Data Mining
Association Analysis: Basic Concepts and Algorithms

Lecture Notes for Chapter 6

정보보호 및 지능형 IoT 연구실
Information Security & Intelligent IoT

Contents

Rule Generation

Rule Generation from frequent itemset

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  - If {A,B,C,D} is a frequent itemset, candidate rules:

    ABC →D, ABD →C, ACD →B, BCD →A,
    A →BCD, B →ACD, C →ABD, D →ABC
    AB →CD, AC → BD, AD → BC, BC →AD,
    BD →AC, CD →AB,

- If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)

Rule Generation from frequent itemset

- How to efficiently generate rules from frequent itemsets?
  - 신뢰도(confidence)는 anti-monotone 성질을 가지지 않는다. → Apriori 특성 사용이 어려움

    c(ABC →D) can be larger or smaller than c(AB →D)

  - 그러나, 동일한 항목집합에서 생성된 규칙에 대해서는 anti-monotone 성질이 성립
  - (That is, confidence is anti-monotone w.r.t number of items on the RHS of the rule, or monotone w.r.t. the LHS of the rule)
  - e.g., L = {A,B,C,D}:

$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

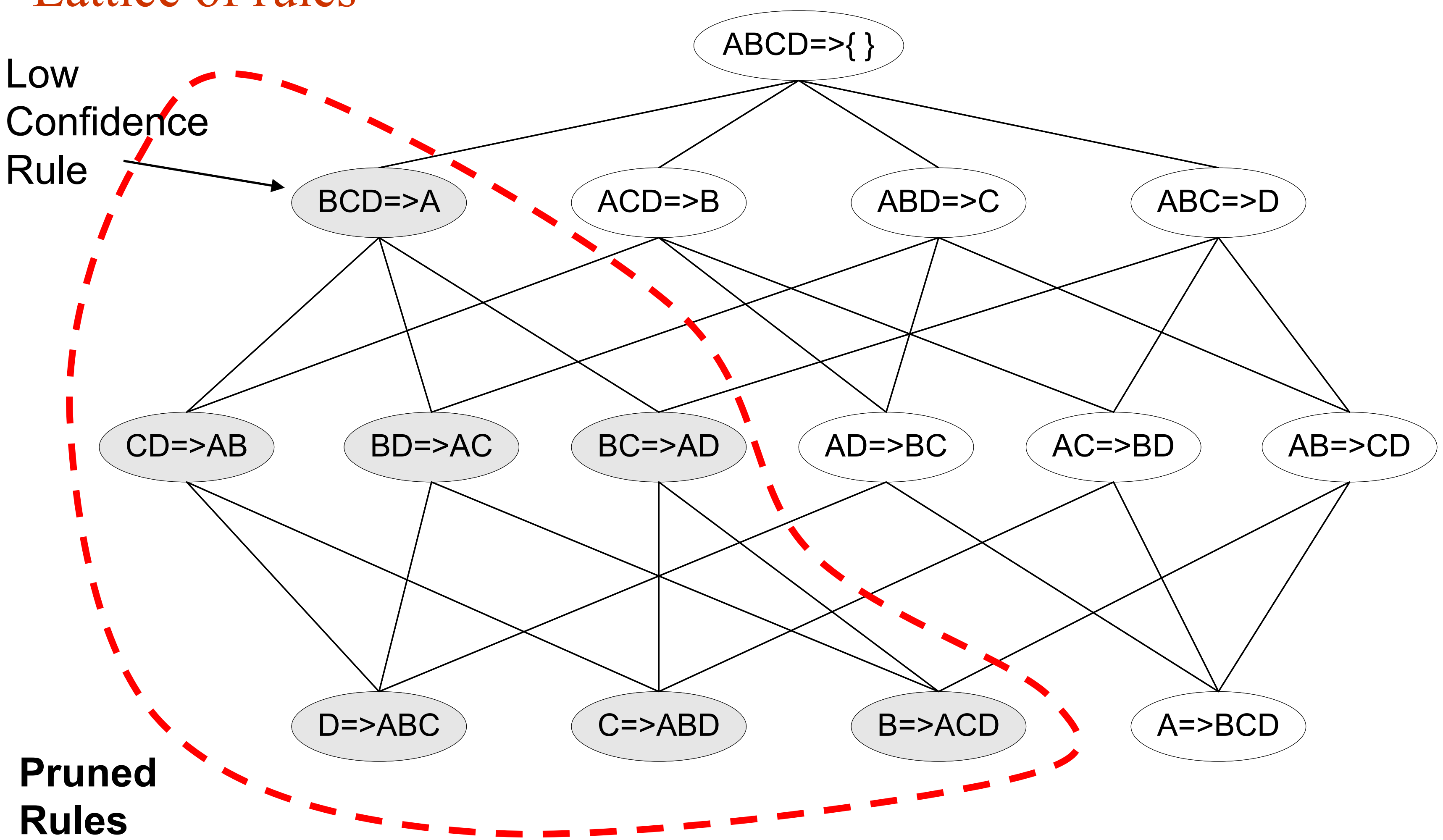
$$c(ABC \rightarrow D) = \frac{\sigma(\{A,B,C,D\})}{\sigma(\{A,B,C\})}$$

