클러스터링을 고려한 다차원척도법의 개선: 군집 지향 척도법*

Improved Multidimensional Scaling Techniques Considering Cluster Analysis: Cluster-oriented Scaling

이재윤(Jae Yun Lee)**

초 록

개체들 사이의 관계를 저차원 공간에 매핑하는 다차원척도법을 수행하기 위한 다양한 방법과 알고리즘이 개발되어왔다. 그러나 PROXSCAL이나 ALSCAL과 같은 기존의 기법들은 50개 이상의 개체를 포함하는 데이터 집합을 대상으로 개체 간의 관계와 군집 구조를 시각화하는데 있어서 효과적이지 못한 것으로 나타났다. 이 연구에서 제안하는 군집 지향 척도법 CLUSCAL(CLUster-oriented SCALing)은 기존 방법과 달리 입력되는 데이터의 군집 구조를 고려하도록 고안되었다. 50명의 저자동시인용 데이터와 85개 단어의 동시출현 데이터에 대해서 적용해본 결과 제안한 CLUSCAL 기법은 군집 구조를 잘 식별할 수 있는 MDS 지도를 생성하는 유용한 기법임이 확인되었다.

ABSTRACT

There have been many methods and algorithms proposed for multidimensional scaling to mapping the relationships between data objects into low dimensional space. But traditional techniques, such as PROXSCAL or ALSCAL, were found not effective for visualizing the proximities between objects and the structure of clusters of large data sets have more than 50 objects. The CLUSCAL(CLUster-oriented SCALing) technique introduced in this paper differs from them especially in that it uses cluster structure of input data set. The CLUSCAL procedure was tested and evaluated on two data sets, one is 50 authors co-citation data and the other is 85 words co-occurrence data. The results can be regarded as promising the usefulness of CLUSCAL method especially in identifying clusters on MDS maps.

키워드: 군집 지향 척도법, 다차원척도법, 클러스터링, 군집분석, 정보시각화
cluster-oriented scaling, CLUSCAL, MDS, multidimensional scaling, clustering, cluster analysis, information visualization

---

* 본 연구는 2009학년도 경기대학교 학술연구비(일반연구과제) 지원에 의하여 수행되었음.
** 경기대학교 문헌정보학과 부교수(memexlee@kgu.ac.kr)
■ 논문접수일자 : 2012년 5월 7일 ■ 최초심사일자 : 2012년 5월 15일 ■ 게재확정일자 : 2012년 5월 30일
■ 정보관리학회지, 29(2), 45-70, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.2.045〕

1. 서 론

동시인용분석(Small, 1973)과 저자동시인용분석(White & Griffith, 1981)을 비롯한 지적 구조 분석이나 연구동향 분석이 활성화되면서 최근까지 다양한 분석 사례가 발표되었다. Small (1973)이 동시인용분석을 제안했을 때에는 지적 구조를 문헌 간의 네트워크로 시각화하였으나 White와 Griffith(1981)가 다차원척도법과 군집분석을 사용한 이후 동일한 방법이 다수의 연구에서 채택되었다. 2차원 지도로 표현된 지적 구조는 다차원척도법으로 저자나 단어와 같은 개체를 배치하고 군집분석의 결과에 따라서 군집을 표현한다.

최근 발표된 지적구조 분석 연구에서 다차원척도법을 사용한 연구를 살펴보면 48명의 저자 동시인용분석을 수행한 곽선영과 정은경(2012), 37명의 저자동시인용분석을 수행한 문주영(2011), 46명의 저자프로파일링분석과 저자동시인용분석을 수행한 유종덕과 최은주(2011), 46명의 저자동시인용분석과 43명의 저자서지결합분석을 수행한 김희전과 조현양(2010) 등이 있다. 그런데 이들 연구는 모두 MDS 지도에 표현한 개체의 수가 50개 미만이라는 공통점이 있다. 시각적으로 표현해야 하는 개체가 너무 많을 경우에는 아예 다차원척도법이 수행되지 않기 때문에 개체 대신 개체가 모인 군집이나 요인 사이의 관계를 다차원척도법으로 나타낸 연구도 있었다(박재신, 정영미, 2010; 이재윤, 김희전, 유종덕, 2010; 조선례, 이재윤, 2012).

50개 이상의 개체를 대상으로 지적구조 분석을 수행한 최근 연구로는 98개 키워드의 동시출현분석과 62명의 저자동시인용분석을 수행한 Choi와 Lee(2011), 52개 디스크립터의 동시출현분석을 수행한 Park과 Kim(2011), 60개 키워드의 동시출현분석을 수행한 Choi(2011), 59개 저자 키워드와 55개 저자 키워드의 동시출현 네트워크 분석을 수행한 조재인(2011), 53명의 저자에 대한 프로파일링 분석을 수행한 김판준(2011) 등이 있다. 이들은 모두 개체 사이의 관계를 네트워크로 시각화하였다.

개체 수가 50개 이상으로 많은 경우에 다차원척도법을 사용하지 않고 네트워크로 표현한 이유는, 다차원척도법(MDS)이 국지적인 세부 구조를 제대로 정확하게 표현하지 못하는 단점이 있기 때문이다(Börner et al., 2003; Chen, 2006). 개체들 사이의 복잡한 고차원 관계를 2차원의 평면에 표현하면서 일부 정보가 손실되는데, 함께 사용되는 군집분석이 동일한 군집에 속한 개체들이 어떤 것인지에 대한 국지적인 정보만을 산출하기 때문이다. 전체적인 지적 구조를 보여주는 MDS 지도 위에 국지적인 관계를 부각시키는 군집을 표현해보면 다른 군집에 속한 개체가 뒤섞여서 제대로 군집을 구분하기 어려운 경우도 나타난다. 특히 MDS 지도에 표현할 저자나 단어가 50개 이상으로 많으면 이런 현상이 흔히 나타나며, 군집의 영역이 서로 배타적으로 표시되지 않기 때문에 지적 구조의 파악이 제대로 이루어질 수 없다. 이런 점을 고려하여 SPSS를 이용해서 다차원척도법을 수행할 때 어떤 절차와 설정으로 수행하는 것이 최선인가를 살펴본 연구(이재윤, 2007)도 있었다. 그러나 선행연구에서 제시된 절차와 설정을 따르더라도 개체의 수가 50개 보다 많으면 지적 구조를 표현하는데 한계가 있다.

이 연구에서는 국지적 구조를 표현하는 능력

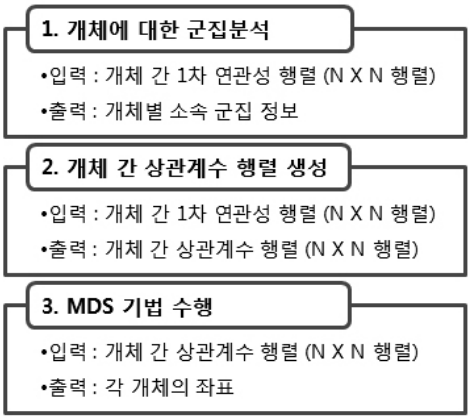
이 약한 다차원척도법의 한계를 극복하고, 표현해야 하는 개체의 수가 많을 경우에도 군집분석과 잘 조화되는 개선된 다차원척도법을 개발하고자 한다. 군집분석 결과와 가장 잘 어울리는 처리 절차를 파악하고자 하였던 선행 연구(이재윤, 2007)가 기존 방법 중에서 최선의 경우를 찾고자 하는 수동적인 접근이었다면, 이 논문은 기존 방법을 대체할 수 있는 새로운 기법을 개발하는 능동적인 접근이라고 할 수 있다.

다차원척도법을 수행한 결과가 군집분석 결과와 잘 어울리기 위해서는 우선 군집분석을 수행하여 각 개체의 소속 군집을 결정한 후, 이 정보를 각 개체의 MDS 지도 좌표 산출에 반영하는 방법을 생각해볼 수 있다. 즉, 각 개체의 좌표를 결정할 때 군집분석으로 결정된 소속 군집을 고려하는 것이다. 이와 같이 군집분석 결과로 도출된 각 개체의 소속 군집 정보를 반영하는 새로운 다차원척도법으로 이 연구에서는 '군집 지향 척도법'(CLUster-oriented SCALing; 약자로는 CLUSCAL)을 제안하고자 한다. 새로 개발되는 CLUSCAL은 PROXSCAL이나 ALSCAL과 같은 기존의 MDS 기법을 활용하되 사전에 군집분석을 수행하여 각 개체의 소속 군집 정보를 획득한 다음 이를 다차원척도법에 반영하도록 고안되었다. CLUSCAL의 수행을 위한 프로그램은 파이썬 언어로 구현하였으며 기존의 MDS 기법은 SPSS v.20을 활용하였다.

2. 군집을 고려한 다차원척도법 CLUSCAL

2.1 전통적인 다차원척도법의 수행 절차와 문제점

지적 구조 분석을 위해서 다차원척도법과 군집분석을 수행하는 절차는 〈그림 1〉과 같으며 White와 Griffith(1981)의 연구 이후 거의 변화가 없었다. 다차원척도법과 군집분석을 각각 수행한 이후에 MDS 지도에 군집 구분을 선으로 표시하여 지적 구조를 표현하는 지도를 완성



〈그림 1〉 다차원척도법과 군집분석을 별도로 수행하는 기존 절차

한다. 별도로 수행된 다차원척도법과 군집분석의 결과를 결합하는 과정에서 불일치가 발생할 여지가 생긴다.

개체 수가 50개 이상인 데이터에 대해서 실제로 다차원척도법으로 생성한 MDS 지도에 군집분석 결과를 표현했을 때 발생하는 문제를 확인해보기 위해서 〈표 1〉과 같이 일반적으로 지적 구조 분석에 가장 많이 활용되는 기법인 저자동시인용분석과 단어동시출현분석 사례를 한 가지씩 준비하였다. AC50은 이은숙과 정영미(2002)의 연구에서 사용된 저자동시인용분석 자료로서 국내 컴퓨터과학 분야 주요 저자 50명을 대상으로 한 것이다. CW85는 이재윤과 정주희(2006)의 연구에서 사용된 단어동시출현 자료로서 국내 인지과학분야 학술논문에서 추출한 주요 표제어 85개를 대상으로 한 것이다.

