다차원 축척 기법

백병원 도서실

이 인 순

Multidimensional Scaling

In-Soon Lee

Inje University Seoul Paik Hospital Library

I. 머 리 말

문헌 정보학(Library and Information Science)이란 전통적인 도서관학(Library Science)과 최근에 발달한 정보학(Information Science) 분야가 결합된 학문으로 첫째, 문헌이나 정보와 관련된 관찰 가능한 모든 사실이나 현상을, 둘째로 논리적이고 과학적인 방법을 이용하여, 세째로 보편 타당한 지식체계를 형성하기 위한 이론 및 실무에 관한 연구를 목적으로 형성된 학제적인 성격을 띤 독자적 학문(independent discipline) 분야로 정의된다.

이러한 정의를 기준으로 문헌정보학의 연구대상은 크게 전문봉사적 측면, 기술적 측면, 과학적 측면의 3가지로 구분 되며, 학제적인 특성을 지닌 문헌정보학의 연구는 관련 학문의 타당한 방법론이나 접근법을 원용함으로써 보다 심도있는 연구를 가능하게 한다[1].

계량 정보학은 이중 과학적 측면으로 특히 연구 활동이 활발한 분야는 인용 분석 분야이다. 인용분석을 위한 새로운 기법과 척도가 창안되고 computer등의 도구를 사용하여 인용계수의 분석, 서지적 연구방향에서 최근에는 저자동시인용을 통하여 학문의 특정분야를 규명하려고하는 연구들이 수행되고 있다[2]. 본고에서는 이 중 computer program을 이용한 다차원 축척기법(Multidimensional Scaling)에 대하여 말하고자 한다.

II. 다차원 축척 기법의 개념

우리가 일상생활에 사용하고 있는 지도는 관심있는 지점들의 상대적 위치를 나타내고 있다. 이는 두 지점간의 거리를 나타낸 것으로 거리 대신 걸리는 시간으로 지도를 작성할 수도 있다. 한걸음 더 나아가 두 지점의 유사성을 나타내는 지도도 가능한데 MDS 기법은 이 유사성에 따라 각각의 사물을 공간상에 점으로 나타냄으로써 데이타와 데이타 구조의 이해를 보다 쉽게해 주는 다변량 통계기법중의 하나로 군집 분석과는 기본적으로 다른 목적과 전제에서 출발한다. 군집 분석은 특정 평가차원을 제시하여 대상물을 평가한 측정치를 가지고서 유사성의 정도에 따라 몇개의 집단으로 분류하는 기법인데 반해, MDS는 대상들간의 유사성 또는 선호도를 아무런 기준을 제시하지 않고 평가하게 하여 평가자가 대상물을 평가하는데 사용하는 기준을 발견하고 각 기준에 따라 대상물들이 갖는 측정값을 찾아내는데 그 목적이 있다[3].

MDS기법은 정치학, 사회학, 심리학 및 광고 전략 등에 복잡한 다변량 데이터를 분석하여 이를 종합적으

1) 정동열. 문헌정보학 연구방법론. 서울, 구미무역 출판부, *p 1992; 16-17.*

2) 김영민. 인용문헌을 이용한 검색에 관한 연구. 연세 대학교 대학원 석사학위논문(미간행), *1985.*

3) 오택섭. 사회과학 데이터 분석법. 서울, 나남. *pp. 457-484, 1984.*

로 일목 요연하게 볼 수 있는 방법으로 널리 쓰이고 있으며 데이타의 이면에 잠재하고 있는 구조적 요인을 찾는 방법으로 이용된다[4]. 특히 경영학에서 positioning map을 작성하는데 효과적으로 활용되고 있다. positioning map이란 사람들이 대상에 대해 느끼고 있는 내면적 기준을 심리적인 공간으로 나타낸 것으로 각 대상들을 상대적으로 좌표화하여 기하학적으로 형상화하는 기법이다. 이는 각 대상들을 동일 공간에 위치시켜 봄으로써 여러 가지 속성 차원에서 강점과 약점을 파악하는데 매우 유용한 정보를 제공해주는 도구로 활용될 수 있는데 이러한 positioning map을 작성하기 위한 방법이 다차원 축척이다.

