동국대학교

dongguk UNIVERSITY

2020년 - 데이터 청년 캠퍼스

데이터사이언스 기반 지능소프트웨어 과정

데이터 사이언스 개론

2) 비지도 학습 (Unsupervised Learning)

학습 내용

■ 비지도 학습 (Unsupervised Learning) : 4가지

- K-평균 군집화 (k-means Clustering)
- 주성분 분석 (Principal Component Analysis)
- 연관 규칙 (Association Rules)
- 소셜 네트워크 분석 (Social Network Analysis)

■ 지도 학습 (Supervised Learning) : 6가지

- 회귀 분석 (Regression Analysis)
- K-최근접 이웃 (k-Nearest Neighbors)
- ...

■ 강화 학습 (Reinforcement Learning) : 1가지

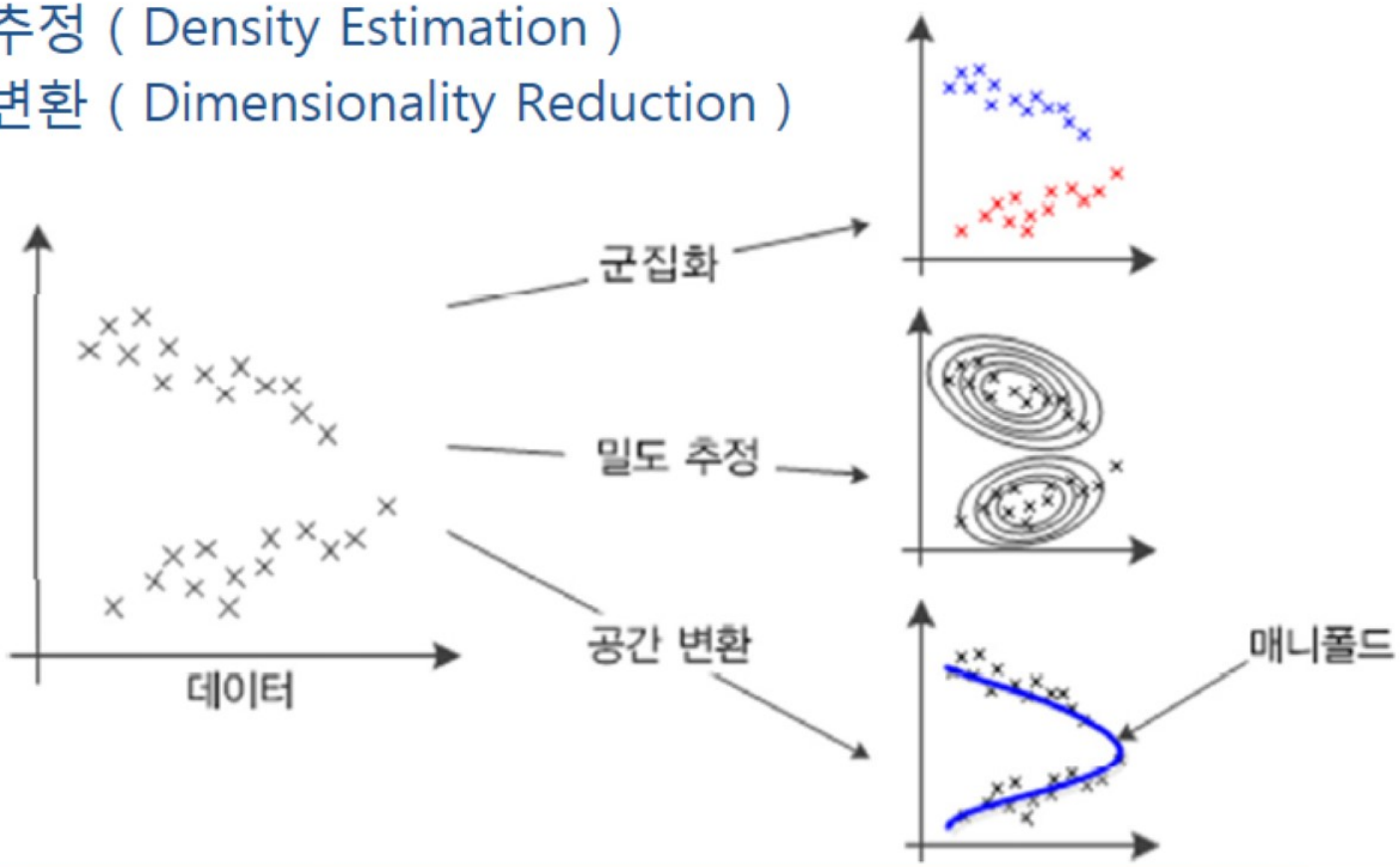
- 멀티-암드 밴딧 (Multi-Armed Bandits)

비지도 학습 (Unsupervised Learning)

- 데이터 셋 (Dataset) 에 숨겨진 패턴 (Pattern)을 찾고자 할 때 사용
- 레이블 (Label) 이 없는 샘플 (Sample) 데이터 셋을 이용
- 비지도 : 패턴이 없는지 / 있는지 / 있다면 무슨 모양인지 모르기 때문
  - 주로 어떤 상품들이 함께 팔리는가?
  - 고객들의 구매 성향이 어떠한가?
  - 새로운 고객은 구매 성향이 어떠할 것인가? / 어떤 상품을 구매할 예정인가?