저자동시인용분석인 AC50과 단어동시출현분석인 CW85의 두 데이터에 대해서 〈그림 1〉의 절차에 따라 통계분석 프로그램인 SPSS를 이용해 다차원척도법과 군집분석을 수행하고 생성된 MDS 지도에 군집을 표시한 결과가 〈그림 2〉와 〈그림 3〉이다. 다차원척도법을 수행할 때에는 이재윤(2007)의 분석결과에 따라서 가장 좋은 품질의 MDS 지도가 생성될 수 있도록 PROXSCAL 프로시저를 선택하고 피어슨 상관계수 행렬을 입력한 후 z점수로 표준화한 다음 제곱유클리드거리를 산출하는 3ZP 방식을 따랐다. 군집분석에서 클러스터링 기법과 적정 군집 수는 두 데이터를 각각 수집하고 분석한 선행 연구(이은숙, 정영미, 2002; 이재윤, 정주희, 2006)에서 정한 대로 따랐다.

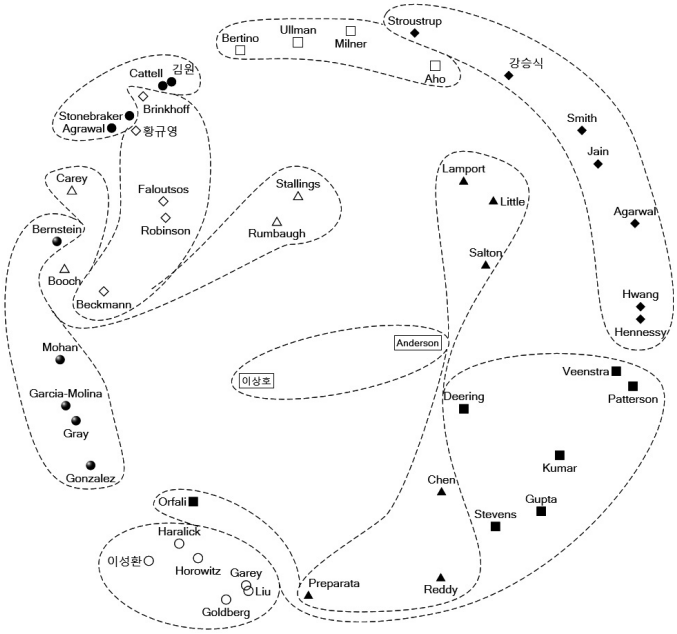
50명의 저자에 대한 동시인용분석 결과인 〈그림 2〉에서는 특정 군집에 속한 개체들이 다른 군집의 경계를 넘어 멀리 떨어져 위치한 경우가 전체 열 군집 중 세 군집으로 나타났다. 왼쪽에 빈 삼각형 표식으로 표현된 4개 개체가 포함된 군집과 중앙의 오른쪽 아래에 검은 사각형 표식으로 표현된 7개 개체가 포함된 군집은 U자형으로 좌우가 분리된 것처럼 보이며, 중앙 우측의 검은 삼각형 표식으로 표현된 6개 개체로 구성된 군집은 위에서 아래로 길게 늘어져서 표현되었다.

85개의 단어에 대한 동시출현분석 결과인 〈그림 3〉에서는 중앙의 작은 두 군집(빈 마름모꼴 표식으로 표현된 '의미'가 포함된 군집과 검은 마름모꼴 표식으로 표현된 '알고리즘'이 포함된 군집)을 제외한 대부분의 군집이 뒤틀리게 표현되었다. 특히 빈 사각형 표식으로 표현된 오른쪽 위의 군집('표상' 포함)은 소속 개체인 '텍스트'가 멀리 떨어진 왼쪽 윗부분에 배치되어 다른 군집들 사이에 끼인 형상이 되었다. 왼쪽 위의 검은 사각형 표식으로 표현된 군집('구조' 포함)과 그 오른쪽의 빈 동그라미 표식으로 표현된 군집('시스템' 포함)은 서로 뒤엉킨 형태가 되었다.

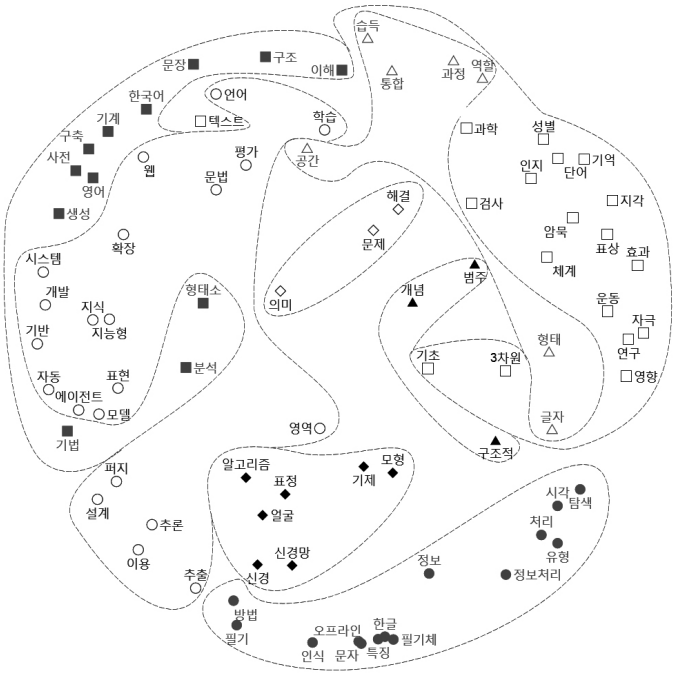
이상과 같이 전통적인 지적구조 분석 연구에

〈표 1〉 다차원척도법의 품질 평가를 위한 실험 데이터

명칭	유형	개체 수	내용	원자료 출처
AC50	저자동시인용	50	국내 컴퓨터과학 분야 주요 저자	이은숙, 정영미(2002)
CW85	단어동시출현	85	국내 인지과학분야 학술논문의 주요 표제어	이재윤, 정주희(2006)



〈그림 2〉 PROXSCAL(AC50, Ward기법 10군집)



〈그림 3〉 PROXSCAL(CW85, 평균연결기법 8군집)

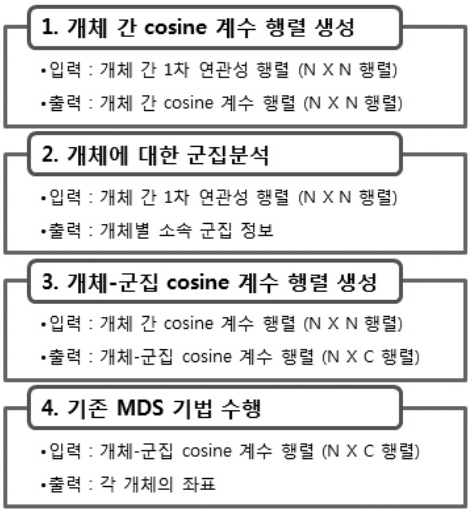
서 사용된 다차원척도법과 군집분석 결과는 개체 수가 50개 이상인 경우에 적절하게 구조를 표현하지 못하며 개체가 배치된 위치가 소속 군집을 제대로 반영하지 못하는 경우가 많음을 알 수 있다. 개체 수가 50개인 〈그림 2〉보다 개체 수가 85개인 〈그림 3〉에서 더 많은 문제가 발견되었으므로 표현해야 하는 개체 수가 많을수록 이런 문제점은 더욱 심각할 것으로 예상된다.

2.2 군집 지향 척도법 CLUSCAL

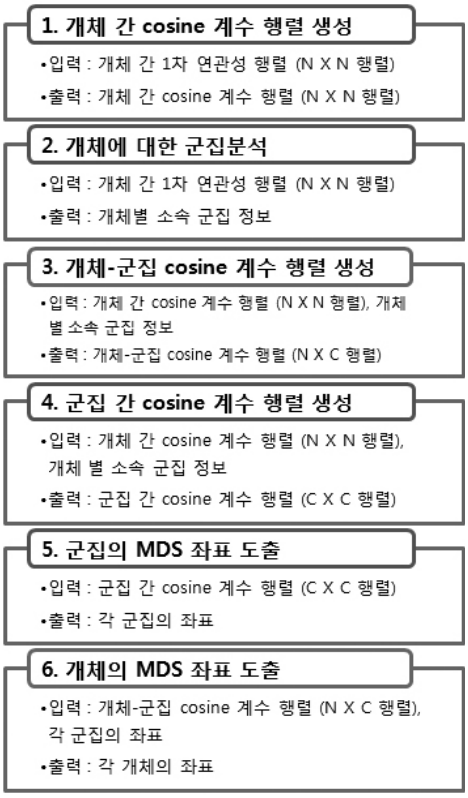
다차원척도법을 수행할 때 각 개체가 어느 군집에 소속되었는가에 따라서 동일 군집에 속한 개체들의 좌표를 가깝게 설정한다면 MDS 지도 상에서 군집 정보를 손쉽게 식별할 수 있다. 이 연구에서 제안하는 군집 지향 척도법 CLUSCAL은 각 개체가 어느 군집에 속했는지를 고려하여 다차원척도법을 수행하도록 고안되었다. 개체의 소속 군집 정보를 반영하는 시기에 따라서 기존 MDS 기법을 수행하기 전에 반영하는 CLUSCAL-A와 기존 MDS 기법을 수행한 이후에 반영하는 CLUSCAL-B로 구분한다.

CLUSCAL 기법 중에서 첫 번째인 CLUSCAL-A는 〈그림 4〉와 같이 기존 MDS 기법을 수행하기 전에 개체의 소속 군집 정보를 반영한다. CLUSCAL-A 기법이 기존의 방식과 다른 점은 다차원척도법의 입력 데이터로 개체 간 관계 행렬을 사용하는 대신 개체-군집 코사인 계수 행렬을 사용하는 것이며, 입력 데이터에 대해서 PROXSCAL이나 ALSCAL과 같은 기존 MDS 기법을 적용하는 것은 동일하다.

CLUSCAL-B 기법에서는 〈그림 5〉와 같이 일단 군집분석을 수행하고 군집 간 연관성을 이용해서 군집의 MDS 지도를 산출한 이후 각 군집에 속한 개체의 좌표를 결정한다. CLUSCAL-B 기법의 3단계까지는 CLUSCAL-A 기법과 동



〈그림 4〉 CLUSCAL-A 수행 절차



〈그림 5〉 CLUSCAL-B 수행 절차

일하다. CLUSCAL-A 기법에서는 3단계에서 도출된 개체-군집 코사인 계수 행렬을 MDS 기법의 입력 데이터로 사용하였지만, CLUSCAL-B 기법에서는 군집 간 코사인 계수 행렬을 생성하여 MDS 기법의 입력 데이터로 사용한다. 각 군집의 좌표를 결정한 이후에는 개체와 군집 사이의 코사인 유사도를 활용하여 개체의 좌표를 결정한다.