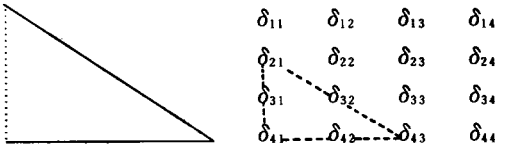
학문을 하는데 있어 즉 새로운 개념이 창조되고 그 개념이 학자들간에 받아들여지게 되는 과정은 보이지 않는 구조를 초래하는데 이러한 구조를 발견, 관찰할 것인가에 대한 관심은 학문에 대한 학문으로써 또 다른 연구 분야를 형성하는데 학문의 지적 구조 이해에 다차원적 지식 공간에서의 구조적 측면의 시각이 보다 융통성있고, 유용하다고 할 수 있다. 즉 어떤 개념을 다차원 공간에 벡터로 나타낼 수 있다면 지식 공간의 상호 관련된 주제 분야는 서로 공통적 개념 벡터를 갖게 되므로 개념들의 관계를 구조적 측면으로 기술하는 것이 정보 검색 과정을 보다 융통성있게 해 준다는 것이다. 이러한 구조의 변화에 대한 연구는 지적변화와 관련된 사회적 관계와 커뮤니케이션 패턴의 변화에 촛점을 둔 것과 학자들이 공식적으로 발표한 연구 문헌을 계량학적으로 분석하는 것이다. 후자의 경우 동시 인용 기법, 다차원 축척 기법, 군집 분석, 인자 분석등의 학문 영역의 지적 구조를 분석하는 보다 객관적인 방법론으로 발전되었다[5].

III. 다차원축척기법의 용어

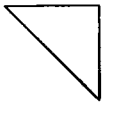
1. 자료와 관계되는 용어

- 대상(Object): 사물이나 사건으로 주로 n으로 표시 예: 사과
- 자극(Stimulus): 대상을 느끼는 것 예: 사과 맛
- 속성(Attribute): 자극의 특질 예: 달다
- 상이도계수(Proximity): 두 대상이 어느 정도 유사한지 또는 상이한지를 나타내는 수치로서 근접도에 의한 의견을 직접 조사하거나 또는 간접적으로 변인들간의 상관 관계를 구하여 얻을 수 있다 (여기서 유사도와 상이도를 같이 설명하는 것은 예를 들여 두 대상이 가장 비슷한 경우 가장 작은 수치를 주었다면 computer 입력時 regression =ASCENDING 가장 비슷하지 않은 경우 가장 작은 수치를 주었다면 입력時 regression=DESCENDING으로 저장된다). 일반적으로 δ를 symbol로 사용한다.
- 기본 행력(Data matrix): 대상을 둘씩 짝지은 행렬이므로 대상이 n개일때 n×n의 행렬이 생기는데 그 형태에 따라

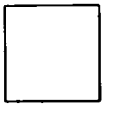
DATA, LOWER-HALF MATRIX



PATA, UPER-HALF MATRIX.



DATA, MATRIX가 있다.



- 자료의 측정 형태: 유사성 행렬이 어떤 척도를 이용하여 생성되었는가에에 따라 METRIC형태와 NON-METRIC형태로 대별될 수 있는데 METRIC형태의 자료는 등간 척도, 비율 척도에 의해 얻어진 자료를, NON-METRIC자료는 서열 척도에 의해 얻어진 자료를 말한다.