$$c(AB \rightarrow CD) = \frac{\sigma(\{A,B,C,D\})}{\sigma(\{A,B\})}$$

$$c(A \rightarrow BCD) = \frac{\sigma(\{A,B,C,D\})}{\sigma(\{A\})}$$

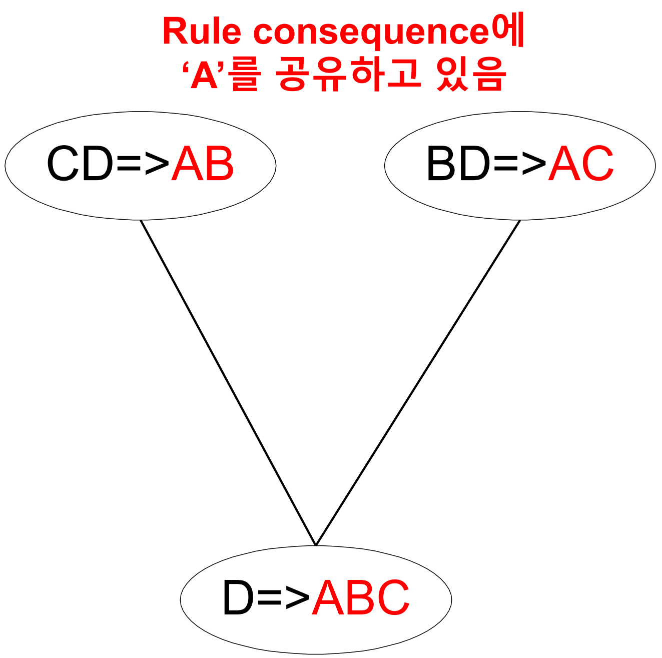
Rule Generation for Apriori Algorithm

Lattice of rules



Rule Generation for Apriori Algorithm

- Candidate rule is generated by merging two rules that share the same prefix in the rule consequent



- join(CD→AB,BD→AC) would produce the candidate rule D → ABC
- Prune rule D →ABC if the exists a subset (AD→BC) that does not have high confidence
- Essentially, we are doing Apriori on the RHS