2.2.1 CLUSCAL-A 기법

CLUSCAL-A 기법에서는 다차원척도법의 입력 데이터로 개체 간 관계 행렬을 사용하는 대신 개체-군집 코사인 계수 행렬을 사용한다. 개체-군집 코사인 계수 행렬은 각 개체와 각 군집 사이의 코사인 계수로 구성되며 각 군집과 그 군집에 속한 개체와의 코사인 계수는 1.0으로 설정한다. 예를 들어 개체 n_a와 군집 C_j 사이의 코사인 계수 $\cos(n_a,C_j)$는 군집 C_j에 속한 각 개체들과 개체 n_a 사이의 코사인 계수를 평균하여 산출한다. 이는 군집 C_j의 센트로이드와 개체 n_a 사이의 코사인 계수를 구하는 것과 같다. 수식으로 표현하면 다음과 같다.

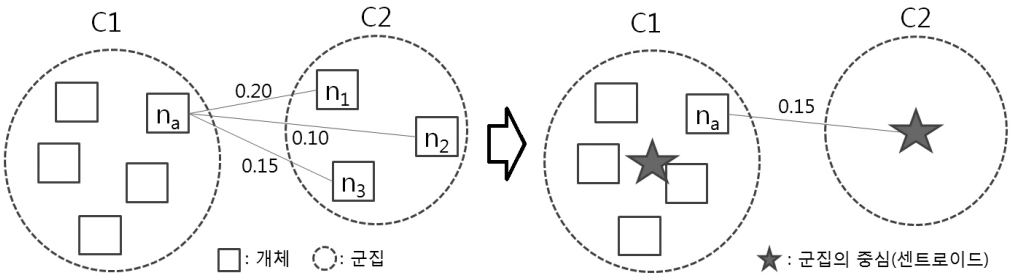
$$\cos(n_a, C_j) = \frac{1}{size(C_j)} \sum_{k=1}^{size(C_j)} \cos(n_a, n_k), \quad \forall n_k \in C_j$$

예를 들어 〈그림 6〉과 같이 군집 C_1에 속한 개체 n_a와 군집 C_2에 속한 개체 n_1, n_2, n_3 사이의 코사인 계수가 각각 0.20, 0.10, 0.15라고 하면 개체 n_a와 군집 C_2 사이의 코사인 계수는 세 값을 더해서 3으로 나눈 0.15가 된다. 개체 n_a와 군집 C_1 사이의 코사인 계수는 n_a가 군집 C_1의 소속 개체이므로 1.0으로 처리한다.

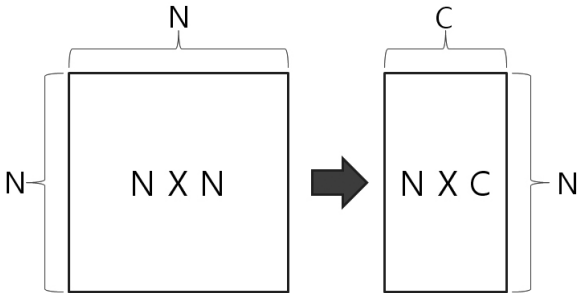
전체 개체 수가 N개일 때 개체 간 관계 행렬은 크기가 N×N인 행렬이 되는데, CLUSCAL-A 기법에서 입력 데이터로 사용하는 개체-군집 행렬은 군집 수가 C개일 때 크기가 N×C인 행렬이 된다. 일반적으로 군집 수 C가 개체 수 N보다 매우 적으므로 〈그림 7〉과 같이 N개의 개체를 N차원 벡터로 표현한 행렬로부터 N개의 개체를 C차원 벡터로 표현한 행렬로 축소하게 된다. 결국 CLUSCAL-A는 잠재의미분석(Latent Semantic Analysis; Deerwester et al., 1989)과 유사한 일종의 차원축소(dimension reduction)를 수행하는 기법이라고 할 수 있다.

2.2.2 CLUSCAL-B 기법

CLUSCAL-B 기법에서는 앞의 〈그림 5〉와 같이 먼저 군집을 도출하고 군집의 MDS 좌표를 구한 다음 개체와 군집 사이의 코사인 유사도를 활용하여 개체의 좌표를 결정한다. 군집의 좌표로부터 개체의 좌표를 도출하는 방법과 개체의 좌표를 미세 조정하는 방식에 따라서 CLUSCAL-B1에서 CLUSCAL-B4까지의 네 종류로 세분된다. 따라서 이후에 설명하는 네 가지 CLUSCAL-B 기법은 〈그림 5〉의 다섯



〈그림 6〉 개체 간 코사인 계수로부터 개체-군집 간 코사인 계수를 산출하는 개념도



〈그림 7〉 개체 간 코사인 계수 행렬에서 개체-군집 코사인 계수 행렬로의 차원 축소

번째 단계까지는 동일한 방법으로 수행되며 여섯 번째 단계의 세부 내용이 달라진다.

1) CLUSCAL-B1

CLUSCAL-B1에서는 〈그림 5〉의 다섯 번째 단계까지 수행해서 군집의 좌표를 도출한 다음, 세 번째 단계에서 미리 구해둔 개체와 군집 사이의 유사도(코사인 계수)에 비례하여 각 군집의 좌표를 개체의 좌표에 반영한다. 구체적으로 개체 n_i의 X축 좌표 $X(n_i)$와 Y축 좌표 $Y(n_i)$를 도출하기 위한 공식은 아래와 같다.

$$X(n_i) = \sum_j \{\cos(n_i, C_j) \times X(C_j)\}$$
$$Y(n_i) = \sum_j \{\cos(n_i, C_j) \times Y(C_j)\}$$

이 공식은 개체와 각 군집 사이의 코사인 계수 $\cos(n_i,C_j)$에 비례하여 군집의 좌표 $X(C_j)$와 $Y(C_j)$를 각각 축별로 합산하도록 되어 있다. 개체 n_i의 좌표를 도출할 때, 개체 n_i와의 코사인 계수가 0.2인 군집의 좌표는 20%가 반영되며, 개체 n_i의 소속 군집 $C(n_i)$는 개체 n_i의와의 코사인 계수가 1.0이므로 좌표값이 100% 반영된다. 따라서 대부분의 개체는 소속 군집에 가까운 위치로 배치되며 개체마다 타 군집 중에서 비교적 유사도가 높은 군집을 향한 방향으로 조금씩 옮겨서 자리잡게 된다.

예를 들어 4개 군집이 존재하는 데이터에서 군집 간 코사인 계수 및 개체 n_1과 각 군집 사이의 코사인 계수가 〈표 2〉와 같고 각 군집의 MDS 좌표가 〈표 3〉과 같이 산출되었을 때, 개체 n_1의 X축 좌표 $X(n_1)$과 Y축 좌표 $Y(n_1)$을 산출해보면 각각 다음과 같다.

$$\begin{aligned} X(n_1) &= \cos(n_1,C_1)\times X(C_1)+\cos(n_1,C_2)\times X(C_2)] \\ &\quad +\cos(n_1,C_3)\times X(C_3)+\cos(n_1,C_4)\times X(C_4) \\ &= 1.00\times 0.7 + 0.20\times 0.6 + 0.15\times -0.9 + 0.10\times -0.4 \\ &= 0.645 \end{aligned}$$

$$\begin{aligned} Y(n_1) &= \cos(n_1,C_1)\times Y(C_1)+\cos(n_1,C_2)\times Y(C_2) \\ &\quad +\cos(n_1,C_3)\times Y(C_3)+\cos(n_1,C_4)\times Y(C_4) \\ &= 1.00\times 0.8 + 0.20\times -0.4 + 0.15\times 0.4 + 0.10\times -0.8 \\ &= 0.700 \end{aligned}$$

〈그림 8〉에서는 네 군집과의 코사인 계수에 비례한 만큼 좌표를 반영하여 산출한 개체 n_1의

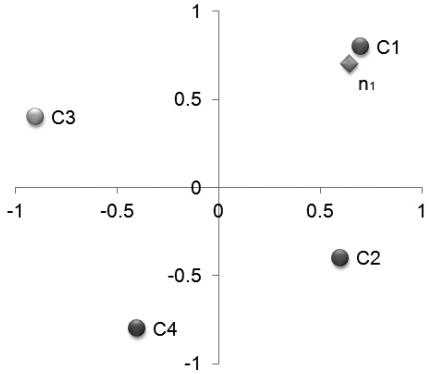
〈표 2〉 가상 사례 - 4개 군집 간 코사인 계수 및 개체 n_1과 각 군집 간의 코사인 계수

	C_1	C_2	C_3	C_4
C_1	1.00	0.30	0.10	0.05
C_2	0.30	1.00	0.08	0.25
C_3	0.10	0.08	1.00	0.15
C_4	0.05	0.25	0.15	1.00

	n_1
C_1	1.00
C_2	0.20
C_3	0.15
C_4	0.10

〈표 3〉 가상 사례 - 4개 군집의 MDS 좌표

	C_1	C_2	C_3	C_4
X	0.7	0.6	-0.9	-0.4
Y	0.8	-0.4	0.4	-0.8



〈그림 8〉 가상 사례에 대해서 CLUSCAL-B1으로 n_1의 좌표를 도출한 결과

좌표가 소속 군집인 C_1 부근임을 보여주고 있다. 이 그림에서도 볼 수 있듯이 CLUSCAL-B1으로 산출한 개체의 좌표는 대체로 소속 군집의 좌표에서 MDS지도의 원점 방향으로 이동하는 경향이 있다.

