

Minería de datos aplicada

Reglas de
Asociación

Minería de reglas de asociación (Patrones)

La gaseosa es comprada junto a las ¿papitas? ¿choclos?

Los productos de limpieza son comprados conjuntamente con cuales

¿Como la demografía de los vecinos afecta lo que compramos?

¿Si cambia el precio de algunos productos como cambia la cantidad de otros?



Minería de reglas de asociación (Patrones)

Dado un conjunto de transacciones, encontrar las reglas que predicen la aparición de un artículo basado en las ocurrencias de otros elementos en la transacción

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ejemplos de reglas de asociación

$\{\text{Diaper}\} \rightarrow \{\text{Beer}\},$
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\},$
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\},$

!La implicación significa co-ocurrencia no causalidad!

Minería de reglas de asociación (Patrones)

Conjunto de items (itemset)

- Una colección de 1 o más items
 - Ejemplo: {Milk, Bread, Diaper}
- k-itemset
 - Un conjunto de items que contiene k items

Suporte absoluto - Support count (σ)

- Frecuencia de las co-ocurrencias de un conjunto de items
- E.g. $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$

Soporte relativo -Support

- La fracción de las transacciones que contiene un conjunto de items
- E.g. $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$

Conjunto de items frecuente

- Un conjunto de items que son mayores o iguales a un umbral (*minsup*)

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Minería de reglas de asociación (Patrones)

Regla de asociación

- Aun expresión de la implicación de la forma $X \rightarrow Y$, donde X y Y son conjunto de ítems
- Ejemplo:
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

Métricas de evaluación de reglas

- Soporte (s)
 - Fracción de las transacciones que contienen a X y a Y
- Confianza (c)
 - Medide la frecuencia con la que el conjunto de ítems de Y aparecen en las transacciones que tiene X

Ejemplo:

$$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$$

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

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

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

Dado un conjunto de transacciones T, El objetivo de la minería de reglas de asociaciones es encontrar todas las reglas que:

- Soporte $\geq \text{minsup}$
- Confianza $\geq \text{minconf}$

Enfoque de fuerza bruta:

- Listar todas las posibles reglas de asociación
 - Computar el soporte y la confianza de cada regla
 - Podar reglas que fallen en los umbrales minsup y minconf
- ⇒ **!Computacionalmente muy costoso!**

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ejemplos de reglas:

$\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$ ($s=0.4$, $c=0.67$)
 $\{\text{Milk, Beer}\} \rightarrow \{\text{Diaper}\}$ ($s=0.4$, $c=1.0$)
 $\{\text{Diaper, Beer}\} \rightarrow \{\text{Milk}\}$ ($s=0.4$, $c=0.67$)
 $\{\text{Beer}\} \rightarrow \{\text{Milk, Diaper}\}$ ($s=0.4$, $c=0.67$)
 $\{\text{Diaper}\} \rightarrow \{\text{Milk, Beer}\}$ ($s=0.4$, $c=0.5$)
 $\{\text{Milk}\} \rightarrow \{\text{Diaper, Beer}\}$ ($s=0.4$, $c=0.5$)

Observaciones:

- Todas las reglas son binarias del mismo conjunto de items:
 $\{\text{Milk, Diaper, Beer}\}$
- Las reglas originarias del mismo conjunto de items tienen el mismo soporte pero diferente confianza
- Así, podemos desvincular los requerimiento de soporte y confianza

Minería de reglas de asociación

Aproximación por dos pasos:

