

## Chapter. 12

어디서 많이 봤던 패턴이다 싶을 때: 빈발 패턴 탐색

# | 연관규칙 탐색

FAST CAMPUS  
ONLINE  
데이터 탐색과 전처리 I

강사. 안길승

# I 연관규칙이란?

- “A가 발생하면 B도 발생하더라”라는 형태의 규칙으로, **트랜잭션 데이터**를 탐색하는데 사용
  - A: 부모 아이템 집합 (antecedent)
  - B: 자식 아이템 집합 (consequent)
  - A와 B는 모두 공집합이 아닌 집합이며,  $A \cap B = \emptyset$ 을 만족함 (즉, 공통되는 요소가 없음)
- 규칙 예시

| 분야     | 규칙                       | 해석   | 활용 예시                       |
|--------|--------------------------|--|-----------------------------|
| 마트     | {맥주, 종이컵} → {땅콩}         | 맥주와 종이컵을 <b>구매</b> 하면,<br>땅콩도 <b>구매</b> 하더라              | 맥주, 종이컵, 땅콩을 한 공간에<br>배치하자! |
| 인터넷 쇼핑 | {광고 팝업}<br>→ {다른 사이트 접속} | 광고 팝업이 뜨면 ( <b>이벤트</b> ),<br>다른 사이트로 접속하더라 ( <b>행동</b> ) | 불필요한 광고 팝업을<br>최대한 줄이자      |

# I 연관규칙 탐색이란?

- 트랜잭션 데이터에서 **의미있는 연관규칙을 효율적**으로 탐색하는 작업

| 거래 ID | 구매 아이템    |
|-------|-----------|
| 1     | {A, B}    |
| 2     | {B, C, D} |
| 3     | {A, B, F} |
| 4     | {B}       |
| ⋮     | ⋮         |

트랜잭션 데이터



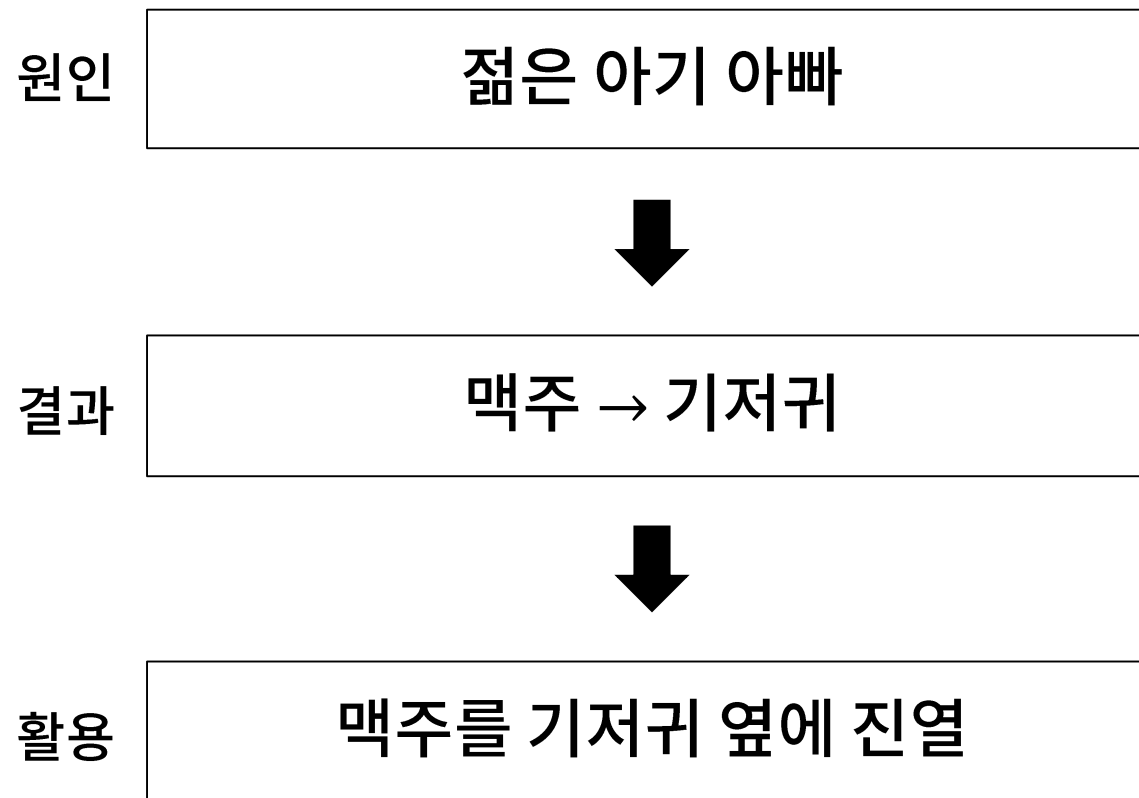
| 규칙 ID | 규칙           | 점수 |
|-------|--------------|----|
| 1     | {A} → {B}    | 85 |
| 2     | {B} → {C}    | 80 |
| 3     | {A, B} → {C} | 60 |
| 4     | {A} → {B}    | 55 |
| ⋮     | ⋮            | ⋮  |

연관 규칙 목록

- 아이템의 개수가  $n$ 개라면, 생성 가능한 연관 규칙의 개수는  $\sum_{k=2}^n \binom{n}{k} \times (2^k - 2)$ 로 매우 많음
- 아이템이 **20개**만 되더라도 **34억 개** 가량의 규칙이 생성되므로 **효율적인 탐색**이 필수적임