비지도 학습 전 해야 할 일

- 비지도 학습으로 문제를 해결하기 위해서는 다음 문제로 접근해야 함
  - 군집화 (Clustering)
  - 밀도 추정 (Density Estimation)
  - 공간 변환 (Dimensionality Reduction)



k-평균 군집화 (k-means Clustering)

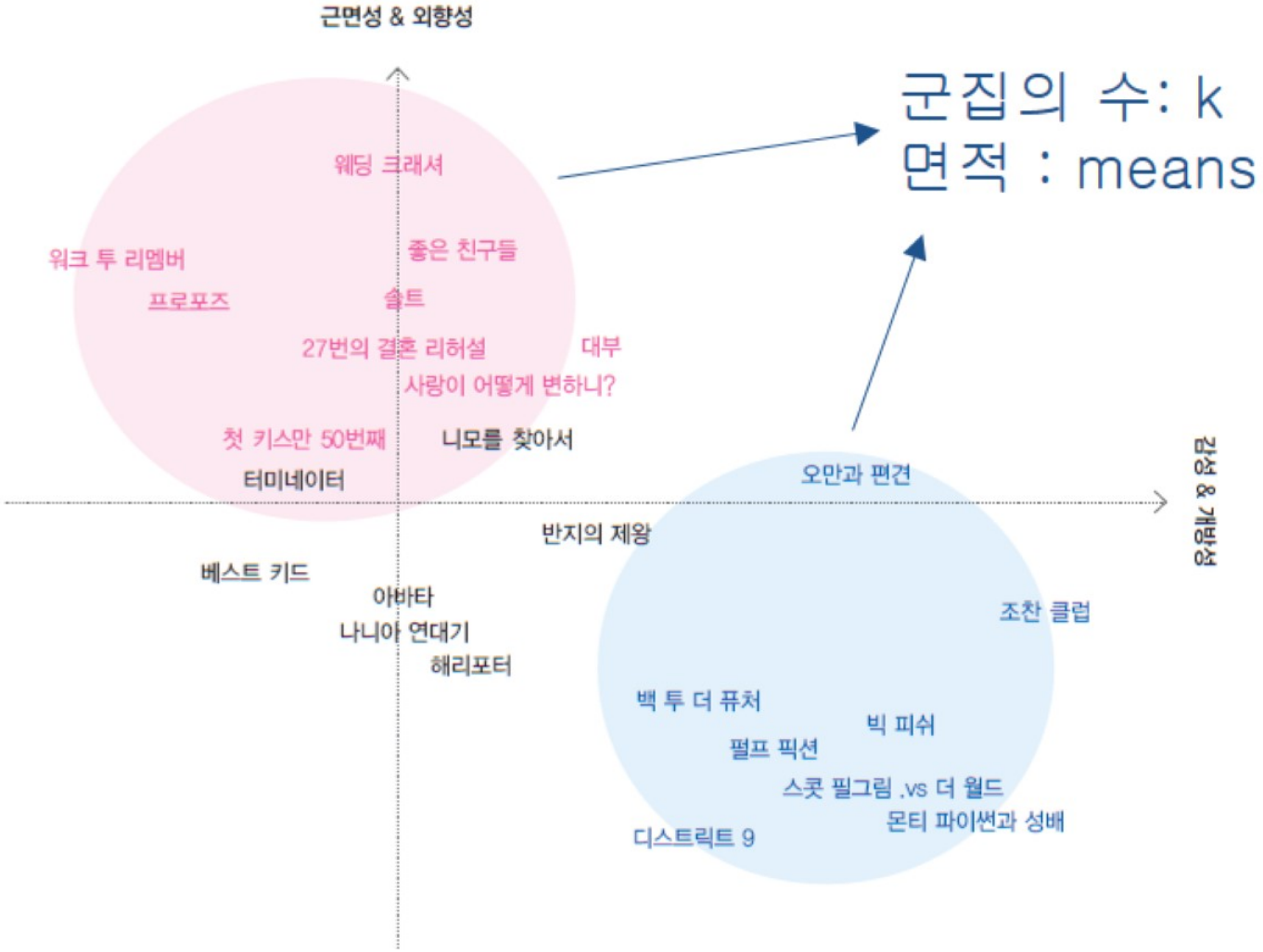
취미가 영화보기인 사람들의 성향

- SNS 사용자들에게 설문조사를 실시
  - 외향성
  - 근면성
  - 감성
  - 개방성

- 위 성격의 설문조사 결과
  - 감성적일 수록 개방성이 높음 (***X 축***)
  - 근면할 수록 외향적인 경우가 많음 (***Y 축***)

- 2차원 차트를 기반으로 SNS 에서 "좋아요" 한 영화를 매칭

k-평균 군집화 (k-means Clustering)



k-평균 군집화 (k-means Clustering)