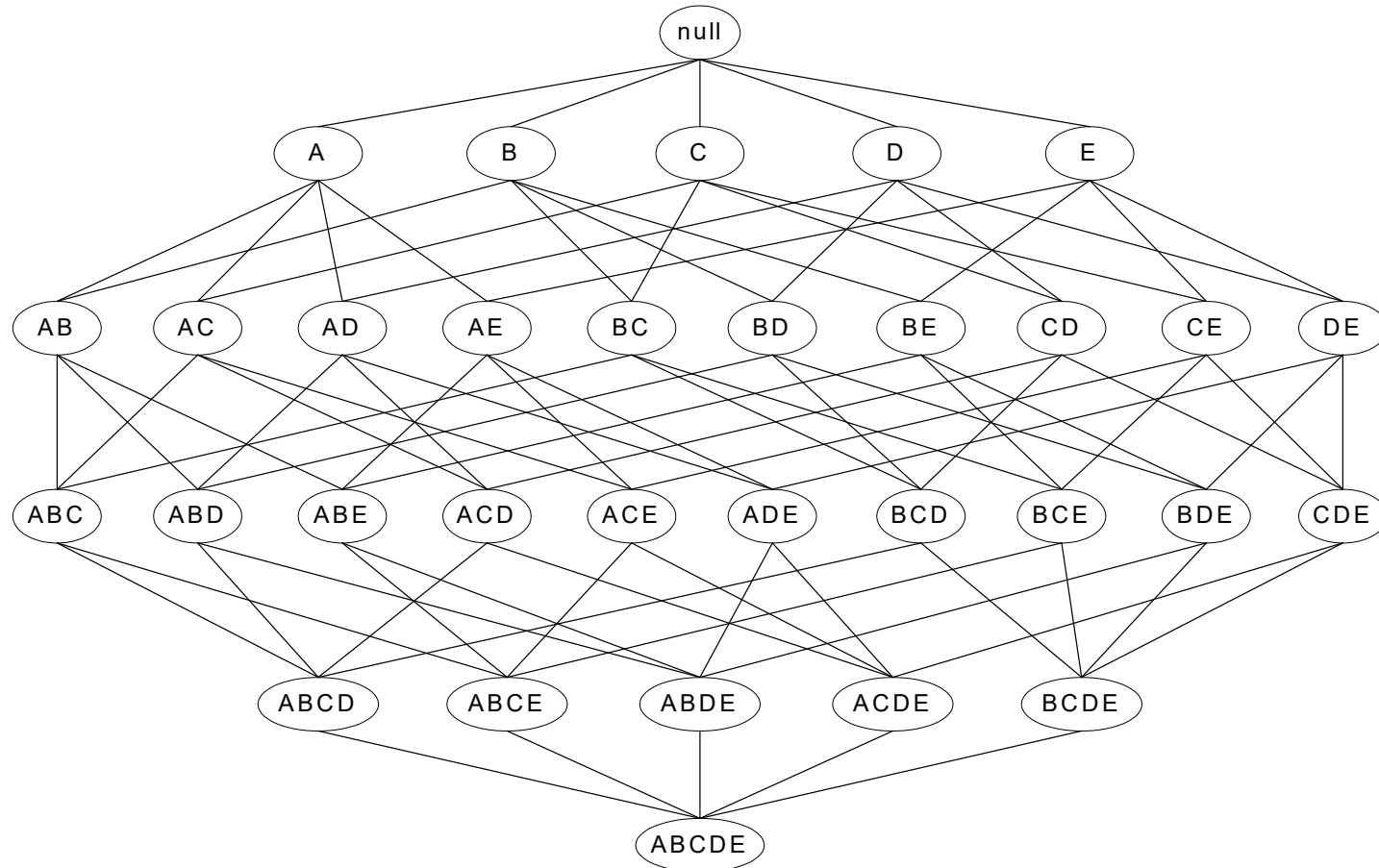
1. Generación de conjunto de ítems frecuentes

