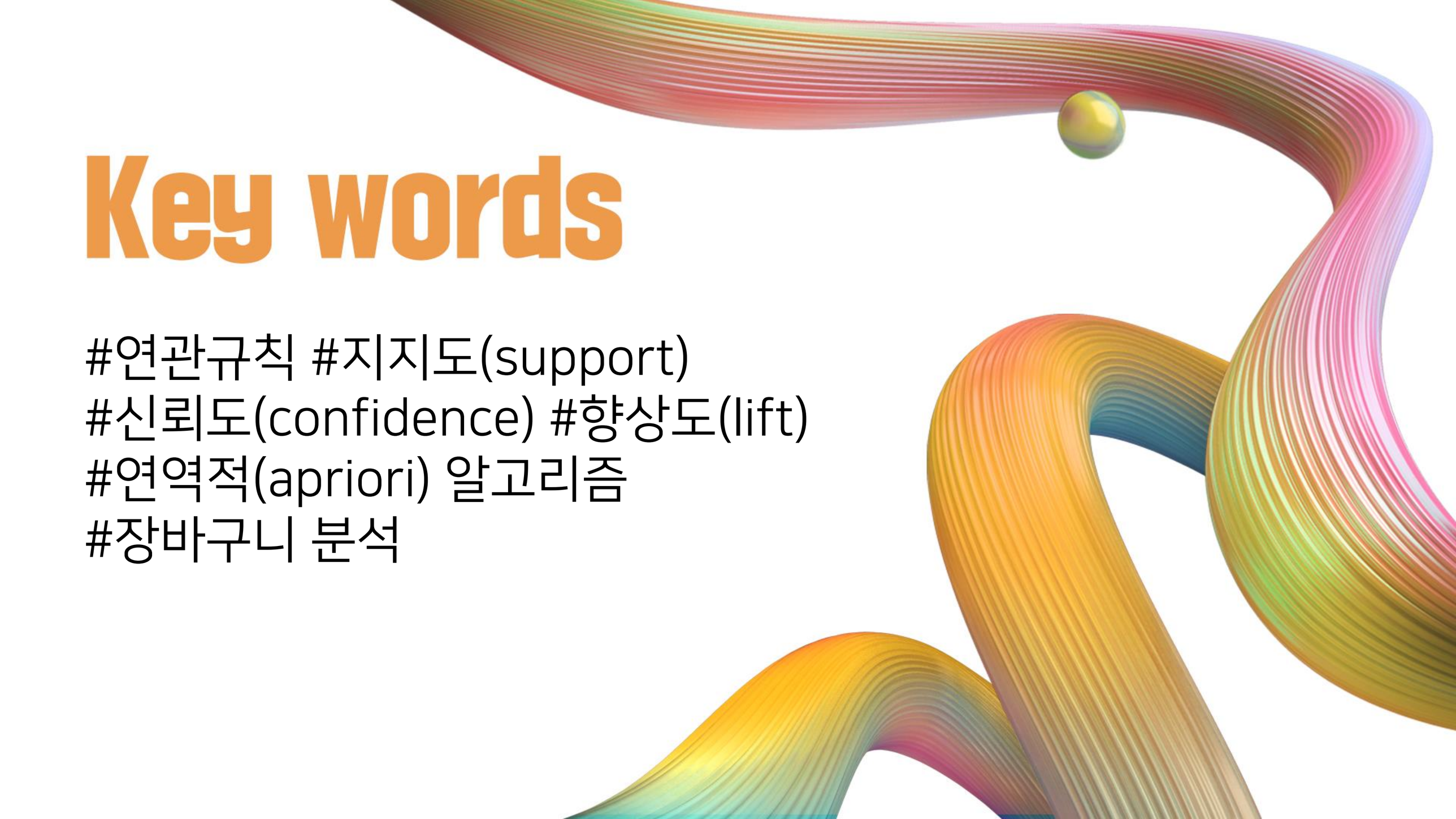
추천: 연관성 분석
(Association Rule)