Contents

Maximal itemset/closed itemset

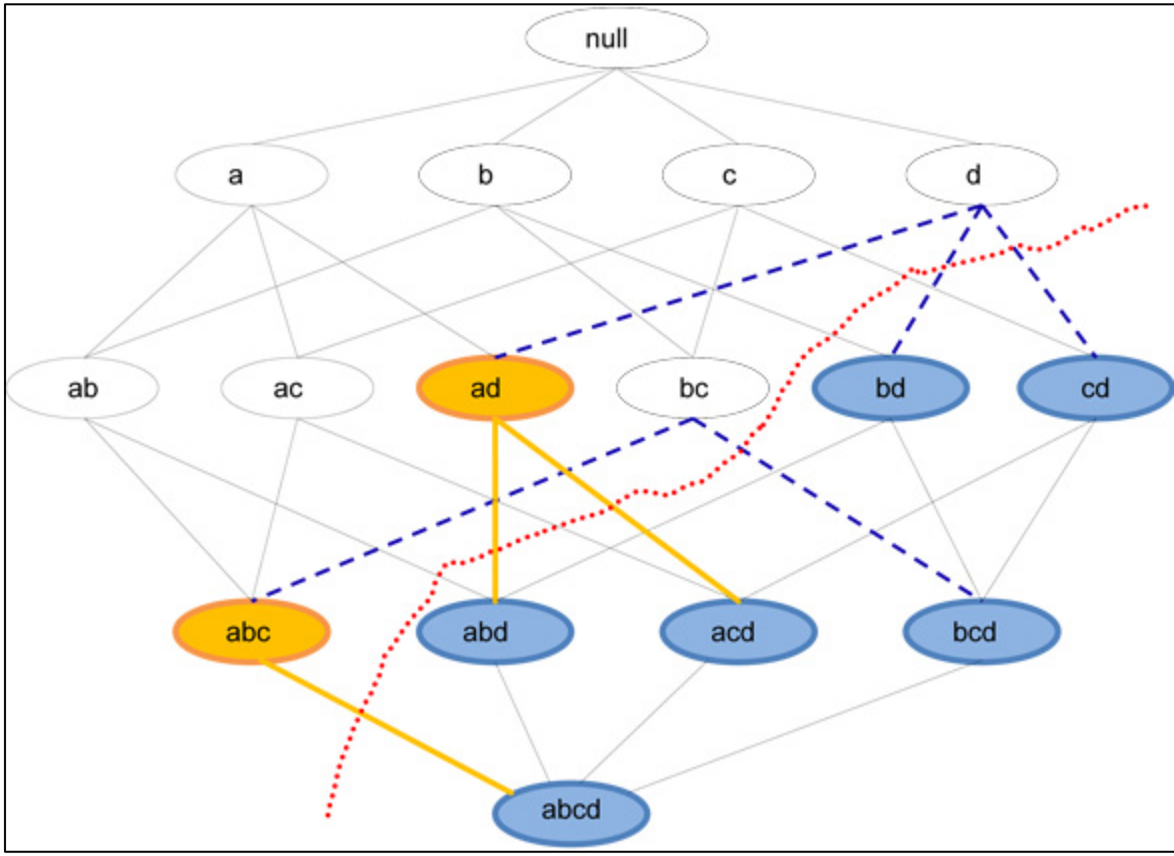
Maximal Frequent Itemset

- An itemset is maximal frequent(최대 빈발 항목집합) if none of its immediate supersets are frequent
- That is, this is a frequent itemset which is not contained in another frequent itemset.
- 찾는 방법
  - 먼저 Infrequent와 frequent itemset 사이의 border에 있는 frequent itemset 찾음
  - 모든 immediate supersets을 찾음
  - 만약 immediate superset 모두가 frequent 하지 않으면, 해당 itemset은 maximal frequent함
    - ✓ 예: Items: a, b, c, d, e
    - ✓ Frequent Itemset: {a, b, c}
    - ✓ {a, b, c, d}, {a, b, c, e}, {a, b, c, d, e} are not Frequent Itemset.
