지적 구조 분석을 위한 새로운 클러스터링 기법에 관한 연구*

A novel clustering method for examining and analyzing the intellectual structure of a scholarly field

이 재 윤(Jae-Yun Lee)**

초 록

패스파인더 네트워크를 사용하여 지적 구조의 분석과 규명을 시도한 여러 연구가 발표되었다. 패스파인더 네트워크는 다차원척도법에 비해서 여러 장점을 가지고 있지만 구축 알고리즘의 복잡도가 매우 높아서 실행 시간이 오래 걸리며, 전통적인 지적 구조 분석에 유용하게 사용되어온 군집분석을 함께 적용하기가 어려운 것이 단점이다. 이 연구에서는 이와 같은 패스파인더 네트워크의 약점을 보완할 수 있는 새로운 기법으로 병렬 최근접 이웃 클러스터링(PNNC) 기법을 제안하였다. PNNC 기법의 클러스터링 성능을 전통적인 계층적 병합식 클러스터링 기법들과 비교해본 결과 효과성과 효율성 양면에서 기존 기법보다 우세한 것으로 확인되었다.

ABSTRACT

Recently there are many bibliometric studies attempting to utilize Pathfinder networks(PFNets) for examining and analyzing the intellectual structure of a scholarly field. Pathfinder network scaling has many advantages over traditional multidimensional scaling, including its ability to represent local details as well as global intellectual structure. However there are some limitations in PFNets including very high time complexity. And Pathfinder network scaling cannot be combined with cluster analysis, which has been combined well with traditional multidimensional scaling method. In this paper, a new method named as Parallel Nearest Neighbor Clustering (PNNC) are proposed for complementing those weak points of PFNets. Comparing the clustering performance with traditional hierarchical agglomerative clustering methods shows that PNNC is not only a complement to PFNets but also a fast and powerful clustering method for organizing informations.

키워드: 계량서지학, 지적 구조, 패스파인더 네트워크, 클러스터링, 군집분석, 병렬최근접이웃클러스터링, PNNC, Bibliometrics, Intellectual structure, Pathfinder networks, Clustering, Cluster analysis, Parallel nearest neighbor clustering

---

* 이 논문은 2006년 한국정보관리학회 추계학술발표회(2006년 11월 17일 연세대학교 새천년관)에서 발표한 것을 수정, 보완한 것임.

** 경기대학교 문헌정보학전공 조교수(memexlee@kgu.ac.kr)

1. 서론

최근 들어 계량서지적 분석에 대한 관심이 증가하고 있다. 정보학 분야의 전통있는 학술지인 *Information Processing & Management*에서는 2005년의 마지막 호를 계량정보학 특집호로 발간한 데 이어서 2006년 마지막 호를 또다시 계량정보학 특집호로 발간하는 파격적인 행보를 보였다. 각각 스무 편에 가까운 많은 논문이 실린 두 특집호의 편집은 모두 L. Egghe가 담당하였다. 그는 편집자 논설에서 이와 같이 해를 거듭하여 특집호를 발간하게 된 이유로, 인터넷을 비롯한 네트워크에 대한 양적 연구가 등장한 것을 비롯하여 계량정보학의 영역이 다학문적으로 확대되면서 성장하였기 때문이라고 언급하였다(Egghe 2005; 2006). 그는 이와 같은 계량정보학 분야의 성장이 *Journal of the American Society for Information Science and Technology*와 같은 다른 주요 학술지에도 영향을 끼치고 있다고 지적하였다.

계량서지적 분석은 학문 연구 영역으로서만 아니라 도서관 실무와 관련하여서도 주목받기 시작하고 있다. 독일 Central Library of Research Centre Jülich에 근무하는 R. Ball과 D. Tunger는 최근 *Scientometrics*에 발표한 논문에서 계량서지적 분석을 도서관의 업무로 육성하자고 주장하였다(Ball & Tunger 2006). 이들은 디지털도서관의 보편화로 인하여 사서의 전문성에 대한 도전이 더 커진 현 상황에서, 연구도서관 사서가 계량서지적 분석을 통해서 정보의 부가가치를 높여서 소속 연구 기관의 연구 투자 정책 결정과 연구개발 방향 설정, 그리고 연구 성과 평가에 기여해야 한다고 주장하였다. 이들이 계량서지적 분석이 연구도서관의 업무로서 적절하다고 본 이유는 다음과 같다(Ball & Tunger 2006).

첫째, 성과 지향적인 자금 할당이 빈번해짐으로 인해서 연구 개발 투자를 위한 정책 판단에 기여할 수 있는 믿을 만한 계량적인 데이터의 필요성이 절실해졌다. (분석 필요성의 증가)

둘째, 대부분의 학술 논문이 이제는 데이터베이스 형태로 제공되며 이들은 모두 도서관에서 다루고 있으므로 분석 대상 자료가 풍부해졌다. (분석 대상의 증가)

셋째, 연구 도서관은 분석 대상인 학술 데이터베이스와 함께 분석 지원 인력인 주제 전문 연구자를 확보하고 있다. (분석 지원 인력의 확보)

넷째, 일부 국가에는 계량서지적 분석을 주 임무로 하는 기관이 별도로 존재하기는 하지만 개별 연구소를 위한 분석까지 상시적으로 수행할 능력은 없다. (기존 역할 수행자의 부재)

여기에 한 가지 이유를 더하자면, 텍스트 마이닝과 같은 기법이 발전하여 비블리오마이닝(bibliomining)이라고 불리기까지 할 정도로 분석 기법이 발전한 것을 들 수 있

다.

결국 개별 연구 기관에 특화된 계량서지적 분석의 필요성이 증가한 상황에서 분석을 가능하게 하는 자원을 확보하고 있는 도서관이 발전된 관련 기술을 도입하여 새로운 서비스를 모 기관에 제공하는 것은, 도서관과 사서의 블루 오션 개척이라는 측면에서도 적극적으로 검토해볼 만한 일이다. R. Ball과 D. Tunger은 실제로 자신들이 근무하는 Central Library of Research Centre Jülich에서 2003년부터 기초적인 빈도분석에서부터 시작하여 성과 분석과 연구 동향 분석 서비스를 단계적으로 개발해나가고 있다. 계량서지적 분석 서비스를 도서관에 도입하자는 이런 주장은 Hjørland(2002)가 정보전문가의 역할로 11가지를 제시하면서 그중 하나로 계량서지적 분석을 언급한 것과도 일맥상통하는 것이다.

이와 같이 연구와 실무 양면에서 관심이 증가하고 있는 계량서지적 분석을 위한 기법으로는 Egghe(2005; 2006)도 언급한 것처럼 양적인 네트워크 분석이 최근 새로운 기법으로 부각되고 있다. 특히 지적 구조의 시각적인 표현을 통한 분석을 위해서는 여러 가지 네트워크 표현 기법이 도입되고 있다(이재윤 2006). 학문 분야나 연구 분야의 지적 구조를 분석하는 전통적인 연구에서는 White와 Griffith(1983)의 연구에서 도입된 다차원척도법과 군집분석을 활용하여 시각적인 표현과 분석을 수행해왔으나, 최근 들어서 인지심리학 분야에서 개발된 패스파인더 네트워크 알고리즘으로 다차원척도법을 대신하거나 보완하는 경우가 증가하고 있다. 패스파인더 네트워크 알고리즘은 다차원척도법과 마찬가지로 전체 구조를 잘 표현할 뿐 아니라 다차원척도법보다 세부 구조의 표현에 유리한 것으로 알려져 있다.

