Association Rule

기본 개념

고객들이
자주 같이 구매하는
품목은 무엇인가?

연관규칙분석

- 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙
- IF X then Y $(X \Rightarrow Y)$
- 자료에 존재하는 항목(item)들 간의 If-then 식의 연관 규칙 발견
- 비지도학습(unsupervised learning)의 일종
- 상품의 구매, 서비스 등 일련의 거래 또는 사건들, 상품 추천, 웹페이지간의 링크에 대한 분석
- 장바구니분석라고도 부름(유통업)

평가척도 - 지지도 (Support)

- 관련성이 있다고 판단되는 품목들을 포함하고 있는 거래나 사건의 확률
- 지지도 $(A \Rightarrow B) = \dfrac{A, B\text{를 모두 포함하는 거래의 수}}{\text{전체 거래의 수}}$
- 지지도$(A \Rightarrow B)$ = 지지도$(B \Rightarrow A)$: 상호 대칭적
- A나 B 둘 중 하나라도 포함될 확률이 낮은 경우 지지도가 작게 추정
- 실제 연관성이 높더라도 잘 잡아내지 못함
- 전체적인 구매도에 대한 경향 파악

평가척도 - 신뢰도 (Confidence)

- A를 구입하였을 경우 B를 구입하는 확률(조건부 확률)
- 신뢰도 $(A \Rightarrow B) = \dfrac{A, B\text{를 모두 포함하는 거래의 수}}{A\text{를 포함하는 거래의 수}} = P(B|A)$
- 신뢰도$(A \Rightarrow B) \neq$신뢰도$(B \Rightarrow A)$: 비대칭적
- 둘 중 하나가 포함될 확률이 낮은 경우 연관성을 잘 찾아내지 못하는 지지도의 단점을 보완

평가척도 - 예

맥주	맥주	맥주	맥주	콜라	콜라	컵라면	맥주	맥주	맥주	컵라면	맥주
치킨	치킨	치킨	치킨	맥주	맥주	김치	치킨	치킨	치킨	김치	치킨

- $\{\text{컵라면}\} \Rightarrow \{\text{김치}\}$
- 지지도 : 컵라면과 김치의 판매가 동시에 일어날 확률

$$\frac{\text{컵라면과 김치를 모두 포함하는 거래의 수}}{\text{전체 거래의 수}} = \frac{2}{12} = \frac{1}{6}$$

- 신뢰도 : 컵라면을 구입하였을 경우 김치를 구입하는 확률

$$\frac{\text{컵라면과 김치를 모두 포함하는 거래의 수}}{\text{컵라면을 포함하는 거래의 수}} = \frac{2}{2} = 1$$

평가척도 - 향상도 (Lift)

- 우연에 의한 발생에 비해 연관성이 강한지 나타내는 측도
- 향상도$(A \Rightarrow B) = \dfrac{P(B|A)}{P(B)} = \dfrac{P(A \cap B)}{P(A)P(B)}$
- 향상도 $= 1$: A와 B가 독립
- 향상도 > 1 : 규칙이 의미가 있음
- 향상도 < 1 : 규칙이 의미가 없음

평가척도 - 예

맥주	맥주	맥주	맥주	콜라	콜라	컵라면	맥주	맥주	맥주	컵라면	맥주
치킨	치킨	치킨	치킨	맥주	맥주	김치	치킨	치킨	치킨	김치	치킨

- $\{\text{컵라면}\} \Rightarrow \{\text{김치}\}$
- 향상도 : 컵라면과 김치의 판매가 동시에 일어날 확률

$$\frac{\text{컵라면을 구입하였을 경우 김치를 구입하는 확률}}{\text{김치를 구입하는 확률}} = \frac{1}{2/12} = 6$$