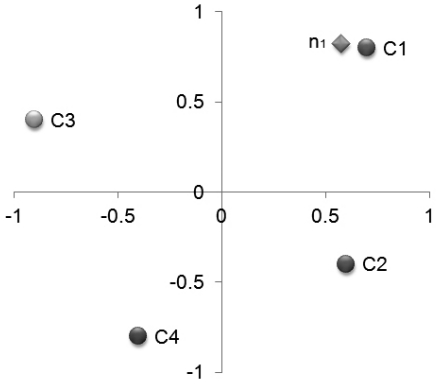
2) CLUSCAL-B2

CLUSCAL-B1에서는 소속 군집을 비롯해서 모든 군집의 좌표를 코사인 계수 크기에 비례해서 반영하였다. CLUSCAL-B2에서는 각 개체가 속한 소속 군집의 좌표를 기점으로 하여 개체의 초기 좌표를 정한 다음, 소속 군집이 아닌 타 군집과 가지는 코사인 계수에 비례해서 위치를 이동시키되, 소속 군집 C_i가 타 군집 C_j와 가지는 코사인 계수보다 개체 n_i가 타 군집 C_j와 가지는 코사인 계수가 더 크면 차이에 비례해서 C_j의 좌표 쪽으로 더 이동시키고 더 작으면 C_j의 좌표와 반대 방향으로 이동시킨다. 만약 소속 군집의 중심과 동일한 개체인 경우에는, 개체와 타 군집과의 코사인 계수가 소속 군집과 타 군집 간의 코사인 계수와 동일하므로 위치도 추가 이동없이 초기 좌표로 지정된 군집의 좌표가 된다. X좌표가 $X(C_i)$이고 Y좌표가 $Y(C_i)$인 군집 C_i에 속한 개체 n_i의 X축 좌표 $X(n_i)$와 Y축 좌표 $Y(n_i)$는 각각 다음 공식으로 산출한다.

$$X(n_i) = X(C_i) + X\text{축 이동 거리}$$
$$= X(C_i) + \sum_{j \neq i} [\{\cos(n_i, C_j) - \cos(C_i, C_j)\} \times \{X(C_j) - X(C_i)\}]$$

$$Y(n_i) = X(C_i) + Y\text{축 이동 거리}$$
$$= Y(C_i) + \sum_{j \neq i} [\{\cos(n_i, C_j) - \cos(C_i, C_j)\} \times \{Y(C_j) - Y(C_i)\}]$$

〈표 2〉와 〈표 3〉의 가상 사례에 대해서 CLUSCAL-B2를 적용하면 다음과 같이 개체 n_1의 X좌표값과 Y좌표값이 산출되고, 네 군집과 함께 지도에 표시해보면 〈그림 9〉와 같이 개체 n_1은 소속 군집인 C_1의 위치로부터 X축 상으로 -0.125만큼 이동(왼쪽으로 0.125 이동)하고 Y축 상으로 0.02만큼 이동(위쪽으로 0.02 이동)하여 배치된다. 이 그림에서도 볼 수 있듯이 CLUSCAL-B2로 산출한 개체의 좌표는 대체로 소속 군집의 좌표를 중심으로 주변에 분포하는 경향이 있다.



〈그림 9〉 가상 사례에 대해서 CLUSCAL-B2로 n_1의 좌표를 도출한 결과

$$X(n_i) = X(C_i) + \sum_{j \neq i}\{(\cos(n_i, C_j) - \cos(C_i, C_j)) \times (X(C_j) - X(C_i))\}$$
$$= 0.7 + (0.20 - 0.30) \times (0.6 - 0.7) + (0.15 - 0.10) \times (-0.9 - 0.7) + (0.10 - 0.05) \times (-0.4 - 0.7)$$
$$= 0.7 + (-0.125) = 0.575$$

$$Y(n_i) = Y(C_i) + \sum_{j \neq i}\{(\cos(n_i, C_j) - \cos(C_i, C_j)) \times (Y(C_j) - Y(C_i))\}$$
$$= 0.8 + (0.20 - 0.30) \times (-0.4 - 0.8) + (0.15 - 0.10) \times (0.4 - 0.8) + (0.10 - 0.05) \times (-0.8 - 0.8)$$
$$= 0.8 + 0.020 = 0.820$$

3) CLUSCAL-B3

앞에서 살펴본 CLUSCAL-B2 기법은 처리 대상 데이터의 코사인 계수 수준에 따라서 기점으로부터의 이동거리가 좌우되기 때문에 각 개체가 소속 군집의 좌표로부터 이동하는 거리가 너무 짧아서 군집별로 개체가 한 덩어리로 뭉쳐 잘 구분되지 않을 위험이 있다. 반대로 개체가 소속 군집의 좌표로부터 너무 많이 이동하면 다른 군집의 개체들과 섞여버리는 현상도 나타날 수 있다.

CLUSCAL-B3는 이런 문제를 해결하기 위해서 각 군집 소속 개체가 다른 군집의 영역을 침범하지 않고 자리 잡을 수 있는 범위인 군집의 반경을 미리 산출해두고 개체의 이동거리를 이 반경 이내로 설정하는 방법이다. 군집 C_j의 반경 rad(C_j)은 가장 가까운 군집과의 직선거리의 절반으로 산출한다.

군집의 좌표를 기점으로 하여 CLUSCAL-B2와 같이 X축과 Y축 별로 이동거리를 산출한 다음 개체 n_i가 기점으로부터 이동한 총 거리 d(n_i)($= \sqrt{X\text{축 이동거리}^2 + Y\text{축 이동거리}^2}$)를 계산한다. 그 후 수정된 이동거리 d′은 n_i가 소속된 군집 C(n_i)의 상대적인 크기인 $RS(C(n_i))$와 군집 C(n_i)에 포함된 각 개체들의 이동거리 중 최댓값 대비 해당 개체의 이동거리의 비율인 $RD(n_i)$을 각각 개체 n_i가 소속된 군집의 반경 rad(C(n_i))에 곱해서 산출한다.

$$d' = rad(C(n_i)) \times \alpha \times RS(C(n_i)) \times RD(n_i)$$

이 공식에서 α(alpha) 항은 군집의 반경을 어느 정도까지 개체를 배치하는데 사용할 것인가를 결정하는 파라미터로서 최대 이동거리를 결정한다. α항에는 0에서 1 사이의 값을 지정할 수 있는데, 군집의 반경을 폭넓게 활용하기 위해서 1에 가까운 높은 값을 설정하는 것이 바람직하다. 개체 n_i가 소속된 군집 C(n_i)의 상대적

인 크기인 RS(C(n_i))는 다음과 같이 가장 큰 군집의 개체 수로 개별 군집의 개체 수를 나눈 것이다.

$$RS(C(n_i)) = \frac{|C(n_i)|}{\max(|C_j|)}$$

개체 n_i의 상대적 이동거리를 반영하는 항인 RD(n_i)는 다음과 같이 해당 군집 내에서 기점으로부터의 거리가 가장 먼 개체의 이동거리로 개별 개체의 이동거리를 나눈 것이다.

$$RD(n_i) = \frac{d(n_i)}{\max(d(n_j))}, \quad \forall n_j \in C(n_i)$$

위와 같은 과정을 거쳐 수정된 이동거리 d′이 산출되면 해당 개체가 기점으로부터 이동하던 원래의 방향 그대로 d가 아닌 d′ 만큼 이동하도록 X, Y축 좌표를 산출한다.

4) CLUSCAL-B4

소속 개체 사이의 연관성이 매우 밀접한 군집에서는 여러 개체의 이동거리가 지나치게 짧게 산출되어 기점에 몰려서 배치되는 현상이 나타날 수 있다. 이렇게 되면 원래 군집의 좌표 부근에 여러 개체가 뭉쳐지게 되므로 군집 내 개체들을 시각적으로 구분하기 어려워진다.

CLUSCAL-B4는 이와 같이 군집의 원래 좌표인 기점에 여러 개체가 몰리는 현상을 방지하기 위해서 최소 이동 거리를 보장하도록 고안되었다. 수정된 이동거리 d′을 산출하는 CLUSCAL-B3의 공식을 간단히 수정하여 최소 이동거리 비율을 나타내는 β(beta) 항을 추가하였다. 또한 군집의 상대적 크기를 나타내는 RS 항과 개체의 상대적인 이동거리를 나타내는 RD 항을 제곱하여 반영함으로써 두 값이 지나치게 큰 경우에 군집으로부터 다소 멀리 위치하는 개체가 발생할 가능성을 감소시켰다. CLUSCAL-B4에서 수정된 이동거리 d′을 산출하는 공식은 다음과 같다.

$$d' = rad(C(n_i)) \times \left(\beta + (\alpha - \beta) \times RS(C(n_i))^2 \times RD(n_i)^2\right)$$

CLUSCAL-B4에서 도입된 최소 이동거리 파라미터인 β항도 최대 이동거리 파라미터인 α항처럼 0에서 1 사이로 지정할 수 있다. 하지만 1에 가까운 값으로 지정하는 α항과 달리 β항은 0에 가까운 값으로 지정해야 한다. 대체로 군집 반경의 10% 정도에 해당하는 0.1 이내로 지정하면 무난하다.

3. CLUSCAL 기법의 적용 결과

3.1 MDS 지도의 시각적 평가

CLUSCAL 기법을 적용한 결과를 전통적인 다차원척도법의 결과와 비교해보기 위해서 2.1절의 〈표 1〉에서 소개된 AC50 데이터와 CW85 데이터를 대상으로 CLUSCAL 기법 다섯 가지를 적용해보았다. CLUSCAL 기법을 수행하는 과정에서 필요한 기존 MDS 기법으로는 PROXSCAL 기법을 사용하였다. 앞의 2장에서 기존의 PROXSCAL 기법만을 적용한 결과인 〈그림 2〉, 〈그림 3〉과 비교하기 위해서 다섯 가지 CLUSCAL 기법을 두 데이터에 대해서 적용한 10가지 결과를 〈그림 10〉에서 〈그림 19〉까지 제시하였다.

이 절에서는 CLUSCAL 기법으로 생성된 다섯 종류의 MDS 지도에 대해서 시각적 평가를 수행하였다. 시각적 평가에서는 군집의 구분력과 개체의 구분력을 기준으로 살펴보았다. 군집의 구분력은 한 군집에 소속된 개체가 타 군집의 개체들과 잘 구분되는 정도이고, 개체의 구분력은 개체와 개체가 너무 뭉치지 않고 적절히 구분될 수 있도록 떨어져서 배치되는 정도를 의미한다. 각 개체를 서로 멀리 띄어 놓으면 개체의 식별이 용이하지만 타 군집의 개체들과 뒤섞일 위험이 높아지고, 반대로 동일 군집의 개체끼리 가깝게 모아놓으면 군집의 구분은 용이하지만 여러 개체가 뭉치게 되어 각 개체를 식별하기 어려울 위험이 높아진다. 따라서 시각적으로 잘 만들어진 MDS 지도라면 군집의 구분력과 개체의 구분력이 적절히 조화롭게 확보되어야 한다.