- Generar todos los itemsets cuyo Soporte \geq minsup

2. Generación de reglas

- Generar reglas de alta confianza para cada conjunto frecuente

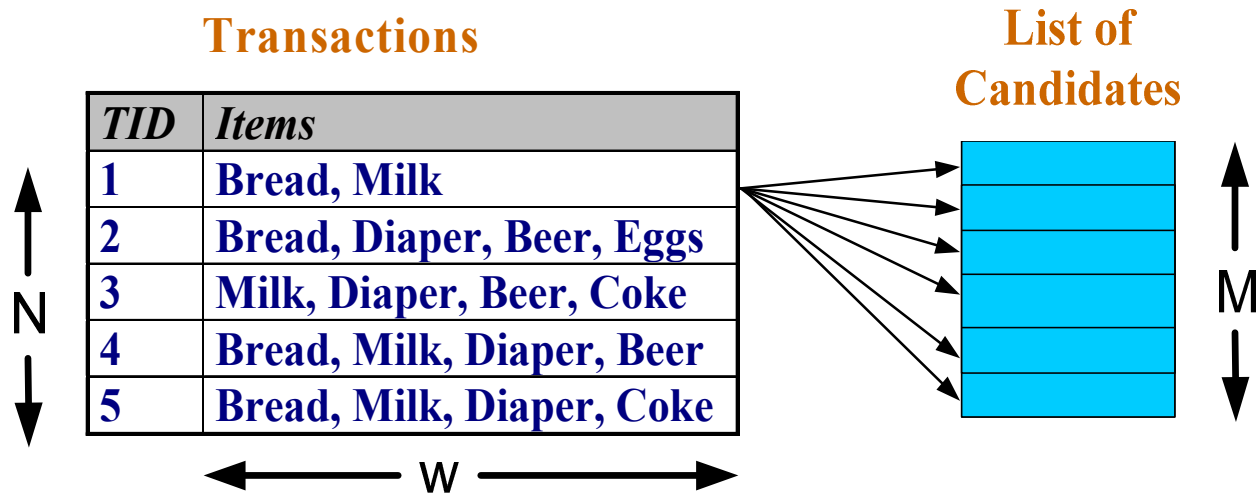
Generación del conjunto de ítems frecuentes



Generación del conjunto de ítems frecuentes

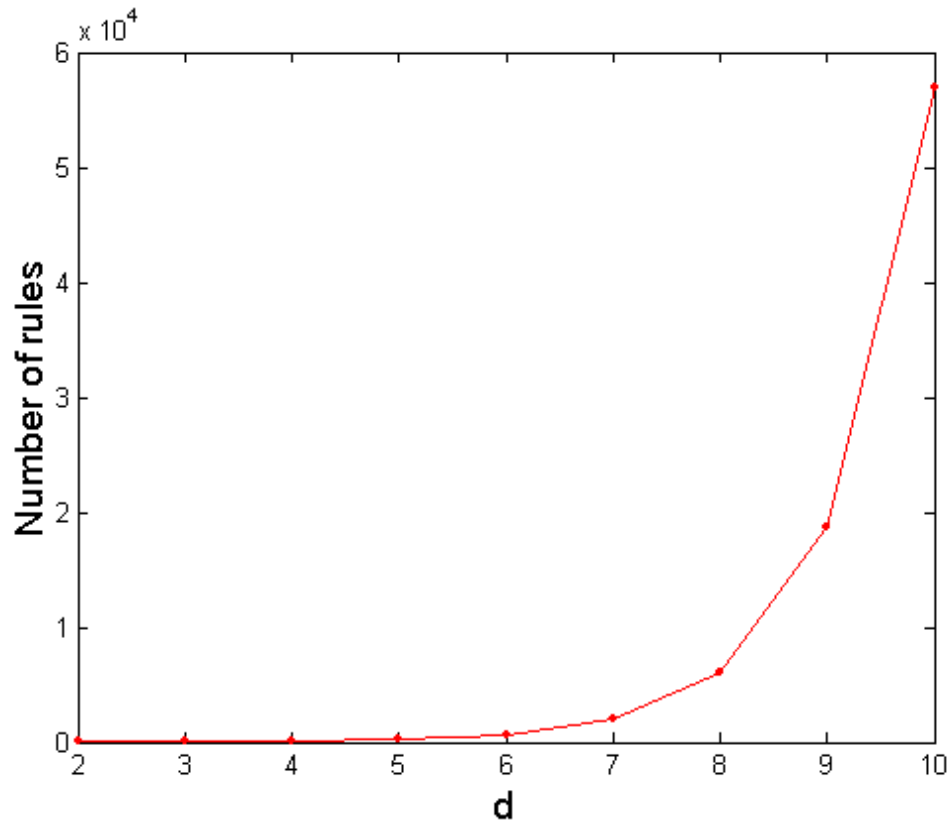
Enfoque de fuerza bruta:

- Cada itemset en el entramado es un **candidato** a itemset frecuente
- Se cuente el soporte de cada candidato escaneando la base de datos



- Emparejar cada transaction contra cada candidato
- Complejidad $\sim O(NMw) \Rightarrow$ **Costosa desde $M = 2^d$!!!**

Generación del conjunto de ítems frecuentes



Dado d únicos items:

- # tota de reglas de items = 2^d
- # tota de reglas de asociación:

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

Sí $d=6$, $R = 602$ rules

Generación del conjunto de ítems frecuentes

Reducir el **numero de candidatos**(M)

Búsqueda completa: $M=2^d$

Usar técnicas de poda para reducir M

Reducir el **numero de transacciones** (N)

Usado por DHP y algoritmos vertical-based mining

Reducir el **numero de comparisons** (NM)

Usar estructuras de datos eficientes para guardar las transacciones

No necesita comparar cada transacción contra las otras

Generación del conjunto de ítems frecuentes

Principio Apriori:

Si un itemset es frecuente , entonces todos sus subconjuntos deberían ser frecuentes

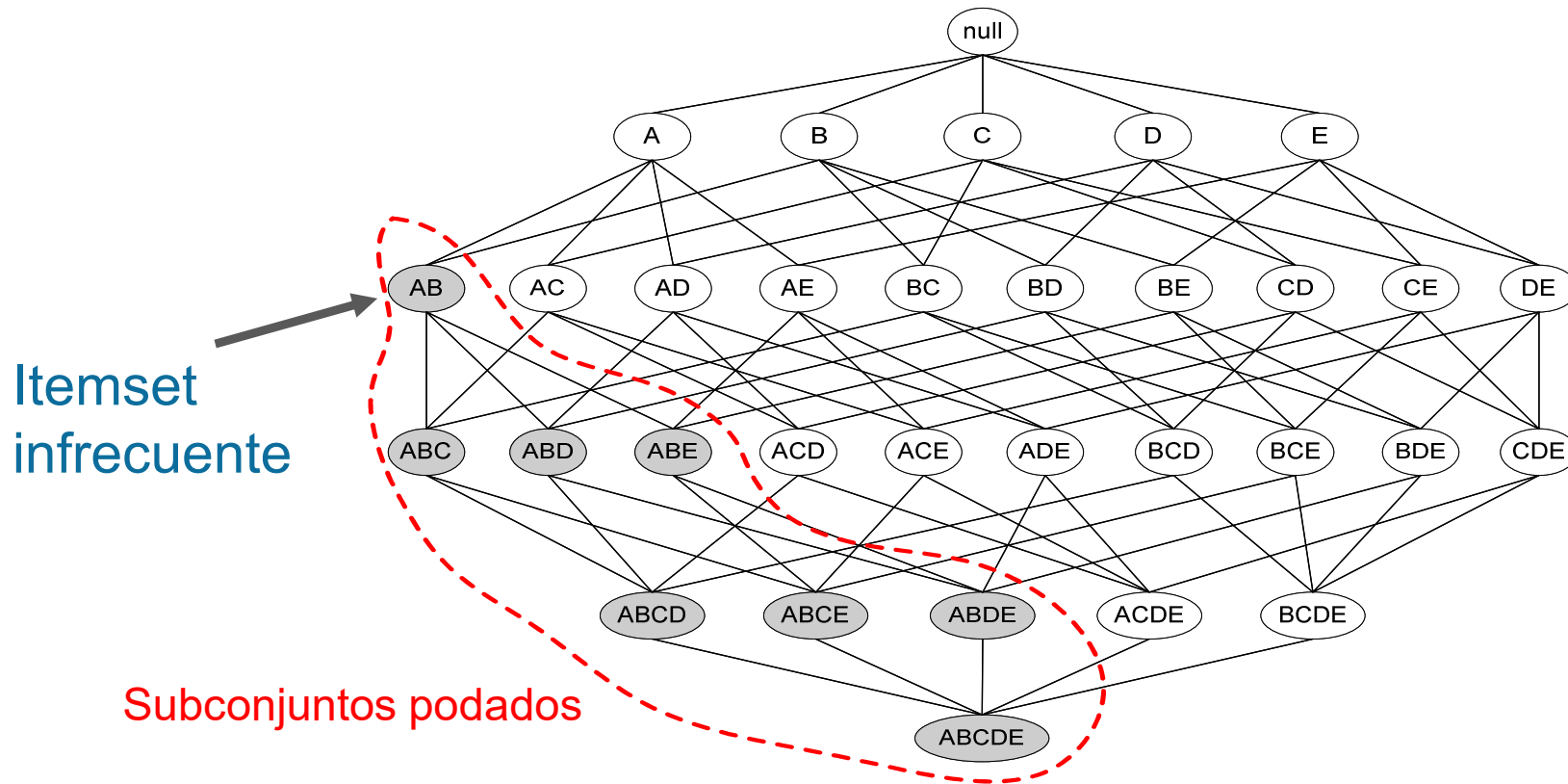
EL principio a priori se sostiene gracias a la siguiente propiedad de la medida de soporte:

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

El soporte de un itemset nunca excede el soporte de sus subconjuntos:

Esto es conocido como la propiedad **anti monotona** del soporte

Generación del conjunto de ítems frecuentes



Generación del conjunto de ítems frecuentes

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

Par (2-itemsets)

(No necesita generar candidatos que involucren a Coke o Eggs)



Tripleta (3-itemsets)

Item set	Count
{Bread,Milk,Diaper}	3



Soporte minimo = 3

Si cada subconjunto es considerado,
 ${}^6C_1 + {}^6C_2 + {}^6C_3 = 41$
 Basado en e soporte podamos,
 $6 + 6 + 1 = 13$

Método:

Sea $k = 1$

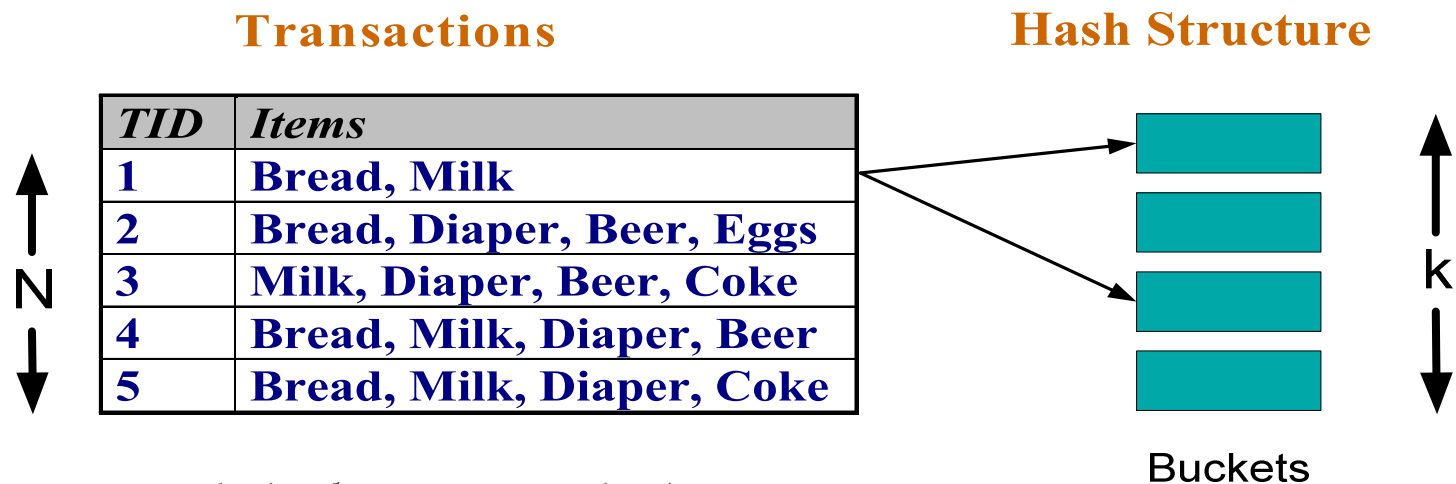
Generar conjuntos de elementos frecuentes de longitud 1