- 향상도가 6으로 1에 비해 매우 큰 값을 보이므로 이는 우연에 의한 일이 아닐 가능성이 큼

연관규칙분석 - 예제

ID	판매상품
1	소주, 콜라, 맥주
2	소주, 콜라, 와인
3	소주, 주스
4	콜라, 맥주
5	소주, 콜라, 맥주, 와인
6	주스

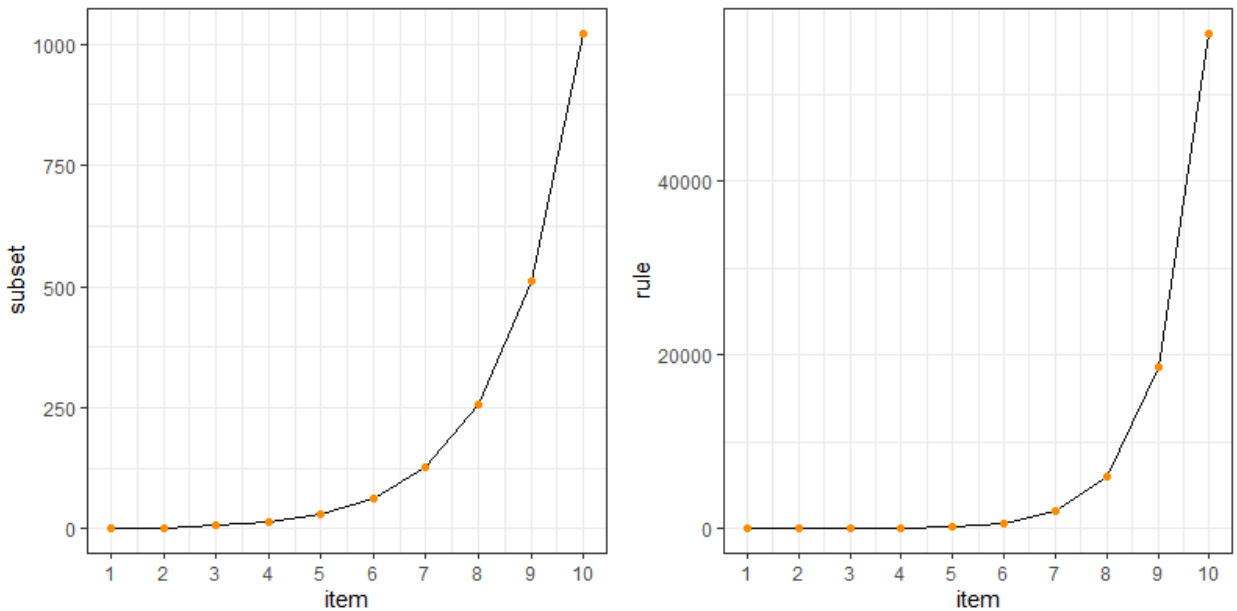
지지도 50% 이상인 규칙	해당 transaction	신뢰도
소주 ⇒ 콜라	1,2,5	75%
콜라 ⇒ 맥주	1,4,5	75%
맥주 ⇒ 콜라	1,4,5	100%

- 연관규칙 : 맥주를 구입한 사람은 모두 콜라를 구매한다. (100%)
- 지지도 : 이러한 경향을 갖는 사람은 전체의 50%정도이다.
- 향상도 $= \frac{P(\text{콜라}|\text{맥주})}{P(\text{콜라})} = \frac{1}{4/6} = 1.5$

  - 맥주 구매 시 콜라를 구입하게 될 가능성은 맥주 구매가 전제되지 않았을 경우보다 1.5배 높아진다.

