지적 구조 분석을 위한 MDS 지도 작성 방식의 비교 분석

A Comparison Analysis of Various Approaches to Multidimensional Scaling in Mapping a Knowledge Domain's Intellectual Structure

이 재 윤(Jae-Yun Lee)*

목 차

초 록

다차원척도법(MDS)은 지적 구조의 시각화를 위해서 오랫동안 사용되어 왔다. 그러나 MDS는 지적 구조를 시각적으로 표현하는데 있어서 세부 구조의 표현력이 취약하다는 약점을 가지고 있다. 이 연구에서는 상관계수 행렬의 가공 방식과 MDS 알고리즘을 조합한 여섯 가지 MDS 지도 작성 방식을 파악한 다음, 실제 지적 구조 데이터에 적용하여 비교해보았다. 실험 결과에서 가장 나쁜 방식으로 파악된 것은 가장 널리 사용되고 있는 방식으로서, 상관계수행렬로부터 유클리드 거리를 산출한 후 ALSCAL 알고리즘으로 MDS 지도를 작성하는 방식인 것으로 나타났다. 반면에 가장 좋은 방식은 상관계수를 z점수로 표준화하여 유클리드 거리를 산출한 후 PROXSCAL 알고리즘를 사용하는 방식이었다. 결론적으로 MDS 처리 과정을 주의깊게 구성한다면 더 구체적이고 명확한 지적 구조를 파악할 수 있음이 확인되었다.

ABSTRACT

There has been many studies representing intellectual structures with multidimensional scaling(MDS). However MDS configuration is limited in representing local details and explicit structures. In this paper, we identified two components of MDS mapping approach: one is MDS algorithm and the other is preparation of data matrix. Various combinations of the two components of MDS mapping are compared through some measures of fit. It is revealed that the conventional approach composed of ALSCAL algorithm and Euclidean distance matrix calculated from Pearson's correlation matrix is the worst of the compared MDS mapping approaches. Otherwise the best approach to make MDS map is composed of PROXSCAL algorithm and z-scored Euclidean distance matrix calculated from Pearson's correlation matrix. These results suggest that we could obtain more detailed and explicit map of a knowledge domain through careful considerations on the process of MDS mapping.

키워드: 지적구조 분석, 다차원척도법, MDS 지도, 군집분석, 계량서지학
Knowledge Domain Analysis, Multidimensional Scaling, MDS Map, Cluster Analysis, Bibliometrics

* 경기대학교 인문대학 문헌정보학전공 조교수(memexlee@kgu.ac.kr)
논문접수일자 2007년 5월 22일
게재확정일자 2007년 6월 10일

1. 서 론

학문 영역 등의 지적 구조를 분석하기 위해서는 인용 정보, 서지 정보, 웹 링크 정보와 같은 단서 정보를 준비한 후 이를 통계적인 다변량 분석 기법으로 처리하는 것이 일반적이다. 계량서지적 분석에 흔히 응용되는 다변량 분석 기법으로는 다차원척도법, 요인분석, 주성분분석, 패스파인더 네트워크, SOM(self-organizing maps), 군집분석 등이 있다. 이중에서 군집분석을 제외한 나머지 기법을 Börner 등(2003)은 차원 축소(dimensionality reduction) 기법이라고 부르기도 하였다. 이와 같은 차원 축소 기법의 역할에 대해서 Börner 등(2003)은 (1) 다변량 자료를 2차원 표면에 표시해야 하며 (2) 대량의 자료를 한정된 공간과 해상도의 제약 하에서 표시해야 하기 때문이라고 설명하였다. 이는 지적 구조 분석에서 차원 축소 기법의 역할을 시각적 표현 도구로 간주한 것이다.

지적 구조 분석을 위한 다변량 분석 기법 중에서 가장 널리 사용되고 있는 것이 바로 다차원척도법(Multidimensional Scaling; 이하 MDS로 약칭)과 군집분석이다. White와 Griffith (1981)가 저자동시인용 분석을 처음 제안할 때에는 이 두 분석기법 이외에 요인분석도 수행하였다. 그러나 요인분석의 분석 결과가 군집분석과 어느 정도 유사할 뿐만 아니라, 군집분석이 MDS 지도와 결합하여 결과를 제시하기가 더 용이하기 때문에 요인분석은 이후의 저자동시인용 분석 연구에서 생략되는 경우가 많았다.

특히 MDS는 최근까지도 지적 구조 분석 연구(이수범, 권영순 2006; 이재윤, 문정순, 김희정 2007; Moya-Anegón et al. 2006; Schildt & Mattsson 2006; Vaughan & You 2006)에서 널리 사용되고 있을 뿐만 아니라, 문헌집단 브라우징 인터페이스(Chalmers, & Chitson 1992; McQuaid et al. 1999)나 온라인 목록 주제 브라우징 인터페이스(Herrero-Solana et al. 2006), 또는 문헌내 브라우징 인터페이스(한승희, 이재윤 2004) 등과 같은 용도로도 응용되고 있다.

MDS를 위한 알고리즘으로는 MDSCAL, KYST, POLYCON, ALSCAL, PROXSCAL 등을 비롯하여 여러 가지가 개발되어 있다. 이중에서 White와 Griffith(1981)가 지적 구조 분석에 처음 사용한 알고리즘은 MDS의 선구자인 Kruskal(1964)이 개발한 MDSCAL이었다. 그러나 대표적인 사회과학 통계처리 패키지인 SPSS가 PC용으로 개발되어 대중화된 이후에는 SPSS에서 지원되는 ALSCAL 알고리즘을 주로 사용하게 되었다. 근래에는 SPSS에서 지원하는 다른 MDS 알고리즘인 PROXSCAL을 사용한 연구(Leydesdorff & Vaughan 2006; Vaughan & You 2006)도 등장하였다.

이처럼 1981년 이후 오랜 기간 동안 지적 구조 분석을 위해 사용되어온 MDS에 대해서 최근 시각적인 표현도구로서의 약점이 지적되기 시작하였다. 주로 지적되는 사항은 도출된 차원을 해석할 수 있는 간단한 규칙이 없다는 점, 그리고 국지적인 세부구조를 표현하는 데 한계가 있다는 점 등이다(Börner et al. 2003; Chen 2006).

MDS로 국지적인 세부 구조가 부정확하게 표현될 경우에 나타나는 대표적인 문제는 군집분석 결과를 MDS 지도에 영역으로 표시할 때 구불구불하게 그려지거나 심지어는 군집끼리 겹쳐지는 현상이다. 이는 특히 지도에 표현할

저자나 웹 사이트와 같은 노드가 많을 경우에 흔히 나타나는데, 군집의 영역이 서로 배타적으로 표시되지 않기 때문에 지적 구조의 파악이 제대로 이루어지지 않게 된다.

이런 지적과 문제점이 있음에도 불구하고 MDS를 사용하여 저자동시인용이나 동시링크 분석 등을 수행하는 지적구조 분석 연구에서는 SPSS를 이용한 MDS 분석과 그 결과로 생성되는 MDS 지도의 품질에 대해서 의문을 제기하거나 별도의 검증을 시도한 사례는 없었다.

따라서 이 연구에서는 우선 지적 구조 분석을 위한 MDS의 개념과 분석 절차에 대해서 살펴본 후, 지적 구조 분석을 위해 도출된 MDS 지도를 평가할 수 있는 새로운 합치도 지수를 제안하고, 전통적인 MDS 지도 작성 방식 이외에 SPSS를 이용하여 수행할 수 있는 다양한 작성 방식을 제시하였다. 그리고 새로운 합치도 지수와 함께 MDS 지도의 품질을 측정하는 여러 방법을 제안한 다음 이를 이용하여 여러 방식으로 생성된 MDS 지도의 품질을 비교하는 실험을 통해서 지적 구조 분석을 위한 최적의 MDS 지도 작성 방식을 파악하였다.

2. MDS의 개념과 절차

넓게 보면 자료로부터 도출된 잠재적 차원으로 이루어진 공간에 주어진 자료를 시각적으로 표현하는 방법을 모두 MDS로 간주할 수도 있다(박광배 2000, 16). 이런 관점에서는 군집분석, 요인분석, 대응일치분석(correspondence analysis), 패스파인더 네트워크 스케일링(pathfinder network scaling)과 같은 다변량 자료를 대상으로 시각적인 해석을 돕는 여러 기법이 모두 포함될 수 있다.

좁은 의미의 MDS는 대상 자료간의 유사성(혹은 비유사성) 측정 결과를 저차원 공간에 점간의 거리로 나타내는 기법이다(Borg & Groenen 2005, 3). Cox와 Cox는 일반적으로 유클리드 공간에 대상을 점으로 나타내고 대상 사이의 비유사성에 가급적 부합되도록 점간의 거리를 설정하는 기법이 좁은 의미의 MDS라고 더 제한적으로 정의하였다(Cox & Cox 2001, 1).

MDS는 입력변수의 유형에 따라서 계량적 MDS(metric MDS)와 비계량적 MDS(nonmetric MDS)로 나뉘어진다. 계량적 MDS는 입력 자료를 비율 변수로 간주하여 처리하는 경우로서 ratio MDS라고 부르기도 한다. 비계량적 MDS는 입력 자료를 서열 변수로 간주하여 처리하는 경우로서 ordinal MDS라고도 부른다(Borg & Groenen 2005). 계량적 MDS에서는 대상 자료간 비유사성의 크기에 비례하도록 저차원 공간에서의 거리를 산출하는 것이 목표이다. 반면에 비계량적 MDS에서는 대상 자료간 비유사성의 서열이 유지되도록 저차원 공간에서의 거리를 산출하는 것을 추구한다.