CLUSCAL 기법 중에서 첫 번째인 CLUSCAL-A를 AC50 데이터에 적용한 결과인 〈그림 10〉과 CW85 데이터에 대해서 적용한 결과인 〈그림 11〉을 살펴보면 대부분의 군집이 다른 군집과 엉키지 않고 뚜렷하게 구분된다는 것을 알 수 있다. 다만 〈그림 10〉에서는 위쪽의 '이상호'와 'Anderson'으로 구성된 작은 군집에 속한 두 개체가 비교적 멀리 떨어져서 배치되었으며, 아래쪽의 빈 삼각형 표식으로 표현된 군집('Booch'가 포함된 군집)은 다른 군집들 사이에 끼인 형태로 늘어져 있다. 이들 군집에 속한 개체들은 군집을 선으로 표시하지 않을 경우에는 동일한 군집에 속한 것을 알아보기 어렵다. CW85의 결과인 〈그림 11〉에서는 8개 군집 중에서 왼쪽의 검은 사각형 표식으로 표현된 군집에서 '표상'과 같은 일부 개체가 이웃 군집보다 더 먼 위치에 배치되는 현상이 나타났다. 이와 같이 CLUSCAL-A에 의한 MDS 지도는 군집끼리 엉키는 현상은 나타나지 않고 대부분의 군집에서 소속 개체끼리 한 군집임을 시각적으로 확인할 수 있었으나, 소속 개체끼리 동일 군집에 속한다는 것을 확인하기 어려운 군집도 일부 확인되었다.

군집의 좌표를 먼저 결정한 후 개체와 군집 사이의 코사인 계수에 비례하여 각 군집의 좌표를 개체의 좌표에 반영하는 CLUSCAL-B1을 AC50 데이터에 적용한 결과는 〈그림 12〉와 같으며, CW85 데이터에 적용한 결과는 〈그림 13〉과 같다. 〈그림 12〉는 CLUSCAL-A를 적용했던 〈그림 10〉에 비해서 모든 군집의 소속 개체가 가까운 자리에 위치하고 있으므로 군집 구분선이 없더라도 소속 군집을 쉽게 알아볼 수 있다. 그러나 〈그림 13〉의 경우에는 동일 군집에 속한 개체끼리 지나치게 가깝게 배치되어서 군집을 식별하기는 용이하지만 군집 소속 개체를 서로 구분하기가 곤란하다. 개체 수가 50개 정도인 AC50 데이터에 대해서는 문제가 없었으나 개체 수가 85개로 훨씬 많은 CW85 데이터에 대해서는 CLUSCAL-B1이 효과적이지 않은 것으로 나타났다

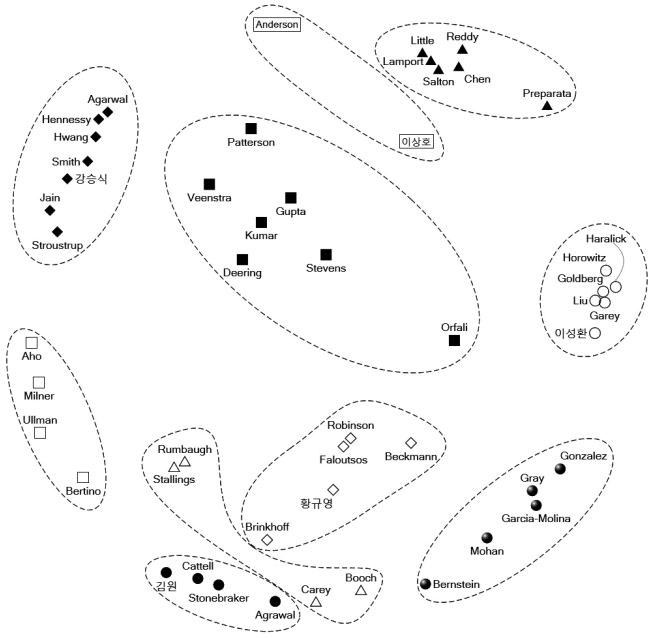
각 개체가 속한 소속 군집의 좌표를 기점으로 하여 타 군집과의 상대적인 연관성에 비례해서 개체의 좌표를 정하는 CLUSCAL-B2를 AC50 데이터에 적용한 결과는 〈그림 14〉와 같고, CW85 데이터에 적용한 결과는 〈그림 15〉와 같다. 〈그림 14〉는 CLUSCAL-B1의 결과인 〈그림 12〉와 유사하지만 위쪽의 'Bertino'(빈 사각형 표식)와 그 왼쪽의 '김원'(검은 동그라미 표식)은 소속 군집이 다름에도 불구하고 근접하게 배치

되었으며, 왼쪽의 'Berstein'(입체 원 표식)은 소속 군집보다 위에 있는 군집(빈 삼각형 표식)에 더 가깝게 자리잡고 있다. CW85 데이터에 적용한 결과인 〈그림 15〉는 CLUSCAL-B1의 결과인 〈그림 13〉보다는 각 개체 사이의 간격이 조금씩 벌어져 있으나 겹쳐진 개체가 여전히 적지 않으며 군집 사이의 간격은 지나치게 벌어져 있다. 이와 같이 CLUSCAL-B2는 개체 수가 많은 CW85 데이터에 적용했을 때 개체가 뭉치는 문제점이 CLUSCAL-B1보다는 완화되었으나 완전히 해소되지는 못했다.

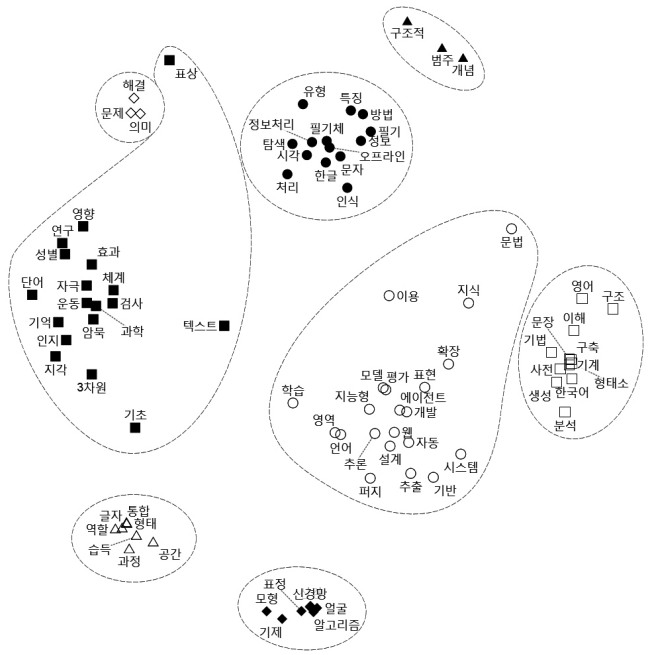
CLUSCAL-B1이나 CLUSCAL-B2에서 동일 군집에 속한 개체가 뭉치는 현상이 발생하는 것을 대비해서 군집의 범위를 미리 산출하여 개체의 이동거리를 확보하는 방법이 CLUSCAL-B3이다. 〈그림 16〉은 개체의 수가 50개인 AC50 데이터를 대상으로 CLUSCAL-B3를 적용한 결과이다. 이때 최대 이동거리 비율 α는 1.0으로 설정하였다. CLUSCAL-B1을 적용했던 〈그림 12〉와 마찬가지로 모든 군집의 소속 개체가 가까운 자리에 위치하고 있으므로 군집 구분선이 없더라도 소속 군집을 쉽게 알아볼 수 있다. 개체의 수가 이보다 훨씬 많은 CW85 데이터에 대해 CLUSCAL-B3를 적용한 결과인 〈그림 17〉에서는 CLUSCAL-B1이나 CLUSCAL-B2를 적용했던 결과와 달리 각 개체가 군집별로 덩어리로 뭉치지 않고 다른 군집과의 경계에 가까운 위치까지 폭넓게 배치되었다. 다만 위쪽 검은 동그라미 표식으로 표시된 군집에 속한 '방법'이나 아래쪽 빈 동그라미 표식으로 표시된 군집에 속한 '이용', 검은 사각형 표식으로 표시된 군집에 속한 '표상'과 같이 일부 개체는 소속 군집의 개체보다 이웃 군집에 속한 개체와 더 가깝게 표시되는 경우도 나타났다. 이상과 같이 CLUSCAL-B3를 적용해본 결과 개체가 뭉치는 현상은 해결한 반면에 개체의 수가 많은 데이터에서는 이웃 군집에 너무 가깝게 배치되는 개체가 일부 나타났다.

CLUSCAL-B4는 군집의 소속 개체가 기점에 몰리는 현상과 일부 개체가 소속 군집보다 다른 군집에 가깝게 배치되는 현상을 방지하도록 고안된 것이다. 개체의 수가 50개인 AC50 데이터를 대상으로 CLUSCAL-B4를 적용한 결과는 〈그림 18〉과 같다. 이때 α는 0.8로, β는 0.1로 지정하였다. 앞의 〈그림 12〉나 〈그림 16〉과 마찬가지로 모든 군집의 소속 개체가 가까운 자리에 위치하고 있으며 군집 구분선이 없더라도 소속 군집을 쉽게 알아볼 수 있다. CW85 데이터에 대해 CLUSCAL-B4를 적용한 결과인 〈그림 19〉에서는 각 개체가 타 군집의 개체와 판연하게 구분되도록 배치되었고 군집별로 중심 부근에 뭉치는 현상도 없었다. 〈그림 17〉의 CLUSCAL-B3 적용 결과에서 소속 군집보다 타 군집에 가깝게 배치되었던 '방법', '이용', '표상'의 세 개체는 모두 〈그림 19〉에서 소속 군집에 더 가까운 위치로 자리를 잡았다.

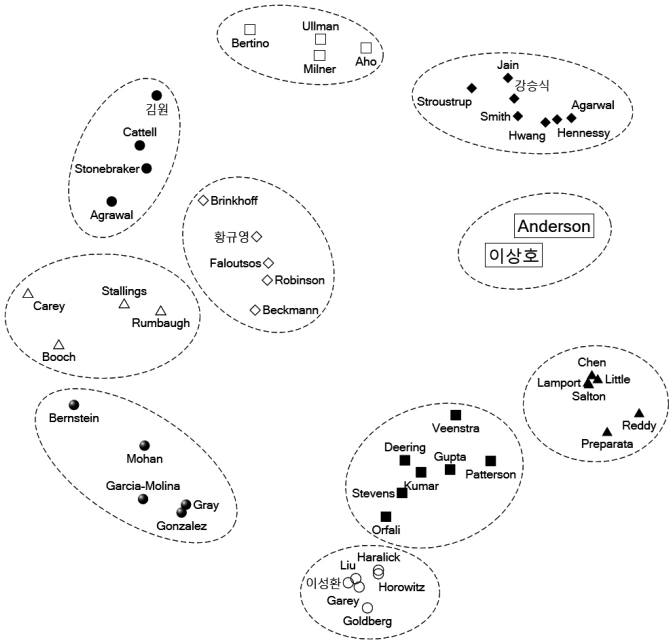
결국 AC50 데이터나 CW85 데이터 모두 CLUSCAL-B4를 적용한 결과에서 개체의 구분력과 군집의 구분력이 가장 뛰어난 것으로 나타났다. 다만, CLUSCAL 기법에 의한 MDS 지도에서 각 군집의 상대적인 위치가 PROXSCAL에 의한 MDS 지도에 나타난 것과 다소 달라지는 현상이 보였으며, CLUSCAL-A보다 CLUSCAL-B 기법에서 PROXSCAL과의 차이가 더욱 두드러졌다.