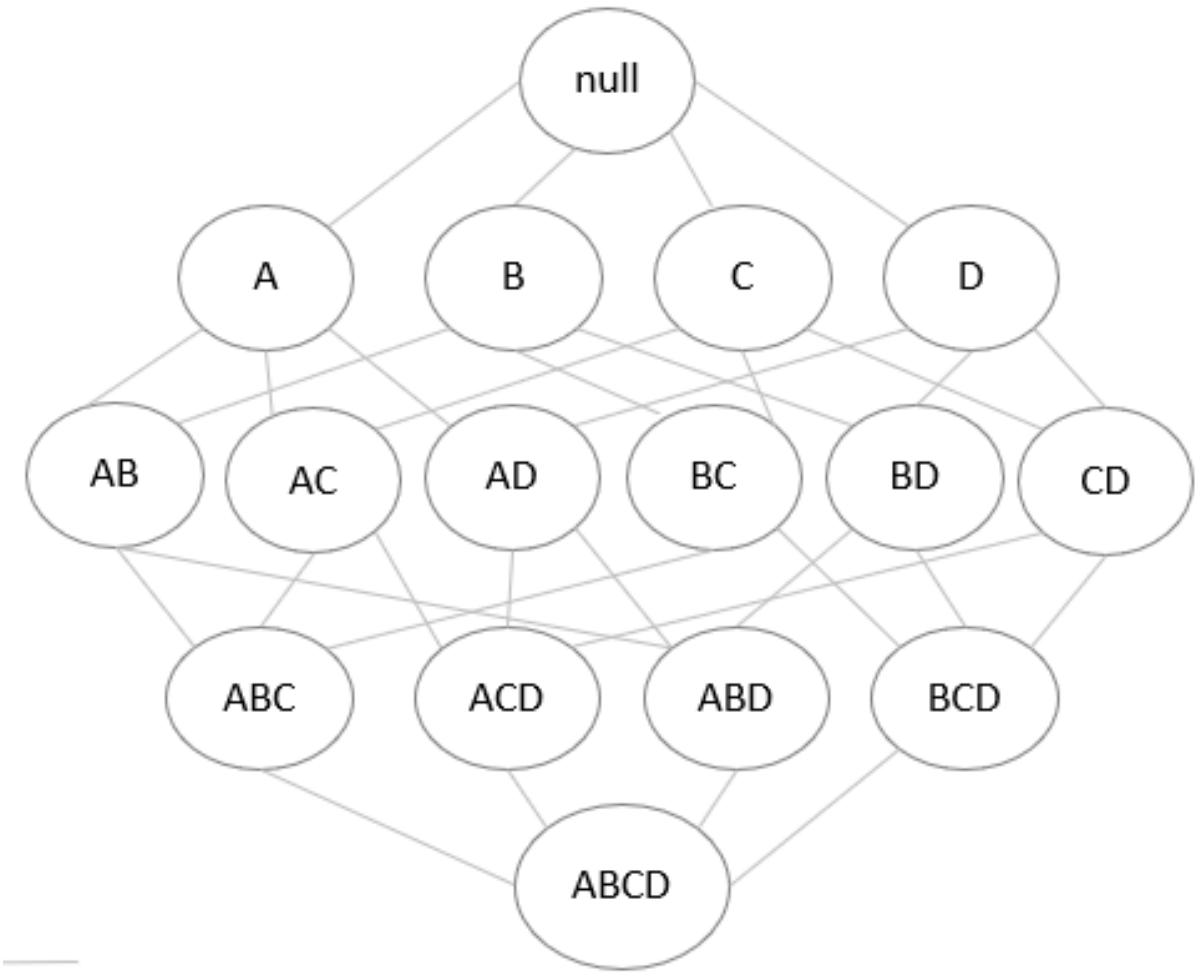
연관규칙분석

- 품목이 많은 경우 모든 연관 규칙을 찾는 것은 불가능
- 예를 들어 k개의 품목(item)이 있다면
  - 품목집합 수 : $2^k - 1$
  - 연관규칙 수 : $3^k - 2^{k+1} + 1$



효과적인 알고리즘 필요

- $k = 4$



Apriori 알고리즘

- 빈발항목집합: 최소 지지도 이상을 갖는 항목 집합
- 모든 항목 집합에 대한 측도(지지도, 신뢰도, 향상도)를 구하는 대신 최소 지지도 이상의 빈발항목집합만 찾아내서 연관 규칙을 계산
- 기본원리 :
  - 빈발항목집합 $\Rightarrow$ 개별 부분집합도 빈발
  - 비빈발항목집합 $\Rightarrow$ 모든 상위집합이 비빈발
- 연관규칙 마이닝 분석

  1. 빈발항목 찾기 $\rightarrow$ 2. 빈발항목에서 강한 연관규칙 생성

Apriori 알고리즘 - 예제

- (예제) 빈발패턴 생성

TID	구매항목
T10	A,B,E
T20	B,D
T30	B,C
T40	A,B,D
T50	A,C
T60	B,C
T70	A,C
T80	A,B,C,E
T90	A,B,C

Apriori 알고리즘 - 예제

- (예제) 빈발패턴 생성

TID	구매항목
T10	A,B,E
T20	B,D
T30	B,C
T40	A,B,D
T50	A,C
T60	B,C
T70	A,C
T80	A,B,C,E
T90	A,B,C

C_1

항목	지지도
A	6
B	7
C	6
D	2
E	2

L_1

항목	지지도
A	6
B	7
C	6
D	2
E	2

C_2

항목
{A,B}
{A,C}
{A,D}
{A,E}
{B,C}
{B,D}
{B,E}
{C,D}
{C,E}
{D,E}

C_2

항목	지지도
{A,B}	4
{A,C}	4
{A,D}	1
{A,E}	2
{B,C}	4
{B,D}	2
{B,E}	2
{C,D}	0
{C,E}	1
{D,E}	0