일반적으로 MDS의 분석 대상이 되는 조사 자료나 심리 측정 자료는 비율 변수보다 서열 변수인 경우가 많기 때문에 비계량적 MDS가 많이 사용된다. 지적 구조 분석에 흔히 활용되는 동시인용 자료는 비율 변수로 간주할 수도 있지만, 동시출현빈도 행렬로부터 산출된 상관계수는 더하거나 빼는 연산이 불가능하고 순위만 의미가 있는 서열 변수에 해당한다. White와 Griffith(1981)의 제안 이후 대부분의 지적 구조 분석 연구에서는 상관계수 행렬을 MDS

의 입력 자료로 삼고 있으므로 지적 구조 규명을 위해서는 비계량적 MDS를 사용해야 한다.

MDS를 수행하는 알고리즘은 여러 가지가 존재하지만 SPSS의 이용이 확산되면서 지적 구조의 분석 용도로는 ALSCAL 알고리즘이 널리 사용되고 있다. 비계량적 MDS를 수행하는 ALSCAL 알고리즘의 처리 절차를 간단히 정리하면 다음과 같다(박광배 2000; Borg & Groenen 2005).

① 원래의 근접성 값에 합산상수라는 특정한 값을 일률적으로 더해서 거리모형에 부합되도록 변환한다.

② 변환된 근접성 값으로 고전척도법을 수행하여 각 변수의 초기 좌표와 시초거리(initial distance)를 산출한다.

③ 변환된 근접성 값이 시초거리와 같은 단위로 표현될 수 있도록 일종의 회귀분석을 통해서 상위(disparity)라는 값으로 단조변환(monotone transformation) 한다. 이 과정을 최적화 절차라고 부르며 이때 적용되는 회귀분석 기법이 단조 회귀(monotone regression)이다. 만약 단조 회귀 대신 일반적인 선형 회귀(linear regression)를 적용하면 계량적 MDS가 된다.

④ 상위와 시초거리의 일치 정도를 나타내는 스트레스 함수값을 측정한다. 이때 스트레스 함수값이 미리 설정한 기준 이상이면 편미분에 의해서 각 변수의 새로운 좌표를 산출하는 과정을 반복하고, 기준 미만이면 현재 좌표에 의한 배치도를 그리고 종료한다.

3. 국지적 구조 표현 능력을 고려한 합치도 지수 개발

3.1 국지적 구조 표현 능력의 중요성

지적 구조 분석에 흔히 사용되고 있는 MDS 알고리즘인 ALSCAL은 통계처리 프로그램인 SPSS의 ALSCAL 프로시저로 구현되어 있다. SPSS 패키지가 널리 보급되었고 사용하기 쉬우므로 ALSCAL 프로시저는 지적 구조 분석뿐만 아니라 일반적인 다차원척도분석을 위해서도 빈번히 사용되고 있다.

그런데 ALSCAL 알고리즘은 비유사성(거리)이 큰 관계를 비유사성이 작은 관계보다 더 잘 표현하는 특성을 가졌다(Borg & Groenen 2005, 551). 이는 관계가 먼 변수간의 거리로 구성되는 전체적인 배치 상태는 잘 나타내는 반면에, 관계가 가까운 변수간의 거리로 구성되는 국지적인 배치 상태는 상대적으로 잘 나타내지 못함을 뜻한다. 그 이유는 상위와 시초거리의 일치 정도를 나타내는 스트레스 함수 공식에서 여타 MDS 알고리즘은 상위와 거리의 차이를 제곱하여 합산하는 반면에, ALSCAL 알고리즘은 상위의 제곱과 거리의 제곱의 차이를 다시 제곱해서 합산하기 때문이다(Borg & Groenen 2005, 551). 이와 같은 ALSCAL 알고리즘의 스트레스 함수를 특별히 S-STRESS라고 부른다(Takane, Young, & De Leeuw, 1977). 큰 값과 작은 값의 차이는 각각을 제곱하여 산출하면 더 커지기 마련이므로 S-STRESS는 비유사성이 작은 관계(가까운 관계)의 차이보다는 큰 관계(먼 관계)의 차이를 더 민감하게 반영한다. 따라서 ALSCAL 알고리즘은 국지

적인 변수 배치가 상대적으로 부정확할 수밖에 없다.

정보 시각화를 통한 정보 분석 분야의 권위자인 C. Chen은 최근 주목받고 있는 시각적 표현 기법인 패스파인더 네트워크가 MDS에 비해서 좋은 점은 국지적 구조를 더 정확하게 표현하는 능력이라고 설명하였다(Chen 2006, 51). 이처럼 MDS는 국지적 구조의 표현 능력이 떨어지는 것이 큰 단점으로 인식되고 있는데, MDS 중에서도 ALSCAL 알고리즘은 그 정도가 더 심하다고 할 수 있다.

지적 구조 분석에 있어서 가까운 변수간의 관계를 잘 나타내는 능력이 얼마나 중요한가를 구체적으로 확인하기 위해서 인위적으로 가상의 변수 9개로 구성된 배치도를 만들어보았다. 〈표 1〉에서 첫 번째 x, y 좌표는 변수간의 관계가 이상적으로 반영된 경우이고, 두 번째 x, y 좌표는 변수간 거리 중에서 먼 거리가 왜곡되어 배치된 경우이며, 세 번째 x, y 좌표는 가까운 거리 일부가 왜곡되어 배치된 경우이다. 세 가지 경우를 2차원 지도로 나타낸 그림이 각각 〈그림 1〉, 〈그림 2〉, 〈그림 3〉이다. 9가지 변수 사이의 비유사도를 왜곡이 없는 이상적인 경우의 좌표를 기준으로 산출하면 〈표 2〉와 같은 유클리드 거리 행렬로 나타낼 수 있다. 이런 원래의 변수간 거리와 함께 좌표가 왜곡된 경우의 변수간 거리를 비교하는 산점도를 그린 것이 〈그림 4〉와 〈그림 5〉이다. 거리변환 산점도는 점의 배열이 일직선에 가까울수록 올바르게 변환되었음을 나타낸다.

먼 거리가 왜곡된 경우는 변수 9개 중에서 F, G, H, I의 네 변수의 위치에 오류가 나타나서 F와 G가 속한 군집3과 H와 I가 속한 군집4의 위치에 〈그림 2〉와 같은 큰 변화가 나타난 상태이다. 원래 거리와 왜곡된 거리의 변환 산점도를 그린 〈그림 4〉를 보면 원래 거리가 가까운 경우는 변화가 없으나 먼 거리가 상당수 왜곡되어 있음이 드러난다. 그러나 이런 변화가 〈그림 2〉의 2차원 지도에서 전체적인 구조를 변화시키지는 못한다. 각 변수의 소속 군집은 변화하지 않아서 군집1이 가장 크다는 사실은 그대로이며 전반적인 군집의 배치 상태도 달라졌다는 것을 인식하기 어렵다.

반면에 가까운 거리가 왜곡된 경우는 변수 9개 중에서 B와 C의 두 변수의 위치만 약간 왜곡된 경우이며, 〈그림 5〉의 거리 변환 산점도를 살펴보아도 가까운 거리와 중간 거리 일부만 조금 왜곡된 것으로 나타난다. 그럼에도 불구하고 〈그림 3〉과 같이 변수 B의 소속 군집이 달라지면서 가장 큰 군집이 군집1에서 군집2로 바뀌게 된다. 먼 거리가 대폭 변화한 〈그림 2〉와 달리 가까운 거리가 소폭 변화한 〈그림 3〉에서 오히려 전반적인 구조가 달라지는 결과가 된다. 〈그림 2〉와 〈그림 3〉이 가상의 지적 구조를 해석하기 위한 MDS 지도라고 가정한다면 군집 구조가 바뀌는 〈그림 2〉보다는 〈그림 3〉이 원래의 구조를 더 잘 반영한다고 말할 수 있다.

이와 같이 지적 구조 분석에 있어서 이상적인 구조를 파악하기 위해서는 MDS 지도 상에서 먼 변수간의 관계보다 가까운 변수간의 관계를 제대로 반영하는 것이 더 중요하다.

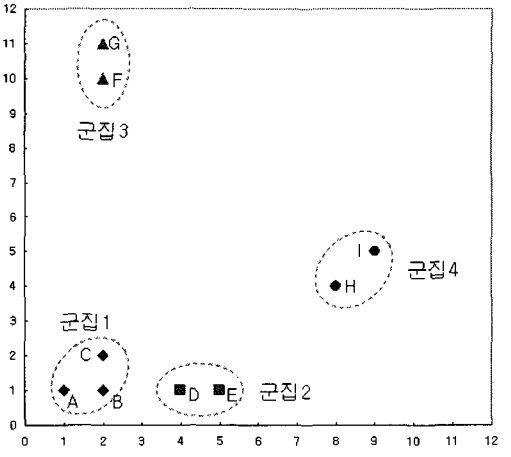
〈표 1〉 가상 변수 9개의 소속 군집과 2차원 지도상 좌표

변수	소속 군집	원래 좌표(이상적 배치)		먼 거리가 왜곡된 경우		가까운 거리가 왜곡된 경우	
		x좌표	y좌표	x좌표	y좌표	x좌표	y좌표
A	군집 1	1	1	1	1	1	1
B	군집 1	2	1	2	1	3	1
C	군집 1	2	2	2	2	1	2
D	군집 2	4	1	4	1	4	1
E	군집 2	5	1	5	1	5	1
F	군집 3	2	10	2	6	2	10
G	군집 3	2	11	2	7	2	11
H	군집 4	8	4	11	4	8	4
I	군집 4	9	5	10	5	9	5