2. 공간(space)과 관계되는 용어

- 점(Point): 자극이 공간상에 차지하는 절대적인

4) 이재창. 박정섭. 다차원 축척*(Multidimensional Scaling)*기법. 응용 통계. *1986; 1: 62-80.*

5) 김영민. 인용문헌을 이용한 검색에 관한 연구. 연세 대학교 대학원 석사학위논문*(미간행), 1985.*

위치

- 차원(Dimension): 다차원 공간을 형성하고 있는 축
- 공간(Space): 축에의해 잠재적으로 각 점들이 위치할 수 있는 영역
- 형상(Configuration): 각 point가 기하학적으로 배열된 지도(map)
- 방향(Direction): MDS기법으로 만들어진 지도는 형상만이 의미를 가지고 있으며 방향에는 의미가 없어므로 회전시켜도 무방하다.
- 거리(Distance): 근접 관계를 나타낸 δ와 d가 지도상에서 어떻게 유지되는가가 중요한데 Shepard-Kruskal의 방법은 d와 δ가 타점에서 단조관계를 유지해야 가능하다고 하였다.
- Stress: 유사성 행렬로 부터 축척된 지도가 본래의 자료를 어느 정도 재현하였는가를 나타내는 목적 함수.

축척 차원이 증가하면 스트레스 값은 감소하므로 축척 공간에서 임의의 n개점을 출발 좌표(Starting configuration)로 하거나 computer내부의 난수(random number)를 이용하여 출발 좌표를 만든다.

특정함수 f에 대해 $d_{rs}=f(\delta_{rs})$인 관계라면 차이는 $f(\delta_{rs})-d_{rs}$가 되어 목적함수 스트레스는=

$\sqrt{\sum_r\sum_s(f(\delta_{rs})-d_{rs})^2/\sum_r\sum_s d_{rs}\sum_r\sum_s d_{rs}^2}$ 은 Scalefactor로 축척요소를 의미한다.

- 반복(Iteration): 스트레스 값이 일정한 수준에 이를때까지 계속하여 최적함수 f를 찾도록 한다.

스트레스와 반복의 의미를 살펴보면 그림 1~2의 지형에 낙하하여 가장 낮은곳을 찾는다면 Q점을 찾겠지만 임의의 여러지점에 낙하하면 결국 가장 낮은 지점을 찾아갈 것이다.

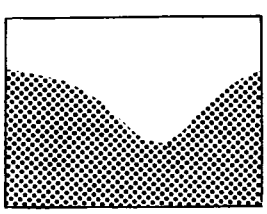


그림 1-1.

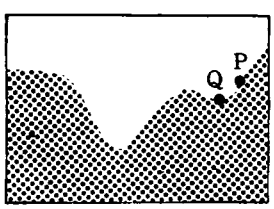


그림 1-2.

IV. 절 차

MDS기법을 이용하려면 둘씩 짝지은 개체간의 유사성(또는 상이성)자료를 토대로 m차원 공간상 이들 n개의 개체를 좌표화하여 기하학적으로 형상(configuration)화하는 기법이므로 유사성행렬(Similiarity-Dissimilarity matrix)이 직접 주어진 경우가 아니면 적절한 척도를 이용하여 다변량 자료를 유사성(Similarity)또는 상이성(Dissimilarity)행렬로 만들어야 한다. 분석 목적에 타당한 척도에 의한 유사성 행렬을 구해야 하는데 그 행렬수는

$nC2=\frac{n!}{2!(n-2)!}=n(n-1)/2$이고 구하는 방법은

i) 두 개체씩 짝 지은 모든 짝들이 각각에 대한 유사성의 평균을 구하여 이를 행렬형태로 배열하는 방법

ii) 비교적 개체의 수가 많은 경우(약 50~100개)사용되는 방법은 피험자로 하여금 개체들을 유사한것끼리 집단으로 분류하게 하는것

iii) 두 개씩 짝 지워진 개체들을 유사성의 순서에 따라 가장 유사한 짝들에서 가장 상이한 짝들에 이르기까지 순서를 매기게하여 얻는것

iv) 피험자로 하여금 제시된 두 개체가 같은 것이냐 다른 것이냐를 판단하도록 하는 방법

v) 개체들의 조를 만들고 이 조로 부터 <조 의 조>를 만들어 비교하는 방법이 있다.