L_2

항목	지지도
{A,B}	4
{A,C}	4
{A,E}	2
{B,C}	4
{B,D}	2
{B,E}	2

C_3

항목
{A,B,C}
{A,B,E}

C_3

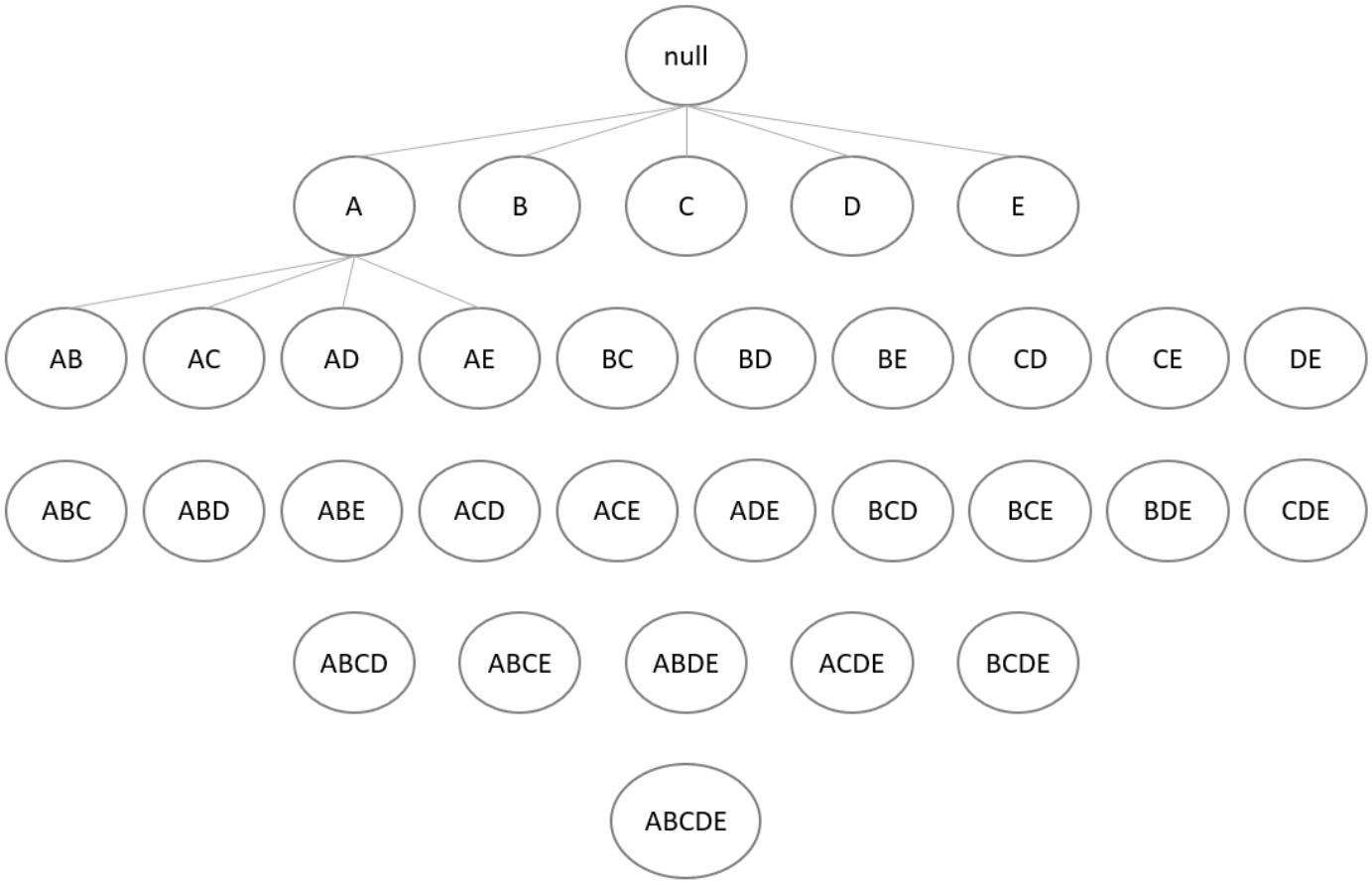
항목	지지도
{A,B,C}	2
{A,B,E}	2

L_3

항목	지지도
{A,B,C}	2
{A,B,E}	2

Apriori 알고리즘 - 예제

- Apriori 알고리즘



Apriori 알고리즘 - 예제

- (예제) 연관규칙 생성

TID	구매항목
T10	A,B,E
T20	B,D
T30	B,C
T40	A,B,D
T50	A,C
T60	B,C
T70	A,C
T80	A,B,C,E
T90	A,B,C

항목	지지도
A	6
B	7
C	6
D	2
E	2

항목	지지도
{A,B}	4
{A,C}	4
{A,E}	2
{B,C}	4
{B,D}	2
{B,E}	2

빈발패턴	지지도
{A,B,C}	2
{A,B,E}	2

{A,B,E} => {A}, {B}, {E}, {A,B}, {A,C}, {B,E}

- 신뢰도 구하기

{A} ⇒ {B,E} : 신뢰도 = 2/6 = 0.33

{B} ⇒ {A,E} : 신뢰도 = 2/7 = 0.29

{E} ⇒ {A,B} : 신뢰도 = 2/2 = 1

{A,B} ⇒ {E} : 신뢰도 = 2/4 = 0.5

{A,E} ⇒ {B} : 신뢰도 = 2/2 = 1

{B,E} ⇒ {A} : 신뢰도 = 2/2 = 1

FP-Growth Algorithm

- 연관규칙을 Tree구조를 활용하여 나타내는 방식으로 트리와 노드 링크라는 자료구조를 활용