Repita hasta que se identificaron nuevos conjuntos de elementos frecuentes

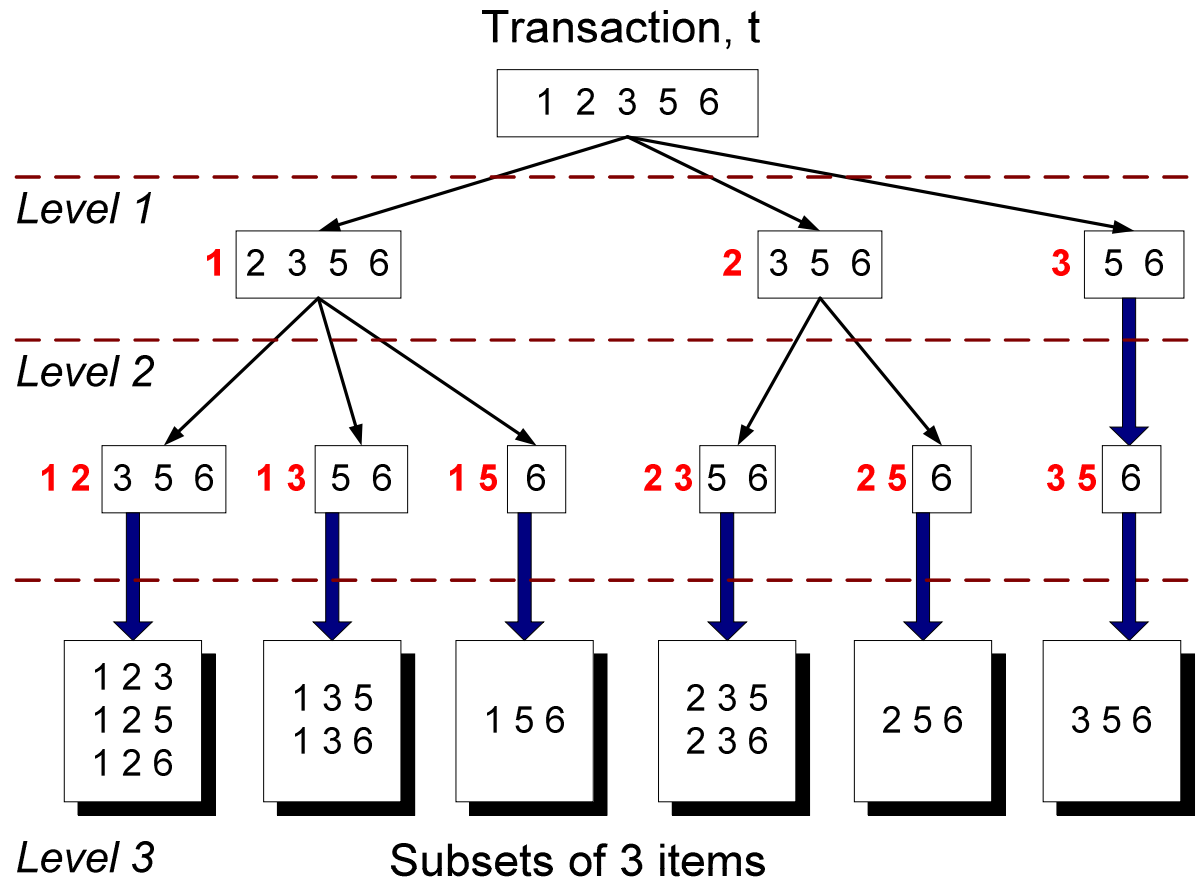
- Generar longitud $(k + 1)$ conjuntos de elementos candidatos de longitud k conjuntos de elementos frecuentes
- Poda los conjuntos de elementos candidatos que contienen subconjuntos de longitud k que son poco frecuentes
- Contar con el apoyo de cada candidato mediante el escaneo de la DB
- Eliminar a los candidatos que son poco frecuentes, dejando sólo las que son frecuentes

