데이터 전처리
Data Preprocessing

09
연관 분석(Association Analysis)

목차

1. 연관 규칙
2. 빈발 항목집합 생성
3. 연관 규칙 생성
4. 빈발 항목집합 표현
5. 연관 패턴 평가

1. 연관 규칙

연관 규칙 마이닝

- 트랜잭션 집합이 주어지면, 트랜잭션에서 다른 항목들의 발생을 기반으로 항목의 발생을 예측하는 규칙 찾기

시장-바구니 거래Market-Basket transactions

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

연관 규칙의 예

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

함축Implication은 공동 출현을 의미하며,
인과관계causality를 의미하지 않음

정의: 빈발 항목집합Frequent Itemset

- 항목집합Itemset
  - 하나 이상의 항목 모음
  - 예: {Milk, Bread, Diaper}
  - k-항목집합k-itemset: k개 항목들이 포함된 항목 집합
- 지지 횟수Support count(σ)
  - 항목집합의 출현 빈도
  - 예: σ({Milk, Bread, Diaper}) = 2
- 지지도Support
  - 항목 집합을 포함하는 트랜잭션의 비율
  - 예: s({Milk, Bread, Diaper}) = 2/5
- 빈발 항목집합Frequent Itemset
  - *minsup* 임계값보다 크거나 같은 항목집합

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

정의: 연관 규칙Association Rule

- 연관 규칙Association Rule
  - X → Y 형식의 함축적 표현식, 여기서 X와 Y는 항목집합
  - 예제: {Milk, Diaper} → {Beer}
- 규칙 평가 척도Rule Evaluation Metrics
  - 지지도Support (s): X와 Y 모두를 포함하는 트랜잭션의 비율
  - 신뢰도Confidence (c): X가 포함된 트랜잭션에서 Y에 있는 항목이 나타나는 빈도 측정

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

예제:

$$\{Milk, Diaper\} \Longrightarrow \{Beer\}$$

$$s = \frac{\sigma(Milk, Diaper, Beer)}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(Milk, Diaper, Beer)}{\sigma(Milk, Diaper)} = \frac{2}{3} = 0.67$$

연관 규칙 마이닝 작업

- 트랜잭션 집합 T가 주어질 때, 연관 규칙 마이닝의 목표는 다음 조건을 갖는 모든 규칙을 찾는 것
  - support ≥ *minsup* 임계값
  - confidence ≥ *minconf* 임계값

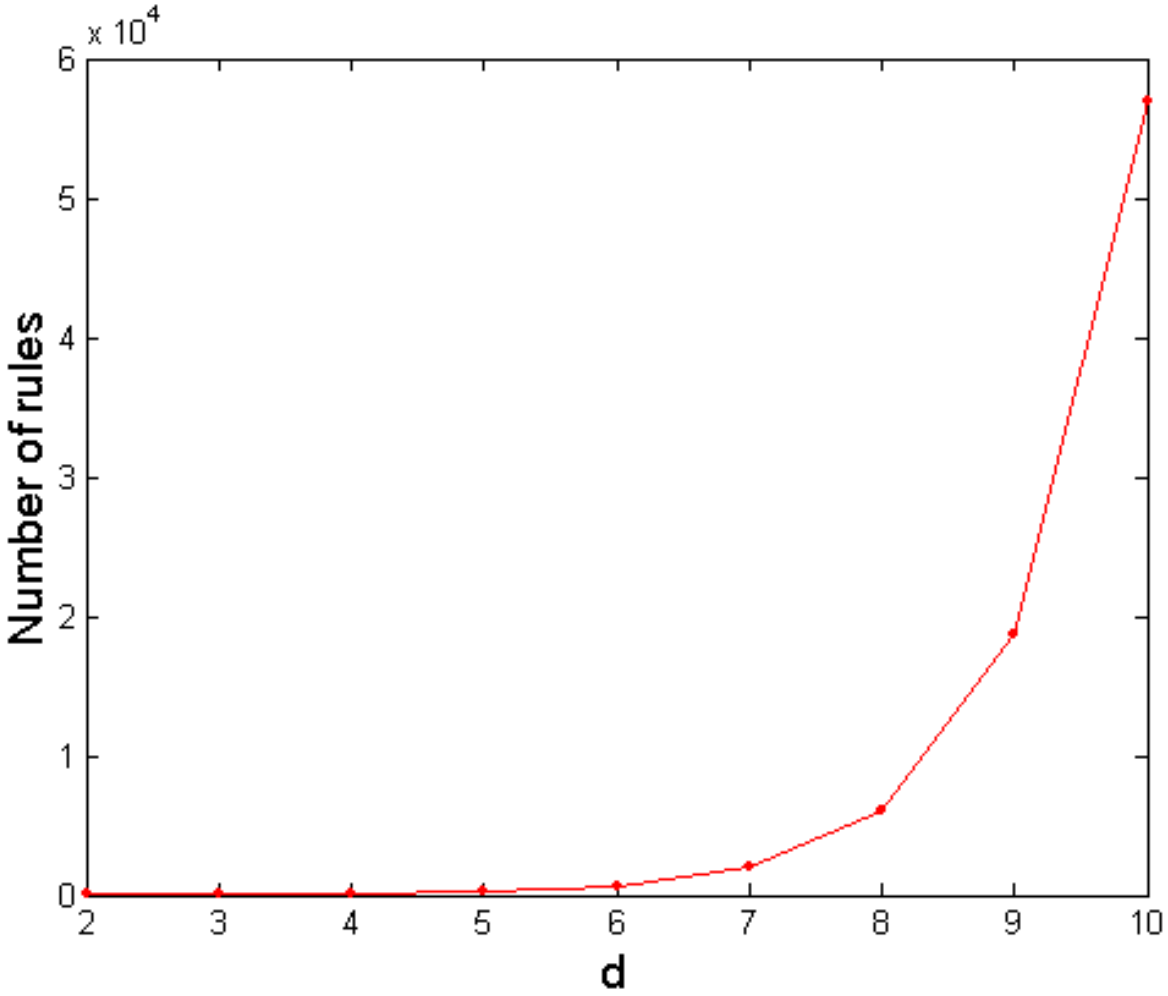
- 무차별 대입 접근Brute-force approach:
  - 가능한 모든 연관 규칙 나열
  - 각 규칙에 대한 지지도와 신뢰도 계산
  - *minsup*와 *minconf* 임계값에 포함 안된 규칙 삭제 ⇒ 계산 금지Computationally prohibitive

계산 복잡도

- 주어진 d개의 고유 항목:
  - 총 항목집합의 수 = 2^d
  - 가능한 연관 규칙의 전체 수:

$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$= 3^d - 2^{d+1} + 1$$

  - 만약 d = 6이면, R = 602 규칙



마이닝 연관 규칙

- 규칙 예제:
  - {Milk,Diaper} → {Beer} (s=0.4, c=0.67)
  - {Milk,Beer} → {Diaper} (s=0.4, c=1.0)
  - {Diaper,Beer} → {Milk} (s=0.4, c=0.67)
  - {Beer} → {Milk,Diaper} (s=0.4, c=0.67)
  - {Diaper} → {Milk,Beer} (s=0.4, c=0.5)
  - {Milk} → {Diaper,Beer} (s=0.4, c=0.5)

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

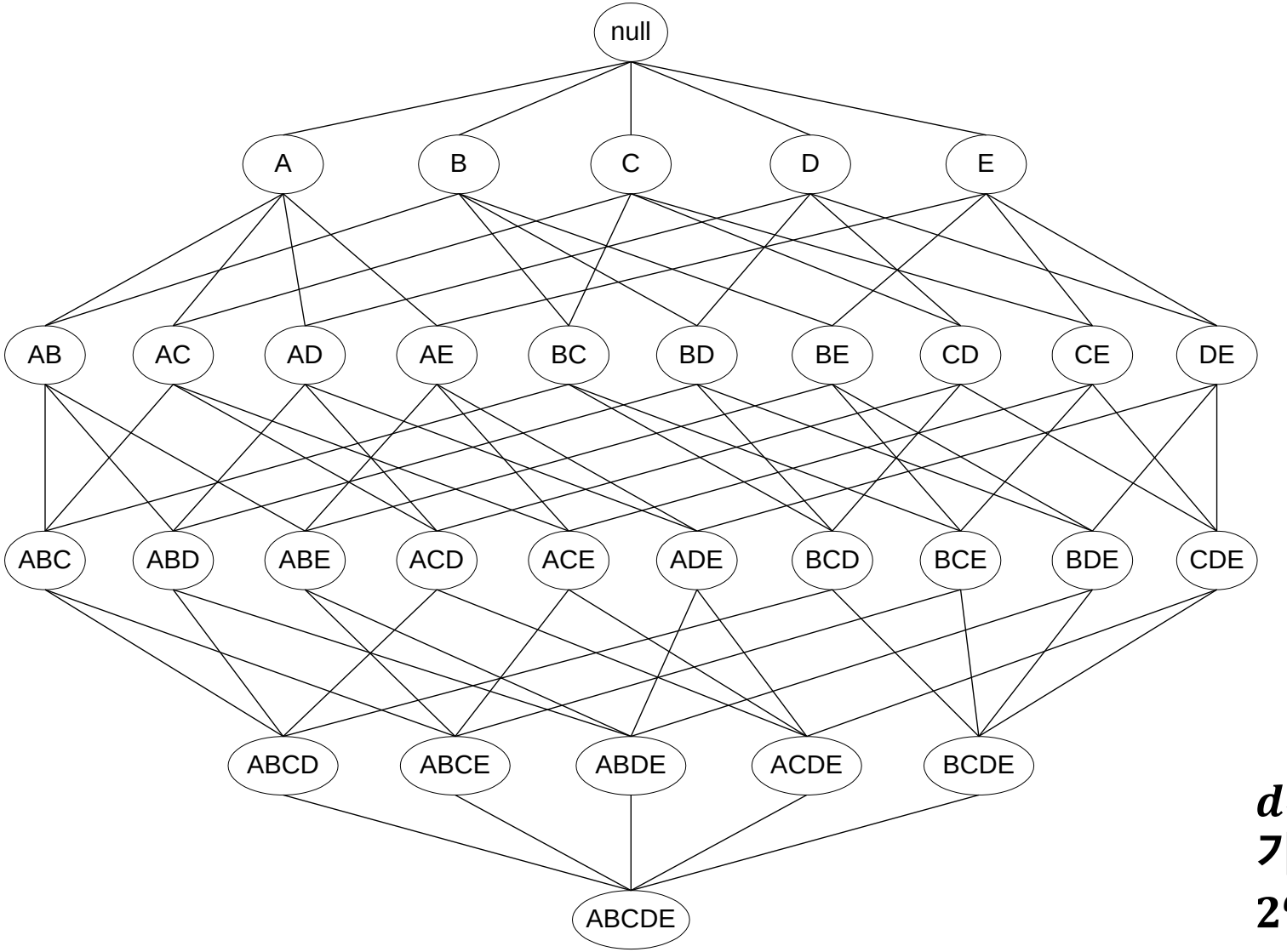
- 관찰Observations:
  - 위에 모든 규칙은 동일한 항목집합{Milk, Diaper, Beer}의 이진 분할binary partitioning
  - 같은 항목집합에서 비롯된 규칙은 동일한 지지도를 갖지만 다른 신뢰도를 가질 수 있음
  - 따라서, 지지도와 신뢰도 요구사항을 분리할 수 있음

마이닝 연관 규칙

- 2단계 접근:
  - 빈발 항목집합 생성: $support \geq minsup$인 모든 항목집합 생성
  - 규칙 생성: 각 빈발 항목집합에서 높은 신뢰 규칙 생성, 여기서 각 규칙은 빈발 항목집합의 이진 분할
- 빈발 항목집합 생성은 여전히 계산적으로 비용이 높음

2. 빈발 항목집합 생성

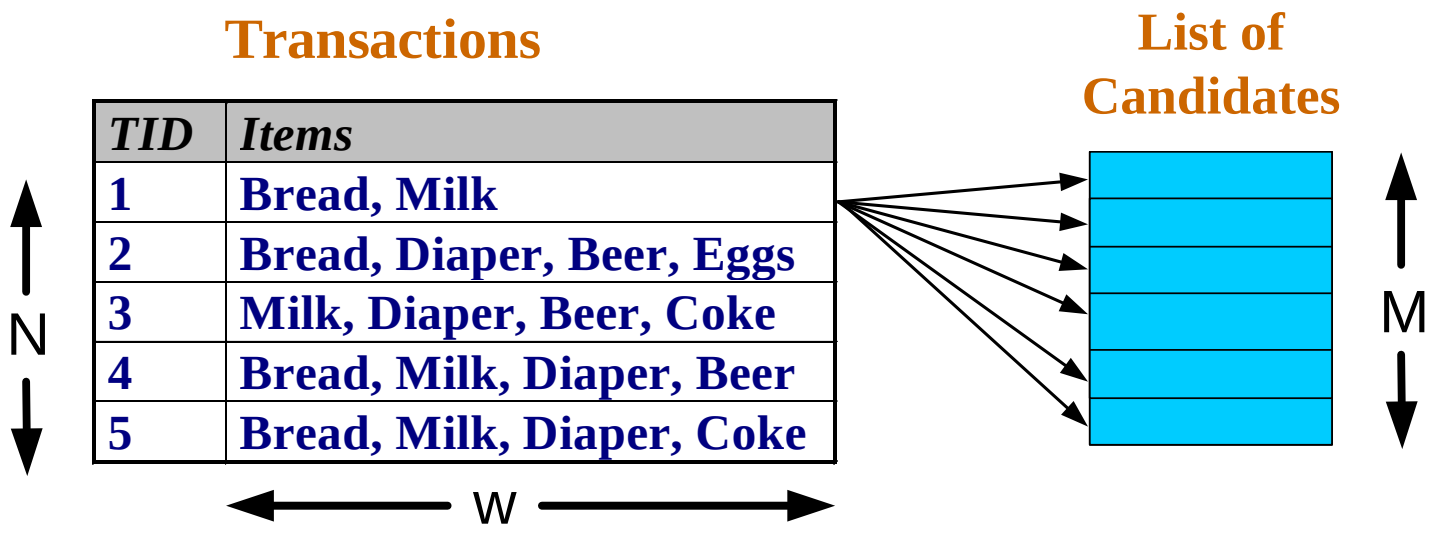
빈발 항목집합 생성



d개의 항목이 주어지면, 가능한 후보 항목집합이 2^d개

빈발 항목집합 생성

- 무차별 대입 접근Brute-force approach:
  - 격자의 각 항목집합은 빈발 항목집합 후보
  - 데이터베이스를 스캔하여 각 후보의 지지 횟수 계산
  - 모든 후보와 각 트랜잭션을 매치
  - 복잡도 ~ O(NMw) => $M = 2^d$ 이므로 비쌈



빈발 항목집합 생성 전략

- 후보 수 감소 (M)
  - 완전한 검색: $M=2^d$
  - 가지치기pruning 기술을 사용하여 M을 감소

- 트랜잭션 수 감소 (N)
  - 항목집합의 크기가 커짐에 따라 N의 크기 감소
  - DHP 및 수직 기반 마이닝 알고리즘이 사용됨

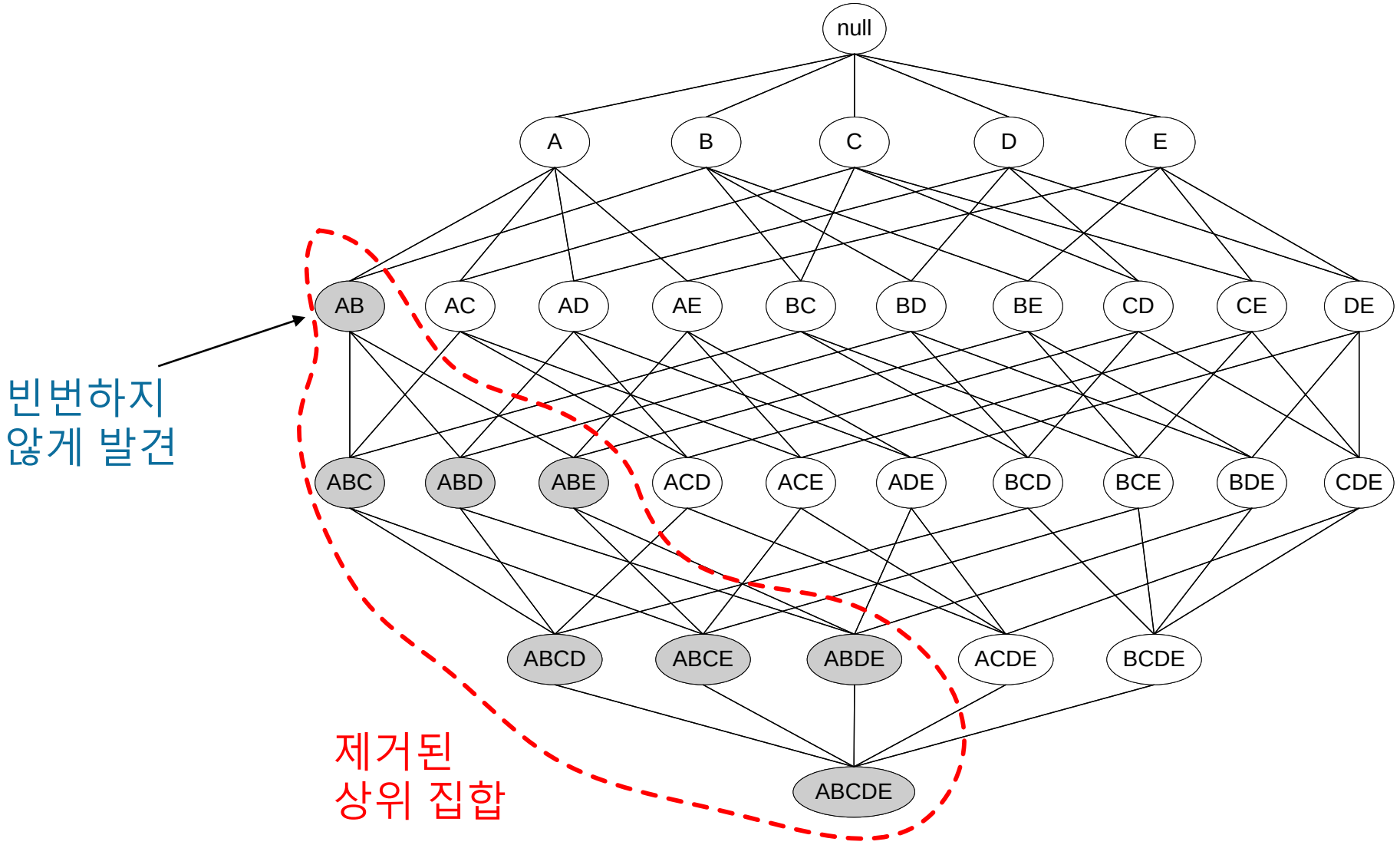
- 비교 횟수 감소 (NM)
  - 효율적인 데이터 구조를 사용하여 후보 또는 트랜잭션 저장
  - 모든 트랜잭션과 모든 후보들을 매치할 필요는 없음

후보 수 감소

- *Apriori* 원리: 항목집합이 빈발하게 있는 경우 모든 부분집합도 빈발해야 함
- *Apriori* 원리는 다음과 같은 지지도 값 특성에 따라 유지
  - 항목집합의 지지도는 부분집합의 지지도를 결코 초과하지 않음
  - 지지도의 안티-모노톤anti-monotone 특성으로 알려져 있음

$$\forall X, Y : (X \subset Y) \Rightarrow s(X) \geq s(Y)$$

Apriori 원리 설명



Apriori 원리 설명

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

Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

최소 지지도Minimum Support = 3

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 4 = 16$$

Apriori 원리 설명

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

Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

최소 지지도Minimum Support = 3

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 4 = 16$$

Apriori 원리 설명

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset
{Bread,Milk}
{Bread,Beer}
{Bread,Diaper}
{Milk,Beer}
{Milk,Diaper}
{Beer,Diaper}

Pairs (2-itemsets)

(Coke 또는 Eggs가 포함된 후보들을 생성할 필요 없음)