- Apriori의 연산 속도를 개선하기 위해 등장
- Candidate를 만들지 않음으로써 그를 위한 시간과 메모리 절약 가능
- 전체 DB를 2번만 scan함으로써 실행시간을 단축

FP-Growth Algorithm - 예제

- 예제) 빈발패턴 생성

TID	구매항목
T10	A,B,E
T20	B,D
T30	B,C
T40	A,B,D
T50	A,C
T60	B,C
T70	A,C
T80	A,B,C,E
T90	A,B,C

→

항목	지지도개수
A	6
B	7
C	6
D	2
E	2

빈발항목 수집

→

항목	지지도개수
B	7
A	6
C	6
D	2
E	2

내림차순 정렬

→

TID	구매항목
T10	B,A,E
T20	B,D
T30	B,C
T40	B,A,D
T50	A,C
T60	B,C
T70	A,C
T80	B,A,C,E
T90	B,A,C

FP-Growth Algorithm - 예제

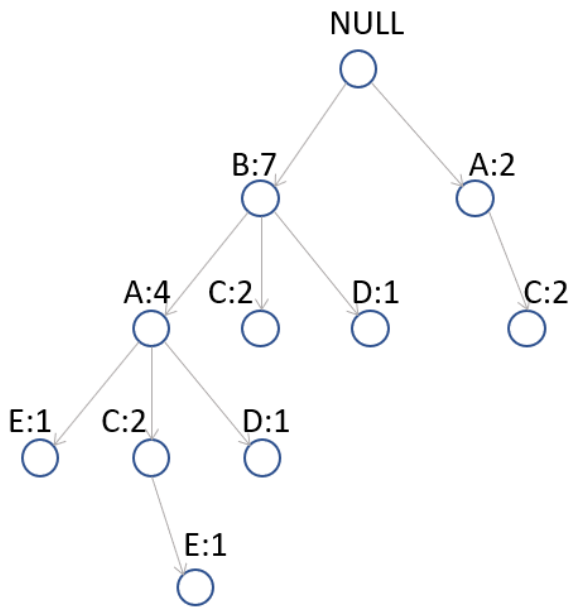
- 예제) 빈발패턴 생성

TID	구매항목
T10	B,A,E
T20	B,D
T30	B,C
T40	B,A,D
T50	A,C
T60	B,C
T70	A,C
T80	B,A,C,E
T90	B,A,C

FP-Growth Algorithm - 예제

- 예제) 빈발패턴 생성

항목	지지도
B	7
A	6
C	6
D	2
E	2



항목	빈발패턴
E	{B,E : 2}, {A,E : 2}, {A,B,E : 2}
D	{B,D : 2}
C	{B,C, : 4}, {A,C : 4}, {B,A,C : 2}
A	{B,A : 4}