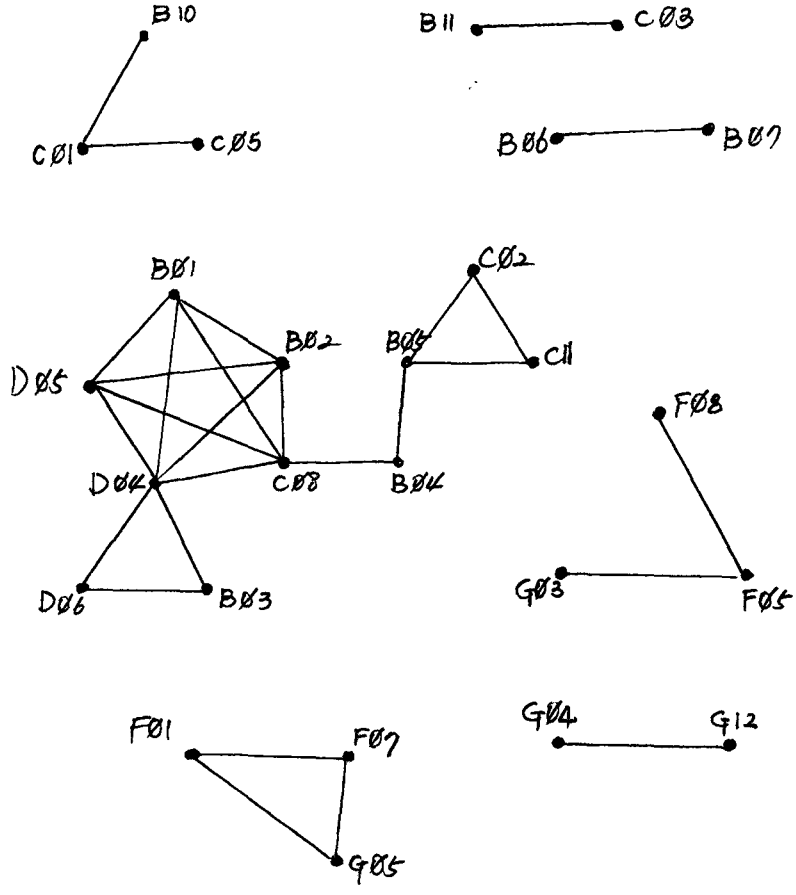

그림 4-9 실험문헌 집단의 서지결합 그래프

그림 4-9 의 서지결합에서 문헌 C08을 제거했을 경우 그래프가 크게 둘로 나뉘어짐을 볼 수 있다. 이것은 문헌 C08이 그 집단의 주제적으로 핵심적 역할을 수행함을 설명해 준다.

Ⅴ. 결론 및 제언

앞에서 분석한 실험의 결과와 본 실험에서 제시한 가설을 놓고 볼 때, 실험집단의 선정이 애매함을 알 수 있으나 나름대로 분석에 대한 해석은 가능하다고 본다.

본 실험에서 얻은 결과는 앞으로 정보검색의 새로운 지표와 방법으로 설정할 수 있으며 동시인용과 서지결합이 보여주는 주제적 검색 시스팀에 대한 보완적 역할은 매우 크다고 본다. 특히, 실험 데이터를 특정 연대로 한정하지 않고 일반적인 검색 시스팀을 통한 주제 검색에서 일차 선별된 문헌을 대상으로 이차적 검색을 시도할 때 동시인용 또는 서지결합의 기법을 쓴다면 보다 현실적이고 실용적인 결과를 얻을 수 있으리라 믿는다.

참 고 문 헌

Berge, Claude. The Theory of Graphs and Its Applications [translated by Alison Doig]. Westpoint, CT : Greenwood, 1962.

Garfield, E. "Citation Analysis As a Tool in Journal Evaluation." Science 178 (1972): 471-479.

Gerson, Gordon M. "Cliqueing — a Technique For Producing Maximally Connected Clusters." JASIS 29 (1978):125-129.

Jeong, Jun Min. The Ecology of the Scientific Literature and Information Retrieval (Unpublished PhD Dissertation Case Wester Reserve Univ., 1985)

Kessler, M. M. "Bibliographic Coupling Between Scientific Papers." American Documentation 14 (1963): 10-25.

Margolis, J. "Citation Indexing and Evaluation of Scientific Papers." Science 155 (1967): 1213-1219.

Small, Henry. "Co-Citation in The Scientific Literature: A New Measure of The Relationship Between Two Documents." JASIS 24 (1973): 265-269.

The Development of The Information Retrieval System By The Scientific Communication Network

Jun Min Jeong*

ABSTRACT

The paper suggests newly conceptualized information retrieval system on the notion of citation analysis. The paper also criticizes the traditional information retrieval techniques using Boolean logic.

The underlying assumption of this paper is that any pair of papers cited by one paper could be strongly related each other in meaning (Co-citation Analysis). And also any two papers to share same references could be similar each other (Bibliographic Coupling). By using graph algorithm, the networks of two kinds of the papers (the citing group, the cited group) is made in the fields of the genetics and the information and library science.

The results say that the maps or networks for cited and citing groups can be useful when applied to the paper set made by the broad searching by subjects or keywords.

* Associate Professor
Dept. of Library & Information Science, Chonnam National University