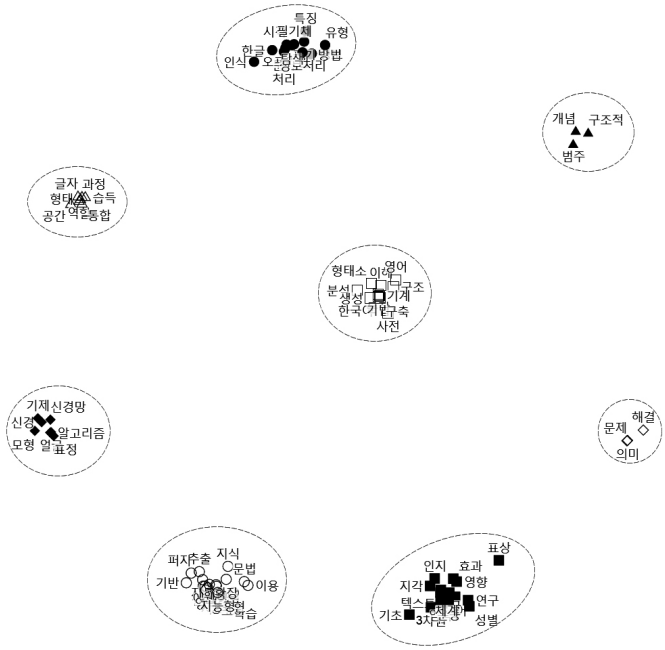
〈그림 10〉 CLUSCAL-A(AC50, Ward기법 10군집)



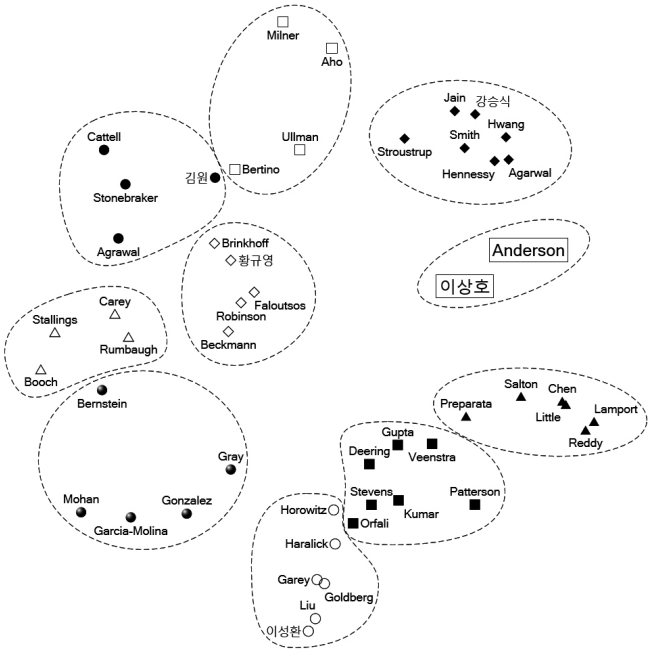
〈그림 11〉 CLUSCAL-A(CW85, 평균연결기법 8군집)



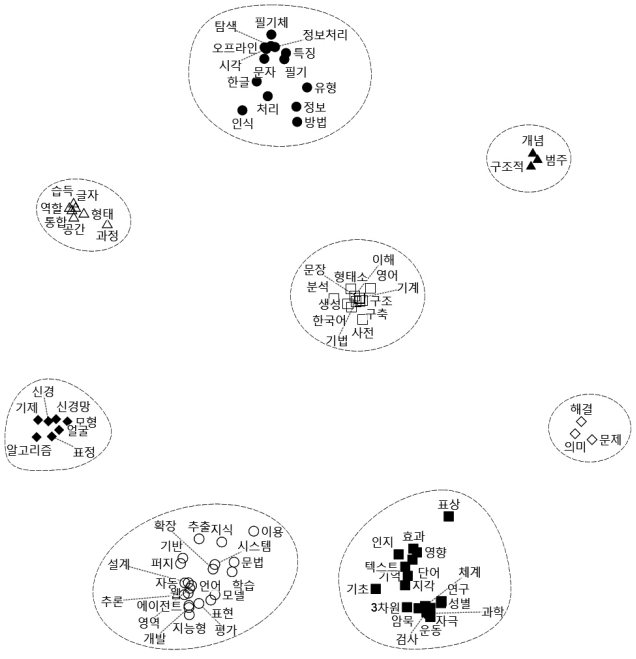
〈그림 12〉 CLUSCAL-B1(AC50, Ward기법 10군집)



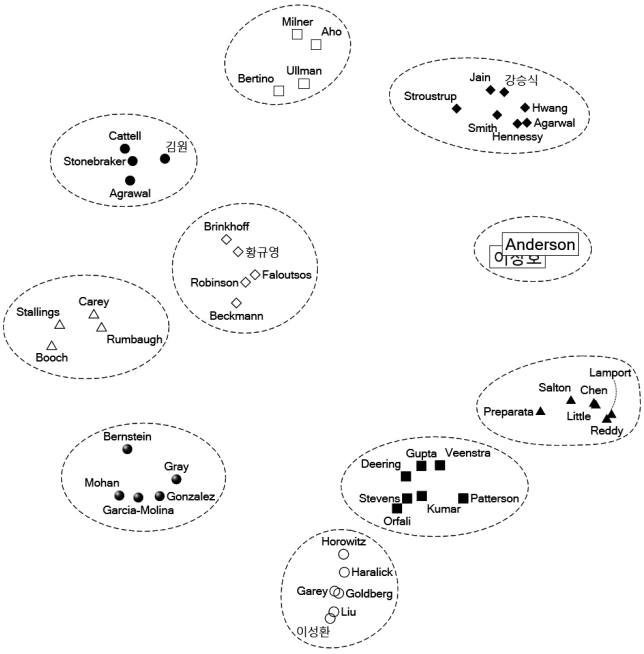
〈그림 13〉 CLUSCAL-B1(CW85, 평균연결기법 8군집)



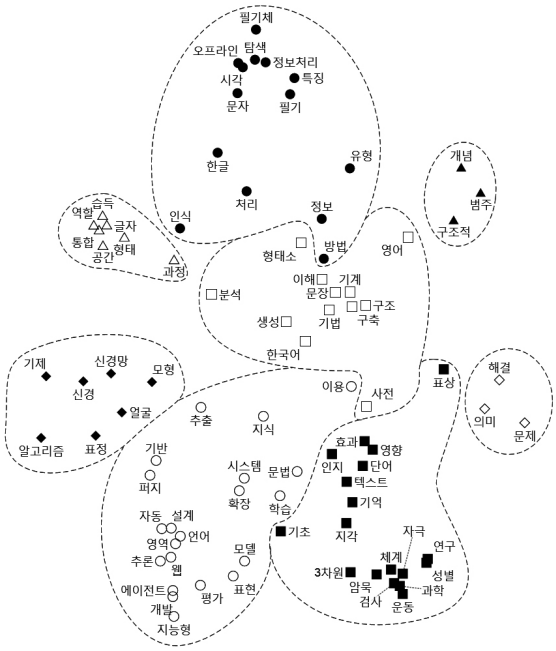
〈그림 14〉 CLUSCAL-B2(AC50, Ward기법 10군집)



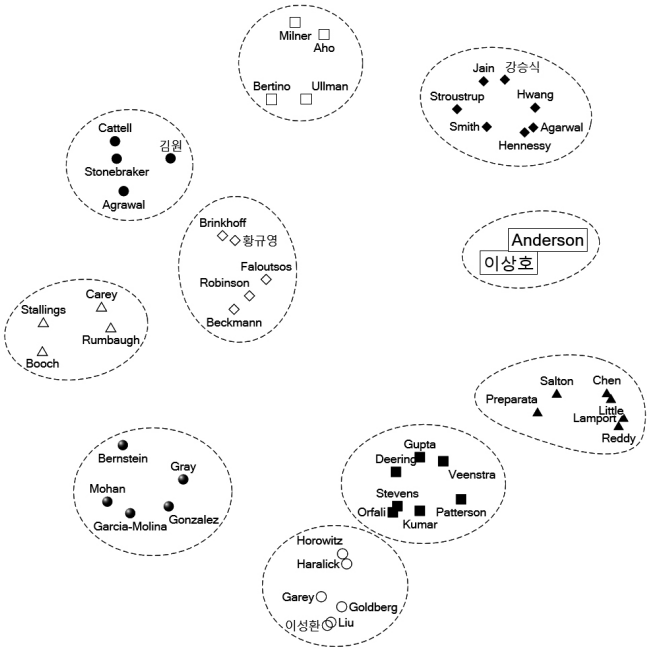
〈그림 15〉 CLUSCAL-B2(CW85, 평균연결기법 8군집)



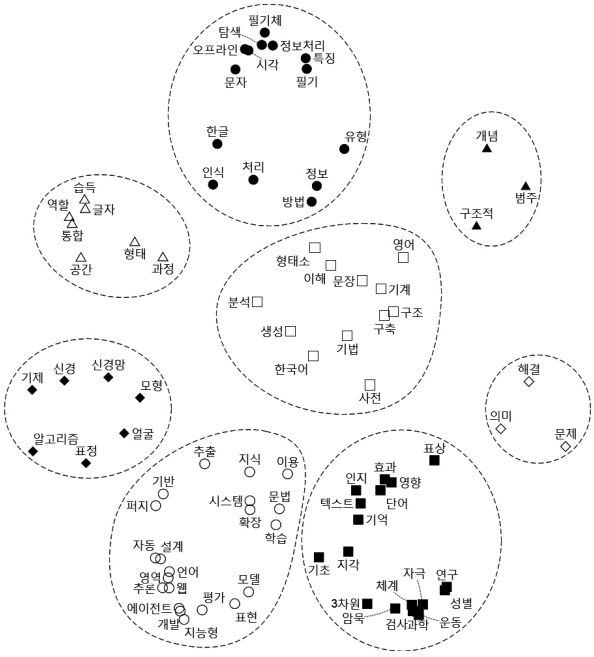
〈그림 16〉 CLUSCAL-B3(AC50, Ward기법 10군집)



〈그림 17〉 CLUSCAL-B3(CW85, 평균연결기법 8군집)



〈그림 18〉 CLUSCAL-B4(AC50, Ward기법 10군집)



〈그림 19〉 CLUSCAL-B4(CW85, 평균연결기법 8군집)

3.2 MDS 지도의 품질 측정

앞 절에서 시각적 평가를 수행한 결과 CLUSCAL 기법이 개체의 구분력과 군집의 구분력이 모두 조화롭게 높아서 MDS 지도에서 군집별로 개체를 적절히 표시하는 것으로 확인되었다. 그러나 CLUSCAL 기법에 의한 MDS 지도가 눈으로 살펴보기에는 좋을지 몰라도 산출된 좌표에 따른 개체 사이의 관계가 원래 개체 간의 관계를 지나치게 왜곡하는 것은 아닌지 검증할 필요가 있다.