유사성 행렬이 어떤 척도로 생성되었는가에 따라 고전적 축척과(Classical Scaling 또는 Metric Scaling) 순서적 축척(Ordinal Scaling)으로 나뉘어지지만 응답자로 부터 얻기 쉬운 자료의 형태인 non-metric자료를 이용하는 경우가 많다. 여기에서 고전적 축척이란 자료의 단위에 따라 결과가 달리 나오므로 자료의 표준편차를 구하여 즉 자료를 표준화하여 행렬을 구하는 것이고, 순서적 추척이란 $d_{ij}<d_{kl}\Rightarrow\delta_{ij}<\delta_{kl}$의 순서로 나열된 행렬을 의미하여 d_{ij}는 R^k공간에서의 i번째 점과 j번째 점의 거리를 의미한다. Euclidan 거리로 구성된 유사성 행렬은 Kruskal and Wish(1981)에 의하면 두 방법이 비슷한 축척을 한다고 하였다. 유사성 행렬을 얻어야 하는 경우 분석의 목적에 따라 적절한 유사도를 측정하는 척도를 결정해야 하는데 이 척도들은 몇가지 기본 가정을 만족

시켜야 한다. 살펴 보면

δ를 개체 r과 개체s의 거리(상이도)라 할때

(i) $\delta >= 0$(모든 r, s에 대해)

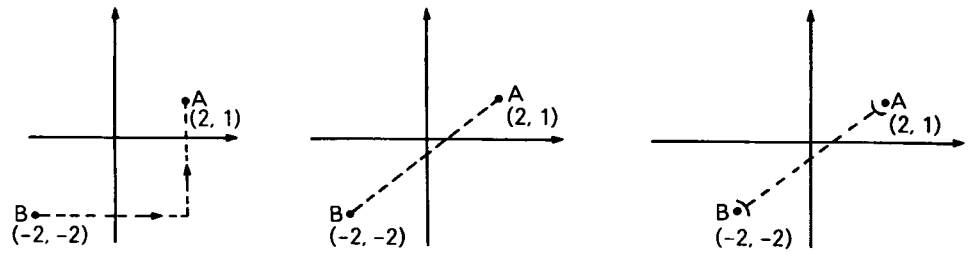
(ii) $\delta rr = 0$(모든 r에 대해)

(iii) $\delta rs = \delta sr$(모든 s, r에 대해)

(i)과 (ii)는 당연한 일이나(iii)을 일반화하는데는 문제가 있으므로 Minkowski metric의 특수한 경우인 유크리디언 거리 $[\delta rs = \{\sum_{j=1}^{p} (xrj - xsj^2)\}^{\frac{1}{2}}]$는 위 조건을 만족 시킨다.

$$\delta_{rs} = [\sum_{j=1}^{p} (X_{rj} = X_{sj})^k]^{1/k}$$

참고로 r차원에 대한 거리개념을 알아보면 1차원일 때는 그림 2의 City Block과 같은 개념이며 2차원일 때는 Euclidean, 3차원일때는 Minkowski의 그림과 같이 나타난다.



City Block
$D_{AB} = 4 + 3 = 7$

Euclideon
$D_{AB} = (4^2 + 3^2)^{\frac{1}{2}} = 5$

Minkowski r=3
$D_{AB} = (4^3 + 3^3)^{\frac{1}{3}} = 4.57$

*Minkowski의 r=3은 임의의 수로 r<2

그림 2.

유사성 행렬로 부터 축척된 지도가 본래의 자료를 어느 정도 재현하였는가를 목적함수(object function)로 축척하고 이 목적 함수를 stress라고 부르며 0에서 1사이의 값을 갖는데 수치가 0에 가까울수록 실제 상태에 가깝게 재현된다. 스트레스 값에 대한 명쾌한 기준치는 없으나 Kruskal은 다음과 같은 값과 의미를 제시하고 있다.

표 1.