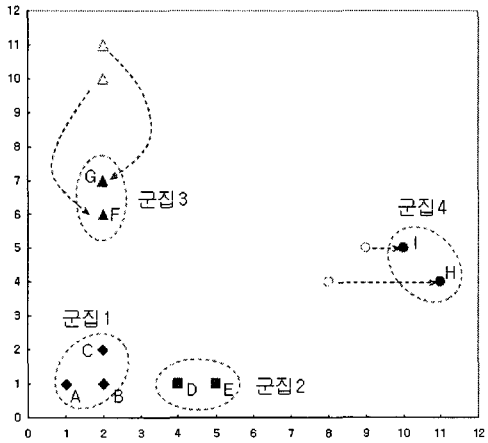
* 음영 부분은 위치가 왜곡된 변수의 좌표

〈표 2〉 9가지 가상 변수의 원래 좌표에 의한 유클리드 거리 행렬

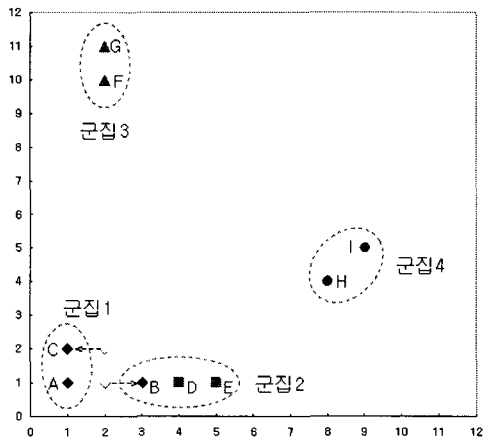
	A	B	C	D	E	F	G	H	I
A	0.0	1.0	1.4	3.0	4.0	9.1	10.0	7.6	8.9
B	1.0	0.0	1.0	2.0	3.0	9.0	10.0	6.7	8.1
C	1.4	1.0	0.0	2.2	3.2	8.0	9.0	6.3	7.6
D	3.0	2.0	2.2	0.0	1.0	9.2	10.2	5.0	6.4
E	4.0	3.0	3.2	1.0	0.0	9.5	10.4	4.2	5.7
F	9.1	9.0	8.0	9.2	9.5	0.0	1.0	8.5	8.6
G	10.0	10.0	9.0	10.2	10.4	1.0	0.0	9.2	9.2
H	7.6	6.7	6.3	5.0	4.2	8.5	9.2	0.0	1.4
I	8.9	8.1	7.6	6.4	5.7	8.6	9.2	1.4	0.0



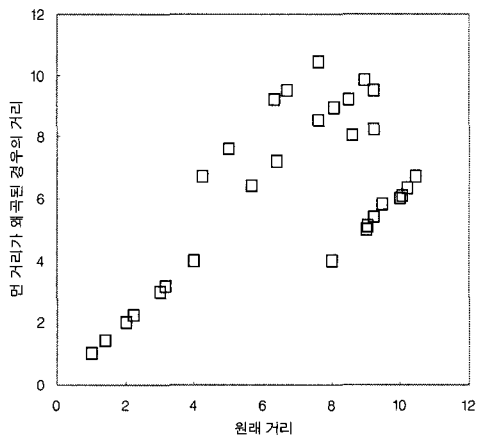
〈그림 1〉 가상 변수의 이상적인 배치도



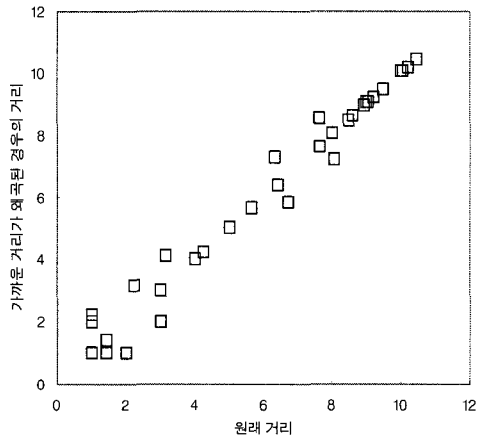
〈그림 2〉 먼 거리가 일부 왜곡된 가상 변수의 배치도



〈그림 3〉 가까운 거리가 일부 왜곡된 가상 변수의 배치도



〈그림 4〉 먼 거리가 왜곡된 경우의 거리 변환 산점도



〈그림 5〉 가까운 거리가 왜곡된 경우의 거리 변환 산점도

3.2 근거리 결정계수

MDS에서 입력된 변수간의 근접성(유사성이나 비유사성)이 최종적으로 도출된 좌표공간 상에서의 거리로 얼마나 잘 변환되었는가를 측정하는 지표를 합치도(fit measures)라고 하며 스트레스, RSQ, 소외계수(coefficient of alienation) 등이 있다. 소외계수는 Guttman(1968)이 제안한 것으로서 상위와 거리 사이의 코사인계수를 구하여 제곱한 다음 1에서 뺀 값의 제곱근을 취한 것인데 최근에는 그다지 사용되지 않는다. 근접성과 거리 사이의 합치도로 주로 사용되는 것은 스트레스(ALSCAL에서는 S-스트레스)와 RSQ이다. 스트레스는 앞에서 설명한 것처럼 상

위와 거리 사이의 차이를 제곱하여 합한 것이고, RSQ는 상위와 거리 사이의 피어슨 상관계수를 구하여 제곱한 것이다. 흔히 상관계수를 제곱한 것을 결정계수(coefficient of determination)라고 부르므로 RSQ는 상위와 거리 사이의 결정계수라고 할 수 있다. 결정계수는 회귀 분석을 비롯한 여러 통계 분석에서 사용되며 흔히 종속변수의 분산 중에서 독립변수에 의해 설명되는 비율로 해석된다(김태근 2006). Schiffman 등은 결정계수가 스트레스보다 해석이 간단하며 데이터가 모델에 얼마나 들어맞는가를 나타내는 가장 좋은 척도라고 하였다(Schiffman, Reynolds, & Young 1981, 175).

그런데 MDS에 의한 지도를 평가하기 위해서는 상위와 거리 사이의 결정계수를 구하는 것보다, 원래 입력 자료에 나타난 변수간의 근접성과 MDS 지도에서의 거리 사이의 결정계수를 구하는 것이 더 정확하며 간단하다. MDS 지도를 평가하기 위한 독립변수로는 원래의 근접성 값에서 변환된 상위보다는 근접성 값 자체가 더 적합하기 때문이다. 앞에서 제시한 가상 사례에 대해서 원래 거리와 왜곡된 거리 사이의 결정계수, 즉 원래 거리와 왜곡된 거리 사이의 피어슨 상관계수의 제곱을 산출해보면 〈표 3〉과 같다.

〈표 3〉 가상 자료에 대한 결정계수 비교

먼 거리가 왜곡된 경우	가까운 거리가 왜곡된 경우
0.550	0.975

〈표 3〉을 보면 가까운 거리가 왜곡된 경우의 결정계수가 0.975로 나타나서 먼 거리가 왜곡된 경우의 결정계수 0.550보다 월등하게 높게 나타났다. 결정계수는 1에 가까울수록 한 변수의 변화가 다른 변수의 변화에 그대로 반영됨을 뜻하므로, 결정계수로만 판단한다면 가까운 거리가 왜곡되어 생성된 〈그림 3〉이 더 바람직하다고 주장할 수 있다. 그러나 〈그림 2〉와 〈그림 3〉의 비교에서 확인하였듯이 지적 구조를 파악하기 위한 MDS 지도라고 가정할 경우에는, 오히려 먼 거리가 왜곡되어 생성된 〈그림 2〉에서 올바른 지적 구조인 〈그림 1〉과 유사한 해석을 얻을 수 있다.

이처럼 MDS 지도의 시각적인 비교 결과와 결정계수를 통한 비교 결과가 상이하게 나타난 이유는 두 방법에서 주로 비교한 대상이 달랐기 때문이다. 시각적인 비교 결과에서는 군집간의 구분을 좌우하는 가까운 거리 위주로 구조를 해석하였지만, 결정계수의 산출에는 모든 변수쌍 간의 거리를 다 비교하였기에 가까운 거리와 먼 거리가 모두 반영된 것이다. 따라서 결정계수와 같은 기존 합치도 판정 방식은 MDS가 지적 구조를 표현하는 능력을 적절히 평가하지 못하는 한계가 있다.

결정계수가 지적 구조의 표현능력을 더 잘 평가할 수 있도록 보완하기 위한 방안으로, 결정계수의 산출에 모든 변수쌍 간의 거리를 반영하지 않고 가까운 변수쌍 간의 거리만 반영하는 방법을 검토해볼 수 있다. 앞의 가상 사례에서 보았듯이 가까운 거리의 변화 정도가 지적 구조의 해석에 미치는 영향이 크기 때문이다. 실제로 〈표 1〉의 가상 자료에 포함된 9개 변수를 조합한 36개 변수쌍 중에서 〈표 2〉에 나타난 변수간의 거리가 가까운 상위 1/3에 속하는 12개 변수쌍만을 대상으로 원래 거리와 왜곡된 거리 사이의 결정계수를 산출해보면

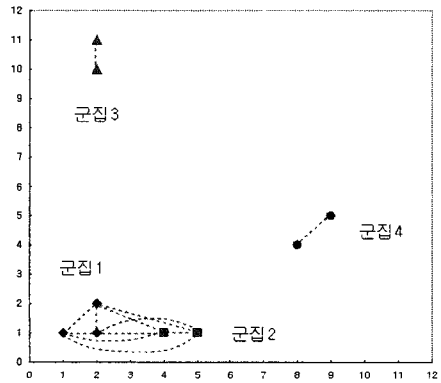
〈표 4〉와 같다.

