지적 구조 분석을 위한 군집분석과 다차원척도법의 결합 방안

An Enhanced Multidimensional Scaling Technique Combined with Clustering Results for Knowledge Domain Analysis

이재윤, 경기대학교, memexlee@kgu.ac.kr

Jae Yun Lee, Kyonggi University

연구동향 분석이나 연구영역 분석에서 널리 사용되고 있는 다차원척도법은 표현할 개체의 수가 많을 경우에 군집분석 결과와 잘 결합되지 못하는 단점이 있다. 이를 해결하기 위해서 군집분석과 다차원척도법을 결합하는 새로운 방법을 제안하고 실제 사례에 적용해보았다.

1. 서론

최근 지적 구조 분석이나 연구동향 분석이 활성화되면서 다양한 영역에서 관련 분석이 이루어지고 있다. 특히 지적 구조를 시각화하는 방법으로는 전통적으로 다차원척도법과 군집분석이 널리 활용되고 있으며, 이 두 가지 방법을 결합하여 2차원 지도를 표현하는 것이 일반적이다.

그런데 다차원척도법(MDS)으로는 국지적인 세부 구조가 부정확하게 표현되는 단점이 있다(Börner et al. 2003; Chen 2006). 개체들 사이의 복잡한 고차원 관계를 2차원의 평면에 표현하면서 일부 정보가 피치못하게 손실되기 때문이다.

따라서 군집분석 결과를 MDS 지도에 영역으로 표시할 때 구불구불하게 그려지거나 심지어는 군집끼리 겹쳐지는 현상이 나타난다. 특히 지도에 표현할 저자나 웹 사이트가 4~50개 이상으로 많으면 이런 현상이 흔히 나타나며, 군집의 영역이 서로 배타적으로 표시되지 않기 때문에 지적 구조의 파악이 제대로 이루어지지 않게 된다.

따라서 이 연구에서는 국지적 구조를 표현하는 능력이 약한 다차원척도법의 한계를 극복하고, 변수의 수가 많을 경우에도 군집분석과 잘 조화되는 개선된 다차원척도법을 개발하고자 한다.

2. 군집분석을 결합한 다차원척도법

이 연구에서 제안하는 클러스터링과 결합한 다차원척도법인 CMDS(Cluster-enhanced MDS)는 군집분석을 수행하여 얻은 군집의 MDS 지도를 먼저 만든 다음, 이 지도의 각 군집의 좌표를 참조하여 개별 노드의 좌표를 결정하는 방식이다. 구체적인 절차는 다음과 같다.

① 군집분석 수행
② 빈도행렬 cosine 정규화
③ 군집간 평균 cosine 산출
④ 평균 cosine행렬에서 군집MDS생성(z점수 유클리드거리)
⑤ 군집과 노드간 cosine 산출
⑥ 군집좌표를 근거로 노드 좌표 산출

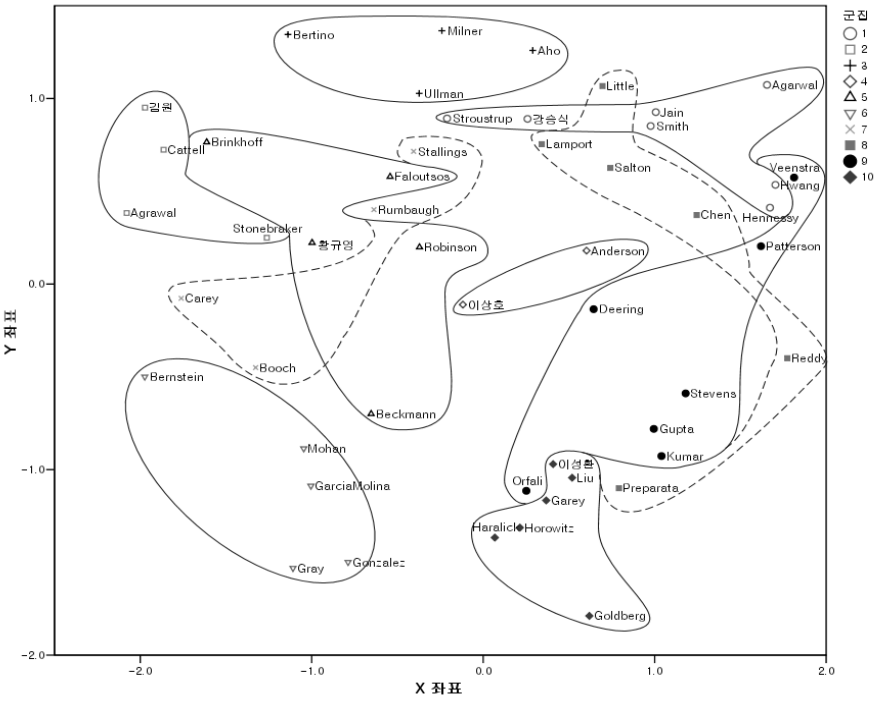
위와 같이 CMDS 기법은 입력 데이터로 주어진 근접성 행렬에 대해서 군집분석을 먼저 수행한 후, 생성된 군집을 단위로 MDS를 수행하여 각 군집의 MDS 좌표를 획득한다. 그 후 각 노드의 좌표는 각 노드와 군집 사이의 유사도에 비례하여 군집의 좌표를 가중평균하여 산출하게 된다.