그런데 지적 구조 분석에 사용되는 패스파인더 네트워크 알고리즘은 계산복잡도가 매우 높아서 분석 대상 노드가 많은 경우에는 처리 시간이 매우 오래 걸리게 된다. 또한 패스파인더 네트워크 상에서 군집을 형성하여 주제를 파악할 수 있는 수단이 없어서 패스파인더 네트워크 알고리즘의 응용에 제약이 되고 있다. 이 때문에 패스파인더 네트워크 알고리즘을 적용한 연구에서도 다차원척도법에 의한 MDS 지도상에 군집분석 결과를 표시하는 전통적인 방식을 병행하는 경우도 적지 않다(Marion & McCain 2001; McCain et al. 2005).

이 연구에서는 이와 같은 패스파인더 네트워크 알고리즘의 약점을 보완할 수 있는 방안으로 병렬 최근접 이웃 클러스터링(Parallel Nearest Neighbor Clustering; PNNC) 기법을 새롭게 제안하고자 한다. 제안하는 기법은 지적 구조 분석을 위한 패스파인더 네트워크를 간단히 구축할 수 있으며 구축된 패스파인더 네트워크 상에서 군집을 분석할 수 있게 해준다.

2. 패스파인더 네트워크 구축을 위한 PNNC 기법

2.1 패스파인더 네트워크

패스파인더 네트워크(PFNet)는 인지심리학자인 R. W. Schvaneveldt가 1980년대에 발표한 일련의 논문과 1990년에 출판한 단행본에서 제시한 네트워크 형성 알고리즘이다. 원래는 다차원척도법처럼 심리학 연구에서 유사성 자료를 분석하기 위한 도구로 개발되었다. 따라서 다차원척도법(Multidimensional Scaling)과 마찬가지로 Scaling을 덧붙여 패스파인더 네트워크 척도법(Pathfinder Network Scaling)이라고도 불리운다(Chen 1998).

패스파인더 네트워크를 계량서지학 분야에 도입한 것은 McCain(1995)이 생물공학 분야의 특허에 대한 연구를 수행하면서 특허항목의 동시분류 분석에 적용한 것이 최초이다. 이후에는 C. Chen이 주로 지적 구조의 시각화를 위해서 패스파인더 네트워크를 적용하는 일련의 연구를 발표하였다. 특히 1998년에 K. McCain과 함께 다차원척도법과 군집분석을 이용해서 정보학분야의 저자동시인용 분석을 수행했던 H. White가 2003년에 동일한 자료에 대해서 패스파인더 네트워크를 생성하여 개선된 연구 결과(White 2003)를 발표하면서 패스파인더 네트워크의 장점이 널리 알려졌고, 이후 여러 연구에서 패스파인더 네트워크 알고리즘을 적용하고 있다(이재윤 2006).

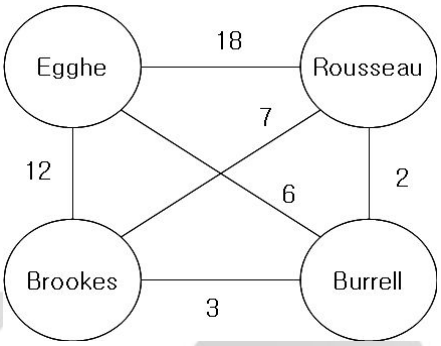
패스파인더 네트워크는 가중치가 있는 모든 링크가 생성된 상태에서 삼각부등식(triangle inequality)을 위반하는 경로를 제거하여 생성되는 네트워크이다(Schvaneveldt 1990). 삼각부등식 위반 여부를 결정하기 위해서는 두 가지 파라미터 q와 r이 필요하다. 파라미터 q는 노드 사이의 경로거리를 산출하는 데 고려하는 최대 링크의 수(거리산출범위)를 뜻한다. q는 2에서 n-1(n은 노드의 총 수)까지 설정한다. q가 2이면 한 노드를 거쳐 우회하는 경로만 고려하면 되고, q가 n-1이면 다른 모든 노드를 거쳐서 멀리 우회하는 경로까지 가능한 경우를 전부 다 고려해야 한다. q가 커질수록 조사대상 경로의 수가 늘어나서 엄격한 조건이 되므로 남는 링크의 수가 줄어든다.

파라미터 r은 민코프스키 거리 공식의 제곱수로서 한 경로를 구성하는 여러 링크가 가지고 있는 가중치를 거리에 반영하는 방법을 뜻한다. 민코프스키 거리는 경로를 구성하는 각 링크의 가중치를 r제곱하여 합한 다음 r제곱근을 취한 값이다. r이 1이면 각 링크 가중치의 합이 그대로 경로의 거리가 되고, r이 무한대이면 경로를 구성하는 링크의 가중치 중 최대값이 경로의 거리가 된다(가중치가 거리이면 최대값, 유사도이면 최소값을 취한다). r이 커질수록 경로 거리가 짧아지므로 역시 엄격한 조건(위반되

기 쉬운 조건)이 되어 남는 링크의 수가 줄어든다(이재윤 2006).

패스파인더 네트워크로 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 2006).

예를 들어 <그림 1>과 같이 네 명의 저자에 대해서 동시인용빈도가 산출되었을 때, 이 저자들 사이의 PFNet(r=∞,q=n-1)을 구하는 경우를 살펴보기로 한다.



<그림 1> 네 명의 저자로 구성된 동시인용 네트워크

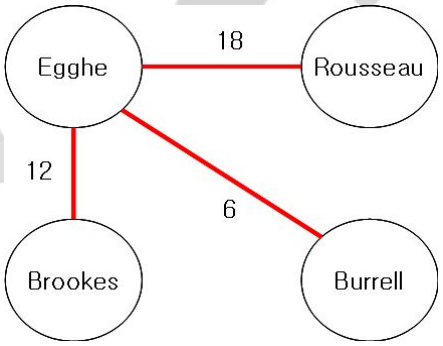
<그림 1>에서 q=n-1=3으로 설정하면 주어진 두 노드 사이의 경로로 검토해야 하는 후보 경로는 가능한 모든 경로가 된다. 예를 들어 Egghe와 Burrell 사이의 경로는 같은 링크를 반복하지 않는 경우에 총 다섯 가지가 존재한다. r=∞이고 링크 가중치가 거리가 아닌 유사도이므로 각 경로의 경로거리는 해당 경로를 구성하는 각 링크의 동시인용빈도 중 최저값이 된다(링크 가중치가 유사도가 아닌 거리이면 최대값을 취하면 된다). Egghe와 Burrell 사이의 직접 링크를 각 이름의 첫 두 글자씩 따서 Eg - Bu 라고 할 때 이 경로의 유사도는 6이 되며, 나머지 네 경로의 유사도가 다음과 같이 모두 6보다 작거나 같으므로 링크 Eg - Bu는 제거되지 않는다.