## I 연관규칙 탐색의 활용 사례: 월마트

- 월마트에서는 엄청나게 많은 영수증 데이터에 대해 연관규칙 탐색을 적용하여, 매출을 향상시킴



# I 연관규칙의 평가 척도

- 지지도 (support): 아이템 집합이 전체 트랜잭션 데이터에서 발생한 비율

$$S(A \rightarrow B) = \frac{N(A,B)}{n}$$

✓  $n$ : 트랜잭션 데이터 크기  
✓  $N(A, B)$ : 트랜잭션 데이터에서 A와 B의 동시 출현 횟수

- 신뢰도 (confidence): 부모 아이템 집합이 등장한 트랜잭션 데이터에서 자식 아이템 집합이 발생한 비율

$$C(A \rightarrow B) = \frac{N(A,B)}{N(A)}$$

✓  $N(A)$ : 트랜잭션 데이터에서 A의 출현 횟수  
✓  $N(A, B)$ : 트랜잭션 데이터에서 A와 B의 동시 출현 횟수

- 지지도와 신뢰도가 높은 연관규칙을 좋은 규칙이라고 판단

## I 연관규칙의 평가 척도 계산 예시

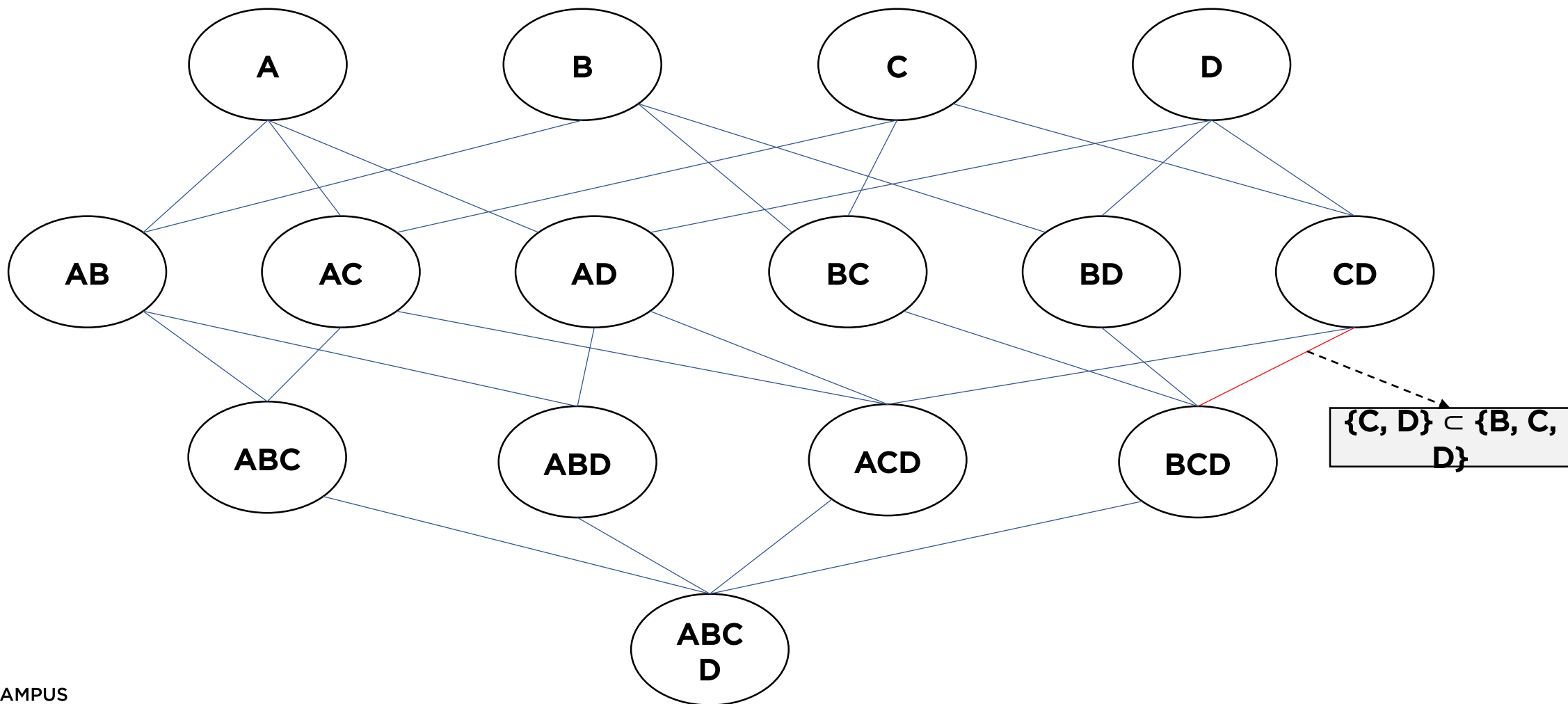
| 거래 ID | 구매 아이템      |
|-------|-------------|
| 1     | {빵, 우유}     |
| 2     | {맥주, 땅콩}    |
| 3     | {빵}         |
| 4     | {빵, 맥주, 땅콩} |
| 5     | {빵, 우유}     |



- $S(\{\text{빵}\} \rightarrow \{\text{우유}\}) = 2 / 5$
- $C(\{\text{빵}\} \rightarrow \{\text{우유}\}) = 2 / 4$

# I 아이템집합 격자

- 아이템 집합과 그 관계를 한 눈에 보여주기 위한 그래프



# I 지지도에 대한 Apriori 원리: 개요

- $S(A \rightarrow B)$ 가 **최소 지지도 (min support)** 이상이면, 이 규칙을 **빈발**하다고 함
- 아이템 집합의 지지도가 최소 지지도 이상이면, 이 집합을 빈발하다고 함
- 지지도에 대한 Apriori 원리: 어떤 아이템 집합이 빈발하면, 이 아이템의 부분 집합도 빈발한다.

$X \subset Y$ 이면,  $S(X) \geq S(Y)$ 가 성립

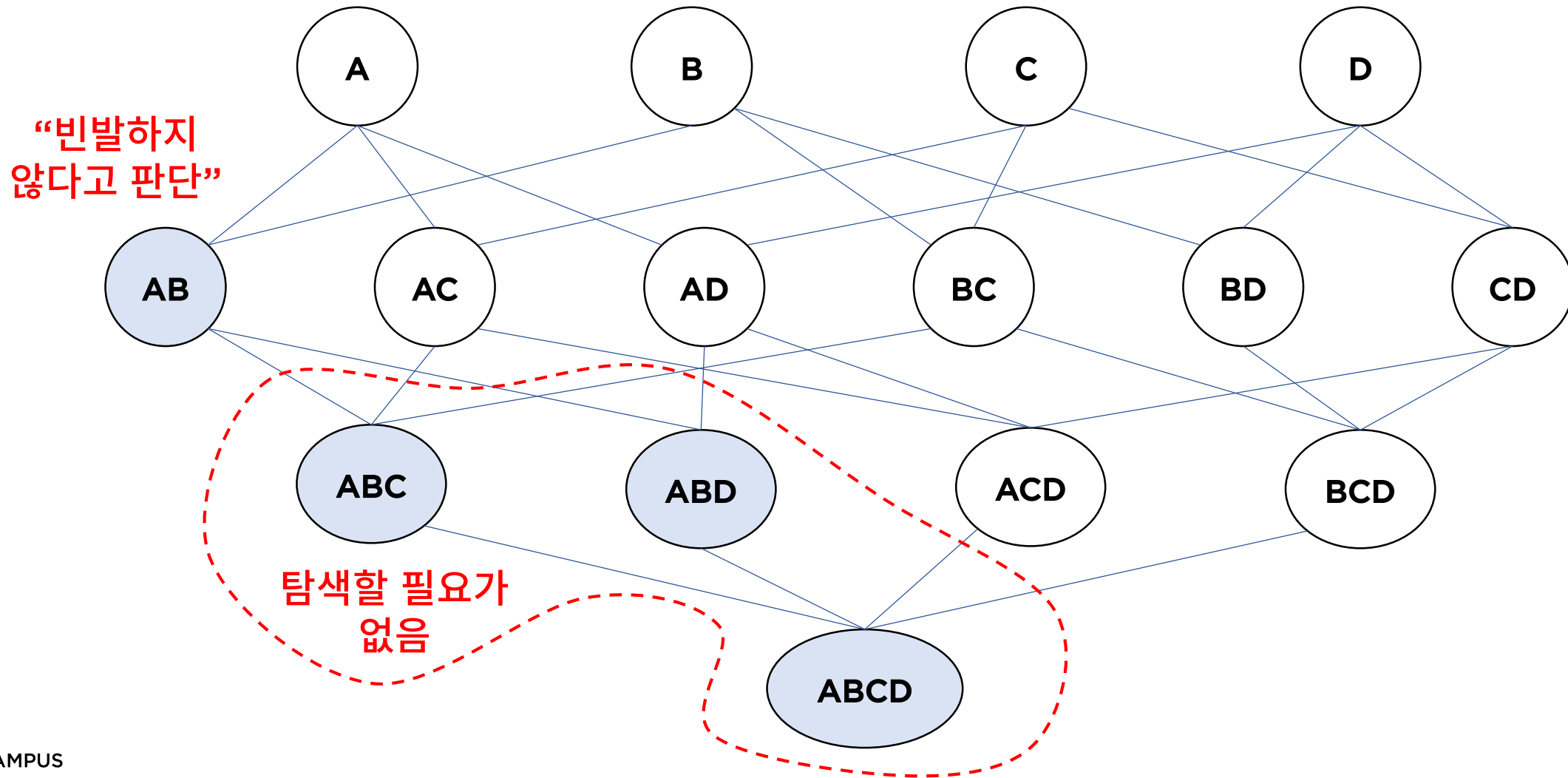
(증명)

$X$ 가  $Y$ 의 부분 집합이면  $N(X) \geq N(Y)$ 가 항상 성립한다.

