지적 구조의 규명을 위한 네트워크 형성 방식에 관한 연구*

A Study on the Network Generation Methods for Examining the Intellectual Structure of Knowledge Domains

이 재 윤(Jae-Yun Lee)**

목 차

초 록

이 연구에서는 지적 구조 분석을 위해서 계량서지적 자료를 시각적으로 표현하는 다양한 네트워크 형성 방식에 대해서 사례와 함께 각각의 특성을 살펴보았다. 기준값 절단 방식, 최근접이웃 그래프, 최소비용 신장트리, 패스파인더 네트워크의 네 가지 네트워크 형성 방식 중에서 전체 구조와 세부 구조의 표현 능력이 모두 뛰어난 패스파인더 네트워크 알고리즘이 최근 가장 활발히 응용되고 있다. 최근접이웃 그래프는 아직까지 계량서지적 분석에 응용된 사례는 없으나 간단한 알고리즘과 클러스터링 능력 등과 같은 지적 구조 규명에 도움이 될 수 있는 몇 가지 장점을 갖추고 있는 것으로 확인되었다. 다차원척도나 군집분석과 달리 네트워크를 이용한 시각화에서는 입력자료의 전처리에 따라서 생성된 지적 구조의 차이가 큰 것으로 나타났다. 이 연구에서 고찰한 여러 네트워크 형성 방식을 적절히 활용함으로써 국내의 지적 구조 규명 연구를 활성화할 수 있을 것이라 기대된다.

ABSTRACT

Network generation methods to visualize bibliometric data for examining the intellectual structure of knowledge domains are investigated in some detail. Among the four methods investigated in this study, pathfinder network algorithm is the most effective method in representing local details as well as global intellectual structure. The nearest neighbor graph, although never used in bibliometic analysis, also has some advantages such as its simplicity and clustering ability. The effect of input data preparation process on resulting intellectual structures are examined, and concluded that unlike MDS map with clusters, the network structure could be changed significantly by the differences in data matrix preparation process. The network generation methods investigated in this paper could be alternatives to conventional multivariate analysis methods and could facilitate our research on examining intellectual structure of knowledge domains.

키워드: 지적 구조, 네트워크 분석, 패스파인더 네트워크, 정보시각화, 계량서지학
Intellectual Structure, Network Analysis, Pathfinder Network, Information Visualization, Bibliometrics

---

* 이 논문은 2005년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음.(KRF-2005-003-H00010)
** 경기대학교 문헌정보학전공 조교수(memexlee@kgu.ac.kr)
논문접수일자 2006년 5월 15일
게재확정일자 2006년 6월 15일

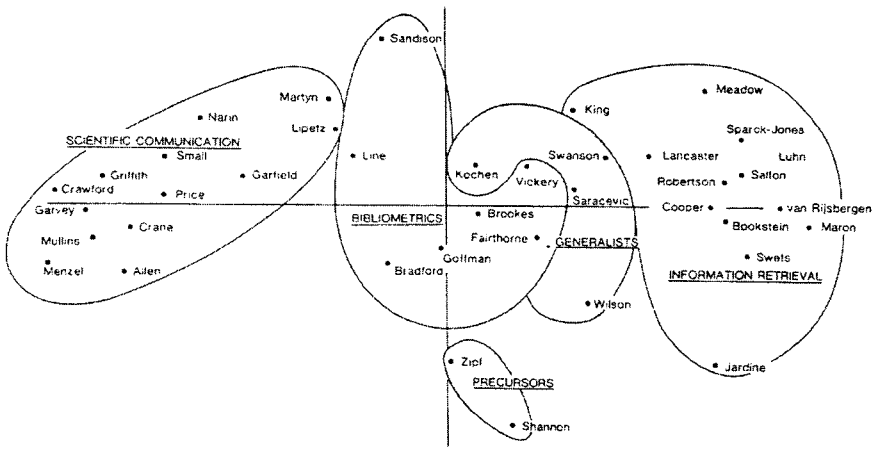
1. 서 론

계량서지학 분야에서는 다양한 방법으로 특정 학문분야 또는 산업분야의 지적 구조를 규명하는 작업을 지속해왔다. 계량서지적 분석의 조사 대상은 그 규모에 따라서 미시적 수준(microlevel), 중간 수준(mesolevel), 거시적 수준(macrolevel)으로 다양하다(Tijssen & van Raan 1994). 미시적 수준의 분석 단위로는 개인이나 문헌, 특허, 웹페이지 등을 들 수 있고, 중간 수준에서는 연구집단, 대학, 회사, 학술지, 웹사이트가 단위가 되며, 학문분야나 국가, 최상위 웹도메인을 단위로 하는 분석은 거시적 수준이라고 할 수 있다.

어떤 대상을 단위로 하여 분석하더라도 지적 구조를 표현하는 단계는 원 자료 추출, 분석단위 결정, 측정, 연관성 산출, 다변량분석, 시각적 표현의 여섯 단계로 나누어 볼 수 있다. 이중에서 다변량분석을 위해서는 주로 군집분석, 요인분석과 같은 기법을 사용되고 있다. 다변량분석 기법 중에서도 다차원척도법(MDS)이나 자기조직신경망(SOM)과 같은 차원축소 기법에서 차원축소를 2차원까지 진행하면 다변량분석의 결과가 바로 시각적 표현이 되기도 한다.

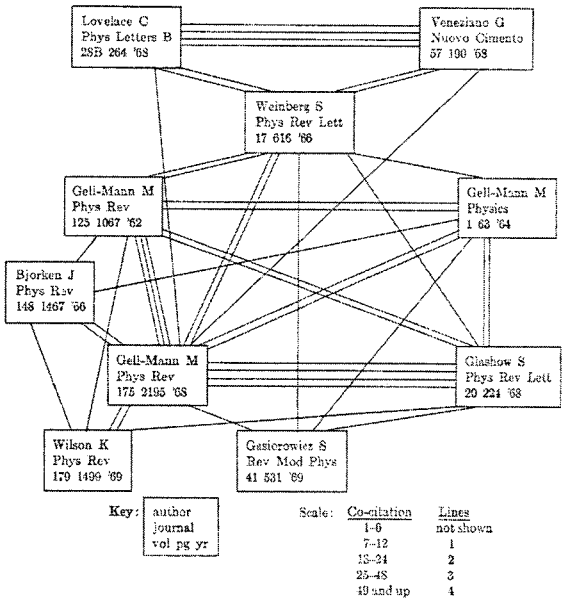
다차원척도법과 군집분석은 White와 Griffith (1981)가 저자동시인용분석에서 사용한 후 지적 구조의 시각적 표현 기법으로 가장 널리 사용되어왔다(그림 1 참조). 비록 분석 대상은 저자동시인용(White & Griffith 1981)을 비롯하여 저널동시인용, 동시출현단어(Callon, Law & Rip 1986), 웹사이트 동시링크(Larson 1996) 등으로 변화해왔지만 표현 방식은 크게 다르지 않았다.

지적 구조를 시각화하는 다른 방법으로는 네트워크 표현을 들 수 있다. 네트워크 표현은 측정 대상을 노드로, 관계를 링크로 나타내는 방식으로서 문헌동시인용분석을 제안한 Small(1973)이 사용한 것에서 알 수 있듯이(그림 2 참조) 다차원척도법보다 지적 구조의 표현에 사용되어온 역사가 오래되었다. 1970년대 말까지는 개별 문헌네트워크와 문헌군집 네트워크를 별도로 그려서



〈그림 1〉 다차원척도법과 군집분석을 이용한 정보학 분야 지적 구조의 표현

출처: White & Griffith(1981)



〈그림 2〉 기준값 절단 방식 - 문헌동시인용

출처: Small(1973)

지적 구조를 분석하였다(Garfield 1979). 그러나 White와 Griffith(1981)가 사용한 다차원척도법과 군집분석은 개별 문헌·저자와 군집을 동시에 표현하면서 전체적인 지적구조를 보여줄 수 있는 장점 때문에 이후 연구에서 널리 사용되었다. 상대적으로 이전의 단순한 네트워크 표현 방식을 사용하는 연구는 줄어들었다.

최근에는 다차원척도법도 지적 구조의 시각적 표현 도구로서 한계가 있음이 여러 연구에서 지적되고 있다. 관련 주장을 정리하면 다음과 같다.

첫째, 표현해야할 개체가 너무 많으면 SPSS와 같은 통계패키지의 다차원척도분석 모듈이 수용할 수 없다(100개가 한계). 설사 수용되더라도 개체가 너무 많으면 개체간 거리의 변형을 뜻하는 스트레스값이 매우 높아져서 원 자료의 왜곡이 심각하다.