① Eg - Bu = **6**
② Eg - Ro - Bu = min(18,2) = **2**
③ Eg - Br - Bu = min(12,3) = **3**
④ Eg - Br - Ro - Bu = min(12,7,2) = **2**
⑤ Eg - Ro - Br - Bu = min(18,7,3) = **3**

한편 이보다 동시인용빈도가 7로 높은 Rousseau와 Brookes 사이의 링크 Ro - Br을 살펴보면 다른 결과가 얻어진다. 역시 둘 사이를 연결하는 경로는 직접 링크를 포함하여 다섯 가지가 되며 각각의 유사도는 다음과 같다.

① Ro - Br = **7**
② Ro - Eg - Br = min(18,12) = **12**
③ Ro - Bu - Br = min(2,3) = **2**
④ Ro - Eg - Bu - Br = min(18,12,3) = **3**
⑤ Ro - Bu - Eg - Br = min(2,6,12) = **2**

계산 결과에서 보듯이 Rousseau와 Brookes 사이의 직접 링크 ①보다 Egghe를 거치는 우회 경로 ②가 유사도 12로 더 가까운 경로가 되므로 삼각부등식을 위반한 직접 링크 ①은 제거된다. 네 노드를 연결하는 여섯 개의 링크에 대해서 이와 같은 검증을 모두 거친 결과 남는 링크는 <그림 2>와 같이 세 개가 되며 이 그림이 바로 네 저자 사이의 동시인용빈도에 의한 PFNet(r=∞,q=n-1)이다.



<그림 2> 네 명의 저자로 구성된 PFNet(r=∞,q=n-1)

계량서지적 분석에 패스파인더 네트워크 알고리즘을 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).

그러나 패스파인더 네트워크 알고리즘의 가장 큰 단점은 알고리즘의 복잡도가 너무 높아서 처리 시간이 오래 걸린다는 점이다(Rossi 2006). Schvaneveldt가 제안한 알고리즘의 복잡도는 $O(N^4)$ 등급으로 매우 높은 편에 속한다(Chen 2006). 더군다나 지적 구조 분석 연구에서는 주요 링크만 남기기 위해서 가장 계산량이 많은 경우인

PFNet(r=∞,q=n-1)을 주로 적용한다. 이와 같이 알고리즘의 높은 복잡도로 인한 느린 수행속도 때문에 수 백 개 이상의 노드에 대해서 패스파인더 네트워크를 구축하는 것은 쉽지 않다.

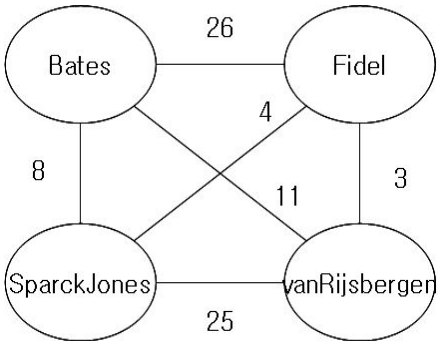
2.2 병렬 최근접 이웃 연결 알고리즘

이 절에서는 널리 사용되면서 가장 제약조건이 엄격한 패스파인더 네트워크인 PFNet(r=∞,q=n-1)이 가진 특별한 성질을 이용하여 간단하고 빠르게 구현할 수 있는 알고리즘을 제시한다.

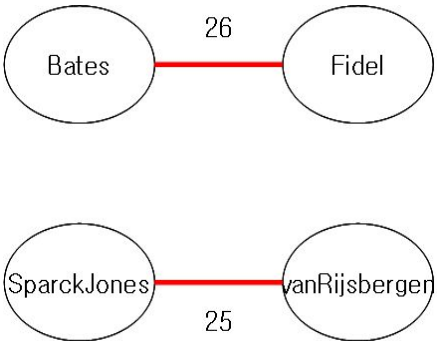
PFNet(r=∞,q=n-1)에서는 각 노드에 연결된 링크 중에서 가장 거리가 짧은(유사도가 높은) 링크에 한해서 남는 것이 보장된다. 왜냐하면 해당 노드와 연결되는 모든 경로 중에서는 최단 링크를 포함하지 않았으면서도 최단 링크보다 거리가 짧은 경로는 존재하지 않기 때문이다. 달리 표현하면, 각 노드별로 최근접 이웃과 연결하는 최단 링크만이 남는 것이 보장된다.

각 노드별로 최근접 이웃과 연결하는 방식은 네트워크 형성 방식 중에서 최근접 이웃 연결이라고 부른다(이재윤 2006). 앞의 <그림 1>에서 각 노드의 최근접 이웃을 찾으면 Egghe는 Rousseau가 가장 가깝고, 나머지 Rousseau, Burrell, Brookes는 모두 Egghe가 최근접 이웃이 된다. 따라서 각자의 최근접 이웃을 연결하는 링크만 남기면 <그림 2>의 PFNet(r=∞,q=n-1)을 얻을 수 있다.

그러나 모든 경우에 최근접 이웃 연결만으로 PFNet(r=∞,q=n-1)을 얻을 수 있는 것은 아니다. 다른 네 명의 저자 사이의 동시인용 네트워크인 <그림 3>의 Bates와 Fidel, 그리고 SparckJones와 van Rijsbergen처럼 서로 최근접 이웃인 노드쌍이 늘어나거나, 세 개 이상의 노드가 최근접 이웃 연결로 고리를 이루는 경우가 늘어나면 전체 N개의 노드가 연결되기 위해 필요한 최소한의 링크 수인 N-1에 못 미치는 링크만 생성된다. 따라서 전체가 연결된 패스파인더 네트워크가 아닌 <그림 4>와 같이 몇 개의 분할된 하위 네트워크가 생성되는 현상이 발생한다.

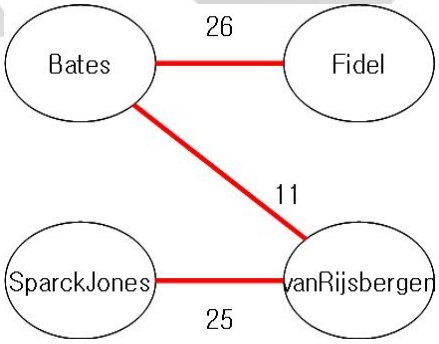


<그림 3> 최근접 이웃 쌍이
여럿인 동시인용 네트워크



<그림 4> 최근접 이웃 연결로
인해 분할된 동시인용 네트워크

이런 경우에는 분할된 각 하위 네트워크끼리의 최근접 이웃을 찾아서 연결해주면 된다. 각 노드 단위로 최단 링크가 남는 것이 보장되는 것과 마찬가지로, PFNet(r=∞,q=n-1)에서는 하위 네트워크 단위에서도 다른 하위 네트워크와의 최단 링크보다 더 짧은 경로는 존재하지 않으므로 이 링크는 남는 것이 보장된다. <그림 4>에 나타난 두 하위 네트워크 사이의 최단 링크(최고 유사도 링크)는 동시인용빈도 11인 Bates와 van Rijsbergen 사이의 링크이므로 이를 연결해주면 <그림 5>와 같은 PFNet(r=∞,q=n-1)이 생성된다.