따라서  $S(X) = \frac{N(X)}{n} \geq S(Y) = \frac{N(Y)}{n}$ 이 성립한다.



# I 지도도에 대한 Apriori 원리: 적용



# I 지도도에 대한 Apriori 원리: 후보 규칙 생성

- Apriori 원리를 사용하여 **모든 최대 빈발 아이템 집합**을 찾은 후, 후보 규칙을 모두 생성함
  - 최대 빈발 아이템 집합: 최소 지도도 이상이면서, 이 집합의 모든 모집합이 빈발하지 않는 집합
  - (예시) {A, B, C}가 빈발한데, {A, B, C, D}, {A, B, C, E} 등이 빈발하지 않으면, {A, B, C}를 최대 빈발 아이템 집합이라고 함
- 만약, {A, B, C}가 최대 빈발아이템 집합이라면, 생성 가능한 후보 규칙은 다음과 같음

| 부모  | 자식     |
|-----|--------|
| {A} | {B}    |
|     | {C}    |
|     | {B, C} |

| 부모  | 자식     |
|-----|--------|
| {B} | {A}    |
|     | {C}    |
|     | {A, C} |

| 부모  | 자식     |
|-----|--------|
| {C} | {A}    |
|     | {B}    |
|     | {B, C} |

| 부모     | 자식  |
|--------|-----|
| {A, B} | {C} |
| {C, A} | {B} |
| {B, C} | {A} |

# I 신뢰도에 대한 Apriori 원리

- 동일한 아이템 집합으로 생성한 규칙  $X_1 \rightarrow Y_1$ 과  $X_2 \rightarrow Y_2$ 에 대해서, 다음이 성립함

$X_1 \subset X_2$ 이면,  $C(X_1 \rightarrow Y_1) \leq C(X_2 \rightarrow Y_2)$ 가 성립

(증명)

$$C(X_1 \rightarrow Y_1) = \frac{N(X_1, Y_1)}{N(X_1)} \text{이고 } C(X_2 \rightarrow Y_2) = \frac{N(X_2, Y_2)}{N(X_2)} \text{이다.}$$

그런데 두 규칙이 동일한 아이템 집합으로부터 생성되었으므로,  $X_1 \cup Y_1 = X_2 \cup Y_2$ 이 성립한다.

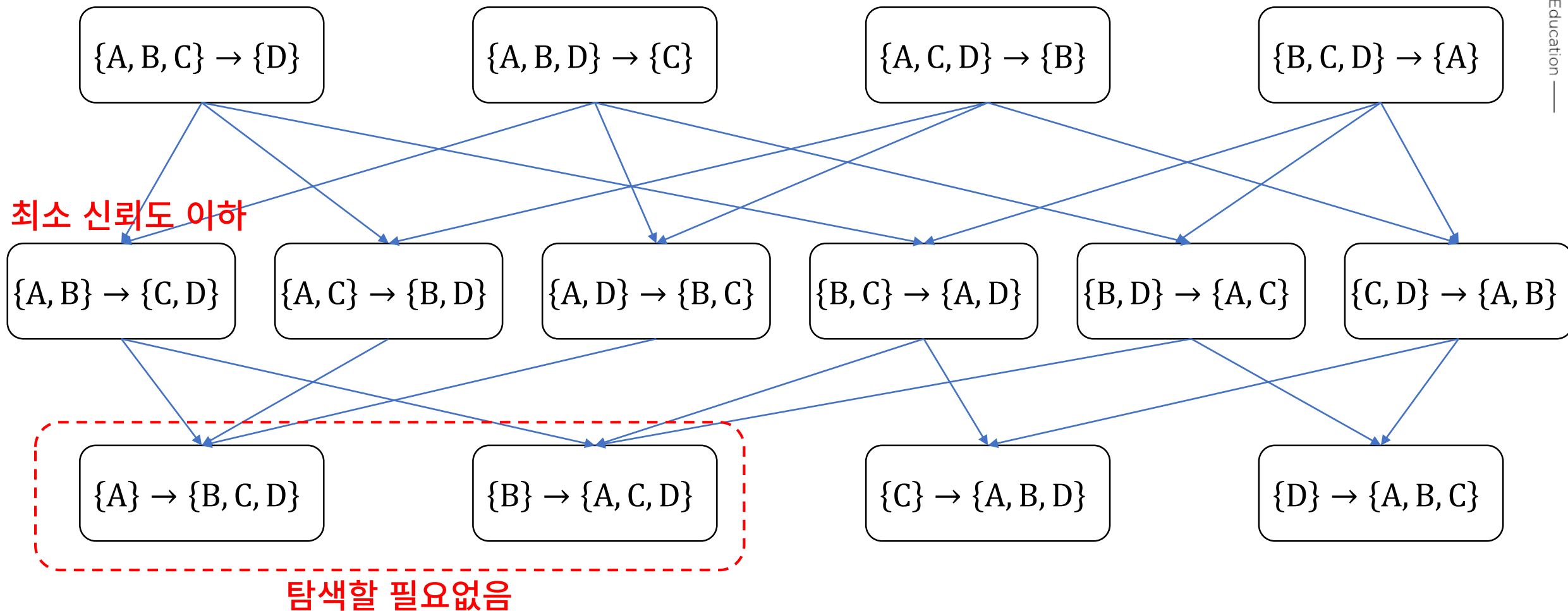
즉,  $N(X_1, Y_1) = N(X_2, Y_2)$ 가 성립한다.

또한,  $X_1 \subset X_2$ 이므로,  $N(X_1) \geq N(X_2)$ 가 성립한다.

따라서  $C(X_1 \rightarrow Y_1)$ 과  $C(X_2 \rightarrow Y_2)$ 의 분자가 같고, 분모는  $C(X_2 \rightarrow Y_2)$ 이 더 작으므로,

$C(X_1 \rightarrow Y_1) \leq C(X_2 \rightarrow Y_2)$ 이 성립한다.