Recuento candidato:

- Escanear la base de datos de transacciones para determinar el soporte de cada conjunto de elementos candidato
- Para reducir el número de comparaciones, almacenar los candidatos en una estructura de hash
 - En lugar de hacer coincidir cada transacción en contra de cada candidato, compararlo con los candidatos que figuran en los recipientes de hash



Dado una transacción t ,
Cuales son los posibles
subconjuntos de tamaño
3?

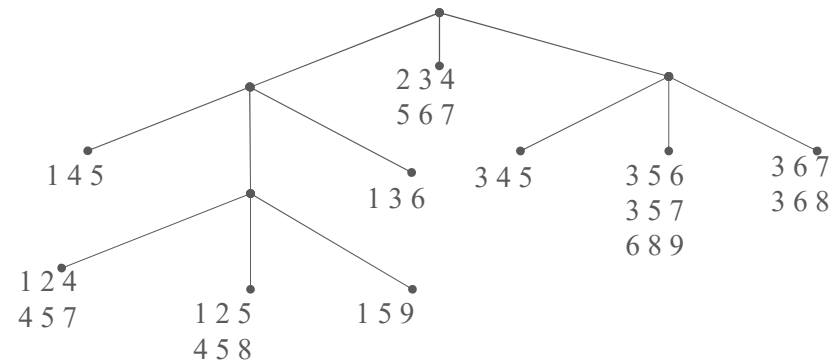
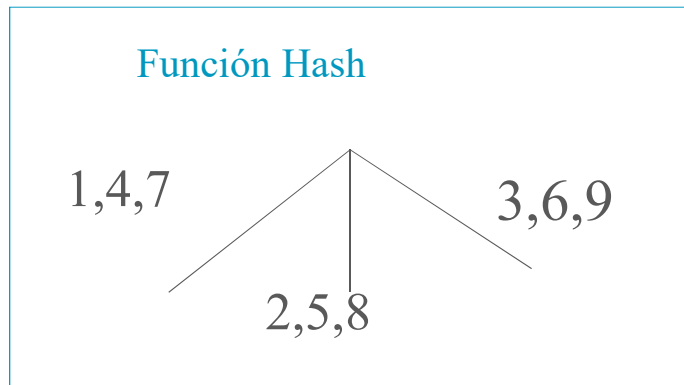


Suponga que tiene 15 itemsets candidatos de longitud 3:

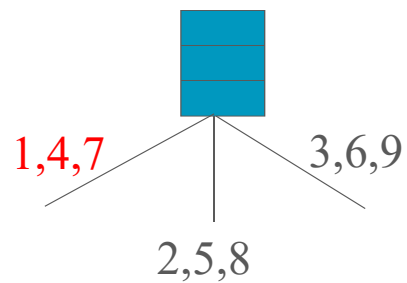
{1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}

Se necesita:

- Función Hash
- Máximo tamaño de la hoja : numero máximo de itemsets guardados en una hoja nodo(Si el numero de itemset candidatos excede el máximo tamaño de hoja, dividir este nodo)