이와 같이 군집의 좌표로 MDS 지도를 생성한 이후에 개별 노드의 MDS 지도를 만드는 방식은 Noyons 등(1998)의 다단계 매핑 방식과 유사하다. 그러나 다단계 매핑 방식에서는 군집의 MDS 지도를 만든 이후에 개별 군집마다 별개의 MDS 지도를 만들었으므로 전체 지적 구조를 한 번에 파악하기가 어려웠다.

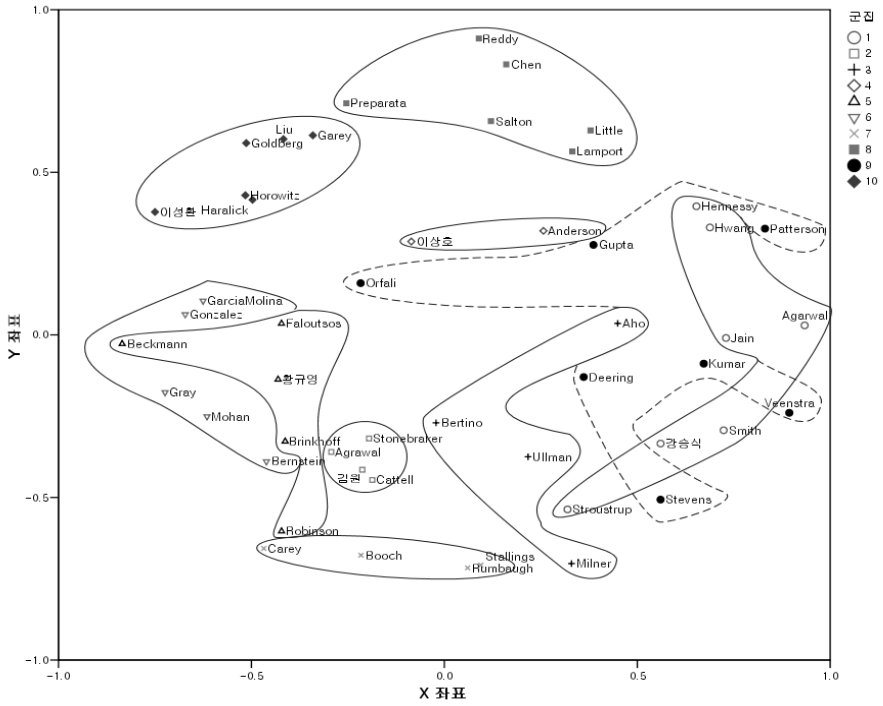
3. CMDS 기법의 적용 결과

선행 연구 중에서 컴퓨터과학 분야 저자 50명의 저자동시인용 분석을 수행한 이은숙(2002)의 연구에서 사용된 저자동시인용 데이터를 시험 대상으로 CMDS 기법을 시험 적용해보았다. 전통적인 다차원척도법인 ALSCAL과 PROXSCAL로 지적 구조를 표현하고 군집을 나타낸 결과는 각각 <그림 1>, <그림 2>와 같다. 이 그림을 보면 군집 간의 경계가 서로를 넘나들면서 복잡하게 얽히게 표현되어서 지적 구조를 제대로 파악하기가 쉽지 않다.

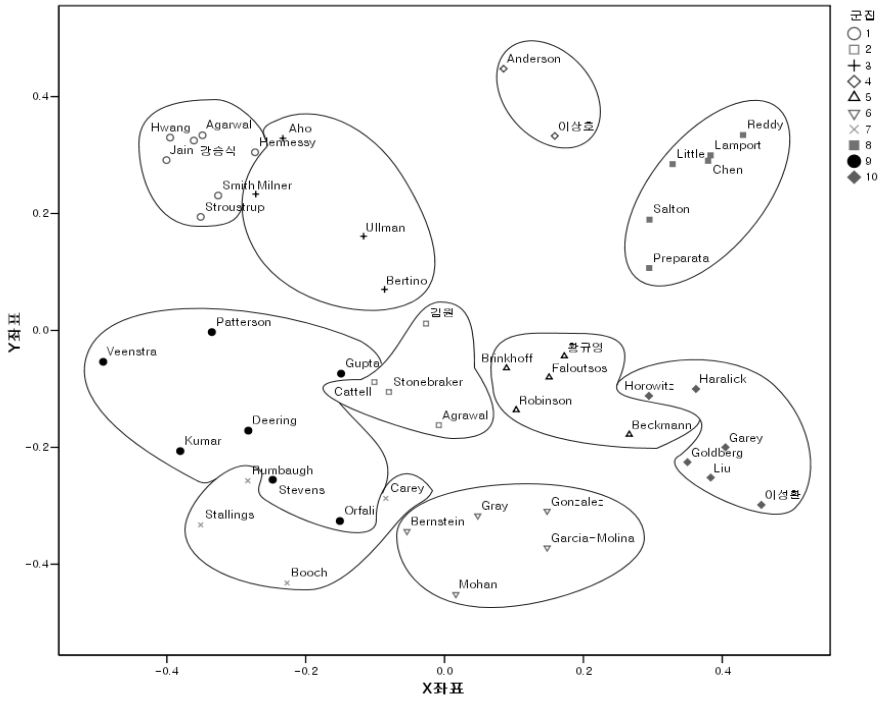
한편 이 연구에서 제안한 CMDS 기법으로 개선한 지적 구조와 군집 표현 결과는 <그림 3>과 같다. 이 그림에서는 군집 사이의 얽힘 현상이 없이 개별 저자들이 명확하게 분리되며, 한 군집 내에서도 각 저자들이 다른 주제 영역과 가까운 정도에 따라서 자리잡고 있는 것으로 나타난다. 이를 통해 CMDS 기법의 장점을 시각적으로 뚜렷하게 확인할 수 있다.