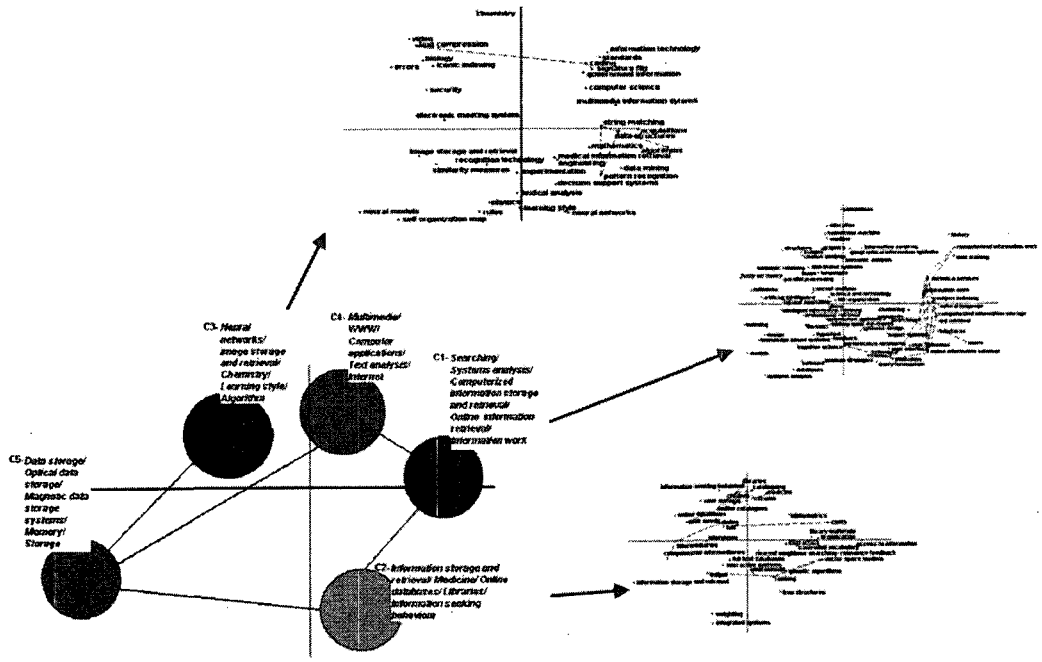
둘째, 수십 개 이상의 개체를 2차원 MDS 지도로 표현할 때 세부구조의 표현력이 매우 떨어진다(Schvaneveldt 1990; Noel et al. 2003). 이는 전체적인 배치는 그럴 듯 하더라도 국지적으로는 개체간의 거리가 원래의 관계를 그대로 반영하지 못한다는 뜻이다. 실제로 2차원 MDS 지도상에서 특정 개체와 가장 가깝게 위치한 개체가 원래의 상관계수로는 가장 가까운 관계가 아닌 경우가 흔하다. 이 때문에 Ding, Chowdhury, Foo(2001)는 전체 MDS 지도는 군집을 단위로 표시하고 각 군집별로 세부적인 MDS 지도를 다시 생성하는 다단계 매핑 전략을 제시한 바 있다(그림 3 참조).

셋째, 표현된 차원을 직관적으로 해석하기가 어렵다(Börner et al. 2003). 다차원척도법의 결과인 MDS 지도에서 가로축과 세로축은 자체적으로 의미를 가지고 있지 않으므로 연구자가 개체의 배치를 보고 임의로 해석해야 한다.

넷째, MDS 지도만으로는 소주제 집단을 식별할 수 없다. 이 때문에 군집분석을 함께 하는 경우가 많은데, 그 경우에도 MDS 지도가 세부구조의 표현력이 떨어지므로 실제 가까운 노드와 지도상 가까운 노드가 일치하지 않아서 〈그림 4〉와 같이 구조를 파악하기 어려울 정도로 어지럽게 구불구불한 군집이 그려지는 경우가 나타난다.

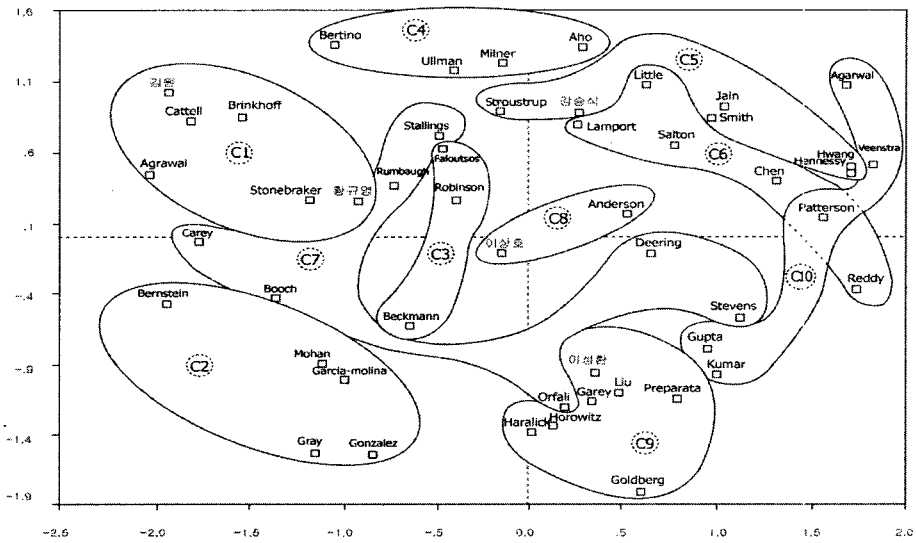
이와 같은 다차원척도법의 한계가 드러나면서 지적 구조의 시각적 표현을 위한 대안으로 네트워크 표현이 다시 주목을 받게 된 것은 인지심리학 분야에서 개발된 패스파인더 네트워크(Pathfinder Network: PFNet으로 약칭) 알고리즘(Schvaneveldt 1990)을 McCain(1995)이 동시분류분석(Co-classificaton analysis)에, Chen(1999), White(2003) 등이 저자동시인용분석에, Marion과 McCain(2001)이 저널동시인용에 도입하면서부터이다.

지적 구조를 네트워크로 표현하기 위해서는 원 자료로부터 나타낼 네트워크를 어떻게 형성할 것인가가 중요하다. 패스파인더 네트워크 알고리즘도 네트워크 형성 방식의 하나라고 볼 수 있다. 이 연구에서는 최근 다차원척도법의 한계가 지적되면서 그 대안으로 다시 부각되고 있는 네트워크 형성 방식에 대해서 사례와 함께 여러 방식의 특성을 살펴보기로 한다. 실제 네트워크 형성 사례로는 기존 연구의 네트워크 그림과 함께, 네트워크 표현을 사용하지 않은 연구의 자료를 네트워크로 표현하여 제시한다.



〈그림 3〉 MDS 지도에서 세부구조 표현을 위한 다단계 매핑 사례

출처: Ding, Chowdhury, Foo(2001)



〈그림 4〉 MDS지도에 표현된 복잡한 군집 구조 사례 - 컴퓨터과학 분야 지적 구조

출처: 이은숙(2003)

2. 네트워크 형성 방식

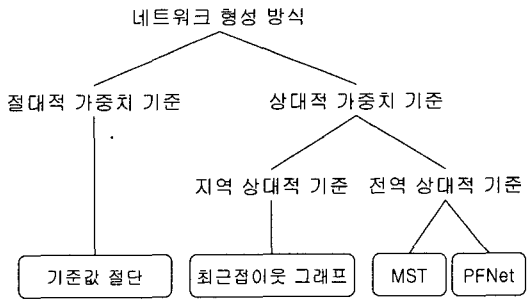
2.1 네트워크 형성 방식의 구분

주어진 연관성(또는 거리) 행렬로부터 네트워크를 형성하는 방식은 전체가 분리된 상태에서 가까운 노드 사이를 이어나가서 네트워크를 만드는 것이라고 할 수 있다. 이와 반대로 최대한의 링크가 연결되어 있는 상태에서 중요하지 않은 링크를 제거하는 것이라고 보는 입장에서는 링크 삭감 알고리즘(link reduction algorithm)이라고 부르기도 한다(Chen & Morris 2003).

어느 쪽으로 보든 간에 이 글에서 다루는 네트워크 형성 방식은 생성할(또는 제거할) 링크를 선택하는 기준에 따라 〈그림 5〉와 같이 구분할 수 있다.

주요한 링크를 선정할 때에는 가중치의 절대적인 기준값을 설정하는 경우와 그렇지 않는 경우가 있을 수 있다. 절대 기준값을 정하지 않을 경우에는 상대적 기준을 적용하게 된다. 기준값 절단 방식은 모든 링크의 가중치를 절대적인 기준으로 삼고 링크 생성 여부를 결정한다. 즉 생성된 링크의 가중치보다 가중치가 높으면서도 생략되는 링크는 있을 수 없다. 이와 달리 상대적 가중치 기준을 적용하는 방식에서는 가중치가 더 높으면서도 상대적인 중요도에 따라서 생략되는 링크가 있을 수 있다. 최근접 이웃 그래프는 각 노드별로 이 노드가 관련된 링크가 가진 가중치의 상대적인 높고 낮음을 기준으로 링크를 선정한다. 최소신장트리(Minimum Spanning Tree: MST로 약칭)나 PFNet은 각 노드별 상황과 전체적인 구조를 함께 고려해서 링크의 중요성을 상대적으로 판단한다.

다음 절부터는 각 방식의 특징에 대해서 사례와 함께 살펴보기로 한다.



〈그림 5〉 네트워크 형성 방식의 구분

2.2 기준값 절단 방식

기준값 절단 방식은 가장 간단한 네트워크 생성 방식으로서 Small(1973)의 연구(그림 2 참조)에서처럼 동시인용네트워크 연구의 초기부터 적용된 방식이다. 이 방식에서는 네트워크 전체적으로 가중치가 일정한 기준값 이상인 링크만 남기기 때문에 모든 링크에 대해서 절대적으로 동일한 기준을 적용하는 방식이다. 주로 문헌동시인용분석(McKechnie et al. 2005; Small 1973; Small & Griffith 1974)이나 공저자분석(Börner 2005; Kretschmer & Aguillo 2004; Nagpaul 2002; Newman 2001; Newman 2004; Otte & Rousseau 2002)에서 많이 사용되어왔다. 1970년대에는 주로 문헌동시인용분석 연구에서 사용되었으나 2000년 이후에는 공저자분석 연구에서 활발히 사용되는 것이 특징이다.