스트레스 값(Stress value)	의미(goodness of fit)
0.2이상	아주 나쁨
0.2	나쁨
0.1	보통임
0.05	좋은 편임
0.025	아주 좋음
0	완벽함

목적 함수는 축척 지도의 차원(m)에 대한 함수가 [스트레스=g(m)]이 되어 함수 g는 축척 차원이 증

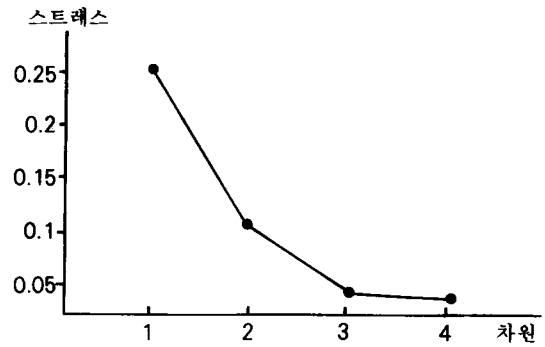


그림 3. x는 elbow pint.
재현 공간의 차원을 1에서 4까지 변화시키면서 대응되는 스트레스 값을 계산한 것으로 elbow point는 2차원으로 Kruskal의 논리에 따라 스트레스 10.6%는 적절한 값으로 2차적 공간에 축척 지도를 그릴 수 있다.

가하면 단조 감소형태이다. 방법은 반복(iteration)에 의하여 더 이상 감소하지 않는 스트레스 값에 도달하면 마지막으로 얻은 좌표들이 축척 지도가 된다. 즉 스트레스 값은 실제의 거리와 fitting된 상대적 거리간의 오차의 정도를 나타내어 주는 것이며 축척 차원의 결정은 함수 g의 기울기가 완만해지는 팔꿈치점(elbow pont)을 선택하여 결정하는 것이 보통이다. 프라이스에 의하면 사회학 분야의 주된 특성이 2차원상에 나타난다고 하였다.

※ **MDS를 실행할 수 있는 computer program**

최근 package로 된 program을 보면 KYST, SAS package의 ALSCAL Procedure, SYSTAT의 MDS 모듈, POLYCON, MULTISCALE MINISSA가 있다.

이들을 기능별로 나누어 보면 표 2와 같다.

표 2.

	Square	Rectangular	
No matrix weights	CMDS MINISSA POLYCON KYST ALSCAL MULTISCALE	CMDU POLYCON KYST ALSCAL	One marrix
	RMDS POLYCON KYST ALSCAL MULTISCALE	RMDU POLYCON KYST ALSCAL	More than one marix
Matrix weights	WMDS INDSCAL ALSCAL MULTISCALE	WMDU ALSCAL	

*ALSCAL: Alternation Least-squares sCaling
*CMDS: Classical MDS
*INDSCAL: Individual Differences multidimensional SCALing
*KYST: Kruskal Young Shepard Torgerson
*RMDS: Replicated MDS
*RMDU: replicated MultiDimensional Unfolding
*WMDS: Weighted MDS

V. 문헌 정보학분야의 다차원축척 기법의 적용 사례 및 문제점

최초의 연구는 Cox(1974)등이 경영학분야의 잡지를 분석 대상으로 삼아 한 잡지가 다른 잡지를 인용하는 패턴에 따라 잡지들의 상호 관계를 2차원의 평면위에 나타내어 서로 관련된 주제 분야를 이해하는데 도움을 줄 수 있는 틀을 제시하였다.

Lenk(1983)는 7개 학문 분야를 대상으로 하여 동시 지명을 바탕으로한 저자들 사이의 유사성을 매핑하였는데 그 결과 동시 지명도가 높은 연구자들이 서로 밀접하게 위치하여 그 분야의 비공식 연구 집단을 형성하였다. 그리고 정보학분야에 대한 동시 인용과의 비교에서는 두 방법간의 차이에도 불구하고 현저하게 비슷한 결과를 얻어 저자의 인용형태가 학문의 사회적 구조와 고도로 상응한다는 것을 입증하였다.

McGrath(1983)는 다차원축척 기법을 도서관분야에 적용시켰는데 대학 도서관에서 대출된도서 수를 각 학과나 전공 분야의 유사성의 척도로 삼아 각 학과사이의 거리를 2차원과 3차원의 지도로 나타내었으며 여기에서 얻은 결과로 각 전공분야에 대한 도서관의 장서 개발과 장서조직, 예산 할당등에 적용 해 볼 수 있는 방법을 거론하였다.