    - ✓ Maximal Frequent Itemsets: {a, b, c}
- Maximal frequent itemset은 아주 긴 빈발 항목집합을 만들 때 유용함
  - 일반적으로 짧은 항목집합은 규칙으로서 큰 의미가 없는 경우가 많음
  - 반면에, 긴 항목집합은 대개가 surprise한 연관규칙을 생성할 수 있음

Maximal Frequent Itemset

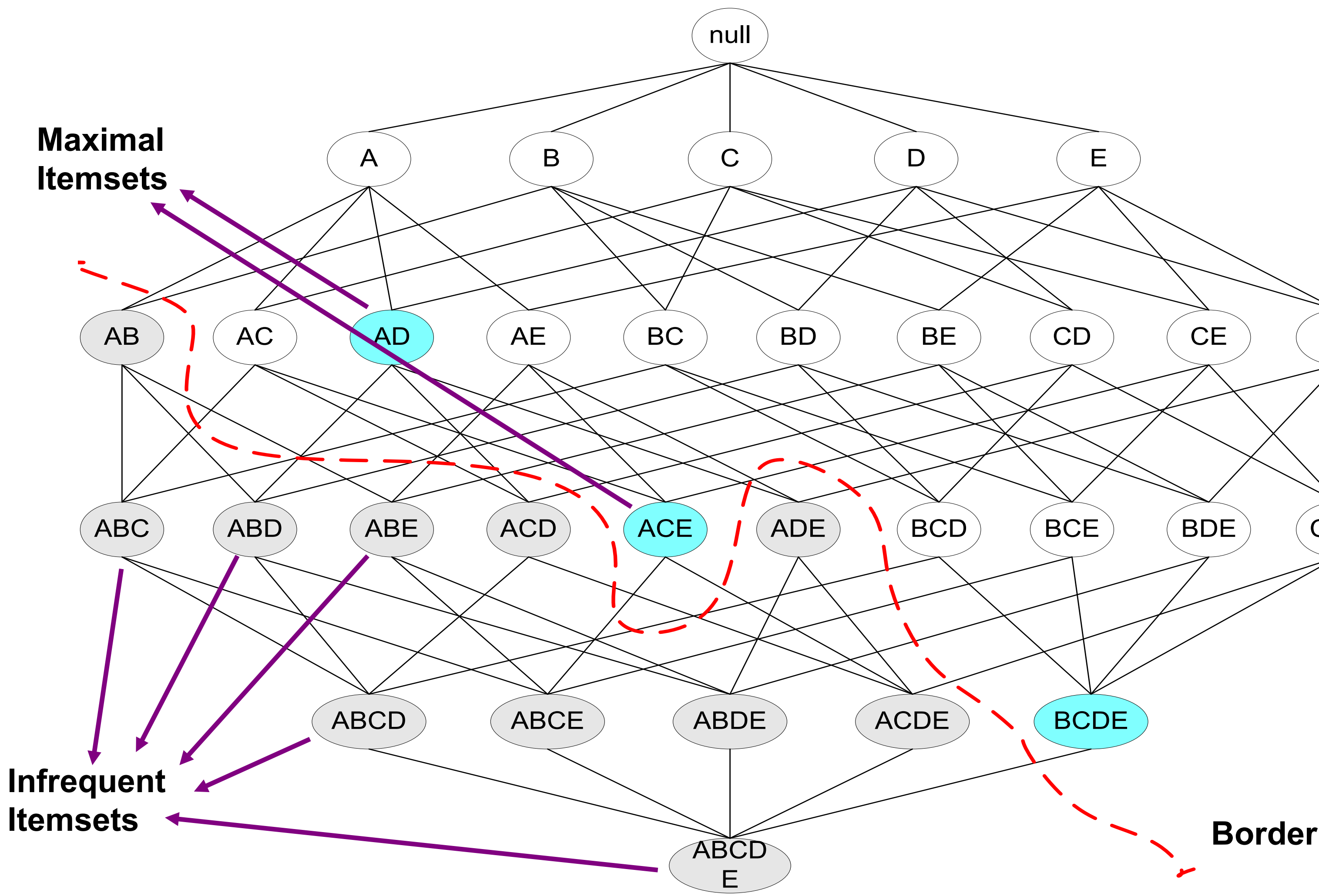
- **Maximal frequent itemset 찾는 예 1:**

- 먼저 Infrequent와 frequent itemset 사이의 border에 d, bc, ad, abc frequent itemset 이 있음을 확인함
- 이들 itemset의 immediate superset을 찾음
  - d의 superset으로 ad, bd, cd 가 있는데, ad는 frequent임. → **d는 maximal 이 되지 못함**
  - Bc는 abc와 bcd를 superset 으로 갖는데, abc가 frequent함 → bc는 maximal되지 못함
  - ad와 abc의 superset은 모두 infrequent 함 → ad와 abc는 모두 maximal임