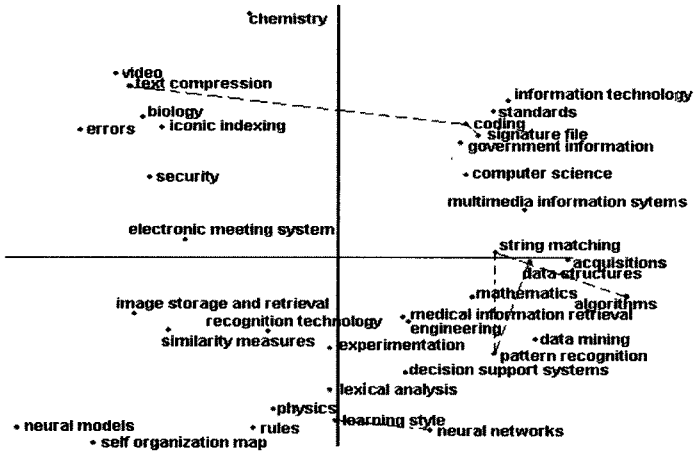
〈그림 2〉에서는 기준값을 7로 하여 이보다 값이 큰 경우만 선을 표시하고 이보다 값이 큰 경우에는 단계적으로 연결선의 수를 늘려서 강한 연결을 표현하였다. 이와 같이 링크 가중치의 차이를 연결선의 개수로 표현하는 방법 이외에도 Börner(2005)와 같이 연결선의 굵기로 나타내는 방법도 사용된다.

기준값 절단 방식에서는 기준값을 낮출수록 링크의 수가 많아져서 그래프가 복잡해진다. 따라서 핵심 저자나 주요 주제와 같은 정보가 부각되지 않는다. 반대로 기준값을 높이면 그래프가 분할되어 서브그래프나 고립 노드가 나타난다. 그 결과 전체적인 연결 흐름이 나타나지 않는 경우가 있다. 따라서 적당한 기준값을 찾는 것이 중요하다.

동시출현단어분석(co-word analysis)에서는 다차원척도법에 의한 MDS지도 상에서 보조수단으로 기준값 절단 방식을 사용하기도 한다(Ding et al. 2001). 〈그림 6〉은 코사인유사도 0.2 이상인 키워드를 점선으로 연결한 상태이다.

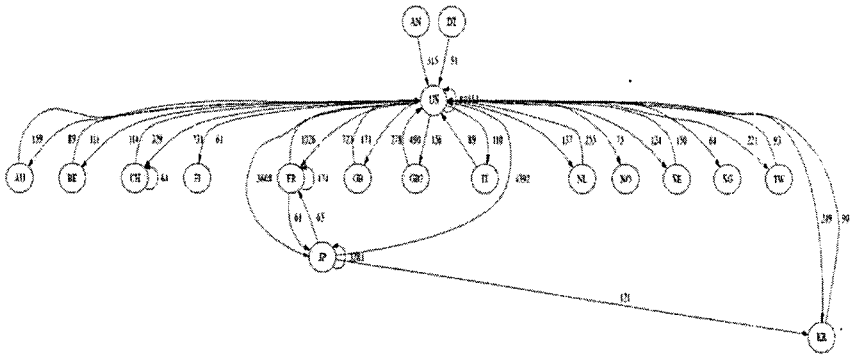
한편 공저자 네트워크 분석에는 무조건 한 편이라도 공저를 하면 링크를 생성하는 경우가 대부분이다(Newman 2001). 동시인용이나 단어동시출현에 비해서 공저라는 사건의 발생확률은 상당히 낮아서 1회라도 발생하는 것 자체가 학술 협력의 지표로서 의의가 있기 때문이다.

동시출현자료를 이용한 경우에는 대개는 고리(loop: 노드 자신에 대한 링크)가 없는 단순 그래프이지만 직접 인용망의 경우에는 드물게 〈그림 7〉처럼 고리를 표현한 경우도 있다.



〈그림 6〉 기준값 절단 방식 - MDS 지도에 표시한 사례

출처: Ding et al.(2001)



〈그림 7〉 기준값 절단 방식 사례 - 나노수준 과학기술분야 국가간 특허인용망(빈도 50이상 연결 기준)

출처: Huang et al.(2003)

2.3 최근접이웃 그래프

최근접이웃 그래프(Nearest Neighbor Graph; NNG로 약칭)는 각 노드마다 일정한 수(대개는 1 또는 작은 수 k)의 가까운 노드를 연결해서 생성된다(Eppstein et al. 1997). k를 붙여서 k-최근접이웃 그래프라고 하는 경우도 있다. 각 노드별로 상대적으로 가까운 노드를 연결하기 때문에 링크 가중치는 국지적인 관점에서 상대적인 기준으로 사용된다. 고려하는 최근접 이웃의 수인 k가 1에 가깝게 적을수록 연결되는 링크의 수가 적어지므로 전체 노드가 하나로 연결되지 않고 여러 개의 서브 그래프로 나누어질 가능성이 높다. 반대로 k가 높으면 링크의 수가 많아서 구조가 복잡해지며 서브 그래프의 수가 줄어들거나 전체가 하나의 그래

프로 연결될 가능성이 높아진다.

최근접이웃 그래프는 주로 사회과학 분야에서 설문이나 관찰을 통해서 인적 네트워크를 조사할 때 사용되어왔다(예: "17대 의원 네트워크 대해부" - 조선일보 2004년 8월 24일자 A4-A6면). 그 이유는 모든 구성원 상호간의 관계를 설문이나 단기간의 관찰로 파악하는 것이 거의 불가능하기 때문이다. 따라서 인적 네트워크를 비롯한 사회 연결망 분석에서는 소수의 가까운 관계 또는 먼 관계에 대해서 집중하여 파악하는 것이 일반적이다.

아직까지 동시인용 자료를 비롯한 계량서지 자료에 최근접이웃 그래프를 적용해본 연구는 없지만, 다음과 같은 몇 가지 장점을 고려할 때 최근접이웃 그래프를 이용한 지적 구조의 규명 작업도 필요하다고 생각된다.

우선 최근접이웃 그래프에서는 기준값 절단 방식과 달리 다른 노드와 연결되지 않는 단독 고립 노드가 나타나지 않는다. 생성되는 가장 작은 단위는 노드 두 개짜리 서브그래프이다. k가 크면 전체가 하나의 그래프로 연결되지만 1이나 그에 가깝게 낮추면 여러 개의 서브 그래프로 분할되기 쉽다. 이 성질을 이용해서 군집을 생성할 수 있다. 그리고 각 노드마다 가까운 노드를 꼽기 때문에 방향성 그래프를 생성할 수 있다. 따라서 지역 중심 노드의 파악이 손쉽다.

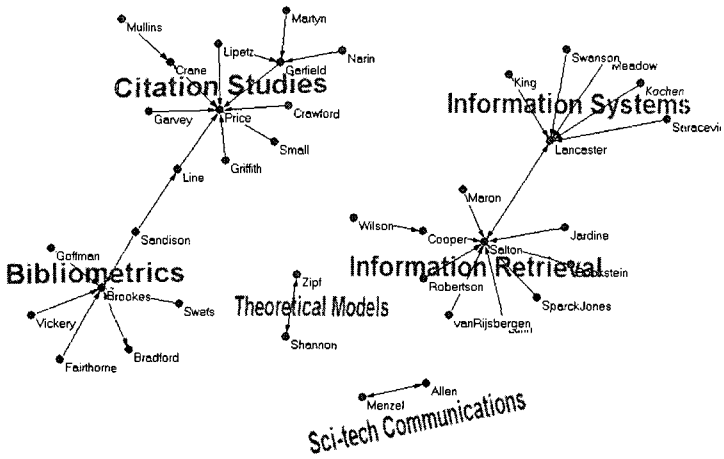
〈그림 8〉은 저자동시인용분석을 제안한 White와 Griffith(1981)의 1970년대 정보학분야 저자동시인용자료를 이용해서 생성한 최근접이웃 그래프(k=1)이다. (a)는 동시인용빈도행렬을 그대로 입력한 경우이고, (b)는 코사인계수를 적용해서 동시인용빈도를 정규화한 행렬을 입력하여 처리한 결과이다. 빈도행렬을 입력자료로 하지 않고 코사인계수나 자카드계수로 정규화한 행렬을 대상으로 하는 경우에도 k값을 낮출 때와 마찬가지로 서브그래프가 많아지는 경향이 있다. 이 경우에는 링크 가중치가 정규화되므로 빈도가 높은 특정 노드로의 집중 경향이 약해지기 때문이다.

〈그림 8〉을 보면 빈도행렬을 이용한(a)는 전체가 큰 두 개의 서브그래프와 작은 서브 그래프 두 개로 구성된다. 큰 서브그래프 두 개에서 화살표가 집중되는 저자는 각각 Salton과 Lancaster, 그리고 Price와 Brookes임이 뚜렷하게 나타난다. 이를 근거로 큰 두 서브그래프의 주제는 정보검색 및 정보시스템 분야, 계량서지학 및 이론정보학 분야로 볼 수 있다. 이와 달리 코사인유사도행렬을 이용한 최근접이웃 그래프인(b)는 9개의 중소규모 서브 그래프가 생성된다. 이를 White와 Griffith(1981)가 실시한 요인분석 결과에 나타난 7개 요인과 비교해보면 대부분 일치하였다. 따라서 최근접이웃 그래프는 소주제를 식별하는 유용한 도구가 될 가능성이 있다.

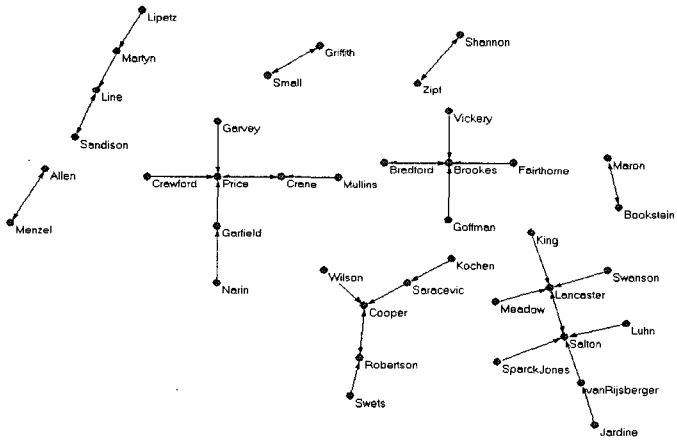
결국 다차원척도법을 사용한 〈그림 2〉와 비교하였을 때 〈그림 8〉의 최근접이웃 그래프에서는 소주제 군집을 파악하면서 동시에 핵심 연구자도 식별할 수 있는 것으로 나타났다. 향후 다양한 자료에 대해서 최근접이웃 그래프를 이용한 지적구조의 규명을 시도해볼 가치가 충분하다고 판단된다.