장점

- 알고리즘이 직관적이고 매우 단순

단점

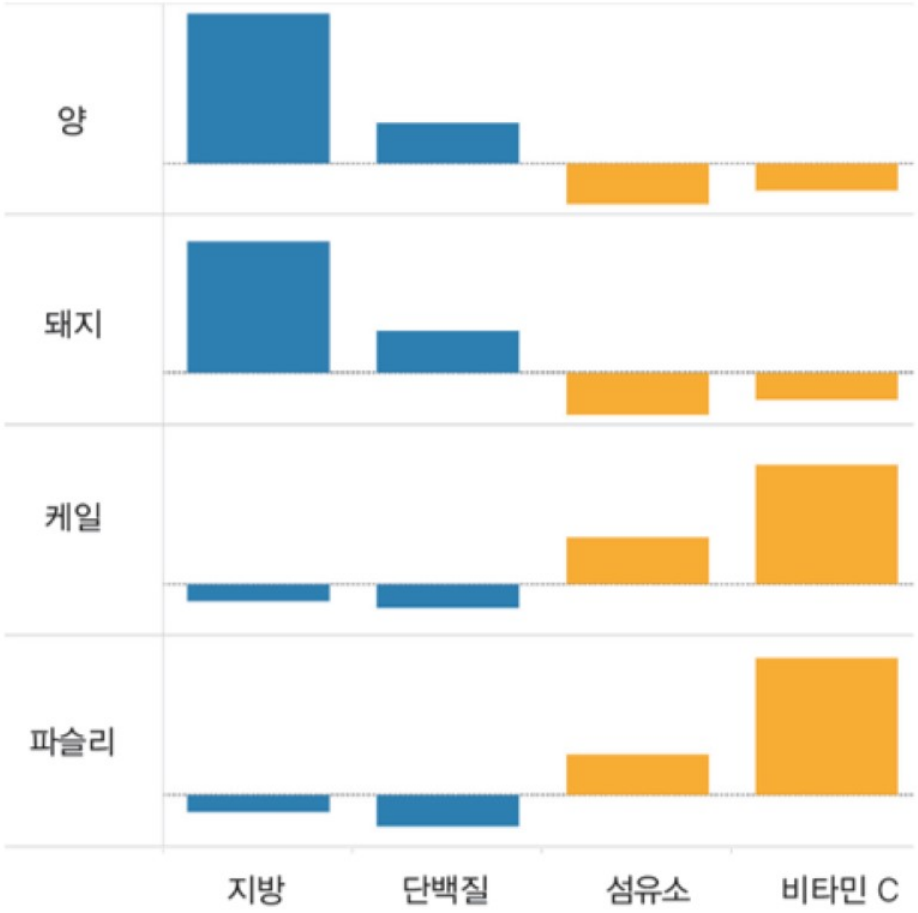
- 각 데이터는 단 하나의 군집에만 속해야 함
- 군집의 모양은 원 혹은 구 형태여야 함
- 군집끼리 겹치지 않아야 함

주성분 분석 (PCA)