〈표 4〉 원래 거리가 가까운 상위 1/3쌍에 대한 결정계수 비교

먼 거리가 왜곡된 경우	가까운 거리가 왜곡된 경우
1.000	0.598

〈표 4〉에서 보듯이 원래 배치도인 〈그림 1〉에서의 거리가 가까운 상위 1/3쌍에 대해서만 결정계수를 산출한다면, 먼 거리가 왜곡된 경우가 1.000으로 나타나서 가까운 거리가 왜곡된 경우의 0.598에 비해서 월등하게 높은 값을 가진다. 이와 같이 전체 변수쌍의 거리가 아닌 가까운 변수쌍의 거리를 대상으로 결정계수를 측정하면 2차원 지도를 시각적으로 해석하여 비교한 결과와 일치하는 결론을 얻게 된다. 이 연구에서는 이와 같은 근접성 상위 1/3쌍에 대해서만 산출한 결정계수를 '근거리 결정계수(local RSQ)'라고 부르기로 한다.

근거리 결정계수 = 피어슨 상관계수(입력 변수간의 근접성, 변환된 변수간의 거리)^2



〈그림 6〉 변수간의 거리가 가까운 상위 1/3 변수쌍을 점선으로 연결한 결과

거리가 가까운 상위 1/3쌍이 어떤 변수쌍인지를 원래 배치도인 〈그림 1〉에 표시하여 확인하면 〈그림 6〉과 같다. 이 그림에 나타나 있듯이 거리가 가까운(유사도가 높은) 상위 1/3 변수쌍은 대부분 동일 군집에 속한 변수 사이의 관계이거나 매우 가까운 군집(군집 1과 군집 2)에 속한 변수 사이의 관계에 해당한다. 군집분석과 결합하여 지적 구조를 규명하기 위한 MDS 지도를 평가할 때에는, 이처럼 가까운 변수 사이의 관계를 비교하여 산출한 근거리 결정계수가 시각적인 해석 결과와의 일치도가 높으므로 일반적인 결정계수보다 더 적합한 평가 지표라고 할 수 있다. 이 연구의 뒷부분에서는 면밀한 평가를 위해서 결정계수와 근거리 결정계수를 함께 측정하여 제시하였다.

근거리 결정계수와 유사한 발상으로 Wu와 Chow(2005)는 각 변수별로 인접 k개 변수와의 거리만을 비교하여 MDS 알고리즘의 품질을 측정한 바 있다. Wu와 Chow(2005)는 k를 4로 하여 평가하였는데, 이 방법 역시 거리가 먼 변수쌍을 제외하는 것이 MDS 지도를 더 올바르게 평가할 수 있다고 판단하여 거리가 가까운 변수쌍만 고려하는 것이다.

4. MDS 지도 작성 방식의 비교 실험

4.1 지적 구조 분석을 위한 MDS 수행 방식 구분

SPSS를 사용하여 지적구조를 분석하는 경우에 수행되고 있는 MDS의 일반적인 적용 절

차를 상세히 구분하면 〈그림 7〉과 같다.

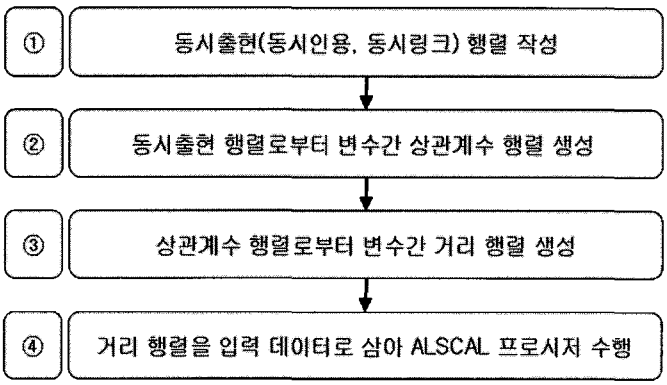
이처럼 단계적으로 각 행렬로부터 변수간의 관계를 다시 산출하는 과정은 근접성 행렬(proximity matrix)의 차수(次數)가 높아지는 것으로 간주할 수 있다. 즉, 원래의 동시출현 행렬은 1차 근접성 행렬, 이로부터 구한 상관계수 행렬은 2차 근접성 행렬, 그리고 여기에서 다시 구한 거리 행렬은 3차 근접성 행렬이 된다.

Vaughan과 You(2006), 그리고 Leydesdorff와 Vaughan(2006)은 SPSS의 ALSCAL 프로시저에서는 거리 행렬만 처리할 수 있을 뿐이고 동시출현빈도나 상관계수와 같은 유사도 행렬을 근접성 행렬로 그대로 처리하지 못한다고 하였으나, 이는 SPSS를 메뉴 방식으로 사용할 때만 해당된다. 이들의 주장과 달리 SPSS를 명령어 방식으로 이용할 경우에는 비계량적 MDS 지정을 위한 하위 명령어인 'LEVEL'에서 다음과 같이 'SIMILAR' 옵션을 지정하면 유사도 행렬을 그대로 근접성 행렬로 처리할 수 있다(SPSS Inc. 2003).

```
ALSCAL
 VARIABLES= A B C D E F G H I
 /SHAPE=SYMMETRIC
 /LEVEL=ORDINAL(SIMILAR)
 ...(이하 생략)
```

이와 달리 SPSS의 PROXSCAL 프로시저는 메뉴 방식에서 입력자료가 유사도 행렬인지 아니면 거리 행렬인지를 지정할 수 있다. 따라서 SPSS를 사용할 때 ALSCAL이나 PROXSCAL 모두 유사도 행렬과 거리 행렬을 그대로 입력하여 근접성 행렬로 사용할 수 있다. 다만 상관행렬을 그대로 근접성 행렬로 사용할 경우에는 설정 항목 중에서 '다음 값 이하의 거리는 결측값으로 처리' 항목의 기본값인 0을 -1로 고쳐서 지정해야 한다.

유사도 행렬을 그대로 근접성 행렬로 간주하여 처리할 수 있으므로 지적구조 분석을 수행할 때 상관계수 행렬로부터 MDS 지도를 도출하는 방식은 〈표 5〉와 같이 여러 가지가 가능하다. 동시인용빈도와 같은 1차 근접성도 MDS의 입력 데이터로 삼을 수 있지만 통상적으로는



〈그림 7〉 지적구조 분석을 위한 상세한 MDS 적용 절차

〈표 5〉 상관계수 행렬에서 MDS 지도를 도출하는 방식의 구분

약 칭	내 용		
	근접성의 차수	근접성	MDS 알고리즘
2CA	[2] 2차 근접성	[C] 상관계수	[A] ALSCAL
2CP	[2] 2차 근접성	[C] 상관계수	[P] PROXSCAL
3UA	[3] 3차 근접성	[U] 상관계수 벡터의 유클리드 거리	[A] ALSCAL
3UP	[3] 3차 근접성	[U] 상관계수 벡터의 유클리드 거리	[P] PROXSCAL
3ZA	[3] 3차 근접성	[Z] z점수화된 상관계수 벡터의 유클리드 거리	[A] ALSCAL
3ZP	[3] 3차 근접성	[Z] z점수화된 상관계수 벡터의 유클리드 거리	[P] PROXSCAL

상관계수 행렬을 이용하고 있으므로 여기서는 입력 데이터의 근접성이 2차 이상인 경우만 제시하였다. 각 방식의 약칭은 내용을 반영하도록 연구자가 임의로 부여하였다.

〈표 5〉의 여섯 가지 방식 중에서 2CA와 2CP는 상관계수 행렬을 그대로 근접성 자료로 삼고 분석을 수행하는 방식이다. 앞에서 설명한 바와 같이 ALSCAL과 PROXSCAL 명령 모두 입력된 행렬을 그대로 근접성 행렬로 처리할 수 있다.

3UA와 3UP는 2차 근접성인 상관계수 행렬로부터 다시 거리 행렬을 산출하게 되므로 3차 근접성으로부터 MDS 지도를 생성하는 방식이다. SPSS를 이용할 때에는 ALSCAL이나 PROXSCAL 메뉴를 선택한 후 나타나는 옵션 설정 대화창의 거리 설정 항목에서 '데이터로부터 거리행렬 계산하기'를 선택한 다음 나머지 항목은 기본 설정을 그대로 두고 수행하는 경우이다.

3ZA와 3ZP는 각각 3UA와 3UP와 거의 동일하지만 상관계수 행렬로부터 유클리드 거리를 산출할 때 상관계수 값을 z점수로 표준화하여 처리하는 방식이다. 상관도를 z점수로 표준화하는 것은 3차 근접성인 유클리드 거리 산출에 있어서 행렬을 구성하는 각 변수의 영향력을 비교적 고르게 반영하는 효과를 가져온다.

이상의 여섯 가지 방식 중에서 SPSS를 사용한 동시인용이나 동시링크 분석 연구를 위해 흔히 사용되어온 것은 3UA이다. 즉, 동시인용 행렬로부터 산출된 상관계수 행렬을 SPSS의 ALSCAL 명령의 입력 데이터로 처리하면서 거리 설정 옵션에서 '데이터로부터 거리행렬 계산하기'를 선택한 다음 나머지 사항은 기본 설정을 그대로 두고 수행하는 경우이다.

이와 같이 상관계수 행렬에서 MDS 지도를 생성하는 방식은 여러 가지가 가능하지만 대부분의 연구에서 3UA 방식을 채택한 이유는 별도로 밝혀져 있지 않다.