<그림 1> 정보학 연구자 50명의 저자동시인용빈도 행렬에 대한 ALSCAL 결과에 WARD 기법에 의한 10개 군집을 표시한 경우



<그림 2> 정보학 연구자 50명의 저자동시인용빈도 행렬에 대한 PROXSCAL 결과에 WARD 기법에 의한 10개 군집을 표시한 경우



<그림 3> 정보학 연구자 50명의 저자동시인용빈도 행렬에 대한 C-MDS 결과에 군집을 표시한 경우 (Ward군집 기법과 PROXSCAL 기법 결합)

<표 1> CMDS 결과와 ALSCAL, PROXSCAL 결과의 품질 비교

평가 척도	비교 기준	ALSCAL	PROXSCAL	CMDS(Ward와 PROXSCAL 결합)
결정계수	저자간 상관계수	0.281	0.403	0.354
	저자간 동시인용빈도	0.096	0.180	0.174
근거리 결정계수	저자간 상관계수	0.184	0.256	0.308
	저자간 동시인용빈도	0.053	0.113	0.125

이상과 같은 CMDS 결과가 원래 변수간의 관계를 지나치게 왜곡하는 것은 아닌지 검증하기 위해서 생성된 MDS 지도의 품질을 측정해보았다.

MDS 지도의 품질 측정을 위해 이 연구에서는 결정계수와 근거리 결정계수(이재윤 2007)를 사용하였다. 결정계수는 입력 행렬에 나타난 노드 사이의 값과 2차원 평면에 표현된 노드 사이의 거리 사이의 상관계수의 제곱을 구한 것으로서, 흔히 종속변수의 분산 중에서 독립변수에 의해 설명되는 비율로 해석된다(김태근 2006). 또한 근거리 결정계수는 전체 변수쌍의 거리가 아닌 근접성 상위 1/3쌍에 대해서만 산출한 결정계수로서, 다차원척도법이 지적 구조를 표현하는 능력을 더 잘 측정하는 척도이다.

<표 1>에 제시한 평가 결과 중에서 원거리까지 감안하는 평가 기준인 결정계수로는 저자간 상관계수나 동시인용빈도와 비슷한 정도가 CMDS 방식이 ALSCAL보다는 월등하며 PROXSCAL에는 약간 못 미치는 것으로 나타났다. 그러나 지적 구조의 파악에 중요한 단서가 되는 가까운 거리 위주의 평가 기준인 근거리 결정계수로는 CMDS 방식으로 생성한 2차원 지도가 가장 품질이 뛰어난 것으로 나타났다.

4. 결론

지적 구조의 분석과 표현에 널리 활용되고 있는 다차원척도법과 군집분석을 결합한 새로운 기법인 CMDS를 제안하였다. 제안된 기법을 정보학 분야 저자동시인용 데이터에 시험 적용해본 결과 기존 기법에 비해서 시각적인 표현 능력이 뛰어나며 지도의 품질도 높은 것으로 확인되었다. 향후에는 다양한 데이터를 대상으로 CMDS 기법의 가능성을 검토하여 적용 분야를 확대할 계획이다.

참고문헌

김태근. 2006.『u-Can 회귀분석』. 서울: 인간과 복지.

이은숙. 2002. 복수저자를 고려한 저자동시인용분석 연구: 정보학과 컴퓨터과학을 대상으로. 석사학위논문, 연세대학교 대학원.

이재윤. 2007. 지적 구조 분석을 위한 MDS 지도 작성 방식의 비교 분석. 『한국문헌정보학회지』, 41(2): 335-357.

Börner, K., C. Chen, and K. Boyack. 2003. "Visualizing knowledge domains". *Annual Review of Information Science & Technology*, 37: 179-255.

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

Noyons, E. C. M. and A. F. J. van Raan. 1998. "Advanced mapping of science and technology." *Scientometrics*, 41(1-2): 61-67.