■ 식품들을 기준에 따라 분류하려 함

- 각 식품들의 함량을 분석
  - 지방 함량
  - 단백질 함량
  - 섬유소 함량
  - 비타민 C

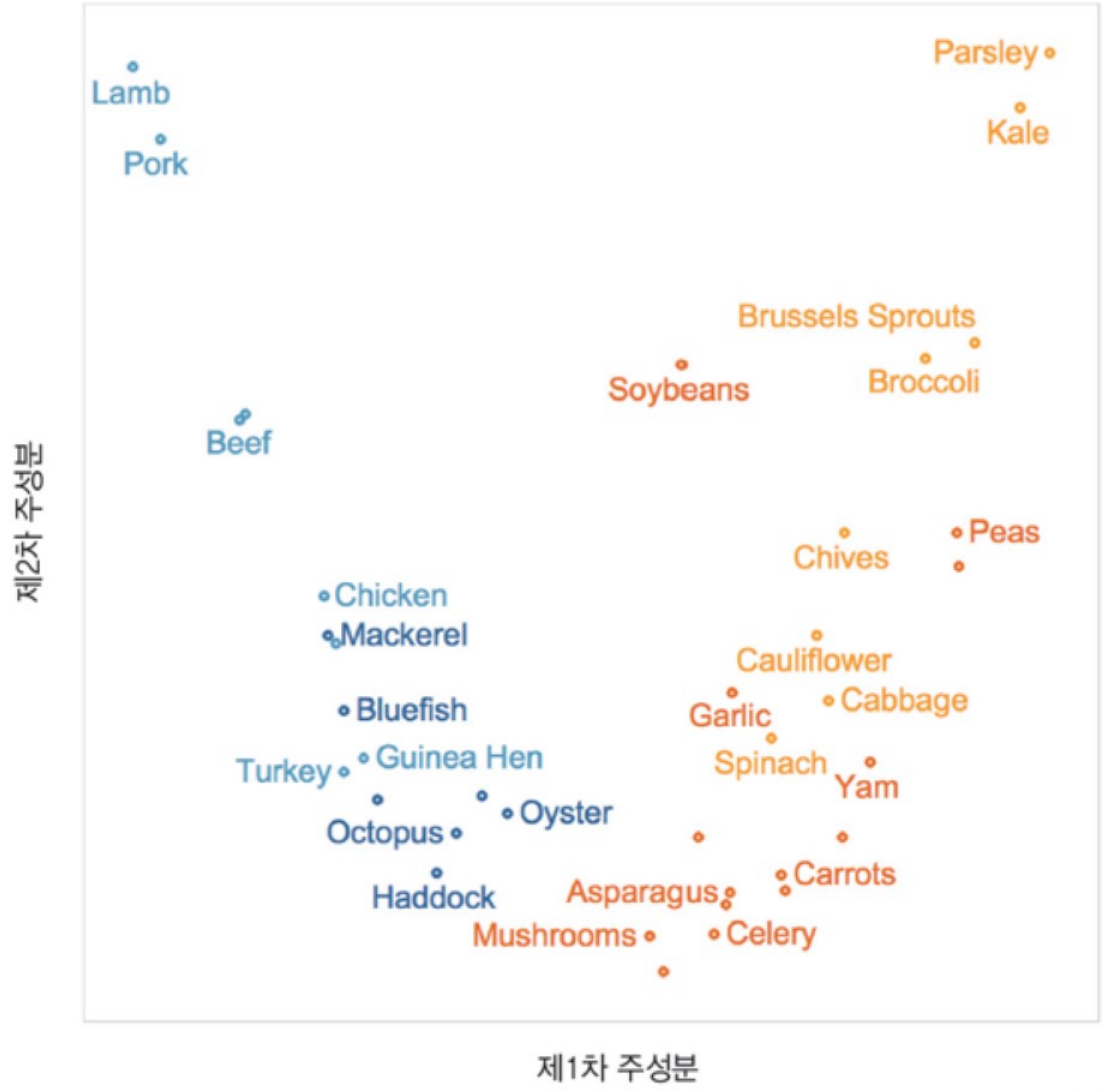
- 각 함량 별로 상관관계 (Correlation) 가 있음
  - 지방과 단백질 함량은 ***양의 상관관계 (Positive Correlation)***
  - 섬유소/비타민 C 과 지방/단백질 함량은 ***음의 상관관계 (Negative Correlation)***



주성분 분석 (PCA)

주성분 분석 결과

- 제1차 주성분은 육류와 채소로 구별 ***(X 축)***
- 제2차 주성분은 지방과 비타민 C로 구별 ***(Y 축)***



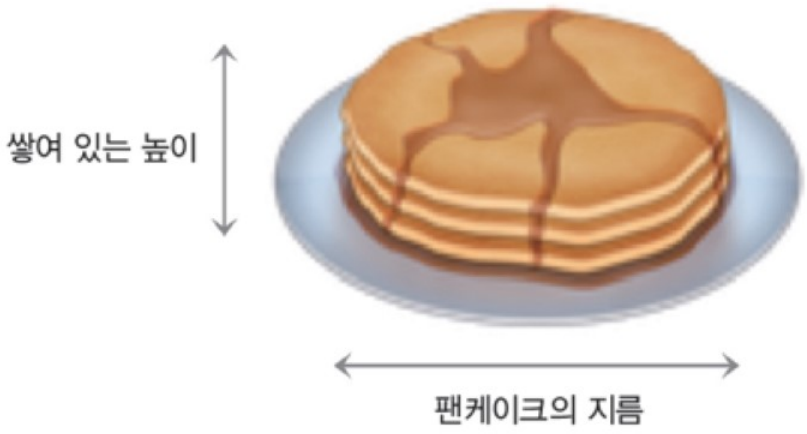
주성분 분석 (PCA)

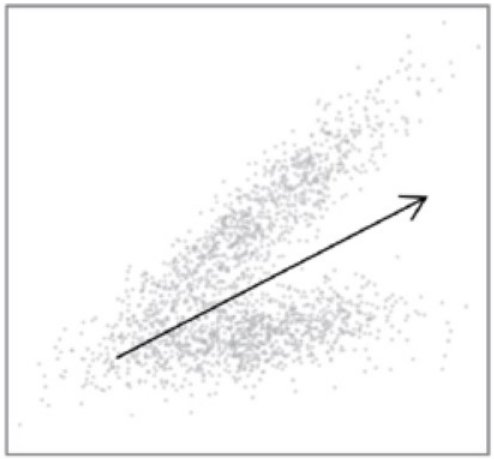
장점

- 변수가 많은 데이터 셋을 분석할 때 유용

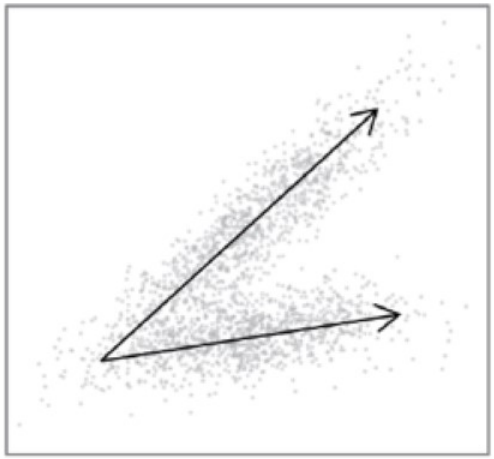
단점

- 데이터 포인트가 가장 넓은 차원이 가장 유용하다는 가정
- 분석 결과에 대한 인과관계 해석이 곤란
- 서로 다른 변수 간에는 최대한 직교 (Orthogonality) 여야 함