# I 신뢰도에 대한 Apriori 원리: 적용



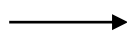
## I 관련 모듈: mlxtend

- apriori 함수를 이용한 빈발 아이템 집합 탐색과 association\_rules 함수를 이용하여 연관규칙을 탐색하는 두 단계로 수행
- mlxtend.frequent\_patterns.apriori(df, min\_support):
  - df: one hot encoding 형태의 데이터 프레임
  - min\_support: 최소 지지도
- mlxtend.frequent\_patterns.association\_rules(frequent\_dataset, metric, min\_threshold):
  - frequent\_dataset에서 찾은 연관 규칙을 데이터 프레임 형태로 반환
  - metric: 연관규칙을 필터링하기 위한 유용성 척도 (default: confidence)
  - min\_threshold: 지정한 metric의 최소 기준치

# mlxtend.preprocessing.TransactionEncoder

- 연관규칙 탐색에 적절하게 거래 데이터 구조를 바꾸기 위한 함수
- 인스턴스 생성 후, `fit(data).transform(data)`를 이용하여 `data`를 각 아이템의 출현 여부를 갖는 `ndarray` 및 `DataFrame`으로 변환

| TID | Items                     |
|-----|---------------------------|
| 1   | Bread, Milk               |
| 2   | Bread, Diaper, Beer, Eggs |
| 3   | Milk, Diaper, Beer, Coke  |
| 4   | Bread, Milk, Diaper, Beer |
| 5   | Bread, Milk, Diaper, Coke |



| TID | Bread | Milk | Diaper | Beer | Eggs | Coke |
|-----|-------|------|--------|------|------|------|
| 1   | 1     | 1    | 0      | 0    | 0    | 0    |
| 2   | 1     | 0    | 1      | 1    | 1    | 0    |
| 3   | 0     | 1    | 1      | 1    | 0    | 1    |
| 4   | 1     | 1    | 1      | 1    | 0    | 0    |
| 5   | 1     | 1    | 1      | 0    | 0    | 1    |

## Chapter. 12

어디서 많이 봤던 패턴이다 싶을 때: 빈발 패턴 탐색

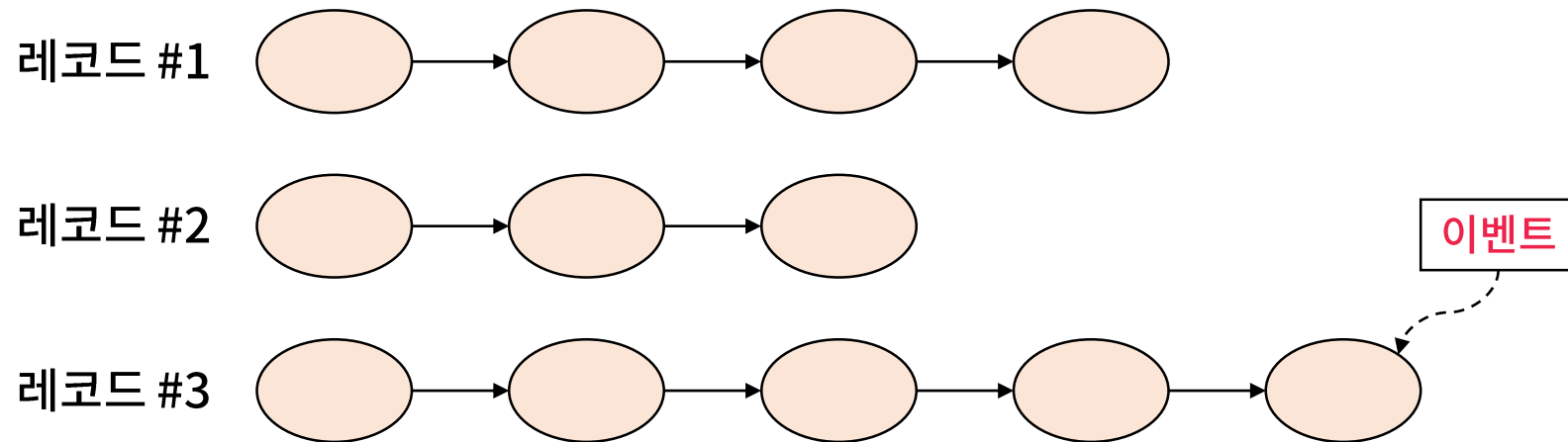
# | 빈발 시퀀스 탐색

FAST CAMPUS  
ONLINE  
데이터 탐색과 전처리 I

강사. 안길승

# I 시퀀스 데이터란?

- 시퀀스 데이터란 각 요소가 **(순서, 값)** 형태로 구성된 데이터로, 분석 시에 **반드시 순서를 고려**해야 함



- 로그 데이터 대부분이 순서가 있는 시퀀스 데이터임
  - 고객 구매 기록
  - 고객 여정
  - 웹 서핑 기록



# I 시퀀스 데이터에서의 빈발 패턴

- 시퀀스 데이터에서의 빈발 패턴은 반드시 순서가 고려되어야 함

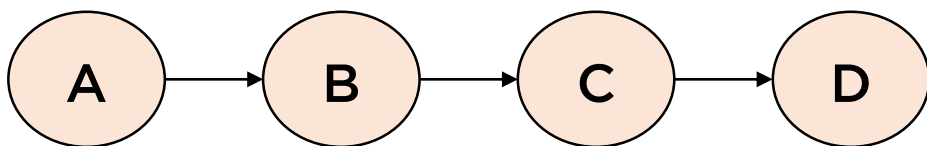
시퀀스 데이터가 아닌 경우에는  
 $\{A, B\} = \{B, A\}$ 가 같음

| ID | 기록        |
|----|-----------|
| #1 | {A, B}    |
| #2 | {B, C, D} |
| #3 | {B, A, F} |
| #4 | {B}       |

- 비시퀀스 데이터에서  $A \rightarrow B$ 가 출현한 ID: {#1, #3}
- 시퀀스 데이터에서  $A \rightarrow B$ 가 출현한 ID: {#1}

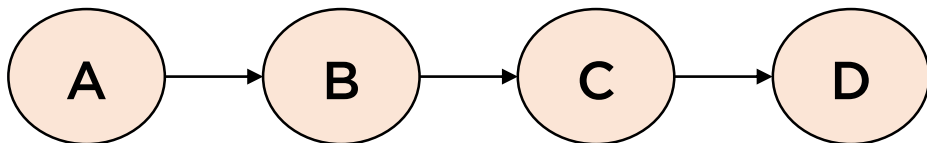
# I 지지도와 신뢰도

- 분석 목적에 따라, 특정 패턴의 등장 여부에 대한 정의가 필요함



- $A \rightarrow B$ 는 출현했음
- $B \rightarrow A$ 와  $E \rightarrow A$ 는 출현하지 않았음
- $A \rightarrow C$ 는 등장했다고 볼 수도, 그렇지 않다고 볼 수도 있음

- 일반적으로, 윈도우 내 (크기  $L$ )에 특정 이벤트가 발생했는지를 기준으로 패턴의 등장 여부를 확인

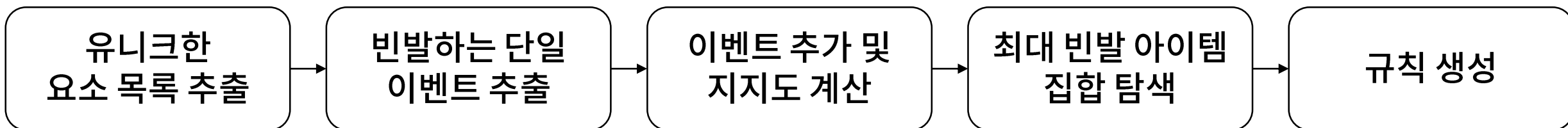


- ( $L = 1$ )  $A \rightarrow B$ 는 등장했으나,  $A \rightarrow C$ 와  $A \rightarrow D$ 는 등장하지 않음
- ( $L = 2$ )  $A \rightarrow B$ 와  $A \rightarrow C$ 는 등장했으나,  $A \rightarrow D$ 는 등장하지 않음
- ( $L = 3$ )  $A \rightarrow B$ ,  $A \rightarrow C$ ,  $A \rightarrow D$  모두 등장함