최소 지지도Minimum Support = 3

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 4 = 16$$

Apriori 원리 설명

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(Coke 또는 Eggs가 포함된 후보들을 생성할 필요 없음)

최소 지지도Minimum Support = 3

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 4 = 16$$

Apriori 원리 설명

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(Coke 또는 Eggs가 포함된 후보들을 생성할 필요 없음)

최소 지지도Minimum Support = 3

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 4 = 16$$

Triplets (3-itemsets)

Itemset
{Beer, Diaper, Milk}
{Beer,Bread, Diaper}
{Bread, Diaper, Milk}
{Beer, Bread, Milk}

Apriori 원리 설명

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(Coke 또는 Eggs가 포함된 후보들을 생성할 필요 없음)

최소 지지도Minimum Support = 3

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 4 = 16$$

$$6 + 6 + 1 + 13$$

Triplets (3-itemsets)

Itemset	Count
{Beer, Diaper, Milk}	2
{Beer,Bread, Diaper}	2
{Bread, Diaper, Milk}	2
{Beer, Bread, Milk}	1

Apriori 알고리즘

- F_k: 빈발 k-항목집합
- L_k: 후보 k-항목집합

1. k=1
2. F_1 = {빈발 1-항목집합} 생성
3. F_k가 없을 때까지 반복
   - **후보 생성**Candidate Generation: F_k 에서 L_{k+1} 생성
   - **후보 가지치기**Candidate Pruning: 빈발하지 않는 길이가 k인 부분집합을 포함한 L_{k+1} 에 있는 후보 항목집합을 가지치기
   - **지지도 계산**Support Counting: DB를 스캔하여 L_{k+1} 의 각 후보 지지도 계산
   - **후보 제거**Candidate Elimination: L_{k+1} 에서 빈발하지 않는 후보 제거, frequent => F_{k+1} 인 것만 남김

후보 생성Candidate Generation: 무차별 대입 방법Brute-force method

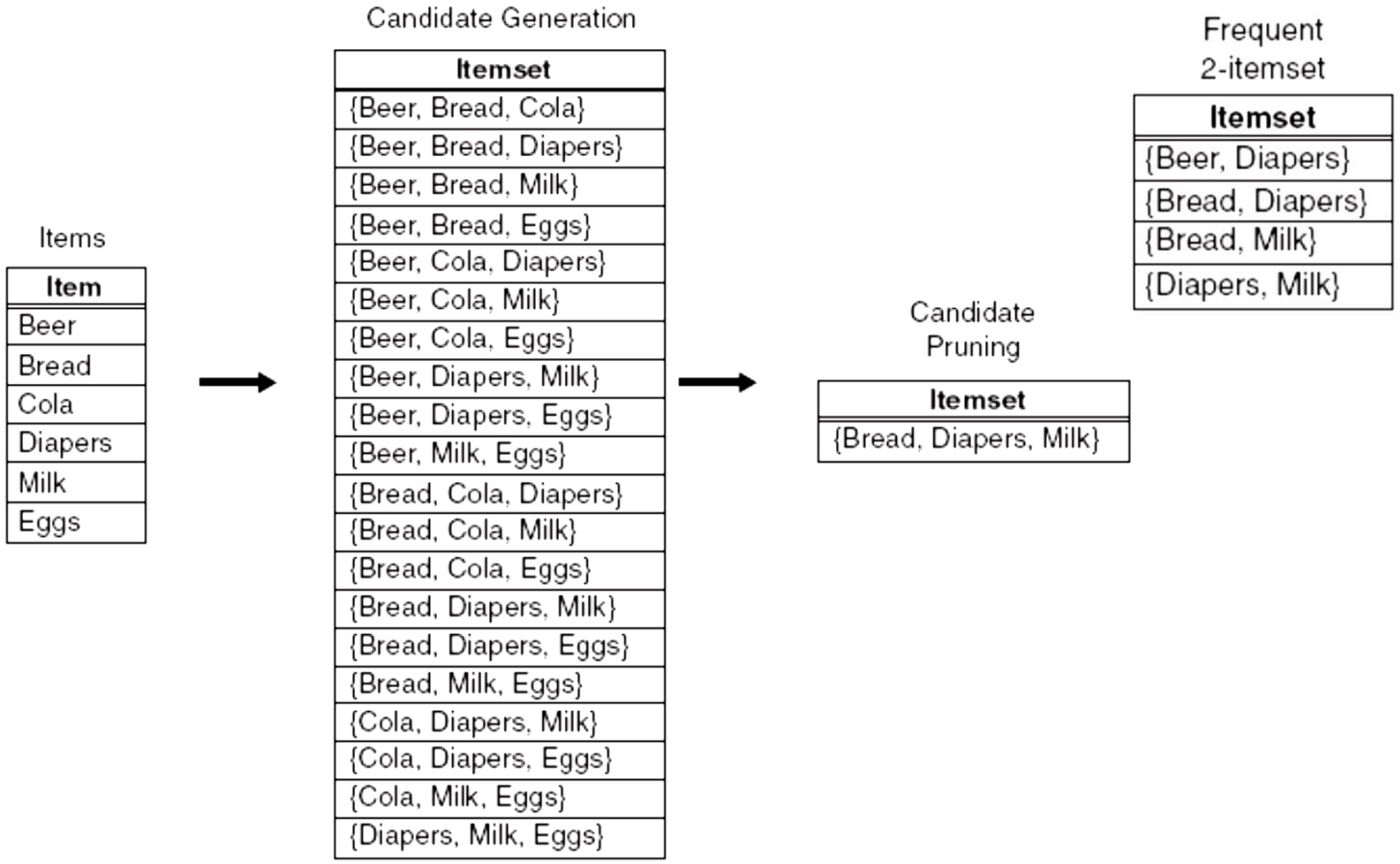


Figure 6.6. A brute-force method for generating candidate 3-itemsets.

후보 생성Candidate Generation: F_{k-1} and F_1 항목집합 합병

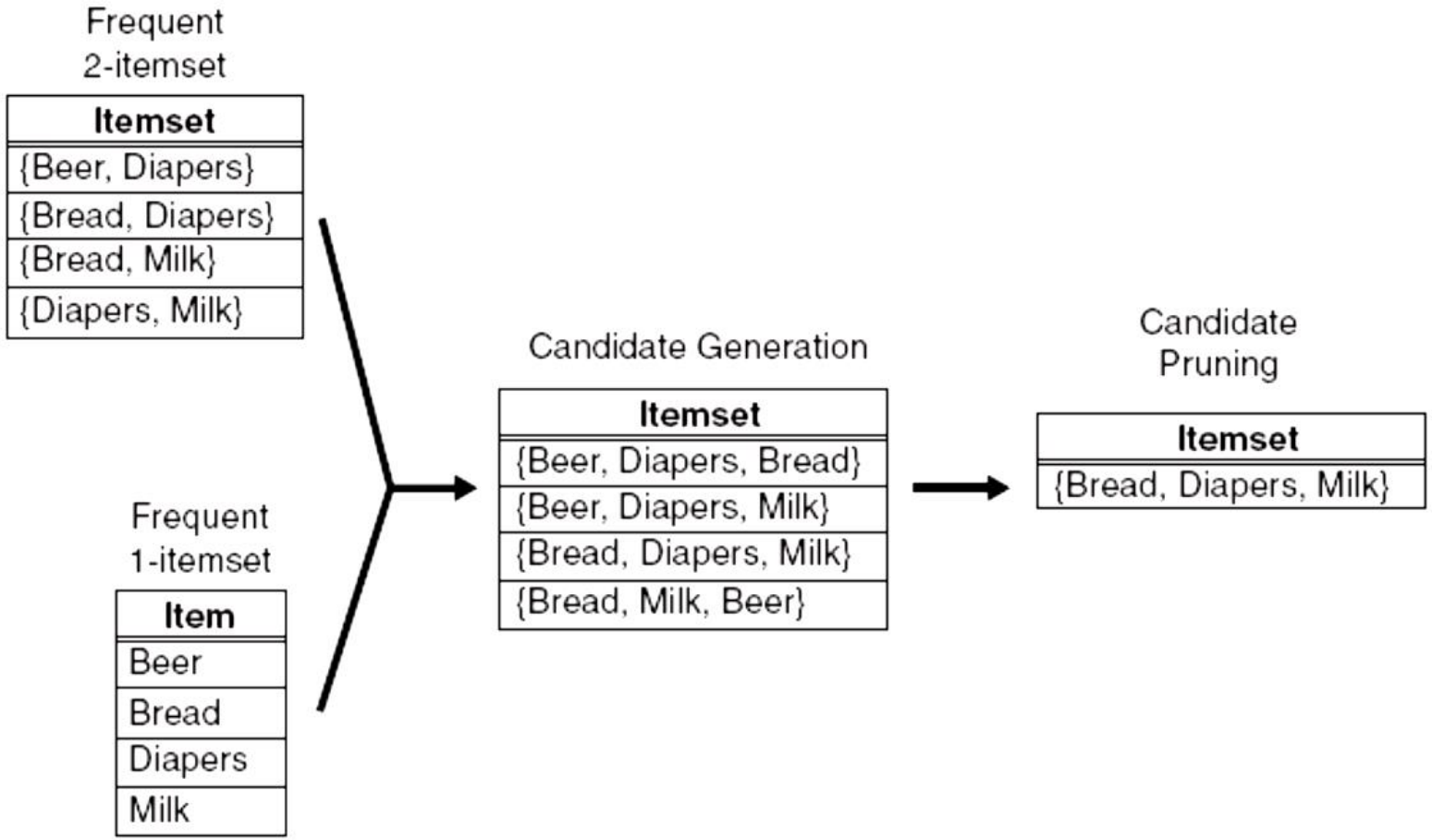


Figure 6.7. Generating and pruning candidate k-itemsets by merging a frequent $(k-1)$-itemset with a frequent item. Note that some of the candidates are unnecessary because their subsets are infrequent.

후보 생성Candidate Generation: $F_{k-1} \times F_{k-1}$ 방법

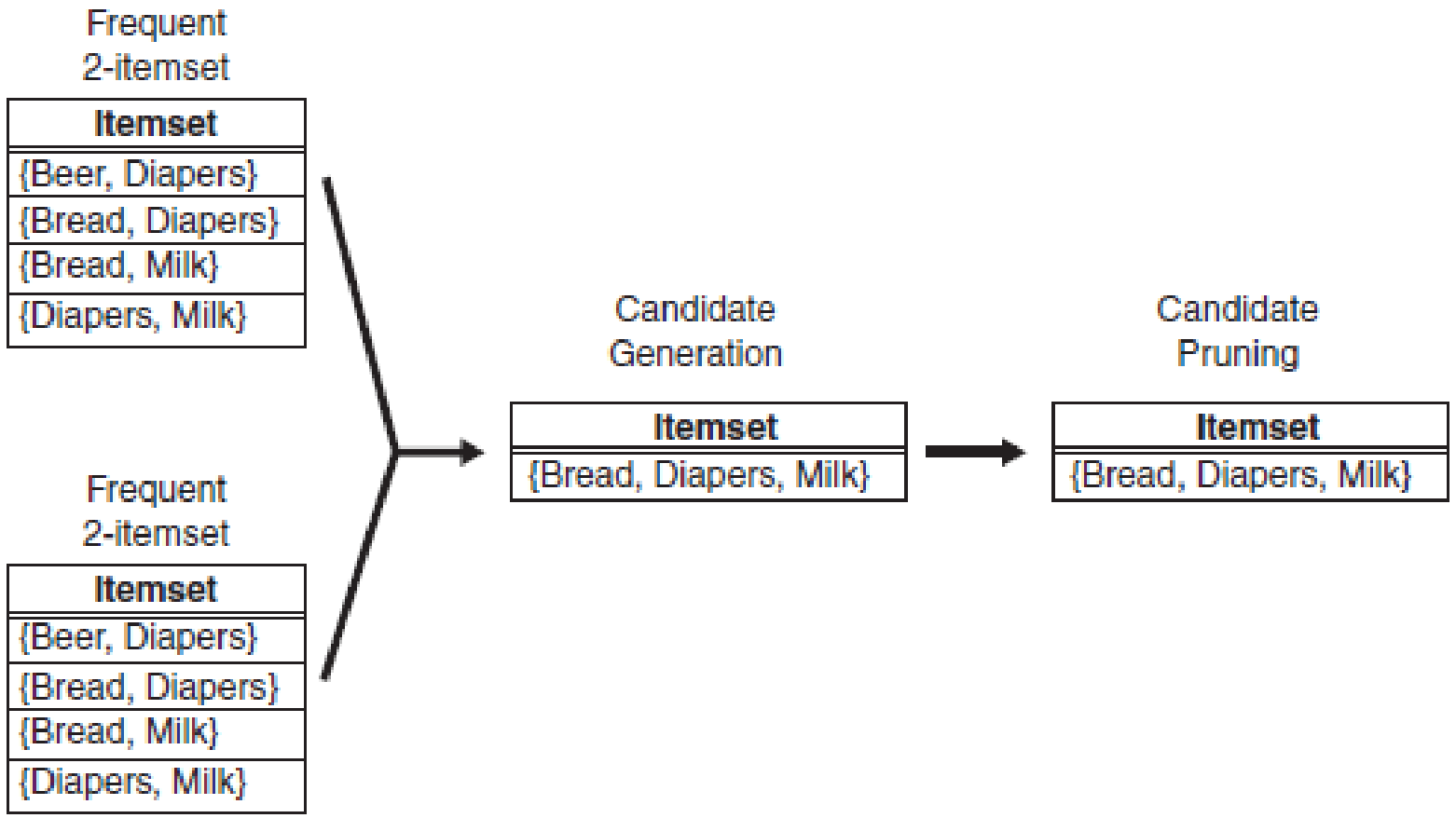


Figure 6.8. Generating and pruning candidate k-itemsets by merging pairs of frequent $(k-1)$-itemsets.

후보 생성Candidate Generation: $F_{k-1} \times F_{k-1}$ 방법

- 첫번째 (k-2) 항목이 동일한 경우 두 개의 빈발 (k-1) 항목집합을 병합

- F_3 = {ABC,ABD,ABE,ACD,BCD,BDE,CDE}
  - Merge(**AB**C, **AB**D) = **AB**CD
  - Merge(**AB**C, **AB**E) = **AB**CE
  - Merge(**AB**D, **AB**E) = **AB**DE

  - 길이 2 대신 길이 1의 접두사prefix만 공유하므로 (**A**BD, **A**CD)는 병합하지 않음

후보 제거Candidate Pruning

- F_3 = {ABC,ABD,ABE,ACD,BCD,BDE,CDE} 빈발 3-항목집합의 집합

- L_4 = {ABCD,ABCE,ABDE} 이전에 생성된 4-항목집합 후보의 집합

- 후보 가지치기
  - ACE와 BCE가 빈발하지 않으므로 ABCE 가지치기
  - ADE가 빈발하지 않으므로 ABDE 가지치기
- 후보 가지치기 후: L_4 = {ABCD}

대체 $F_{k-1} \times F_{k-1}$ 방법

- 첫번째 항목의 마지막 (k-2) 항목이 두번째 항목의 첫번째 (k-2) 항목과 동일한 경우 두 개의 빈발 (k-1)-항목집합을 병합

- F_3 = {ABC,ABD,ABE,ACD,BCD,BDE,CDE}
  - Merge(A**BC**, **BC**D) = A**BC**D
  - Merge(A**BD**, **BD**E) = A**BD**E
  - Merge(A**CD**, **CD**E) = A**CD**E
  - Merge(B**CD**, **CD**E) = B**CD**E

대체 $F_{k-1} \times F_{k-1}$ 방법을 위한 후보 제거

- F_3 = {ABC,ABD,ABE,ACD,BCD,BDE,CDE}는 빈발 3-항목집합의 집합

- L_4 = {ABCD,ABDE,ACDE,BCDE} 이전에 생성된 4-항목집합 후보의 집합

- 후보 가지치기
  - ADE가 빈발하지 않으므로 ABDE 가지치기
  - ACE와 ADE가 빈발하지 않으므로 ACDE 가지치기
  - BCE 때문에 BCDE 가지치기
- 후보 가지치기 후 : L_4 = {ABCD}

Apriori 원리 설명

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(Coke 또는 Eggs가 포함된 후보들을 생성할 필요 없음)

최소 지지도Minimum Support = 3

Triplets (3-itemsets)

Itemset	Count
{Bread, Diaper, Milk}	2

후보 생성을 위해 $F_{k-1}xF_{k-1}$ 방법을 사용하면 하나의 3-항목집합만 생성.
이는 지지도 계산 단계 후에 제거

만약 모든 하위집합이 고려되면,

$${}_6C_1 + {}_6C_2 + {}_6C_3$$

$$6 + 15 + 20 = 41$$

지지도 기반 가지치기support-based pruning,

$$6 + 6 + 1 + 13$$

후보 항목집합의 지지도 계산

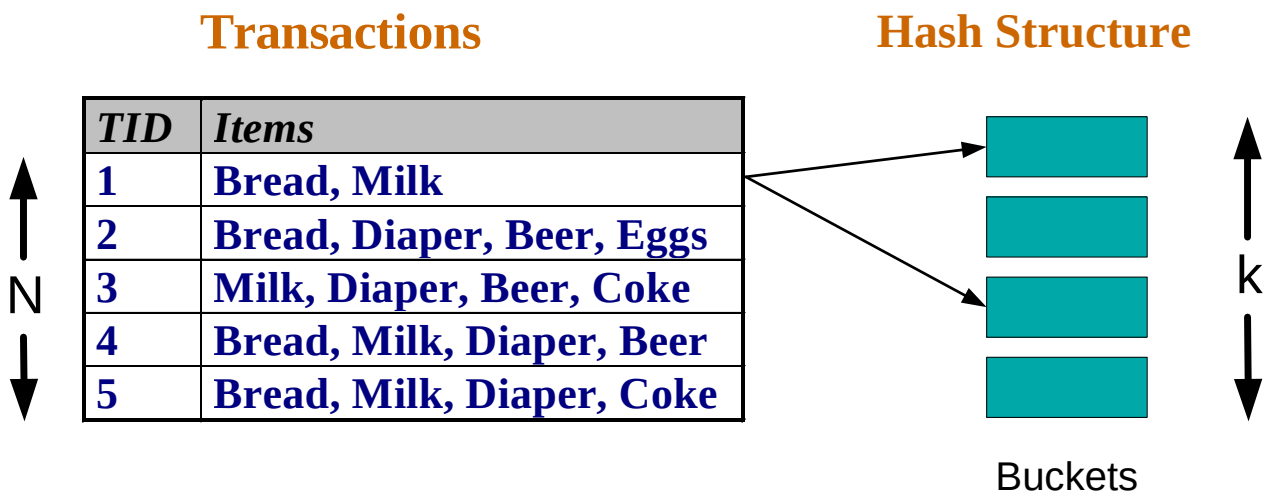
- 트랜잭션 데이터베이스 스캔으로 각 후보 항목집합의 지지도 결정
  - 모든 트랜잭션에 대해 모든 후보 항목집합을 일치시켜야 함(매우 비싼 연산)

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

Itemset
{ Beer, Diaper, Milk}
{ Beer,Bread,Diaper}
{Bread, Diaper, Milk}
{ Beer, Bread, Milk}