2.4 최소신장트리

어떠한 방식으로든 집합내의 모든 구성 노드를 일정한 과정을 거쳐서 차례대로 모두 연결



(a) 빈도행렬을 입력한 결과

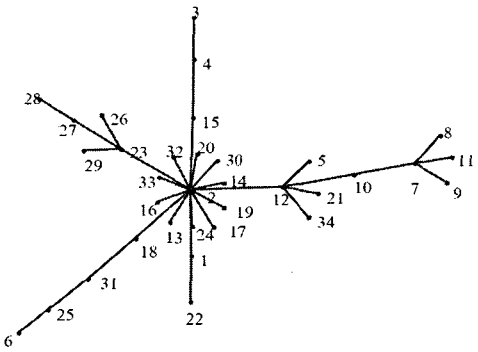


(b) 코사인유사도행렬을 입력한 결과

〈그림 8〉 최근접이웃 그래프(k=1) - 저자동시인용
(White & Griffith(1981)의 자료로 생성한 것임).

시켜주는 방식을 신장트리라고 부른다. 흔히 그래프 탐색기법으로 불리는 깊이우선이나 너비우선 방식도 신장트리를 만드는 알고리즘이다. 신장트리 중에서도 모든 링크의 가중치 합이 가장 작은 트리가 최소(비용)신장트리(Minimum (Cost) Spanning Tree: MST로 약칭)이다.

최소한의 비용으로 전체가 연결되도록 보장하기 위해서 링크 가중치가 아무리 높더라도 전체 구조상 중복된 링크이면 제거하는 상대적인 기준을 적용한다. 생성 결과는 노드 수가 n개인 경우 n-1개의 링크로 연결된 네트워크가 된다. 간단한 구현 알고리즘으로 Kruskal 알고리즘(Kruskal 1956)과 Prim 알고리즘(Prim 1957)이 널리 알려져 있다.



〈그림 9〉 MST 사례 - 문헌동시인용

출처: Noel et al.(2002)

최소신장트리는 시각적 표현을 위해서 흔히 사용되는 방식이지만 계량서지학 분야에서 지적 구조의 시각화를 위한 용도로 사용된 예는 Noel 등(2002), Chen과 Morris(2003) 정도에 불과하다. 대부분의 경우에는 최소신장트리와 유사하면서도 세부구조의 표현능력 면에서 유리한 패스파인더 네트워크가 대신 사용되고 있다. 두 방식의 비교는 다음 절에서 다루기로 한다.

2.5 패스파인더 네트워크

2.5.1 패스파인더 네트워크의 개념

패스파인더 네트워크(PFNet)는 가중치가 있는 모든 링크가 생성된 상태(즉, 기준값 절단 방식에서 기준값을 0초과로 한 경우)에서 삼각부등식(triangle inequality)을 위반하는 경로를 제거하여 생성되는 네트워크이다(Schvaneveldt 1990). 이를 생성하는 알고리즘을 패스파인더 네트워크 알고리즘이라고 하며, 때로는 다변량 분석 기법의 일종으로 간주하여 패스파인더 네트워크 척도법(Pathfinder Network Scaling) 또는 더 간단히 패스파인더 척도법이라고 부르기도 한다.

삼각부등식 위반 여부를 결정하기 위해서는 두 가지 파라미터 q와 r이 필요하다. 파라미터 q는 노드 사이의 경로거리를 산출하는 데 고려하는 최대 링크의 수(거리산출범위)를 뜻한다. q는 2에서 n-1(n은 노드의 총 수)까지 설정한다. q가 커질수록 조사대상 범위가 넓어져서 엄격한 조건이 되므로 남는 링크의 수가 줄어든다. 파라미터 r은 다음과 같은 민코프스키 거리 공식의 제곱수로서 두 노드 n_1과 n_k 사이의 특정 경로를 구성하는 여러 링크가 가지고 있는 가중치를 거리 $w_{n_1 n_k}$에 반영하는 방법을 뜻한다.

$$w_{n_1 n_k} \leq \left(\sum_{i=1}^{k-1} w^r_{n_i n_{i+1}} \right)^{\frac{1}{r}} \quad \forall k = 2, 3, \ldots, q$$

이 공식에서 r이 1이면 각 링크 가중치의 합이 그대로 경로의 거리가 되고, r이 무한대이면 경로를 구성하는 링크의 가중치 중 최대값이 경로의 거리가 된다. r이 커질수록 경로의 길이가 짧아지므로 역시 엄격한 조건(위반되기 쉬운 조건)이 되어 남는 링크의 수가 줄어든다.

예를 들어서 〈그림 10〉과 같이 노드 A에서 노드 C로 연결되는 경로 [a→c]의 제거 여부를 판단하는 문제를 살펴보자. [a→c]의 직접연결 이외에 고려할 수 있는 대안 경로는 [a→b→c]의 간접 경로가 있다. 이때 r=∞이고 q=2이면 [a→b→c]의 민코프스키 거리가 [a→b]의 거리 2와 [b→c]의 거리 3 중에서 최대값인 3이 되므로, 길이가 4인 경로 [a→c]는 삼각부등식 위반으로 제거된다. 이와 달리 r=1이라면 경로 [a→b→c]의 길이는 2+3=5가 되어 이보다 길이가 짧은 경로 [a→c]는 제거되지 않는다.

다시 말해서 PFNet에서의 삼각부등식이 위반되는 경우란, 직접 연결되는 긴 경로보다 여러 개(파라미터인 q개 이내)의 짧은 링크를 통해 간접적으로 연결되는 경로가 존재하는 경우를 말한다.

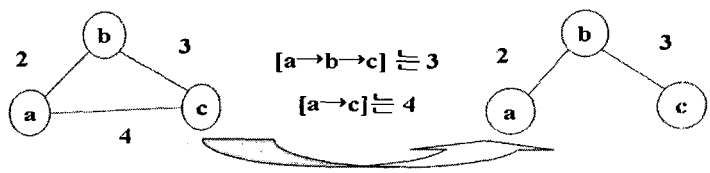
PFNet으로 지적구조를 표현하기 위해서는 흔히 가장 엄격한 조건인 r=∞, q=n-1 로 설정(PFNet(r=∞,q=n-1) 이라고 표기)하여 주요 흐름이 표현되도록 한다(Chen 2006a). PFNet(r=∞,q=n-1)은 가능한 모든 MST를 합친 것과 같다(Schvaneveldt 1990). 즉, MST에서는 링크 가중치가 동률일 때 임의의 링크를 선택하는 대신 동률인 링크를 모두 선택하는 경우가 PFNet에 해당한다. 따라서 MST와 PFNet은 유사한 점이 많다. 만약 각 링크의 가중치가 동률인 경우가 없다면 MST와 PFNet(r=∞, q=n-1)은 같아진다.

이런 이유로 MST와 PFNet은 개별 노드의 전체적인 배치 구조면에서 상당히 비슷한 결과를 보여준다. 결과로 나타난 네트워크에 드러난 전체적인 흐름은 큰 차이가 없다. 그러나 세부 구조에 있어서는 PFNet를 사용한 경우가 MST를 사용한 경우보다 잘 나타나는 경우가 있다. MST가 동률인 링크 중 임의로 하나를 선택하기 때문에 큰 그림은 PFNet과 차이가 없어도 세부 연결정보에서 누락되는 부분이 발생한다.

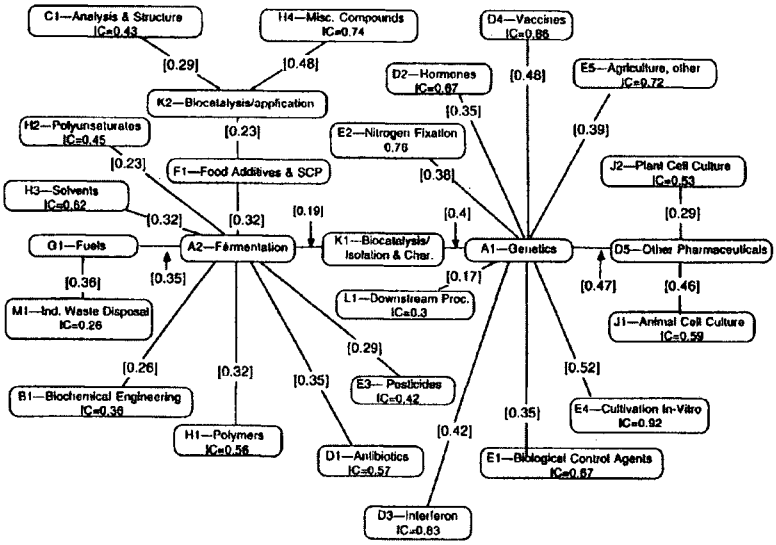
반면에 PFNet이 MST에 비해서 알고리즘이 복잡하며 처리시간과 기억공간을 더 필요로 한다. 삼각부등식을 위반하는지 여부를 행렬 전체에 대해서 검사해야 하기 때문이다.