- 지도도와 신뢰도에 대한 정의는 일반 데이터에 대한 것과 같으나, 출현 횟수를 계산하는 방식이 다름

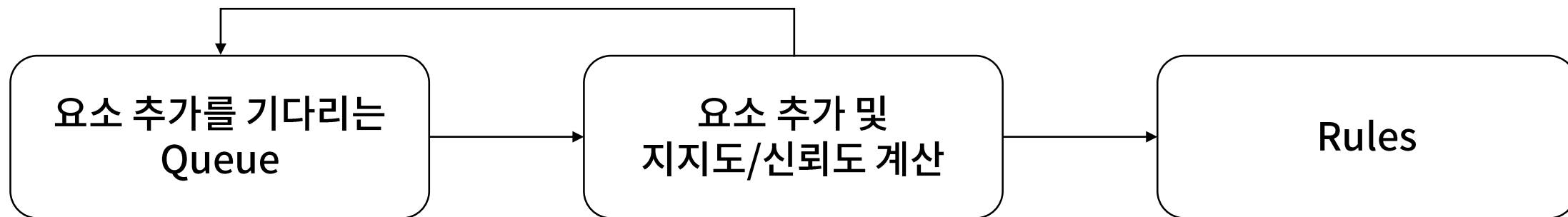
## I 순서를 고려한 연관규칙 탐색

- 시퀀스 데이터에 대한 연관규칙 탐색에 대해서는  $A \rightarrow B$ 와  $B \rightarrow A$ 가 다른 지지도를 갖기 때문에, 같은 항목 집합으로부터 규칙을 생성할 수 없음
- 신뢰도에 대한 apriori 원리는 성립함
- 따라서 개별 요소(이벤트)에 다른 요소를 추가하는 방식으로 규칙을 아래와 같이 **직접** 찾아 나가야 함



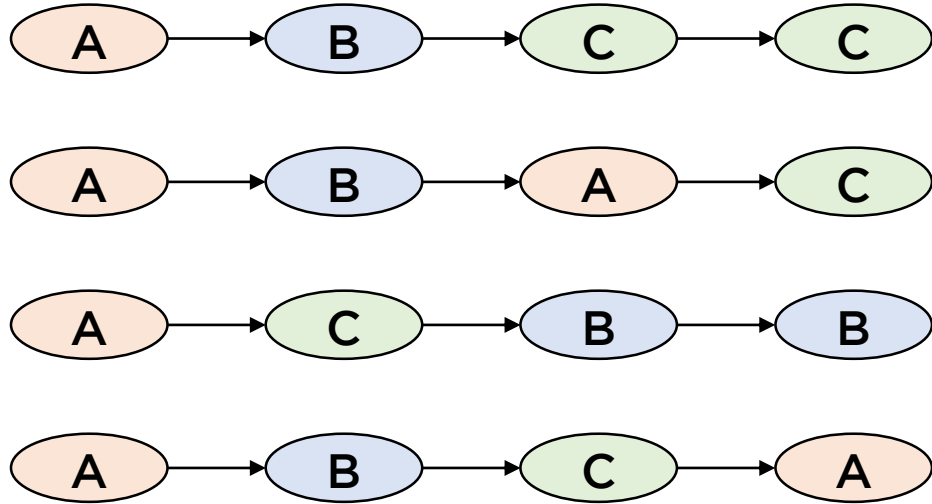
## I 동적 프로그래밍

- 원 문제를 작은 문제로 분할한 다음 **점화식**으로 만들어 **재귀적인 형태**로 원 문제를 해결하는 방식
- 시퀀스 데이터에 대한 연관 규칙 탐사 적용을 위한 동적 프로그래밍 구조



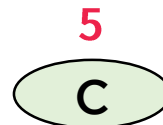
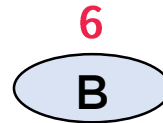
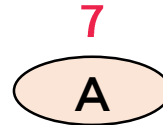
# I 순서를 고려한 연관규칙 탐색 (예시: L = 2, 최소 지지도 = 2)

데이터



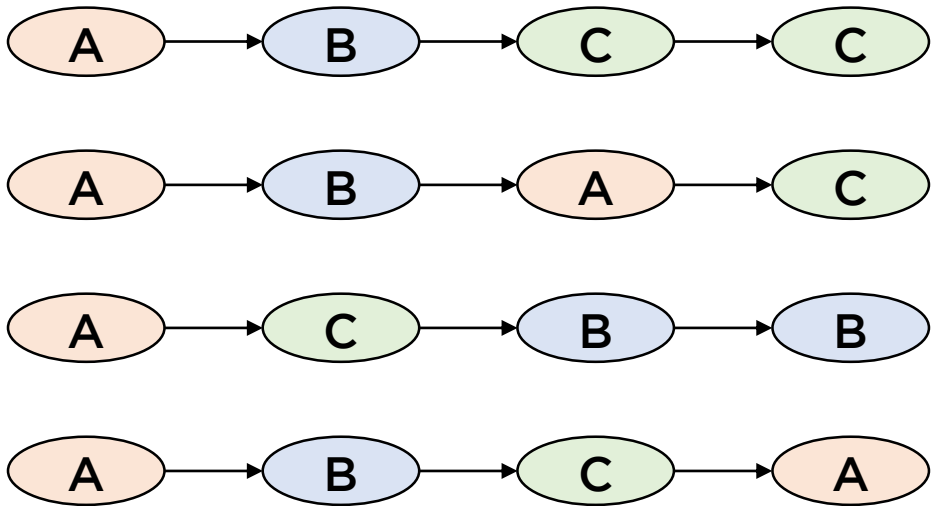
최대 빈발 아이템 집합

탐색 과정

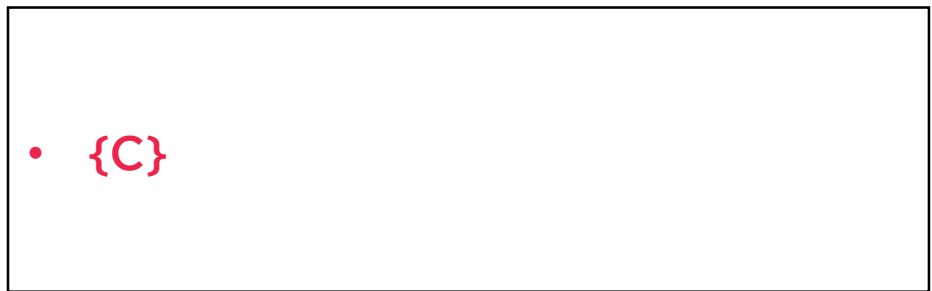


# I 순서를 고려한 연관규칙 탐색 (예시: L = 2, 최소 지지도 = 2)

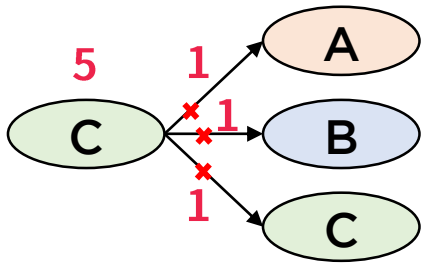
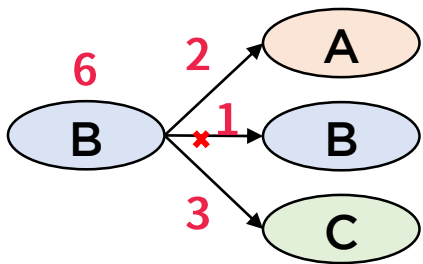
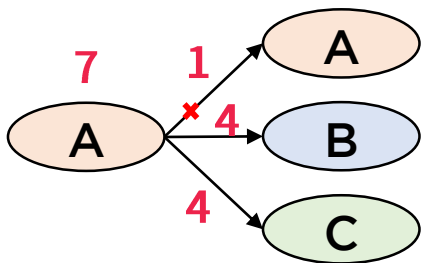
데이터