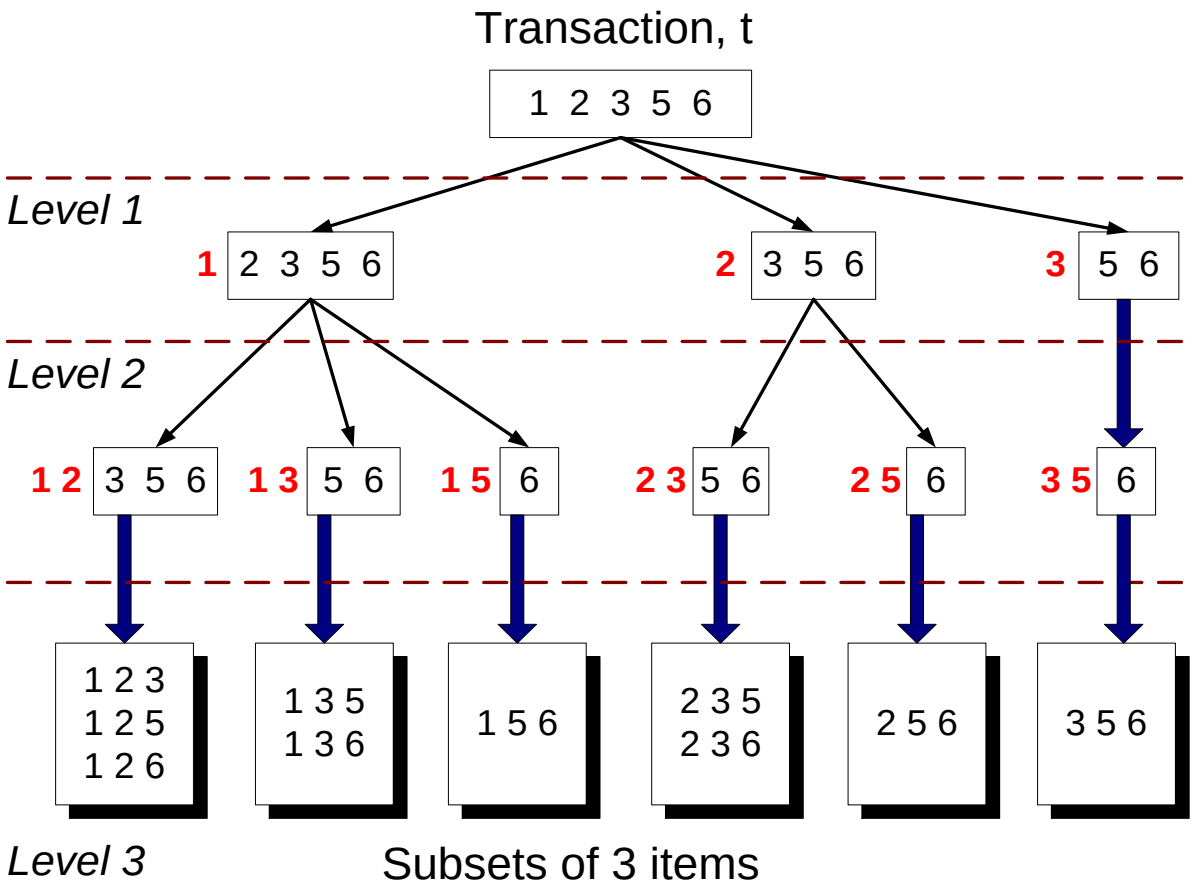
후보 항목집합의 지지도 계산

- 비교 횟수를 줄이기 위해, 후보 항목집합을 해시 구조에 저장
  - 모든 후보에 대해 각 트랜잭션을 대조하는 대신에 해시 버킷에 포함된 후보와 대조



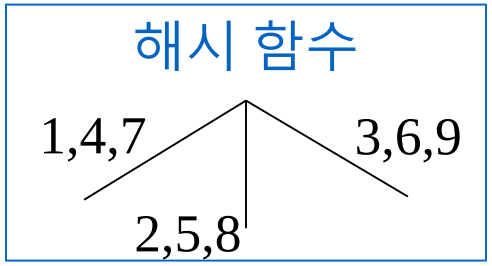
지지도 계산: 예제

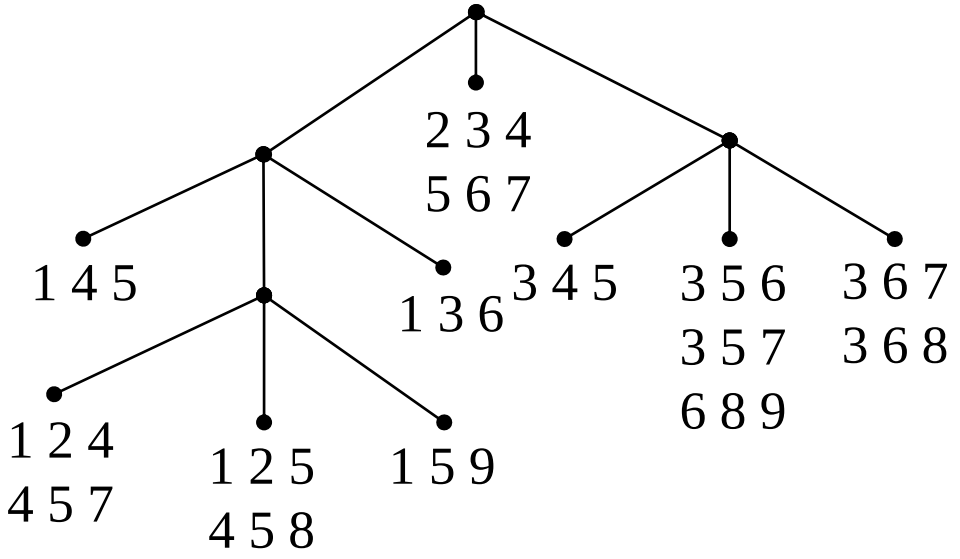
- 길이 3의 후보 항목집합 15개가 있다고 가정: {1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}
- 트랜잭션 (1,2,3,5,6)에서 얼마나 많은 항목 집합을 지원할까?



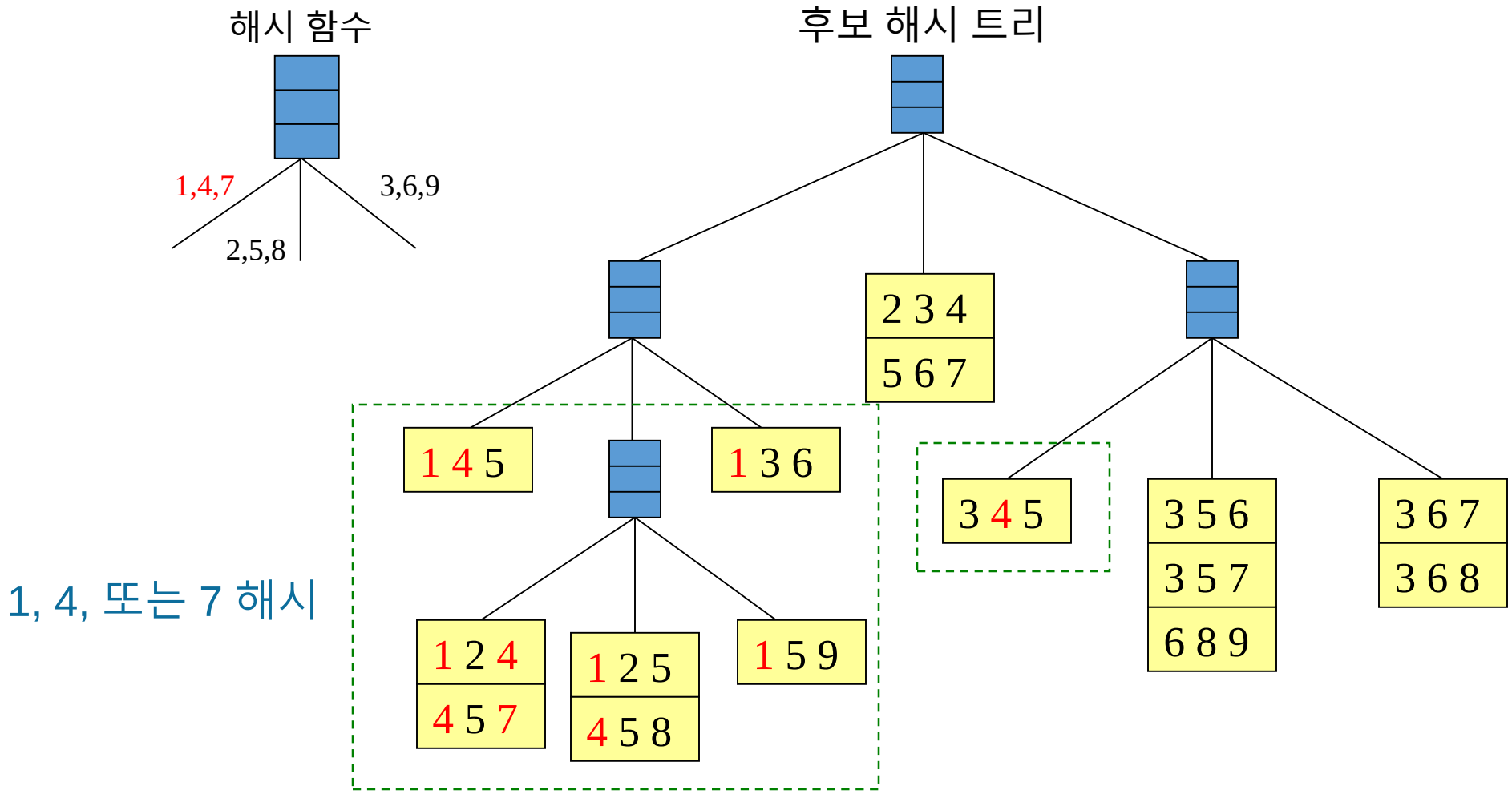
해시 트리를 사용한 지지도 계산

- 길이 3의 후보 항목집합 15개가 있다고 가정:
  {1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}
- 필요 요소
  - 해시 함수
  - 최대 리프 크기: 리프 노드에 저장된 항목집합의 최대 개수 (후보 항목집합 수가 최대 리프 크기를 초과하는 경우 노드 분할)