<그림 5> 2단계 최근접 이웃
연결로 생성된 PFNet(r=∞,q=n-1)

이와 같이 2단계 최근접 이웃 연결에서도 전체 노드가 하나의 네트워크를 형성하지 않고 분할되어 있는 경우에는 다시 하위 네트워크 사이의 최근접 이웃 연결을 반복하면 최종적으로 제거할 수 없는 링크로만 이루어진 PFNet(r=∞,q=n-1)이 만들어진다.

이상의 과정은 각 노드를 최근접 이웃끼리 연결하여 군집을 형성해나가는 일종의 계층적 클러스터링 기법으로 생각할 수 있다. 이를 병렬 최근접 이웃 클러스터링(Parallel Nearest Neighbor Clustering; PNNC) 기법이라고 부르기로 하며 알고리즘으로 정리하면 다음과 같다.

① 각 노드를 모두 크기가 1인 군집으로 간주한다.
② 각 군집의 최근접 이웃 군집을 찾는다(복수 최근접 이웃 허용).
③ 최근접 이웃이 가장 가까운 군집부터 시작하여 모든 군집을 각자의 최근접 이웃 군집과 연결한다.
④ 군집이 하나로 결집되지 않았으면 ②와 ③을 반복한다.

이 알고리즘에서 각 군집별 최근접 이웃 연결을 병렬적으로 수행하지 않고 매 번 최근접 이웃을 연결할 때마다 병합된 군집의 최근접 이웃을 다시 찾으면 잘 알려진 단일연결(single linkage) 클러스터링 기법이 된다. 두 알고리즘은 최근접 이웃과 병합하는 단계를 병렬적으로 수행하느냐, 순차적으로 수행하느냐의 차이밖에 없으므로 PNNC 기법이 단일연결 기법만큼 단순한 알고리즘이라는 것을 알 수 있다. 이와 같이 PNNC 알고리즘은 단일연결 클러스터링 알고리즘과 마찬가지로 단순한 알고리즘이면서도, 지적 구조 분석에 가장 널리 사용되는 패스파인더 네트워크인 PFNet(r=∞,q=n-1)을 간단하고 빠르게 구현할 수 있는 새로운 방식이다.

3. PNNC 기법을 이용한 패스파인더 네트워크의 군집 분석

다차원척도법을 적용한 지적 구조 분석에서는 거의 예외없이 군집분석 결과를 2차원 MDS 지도에 나타내왔다. 지적 지도 상에서 군집은 해당 영역 안에서 세부 주제분야를 파악할 수 있는 유용한 수단이 된다(White 1990).

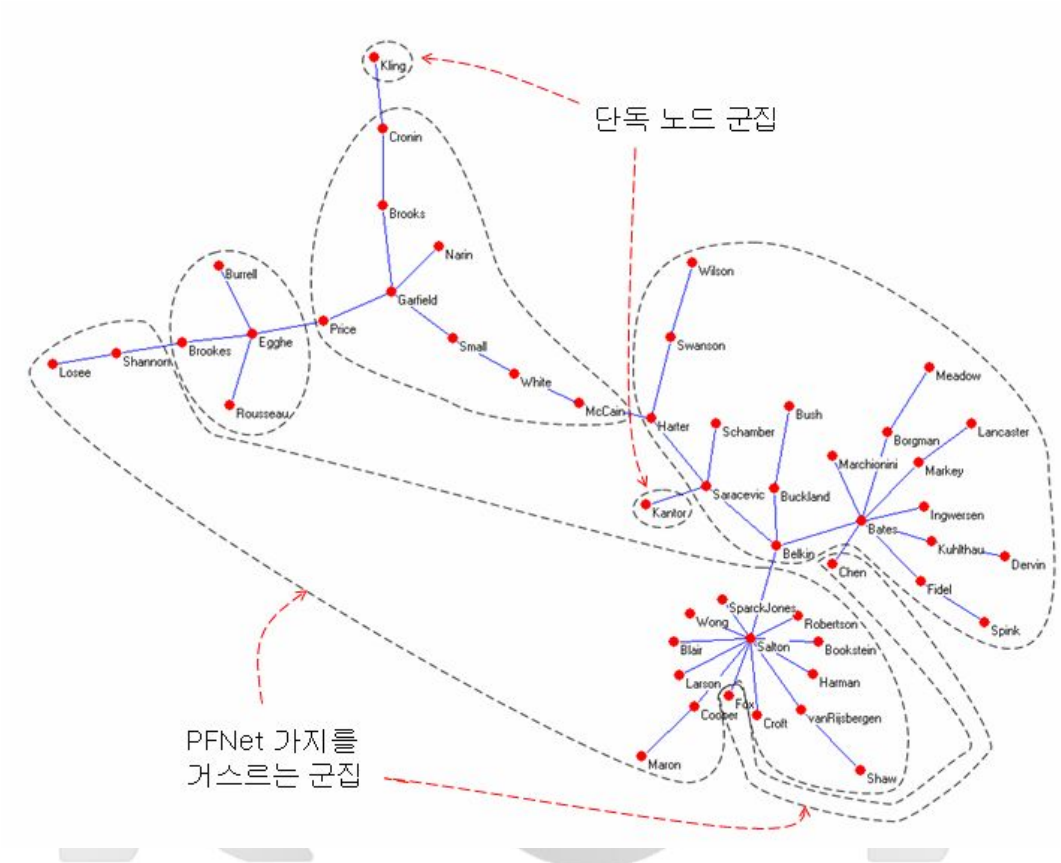
패스파인더 네트워크 상에서도 군집을 식별할 수 있으면 세부 주제 분석에 많은 도움이 된다. 군집이 아닌 소수 핵심 노드 중심의 분석에서는 개별 노드가 가진 주제가 다면적일 수 있으므로 해석이 용이하지 않다. 하지만 노드군을 단위로 분석하면 군집에 속한 여러 노드의 공통 주제를 추출하여 세부 주제를 분명하게 파악하게 되므로 주제분석의 식별력을 높일 수 있다. 또한 생성된 패스파인더 네트워크의 각 링크가 대등한 가중치를 가지는 것이 아니므로 상대적으로 가중치가 약한 링크를 경계로 하는 군집을 생성할 수 있으면 세부 주제의 범위와 특정 노드의 특정 주제 소속 여부를

결정하는 데도 도움이 된다. 이는 지적 구조를 통시적으로 분석하여 시기에 따른 변화를 파악할 때 특히 유용할 것이다.