4.2 실험 설계

지적 구조 분석을 위한 MDS 지도 작성 방식을 비교 평가하기 위해서 우선 판단 기준을 설정한 다음 실험을 위한 실험집합을 준비하였다.

MDS 지도 작성 방식의 평가 기준을 설정하기 위해서는 앞에서 살펴본 MDS의 특징과 한계를 감안해야 한다. 고차원 동시출현 행렬을 처리하는 다변량분석 기법으로서 군집분석과

함께 사용되는 다차원분석 기법이 가져야할 바람직한 조건은 다음과 같다.

① 2차원 지도에서 두 변수 사이의 거리는 두 변수간 피어슨 상관계수와 반비례해야 한다.

② 2차원 지도에서 두 변수 사이의 거리는 두 변수간 동시출현빈도와 가급적 반비례해야 한다.

③ 피어슨 상관계수 행렬에서 상관값이 높은 상위 변수쌍은 상관계수가 높을수록 2차원 지도에서 가깝게 배치되어야 한다.

④ 2차원 지도에서의 거리를 기준으로 생성한 군집이 피어슨 상관계수 행렬로부터 생성한 군집과 같아야 한다.

①번 조건은 대부분의 지적 구조 분석 연구가 White와 Griffith(1981)의 제안 이후 변수간 피어슨 상관계수를 기준으로 수행된다는 점에서 무엇보다도 중요하다. 특히 군집분석이 피어슨 상관계수를 기준으로 수행되므로 군집분석과 MDS의 일관성 유지를 위해서는 꼭 필요한 조건이다. 이 연구에서는 피어슨 상관계수와 2차원 지도상의 거리 사이의 결정계수를 측정하여 ①번 조건의 충족 정도를 판단하기로 한다.

②번 조건은 동시출현빈도가 변수간 관계 측정의 출발점이라는 측면에서 중요하다. 비록 상관계수처럼 정규화가 되지는 않았지만 두 변수간의 동시출현빈도가 높을수록 가까운 관계로 간주하는 것이 합리적이다. 동시출현빈도는 두 변수에 대한 직접적인 관찰에서 1차적으로 드러나는 관계를 반영하는 것이므로 관찰결과(1차 근접성)와 분석결과(MDS 지도에서의 거리)의 일치도는 높을수록 바람직하다. 이 연구에서는 동시출현빈도와 2차원 지도상의 거리 사이의 결정계수를 측정하여 ②번 조건의 충족 정도를 판단하기로 한다.

③번 조건은 상관계수 행렬에 내재된 지적 구조를 제대로 표현하기 위해서는 먼 변수 사이의 관계보다는 가까운 변수 사이의 관계가 잘 반영되는 것이 더 중요하기 때문에 필요한 조건이다. 이 연구에서는 상관계수 행렬에서 상관값이 높은 상위 1/3 변수쌍의 상관계수와, 2차원 지도상에서 이들 변수쌍의 거리 사이의 근거리 결정계수를 산출하여 ③번 조건의 충족 정도를 판단하기로 한다.

④번 조건은 MDS에 의한 표현이 군집분석 결과와 얼마나 조화되는가를 나타낸다. 이는 White와 Griffith(1981)의 연구 이후 대부분의 지적 구조 분석 연구에서 MDS에 의한 MDS 지도에 군집분석으로 생성한 군집을 결합해서 표현하므로 두 분석기법이 조화될 필요가 있다는 점에서 필요한 조건이다. 이 연구에서는 지적 구조 분석에 자주 사용되는 군집분석 방법인 평균연결 기법과 Ward 기법을 사용하여 2차원 지도상의 거리를 입력자료로 삼은 군집생성 결과와 상관계수를 입력자료로 삼은 군집생성 결과를 비교하는 것으로 ④번 조건의 충족 정도를 판단하기로 한다. 군집생성 결과 사이의 비교 척도로는 CSIM(Chung & Lee 2001)과 WACS(정영미, 이재윤 2001)를 사용하였다. 군집분석의 성능 평가를 위한 척도가 여러 가지 제안된 바 있지만, CSIM(Cluster SIMlarity)과 WACS(Weighted Averaged Cluster Similarity)는 그 이름에서도 알 수 있듯이 두 가지 군집생성 결과를 비교하는 입장에서 접근하는 척도이기

때문에 채택하였다.

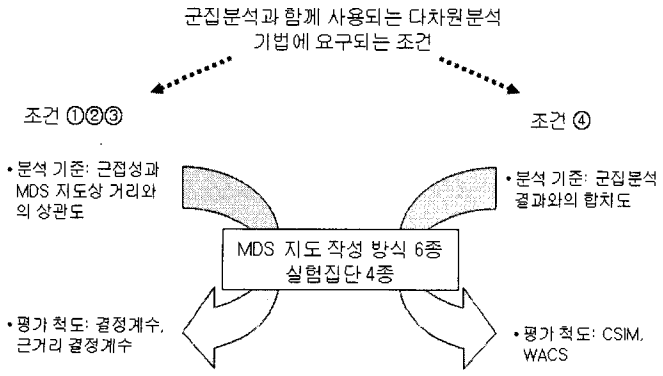
이상의 네 가지 조건 중에서 ①, ②, ③번 조건은 MDS 지도 작성에 사용되는 입력 자료와 지도 생성 결과가 얼마나 부합되는지를 평가하는 것이고, ④번 조건은 동일한 상관계수 행렬을 입력하여 분석한 군집분석 결과와 MDS 지도가 서로 얼마나 부합되는지를 평가하는 것이다.

상관계수 행렬에서 MDS 지도를 생성하는 여섯 가지 방식이 이상과 같은 네 가지 조건을 어느 정도 충족하는지 여부를 판단하기 위해서 〈표 6〉과 같이 네 가지 실험집단을 준비하였다. 이들은 모두 기존 지적 구조 분석 연구에서 사용된 데이터로서 저자동시인용 분석, 웹사이트 동시링크 분석, 단어 동시출현 분석(co-word analysis)을 위해 각각 수집된 자료이다. 가급적 실험 결과가 일반화될 수 있도록 실험집단의 분석 단위를 색인어, 표제어, 웹사이트, 인용저자 등으로 다양하게 구성하였다. 1986년부터 2005년까지 발표된 국내의 지적 구조 분석 연구를 정리한 이재윤(2006)에서 제시된 13편의 논문을 검토해본 결과 대상 집단의 규모가 최하 20에서 최대 71, 평균 47.6인 것으로 나타났으므로 이 연구의 실험집단은 크기 면에서도 대표성이 있다고 판단된다.

군집분석과 함께 사용되는 다차원분석 기법에 요구되는 네 가지 조건을 〈표 6〉의 여섯 가지 실험집단에 대해서 평가하는 실험을 〈그림 8〉과 같이 설계하였다. 조건 네 가지 중에서 ①, ②, ③번 조건을 평가할 때에는 결정계수와 근거리 결정계수로 측정 기준을 삼고, ④번 조건을 평가할 때에는 CSIM과 WACS 척도를 측정 기준으로 삼았다. 생성된 MDS 지도가 원래

〈표 6〉 MDS 지도 작성 방식 평가를 위한 실험집단

실험집단명	성 격	내 용	원자료 출처
CW32	단어 동시출현	국내 문헌정보학 분야 학술논문의 주요 색인어 32개	유영준(2003)
CL36	웹 사이트 동시링크	커뮤니케이션 분야의 주요 웹 사이트 36개	이성숙(2005)
CC50	저자동시인용	국내 컴퓨터과학 분야 주요 저자 50명	이은숙, 정영미(2002)
CW85	단어 동시출현	국내 인지과학분야 학술논문의 주요 표제어 85개	이재윤, 정주희(2006)



〈그림 8〉 MDS 지도 작성 방식의 비교 실험 구성도

의 입력자료인 상관계수 행렬이나 동시출현빈도 행렬과 부합될수록 결정계수와 근거리 결정계수는 1에 가깝게 높아진다. 또한 상관계수 행렬을 입력하여 생성된 군집분석 결과와 MDS 지도가 서로 부합될수록 CSIM과 WACS 척도의 값도 더 높은 값을 가지게 된다. 결정계수, 근거리 결정계수, CSIM, WACS의 네 평가척도의 값의 범위는 모두 최저 0에서 최고 1까지 분포한다.

4.3 실험 결과 분석

4.3.1 입력된 근접성과 MDS 지도상 거리 사이의 상관도

우선 MDS 지도 작성에 사용되는 입력 자료와 지도 생성 결과가 얼마나 부합되는지를 평가하는 세 가지 조건에 대한 실험 결과는 〈표 7〉, 〈표 8〉, 〈표 9〉와 같다. 여기에 더해서 원래의 빈도행렬에 나타난 값과 MDS 지도상 거리 사이의 근거리 결정계수까지 추가로 산출해본 것이 〈표 10〉이다. 〈그림 9〉는 네 가지 실험집단에 대해서 측정한 값의 평균을 MDS 지도 작성 방식별로 비교한 것이다.

대체적으로 입력자료와 처리 방식이 동일할 경우에는 ALSCAL보다 PROXSCAL 알고리즘을 적용한 경우가 더 좋았다. 즉 2CA보다는 2CP가, 3UA보다는 3UP가, 그리고 3ZA보다는 3ZP가 입력자료와 MDS 지도 사이의 부합정도가 더 좋았다. 이는 통상적으로 SPSS를 사용하여 지적 구조에 관한 MDS 지도를 만들 때 사용되어온 ALSCAL보다는 PROXSCAL 알고리즘이 더 적절하다는 것을 뜻한다.