이 절에서는 생성된 MDS 지도의 품질을 나타내는 지표로 근거리 결정계수 LRSQ(Local RSQ; 이재윤, 2007)와 WACS(Weighted Average Cluster Similarity; 정영미, 이재윤, 2001)를 사용하였다. 개체 간의 관계를 2차원 지도에 얼마나 제대로 표현하였는가를 평가하는 근거리 결정계수는 입력 데이터인 코사인 계수 행렬의 전체 개체 쌍 중에서 값이 높은 상위 1/3쌍에 대해서, 입력된 코사인 값과 도출된 MDS 지도에서의 유클리드 거리 사이의 피어슨 상관계수의 제곱을 산출하는 것이다(이재윤 2007). 지적 구조를 표현하는 다차원척도법은 가까운 개체 사이의 관계를 잘 반영하는 것이 중요하기 때문이다. LRSQ는 다음 공식으로 0에서 1 사이로 산출된다.

$$LRSQ = correlation(\text{입력 개체 간의 코사인 계수}, \text{개체 간의 유클리드 거리})^2$$

WACS 척도는 원래 클러스터링 결과의 평가지표로 제안된 것인데, 이 연구에서는 입력 자료인 코사인 계수 행렬로부터 도출된 군집과 MDS 지도에 나타난 개체 사이의 거리 행렬로부터 도출된 군집을 비교하는 용도로 사용하였다. 군집분석을 이용해서 다차원척도법의 결과가 입력 자료의 원래 관계를 얼마나 잘 보존하고 있는가를 판단할 수 있기 때문이다(Green & Rao, 1972, 33; Borg & Groenen, 2005, 108). WACS 척도는 비교 대상인 두 클러스터링 결과에서 군집별로 소속 개체가 얼마나 동일한가를 비교한 결과값이다. 입력 자료에서 도출된 군집을 CI, MDS 지도에서 도출된 군집을 CO라고 하면 개체 수가 n개이고 군집 수가 m개인 경우에 WACS는 다음 공식으로 산출하며 0에서 1 사이의 범위를 가진다.

$$WACS(CI, CO) = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{m}\frac{2\times|CI_i \cap CO_j|}{|CI_i|+|CO_j|}$$

저자동시인용 데이터인 AC50과 단어동시출현 데이터인 CW85를 대상으로 PROXSCAL과 CLUSCAL 기법들을 적용하여 도출된 MDS 지도를 평가한 결과를 〈표 4〉와 〈그림 20〉, 〈표 5〉와 〈그림 21〉에 각각 제시하였다.

〈표 4〉와 〈그림 20〉에서 AC50의 MDS 지도에 대한 평가결과를 보면, 개체 간 관계의 반영 정도인 LRSQ 기준으로는 CLUSCAL-B3가 0.2449로 가장 높고 CLUSCAL-B4가 0.2424로 근소한 차이로 두 번째로 나타났다. CLUSCAL-A는 CLUSCAL-B1을 제외한 CLUSCAL-B 계열 기법보다는 낮은 성능을 보였다. 기존 기법인 PROXSCAL은 LRSQ 값이 0.1819로 가장 낮았다. 군집 표현 능력을 측정한 WACS 기준으로는 CLUSCAL-B2, B3, B4의 세 방법은 모두 1.0으로 완벽하게 군집 구분이 표현되었으며 CLUSCAL-B1은 약간 낮

은 0.9662이고 CLUSCAL-A는 0.8058이었다. 기존 기법인 PROXSCAL은 WACS 값이 가장 좋은 CLUSCAL-B4, B3, B2의 성능에 비해서 절반이 조금 넘는 0.5637로 가장 낮았다. AC50 데이터에 대한 MDS 지도를 평가해본 결과 개체 간 관계의 반영 정도와 군집의 표현 정도에서 CLUSCAL-B3와 CLUSCAL-B4가 가장 우수한 것으로 나타났다.

〈표 5〉와 〈그림 21〉에서 CW85의 MDS 지도에 대한 평가결과를 보면, 개체 간 관계의 반영 정도인 LRSQ는 CLUSCAL-A가 0.2424로 월등하게 가장 높고 CLUSCAL-B4, B3, B2가 차례대로 0.17대의 수치를 보였으며, CLUSCAL-B1은 0.16대로 나타났다. 기존 기법인 PROXSCAL은 LRSQ 값이 0.1444로 가장 낮았다. 군집 표현 능력을 측정한 WACS 기준으로는 CLUSCAL-B1, B2, B4의 세 방법이 모두 1.0으로 완벽하게 군집 구분을 표현하였으며 CLUSCAL-B3는 0.6815로 큰 격차를 보였다. 기존 기법인 PROXSCAL은 WACS 값이 0.4673이어서 가장 좋은 CLUSCAL-B4, B2, B1의 성능에 비해 절반에도 못 미쳤다. 개체 수가 더 많은 CW85 데이터에 대한 MDS 지도를 평가해본 결과 개체 간 관계의 반영 정도는 CLUSCAL-A가 가장 뛰어났으며, 군집의 표현 정도에서는 CLUSCAL-B4, B2, B1 기법이 우수하였다.

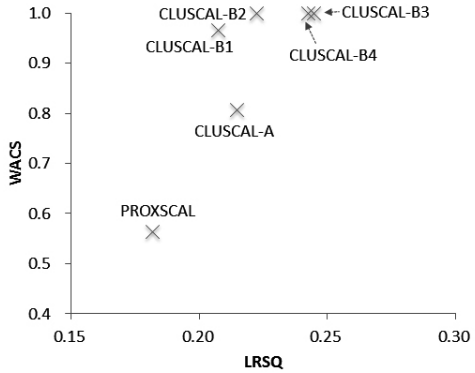
결과적으로 모든 경우에 PROXSCAL보다 CLUSCAL 기법들의 성능이 뚜렷하게 우세한 것으로 우세한 것으로 나타났으며, CLUSCAL 기법들 중에서는 CLUSCAL-B4 기법이 가장 좋은 성능을 보이는 것으로 나타났다. 다시 말해서 두 종류의 데이터에 대해서 전술한 두 가지 평가 척도를 적용해서 산출된 네 가지 평가 수치를 살펴보았을 때, 다른 CLUSCAL 기법들의 성능은 상대적인 편차가 있었다. 그러나 CLUSCAL-B4 기법은 두 종류의 데이터에 대한 평가 척도의 적용 결과에서 모두 1위 또는 2위로 나타나 고르게 좋은 성능을 보였다.

〈표 4〉 MDS 지도 생성 방법의 품질 비교(AC50, Ward기법 10군집)

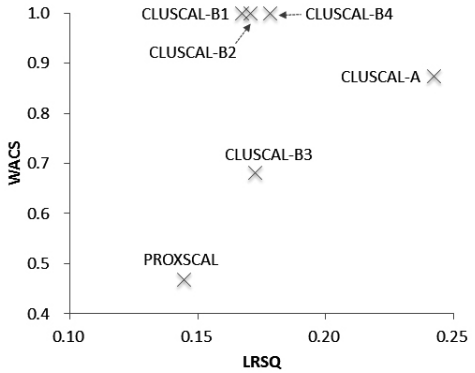
	PROXSCAL	CLUSCAL-A	CLUSCAL-B1	CLUSCAL-B2	CLUSCAL-B3	CLUSCAL-B4
LRSQ	0.1819 (-)	0.2146 (+18%)	0.2073 (+14%)	0.2225 (+22%)	0.2449 (+35%)	0.2424 (+33%)
WACS	0.5637 (-)	0.8058 (+43%)	0.9662 (+71%)	1.0000 (+77%)	1.0000 (+77%)	1.0000 (+77%)

〈표 5〉 MDS 지도 생성 방법의 품질 비교(CW85, 평균연결기법 8군집)

	PROXSCAL	CLUSCAL-A	CLUSCAL-B1	CLUSCAL-B2	CLUSCAL-B3	CLUSCAL-B4
LRSQ	0.1444 (-)	0.2424 (+68%)	0.1672 (+16%)	0.1707 (+18%)	0.1725 (+19%)	0.1784 (+24%)
WACS	0.4673 (-)	0.8734 (+87%)	1.0000 (+114%)	1.0000 (+114%)	0.6815 (+46%)	1.0000 (+114%)



〈그림 20〉 MDS 지도 생성 방법의 품질 비교 (AC50, Ward기법 10군집)



〈그림 21〉 MDS 지도 생성 방법의 품질 비교 (CW85, 평균연결기법 8군집)

4. 결 론

지적 구조 분석에 주로 사용되어온 다차원척도법과 군집분석의 결과가 2차원 지도 상에서 어울리지 않는 현상을 해결하기 위해서 군집 지향 척도법 CLUSCAL을 개발하였다. CLUSCAL은 PROXSCAL이나 ALSCAL과 같은 기존의 MDS 기법을 활용하되 사전에 군집분석을 수행하여 각 개체의 소속 군집 정보를 획득한 다음 이를 다차원척도법에 반영하도록 고려하였으며, 개체의 소속군집 정보를 반영하는 시기에 따라서 기존 MDS 기법을 수행하기 전에 반영하는 CLUSCAL-A와 기존 MDS 기법을 수행한 이후에 반영하는 CLUSCAL-B 계열 기법 네 가지를 포함하여 총 다섯 종류를 제안하였다.