Key words

#연관규칙 #지지도(support)
#신뢰도(confidence) #향상도(lift)
#연역적(apriori) 알고리즘
#장바구니 분석



연관성 분석 개요

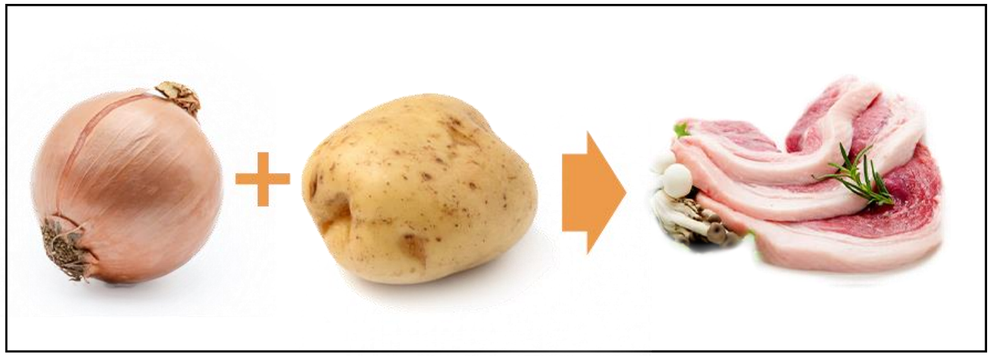
연관성 분석(Association Analysis)

- 연관성 규칙을 통해 하나의 거래나 사건에 포함되어 있는 둘 이상의 품목 간 상호 연관성을 발견하는 과정.
- 고객이 동시에 구매하는 상품 간의 관계를 분석한다는 의미에서 장바구니 분석(Market basket analysis)라고도 함.

연관성 분석 개요

연관규칙(Association Rule)

- 항목들 간의 **if** item A → **then** item B 형태로 표현되는 유용한 패턴.
- 예. {onion, potato} ⇒ {meat}



연관규칙의 탐색 및 평가

고객의 구매시점에 기록된 트랜젝션 자료를 분석.

- 장바구니 하나에 포함된 품목(item)들의 집합 형태로 주어짐.

영수증번호	구매 상품(item)
1	우유, 감자, 시리얼
2	와인, 치즈
3	시리얼, 우유
4	우유, 감자, 소고기
5	맥주, 와인, 치즈

연관규칙의 탐색 및 평가

연관규칙을 파악하기 위한 척도

- 지지도(support) 객관적 척도
  - 전체 구매 건수 가운데 상품 A와 B를 동시에 구매한 비율로, P[A * B]로 나타냄.
  - 지지도(A→B) : $\frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{전체 거래 수}}$
  - 상품 A 하나에 대한 지지도는 지지도(A) : $\frac{\text{A의 거래 수}}{\text{전체 거래 수}}$

연관규칙의 탐색 및 평가

연관규칙을 파악하기 위한 척도

- 신뢰도(confidence)
  - 상품 A를 구매한 건수 가운데 B도 같이 구매한 비율로, 조건부 확률, P[B|A]로 나타냄.
  - 신뢰도(A→B) : $\frac{\text{A와 B가 동시에 포함된 거래 수}}{\text{A의 거래 수}} = \frac{\text{지지도}_{(A \to B)}}{\text{지지도}_{(A)}}$

$$\frac{P(A \cap B)}{P(A)}$$

연관규칙의 탐색 및 평가

연관규칙을 파악하기 위한 척도

- 향상도(lift)

- 전체에서 상품 B를 구매한 비율에 비해, A를 구매한 고객이 B를 구매한 비율이 몇 배인가를 의미하며, P[B|A]/P[B] 로 나타냄.

- 향상도(A→B) : $\frac{\text{신뢰도}_{(A \rightarrow B)}}{\text{지지도}_{(B)}}$

- 향상도의 해석
  - 향상도 (A→B) = 1 : 상품 A와 상품 B의 구매는 상호 연관성이 없음.
  - 향상도 (A→B) > 1 : 상품 A와 상품 B의 구매는 양(+)의 영향력이 있음.
  - 향상도 (A→B) < 1 : 상품 A와 상품 B의 구매는 음(-)의 영향력이 있음.