a) PCA로 찾은 성분

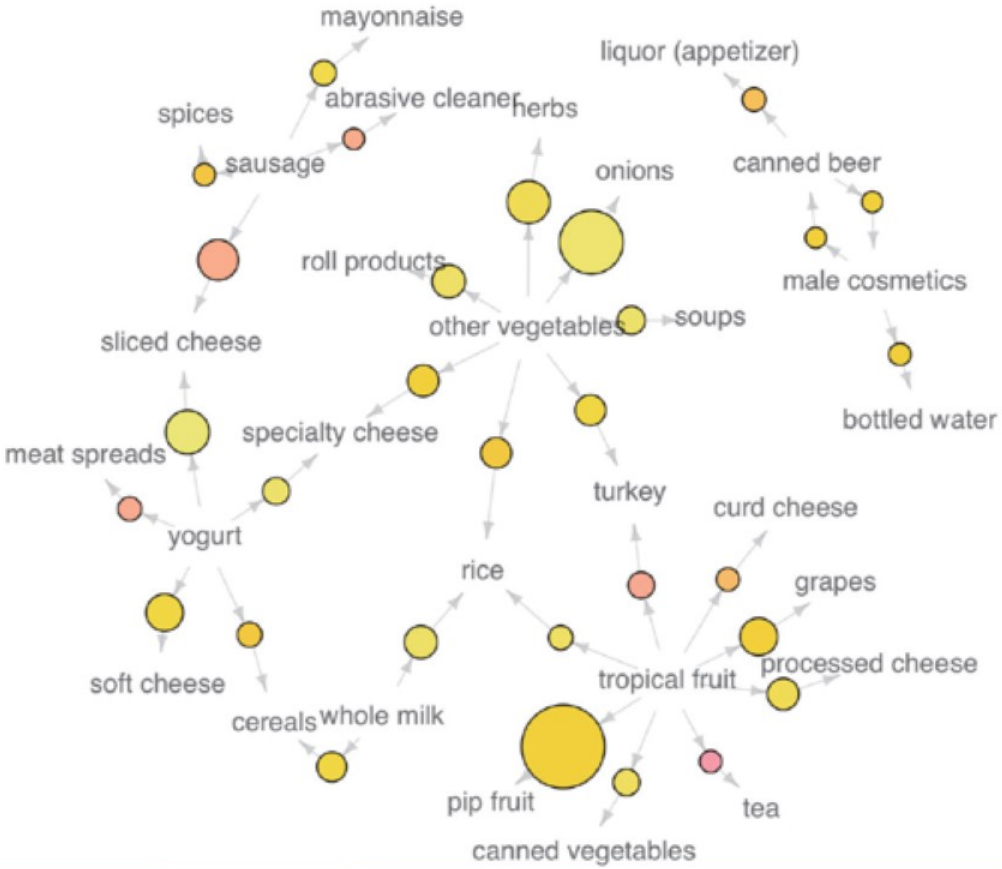


b) ICA로 찾은 성분

연관 규칙 (Association Rules)

식료품점에서 각 품목간 연관성을 찾고자 함

- 신뢰도가 0.9% 이상, 향상도가 2.3 이상인 품목들
- 원의 크기가 클수록 지지도가 높고, 붉은 원은 향상도가 높음



- 가장 빈번한 거래는 *씨앗 과일 (pip fruit)* 과 *열대 과일 (tropical fruit)*
- 두 번째로 빈번한 거래는 *양파 (onions)* 와 *채소 (other vegetables)*
- *슬라이스 치즈 (sliced cheese)* 를 구매하면 *소시지 (sausage)* 를 구매할 가능성이 큼
- *차 (tea)* 를 구매하면 *열대 과일 (tropical fruit)* 을 구매할 가능성이 큼

연관 규칙 (Association Rules)

■ 지지도, 신뢰도, 향상도

- **지지도 (Support)** : 데이터 셋에서 특정 품목의 집합 (Item Set) 비율

$$\text{Support}\{🍎\} = \frac{4}{8}$$

- **신뢰도 (Confidence)** : 품목 A 가 존재할 때, 품목 B 가 나타나는 비율

$$\text{Confidence}\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\{🍎, 🍺\}}{\text{Support}\{🍎\}}$$

- **향상도 (Lift)** : 품목 A 와 B 가 각각 잘 팔리는데, 그 중 품목 A 와 B 가 같이 팔리는 비율

$$\text{Lift}\{🍎 \rightarrow 🍺\} = \frac{\text{Support}\{🍎, 🍺\}}{\text{Support}\{🍎\} \times \text{Support}\{🍺\}}$$

연관 규칙 (Association Rules)