또한 알고리즘이 동일할 경우에는 상관계수를 z점수화하여 유클리드 거리를 산출하는 방식이 가장 좋고 상관계수 행렬을 그대로 입력한 2C? 방식이 그 다음이었으며, 상관계수 행

〈표 7〉 상관계수값과의 결정계수(괄호안은 순위)

방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.1076(5)	0.1475(3)	0.0960(6)	0.1401(4)	0.1527(2)	0.1890(1)
CL36	0.5263(4)	0.5678(1)	0.3607(6)	0.4018(5)	0.5359(3)	0.5655(2)
CC50	0.3796(4)	0.4029(2)	0.2812(6)	0.2973(5)	0.3916(3)	0.4058(1)
CW85	0.2514(4)	0.2656(2)	0.1372(6)	0.1586(5)	0.2531(3)	0.2683(1)
평균	0.2946(4)	0.3271(2)	0.2050(6)	0.2381(5)	0.3169(3)	0.3427(1)

〈표 8〉 빈도값과의 결정계수(괄호안은 순위)

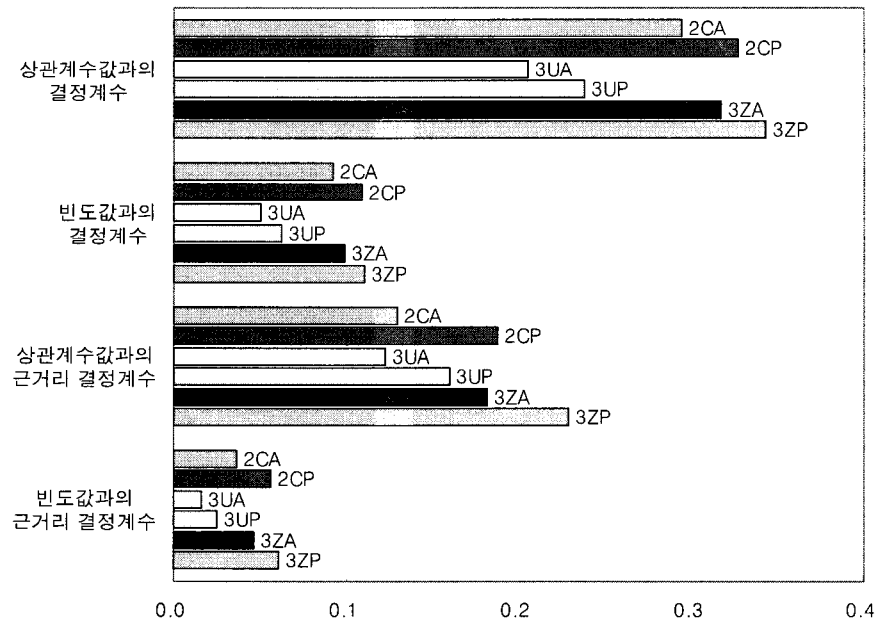
방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.0435(5)	0.0619(3)	0.0294(6)	0.0503(4)	0.0684(2)	0.0894(1)
CL36	0.1000(2)	0.1075(1)	0.0572(6)	0.0636(5)	0.0903(4)	0.0949(3)
CC50	0.1555(4)	0.1800(1)	0.0960(6)	0.1038(5)	0.1628(3)	0.1698(2)
CW85	0.0866(3)	0.1018(1)	0.0300(6)	0.0402(5)	0.0860(4)	0.0949(2)
평 균	0.0920(4)	0.1089(2)	0.0499(6)	0.0624(5)	0.0991(3)	0.1101(1)

〈표 9〉 상관계수값과의 근거리 결정계수(괄호안은 순위)

방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.1238(3)	0.1137(4)	0.0705(6)	0.0991(5)	0.1806(2)	0.1976(1)
CL36	0.2326(6)	0.4280(1)	0.2619(5)	0.3167(3)	0.3063(4)	0.3946(2)
CC50	0.1450(6)	0.2559(2)	0.1840(5)	0.2221(3)	0.2042(4)	0.2846(1)
CW85	0.0496(5)	0.0566(4)	0.0378(6)	0.0621(3)	0.0748(2)	0.0955(1)
평 균	0.1292(5)	0.1882(2)	0.1226(6)	0.1596(4)	0.1814(3)	0.2292(1)

〈표 10〉 빈도값과의 근거리 결정계수(괄호안은 순위)

방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.0592(4)	0.0604(3)	0.0306(6)	0.0473(5)	0.1112(2)	0.1308(1)
CL36	0.0394(2)	0.0670(1)	0.0150(6)	0.0176(5)	0.0222(4)	0.0283(3)
CC50	0.0696(4)	0.1130(2)	0.0525(6)	0.0637(5)	0.0928(3)	0.1202(1)
CW85	0.0032(4)	0.0117(2)	0.0004(6)	0.0008(5)	0.0057(3)	0.0119(1)
평 균	0.0363(4)	0.0563(2)	0.0161(6)	0.0248(5)	0.0465(3)	0.0607(1)



〈그림 9〉 MDS 지도 작성에 사용되는 입력 자료와 지도 생성 결과 사이의 부합 정도 평가 결과(4개 실험집단 결과값의 평균)

렬에서 유클리드 거리를 산출하는 방식이 가장 나쁜 것으로 나타났다(ALSCAL에서는 3ZA 〉 2CA 〉 3UA; PROXSCAL에서는 3ZP 〉 2CP 〉 3UP).

측정 결과 평균적으로 입력자료와 MDS 지도 사이의 부합 정도가 가장 좋은 MDS 지도 작성 방식은 3ZP이고 가장 나쁜 방식은 3UA인 것으로 나타났다. 3ZP는 상관계수 행렬을 입력자료로 하되 z점수로 정규화한 다음 벡터 간 유클리드 거리를 산출하여 PROXSCAL 알고리즘으로 MDS 지도를 만드는 방식이다. 결정계수와 근거리 결정계수가 가장 낮은 3UA는 지적구조 분석에 전통적으로 사용되어온 방식으로서, 동시인용 행렬로부터 산출된 상관계수 행렬을 SPSS의 ALSCAL 명령의 입력 데이터로 처리하면서 거리 설정 옵션에서 '데이터로부터 거리행렬 계산하기'를 선택한 다음 나머지 사항은 기본 설정을 그대로 두고 수행하는 경우이다. 이와 같은 전통적인 방식이 가장 나쁜 결과를 보였다는 것은 지적 구조 분석을 위한 MDS의 기존 적용 방식을 재검토해야 함을 시사한다.

4.3.2 군집분석 결과와 MDS 지도 사이의 합치도

군집분석 결과가 MDS 지도 생성 결과와 얼마나 합치되는가 여부를 평가하기 위해서 지적구조 분석에 흔히 사용되는 군집분석 기법인 Ward 기법과 평균연결(ALINK) 기법의 두 가지를 사용하였다. 상관계수 행렬을 입력하여 생성한 군집과 MDS 지도 상에서의 거리 행렬을 입력하여 생성한 군집 사이의 유사한 정도를 CSIM과 WACS 척도로 측정한 결과는 〈표 11〉~〈표 14〉와 같다. 〈그림 10〉에는 네 가지

〈표 11〉 Ward 클러스터링 결과 - CSIM 기준(괄호안은 순위)

방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.1805(6)	0.3358(4)	0.3866(2)	0.4655(1)	0.2846(5)	0.3622(3)
CL36	0.6241(3)	0.6863(1)	0.4390(6)	0.5591(5)	0.6160(4)	0.6517(2)
CC50	0.3750(6)	0.4255(2)	0.3982(4)	0.3891(5)	0.3983(3)	0.4796(1)
CW85	0.3877(5)	0.3508(6)	0.4093(4)	0.4226(2)	0.4149(3)	0.4508(1)
평균	0.3918(6)	0.4496(3)	0.4083(5)	0.4591(2)	0.4284(4)	0.4861(1)

〈표 12〉 Ward 클러스터링 결과 - WACS 기준(괄호안은 순위)

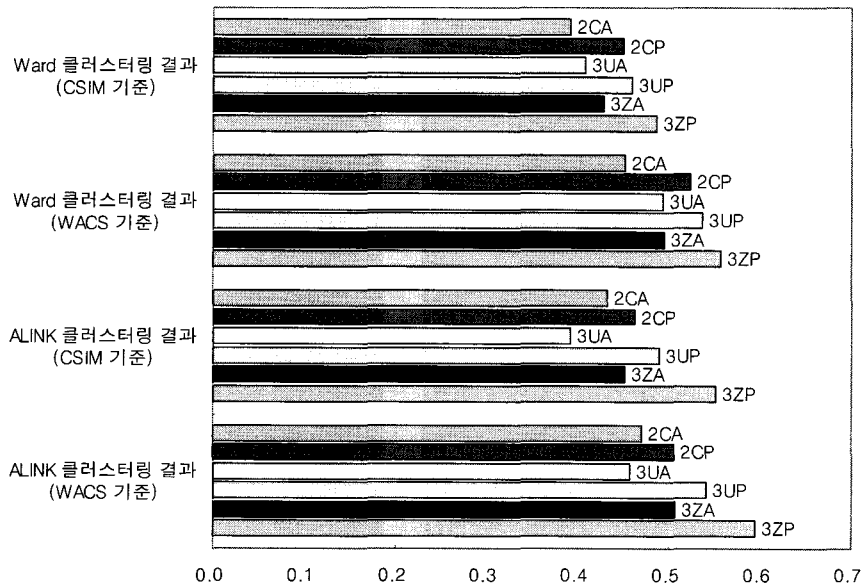
방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.2902(6)	0.4892(3)	0.4803(4)	0.5581(1)	0.3998(5)	0.4958(2)
CL36	0.6173(4)	0.6791(1)	0.5163(6)	0.6064(5)	0.6367(3)	0.6711(2)
CC50	0.4813(6)	0.5225(2)	0.5088(3)	0.5031(4)	0.5019(5)	0.5730(1)
CW85	0.4178(5)	0.4014(6)	0.4651(3)	0.4796(2)	0.4424(4)	0.4849(1)
평균	0.4516(6)	0.5230(3)	0.4926(5)	0.5368(2)	0.4952(4)	0.5562(1)