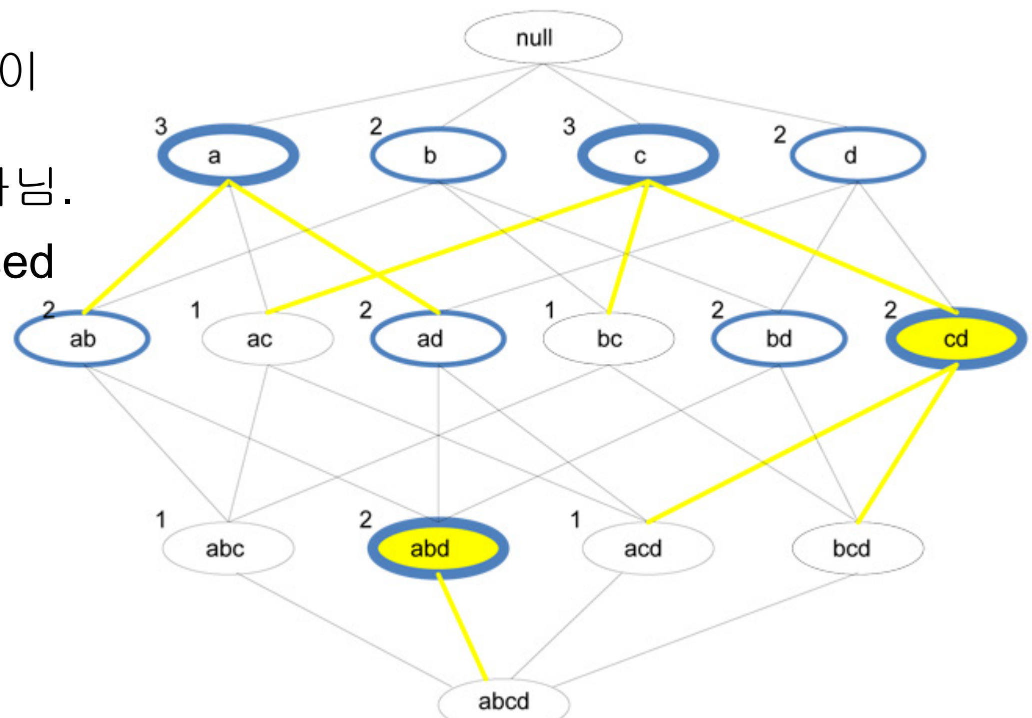
Maximal Frequent Itemset

- **Maximal frequent itemset 찾는 예 2:**



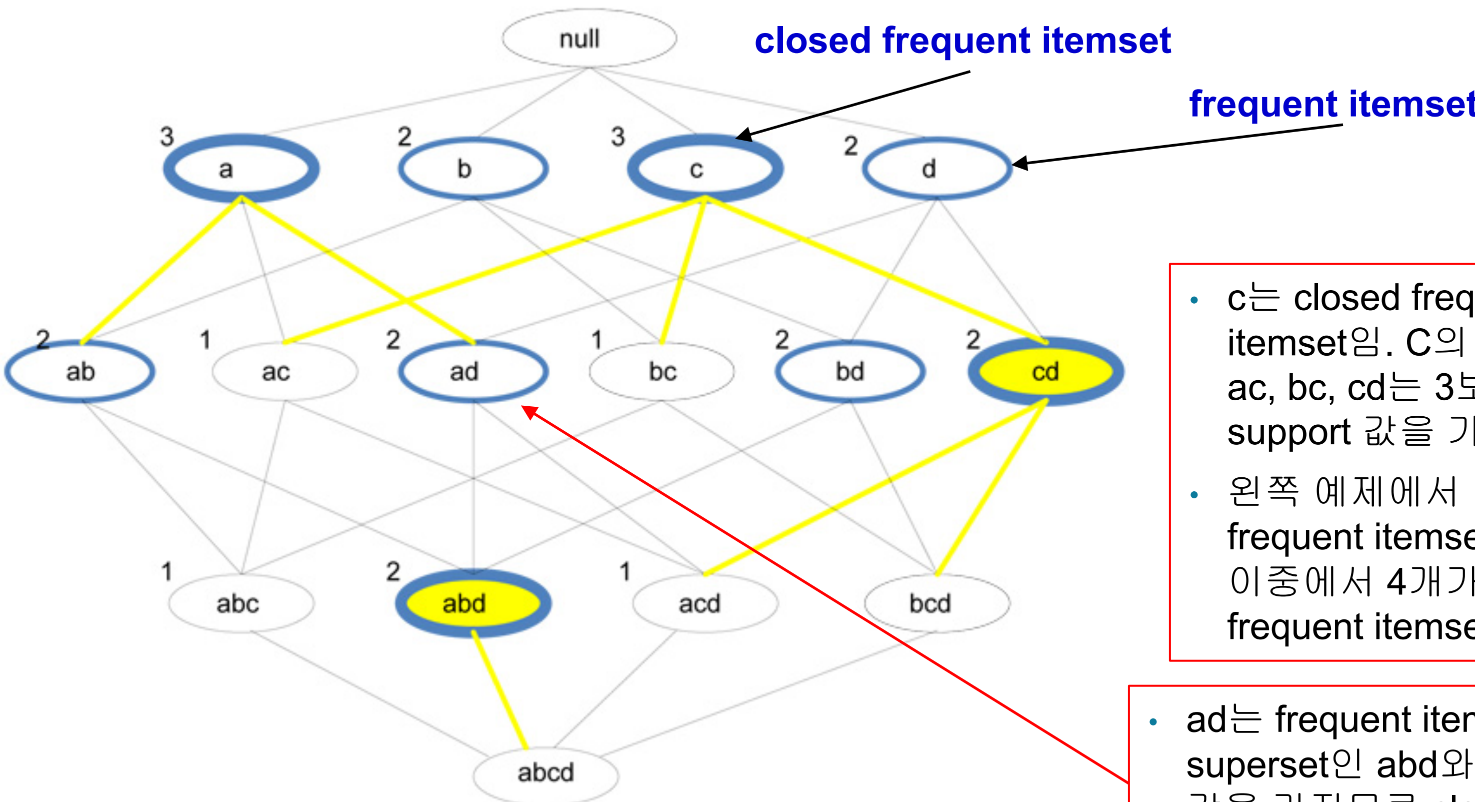
Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the original itemset
- That is, this is a set of items which is as large as it can possibly be without losing any transactions
- Closed itemset이 frequent 하면 closed frequent itemset임
- 예 closed frequent itemset 찾는 방법