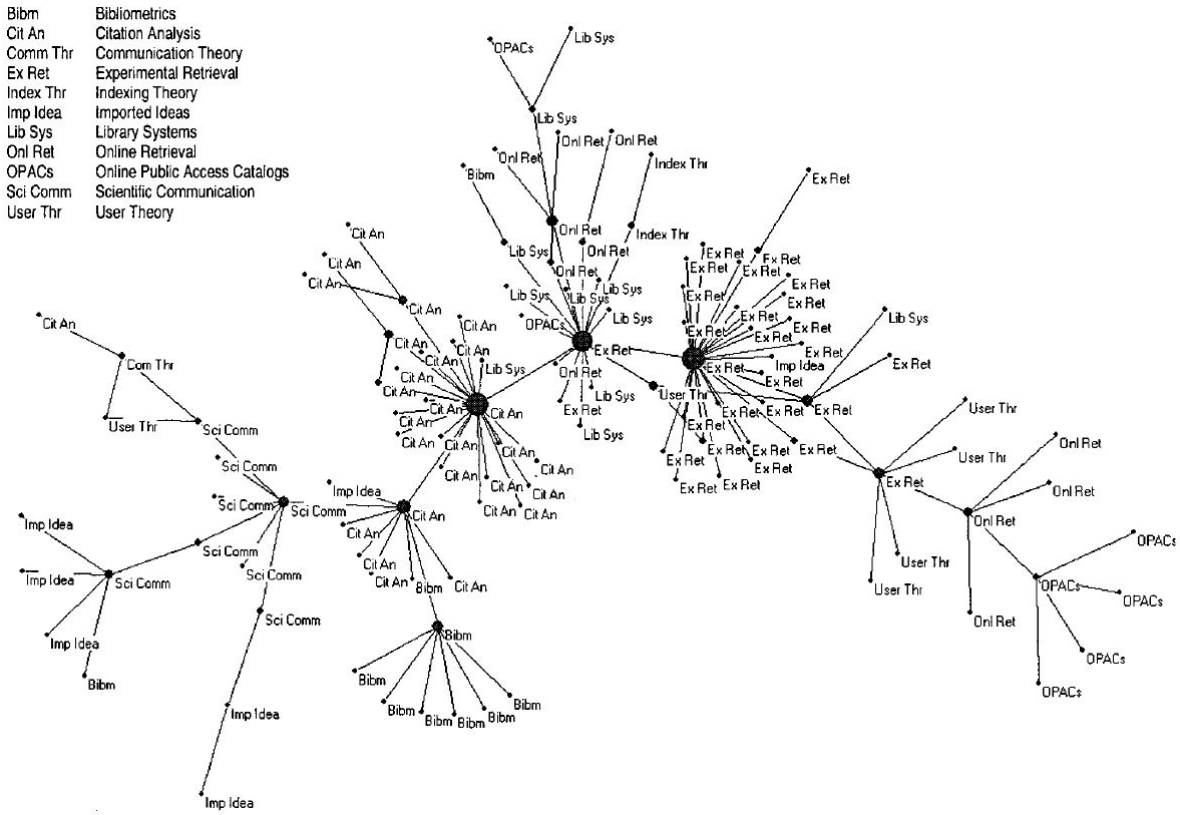
그러나 패스파인더 네트워크를 사용하여 지적 구조를 분석하는 기존 연구에서는 패스파인더 네트워크 상에 군집을 표시하지 못했다. 그 이유는 MDS 지도 상에서는 노드간의 연결 정보가 없으므로 다소 복잡하더라도 군집을 그릴 수 있는 반면에, 이미 노드간 연결 정보가 제시되어 있는 패스파인더 네트워크에서는 링크를 무시하고 노드를 군집으로 묶는 것이 매우 부자연스럽기 때문이다. 이런 이유로 McCain 등(2005)은 소프트웨어 공학 분야의 저자동시인용분석을 수행하면서 패스파인더 네트워크는 스타 노드라고 할 수 있는 핵심 저자를 파악하는 용도로 사용하였고, 세부 주제의 분포와 각 저자의 소속 주제 파악을 위해서는 MDS 지도와 군집분석을 사용하였다. Chen(2006)도 네트워크 표현과 군집분석이 잘 결합되지 않는 것을 해결해야 될 과제 중 하나로 언급하고 있다.

이은숙과 정영미(2002)의 정보학 분야 저자동시인용 자료를 대상으로 패스파인더 네트워크에서 군집 분석의 어려움을 알아보기로 한다. 이은숙과 정영미(2002)는 *Journal of the American Society for Information Science*에 1990년부터 2000년까지 11년간 수록된 논문에서 14회 이상 인용된 50명을 대상으로 동시인용 정보를 산출하였다. 본 연구에서는 이 동시인용행렬을 코사인계수로 정규화하고 이로부터 PFNet(r=∞,q=n-1)을 생성한 다음, 여기에 평균연결기법으로 생성한 군집을 <그림 6>과 같이 그려보았다. 이 그림에서 보듯이 기존의 군집분석 기법은 패스파인더 네트워크의 연결 가지를 거스르는 군집을 형성하게 되므로 주제 분포를 파악하기가 곤란하다. 또한 <그림 6>의 Kling이나 Kantor처럼 단일 노드로 구성된 군집이 나타나는 경우도 주제 해석에 도움이 안된다. 이런 문제는 평균연결기법이 아닌 단일연결, 완전연결, Ward 기법 모두에 해당된다.

이런 이유 때문에 White(2003)는 주성분분석(Principle Component Analysis) 결과를 패스파인더 네트워크에 반영하여 <그림 7>과 같이 주제 분포를 파악하고자 하였다. 이 그림에서 어느 정도는 주제 분포가 드러나지만, 이런 결과를 얻기 위해서는 121명의 분석 대상 저자 중에서 80명이 둘 이상의 요인에 적재되었으므로 이런 복수 주제에 해당하는 저자의 주제 판정은 해석에 유리한 쪽으로 임의적으로 판정하였다. 이와 같은 자의적인 판단을 동원했음에도 불구하고 <그림 7>에서 온라인 검색(Onl Ret) 주제와 같이 동일 주제에 속한 저자가 한 지역에 모이지 않고 흩어지는 경우가 종종 나타난다. 따라서 주성분분석으로 패스파인더 네트워크 위에 주제 영역을 표시하는 것은 평균연결 기법으로 주제 군집을 표시하는 것과 마찬가지로 불완전하다는 것을 알 수 있다.