〈표 13〉 ALINK 클러스터링 결과 - CSIM 기준(괄호안은 순위)

방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.2413(6)	0.2857(4)	0.2847(5)	0.3540(2)	0.3158(3)	0.3768(1)
CL36	0.6227(4)	0.7395(3)	0.5412(6)	0.8084(2)	0.6160(5)	0.8949(1)
CC50	0.4471(3)	0.3794(5)	0.3469(6)	0.4215(4)	0.4521(2)	0.5098(1)
CW85	0.4161(4)	0.4446(1)	0.3974(5)	0.3712(6)	0.4192(3)	0.4213(2)
평균	0.4318(5)	0.4623(3)	0.3925(6)	0.4888(2)	0.4508(4)	0.5507(1)

〈표 14〉 ALINK 클러스터링 결과 - WACS 기준(괄호안은 순위)

방식 / 실험집단	2CA	2CP	3UA	3UP	3ZA	3ZP
CW32	0.3085(6)	0.3962(4)	0.3928(5)	0.4695(2)	0.4331(3)	0.4814(1)
CL36	0.6306(5)	0.6952(3)	0.5888(6)	0.7591(2)	0.6367(4)	0.8692(1)
CC50	0.5125(3)	0.4762(5)	0.4351(6)	0.5077(4)	0.5257(2)	0.5776(1)
CW85	0.4297(4)	0.4523(1)	0.4082(6)	0.4244(5)	0.4317(3)	0.4513(2)
평균	0.4703(5)	0.5050(4)	0.4562(6)	0.5402(2)	0.5068(3)	0.5949(1)



〈그림 10〉 상관계수 행렬을 입력한 군집분석 결과와 MDS 지도의 부합 정도 평가 결과(4개 실험집단 결과값의 평균)

실험집단에 대해서 측정한 CSIM과 WACS 값의 평균을 비교하였다.

두 가지 군집화 기법과 두 가지 평가 척도를 조합한 네 경우에서 모두 ALSCAL을 적용한 결과보다 PROXSCAL 알고리즘을 적용한 결과의 MDS 지도가 군집분석 결과와 더 부합되는 것으로 나타났다(2CA > 2CP; 3UA > 3UP; 3ZA > 3ZP). 앞서 살펴본 근접성과 MDS 지도상 거리 사이의 상관도 분석에서와 마찬가지로 지적 구조 분석을 위해서는 통상적인 ALSCAL 알고리즘보다는 PROXSCAL 알고리즘이 더 적절함을 알 수 있다.

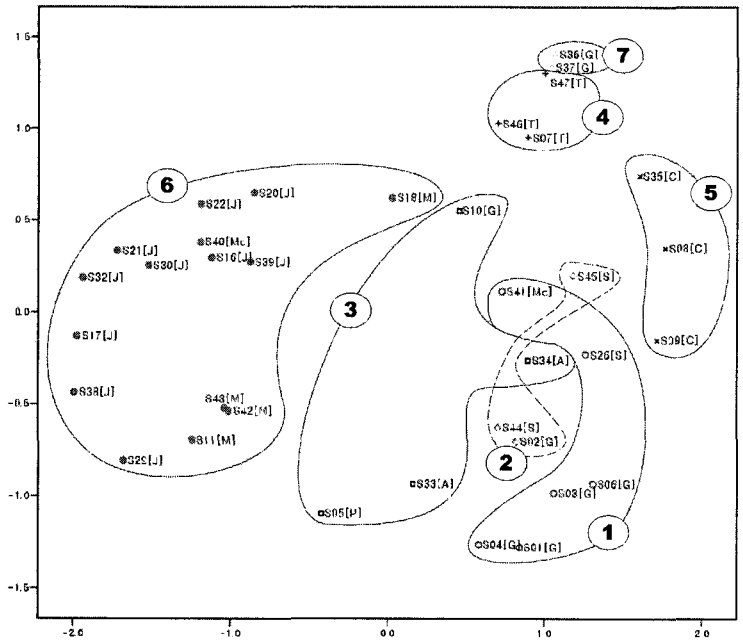
동일한 MDS 알고리즘을 사용하고 입력자료의 처리 방식을 달리했을 때의 성능 차이는 Ward 기법과 평균연결 기법이 약간 다르게 나타났다. Ward 기법으로 군집을 생성한 경우에는 상관계수를 z점수화하여 유클리드 거리를 산출한 3Z 계열 방식이 가장 좋고 상관계수 행렬에서 유클리드 거리를 산출하는 3U 계열 방식이 그 다음이었으며, 상관계수 행렬을 그대로 입력한 2C 계열 방식이 가장 나쁜 것으로 나타났다(ALSCAL에서는 3ZA > 3UA > 2CA; PROXSCAL에서는 3ZP > 3UP > 2CP). 평균연결 기법으로 군집을 생성한 경우에는 앞 절의 근접성과 MDS 지도상 거리 사이의 상관도 분석에서와 마찬가지로 상관계수를 z점수화하여 유클리드 거리를 산출한 3Z 계열 방식이 가장 좋고 상관계수 행렬을 그대로 입력한 2C 계열 방식이 그 다음이었으며, 상관계수 행렬에서 유클리드 거리를 산출하는 3U 계열 방식이 가장 나쁜 것으로 나타났다(ALSCAL에서는 3ZA > 2CA > 3UA; PROXSCAL에서는 3ZP > 2CP > 3UP).

군집화 기법 두 가지 모두에서 가장 좋은 MDS 방식은 3ZP 방식으로 나타나서 앞 절에서 기준 ①, ②, ③을 대상으로 분석한 결과와 같았다. 반면에 3UA 방식은 Ward 기법을 사용한 군집분석 결과와의 합치도는 뒤에서 두 번째로 낮았고, 평균연결 기법을 사용한 군집분석 결과와의 합치도는 가장 낮은 것으로 나타났다.

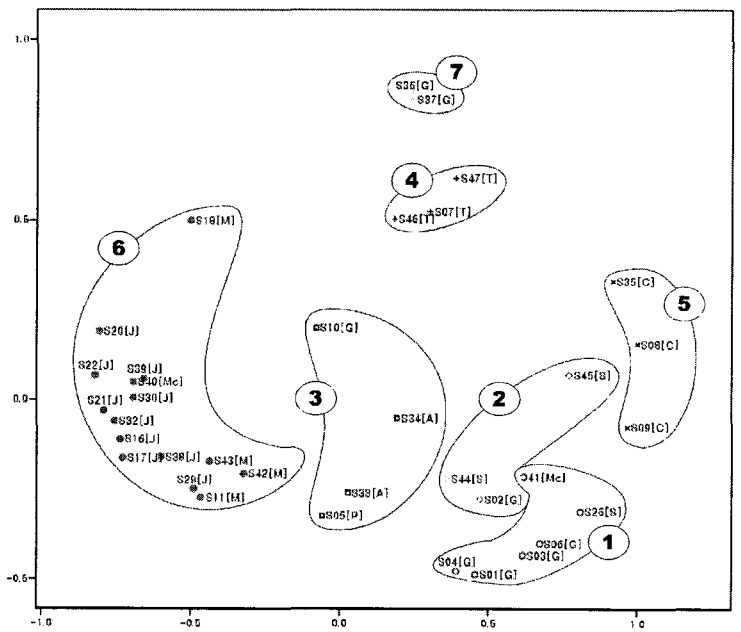
이상의 결과에서 보듯이 지적 구조 분석을 위해 흔히 사용되어온 3UA 방식은 두 가지 평가기준인 입력 자료인 근접성 행렬과의 상관도와 군집분석 결과와의 합치도 모두에서 평가 대상 MDS 지도 작성 방식 중에서 최하에 가까운 것으로 나타났다. 반면에 상관계수를 z점수로 표준화한 다음 유클리드 거리를 산출해서 PROXSCAL 알고리즘으로 MDS 지도를 작성하는 3ZP 방식은 모든 평가 기준과 척도에서 항상 최고인 것으로 나타났다.

3UA 방식과 3ZP 방식의 차이를 시각적으로 확인하기 위해서 <그림 11>과 <그림 12>를 제시하였다. 두 그림 모두 실험 집단 중에서 웹사이트 동시링크 자료인 CL36 실험집단을 대상으로 생성한 MDS 지도이며, 상관계수 행렬을 입력하여 Ward 기법으로 생성한 군집을 표시한 것이다. 다만 <그림 11>에서는 3UA 방식, 그리고 <그림 12>에서는 3ZP 방식으로 MDS 지도를 생성하였다. 두 그림의 비교에서 관찰되는 사항은 다음과 같다.

첫째, 3ZP 방식에 의한 <그림 12>에 나타난 각 군집이 3UA 방식에 의한 <그림 11>에 나타난 각 군집에 비해서 더 결집되어 있다. 즉, 군집을 둘러싼 곡선의 영역이 <그림 12>에서 더 좁게 나타난다.