2.5.2 패스파인더 네트워크의 응용

PFNet은 원래 인지심리학 분야에서 개발하였으나 McCain(1995)이 생물공학 분야의 특허에 대한 연구를 수행하면서 특허항목의 동시분류 분석에 적용한 이후 계량서지학 분야에 도입되었다(그림 11 참조). McCain은 이후에 소프트웨어공학 분야 저널동시인용분석에도 PFNet을 적용하였다(Mario & McCain 2001). 특히 1998년에 McCain과 함께 다차원척도법과 군집분석을 이용해서 정보학분야의 저자동시인용 분석을 수행했던 White가, 2003년에 동일한 자료에 대해서 PFNet을 적용하여 개선된



〈그림 10〉 r=∞, q=2일 때 삼각부등식 위반에 따른 링크 제거의 예



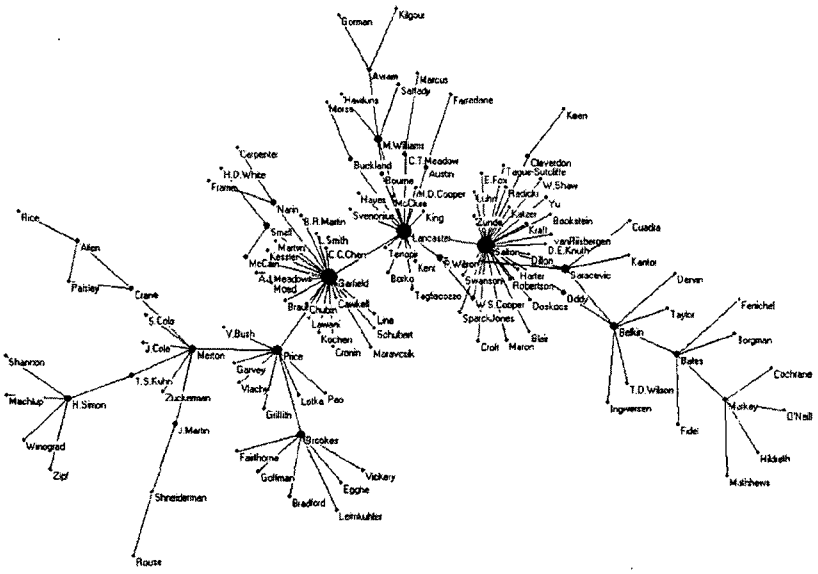
〈그림 11〉 PFNet 사례 - 생물공학분야 특허항목의 동시분류 분석

출처: McCain(1995)

연구 결과(White 2003)를 발표한 것이 PFNet의 장점이 더 넓게 알려지는 계기가 되었다(그림 12 참조). 한편 1990년대 후반 이후 정보시각화를 연구하는 C. Chen이 일련의 연구를 통해서 PFNet 관련 논문을 발표하고 계량서지적 자료를 PFNet으로 분석할 수 있는 공개 소프트웨어인 CiteSpace(Chen 2006)를 개발하면서 PFNet 기법의 보급에 기여하였다. Chen(2004)은 특히 PFNet을 그대로 제시하는 것에 그치지 않고 각 노드의 역할을 〈그림 13〉에서와 같이 전환점 역할 노드, 피벗 노드, 허브 노드 등으로 구분하여 파악할 것을 제안하였다.

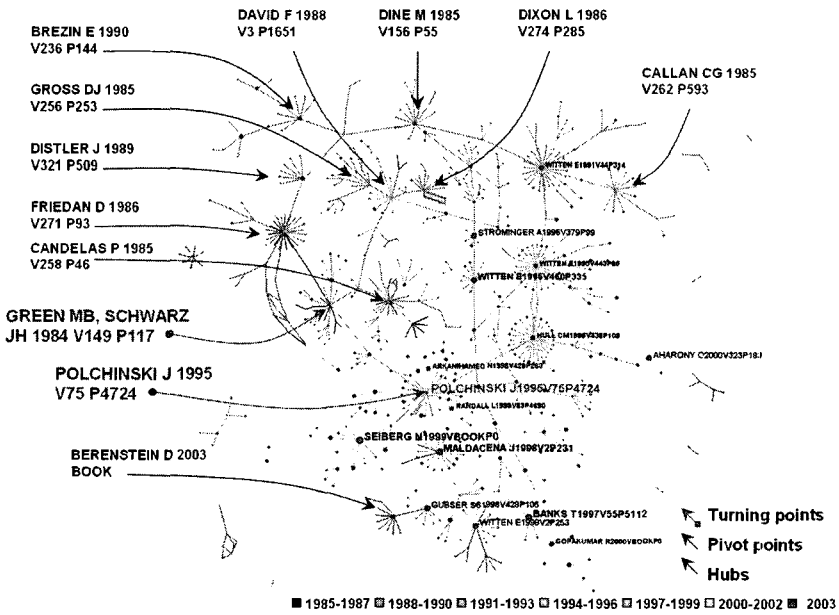
계량서지적 분석에 PFNet을 적용해본 연구자들은 다차원척도법에 비해서 세부구조가 잘 드러날 뿐만 아니라 전체적인 구조도 더 뚜렷하게 제시해주는 것으로 인정하고 있다(Börner et al. 2003; Chen 2003; White 2003). 그에 따라 다차원척도법을 대체하거나 보완하는 용도로 PFNet을 사용하는 경우가 증가하고 있다.

한편 PFNet은 원래 인지심리학 분야에서 개발되었던 만큼 인지적인 연상 구조나 연결구조를 잘 반영하는 특성 때문에 검색이나 브라우징을 위한 이용자 인터페이스에 적용되는 경우도 적지 않다. Fowler, Wilson & Fowler (1992)가 처음으로 용어의 PFNet과 문헌의 PFNet을 검색 시스템의 인터페이스로 구현해 본 이후, 디지털도서관 시스템의 인터페이스로 저자동시인용 PFNet 지도를 사용하거나(White, Buzydlowski & Lin 2000) 용어동시출현 PFNet 지도를 사용한 시스템(Buzydlowski, White & Lin 2002)이 제안된 바 있다. 또한 내용기반 이미지 브라우징 시스템의 인터페이스로 이미지의 PFNet을 제안한 연구도 있다(Gagaudakis, Rosin & Chen 2000; Mukhopadhyay, Ma & Sethi 2004 - 〈그림 14〉 참조)



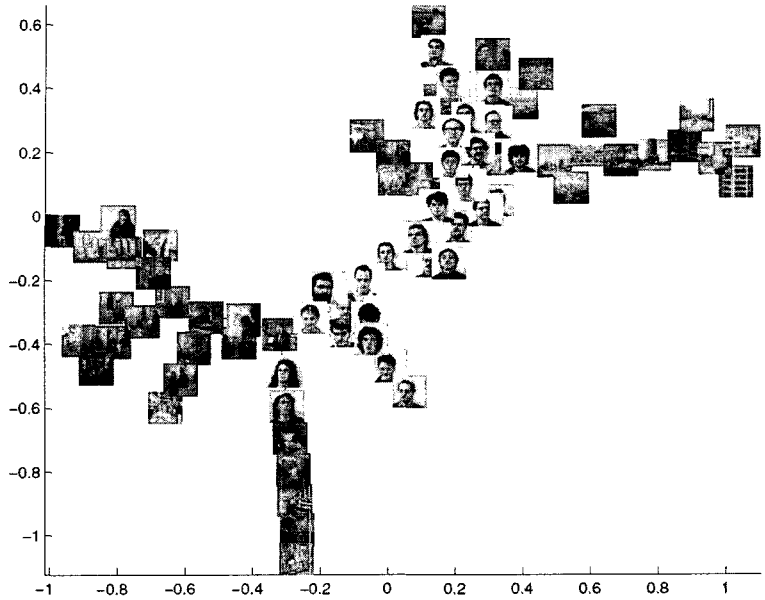
〈그림 12〉 PFNet 사례 - 정보학 분야 저자동시인용

출처: White(2003)



〈그림 13〉 PFNet 사례 - 초끈이론분야 문헌동시인용

출처: Chen(2004)



〈그림 14〉 PFNet 사례 - 내용기반 이미지 브라우징 시스템 이용자 인터페이스

출처: Mukhopadhyay, Ma & Sethi(2004)

3. 행렬값 가공처리가 네트워크 구조에 미치는 영향

3.1 행렬값의 가공 방식 구분

앞에서 살펴본 여러 네트워크 방식을 적용하기 위해서는 동시인용빈도와 같은 행렬 자료를 준비해서 입력해야 한다. 이때 행렬의 수치를 가공하는 방법에 따라서 생성되는 네트워크의 형태가 달라진다. 2장의 〈그림 8〉에 나타난 최근접이웃 그래프의 예를 보아도 빈도 행렬을 입력했을 경우와 코사인유사도로 정규화한 행렬을 입력한 경우에 다소 다른 구조가 만들어짐을 알 수 있다.

행렬값의 가공 방법에 따라서 지적 구조의 분석을 위한 입력 행렬은 1차 연관성 행렬과 2차 연관성 행렬로 나눌 수 있다.

1차 연관성 행렬은 빈도값을 그대로 이용하거나, 각자의 출현빈도(보정한 대각선값)를 감안하여 코사인계수 등으로 정규화한 값으로 구성된 행렬이다. 정규화 여부와 상관없이 1차 연관성 행렬은 양자간의 직접 동시출현 정도를 반영한 행렬이다.

2차 연관성 행렬은 White와 Griffith(1981)가 제안한 바와 같이 1차 연관성 행렬을 다시 가공하여 생성되는 연관성 행렬이다. 1차 연관성 행렬의 개별 프로파일 벡터간의 상관도를 피어슨 상관계수 등으로 한 차례 다시 계산하므로, 2차 연관성 행렬은 제삼자와의 동시출현 패턴의 유사함을 측정한 행렬이다. 따라서 2차 연관성을 프로파일 유사도라고 부르기도 한다.