<그림 6> 패스파인더 네트워크에 평균연결기법에 의한 군집을 표시한 결과



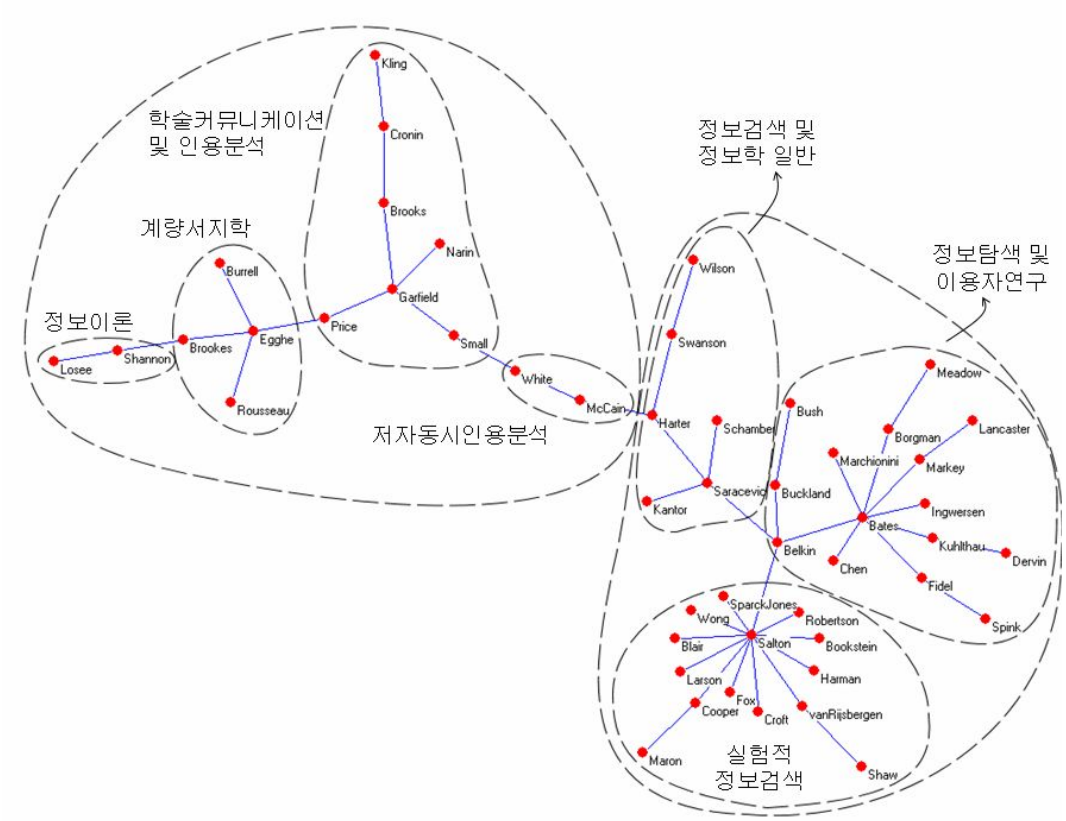
<그림 7> 정보학 분야 저자 121명의 PFNet(r=∞,q=n-1)에 표시한 주성분

출처: White(2003)

앞에서 제안한 PNNC 기법은 본질적으로 계층적 클러스터링 기법에 해당하므로 이를 통해 생성되는 패스파인더 네트워크를 여러 개의 군집으로 분할하는데 사용할 수 있다. 즉, 패스파인더 네트워크와 군집분석을 결합하여 지적 구조를 파악하는 것이 가능하다.

이은숙과 정영미(2002)의 저자동시인용 자료에 대해서 PNNC 기법을 적용하면 <그림 8>과 같이 2단계 만에 PFNet(r=∞,q=n-1)이 생성된다. 이 그림에서는 군집이 패스파인더 네트워크의 줄기를 거스르지 않고 자연스럽게 형성되는 것을 볼 수 있다. 또한 정보학 영역을 크게는 두 개의 군집으로, 작게는 일곱 개의 군집으로 나눈 결과를 얻었다. 소속된 저자를 살펴보면 크게 분리된 두 군집 중에서 왼쪽은 계량정보학 및 정보 이론 영역에 해당하고, 오른 쪽은 정보검색 및 정보시스템 영역에 해당함을 알 수 있다. 이와 같이 PNNC 기법을 이용하여 패스파인더 네트워크를 형성하면 자연스럽게 군집분석을 병행할 수 있다.

PNNC 기법을 이용한 군집분석이 가지는 또 다른 장점은 <그림 8>에서처럼 군집의 수가 자동적으로 결정된다는 점이다. 앞에서 제시한 병렬 최근접 이웃 클러스터링 알고리듬에서 ②단계와 ③단계를 반복할 때마다 각 군집이 최근접 이웃과 결합하여 적절한 수의 군집을 형성한다. 전통적인 군집분석에서 군집의 수를 결정하는 기준이 다소 자의적이었던 점을 감안하면 PNNC 기법이 반복 처리 단계마다 군집의 수를 자동으로 산출해주는 것은 연구자의 고민거리를 덜어줄 수 있는 특성이다. 물론 필요한 경우에는 PNNC 기법에서도 원하는 수의 군집을 생성하고 중단할 수 있다.



<그림 8> 패스파인더 네트워크에 PNNC 기법에 의한 군집을 단계별로 표시한 결과

인용빈도 자료 출처: 이은숙 & 정영미(2002)

4. 클러스터링 성능의 검증

PNNC 기법이 비록 패스파인더 네트워크의 간편한 구축과 군집 파악을 동시에 가능하게 하는 장점을 가지고 있지만, 클러스터링 기법으로서의 성능이 기존 기법에 미치지 못한다면 지적 구조의 분석에 반드시 유용하다고만 할 수는 없다. 전통적으로 저자동시인용 등의 계량서지적 분석에서는 군집을 파악하기 위해서 평균연결 기법이나 Ward 기법을 사용해왔다. 이들 전통적인 기법으로 생성한 군집이 더 지적 구조를 잘 나타낸다면 패스파인터 네트워크 상에서의 분석 이외에 별도로 군집 분석을 수행하는 것도 검토해야 한다.

여기서는 순수하게 클러스터링 기법으로서 PNNC의 성능을 전통적인 계층적 클러스터링 기법인 단일연결(SLINK), 완전연결(CLINK), 평균연결(ALINK), 와드(Ward) 기법과 비교해보기로 한다.

클러스터링 성능 비교 실험을 위해서 <표 1>과 같은 여섯 가지 실험집단을 이용하였다. 이중에서 CQS49-11C와 HQS104-11C는 정보검색용 실험집단인 CACM과 HANTEC 1.0에 포함된 11개 질의에 대한 적합문서로 구성한 것이며(이재윤, 정진아 2005), KTSET195-6C와 TREND193-8C는 분류정보가 포함된 실험집단인 KTSET 1.0의 논문초록과 HANTEC 1.0의 과학기술문헌속보 중에서 각각 6개 주제와 8개 주제에 속하는 문서들로 구성하였다. Web32-6C와 Web36-8C는 커뮤니케이션 분야 웹 사이트의 동시링크 분석 연구인 정동렬과 최은미(1999), 그리고 이성숙(2005)의 연구에서 다룬 커뮤니케이션 분야 웹 사이트 32개와 36개로 각각 구성된 집합이다. 웹 사이트의 소속 군집은 두 연구에서 연구자들이 지정한 각 웹 사이트의 주제를 약간 보완하여 결정하였다.

<표 1> 클러스터링 성능 실험을 위한 실험집단

실험집단	노드 수	군집 수	내 용	유사도 산출
CQS49-11C	49	11	영어 논문 초록	문서 벡터간 코사인
HQS104-11C	104	11	한국어 논문 초록/ 기사 원문	문서 벡터간 코사인
KTSET195-6C	195	6	한국어 논문 초록	문서 벡터간 코사인
TREND193-8C	193	8	한국어 기사 원문	문서 벡터간 코사인
Web32-6C	32	6	웹 사이트	동시링크
Web36-8C	36	8	웹 사이트	동시링크

클러스터링 성능 평가는 평가 척도마다 다소 달라질 수 있으므로 다양한 척도에 의한 평가 결과를 망라하여 살펴보기 위해서 주요한 다섯 가지 평가 척도를 적용해보았

다. 사용한 척도는 CSIM, WACS, F_1, MI(상호정보량)를 상한값을 이용해 0에서 1사이로 정규화한 NMI, 카이제곱정보량을 상한값을 이용해 0에서 1 사이로 정규화한 NCHI이다. 각 평가 척도에 대한 상세한 설명은 정영미, 이재윤(2001)에서 다루고 있으며 이 논문에서는 생략하였다. 군집의 수는 각 실험집단의 원래 군집 수에 맞추어 생성한 결과를 대상으로 성능을 평가해보았다.