최대 빈발 아이템 집합

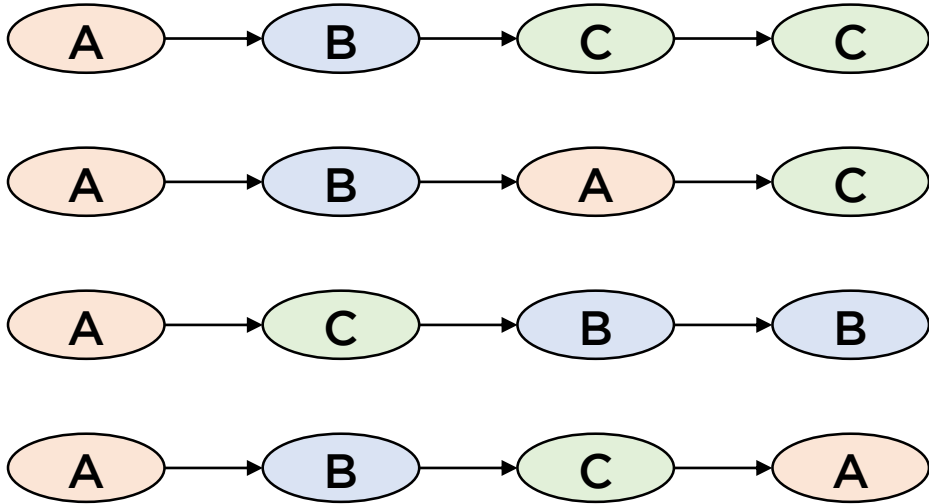


탐색 과정



# I 순서를 고려한 연관규칙 탐색 (예시: L = 2, 최소 지지도 = 3)

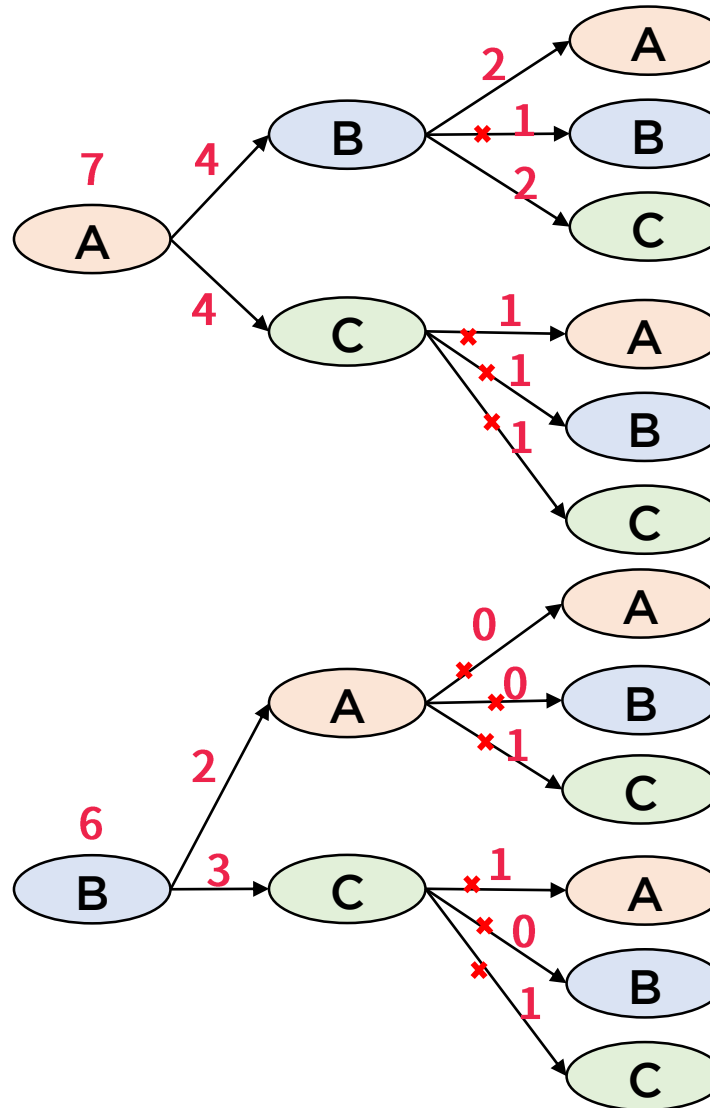
데이터



최대 빈발 아이템 집합

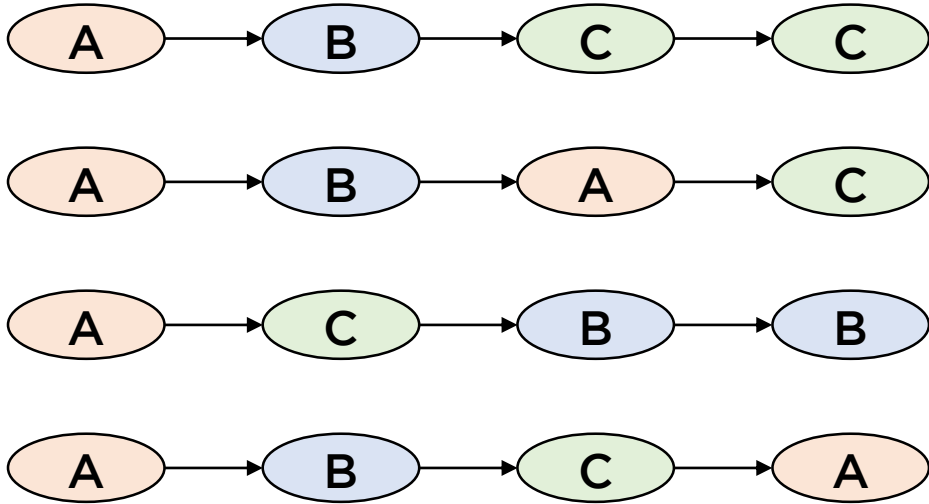
- {C}
- {B → A}
- {B → C}
- {A → C}

탐색 과정



# I 순서를 고려한 연관규칙 탐색 (예시: L = 2, 최소 지지도 = 3)

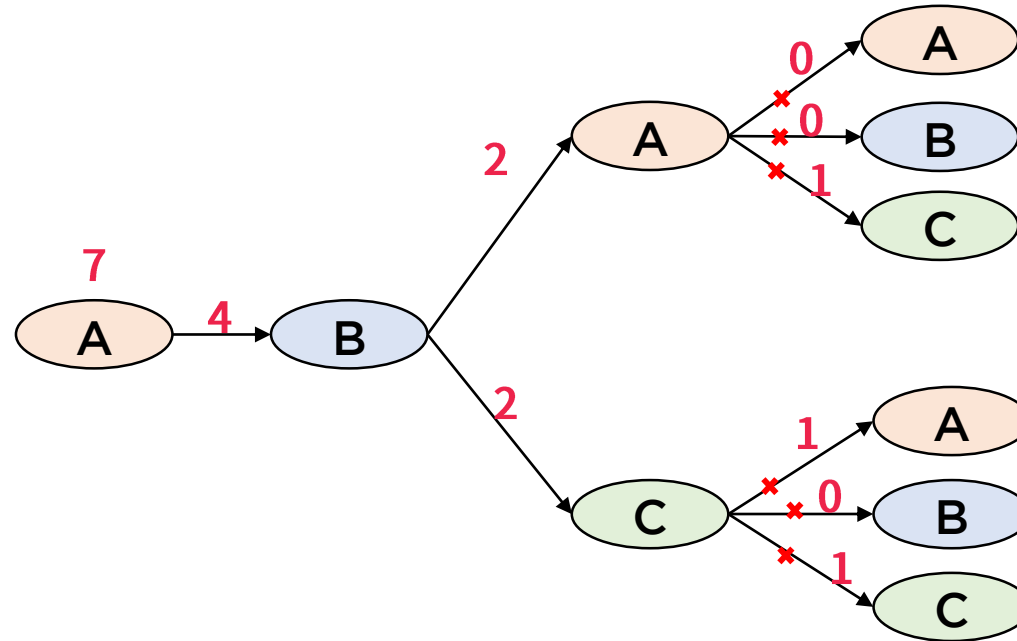
데이터



최대 빈발 아이템 집합

- {C}
- {B → A}
- {B → C}
- {A → C}
- {A → B → A}
- {A → B → C}

탐색 과정





I 순서를 고려한 연관규칙 탐색 (예시:  $L = 2$ , 최소 지지도 = 3)

| 최대 빈발 아이템 집합                    | 탐색 순서 1                           | 탐색 순서 2                           |
|---------------------------------|-----------------------------------|-----------------------------------|
| $C$                             | None                              | None                              |