2차 연관성 행렬을 생성할 때 1차 연관성 행

렬의 대각선값을 처리하는 방법은 두 가지로 나뉜다. McCain(1990)과 같이 대각선값을 결측치로 처리하면 이에 상응하는 양자간의 동시출현빈도가 연관성 산출에 전혀 반영되지 않기 때문에 순수한 2차 연관성을 측정하는 셈이다. 반면에 White와 Griffith(1981)처럼 프로파일을 구성하는 값 중에서 가장 큰 값 셋을 더해서 2로 나눈 값으로 보정해서 사용하면 양자간의 직접 동시출현을 어느 정도 반영하는 2차 연관성을 산출하게 된다. 비유적으로 말하자면 순수한 1차 연관성과 순수한 2차 연관성의 중간이므로 1.5차 연관성이라고 할 수 있을 것이다. McCain(1990)은 대각선값을 보정한 경우와 결측치로 처리한 경우를 비교해본 결과, 다차원척도법, 군집분석, 요인분석 모두 두 경우의 차이가 미미했다고 보고한 바 있다.

연관성 산출 방식에 따른 행렬값의 차이를 직접 확인하기 위해서 White와 Griffith(1981)의 정보학분야 저자동시인용 자료로부터 여러 방법으로 가공한 수치로 구성된 행렬간의 상관도를 구해보면 〈표 1〉과 같다. 이 표를 보면 1차 연관성 행렬인 빈도행렬은 역시 1차 연관성 행렬인 코사인유사도행렬과 가장 비슷하고, 2차 연관성 행렬인 대각선값 보정 후 상관계수 행렬은 역시 2차 연관성 행렬인 대각선값 결측처리 후 상관계수 행렬과 가장 비슷함을 알 수 있다. 2차 연관성 중에서도 행렬의 대각선 값을 보정한 경우가 결측으로 처리한 경우보다 원래의 빈도 행렬의 상관도가 높은 것으로 나타났다.

3.2 가공방식에 따른 네트워크 구조의 특성

Chen과 Morris(2003)는 PFNet을 생성할 때에는 입력 자료로 빈도행렬을 사용하는 것보다 자카드계수나 코사인계수와 같은 연관성척도로 정규화한 값을 사용해야 잠재적인 위험을 피할 수 있다고 지적하였다. 그 이유는 빈도행렬을 이용하면 동률값이 상대적으로 흔해서 링크가 더 많이 생성될 수 있으므로 최소한의 주요 링크만 남겨서 흐름을 볼 수 있다는 PFNet의 장점이 많이 희석되기 때문이다.

반면에 MST에 대해서 Noel 등(2002)은 정규화한 값보다는 원 빈도 행렬을 입력해야 한다고 주장하였다. 원래의 값이 아닌 정규화한 값을 사용하면 많은 링크를 받는 핵심 노드가 부각되는 트리 구조가 생성되는 대신에, 노드가 줄지어서는 사슬 효과(chain effect)가 나타나고, 그에 따라서 주요 노드와 관련 주제의 관찰이 어려워진다는 이유에서다.

사슬 효과의 문제는 PFNet에서도 지적된바 있다. White(2003)는 1차 연관성인 빈도행렬을 입력자료로 사용하는 것이 2차 연관성인 상

〈표 1〉 각 행렬간 상관도(spearman 순위상관)

행렬을 구성하는 값의 유형		1차 연관성		2차 연관성(상관계수)	
		빈도	코사인유사도	대각선값 보정	대각선값 결측
1차 연관성	빈도		0.960	0.789	0.672
	코사인유사도	0.960		0.858	0.754
2차 연관성	대각선값 보정	0.789	0.858		0.977
	대각선값 결측	0.672	0.754	0.977	

관계수행렬을 사용하는 것보다 바람직하다고 주장하였다. 그 이유는 상관계수 행렬을 입력할 경우에 핵심저자를 부각시키는 스타노드(다수의 링크가 연결된 노드)가 사라지는 현상이 발생하기 때문이라고 하였다.

White(2003)의 주장에 대해서 유의할 부분은 상관계수 행렬을 산출하는 방법이다. 그가 2차 연관성을 산출할 때 사용한 방법은 대각선값 결측처리 방식이었다. 앞 절에서 보았듯이 빈도행렬의 대각선값을 결측처리한 후 산출된 프로파일 상관계수 행렬은 원래의 빈도행렬과의 상관도가 매우 낮은 형태가 된다. 그 결과 원래의 빈도행렬에서 뚜렷하던 스타노드가 그 특성을 잃으면서 사슬 효과가 심화된 것으로 추정된다. 인용빈도가 높고 다수 저자와 동시인용되어 1차 연관성으로는 스타노드가 될 수 있는 저자가, 동시인용빈도의 높고 낮은 패턴까지 여러 저자들과 유사하지는 않기 때문에 순수한 2차 연관성 행렬에서는 스타노드가 되지 못하는 경우가 발생한다.

행렬값을 가공하는 것이 생성되는 네트워크에 실제로 어떤 결과를 가져오는지 알아보기 위해서 〈표 1〉의 네 가지 행렬을 입력하여 생성된 네 가지 PFNet(r=∞,q=n-1)을 〈그림 15〉에 각각 제시하였다. 이를 위한 PFNet 생성 알고리즘은 Visual Foxpro로 직접 구현하였고, 네트워크 그림을 그리기 위해서 공개용 네트워크 분석 프로그램인 Pajek(Nooy, Mirvar & Batagelj 2005)을 사용하였다.

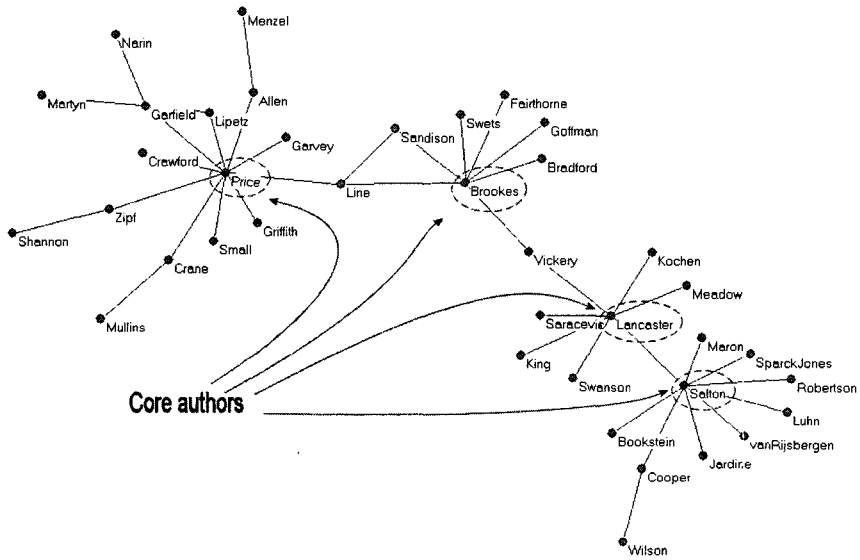
생성된 PFNet을 보면 빈도행렬을 입력한 경우인 (a)에서는 Price를 비롯한 네 명의 핵심저자가 많은 링크를 가지고 있어서 뚜렷하게 식별된다. 반면에 대각선값 결측처리후 산출된 상관계수 행렬을 입력한 경우인 (d)는 사슬 효과가 심하게 나타나서 중심(스타)이라고 할 만한 노드를 찾기가 어렵다. 더군다나 핵심저자였던 Price 등은 간선이 아닌 가지의 끝에 위치하여서 별도로 부각되지 않는다. 같은 2차 연관성이라도 대각선값을 보정한 (c)에서는 사슬 효과는 다소 있지만 원래의 핵심저자가 간선상에서 밀려나지는 않는다. 결국 행렬의 대각선값을 결측처리하지 않는다면 White(2003)의 주장과 달리 2차 연관성 행렬로도 원래의 중심저자를 어느 정도 파악할 수 있는 PFNet 생성이 가능함을 알 수 있다.

구체적으로 〈그림 15〉의 큰 주제 중에서 오른쪽의 정보검색 관련 연구자들을 살펴보면, 빈도행렬을 입력한 (a)에서는 Robertson과 Cooper가 핵심저자인 Salton을 중심으로 모여있다는 정보만 제시되지만, (b)와 (c)에서는 Salton 주위의 저자들 중에서 Robertson과 Cooper는 서로 주제적으로 관련이 깊다는 정보가 드러난다. 역시 큰 주제인 왼쪽의 계량서지학 연구자들에게서도 마찬가지 상황이 나타난다. 빈도행렬을 입력한 (a)에서는 핵심저자인 Price의 위세가 강력하기 때문에 Small과 Griffith 사이의 직접적인 관련성이 나타나지 않지만, (b)와 (c)에서는 Small과 Griffith가 직접 연결됨으로써 소주제(동시인용분석)를 나타냄을 볼 수 있다.