여섯 가지 실험집단에 대해서 PNNC를 비롯한 다섯 가지 클러스터링 기법을 적용한 결과를 다섯 가지 척도로 평가한 결과는 <표 2>와 같다. <표 2>에서 각 실험집단마다 성능이 가장 좋은 기법으로 꼽힌 횟수를 살펴보면 이 연구에서 제안한 PNNC가 17회로 가장 많은 것으로 나타났다. 이외에는 ALINK와 WARD 기법이 4회로 그 다음이었다.

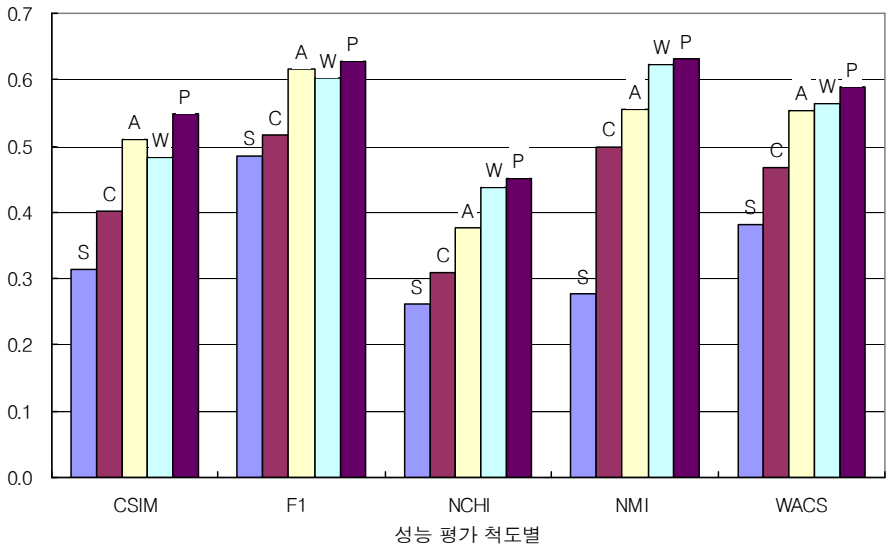
또한 6가지 실험집단에 대한 성능을 평가척도별로 평균해서 산출한 <표 3>과 <그림 9>를 보면 차이가 크지는 않으나 PNNC가 평가척도를 불문하고 가장 평균성능이 좋음을 알 수 있다. CSIM과 F_1 척도로는 ALINK 기법의 평균성능이 2등, WARD 기법의 평균성능이 3등이었으며, 나머지 세 척도로는 WARD 기법이 2등, ALINK 기법이 3등으로 나타났다. 이런 결과를 통해서 이 연구에서 제안한 PNNC 기법에 의해 생성된 군집이 기존의 클러스터링 기법으로 생성된 것에 비해서 품질이 떨어지지 않거나 오히려 더 좋음을 알 수 있다.

<표 2> 5가지 클러스터링 기법을 6개 실험집단에 적용한 결과
(반전된 부분은 5가지 기법 중 최고성능인 경우)

성능척도	실험집단	SLINK	CLINK	ALINK	WARD	PNNC
CSIM	CQS49-11C	0.2122	0.3178	0.4750	0.3721	0.3644
	HQS104-11C	0.2950	0.5152	0.7380	0.6169	0.7784
	KTSET195-6C	0.3035	0.2802	0.3466	0.3486	0.4057
	TREND193-8C	0.2922	0.3631	0.5979	0.6129	0.6443
	Web32-6C	0.4946	0.4275	0.4730	0.4040	0.4722
	Web36-8C	0.2916	0.5081	0.4303	0.5488	0.6264
F_1	CQS49-11C	0.4931	0.4922	0.6020	0.5419	0.5266
	HQS104-11C	0.5394	0.6502	0.7905	0.7595	0.8013
	KTSET195-6C	0.3391	0.3391	0.4202	0.4353	0.4684
	TREND193-8C	0.3327	0.3744	0.6405	0.6248	0.6196
	Web32-6C	0.6437	0.5881	0.6140	0.5608	0.6153
	Web36-8C	0.5606	0.6612	0.6361	0.7008	0.7407
NCHI	CQS49-11C	0.2902	0.1915	0.2877	0.2975	0.2729
	HQS104-11C	0.4793	0.4273	0.6384	0.4390	0.6548
	KTSET195-6C	0.0164	0.0438	0.1369	0.3011	0.3301
	TREND193-8C	0.0200	0.1244	0.3427	0.4527	0.3637
	Web32-6C	0.3325	0.5136	0.3686	0.4497	0.3783
	Web36-8C	0.4358	0.5579	0.4854	0.6942	0.7136
NMI	CQS49-11C	0.3558	0.6254	0.6746	0.6539	0.6319
	HQS104-11C	0.3750	0.7069	0.8298	0.7939	0.8472
	KTSET195-6C	0.0214	0.0706	0.2049	0.3330	0.3896
	TREND193-8C	0.0340	0.4071	0.5450	0.6635	0.5906
	Web32-6C	0.4796	0.4790	0.4769	0.5241	0.5433
	Web36-8C	0.4042	0.6997	0.6071	0.7695	0.7886
WACS	CQS49-11C	0.3340	0.4567	0.5657	0.5057	0.4945
	HQS104-11C	0.4064	0.5786	0.7539	0.6981	0.7834
	KTSET195-6C	0.3004	0.2632	0.3538	0.3868	0.4347
	TREND193-8C	0.2840	0.3753	0.5851	0.6007	0.5871
	Web32-6C	0.5363	0.5291	0.5306	0.5177	0.5266
	Web36-8C	0.4249	0.6077	0.5329	0.6728	0.7155

<표 3> 각 클러스터링 기법의 평균 성능 비교

기법 / 평가척도	SLINK	CLINK	ALINK	WARD	PNNC
CSIM	0.3148	0.4020	0.5101	0.4839	0.5486
F_1	0.4847	0.5175	0.6172	0.6038	0.6286
NCHI	0.2623	0.3098	0.3766	0.4390	0.4522
NMI	0.2783	0.4981	0.5564	0.6230	0.6319
WACS	0.3810	0.4684	0.5537	0.5636	0.5903



<그림 9> 각 클러스터링 기법의 평균 성능 비교
(S=SLINK, C=CLINK, A=ALINK, W=WARD, P=PNNC)

이런 성능 차이가 통계적으로 유의한 지 여부를 알아보기 위해서 비모수검증방법인 Wilcoxon 부호순위검증을 수행하였다. 각 척도별로 PNNC 기법과 기존 클러스터링 기법의 성능 차이를 검증한 결과는 <표 4>와 같다.