  - 먼저 모든 frequent itemset을 찾음
  - 이후, 만약 해당 itemset의 superset이 original frequent itemset과 동일한 support를 가지면 closed itemset 아님.
  - 아니면 해당 original itemset은 closed itemset임



Closed Itemset

- 예 closed frequent itemset 찾는 방법



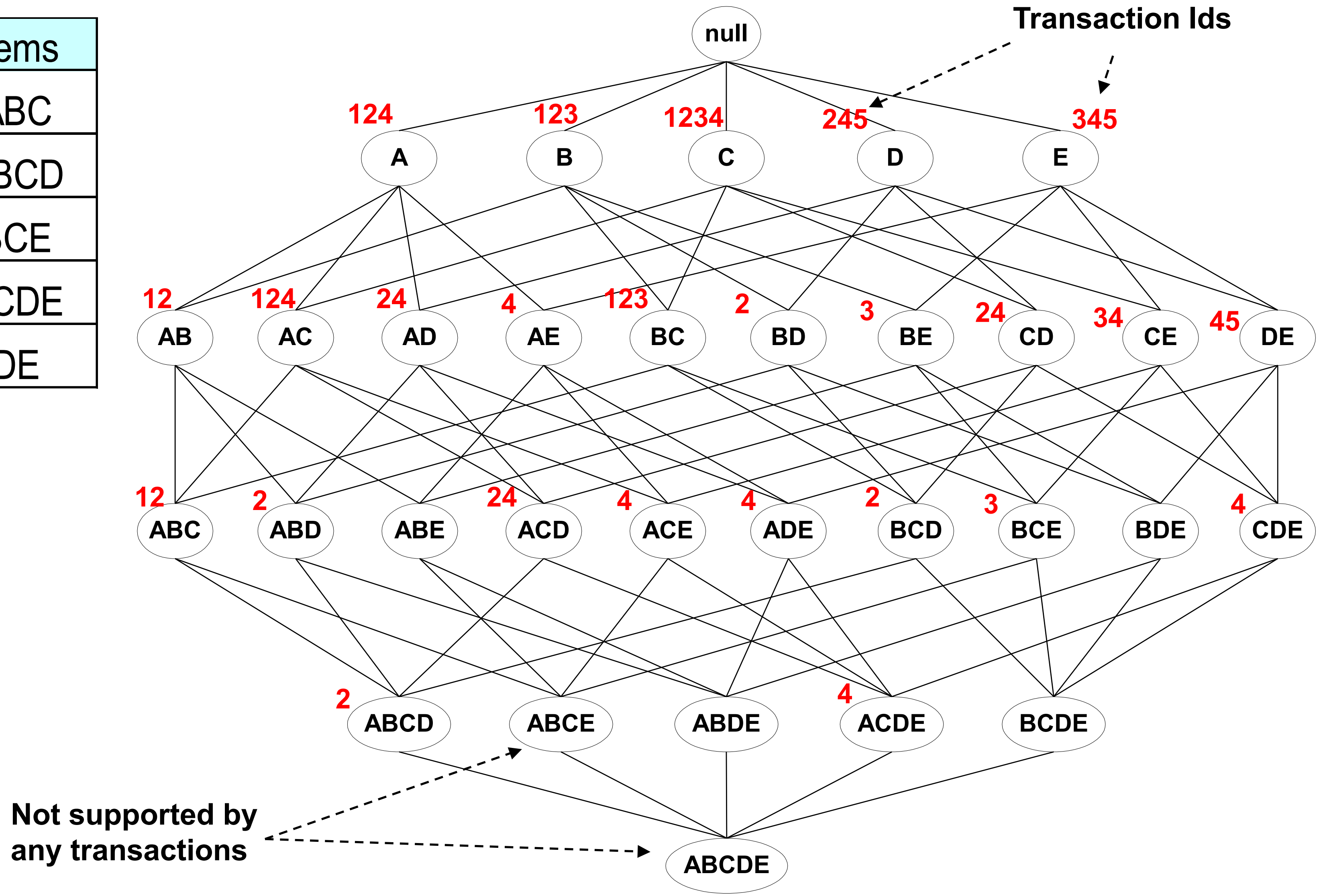
- c는 closed frequent itemset임. C의 superset인 ac, bc, cd는 3보다 작은 support 값을 가지므로
- 왼쪽 예제에서 총 9개의 frequent itemset 이 존재하며, 이중에서 4개가 closed frequent itemset 임