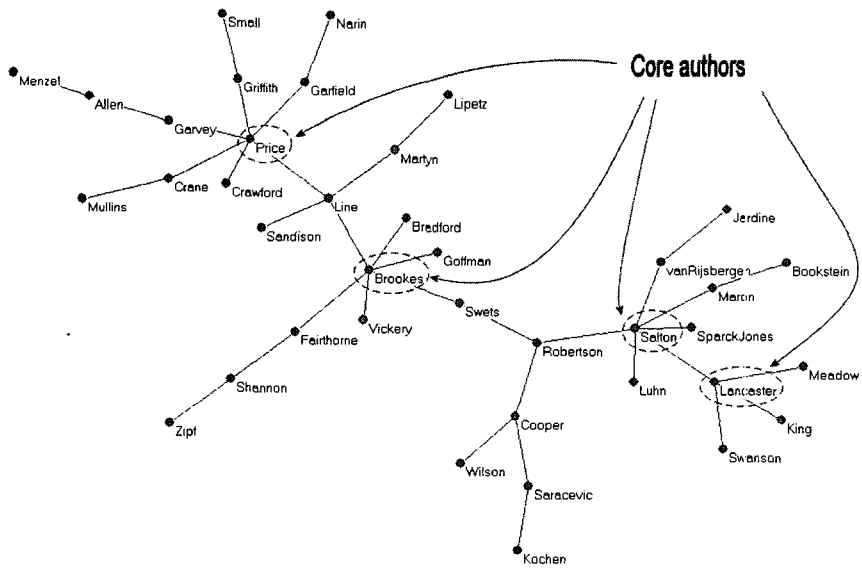
결론적으로 빈도행렬을 입력하면 생성된 네트워크에서 소수의 핵심저자가 부각되어 이들을 중심으로 하는 큰 주제를 파악할 수 있다는 장점이 있다. 반면에 비핵심저자들은 핵심저자를 중심으로 연결될 뿐이므로 비핵심저자들 사이의 관계가 감추어진다는 단점이 있다. 이와 달리 빈도값을 정규화한 행렬이나 대각선값 보정 후의

상관계수 행렬을 입력하면, 핵심저자를 중심으로 하는 큰 주제를 알게 해주는 장점은 다소 감소하지만, 비핵심저자들끼리의 관계가 드러나서 작은 주제도 파악하게 되는 효과를 얻는다.

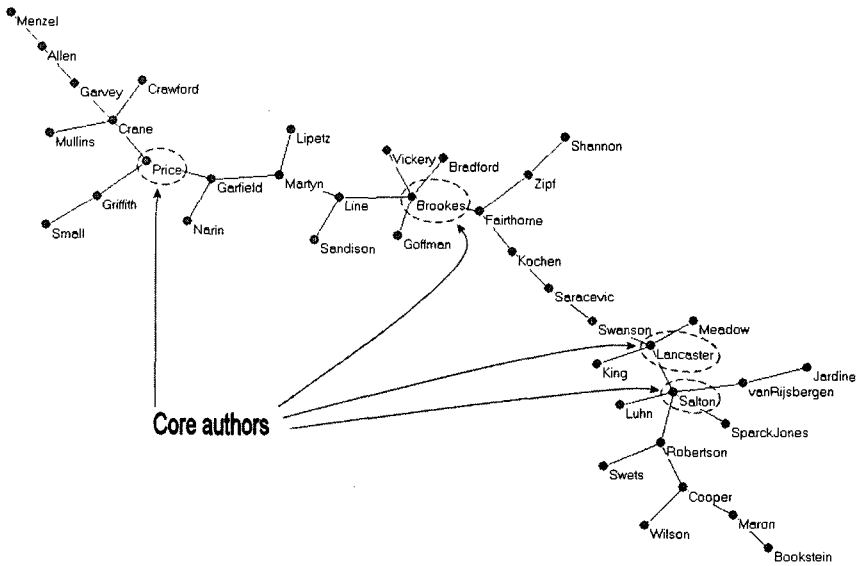
다차원척도법과 같은 기존의 지적 구조 시각화 기법에서는 행렬의 대각선값 처리와 같은 입력 자료의 가공이 결과에 큰 영향을 끼치지 않음이 보고되었으나(McCain 1990), 네트워



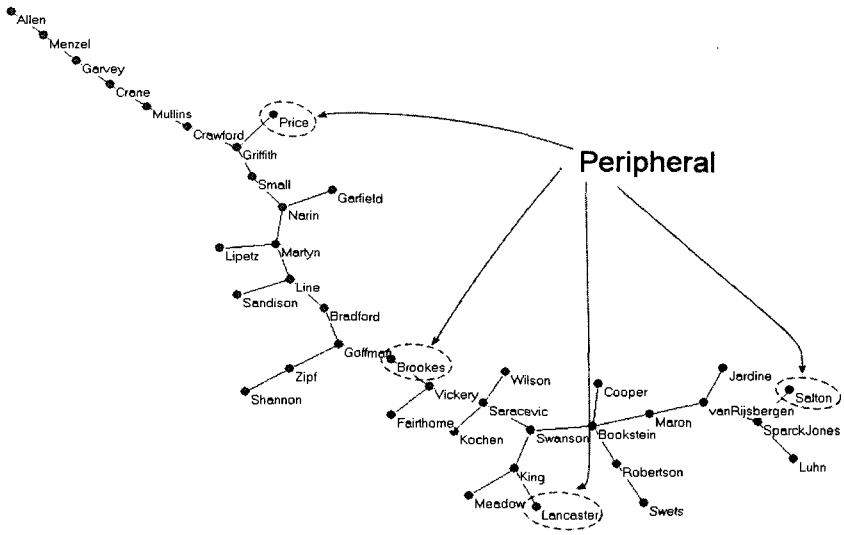
(a) 동시인용빈도행렬로 작성한 PFNet



(b) 동시인용빈도를 정규화한 코사인유사도행렬로 작성한 PFNet



(c) 동시인용빈도행렬의 대각선값 보정후 산출한 상관계수행렬로 작성한 PFNet



(d) 동시인용빈도행렬의 대각선값 결측처리후 산출한 상관계수행렬로 작성한 PFNet

〈그림 15〉 입력행렬의 가공에 따른 PFNet의 사슬 효과 비교
저자동시인용자료 출처: White & Griffith(1981)

크 형성 방식에 있어서는 계량서지 자료의 전처리 방식에 따라 핵심 노드의 입지가 뚜렷하게 달라진다는 점이 확인되었다. 따라서 네트워크 표현을 통한 지적 구조의 규명에 있어서는 분석 목적에 따라서 적절한 자료 가공 방식을 선택하여 활용하는 것이 중요하다.

4. 결론 및 전망

계량서지적 분석에서 지적 구조의 규명을 위한 시각적 표현 방식으로 오랫동안 사용되어온 다차원척도법과 군집분석을 보완 및 대체할 수 있는 네트워크 형성 기법에 대해서 살펴보았다. 이 연구에서 다룬 네 가지 네트워크 형성 방식을 종합적으로 비교하면 〈표 2〉와 같다.

네 가지 네트워크 표현 방식 중에서 기준값 절단 방식은 가장 오래된 전통적인 방식으로서 문헌동시인용분석 연구에 적용되어오다가 최근에는 공저자 분석 연구에서 주로 사용되는 것으로 나타났다. 기준값 절단 방식이 주로 적용되는 이 두 분석기법은 모두 대상이 '문헌'이나 '저자'간의 구체적인 연결을 전제로 하고 있다. 다시 말하면, 전반적인 구조보다는 '어느 문헌/저자가 중요한가?', '특정 문헌/저자와 연결강도가 강한 문헌/저자는?'과 같은 개체 단위의 역할이나 지위에 대한 탐구를 주 목적으로 하는 연구인 셈이다. 그러다보니 다차원척도법과 같이 전반적인 구조를 제시하는 것이 주 목적인 시각화 기법은 잘 사용되지 않았다.

이와 달리 저자동시인용분석이나 동시분류분석과 같은 경우는 '저자'나 '분류항목' 자체를 탐구하기 보다는, 이들이 나타내는 추상적인 개념인 '주제'에 초점을 맞추는 경향이 더 강하다. 따라서 전체 구조 표현에 어울리는 다차원척도법이 오랫동안 널리 사용되어왔다. 1990년대 후반에 들어 다차원척도법의 한계가 지적되면서 지적 구조 표현에 도입된 패스파인더 네트워크 알고리즘은, 이와 같은 전체적인 구조를 표현하는 능력이 뛰어나면서도 개별 개체의 역할이나 지위와 관련된 세부 구조를 비교적 잘 나타내주는 장점을 가지고 있다. 따라서 저자동시인용분석, 동시분류분석, 저널동시인용분석, 문헌동시인용분석 등 거의 모든 계량서

〈표 2〉 네트워크 형성 방식의 종합 비교

	기준값 절단 방식	최근접이웃 그래프	최소비용 신장트리	패스파인더 네트워크
특징	• 절대적 가중치 기준 • 가중치가 일정한 기준값을 넘는 링크만 표시함	• 지역(노드별) 상대적 가중치 기준 • 각 노드별로 가까운 k개의 노드와 연결함	• 전역 상대적 가중치 기준 • 전체 노드의 연결 거리가 최소가 되도록 함	• 전역 상대적 가중치 기준 • 삼각부등식을 위반하는 링크를 제거함
장점	• 알고리즘이 단순함 • 전체구조 파악이 가능함	• 알고리즘이 단순함 • 단독 고립노드가 발생하지 않음 • 군집이 생성됨 • 세부 구조 파악이 용이함	• 전체 구조의 파악이 용이함 • 단독 고립노드가 발생하지 않음	• 전체 구조와 세부 구조의 파악이 모두 용이함 • 단독 고립노드가 발생하지 않음
단점	• 기준값 결정 원칙이 없음 • 단독 고립노드가 발생하기도 함	• 전체 구조 파악이 쉽지 않음	• 링크 가중치가 동률인 경우에 일부 세부 구조가 드러나지 않음	• 알고리즘이 복잡함
전체 노드 연결 여부	• 기준값에 따라 다름	• 대부분의 경우 분리됨	• 항상 전체가 연결됨	• 항상 전체가 연결됨
주된 적용 분야	• 문헌동시인용분석(Small 1973), 공저자분석(Newman 2001), 이외 다수	• 사회연결망 분석에서 주로 사용하고, 계량서지 자료에 적용한 경우는 없음	• 일반적인 정보시각화에서 주로 사용하고 계량서지 자료에 적용한 경우는 Noel et al.(2002) 이외 드뭄	• 동시분류분석(McCain 1995), 저자동시인용분석(Chen 2001), 저널동시인용분석(Marion & McCain 2001), 문헌동시인용분석(Chen 2004), 이외 다수

지적 분석 영역에 패스파인더 네트워크 알고리즘이 적용되고 있다. 그러나 국내에서는 아직까지 패스파인더 네트워크를 사용하여 지적 구조를 규명하는 연구가 없는 만큼 이를 활용하는 시도가 필요할 것이다.