| $B \rightarrow A$               | $B \rightarrow A$                 | None                              |
| $B \rightarrow C$               | $B \rightarrow C$                 | None                              |
| $A \rightarrow C$               | $A \rightarrow C$                 | None                              |
| $A \rightarrow B \rightarrow A$ | $(A \rightarrow B) \rightarrow A$ | $A \rightarrow (B \rightarrow A)$ |
| $A \rightarrow B \rightarrow C$ | $(A \rightarrow B) \rightarrow C$ | $A \rightarrow (B \rightarrow C)$ |

## Chapter. 12

어디서 많이 봤던 패턴이다 싶을 때: 빈발 패턴 탐색

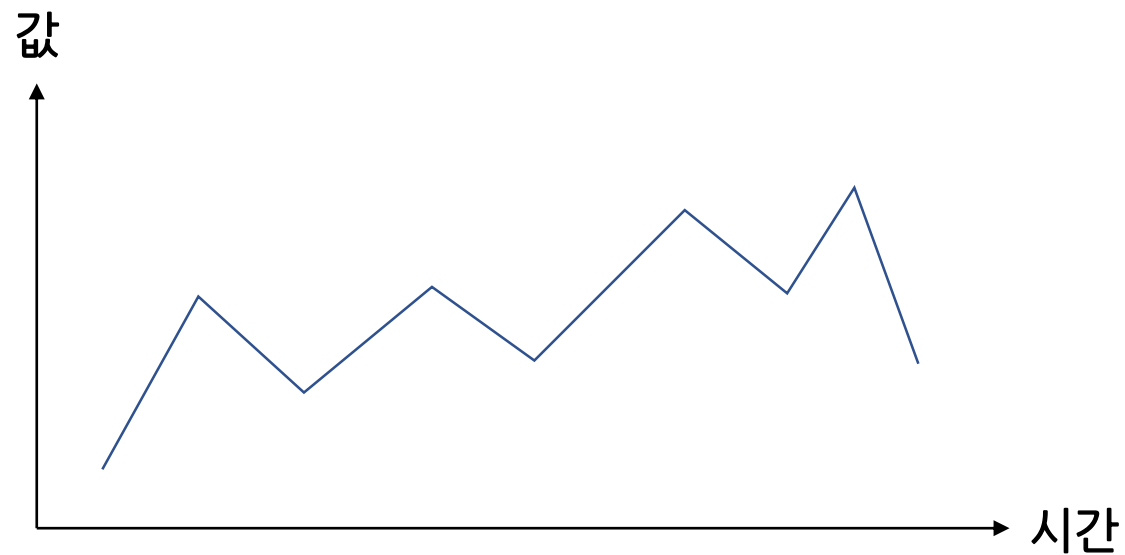
# | 빈발 시계열 패턴

FAST CAMPUS  
ONLINE  
데이터 탐색과 전처리 I

강사. 안길승

# I 시계열 데이터란?

- 시계열 데이터란 각 요소가 **(시간, 값)** 형태로 구성된 데이터로, 반드시 **순서 및 시간을 고려**해야 함



# I 시계열과 시퀀스 데이터의 차이

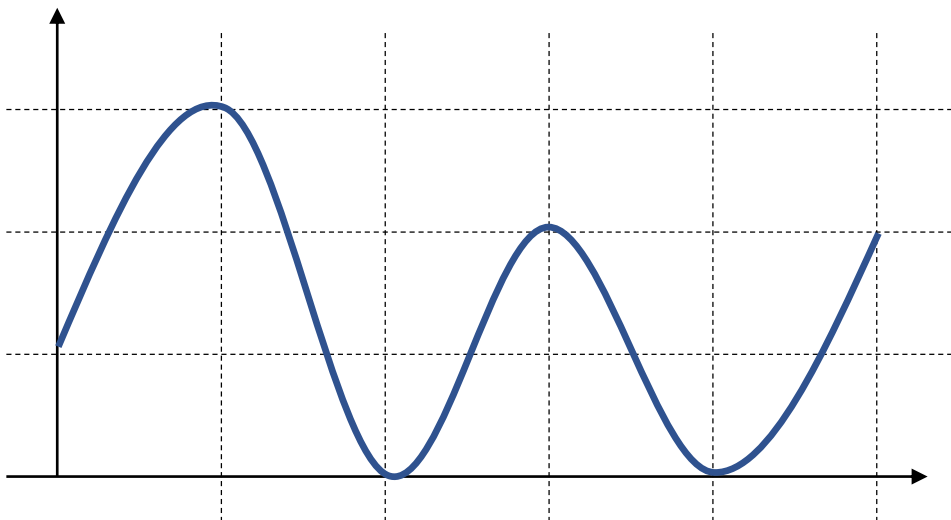
- 시계열 데이터와 시퀀스 데이터는 사용하는 인덱스와 값의 종류로 다음과 같이 구분할 수 있음