장점

- 구체적인 기준 (지지도, 신뢰도, 향상도) 에 따라 각 품목간 빈번한 정도를 수치화할 수 있음

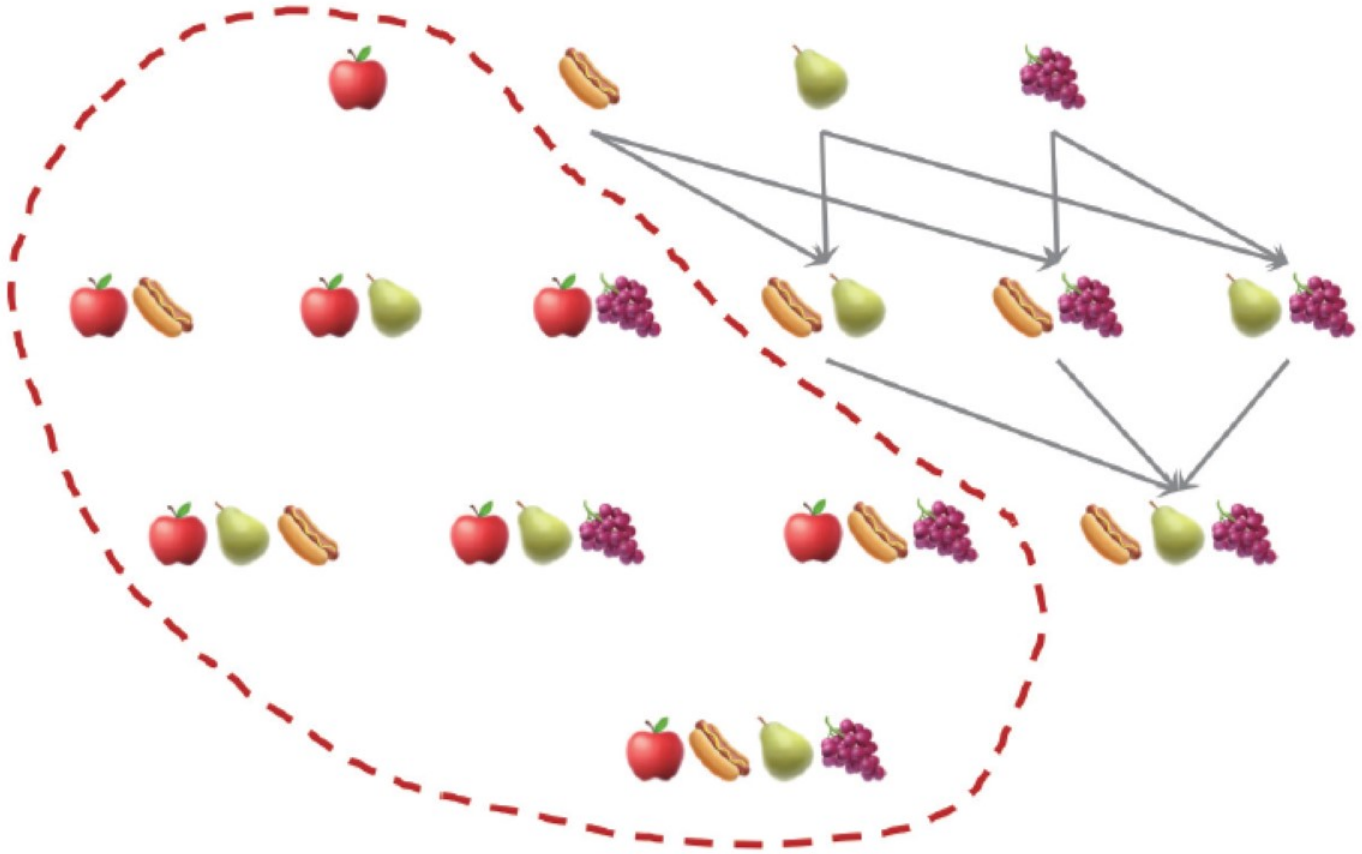
단점

- 각 품목간 거듭제곱 형태로 연산량이 증가함
- 연관에 대한 인과관계 해석이 곤란 (검증 필수)

연관 규칙 (Association Rules)