- ad는 frequent itemset 이지만 superset인 abd와 동일한 support 값을 가지므로 closed 아님

- 파란색 circle은 frequent itemset 임
- 파란색 두꺼운 circle 은 closed frequent itemset 임(closed는 superset과 동일한 support값을 가지지 않아야 함)
- 노란색 색칠된 circle은 maximal frequent itemset 임

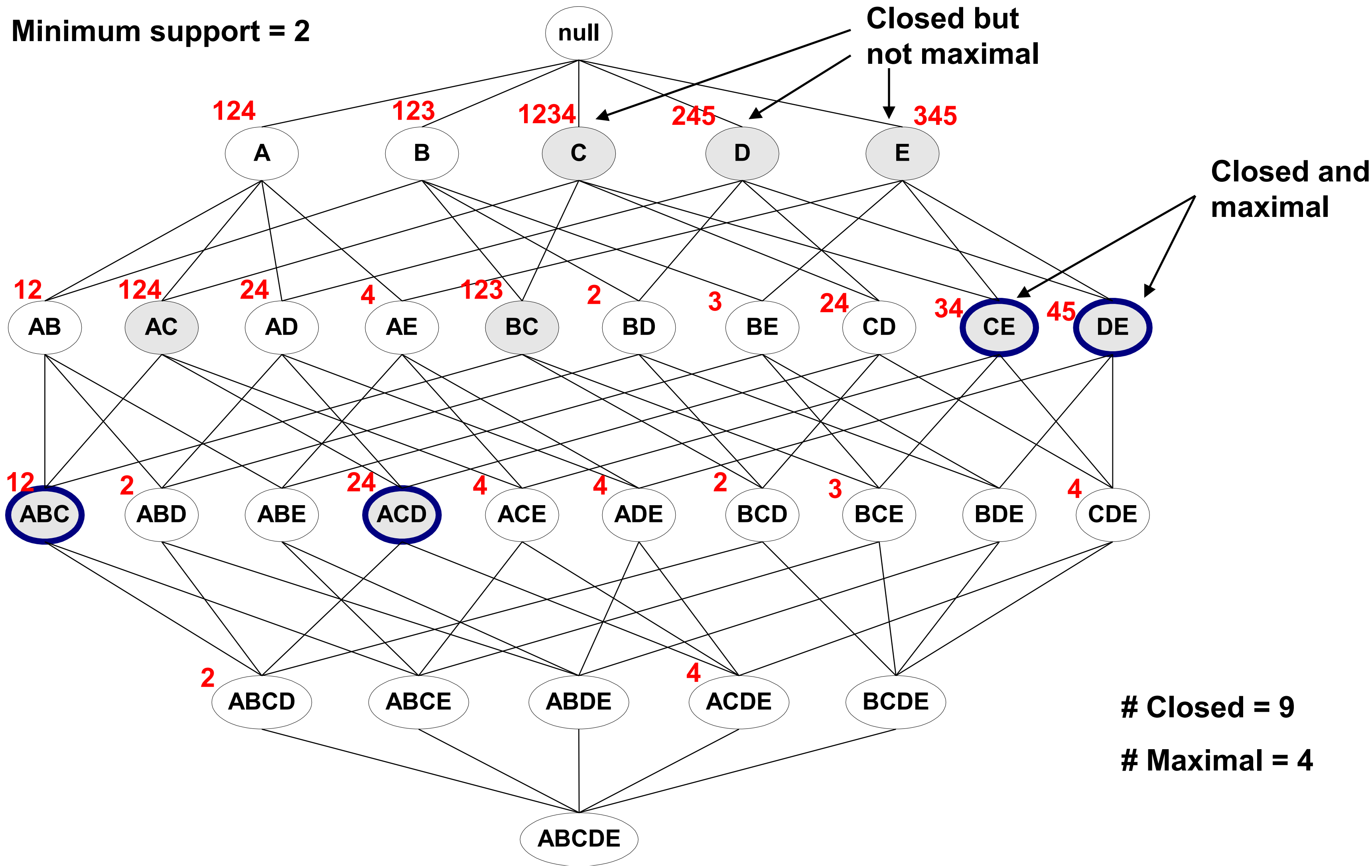
Maximal vs Closed Itemsets

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



Maximal vs Closed Frequent Itemsets

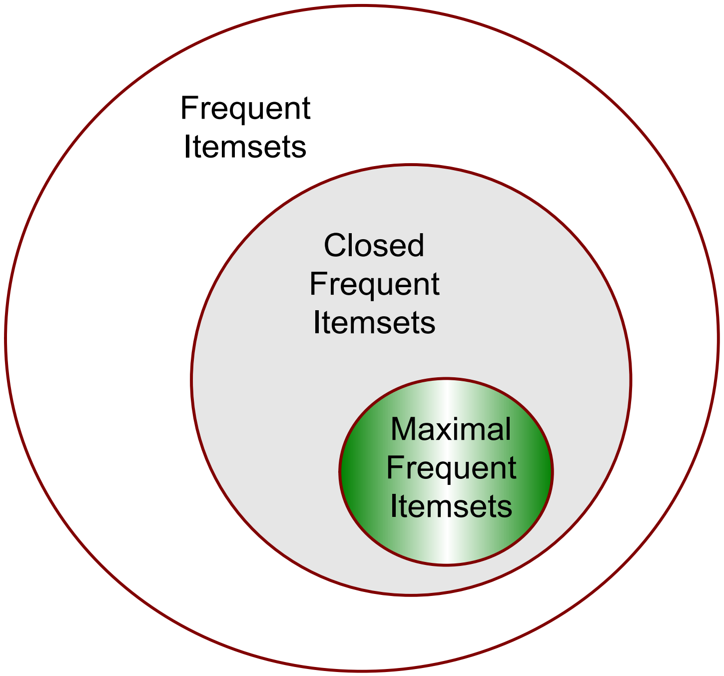
Minimum support = 2



Closed = 9

Maximal = 4

Maximal vs Closed Itemsets



- it is important to point out the relationship between frequent itemsets, closed frequent itemsets and maximal frequent itemsets.
- Closed and maximal frequent itemsets are subsets of frequent itemsets but maximal frequent itemsets are a more compact representation because it is a subset of closed frequent itemsets.
- The diagram to the right shows the relationship between these three types of itemsets.