〈그림 11〉 3UA 방식에 의한 MDS 지도 작성 및 Ward 군집 표시 사례
(CL36 실험집단 대상)



〈그림 12〉 3ZP 방식에 의한 MDS 지도 작성 및 Ward 군집 표시 사례
(CL36 실험집단 대상)

둘째, 3UA 방식에 의한 〈그림 11〉에서는 군집 2번과 같이 다른 군집의 영역을 가로지르는 군집이 나타나서 군집의 상대적인 위치가 모호한 경우가 있다.

셋째, 3UA 방식에 의한 〈그림 11〉에서는 3ZP 방식에 의한 〈그림 12〉에 비해서 군집과 군집 사이의 거리가 그림 위쪽의 군집 4번과 군집 7번처럼 너무 가까운 경우가 있다.

이와 같이 MDS 지도를 시각적으로 살펴본 결과에서도 널리 사용되어온 3UA 방식보다 3ZP 방식이 우월하다는 것을 알 수 있다.

5. 결 론

White와 Griffith(1981)의 연구 이후 4반세기가 넘는 시간이 지나는 동안 지적 구조를 표현하기 위해서는 MDS로 지도를 생성하고 그 위에 군집분석의 결과를 표시하는 것이 일반적이었다. 이런 전통적인 방식은 SPSS의 대중화와 함께 더욱 전형화되었고, 최근에는 홍보학과 같은 타 분야 연구자들이 문헌정보학 분야에서 전형화된 방식을 그대로 답습하여 ALSCAL 알고리즘으로 상관계수 행렬에서 MDS 지도를 만들고 군집을 표시해서 자체 학문분야의 지적 구조를 파악하는 사례(이수범, 권영순 2006)도 발표되었다.

이와 같이 MDS를 사용한 지적 구조의 분석이 오랜 역사를 가졌고 타 분야에서 그대로 수용될 정도로 전형화되었지만, MDS는 지적 구조를 시각적으로 표현하는데 있어서 세부 구조의 표현력이 떨어진다는 약점이 지적되고 있다. MDS로 생성한 지도의 세부 구조 표현 능력 부족 문제는 변수의 수가 많을수록 더 심각해지므로 MDS를 적용할 수 있는 경우가 제한되는 문제가 발생한다. 더군다나 널리 사용되고 있는 ALSCAL 알고리즘은 가까운 거리보다 먼 거리를 더 중요하게 반영하는 특성을 가지고 있으므로 세부 구조의 표현 능력이 더 취약할 가능성이 있다.

이 연구에서는 MDS의 특성을 가상 변수를 통해서 확인한 후, 지적 구조를 분석할 때 가능한 여섯 가지 MDS 지도 작성 방식을 제시하고 각 방식에 의해 생성된 MDS 지도의 품질을 비교하는 실험을 다양한 실제 지적 구조 데이터를 이용하여 수행하였다.

가상 변수를 활용한 분석에서는 MDS 지도상에서 먼 변수간의 관계보다 가까운 변수간의 관계를 제대로 반영하는 것이 더 중요하다는 것을 확인하였다. 이런 점을 감안하여 이후 지적 구조 분석을 위한 MDS 지도의 합치도 측정에 사용할 근거리 결정계수를 제안하였다.

네 가지 실제 지적 구조 데이터를 대상으로 이루어진 실험에서는 다음과 같은 결과를 얻었다.

첫째, 널리 사용되고 있는 ALSCAL 알고리즘보다 PROXSCAL 알고리즘이 더 올바르게 지적 구조를 표현하였다. 이는 ALSCAL 알고리즘의 최적화 함수인 S-스트레스가 상위와 거리의 차이의 제곱을 이용하므로 ALSCAL 알고리즘이 국지적 구조 표현에 더 취약하기 때문으로 풀이된다.

둘째, 알고리즘이 동일할 경우에는 상관계수를 z점수화하여 유클리드 거리를 산출한 다음 MDS 지도를 생성하는 방식이 가장 올바른 지적 구조를 표현하는 것으로 나타났다. 이는 상

관도를 z점수로 표준화하는 것이 3차 근접성인 유클리드 거리 산출에 있어서 행렬을 구성하는 각 변수의 영향력을 비교적 고르게 반영하는 효과를 가져오기 때문으로 생각된다.

셋째, 두 가지 알고리즘과 세 가지 처리 방식을 조합한 여섯 가지 방식 중에서는 상관계수를 z점수화하여 유클리드 거리를 산출한 다음 PROXSCAL 알고리즘으로 생성하는 경우가 가장 좋은 것으로 나타났다.

넷째, 전통적인 방식에 따라 상관계수 행렬을 SPSS의 ALSCAL 명령의 입력 데이터로 처리하면서 거리 설정 옵션에서 '데이터로부터 거리행렬 계산하기'를 선택한 다음 나머지 사항은 기본 설정을 그대로 두고 수행하는 경우는 여섯 가지 방식 중에서 가장 나쁜 편에 해당하는 것으로 나타났다. 여섯 가지 방식 중에서 가장 좋은 경우와 가장 나쁜 경우의 차이는 MDS지도에 군집을 표시한 그림 비교에서도 시각적으로 확인되었다.

이상의 결과를 고려할 때 이후 지적 구조 분석 연구에서는 전통적인 ALSCAL 알고리즘으로 MDS 지도를 작성하던 방식을 벗어나서 PROXSCAL 알고리즘을 사용하면서 변수를 z점수로 표준화하여 처리하는 것이 바람직할 것이다. 이상과 같이 전통적인 방식을 답습하지 않고 MDS 처리 과정을 주의깊게 구성한다면 더 구체적이고 명확한 지적 구조를 파악할 수 있음이 확인되었다. 아울러 이 연구에서 제안한 근거리 결정계수는 지적 구조 분석을 위해 생성된 MDS 지도의 품질을 더 올바르게 측정하는 척도로 활용되리라고 기대된다.

참 고 문 헌

김태근. 2006. 『u-Can 회귀분석』. 서울: 인간과 복지.

박광배. 2000. 『다차원척도법』. 서울: 교육과학사.

유영준. 2003. 『문헌정보학의 지식 구조에 관한 연구』. 박사학위논문, 연세대학교 대학원.

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

이수범, 권영순. 2006. 우리나라 PR 연구의 지적 구조에 대한 탐색적 연구: 저자동시인용분석을 중심으로. 『홍보학 연구』, 10(1): 229-261.

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

이재윤. 2006. 국내 최신 동향 파악을 위한 새로운 지적 구조 분석법. 『제13회 한국정보관리학회 학술대회 논문집』, 145-152.

이재윤, 문정순, 김희정. 2007. 텍스트 마이닝을 이용한 국내 기록관리학 분야 지적구조 분석. 『한국문헌정보학회지』, 41(1): 345-372.

이재윤, 정주희. 2006. 연구자 소속과 표제어 분

석을 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.

한승희, 이재윤. 2004. MDS를 이용한 개별문서의 계층적 지식구조 브라우징 인터페이스 설계. 『정보관리연구』, 35(3): 125-138.

Borg, Ingwer, and Patrick J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*. 2nd ed. New York: Springer Science+Business Media, Inc.

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

Chalmers, Matthew, & Paul Chitson. 1992. "Bead: Explorations in information visualization." *Proceedings of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 330--337.

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

Chung, Young Mee, and Jae Yun Lee. 2001. "A corpus-based approach to comparative evaluation of statistical term association measures." *Journal of the American Society for Information Science and Technology*, 52(4): 283-296.

Cox, T. F., and M. A. A. Cox. 2001. *Multidimensional Scaling*. 2nd ed. Boca Raton, Florida: Chapman & Hall/CRC.

Guttman, L. 1968. "A general nonmetric technique for finding the smallest coordinate space for a configuration of points." *Psychometrika*, 33: 469-504.

Herrero-Solana, V., F. Moya-Anegón, V. Guerrero-Bote, and F. Zapico-Alonso. 2006. "Graphical table of contents for library collections: The application of Universal Decimal Classification codes to subject maps." *Information Technology and Libraries*, 25(1): 43-47.

Kruskal, J. B. 1964. "Nonmetric multidimensional scaling: A numerical method." *Psychometrika*, 29(2): 115-129.

Leydesdorff, Loet, and Liwen Vaughan. 2006. "Co-occurrence matrices and their applications in information science: Extending ACA to the web environment." *Journal of the American Society for Information Science & Technology*, 57(12): 1616-1628.

McQuaid, M. J., T. H. Ong, H. C. Chen, & J. F. Nunamaker. 1999. "Multidimensional scaling for group memory visualization." *Decision Support Systems*, 27(1-2), 163-176.

Moya-Anegón, F., V. Herrero-Solana, and E. Jiménez-Contreras. 2006. "A connectionist and multivariate approach to science maps: the SOM, clustering

and MDS applied to library and information science research." *Journal of Information Science*, 32(1): 63-77.

Schiffman, S. S., M. L. Reynolds, and F. W. Young. 1981. *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*. New York: Academic Press.

Schildt, Henri A., and Juha T. Mattsson. 2006. "A dense network sub-grouping algorithm for co-citation analysis and its implementation in the software tool Sitkis." *Scientometrics*, 67(1): 143-163.

SPSS Inc. 2003. *SPSS® 12.0 Command Syntax Reference*. Chicago, IL: SPSS Inc.

Takane, Y., F. W. Young, and J. De Leeuw. 1977. "Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling method." *Psychometrika*, 42: 7-67.

Vaughan, Liwen, and Justin You. 2006. "Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market." Scientometrics, 68(3): 611-628.

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.

Wu, Sitao, and T. W. S. Chow. 2005. "PRSOM: a new visualization method by hybridizing multidimensional scaling and self-organizing map." *IEEE Transactions on Neural Networks*, 16(6): 1362-1380.