Ariori 원칙

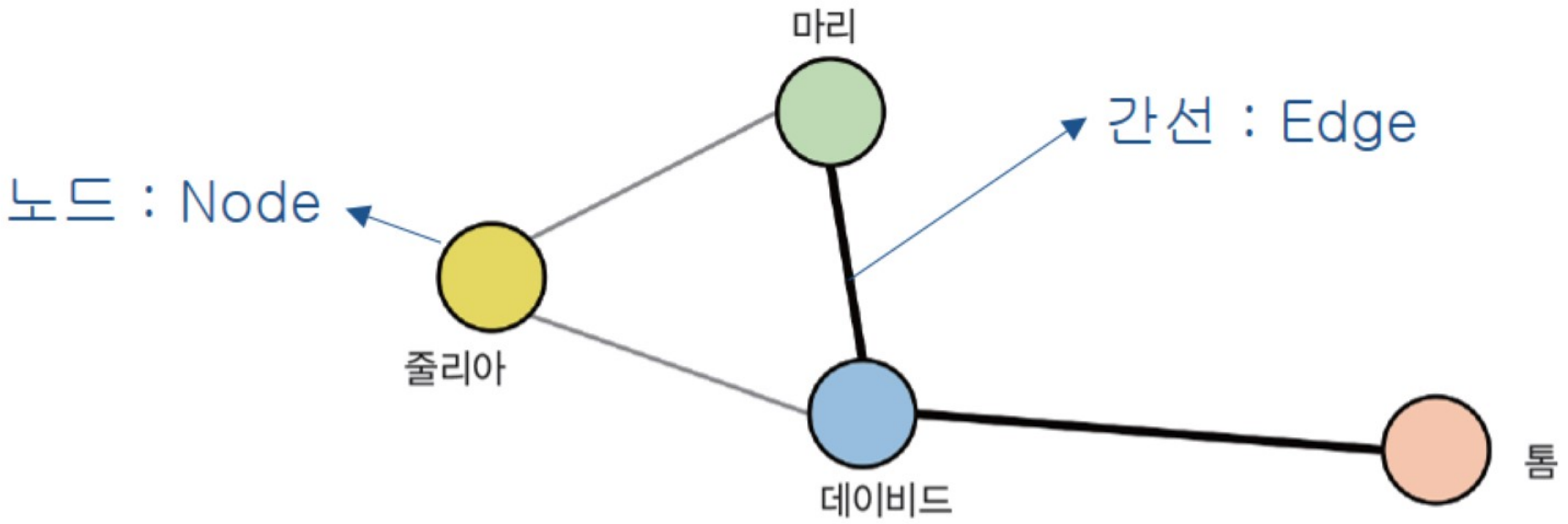
- 특정 품목 집합의 빈도가 낮으면,
  해당 품목을 포함하는 더 큰 특정 품목 집합 또한 낮음



소셜 네트워크 분석 (SNA)

■ 특정 그룹 안에서 영향력 있는 사람과 그 사람이 그룹을 어떻게 이끄는가?

- 특정 인플루언서 (Influencer) 에게 인기 있는 사람은 어떠한 성향을 가지고 있는가?
- COVID-19 관련하여 XX 지역 n 번째 환자의 이동 동선과 겹칠 확률이 높은 사람은 어떠한 사람들인가?
- 탑 / 미드 / 바텀 에 각각 어울리는 사람은 어떠한 성격인가?

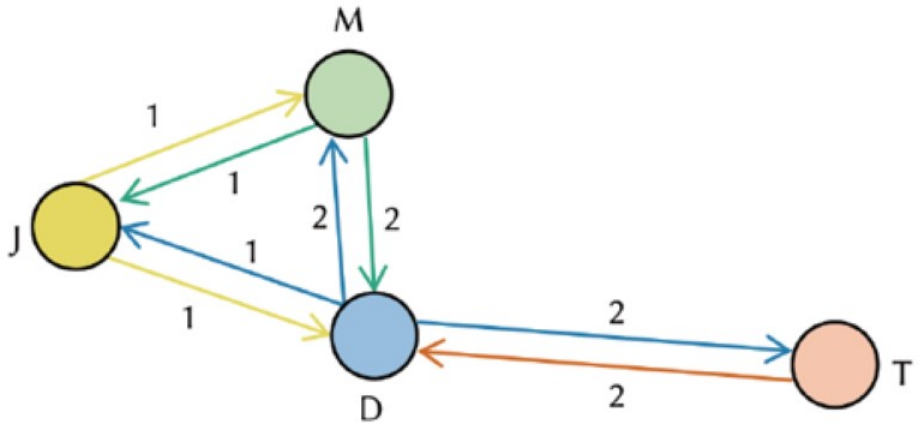


소셜 네트워크 분석 (SNA)

■ 페이지랭크 (PageRank) 알고리즘

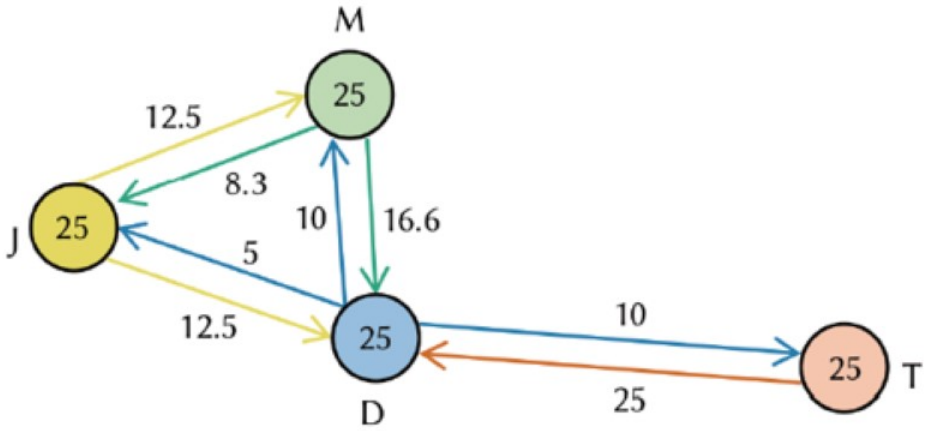
- 구글 (Google) 에서 사용하는 웹 사이트 순위 부여 알고리즘
  - 링크의 수 : 인용 / Citation
  - 링크의 강도
  - 링크의 출처 : 역링크 / Backlink