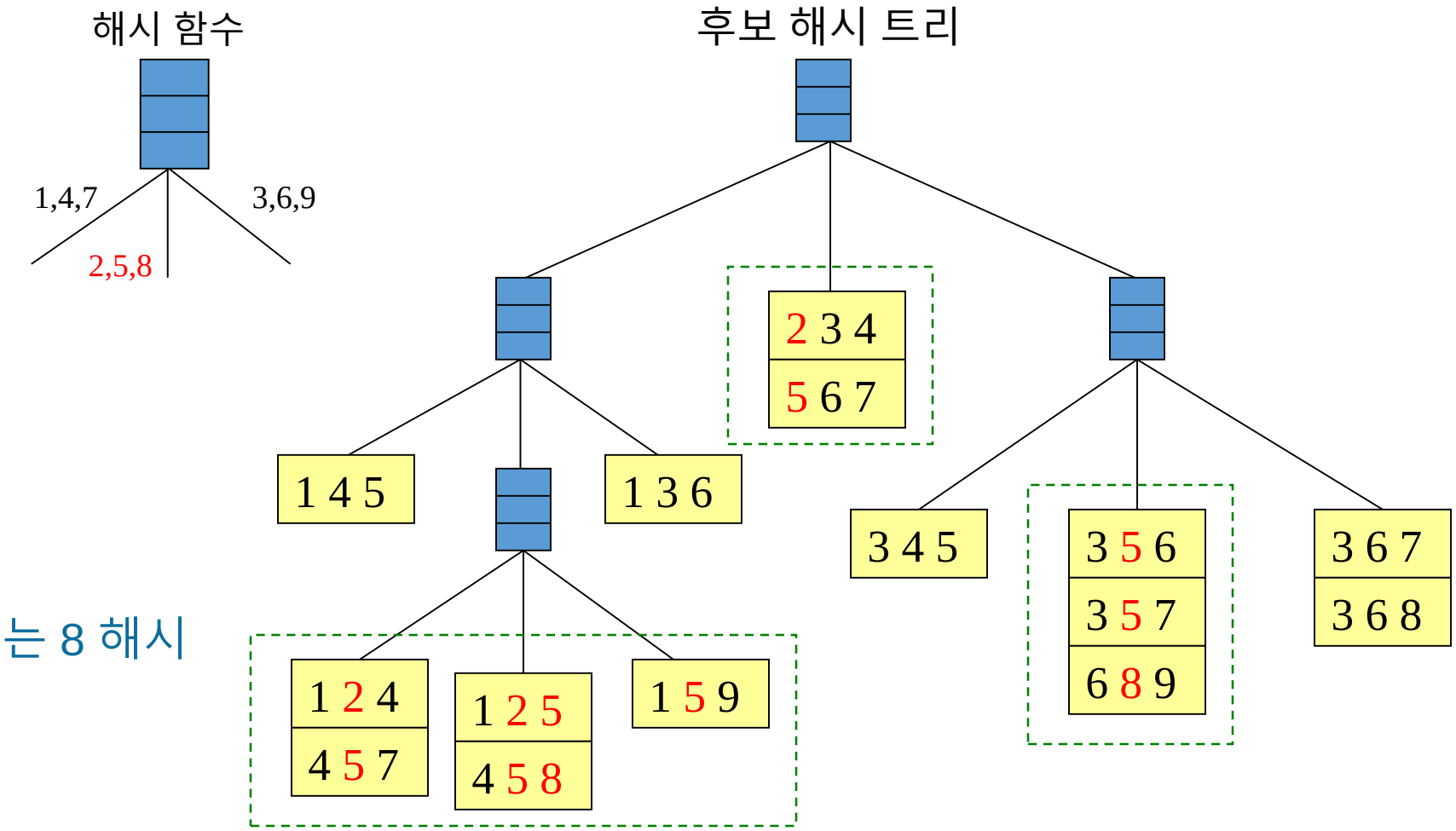




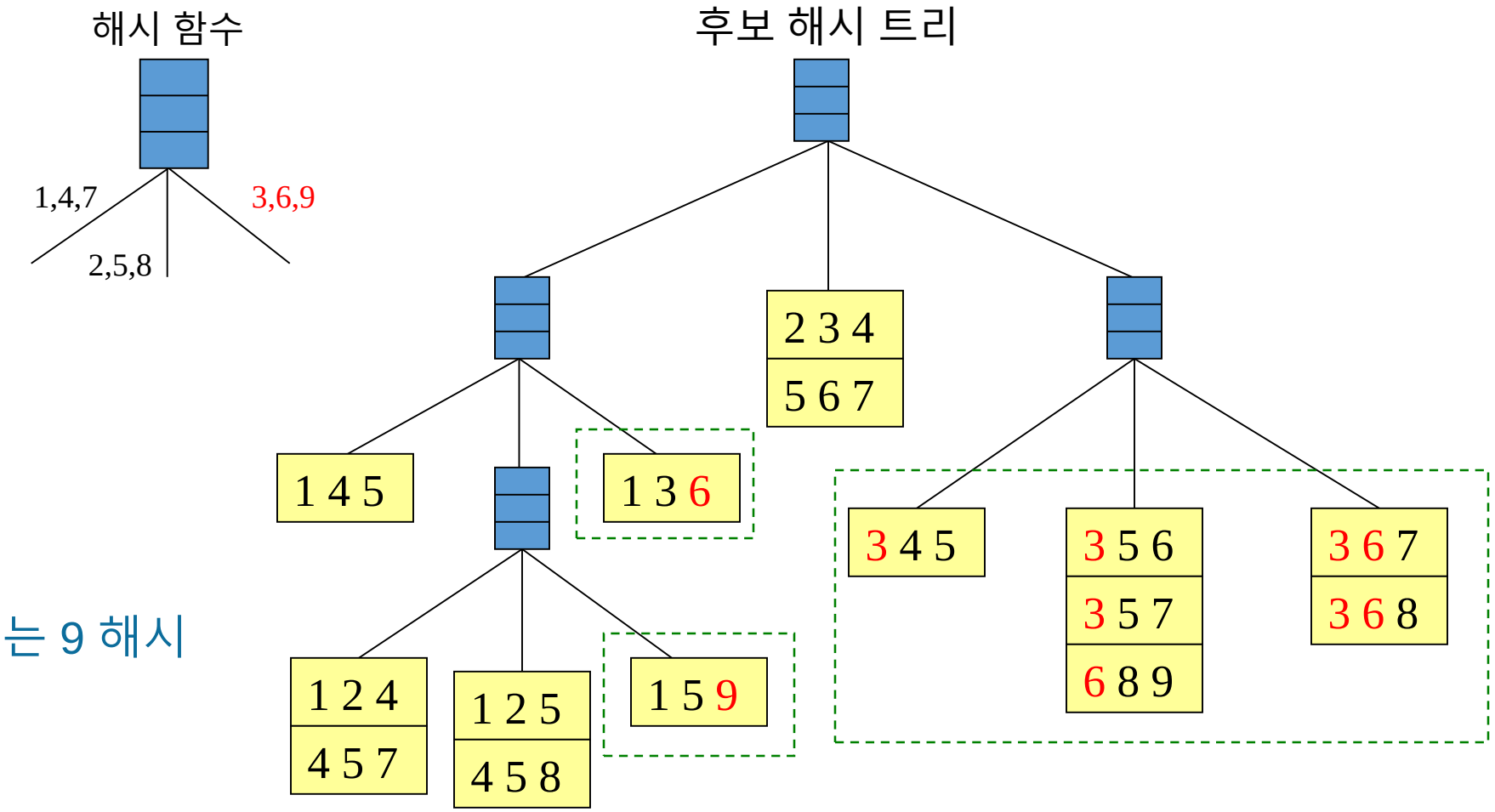
해시 트리를 사용한 지지도 계산



해시 트리를 사용한 지지도 계산

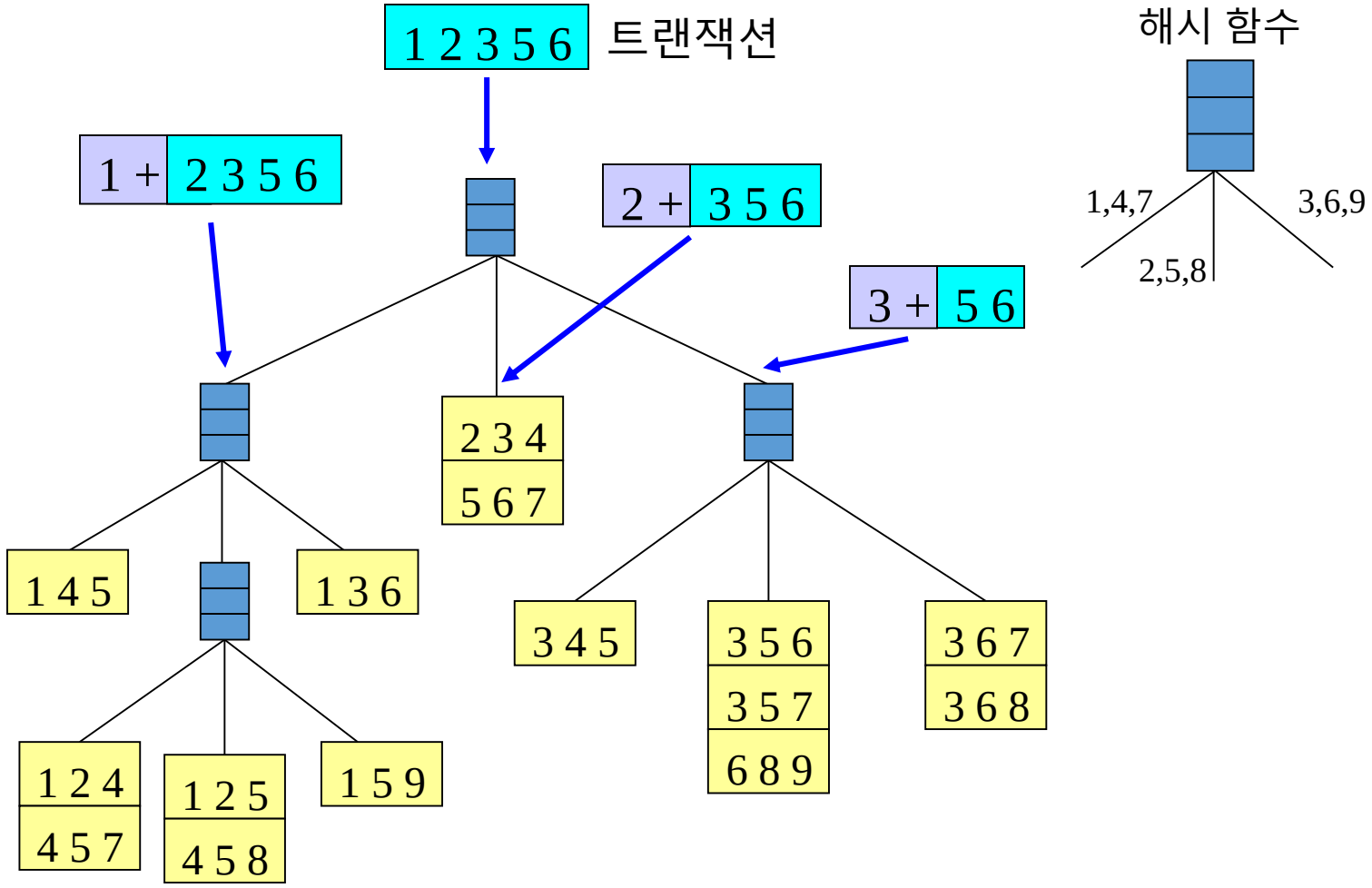


해시 트리를 사용한 지지도 계산

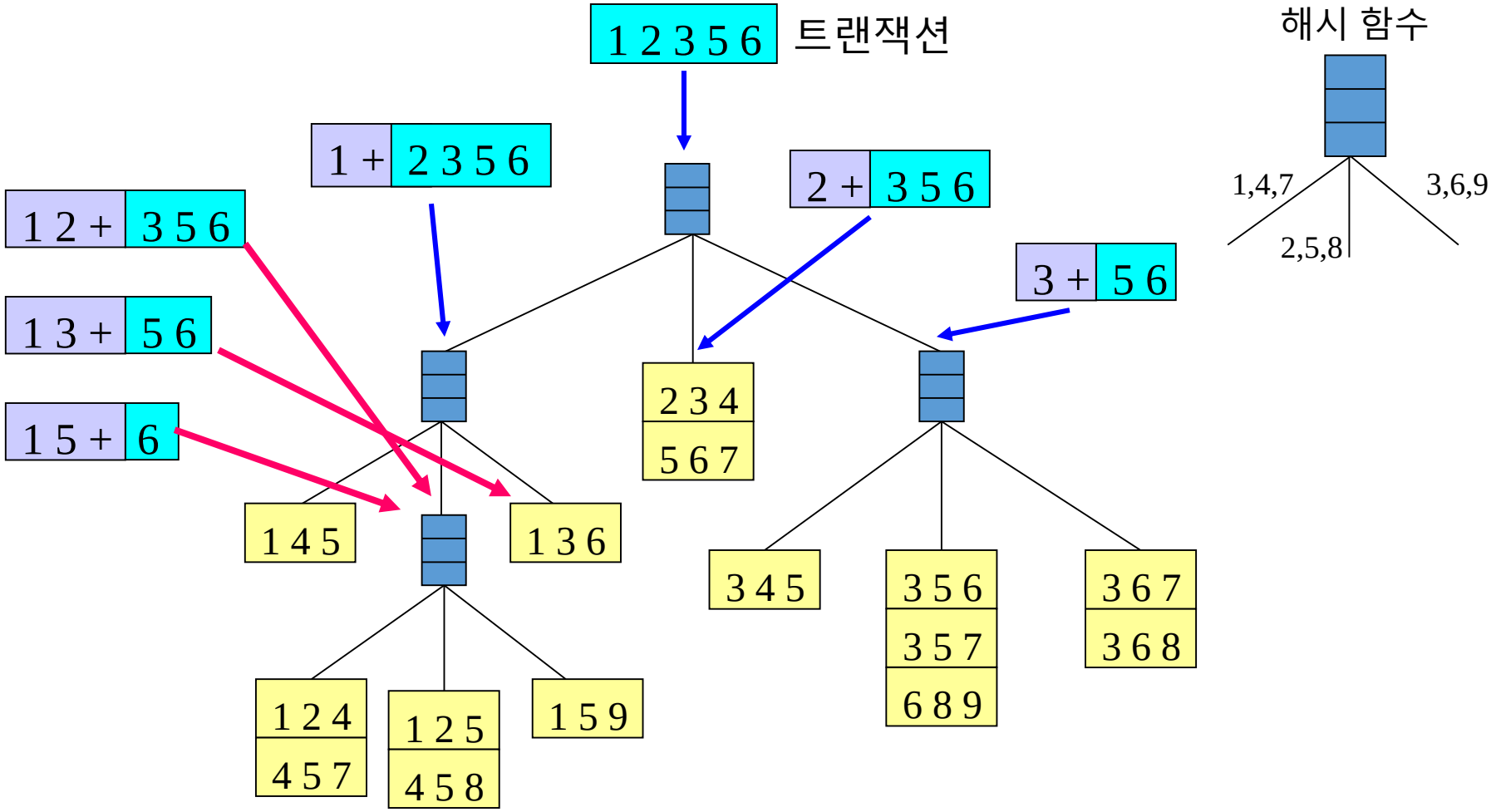


3, 6, 또는 9 해시

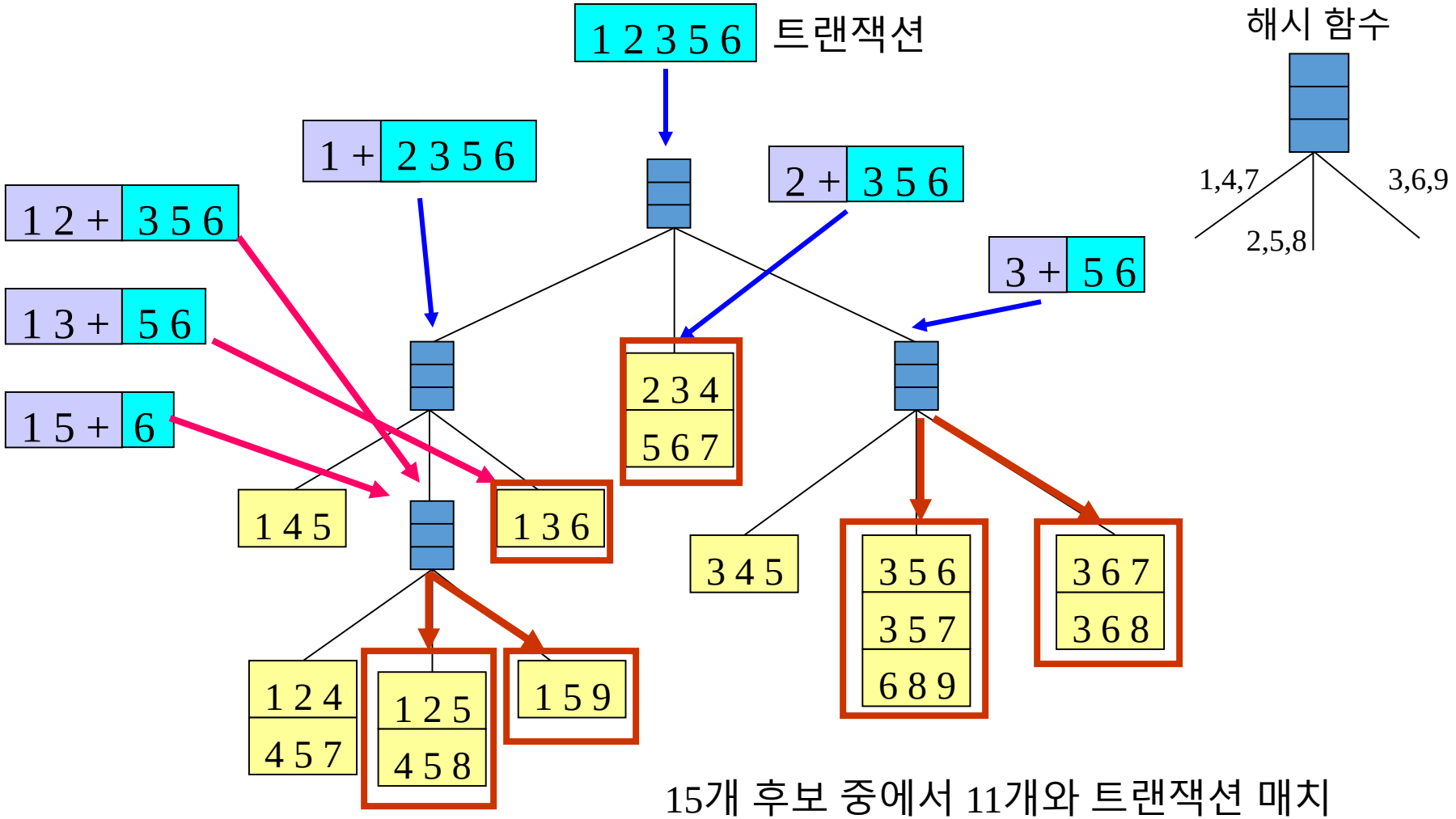
해시 트리를 사용한 지지도 계산



해시 트리를 사용한 지지도 계산



해시 트리를 사용한 지지도 계산



15개 후보 중에서 11개와 트랜잭션 매치

3. 연관 규칙 생성

규칙 생성

- 빈발 항목집합 L이 주어지면, $f \rightarrow L - f$ 가 최소지지도 요구사항을 만족하도록 비어있지 않은 모든 부분집합 $f \subset L$ 을 찾음
- {A,B,C,D}가 빈발 항목집합일 때, 후보 규칙:

$ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,
$A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$
$AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,
$BD \rightarrow AC$, $CD \rightarrow AB$,

- $|L| = k$ 이면, $2k - 2$개의 후보 결합규칙 ($L \rightarrow \varnothing$ 과 $\varnothing \rightarrow L$ 무시)

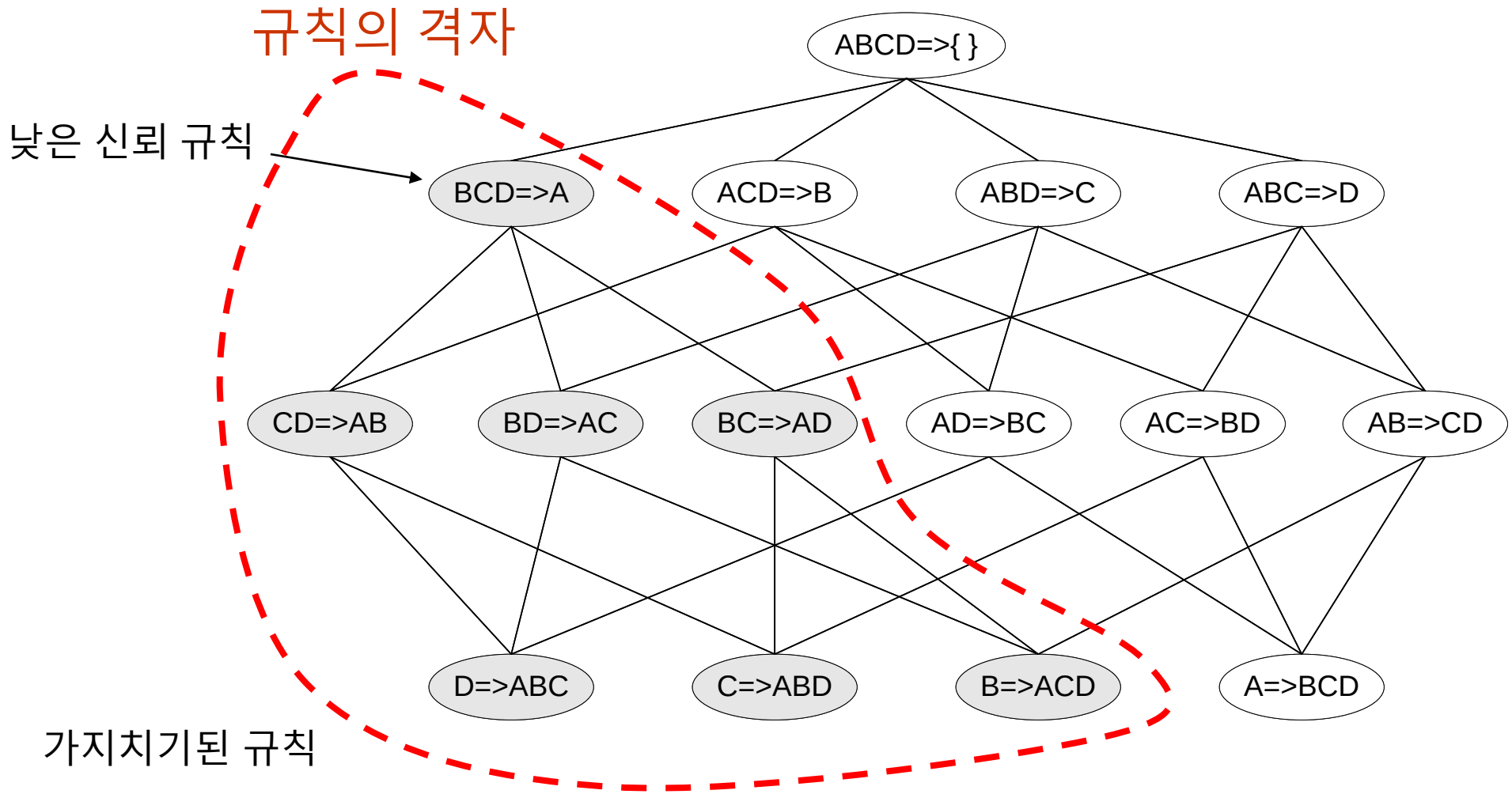
규칙 생성

- 일반적으로 신뢰도에는 안티-모노톤anti-monotone 속성이 없음

  $c(ABC \rightarrow D)$는 $c(AB \rightarrow D)$보다 크거나 작을 수 있음

- 그러나 동일항 항목집합으로부터 생성된 규칙의 신뢰도에는 안티-모노톤속성이 있음
  - 예: {A,B,C,D}는 빈발 4-항목집합이라 가정: $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$
  - 신뢰도는 규칙의 RHS에 있는 항목의 수와 관련하여 안티-모노톤

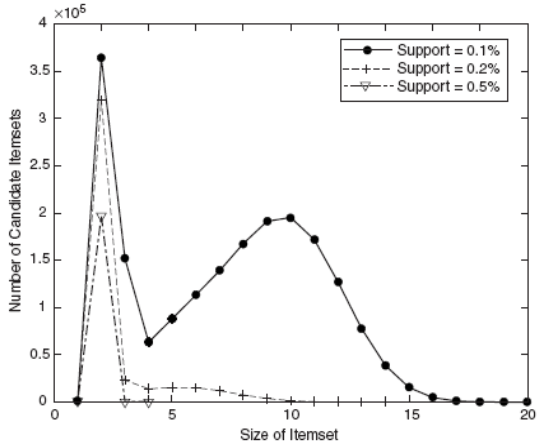
Apriori 알고리즘을 위한 규칙 생성



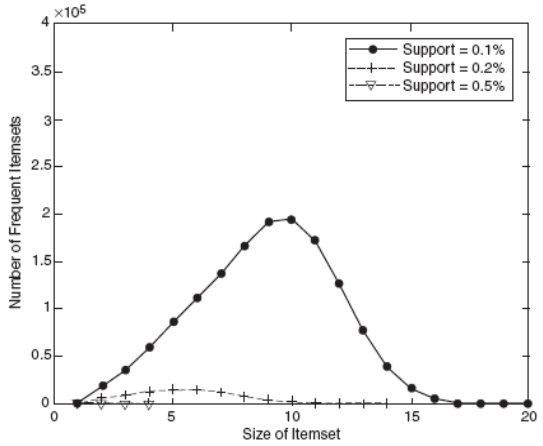
Apriori 복잡성에 영향을 미치는 요인

- 최소 지지도 임계값 선택
  - 지지도 임계값을 낮추면 항목 집합의 빈도가 높아짐
  - 후보의 수와 빈발 항목집합의 최대 길이를 증가시킬 수 있음
- 데이터 집합의 차원 (항목의 수)
  - 각 항목의 지지도 횟수를 저장하는데 더 많은 공간 필요
  - 빈발 항목의 수가 증가하면, 계산 및 입출력 비용도 증가할 수 있음
- 데이터베이스 크기
  - *Apriori*가 다중 패스multiple passes를 하기 때문에 알고리즘의 실행시간이 트랜잭션의 수에 따라 증가할 수 있음
- 평균 트랜잭션 너비
  - 밀집한 데이터 집합과 함께 트랜잭션 너비 증가
  - 빈발 항목집합과 해시 트리의 순회를 증가시킬 수 있음 (트랜잭션의 부분집합 수는 너비에 따라 증가)

Apriori 복잡성에 영향을 미치는 요인

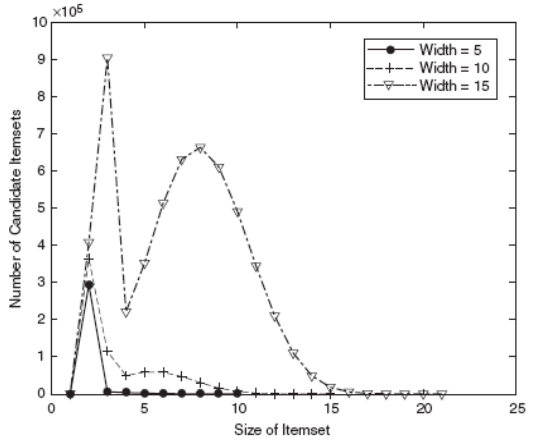


(a) Number of candidate itemsets.

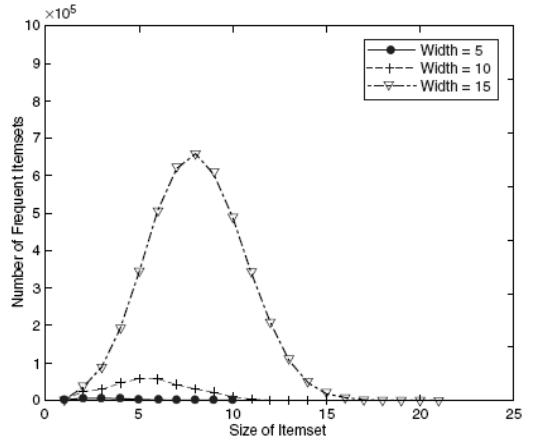


(b) Number of frequent itemsets.

Figure 6.13. Effect of support threshold on the number of candidate and frequent itemsets.



(a) Number of candidate itemsets.



(b) Number of Frequent Itemsets.

Figure 6.14. Effect of average transaction width on the number of candidate and frequent itemsets.