$$P(B) = \frac{P(A \cap B)}{P(A)} = P(B|A)$$

연역적(apriori) 알고리즘

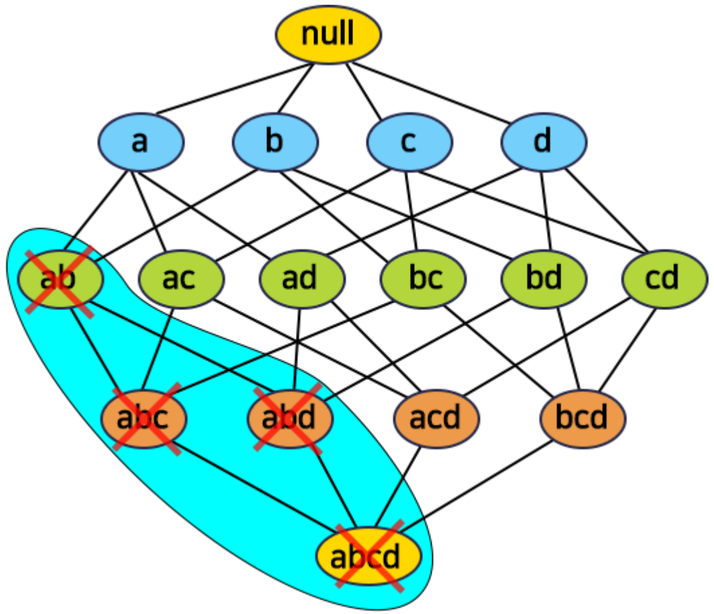
연역적 알고리즘

- 품목들의 집합 별로 지지도, 신뢰도, 향상도 지표를 구해야 하는데,
  품목의 수가 많을 때는 연관규칙의 탐색 비용이 크게 증가함.
- 연역적 알고리즘은 더 이상 탐색하지 않아도 될 품목의 조합을 찾고,
  그 조합을 부분집합으로 갖는 품목의 집합들을 가지치기(pruning)하여,
  효율적인 탐색을 하도록 함.

연역적(apriori) 알고리즘

연역적 알고리즘

- 최소 지지도 가지치기(minimum support pruning, MSP)
  : 어떤 품목(집합)에 대한 지지도가 일정수준
  (문턱기준, threshold criterion)을 넘지 못하면 그 품목(집합)이
  포함된 조합들은 더 이상 탐색하지 않음.



연역적(apriori) 알고리즘

연역적 알고리즘 예시

영수증번호	구매 상품(item)
1	{A, B}
2	{B, C}
3	{A, B, C, D}
4	{A, B, D}
5	{B, D}
6	{B, C, D}
7	{C, D}

- 최소 지지도 = 3/7을 기준으로 분석.

연역적(apriori) 알고리즘

연역적 알고리즘 예시

1) 개별 품목의 지지도 계산

품목	지지도
A	3/7
B	6/7
C	4/7
D	5/7

- 모두 최소지지도 이상임.

연역적(apriori) 알고리즘

연역적 알고리즘 예시

2) 2개씩 짝지어진 품목에 대한 지지도 계산 $_4C_2$

품목	지지도
{A, B}	3/7
{A, C}	1/7
{A, D}	2/7
{B, C}	3/7
{B, D}	4/7
{C, D}	3/7

- {A, C}와 {A, D}는 최소지지도를 넘지 못함(비빈발품목).
  이 품목집합을 포함하는 모든 집합은 비빈발 품목집합으로 처리함.

연역적(apriori) 알고리즘

연역적 알고리즘 예시

3) 2개씩 짝지어진 품목에 대한 지지도 계산

품목	지지도
{B, C, D}	2/7

- {A, B, C}, {A, C, D}는
  비빈발 품목집합이므로 이미 제외되었음.
- {B, C, D}도 최소지지도를 넘지 못함.
  따라서 4개 품목 등은 더 이상 고려되지 않고 알고리즘은 여기서 중단됨.

연역적(apriori) 알고리즘

연역적 알고리즘 예시

4) 규칙 정리

- {A, B}, {B, C}, {B, D}, {C, D}에 대하여 신뢰도, 향상도를 구한 뒤 연관규칙을 판단.

예) 연관규칙 'C→D'의 경우,

- 신뢰도(C→D)=(3/7)/(4/7)=0.75
- 향상도(C→D)=(3/4)/(5/7)=1.05

향상도가 1을 넘기 때문에 품목 C와 품목 D의 구매는 양의 상관이 있다고 해석.