이와 함께 주목할 만한 네트워크 표현 방식은 최근접이웃 그래프이다. 주로 인적 네트워크를 비롯한 사회연결망 연구에서 사용되어온 기법으로서 계량서지 자료에 적용된 사례는 아직 없었다. 그러나 이 연구에서 실제 저자동시인용 자료에 적용해본 결과 단순한 알고리즘이면서도 기준값 절단 방식과 달리 단독 고립노드가 나타나지 않고 군집이 형성되는 등의 장점이 나타났다. 또한 기준값 절단 방식은 기준값의 결정이 시행착오를 통해 임의적으로 이루어질 수 밖에 없다는 단점도 있다. 따라서 최근접이웃 그래프를 지적 구조의 규명을 위한 용도로 적용해보는 후속 연구가 필요하다고 생각된다.

입력자료인 행렬의 가공방식이 네트워크 구조에 미치는 영향을 살펴본 결과, 1차 연관성과 2차연관성을 이용한 경우의 차이가 뚜렷한 것으로 나타났다. 1차 연관성, 그중에서도 빈도값을 그대로 이용한 경우에는 핵심노드 위주의 분석이 가능하였으나 비핵심노드간의 관계는 드러나지 않는 단점이 있었다. 빈도값을 정규화한 코사인계수를 이용하거나, 프로파일 상관계수로 구한 2차 연관성을 이용하면 핵심노드의 입지가 다소 약화되기는 하지만, 비핵심노드간의 관계도 얻을 수 있는 장점이 있었다. 다만 프로파일 상관계수를 구할 때 빈도행렬의 대각선 값을 결측치로 처리하는 것은 핵심노드가 부각되지 않는 사슬효과를 심화시켜서 지적 구조 규명에 도움이 안되는 것으로 나타났다. 네트워크를 형성하기 위한 계량서지 자료의 전처리가 중요하다는 점이 확인된 만큼, 필요에 따라서 적절한 자료 가공 방식을 선택하고 각 방식의 장단점에 따라 상호보완적으로 사용하는 것이 바람직하다.

이 연구에서 살펴본 패스파인더 네트워크를 비롯한 여러 네트워크 형성 기법을 그 특성에 따라 적절히 사용하는 후속 연구를 통해서 국내 계량서지적 분석 연구를 활성화할 수 있을 것이라 기대된다.

참 고 문 헌

이은숙. 2003. 『복수저자를 고려한 저자동시인용분석 연구: 정보학과 컴퓨터과학을 대상으로』. 연세대학교 석사학위논문.

『조선일보』. 2004. 17대 의원 네트워크 대해부. 8월 24일.

Börner, Katy. 2005. "Studying the emergent 'Global Brain' in large-scale co-author networks and mapping the 'Backbone of Science'." Networks and Complex Systems Talk, IUB, February 28th. [online]. [cited 2005.5.9]. 〈http://vw.indiana.edu/talks-spring05/borner.ppt〉.

Börner, K., C. Chen, and K. Boyack. 2003.

"Visualizing knowledge domains." In Blaise Cronin(Ed.), *Annual Review of Informaton Science & Technology*, 37, Medford, NJ: Information Today, Inc., chapter 5, pp.179-255.

Buzydlowski, J. W., H. D. White, and Xia Lin. 2002. "Term co-occurrence analysis as an interface for digital libraries." *Proceedings of the Visual Interfaces to Digital Libraries*, pp.133-144.

Chen, C. 1999. "Visualising semantic spaces and author co-citation networks in digital libraries." *Information Processing & Management*, 35(3): 401-420.

Chen, C. 2003. *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer.

Chen, C. 2004. "Searching for intellectual turning points: Progressive knowledge domain visualization." *Proceedings of the National Academy of Sciences*, 2004, 101: 5303-5310.

Chen, C. 2006a. *Information Visualization: Beyond the Horizon*. 2nd edition. London: Springer.

Chen, C. 2006b. "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature." *Journal of the American Society for Information Science and Technology*, 57(3): 359-377.

Chen, C., and S. Morris. 2003. "Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks." *Proceedings of the IEEE Symposium on Information Visualization (InfoVis'03)*, pp.67-74.

Ding, Y., G. G. Chowdhury, and Schubert Foo. 2001. "Bibliometric cartography of information retrieval research by using co-word analysis." *Information Processing & Management*, 37(6): 817-842.

Eppstein, David, Michael S. Paterson, and Frances F. Yao. 1997. "On nearest neighbor graphs." *Discrete & Computational Geometry*, 17(3): 263-282.

Fowler, R. H., B. A. Wilson, and W. A. L. Fowler. 1992. "Information Navigator: An information system using associative networks for display and retrieval." Technical Report NAG9-551, #92-1, Department of Computer Science, University of Texas Pan American. [online]. [cited 2005.3.19]. <http://www.cs.panam.edu/research/information_navigator.pdf>.

Gagaudakis, G., P. L. Rosin, and C. Chen. 2000. "Using CBIR and pathfinder networks for image database visualization." *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'00)*, Volume 1, pp.1052-1055.

Garfield, E. 1979. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. New York: Wiley-Inter-

science.

Glänzel, W., M. Meyer, M. du Plessis, B Thijs, T. Magerman, B. Schlemmer, K. Debackere, R. Veugelers. 2003. *Nanotechnology - An Analysis based on Publications and Patents.* Report, Steunpunt O&O Statistieken. [online]. [cited 2005.6.5]. <http://www.steunpuntoos.be/nanotech.html>.

Huang, Z., H. Chen, A. Yip, G. Ng, F. Guo, Z.-K. Chen, and M. C. Roco. 2003. "Longitudinal patent analysis for nanoscale science and engineering: Country, institution and technology field." *Journal of Nanoparticle Research*, 5(3/4): 333-363.

Kretschmer, Hildrun, and Isidro P. Aguillo. 2004. "Visibility of collaboration on the Web." *Scientometrics*, 61(3): 405-426.

Kruskal, J. B., Jr. 1956. "On the shortest spanning subtree of a graph and the traveling salesman problem." *Proceedings of the American Mathematical Society*, 7: 48-50.

McCain, K. W. 1995. "The structure of biotechnology R and D." *Scientometrics*, 32(2): 153-175.

Marion, L. S., and K. W. McCain. 2001. "Contrasting views of software engineering journals: Author cocitation choices and indexer vocabulary assignments." *Journal of the American Society for Information Science & Technology*, 52(4): 297-308.

McCain, K. W. 1990. "Mapping authors in intellectual space: A technical overview." *Journal of the American Society for Information Science*, 41(6): 433-443.

McKechnie, E. F., G. R. Goodall, D. Lajoie-Paquette, and H. Julien. 2005. "How human information behaviour researchers use each other's work: a basic citation analysis study." *Information Research*, 10(2), paper 220. [online]. [cited 2006. 1.9]. <http://InformationR.net/ir/10-2/paper220.html>.

Mukhopadhyay, R., A. Ma, and I. K. Sethi. 2004. "Pathfinder networks for content based image retrieval based on automated shape feature discovery." *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering*, pp.522-528.

Nagpaul, P. S. "Visualizing cooperation networks of elite institutions in India." *Scientometrics*, 54(2): 213-228.

Newman, M. E. J. 2001. "Scientific collaboration networks. I: Network construction and fundamental results." *Physical Review E*, 64 p.016131.

Newman, M. E. J. 2004. "Coauthorship networks and patterns of scientific collaboration." *Proceedings of the National Academy of Sciences of the United*

States of America, 101(1): 5200-5205. [online]. [cited 2005.2.8]. <http://www.pnas.org/cgi/content/full/101/suppl_1/5200>.

Noel, S., C.-H. H. Chu, and V. Raghavan. 2003. "Co-citation count vs correlation for influence network visualization." *Information Visualization*, 2(3): 160-170.

Nooy, Wooter de, Andrej Mrvar, and Vladimir Batagelj. 2005. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

Otte, Evelien, and Rousseau, Ronald. 2002. "Social network analysis: a powerful strategy, also for the information sciences." *Journal of Information Science*, 28(6): 441-453.

Prim, R. C. 1957. "Shortest connection networks and some generalizations." *Bell System Techinical Journal*, 36(1): 1389-1401.

Schvaneveldt, R. W.(ed). 1990. *Pathfinder Associative Networks: Studies in Knowledge Organization*. Norwood, NJ: Ablex.

Small, Henry. 1973. "Co-citation in the scientific literature: A new measure of the relationship between publications." *Journal of the American Society for Information Science*. 24: 265-269.

Small, Henry, and Belver C. Griffith. 1974. "The structure of scientific literatures I: Identifying and graphing specialties." *Science Studies*, 4(1): 17-40.

Tijssen, R. W., and A. F. J. van Raan. 1994. "Mapping changes in science and technology." *Evaluation Review*, 18(1): 98-115.

White, H. D. 2003. "Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists." *Journal of the American Society for Information Science & Technology*, 54(5): 423-434.

White, H. D., and B. C. Griffith. 1981. "Author cocitation: A literature measure of intellectual structure." *Journal of the American Society for Information Science*, 32(3): 163-171.

White, H. D., J. Buzydlowski, and Xia Lin. 2000. "Co-cited author maps as interfaces to digital libraries: designing Pathfinder Networks in the humanities." *Proceedings of the IEEE International Conference on Information Visualization*, 19-21 July 2000, pp.25-30.

White, H. D., and K. W. McCain. 1998. "Visualizing a discipline: An author cocitation analysis of information science, 1972-1995." *Journal of the American Society for Information Science*, 49(4): 327-355.