4. 빈발 항목집합 표현

빈발 항목집합의 간결한 표현

- 일부 항목집합은 상위집합supersets과 동일한 지지도를 가지기 때문에 중복

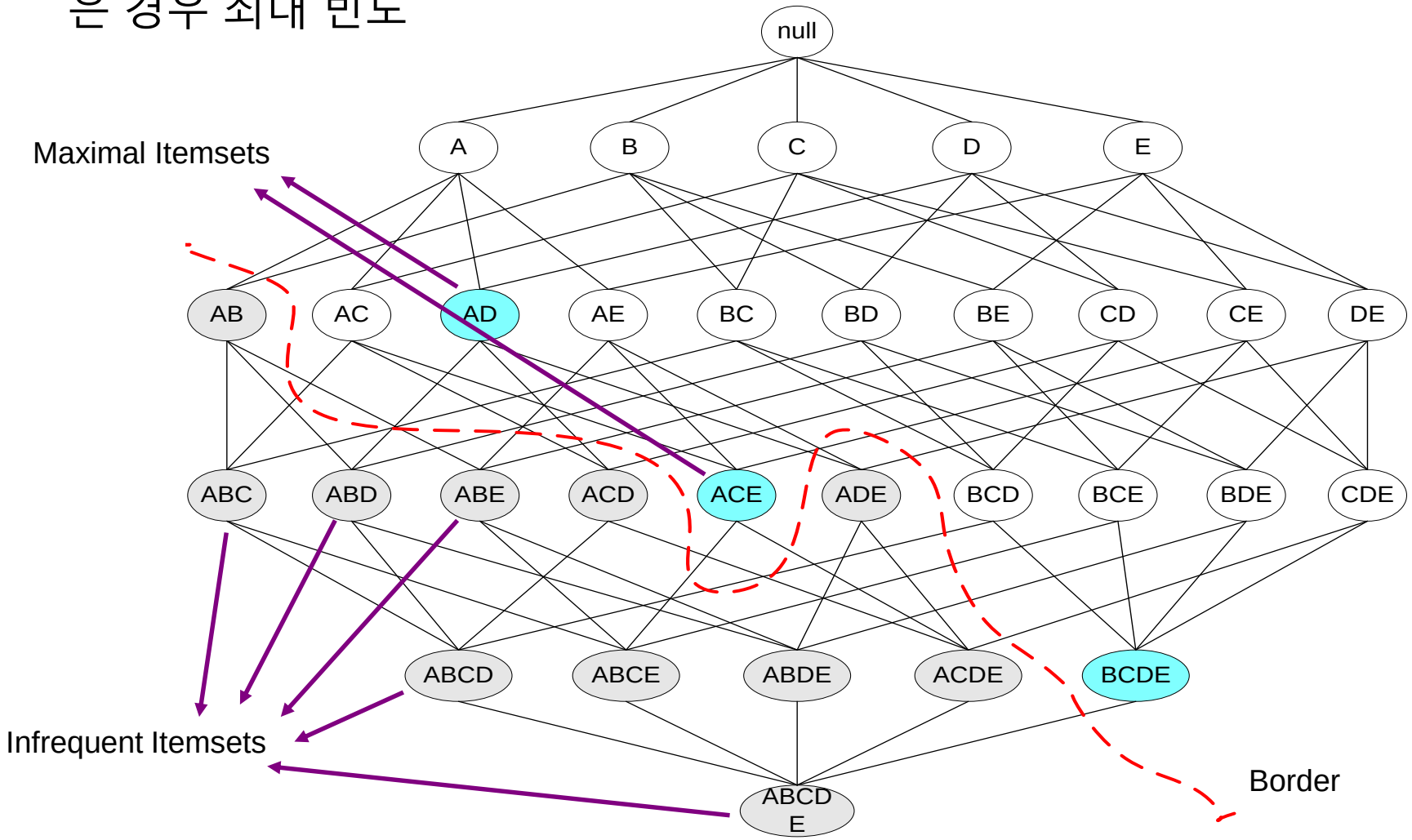
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

- 빈발 항목집합의 수 $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$

- 간결한 표현이 필요

최대 빈발 항목집합Maximal Frequent Itemset

항목 집합은 자주 발생하고 바로 자기 상위집합의 어느 것도 빈번하지 않은 경우 최대 빈도

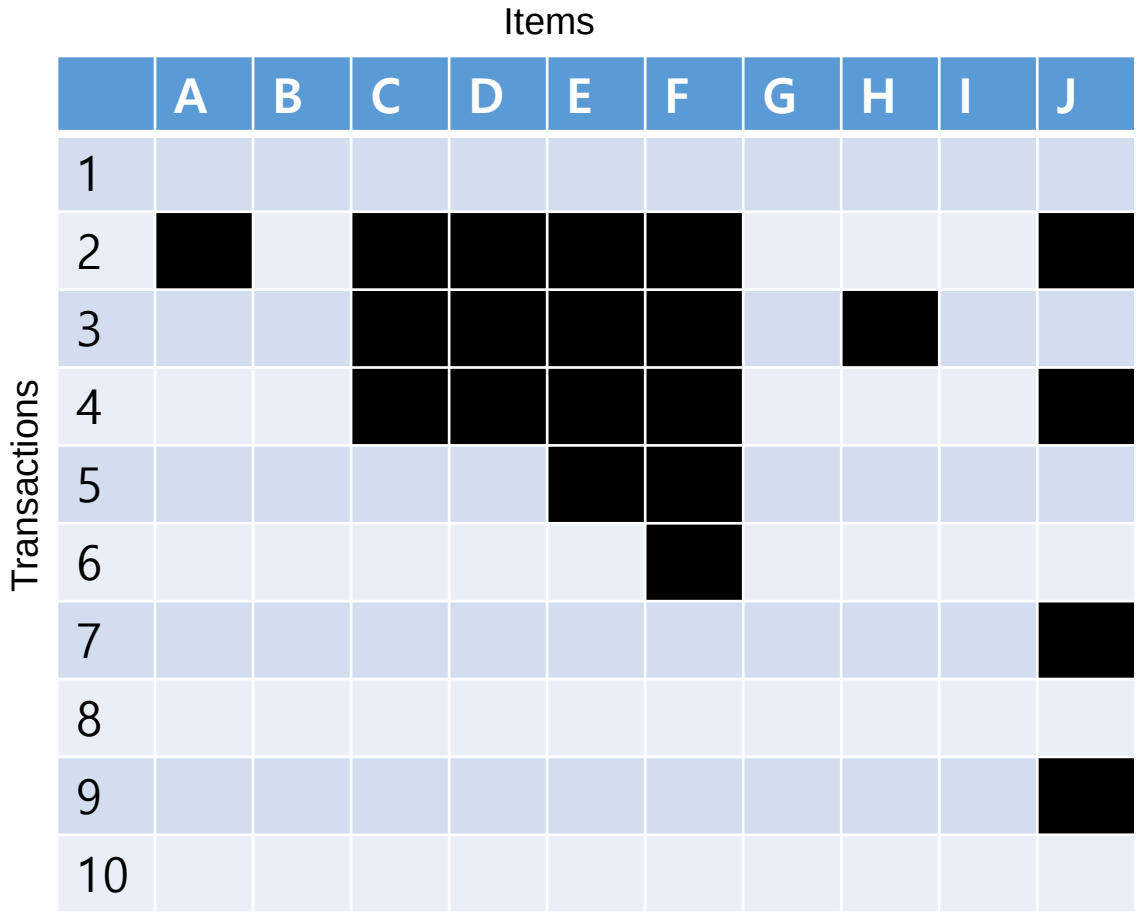


이 데이터의 최대 빈발 항목집합은 무엇인가?

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

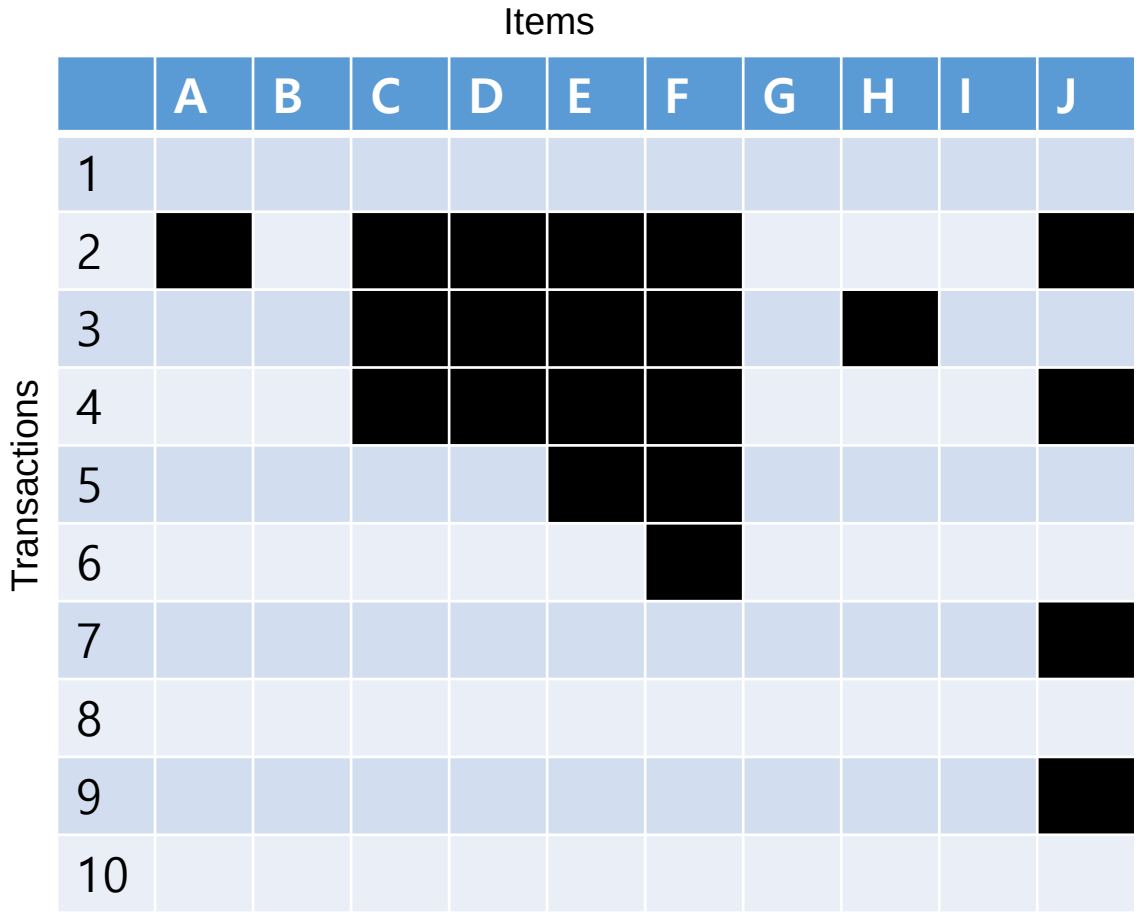
Minimum support threshold = 5

실제 예제



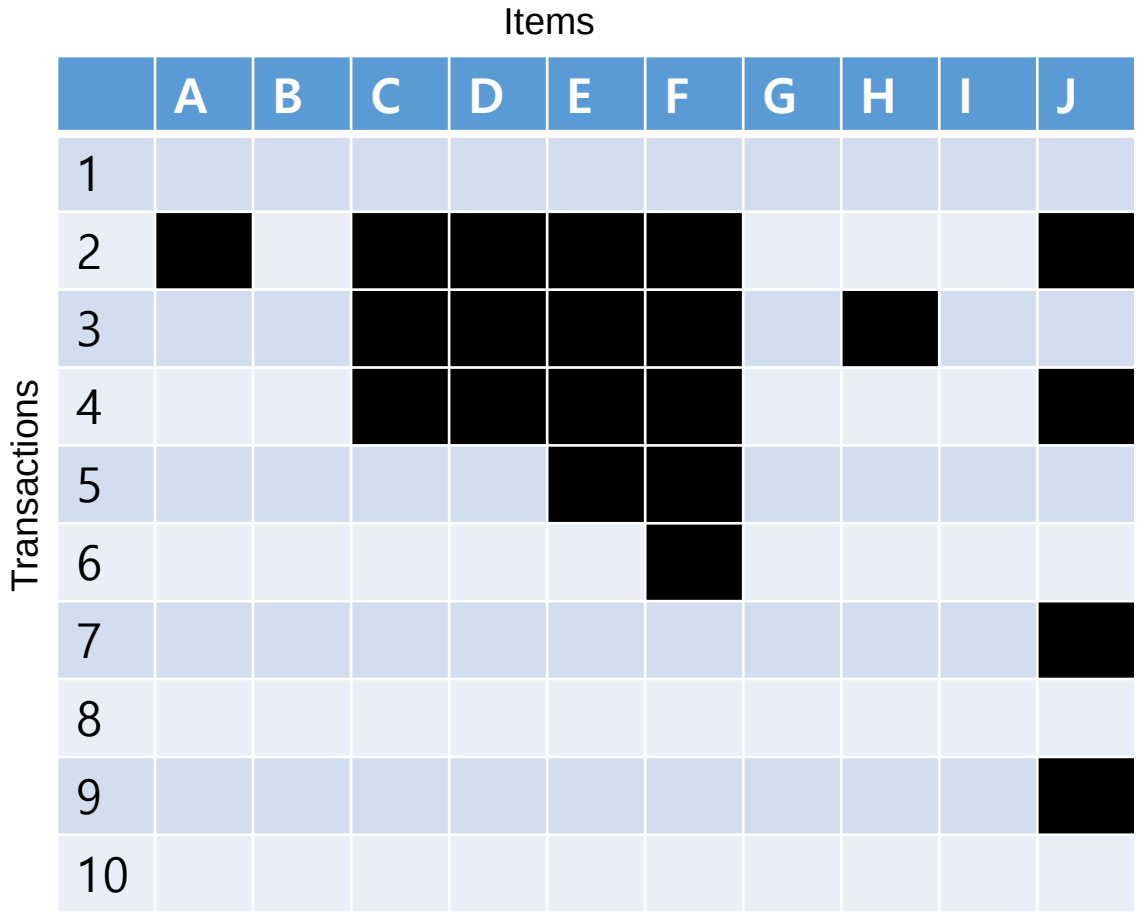
지지도 임계값 (by count) : 5
빈발 항목집합: ?

실제 예제



지지도 임계값 (by count) : 5
빈발 항목집합: {F}

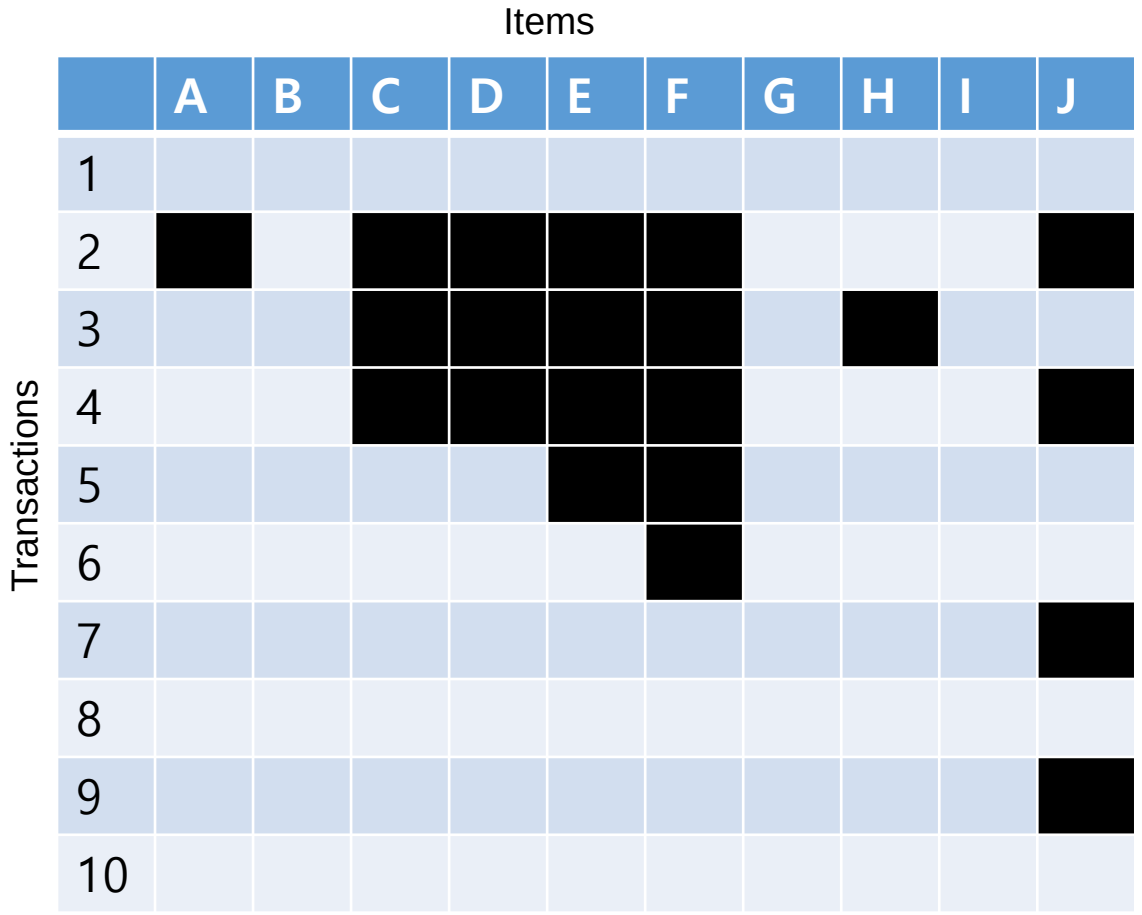
실제 예제



지지도 임계값 (by count) : 5
빈발 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: ?

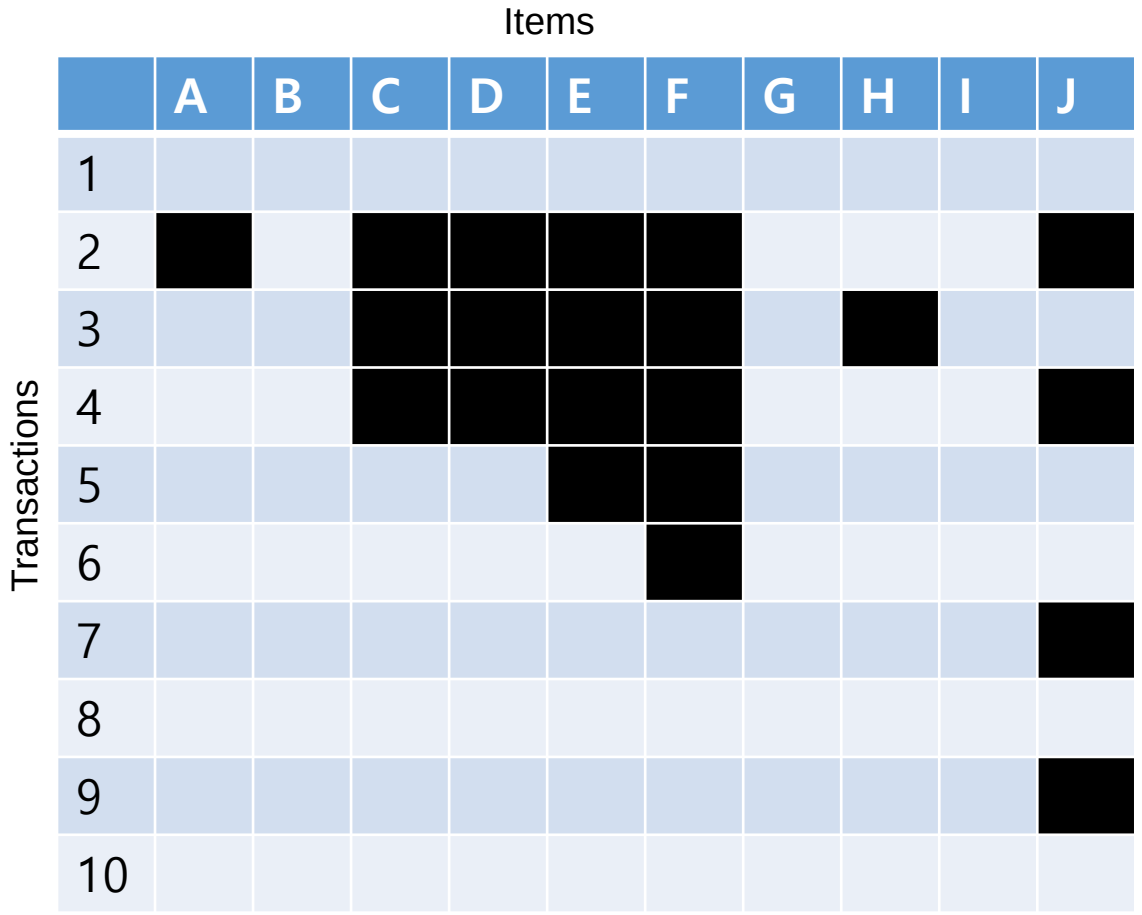
실제 예제



지지도 임계값 (by count) : 5
빈발 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}

실제 예제

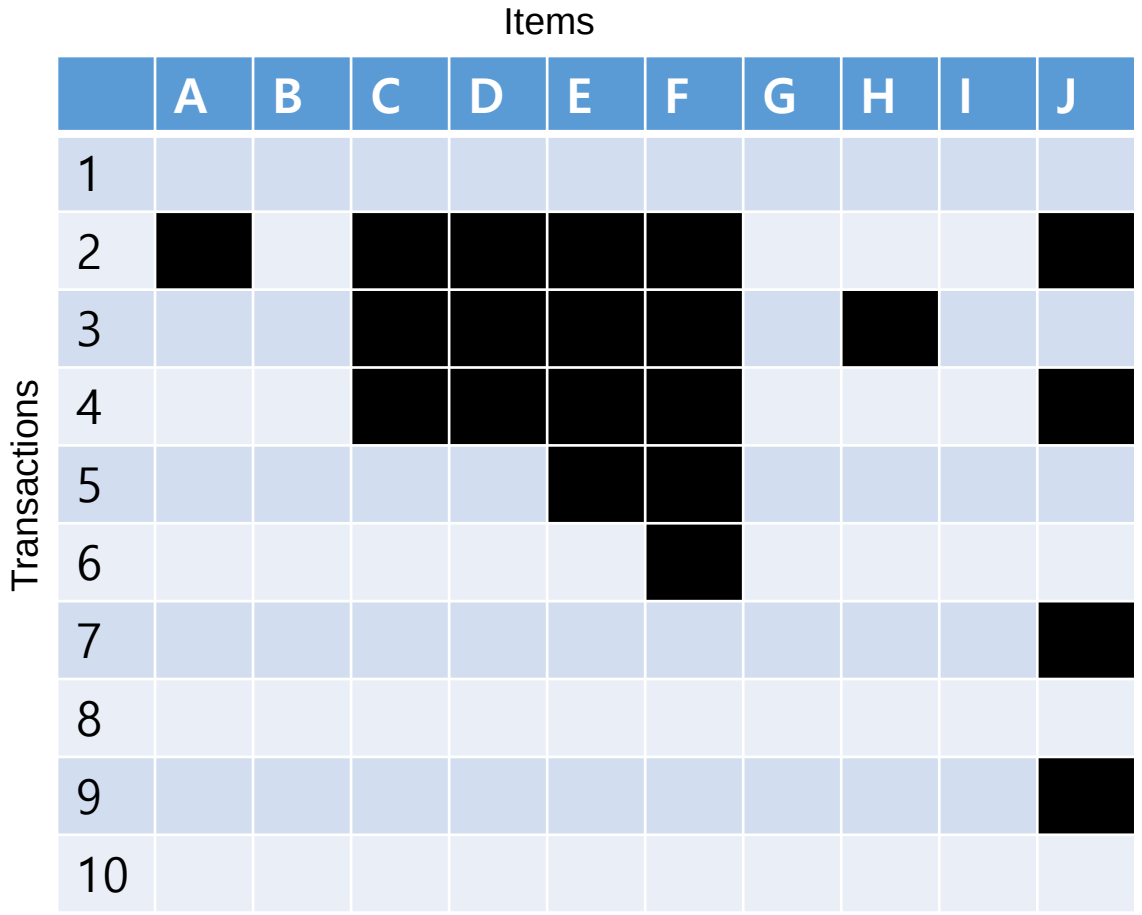


지지도 임계값 (by count) : 5
빈발 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}

지지도 임계값 (by count) : 3
빈발 항목집합: ?