CLUSCAL-A 기법은 개체 간 근접성 행렬 대신 차원축소된 개체-군집 간 코사인 계수 행렬을 입력하여 다차원척도법을 수행하는 방법으로서 비교적 절차가 단순하다는 장점이 있다. 50명의 저자에 대한 저자동시인용분석 결과와 85개의 단어에 대한 동시출현분석 결과를 MDS 지도로 나타내본 결과 시각적 평가에서 일부 개체가 소속 군집이 아닌 다른 군집과 더 가까운 위치에 배치되는 현상이 나타났으며, 군집의 구분이 완벽하지 못하다는 것은 WACS 지표로도 확인되었다. 다만 표현해야 하는 개체의 수가 많을 경우에는 개체 간 관계를 가장 원 데이터에 가깝게 표현하는 것으로 나타났다.

CLUSCAL-B 계열 기법 네 가지는 일단 군집분석을 수행하고 군집 간 코사인 계수를 이용해서 군집의 MDS 지도를 산출한 이후 각 군집의 좌표를 기초로 하여 군집에 속한 개체의 좌표를 결정하도록 고안되었다. 이때 타 군집과의 코사인 계수, 군집의 반경, 군집의 상대적인 크기, 개체의 상대적인 이동 거리 등을 고려하고 최소 이동거리 비율 지정 파라미터로 α, 최대 이동거리 비율 파라미터로 β를 지정하게 하였다. 시각적 평가와 품질평가지표 산출결과에서 모두 CLUSCAL-B4로 도출한 MDS 지도가 가장 안정적이고 품질이 높은 것으로 확인되었다. CLUSCAL-B4는 개체 간 관계를 MDS 지도에 적절히 반영하였으며 군집을 구분해주는 능

력도 매우 뛰어났다. 따라서 어떤 경우에라도 안정적인 품질의 MDS 지도를 얻기 위해서는 CLUSCAL-B4 기법을 사용하는 것이 바람직하다고 판단된다. 다만 표현해야 할 개체의 수가 CW85와 같이 많은 경우에는 개체 간 관계를 잘 표현하는 CLUSCAL-A의 사용도 고려할 필요가 있다.

CLUSCAL 기법이 MDS 지도에서 개체와 군집을 적절히 식별할 수 있도록 보장하지만, CLUSCAL 기법에 의한 MDS 지도에서 각 군집의 상대적인 위치는 PROXSCAL에 의한 MDS 지도에 나타난 것과 다소 달라졌으며, CLUSCAL-A보다 CLUSCAL-B 기법에서 PROXSCAL과의 차이가 더욱 두드러졌다. 물론 PROXSCAL에 의한 MDS 지도가 최적의 배치라고 할 수는 없지만, 지도의 해석이 달라질 가능성이 있다. 경우에 따라서는 기존 기법에 의한 군집 배치를 더 선호할 수도 있으므로 이를 감안하여 CLUSCAL의 추가 개선도 필요하다.

이 연구에서 제안된 CLUSCAL 기법은 지적 구조 분석 과정에서 50개 이상의 개체를 대상으로 다차원척도법과 군집분석을 함께 사용할 때 군집과 개체를 잘 표현할 수 있는 것으로 확인되었다. 다차원척도법과 군집분석은 지적 구조 분석 이외에 사회과학의 여러 분야는 물론이고 자연과학 분야에서도 빈번히 사용되는 기법이므로, 향후 다양한 분야에서 CLUSCAL을 활용하여 개체 간의 관계와 구조를 더욱 정확하게 표현할 수 있을 것으로 기대된다.

참 고 문 헌

곽선영, 정은경 (2012). 복수저자기반 동시인용분석을 활용한 지적구조 분석. 정보관리학회지, 29(1), 115-134. http://dx.doi.org/10.3743/KOSIM.2012.29.1.115

김판준 (2011). 저자 프로파일링 기법을 이용한 국내 독서 연구 영역 분석. 한국비블리아학회지, 22(4), 21-44.

김희전, 조현양 (2010). 저자동시인용분석과 저자서지결합분석에 의한 지적 구조 분석. 정보관리학회지, 27(3), 283-306. http://dx.doi.org/10.3743/KOSIM.2010.27.3.283

문주영 (2011). 2000년대 비서학연구의 저자동시인용분석. 비서학논총, 20(1), 25-44.

박재신, 정영미 (2010). 지구적 환경문제 해결을 위한 학술활동과 환경운동 경향 연구. 정보관리학회지, 27(3), 83-102. http://dx.doi.org/10.3743/KOSIM.2010.27.3.083

유종덕, 최은주 (2011). 저자프로파일링분석과 저자동시인용분석의 유용성 비교 검증. 정보관리학회지, 28(1), 123-144. http://dx.doi.org/10.3743/KOSIM.2011.28.1.123

이은숙, 정영미 (2002). 복수저자를 고려한 동시인용분석 연구: 정보학과 컴퓨터과학을 대상으로. 지식처리연구, 3(2), 1-26. Retrieved from http://jkpm.yonsei.ac.kr/fulltext/v3n2a1.pdf

이재윤 (2007). 지적 구조 분석을 위한 MDS 지도 작성 방식의 비교 분석. 한국문헌정보학회지, 41(2), 335-357. http://dx.doi.org/10.4275/KSLIS.2007.41.2.335

이재윤, 김희전, 유종덕 (2010). 저자프로파일링과 요인분석을 이용한 국내 주거학 분야의 지적 구조 분석. 한국문헌정보학회지, 44(2), 285-308. http://dx.doi.org/10.4275/KSLIS.2010.44.2.285

이재윤, 정주희 (2006). 연구자 소속과 표제어 분석을 통한 국내 인지과학 분야의 학제적 구조 파악. 제13회 한국정보관리학회 학술대회 논문집, 127-134.

정영미, 이재윤 (2001). 지식 분류의 자동화를 위한 클러스터링 모형 연구. 정보관리학회지, 18(2), 203-230.

조선례, 이재윤 (2012). 약학 분야 학술정보서비스를 위한 학술지 동시인용 분석. 정보관리연구, 43(1), 159-185. http://dx.doi.org/10.1633/JIM.2012.43.1.159

조재인 (2011). 네트워크 텍스트 분석을 통한 문헌정보학 최근 연구 동향 분석. 정보관리학회지, 28(4), 65-83. http://dx.doi.org/10.3743/KOSIM.2011.28.4.065

Borg, I., & Groenen, P. J. F. (2005). Modern multidimensional scaling: Theory and Applications (2nd ed). New York: Springer Science+Business Media, Inc.

Börner, K., Chen, C., & Boyack, K. (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.

Choi, Sanghee (2011). An informetric analysis on intellectual structures with multiple features of academic library research papers. Journal of the Korean Society for Information Management, 28(2), 65-78. http://dx.doi.org/10.3743/KOSIM.2011.28.2.065

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407. http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6〈391::AID-ASI1〉3.0.CO;2-9

Green, P. E., & Rao, V. (1972). Applied multidimensional scaling. Hinsdal, IL: Dryden.

Lee, Jae Yun, & Choi, Sanghee (2011). Intellectual structure and infrastructure of informetrics. Journal of the Korean Society for Information Management, 28(2), 11-36. http://dx.doi.org/10.3743/KOSIM.2011.28.2.011

Park, Myung-Kyu, & Kim, Heejung (2011). A bibliometric analysis of the literature on information literacy. Journal of the Korean Society for Information Management, 28(2), 53-63. http://dx.doi.org/10.3743/KOSIM.2011.28.2.053

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between publications. Journal of the American Society for Information Science, 24(4), 265-269.

White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual

structure. Journal of the American Society for Information Science, 32(3), 163-171. http://dx.doi.org/10.1002/asi.4630320302

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cho, Jane (2011). A study for research area of library and information science by network text analysis. Journal of the Korean Society for Information Management, 28(4), 65-83. http://dx.doi.org/10.3743/KOSIM.2011.28.4.065

Chung, Young Mee, & Lee, Jae Yun (2001). Development of a clustering model for automatic knowledge classification. Journal of the Korean Society for Information Management, 18(2), 203-230.

Joe, Seon-Rye, & Lee, Jae Yun (2012). Journal co-citation analysis for library services in pharmaceutics. Journal of Information Management, 43(1), 159-185. http://dx.doi.org/10.1633/JIM.2012.43.1.159

Kim, Hee-Jeon, &, Cho, Hyun-Yang (2010). A study on intellectual structure using author co-citation analysis and author bibliographic coupling analysis in the field of social welfare science. Journal of the Korean Society for Information Management, 27(3), 283-306. http://dx.doi.org/10.3743/KOSIM.2010.27.3.283

Kwak, Sun-Young, & Chung, Eunkyung (2012). Domain analysis on economics by utilizing cocitation analysis of multiple authorship. Journal of the Korean Society for Information Management, 29(1), 115-134. http://dx.doi.org/10.3743/KOSIM.2012.29.1.115

Kim, Pan Jun (2011). Domain analysis of reading research in Korea using author profiling. Journal of the Korean Biblia Society for Library and Information Science, 22(4), 21-44.

Lee, Eun Suk, & Chung, Young Mee (2002). A co-citation analysis of multiple authorship in the subject fields of information science and computer science. Journal of Knowledge Processing, 3(2), 1-26. Retrieved from http://jkpm.yonsei.ac.kr/fulltext/v3n2a1.pdf

Lee, Jae Yun (2007). A comparison analysis of various approaches to multidimensional scaling in mapping a knowledge domain's intellectual structure. Journal of the Korean Society for Library and Information Science, 41(2), 335-357. http://dx.doi.org/10.4275/KSLIS.2007.41.2.335

Lee, Jae Yun, & Jung, Ju Hee (2006). Examining the interdisciplinary structure of Korean cognitive science through analyzing author affiliations and title words. Proceedings of the 13th Annual Conference of Korean Society for Information Management, 127-134.

Lee, Jae Yun, Kim, Hee-Jeon, & Ryoo, Jong-Duk (2010). Examining the intellectual structure of housing studies in Korea with text mining and factor Analysis. Journal of the Korean Society for Library and Information Science, 44(2), 285-308. http://dx.doi.org/10.4275/KSLIS.2010.44.2.285

Moon, Ju Young (2011). Study on intellectual structure of secretarial studies using author co-citation analysis. Journal of Secretarial Science, 20(1), 25-44.

Park, Jae-Shin, & Chung, Young Mee (2010). An informetric study on academic activities and environmental movements in solving global environmental problems. Journal of the Korean Society for Information Management, 27(3), 83-102. http://dx.doi.org/10.3743/KOSIM.2010.27.3.083

Ryoo, Jong-duk, & Choi, Eun-Ju (2011). A comparison test on the potential utility between Author Profiling Analysis (APA) and Author Co-Citation Analysis (ACA). Journal of the Korean Society for Information Management, 28(1), 123-144. http://dx.doi.org/10.3743/KOSIM.2011.28.1.123