Contents

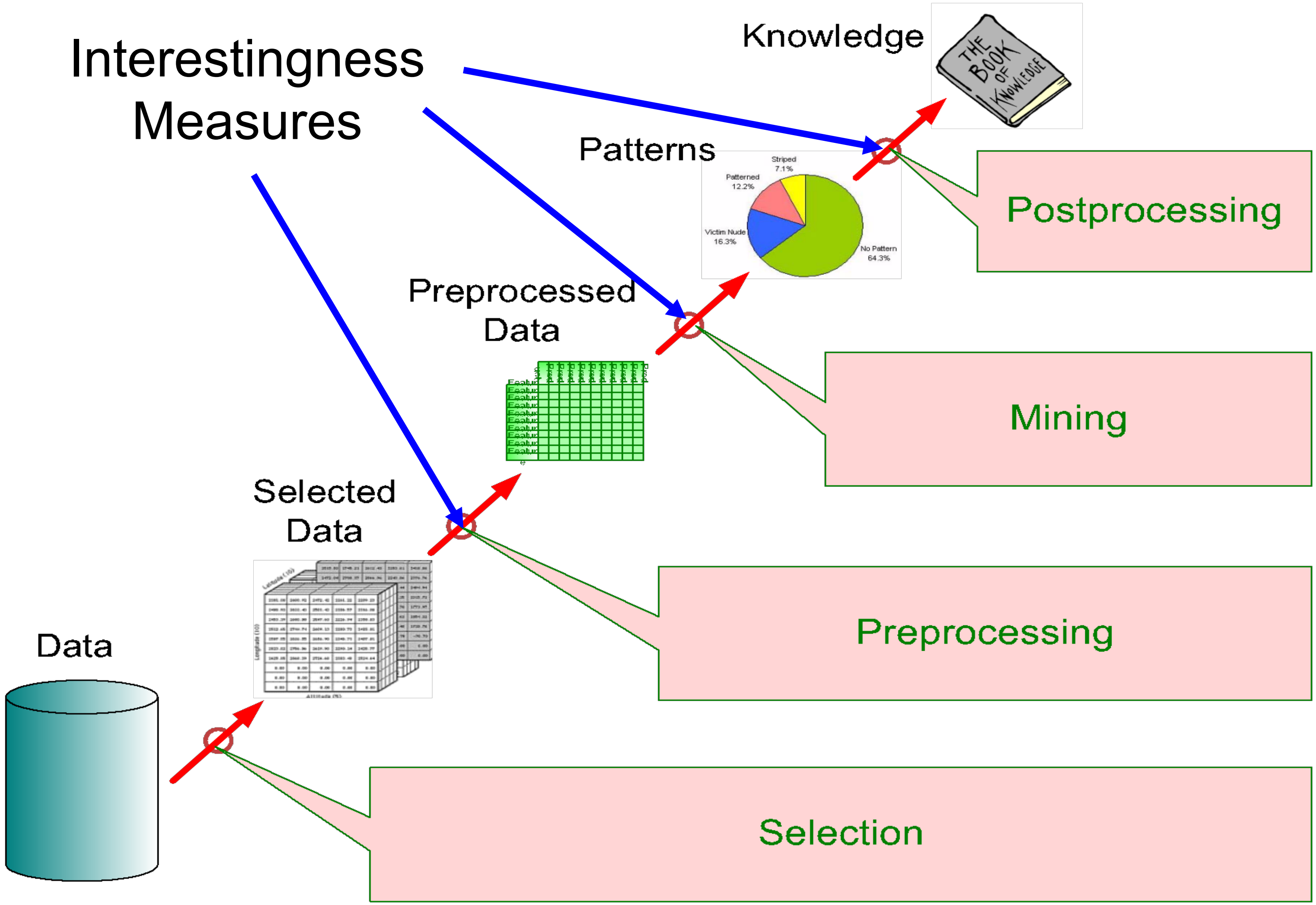
연관 패턴의 평가

연관 규칙 평가(Pattern Evaluation)

- 연관 규칙 생성 알고리즘은 너무 많은 연관 규칙을 생성하는 경향이 있음
  - 생성된 많은 규칙은 유용하지 않음(uninteresting or redundant)
  - 예를 들어, {A, B, C} → {D}와 {A, B} → {D}가 동일한 지지도/신뢰도를 갖는다면, 이들 두 규칙은 redundant 함

- Interestingness measures(유용성 척도)는 유도된 규칙을 제거하거나 순위를 매기는데(prune or rank) 사용됨
- 지지도와 신뢰도(support & confidence)도 유용성 척도에 속함

유용성 척도 활용 시점


Interestingness Measures
Knowledge
THE BOOK OF KNOWLEDGE
Patterns
Striped 7.1%
Patterned 12.2%
Victim Nude 16.3%
No Pattern 64.3%
Postprocessing
Preprocessed Data
Mining
Selected Data
Preprocessing
Data
Selection

Computing Interestingness Measure

- 주어진 규칙 X → Y에 대해, 다음 분할표(contingency table)를 사용하여 다양한 유용성 척도를 계산할 수 있다

Contingency table for X → Y

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

f_{ij}는 support 즉, 빈도수 count 값을 의미함

f_{1+}는 결국 X에 대한 지지도 count를 의미함

X항목이 transaction에 없는 경우 → $\bar{X}$

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

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

신뢰도의 단점(Drawback of Confidence)

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	65	15	80
	80	20	100

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Association Rule: Tea $\rightarrow$ Coffee

- Support(Tea → Coffee) = 15/100 = 15%
- Confidence(Tea → Coffee) = s(Tea U Coffee)/s(Tea) = 15/20= 75%

- 위 신뢰도를 보고 차를 마시는 사람은 coffee를 마시는 경향이 있다고 볼지도 모름
- 하지만, 위 데이터를 보면 차를 마시든 마시지 않든 간에, coffee 를 마시는 사람의 비율은 원래 80%였음
- 즉, 규칙 Tea → Coffee를 통해, 어떤 사람이 차를 마신다는 정보를 통해 커피를 마시는 사람의 정보를 아는 것은 (75%라는 큰 신뢰도 값을 가짐에도) 큰 의미가 없음.

Statistical Independence

- Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)

  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$

  - $P(S \wedge B) = P(S) \times P(B)$ => Statistical independence
  - $P(S \wedge B) > P(S) \times P(B)$ => Positively correlated
  - $P(S \wedge B) < P(S) \times P(B)$ => Negatively correlated

Statistical-based Measures

- Measures that take into account statistical dependence

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

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

$$Note: \frac{P(Y \mid X)}{P(Y)} = \frac{\frac{P(X,Y)}{P(X)}}{P(Y)} = \frac{P(X,Y)}{P(X)P(Y)}$$

Lift와 Interest는 equivalent함

연관 규칙 평가(Pattern Evaluation)

- **Lift** of an association rule: X → Y, lift = P(Y/X)/P(Y))
  - If Lift > 1, then X and Y appear more often together than expected
    - this means that the occurrence of X has a positive effect on the occurrence of Y or that X is positively correlated with Y.
  - If Lift < 1 then, X and Y appear less often together than expected
    - this means that the occurrence of X has a negative effect on the occurrence of Y or that X is negatively correlated with Y
  - If Lift = 1, then X and Y are independent.
    - this means that the occurrence of X has almost no effect on the occurrence of Y

Example: Lift/Interest

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

Association Rule: Tea $\rightarrow$ Coffee

Confidence= P(Coffee|Tea) = 0.75

but P(Coffee) = 0.9

$\Rightarrow$ Lift = P(Coffee/Tea)/P(Coffee) = 0.75/0.9= **0.8333 (< 1, therefore, the Lift is suggesting a slight negative correlation b/w tea drinkers and coffee drinkers)**

There are lots of measures proposed in the literature

Some measures are good for certain applications, but not for others

What criteria should we use to determine whether a measure is good or bad?

What about Apriori-style support based pruning? How does it affect these measures?

#	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)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(A)})\Big)$
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)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
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)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
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)	$\sqrt{P(A,B)}\max(P(B\|A) - P(B), P(A\|B) - P(A))$