실제 예제

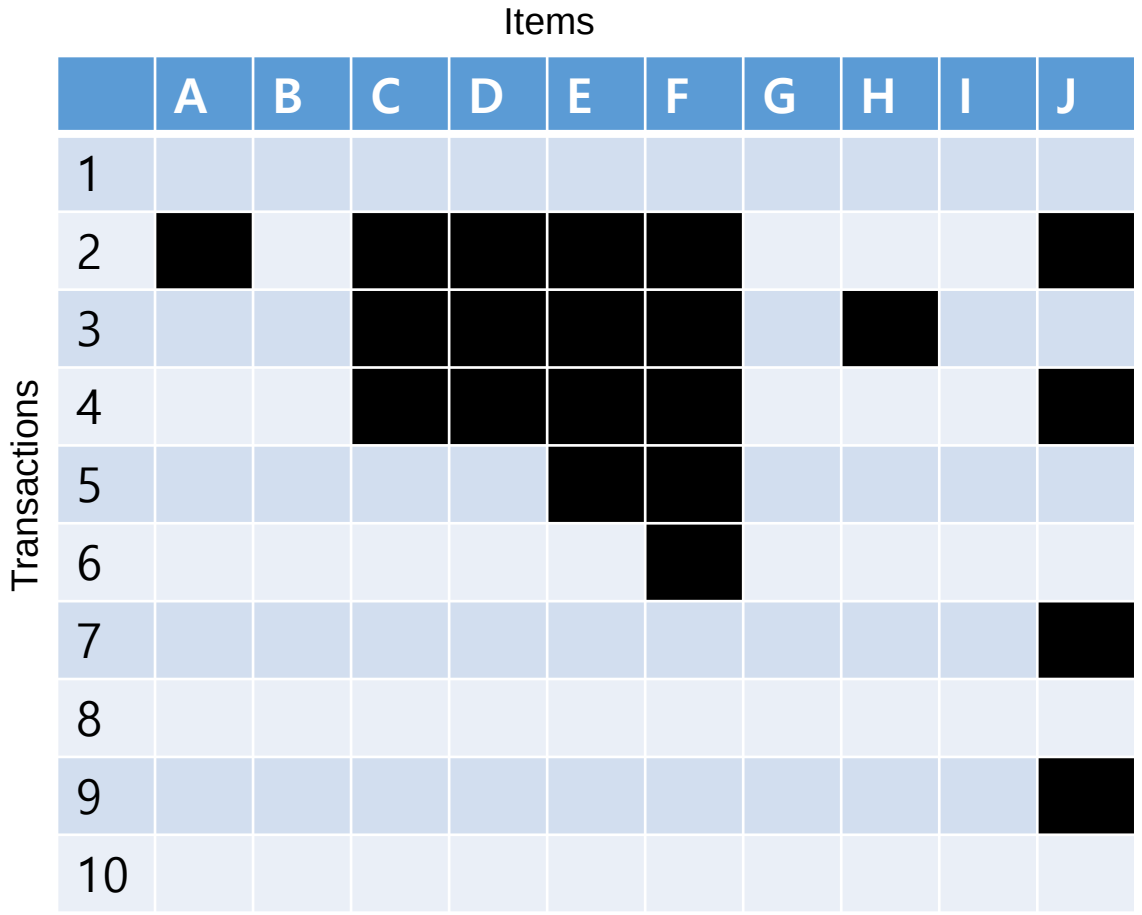


지지도 임계값 (by count) : 5
빈발 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}

지지도 임계값 (by count) : 3
빈발 항목집합:
{C,D,E,F}+{J}의 모든 부분 집합

실제 예제

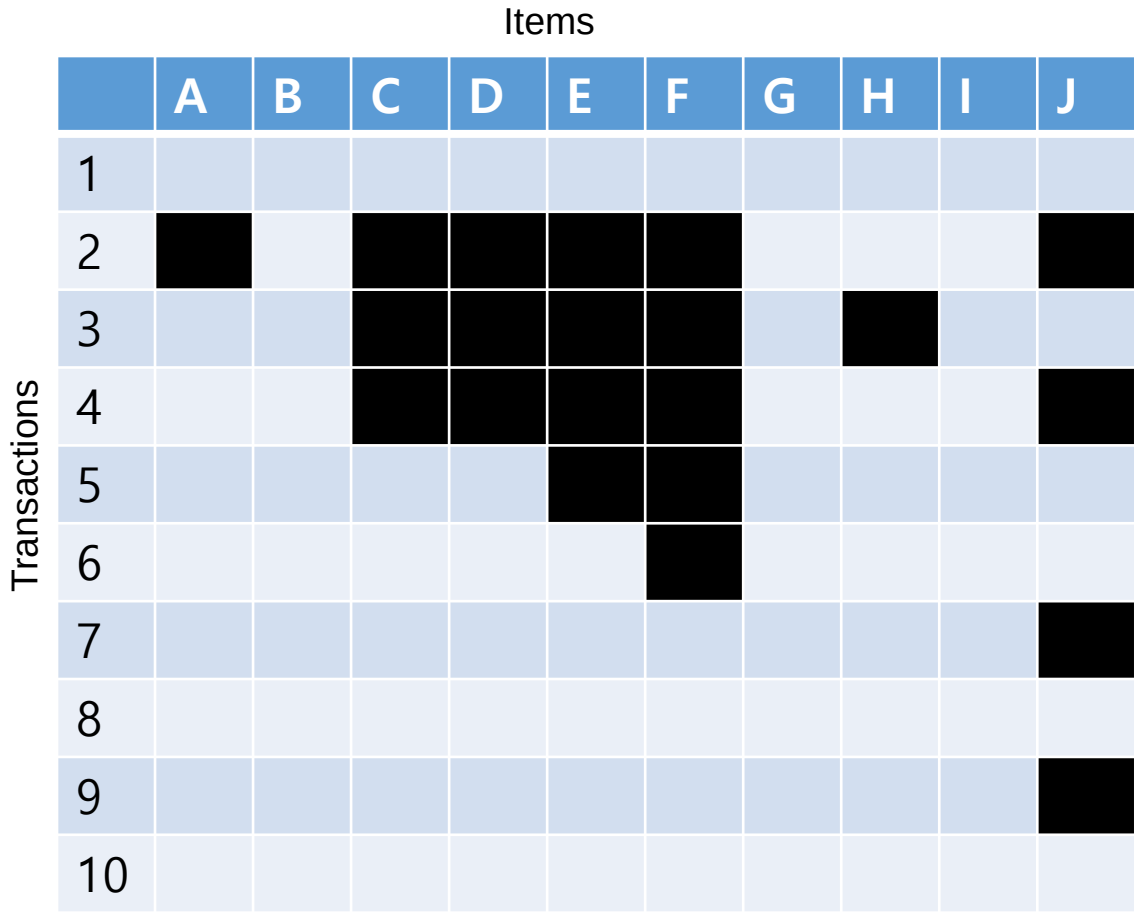


지지도 임계값 (by count) : 5
빈발 항목집합: {F}
최대 항목집합: ?

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}
최대 항목집합: ?

지지도 임계값 (by count) : 3
빈발 항목집합:
{C,D,E,F}+{J}의 모든 부분 집합
최대 항목집합: ?

실제 예제

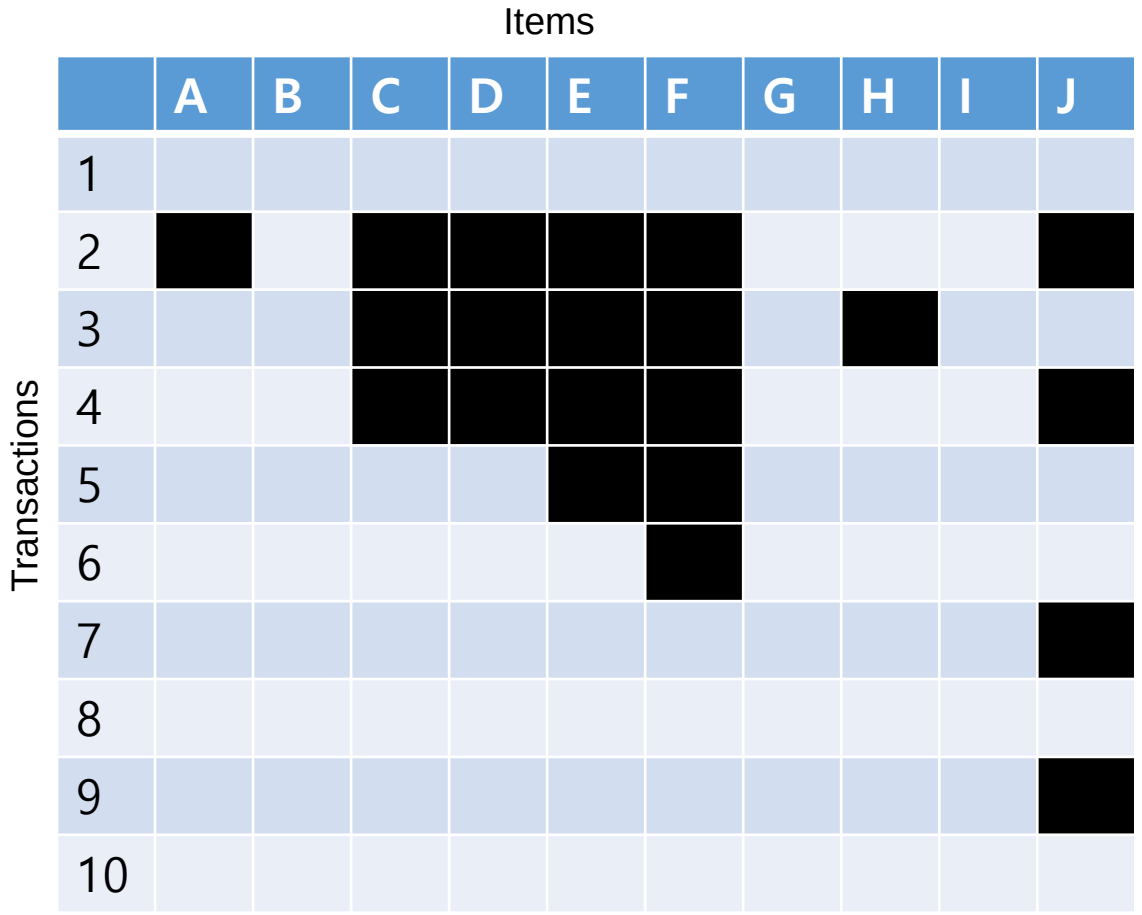


지지도 임계값 (by count) : 5
빈발 항목집합: {F}
최대 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}
최대 항목집합: ?

지지도 임계값 (by count) : 3
빈발 항목집합:
{C,D,E,F}+{J}의 모든 부분 집합
최대 항목집합: ?

실제 예제

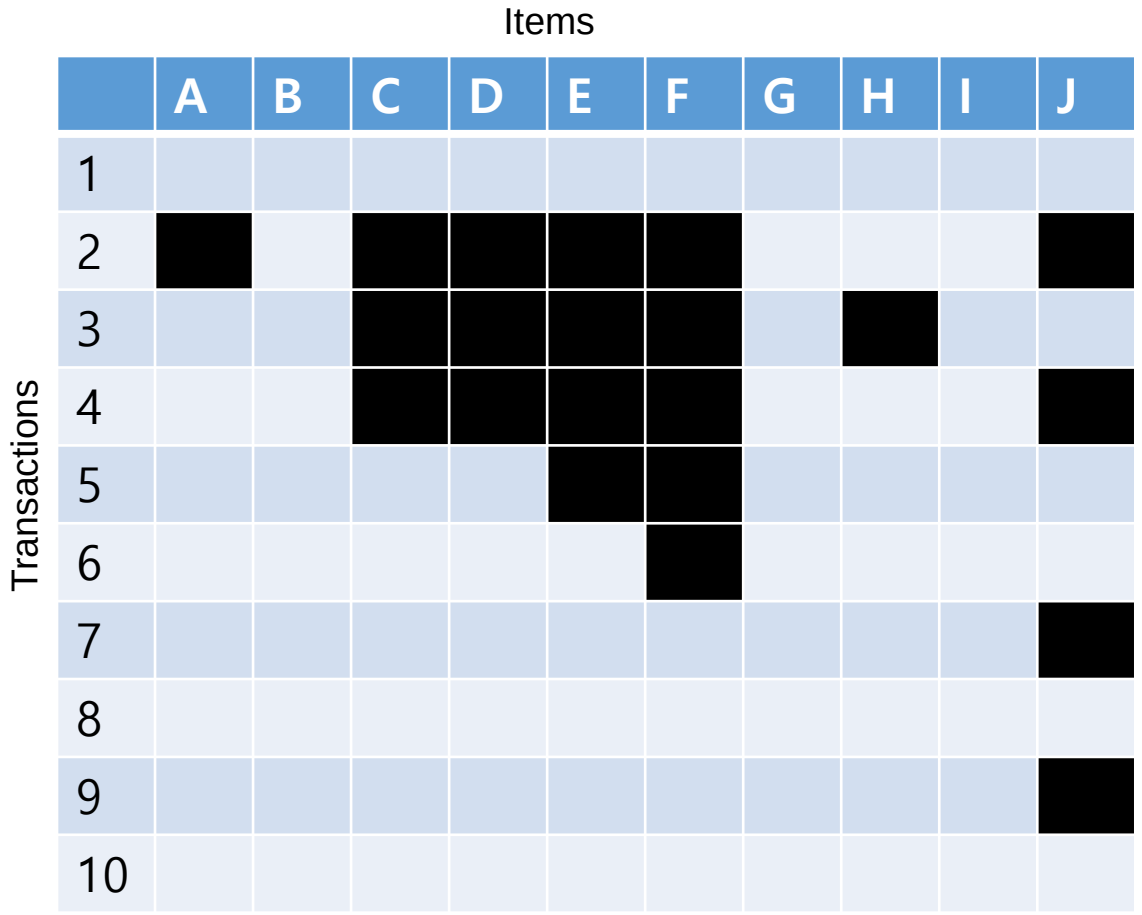


지지도 임계값 (by count) : 5
빈발 항목집합: {F}
최대 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}
최대 항목집합: {E,F},{J}

지지도 임계값 (by count) : 3
빈발 항목집합:
{C,D,E,F}+{J}의 모든 부분 집합
최대 항목집합: ?

실제 예제

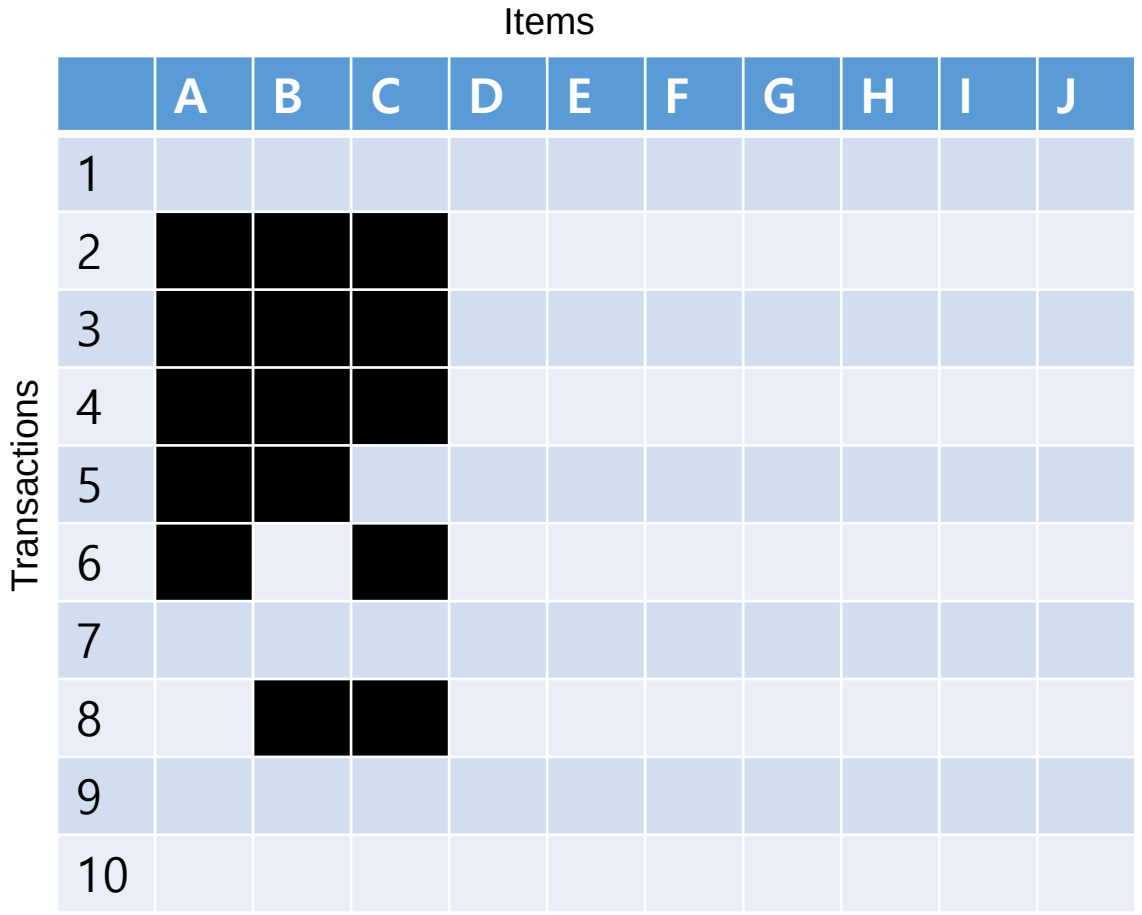


지지도 임계값 (by count) : 5
빈발 항목집합: {F}
최대 항목집합: {F}

지지도 임계값 (by count) : 4
빈발 항목집합: {E},{F},{E,F},{J}
최대 항목집합: {E,F},{J}

지지도 임계값 (by count) : 3
빈발 항목집합:
{C,D,E,F}+{J}의 모든 부분 집합
최대 항목집합: {C,D,E,F},{J}

실제 예제



지지도 임계값 (by count) : 5
최대 항목집합: {A}, {B}, {C}

지지도 임계값 (by count) : 4
최대 항목집합: {A,B},{A,C},{B,C}

지지도 임계값 (by count) : 3
최대 항목집합: {A,B,C}

닫힌 항목집합Closed Itemset

- 항목집합 X의 상위 집합 중 어느 것도 X와 동일한 지지도 횟수를 가지지 않으면, 항목집합 X는 닫혀있음
- 직접 상위집합 중 적어도 하나 이상에 X의 지지도 횟수를 가지면 X는 닫히지 않음