■ 4개의 페이지와 각 페이지별 링크의 수

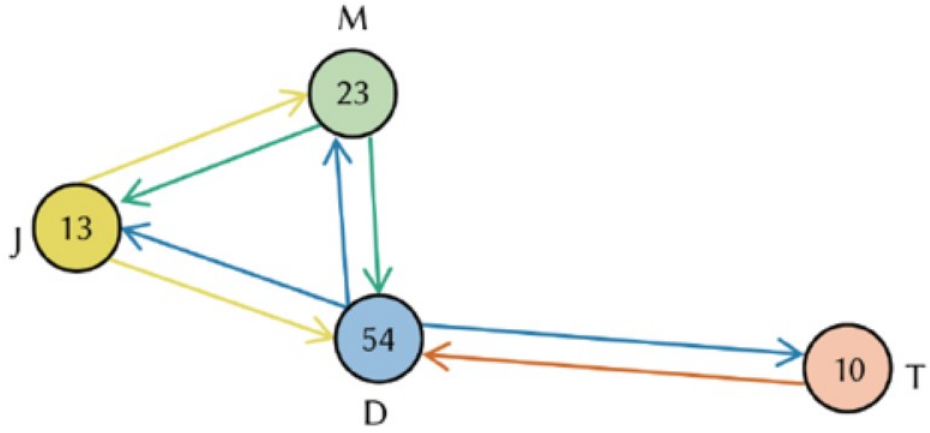


소셜 네트워크 분석 (SNA)

■ 전체 유입자 100명을 기준으로, 그에 따른 간선 가중치



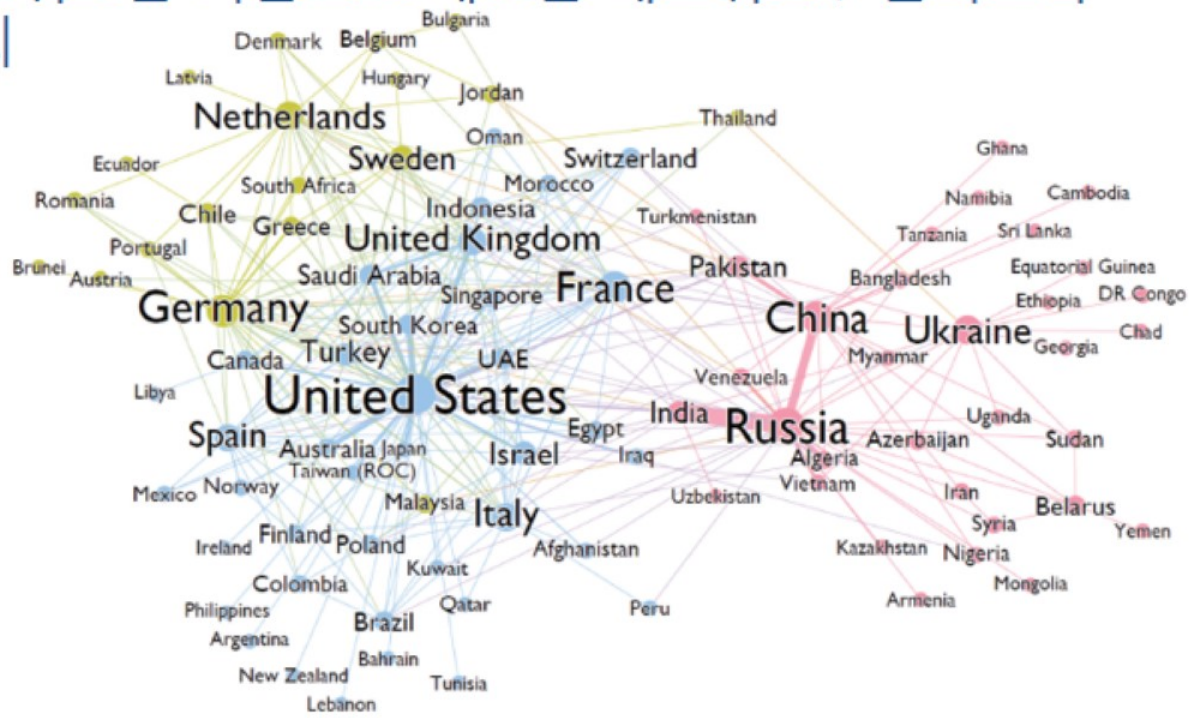
■ 가중치를 고려한 각 노드 점수



소셜 네트워크 분석 (SNA)

장점

- 생성한 네트워크를 기반으로 새로운 네트워크 / 클러스터 생성에 유리



단점

- 오래된 노드 일 수록 점수가 높음 (= 새로운 노드는 불리함)
- 각 노드 간 관계가 고려되지 않음


Q & A