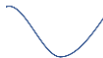
| 구분  | 시계열 데이터 | 시퀀스 데이터 |
|-----|---------|---------|
| 인덱스 | 시간 (순서) | 순서      |
| 값   | 주로 연속형  | 주로 범주형  |

- 다만, 엄밀히 말해서 시계열 데이터도 시퀀스 데이터에 속함

# I 시계열 패턴의 정의

- 시계열의 패턴은 크게 **모양**, **변화**에 의한 패턴과 **값**에 의한 패턴으로 구분할 수 있음

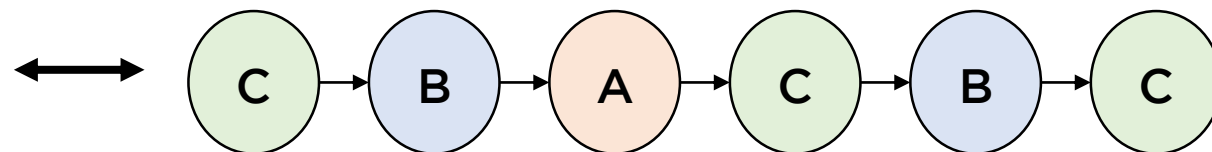
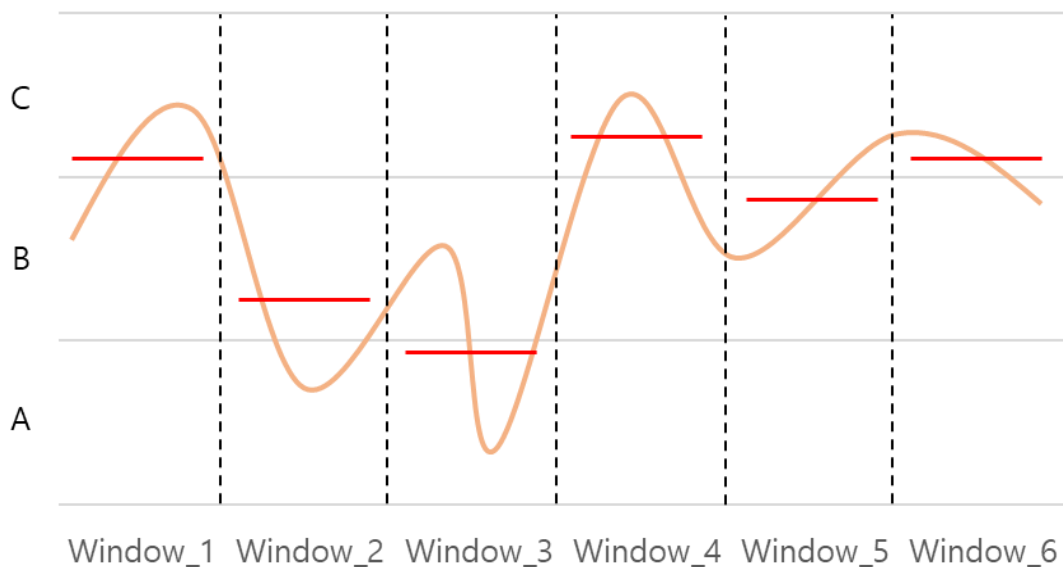


- 모양 패턴: 
- 변화 패턴: 증가 - 감소 - 증가 - 감소
- 값 패턴: 보통 - 큼 - 보통 - 매우 작음 - 보통 ...

- 계절성 혹은 주기성이 있는 경우를 제외하면, 모양 패턴을 찾는 것은 거의 불가능에 가까움

# I SAX: 시계열 → 시퀀스

- 시계열 데이터는 **연속형이라는 특징** 때문에 패턴을 찾으려면 **이산화**가 필요하며, SAX (symbolic aggregate approximation) 를 사용하면 시계열 데이터를 효과적으로 이산화할 수 있음
- SAX는 (1) 윈도우 분할, (2) 윈도우별 대표값 계산, (3) 알파벳 시퀀스로 변환이라는 세 단계로 구성됨



Chapter. 12

어디서 많이 봤던 패턴이다 싶을 때: 빈발 패턴 탐색

# | 머신러닝에서의 빈발 패턴 탐색

FAST CAMPUS  
ONLINE  
데이터 탐색과 전처리 I

강사. 안길승



## I 추천 시스템

- “상품 A를 구매하면 상품 B도 구매할 것이다”라는 유의한 연관 규칙이 있다면, 상품 A를 구매하고 **상품 B를 구매하지 않은 고객**에게 상품 B를 추천해주는 방법에 활용
- (예시) 아마존의 도서 추천



amazon.com<sup>®</sup> [Help](#) | [Close window](#)

### Recommended for You



**Inside Apple: How America's Most Admired--and Secretive--Company Really Works**  
Our Price: **\$9.99**  
Used & new from \$9.99  
[See all buying options](#)

Rate this item

☒ ☆☆☆☆☆

☐ I own it

☐ Not interested

### Because you purchased...



**The Toyota Way : 14 Management Principles from the World's Greatest Manufacturer**  
(Kindle Edition)

☒ ☆☆☆☆☆

☐ This was a gift

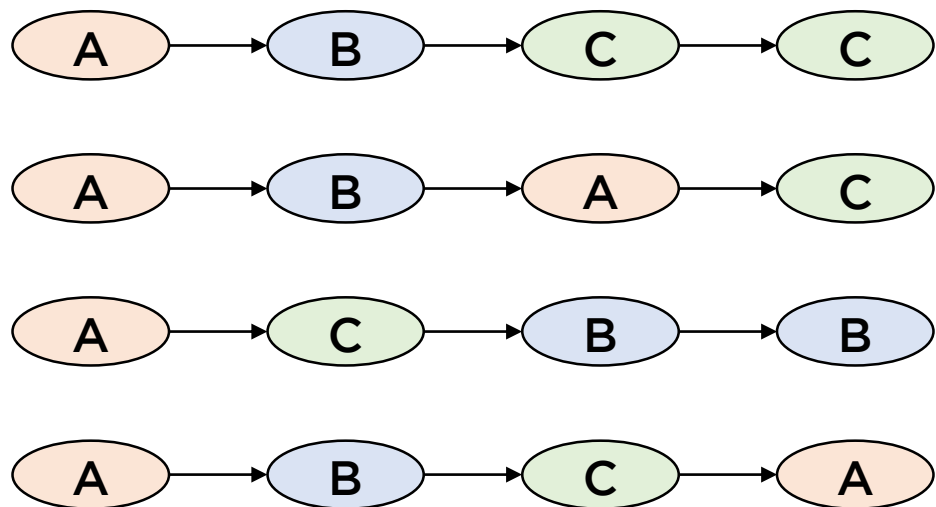
☐ Don't use for recommendations



# I 시계열 및 시퀀스 데이터에서의 특징 추출

- 시계열 및 시퀀스 분류 과제에서 특징을 추출하는데도 활용

데이터



| A → B | B → C | C → A |
|-------|-------|-------|
| 1     | 1     | 0     |
| 1     | 0     | 0     |
| 0     | 0     | 0     |
| 0     | 1     | 1     |

Chapter.

어디서 많이 봤던 패턴이다 싶을 때: 빈발 패턴 탐색

| 감사합니다

FAST CAMPUS  
ONLINE  
데이터 탐색과 전처리 I

강사. 안길승