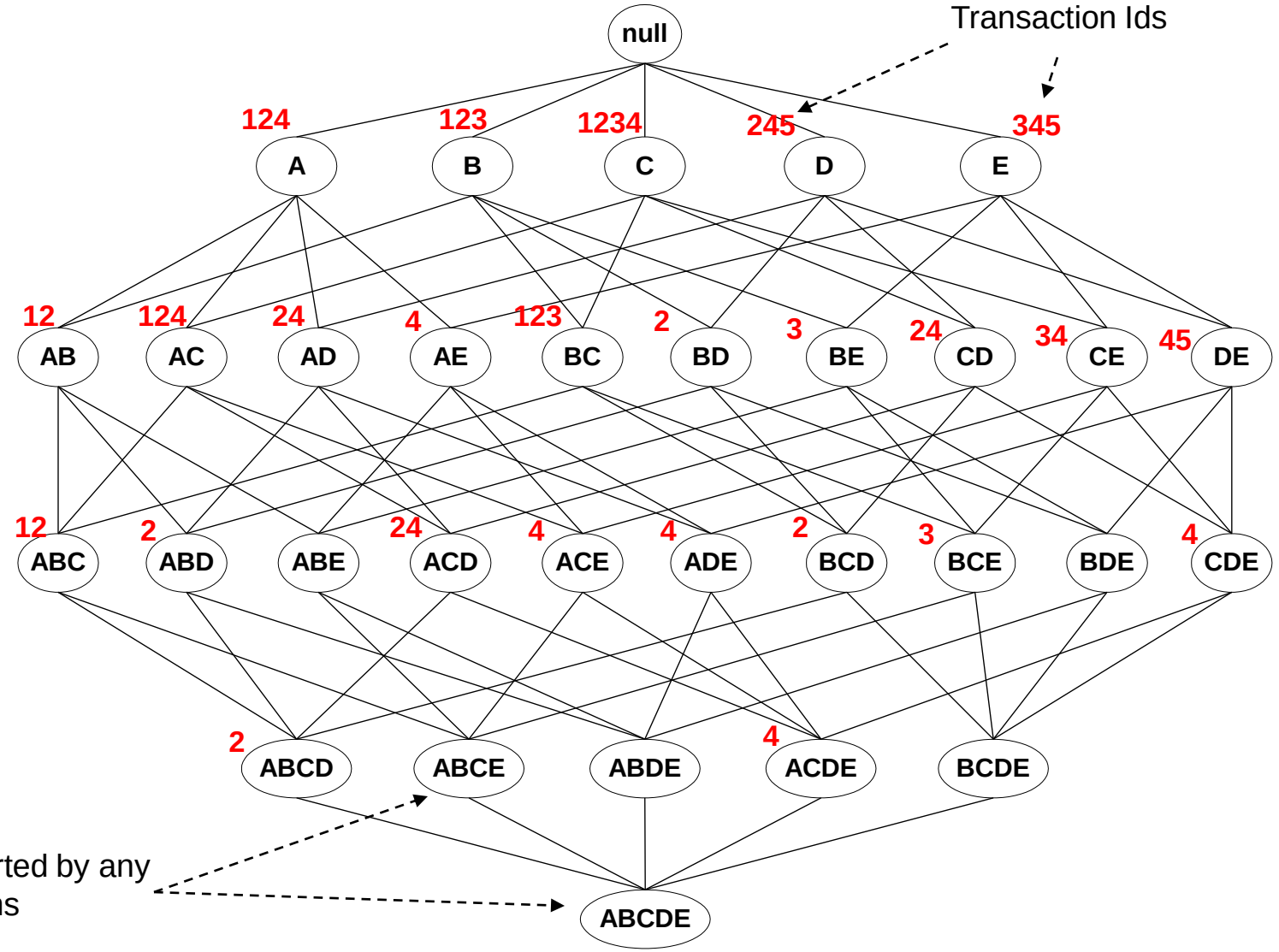
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

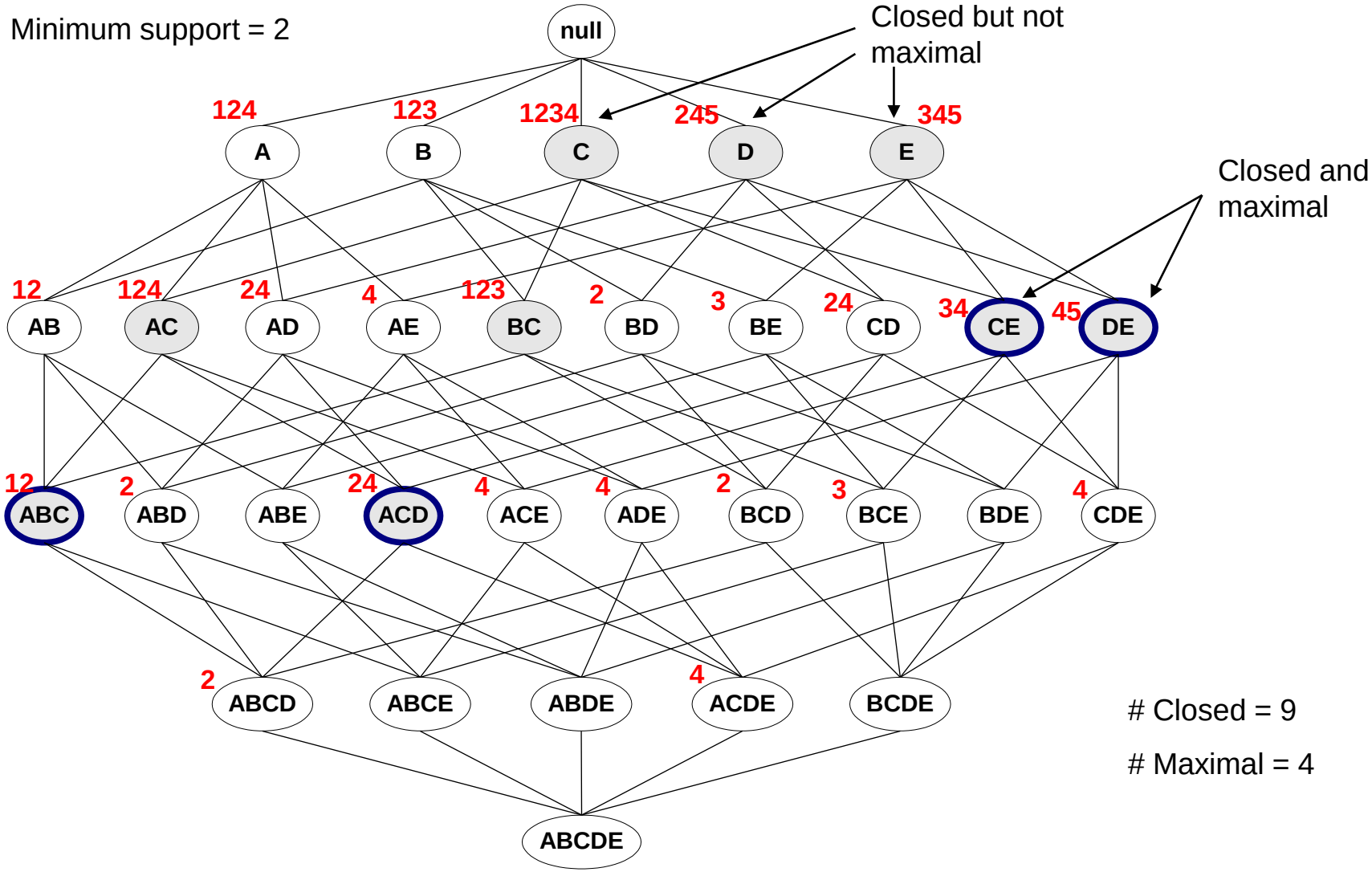
Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	2
{A,B,C,D}	2

최대Maximal vs 닫힌 항목집합Closed Itemsets

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



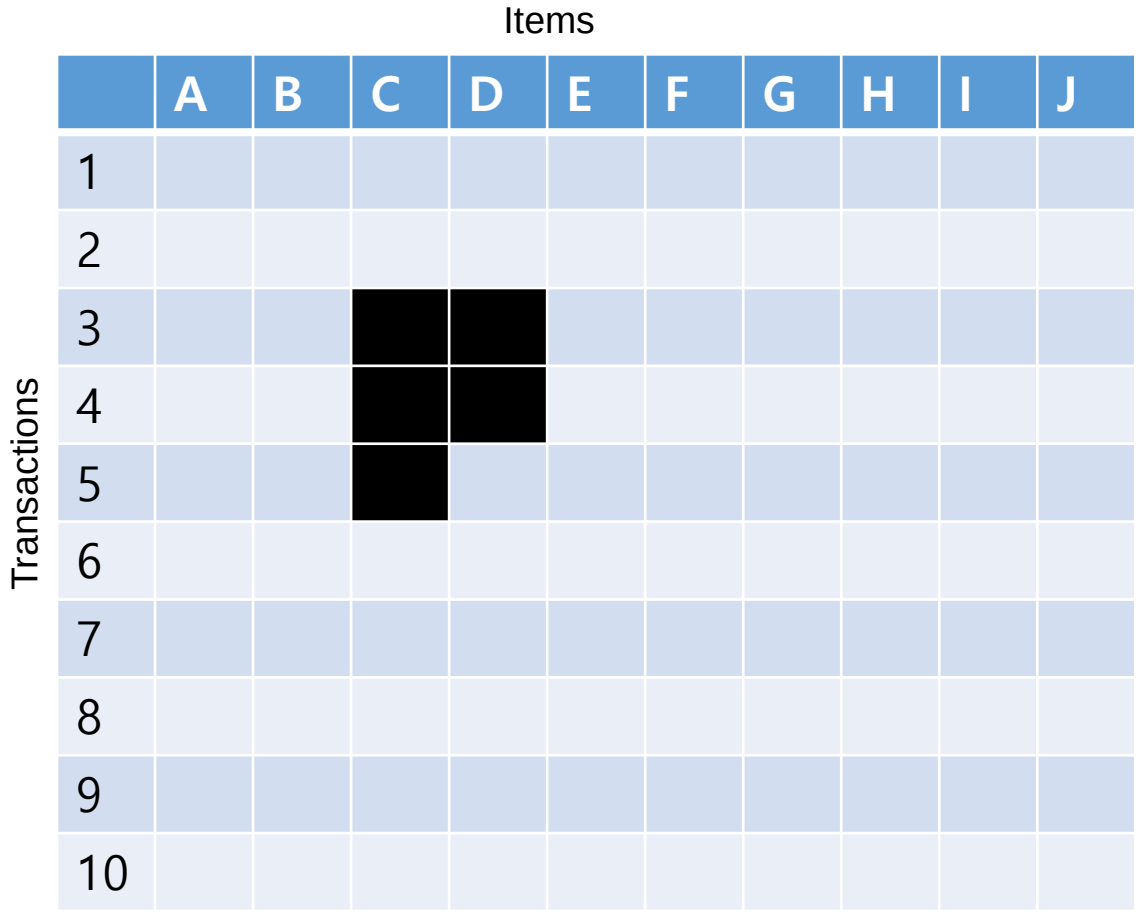
최대Maximal vs 닫힌 항목집합Closed Itemsets



이 데이터의 닫힌 항목집합은 무엇인가?

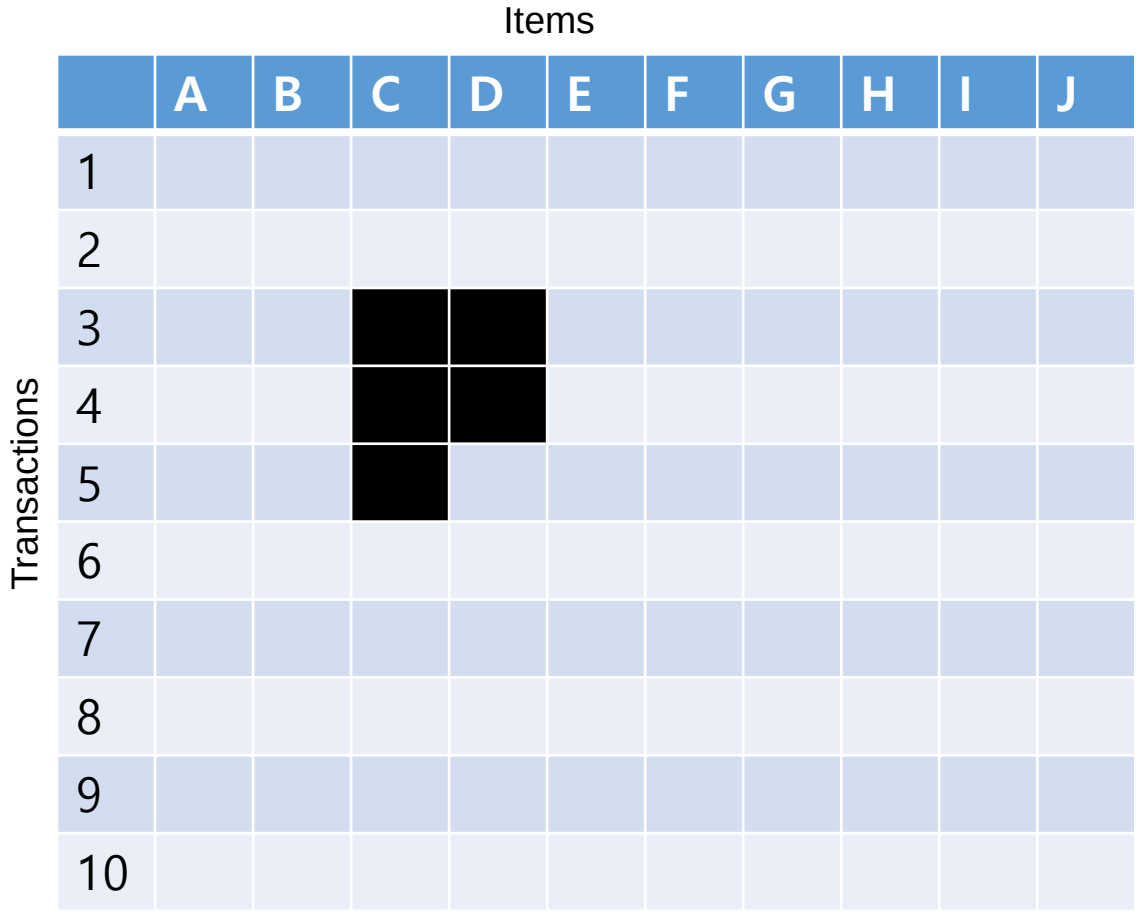
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

예제 1



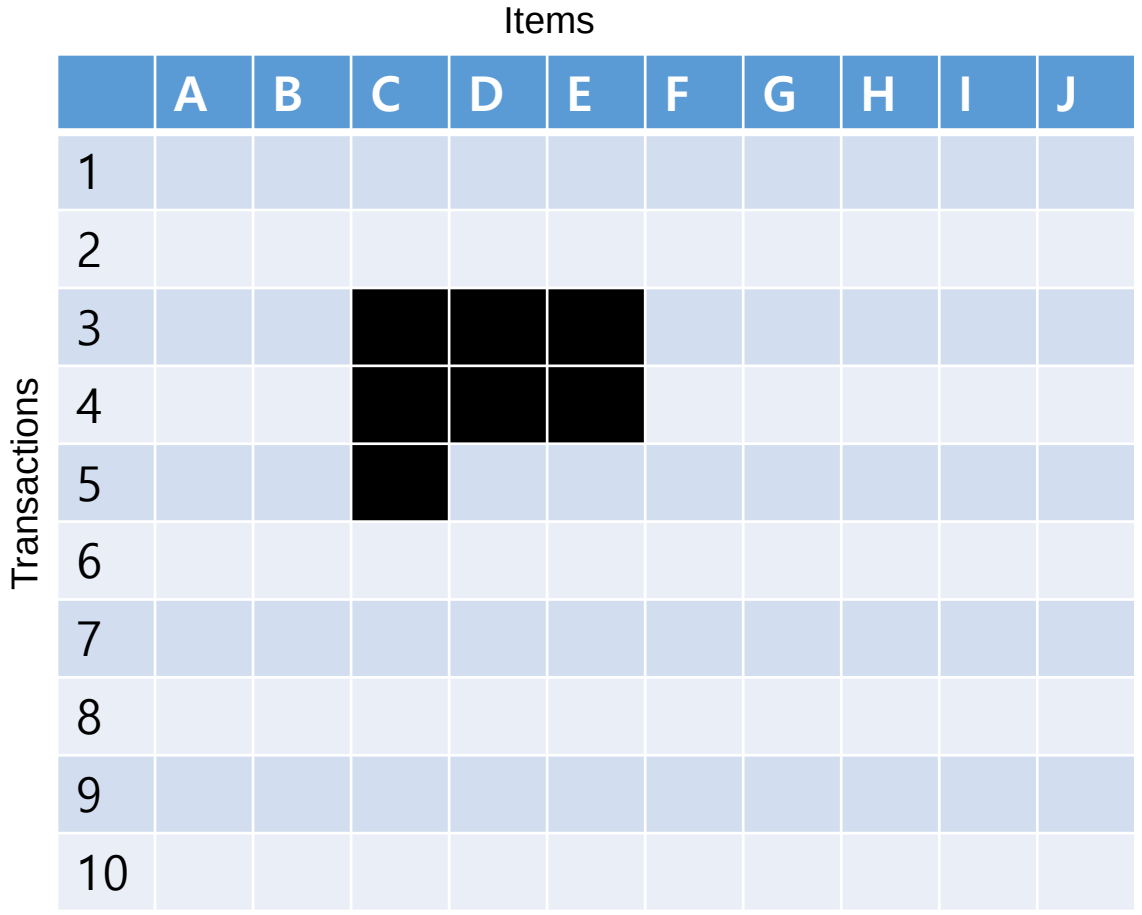
Itemsets	Support (counts)	Closed itemsets
{C}	3	
{D}	2	
{C,D}	2	