또한 그는 OCLC의 회원 도서관을 대상으로 각 도서관의 중복된 소장 도서 문제에 대하여 다차원 축척 기법을 사용하여 접근하였다. 즉 OCLC의 종합 목록에 있는 도서 중, 각 도서관이 공통적으로 소장하고 있는 보유량을 도서관사이의 유사성의 척도로 하여 매핑과 군집 분석을 한 결과 이 기법이 도서관 상호 대차를 위한 적정 규모의 소장 도서를 결정하는데 유용하다고 보고하였다.

한편 Smith(1984)는 한 도서관시스팀의 분관들간의 상호 대출 패턴연구에 다차원 축척의 적용방법이 유용하다는 것을 논의하였다. 즉, 각 분관의 소장 장서 중 상호 대출된 도서 수에 따른 유사성으로 분관들사이의 거리를 재어 매핑한다면, 지도상에 밀접하게 나타난 분관들을 통합하거나 또는 도서관자원의 집중화

6) 김영진, 전계서
7) 오택섭, 전계서

등을 결정하는데 객관적인 척도가 될 수 있다고 하였다.

이 밖에도 White(1984)와 Calhoun은 학생들의 수강 신청 데이타를 토대로 하여 상위 선택 과목을 지도에 나타냄으로써 교과 과정이 어떤 방식으로 운영되어 왔는가를 역사적으로 살펴 보는데에 다차원 축척법을 적용하였다[6].

이처럼 다차원 기법은 학문의 지적 구조 이해에 보다 융통성있고 유용하다고 할 수 있는데 Price는 더욱 구체적으로 지식 공간의 개념을 이차원이 공간에 충분히 나타낼 수 있고 체계적인 배열도 가능하여 저자들의 지도는 한 차원은 주제 성향을 또 다른 차원은 연구 스타일을 반영하여 종축은 각 저자의 하위 주제 분야에 대한 관심의 폭을 나타내고 횡축은 저자들의 연구 방법에 있어서의 태도나 경향으로 지도의 윗 부분에 위치한 저자들은 비교적 여러 하위 분야에 걸쳐 연구 활동을 하거나 인용이 되고 있으며 아래쪽 저자들은 비교적 단일한 주제 분야에서 연구를 하며, 오른쪽 저자들은 이론적 연구 활동을 하는 편이고 왼쪽 저자들은 경험적이고 실증적인 조사 방법 및 연구를 택하여 보다 부드러운 학문 성향을 갖는다고 하였다.

다차원 축척의 문제점은 수학적으로 차원이 아무리 높더라도 계산이 가능하나 3차원 이상인 경우에는 그래프로 나타내는 것이 불가능하며, 어떤 형태이든 computer의 도움 없이는 가장 단순한 형상도 작성하기가 어렵다[7]. 또한 주의 할 점은 각 차원의 의미를 부여 할 경우, 그분야의 전문가의 의견을 들어 차원의 성격을 정하는 일이다. 양자간의 관계를 회귀분석, 판별분석, 상관관계등을 분석하여 상관관계가 가장 높은 속성을 각 차원의 이름으로 결정하는 것이 바람직하다.

참 고 문 헌

1) 김영민. 인용문헌을 이용한 검색에 관한 연구. 연세 대학교 대학원 석사학위논문(미간행), *1985.*

2) 김영진.논문의 동시 인용을 통한 지적 구조의 규명에 관한 연구. 연세 대학교 대학원 석사학위논문(미간행), *1986.*

3) 오택섭. 사회과학 데이터 분석법. 서울, 나남. *pp. 457-484, 1984.*

4) 이재창. 박정섭. 다차원 축척*(Multidimensional Scaling)*기법. 응용 통계. *1986; 1: 62-80.*

5) 정동열. 문헌정보학 연구방법론. 서울, 구미무역 출판부, *p 1992; 16-17.*

6) Egghe L, Rousseau R. *Introduction to Informetrics, Amsterdam: Elsevier, 1990; 105-112.*

7) Shiffman, Susan S, et al. *Introduction to Multidimensional Scaling. New York: Academic Press, 1985.*