<표 4> 클러스터링 성능 차이의 Wilcoxon 부호순위검증 결과

	PNNC				
	F_1 기준	WACS 기준	CSIM 기준	NMI 기준	NCHI 기준
SLINK	<<	<<	<<	<<	<<
CLINK	<<	<<	<<	<<	<
ALINK				<	<
WARD			<<		

* 부등호 두 개(<<)는 PNNC 기법이 95% 유의수준에서 우세한 경우를 뜻하고, 부등호 한 개(<)는 90% 유의수준에서 우세한 경우를 뜻함.

<표 4>에서는 95% 유의수준에서 성능 차이가 있으면 부등호 두 개, 90% 유의 수준에서 차이가 있으면 부등호 한 개로 표시하고, 90% 유의수준에서도 차이가 유의하지 않으면 부등호를 표시하지 않았다. 따라서 부등호가 많을수록 PNNC 기법의 성능 우세가 통계적으로 더 유의하다는 것을 뜻한다.

검증 결과를 보면 PNNC 기법이 SLINK 기법보다는 모든 척도에서, 그리고 CLINK 기법보다는 NCHI를 제외한 나머지 네 척도에서 95% 이상 유의수준에서 우세하게 나타났다. 즉, 이 두 기법보다는 PNNC 기법이 월등하게 우세함이 확인되었다. 한편 WARD 기법보다는 CSIM척도 기준일 때 95% 유의수준에서 PNNC 기법이 우세하게 나타났고, ALINK 기법과 비교하였을 때에는 NMI 척도와 NCHI척도를 기준으로 하였을 때 90% 유의수준에서 PNNC 기법이 우세하다고 말할 수 있는 것으로 나타났다. 결과적으로 다른 기법과 비교해서 PNNC 기법이 최소한 한 가지 이상 평가 척도로는 통계적으로 유의하게 성능이 우월하다는 결과를 얻었다.

5. 결론

이 연구에서는 최근 학문연구와 도서관실무 양쪽에서 관심이 증가하고 있는 계량서지적 분석을 위한 새로운 클러스터링 기법으로 병렬 최근접 이웃 클러스터링(PNNC) 기법을 제안하였다. 제안한 기법은 네트워크의 시각적 표현 기법으로 주목을 받고 있는 패스파인더 네트워크의 특성에서 착안한 것이다. 제안된 PNNC 기법의 특징을 정리하면 다음과 같다.

첫째, 패스파인더 네트워크 중에서 지적 구조 분석에 흔히 사용되는 가장 엄격한 경우인 PFNet(r=∞,q=n-1)을 신속하게 구축할 수 있는 간단한 알고리즘이다.

둘째, PFNet(r=∞,q=n-1) 상에서 네트워크의 줄기에 어울리도록 군집을 생성해주므로 주제 영역의 파악에 유리한 알고리즘이다.

셋째, 분석 대상 자료의 유사도 분포에 따라서 군집의 수를 자동으로 결정해주며, 자료에 따라서는 여러 단계로 적절한 군집의 수가 결정되는 알고리즘이다.

넷째, 기존의 계층적 군집분석 기법 중에서 단일연결 기법과 마찬가지로 단순하면서 실행 효율이 높은 알고리즘이다.

다섯째, 기존의 계층적 군집분석 기법과 비교할 때 성능이 대등하거나 더 좋은 알고리즘이다.

향후에는 다양한 계량서지적 자료에 대해서 PNNC 기법을 적용하는 실험을 통해서 타당성을 검증해보고자 한다. 이를 통해 지적 구조 분석 연구에서 네트워크 표현 및 분석 기법의 도입을 더욱 활성화할 수 있을 것으로 기대된다. 또한 제안된 기법이 보편적인 클러스터링 기법으로서도 효율성과 효과성 양면에서 기존의 계층적 클러스터링 기법을 대신할 가능성이 있는 만큼, 정보검색을 비롯한 여러 영역에 응용해보는 시도도 모색할 계획이다.

참고문헌

이성숙. 2005. 동시링크분석을 이용한 웹정보원의 지적구조 변화에 관한 연구. 『정보관리학회지』, 22(2): 205~228.

이은숙, 정영미. 2002. 복수저자를 고려한 동시인용분석 연구: 정보학과 컴퓨터과학을 대상으로. 『지식처리연구』, 3(2): 1-26.

이재윤. 2006. 지적 구조의 규명을 위한 네트워크 형성 방식에 관한 연구. 『한국문헌정보학회지』, 40(2): 333-355.

이재윤, 정진아. 2005. 계층적 문서 클러스터링을 위한 응집식 기법과 분할식 기법의 비교 연구. 『제12회 한국정보관리학회 학술대회 발표논문집』, pp. 65-70.

정동렬, 최윤미. 1999. 웹 정보원의 동시인용분석에 관한 실험적 연구. 『정보관리학회지』, 16(2): 7-25.

정영미, 이재윤. 2001. 클러스터링 성능 평가를 위한 비편향적 척도의 개발. 『제 8회 한국정보관리학회 학술대회 논문집』, pp. 167-172.

Ball, Rafael, and Dirk Tunger. 2006. "Bibliometric analysis - A new business

area for information professionals in libraries?: Support for scientific research by perception and trend analysis." *Scientometrics*, 66(3): 561-577.

Bollen, Johan, Herbert Van de Sompel, Joan A. Smith, and Rick Luce. 2005. "Toward alternative metrics of journal impact: A comparison of download and citation data." *Information Processing & Management*, 41(6): 1419-1440.

Börner, K., C. Chen, and K. Boyack. 2003. "Visualizing knowledge domains." In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, 37, Medford, NJ: Information Today, Inc., chapter 5, pp. 179-255.

Chen, C. 1998. "Generalised similarity analysis and pathfinder network scaling". *Interacting with Computers*, 10(2): 107-128.

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

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

Egghe, L. 2005. "Expansion of the field of informetrics: Origins and consequences." *Information Processing & Management*, 41(6): 1311-1316.

Egghe, L. 2006. "Expansion of the field of informetrics: The second special issue." *Information Processing & Management*, 42(6): 1405-1407.

Hjørland, Birger. 2002. "Domain analysis in information science: Eleven approaches - traditional as well as innovative." *Journal of Documentation*, 58(4): 422-462.

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

McCain, Katherine W., June M. Verner, Gregory W. Hislop, William Evanco, and Vera Cole. 2005. "The use of bibliometric and knowledge elicitation techniques to map a knowledge domain: Software engineering in the 1990s." *Scientometrics*, 65(1): 131-144.

Rossi, Fabrice. 2006. "Visualization methods for metric studies". In

Proceedings of the International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting, Nancy (France). [online]. [cited 2006.9.5]. <http://eprints.rclis.org/archive/00006047/>.

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

White, H. D. 1990. "Author co-citation analysis: Overview and defense". In C. L. Borgman (Ed.), *Scholarly Communication and Bibliometrics* (pp. 84-106). Newbury Park, CA: Sage.

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: 163-171.