예제 1



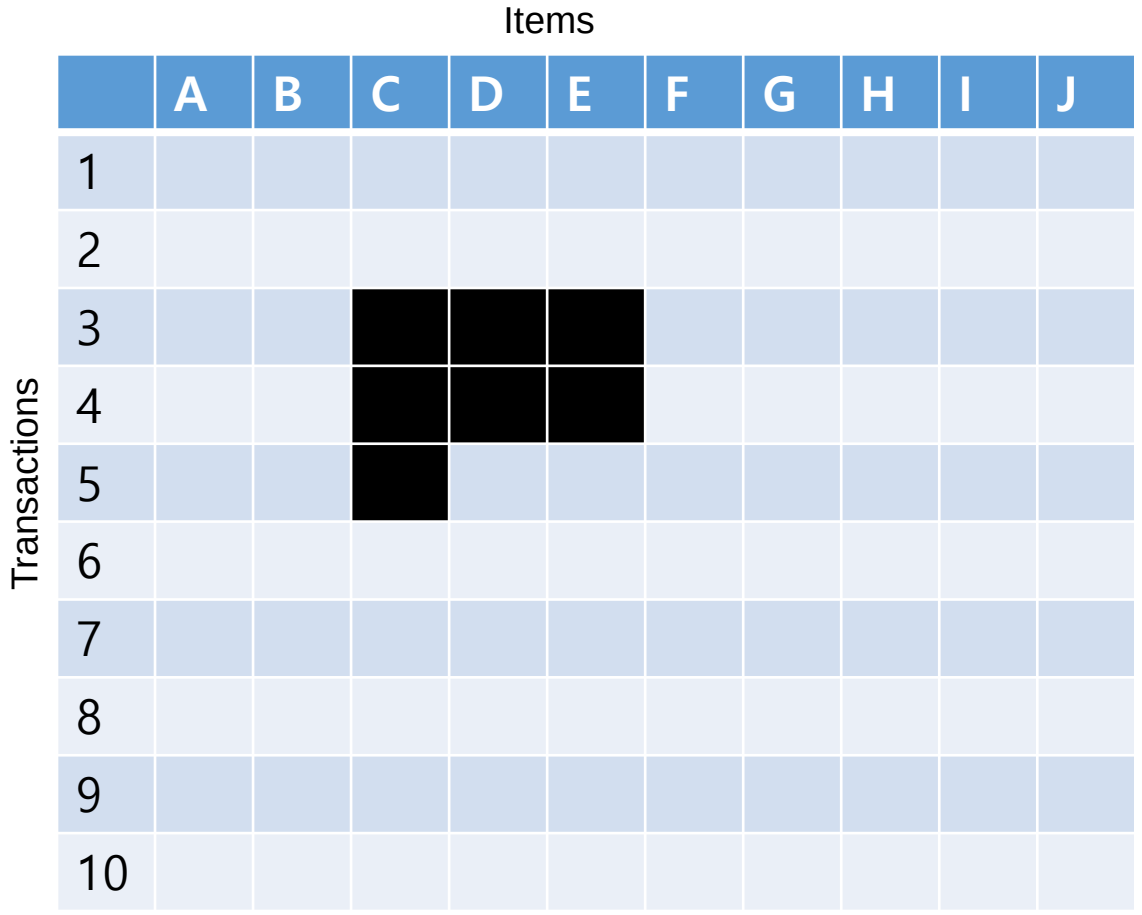
Itemsets	Support (counts)	Closed itemsets
{C}	3	✓
{D}	2	
{C,D}	2	✓

예제 2



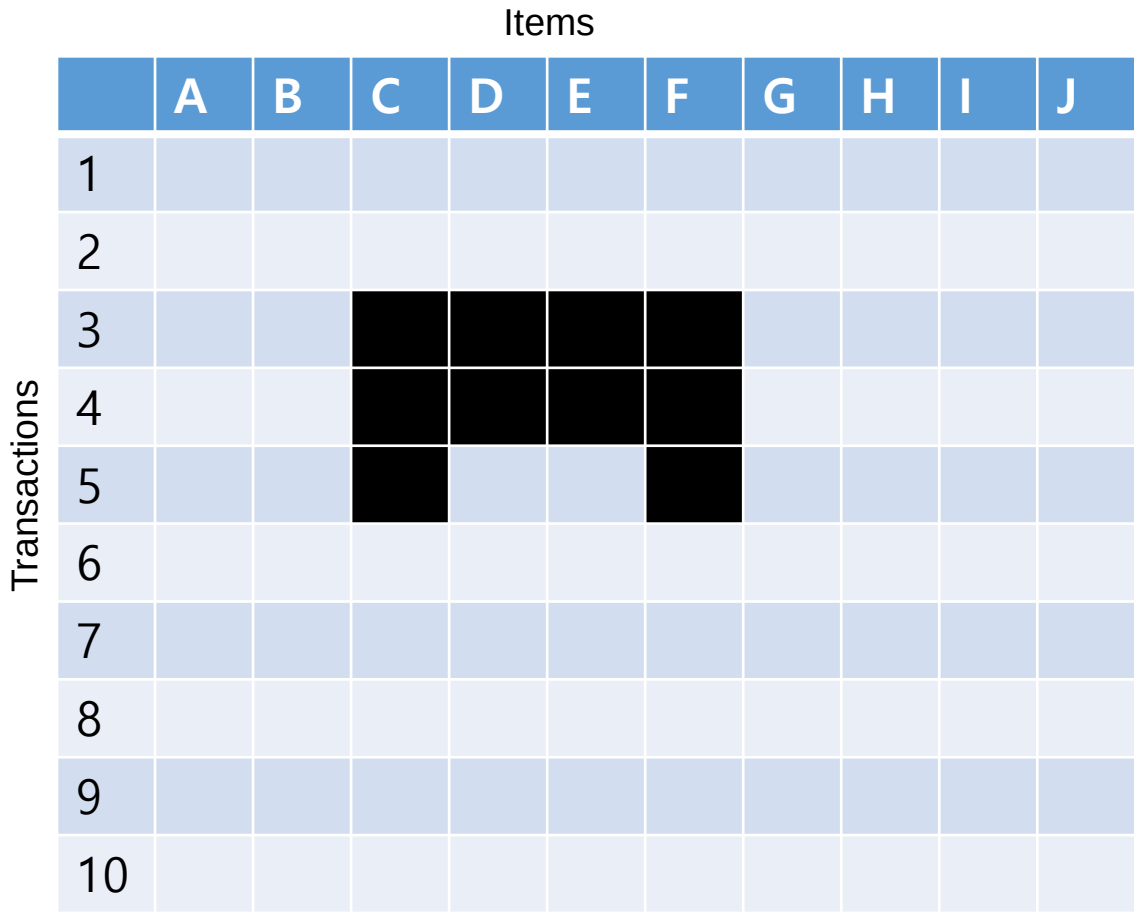
Itemsets	Support (counts)	Closed itemsets
{C}	3	
{D}	2	
{E}	2	
{C,D}	2	
{C,E}	2	
{D,E}	2	
{C,D,E}	2	

예제 2



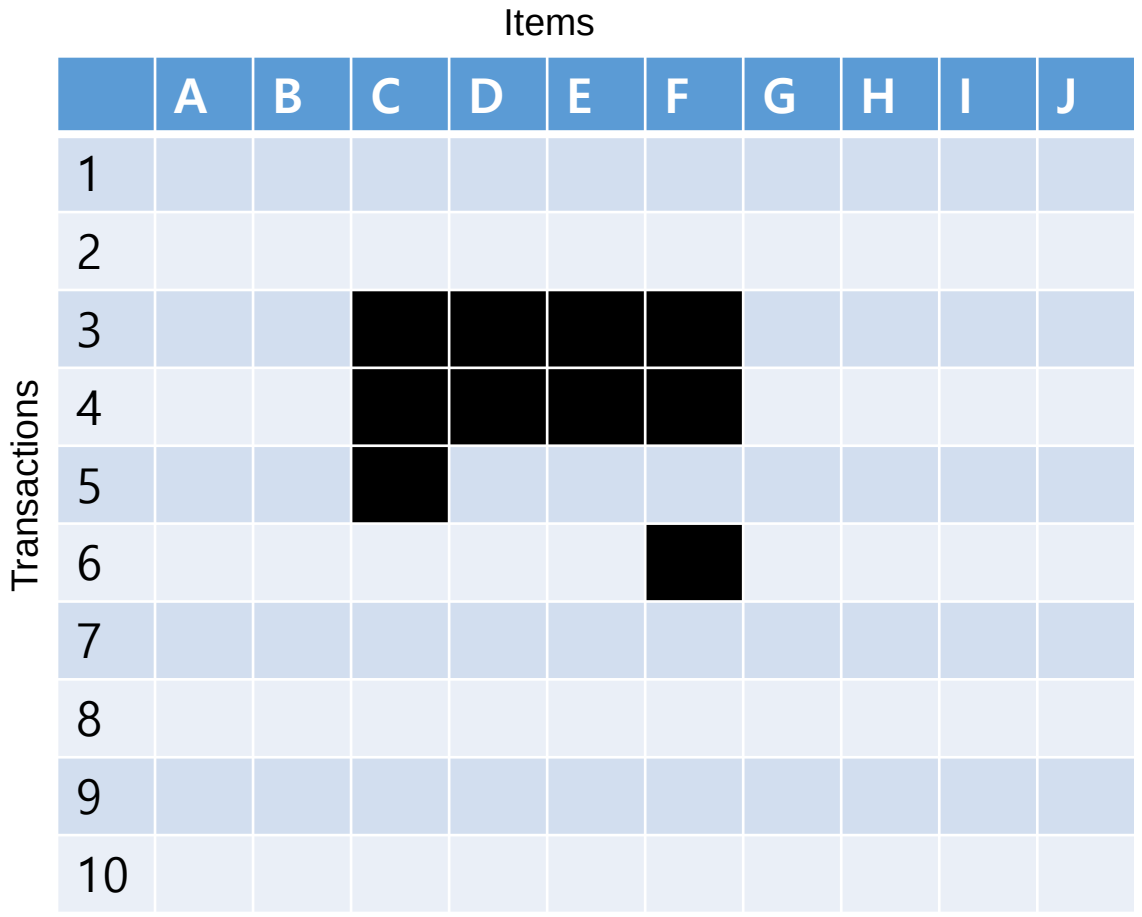
Itemsets	Support (counts)	Closed itemsets
{C}	**3**	✓
{D}	2	
{E}	2	
{C,D}	2	
{C,E}	2	
{D,E}	2	
{C,D,E}	**2**	✓

예제 3



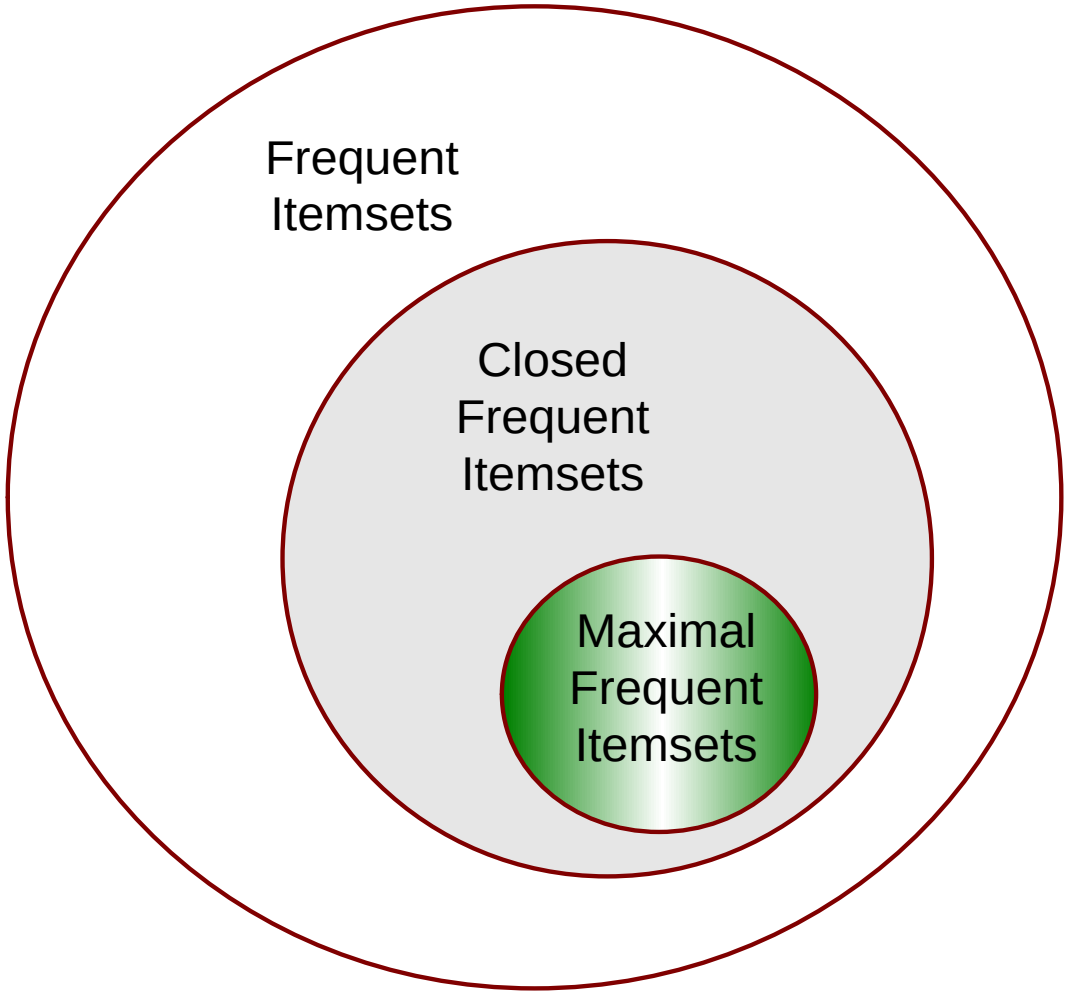
Closed itemsets: {C,D,E,F}, {C,F}

예제 4



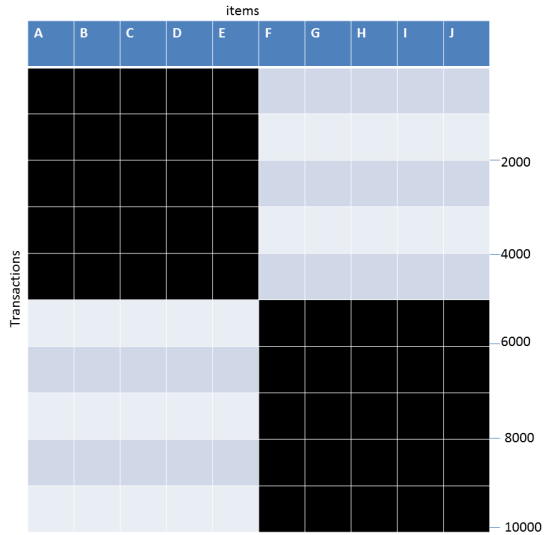
Closed itemsets: {C,D,E,F}, {C}, {F}

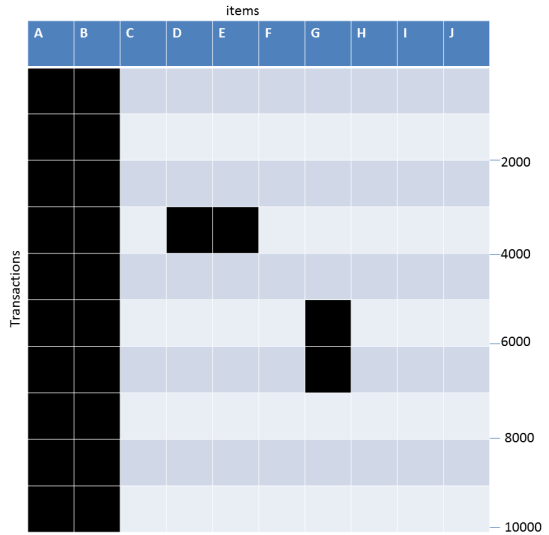
최대Maximal vs 닫힌 항목집합Closed Itemsets

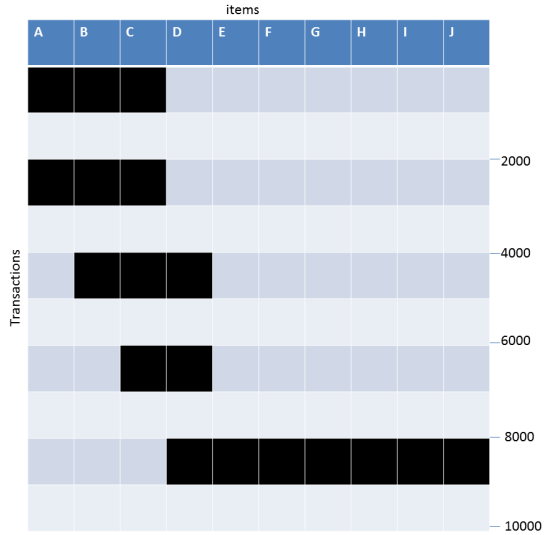


예제 질문

- 트랜잭션 데이터 집합(어두운 셀은 트랜잭션에서 항목이 있음을 나타냄)과 20% 지지도 임계값이 주어진다면, 다음 질문에 답하시오
  - 각 데이터 집합의 빈 항목집합의 수?
  - 어떤 데이터 집합이 가장 자주 항목집합을 생성하나?
  - 어떤 데이터 집합이 가장 빈번하게 항목집합을 생성하나?
  - 어떤 데이터 집합이 최대 지지도가 가장 높은 빈발 항목집합을 생성하나?
  - 어떤 데이터 집합이 다양한 지원 수준(즉, 20%에서 70% 이상까지 범위의 지지도를 갖는 항목을 포함하는 항목집합)을 포함하는 빈발 항목집합을 생성하나?
  - 각 데이터 집합에 대한 최대 빈발 항목집합의 수?
  - 어떤 데이터 집합이 최대 빈발 항목집합을 생성하나?
  - 각 데이터 집합에 대해 닫힌 빈발 항목집합의 수?
  - 어떤 데이터 집합이 닫힌 빈발 항목집합을 가장 많이 생성하나?







5. 연관 패턴 평가

패턴 평가Pattern Evaluation

- 연관 규칙 알고리즘은 많은 수의 규칙을 생성할 수 있음
- 유용성 척도Interestingness Measures를 사용하여 패턴을 잘라내거나 순위를 매길 수 있음
- 원래 공식에서 사용된 유일한 척도는 지지도와 신뢰도

유용성 척도Interestingness Measures 계산

- 주어진 $X \rightarrow Y$ 또는 {X,Y}, 유용성을 계산하는데 필요한 정보는 분할표contingency table에서 얻을 수 있음

분할표

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	N

f_{11}: support of X and Y
f_{10}: support of $\underline{X}$ and $\overline{Y}$
f_{01}: support of $\underline{X}$ and $\underline{Y}$
f_{00}: support of X and $\overline{Y}$

다양한 측정값을 정의하는데 사용

support, confidence, Gini, entropy, etc.

신뢰도의 결점

Customers	Tea	Coffee	…
C1	0	1	…
C2	1	0	…
C3	1	1	…
C4	1	0	…
…			

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence ≅ P(Coffee|Tea) = 15/20 = 0.75

Confidence > 50%, 차를 마시는 사람들은 커피를 마시지 않는 것보다 커피를 마시는 경향이 있음을 의미

규칙은 합리적으로 보임

신뢰도의 결점

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence= P(Coffee|Tea) = 15/20 = 0.75

그러나 P(Coffee) = 0.9, 이는 사람이 차를 마시는 것이 사람이 커피를 마실 가능성을 줄인다는 것을 의미

⇒ P(Coffee|$\overline{\text{Tea}}$) = 75/80 = 0.9375

연관 규칙 측정

- 우리가 정말로 원하는 규칙은 무엇인가?
  - 신뢰도($X \rightarrow Y$)는 충분히 높아야 함
    - X를 사는 사람들이 Y를 사지 않는 것보다 Y를 살 가능성을 높이기 위해
  - 신뢰도 ($X \rightarrow Y$) > 지지도(Y)
    - 그렇지 않으면 규칙 X가 실제로 동일한 트랜잭션에서 항목 Y를 가질 기회를 줄이므로 규칙이 오해가 됨
    - 이 제한을 포착하는 어떤 측정이 있는가?
      - 답변: 많이 있음

통계적 독립성Statistical Independence

- 기준 $\text{confidence}(X \rightarrow Y) = \text{support}(Y)$는 다음과 같음:
  - $P(Y \mid X) = P(Y)$
  - $P(X,Y) = P(X) \times P(Y)$

If $P(X,Y) > P(X) \times P(Y)$: X & Y는 양의 상관 관계가 있음

If $P(X,Y) < P(X) \times P(Y)$: X & Y는 음의 상관 관계가 있음

통계적 의존성을 고려한 측정

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

항목집합에 관심이 있는 동안 리프트lift가 규칙에 사용

$$PS = P(X,Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

예제: Lift/Interest

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

신뢰도= P(Coffee|Tea) = 0.75

그러나 P(Coffee) = 0.9

⇒ Lift = 0.75/0.9= 0.8333 (< 1, 따라서 부정적으로 연관됨)

그래서 가지치기를 위해 confidence/lift를 사용하는 것으로 충분한가?

Lift 또는 Interest

	Y	$\bar{Y}$	
X	10	0	10
$\bar{X}$	0	90	90
	10	90	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

	Y	$\bar{Y}$	
X	90	0	90
$\bar{X}$	0	10	10
	90	10	100

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

통계적 독립성:

If P(X,Y)=P(X)P(Y) => Lift = 1

문헌에 제시된 많은 측정법

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$

Q & A