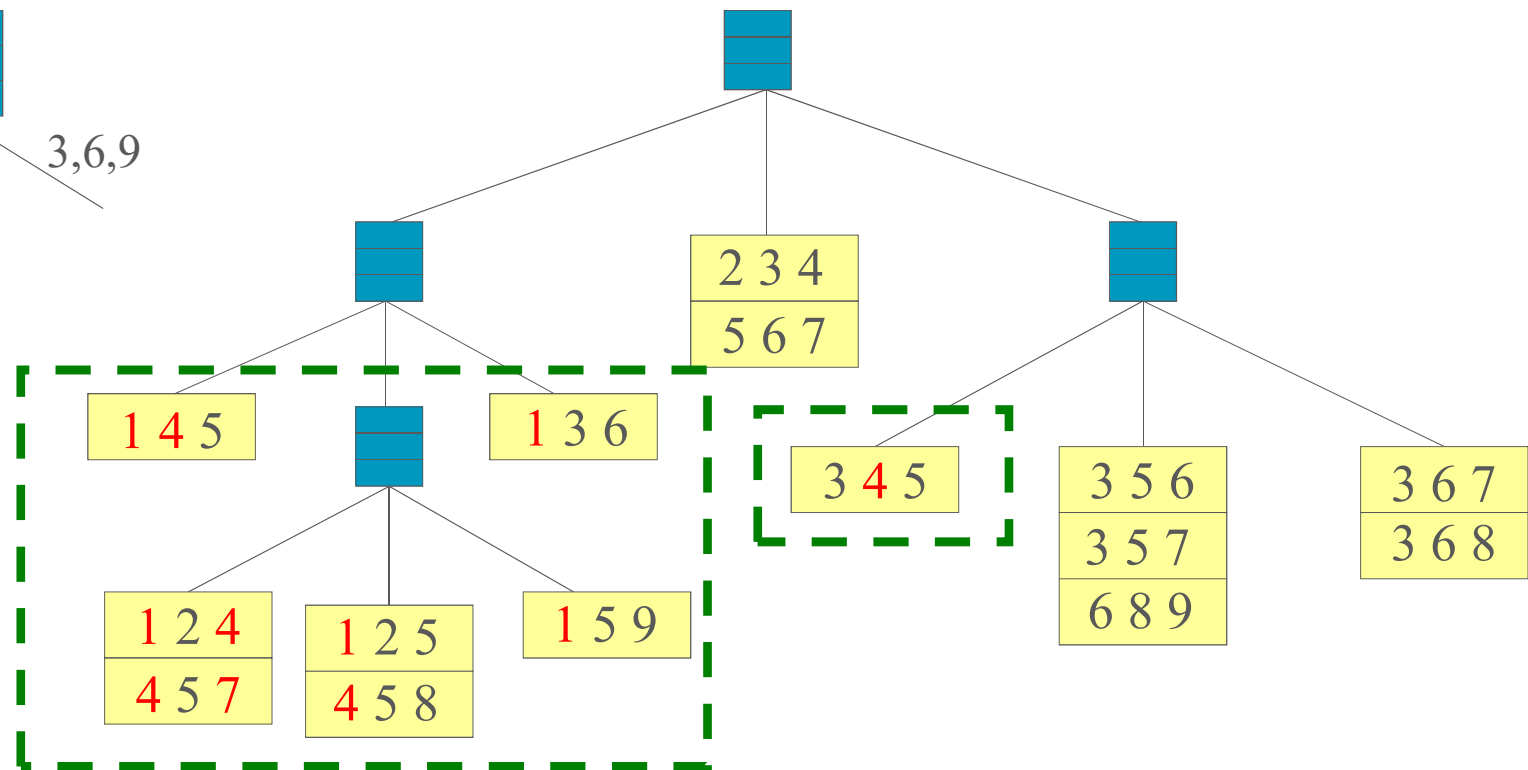


Función Hash

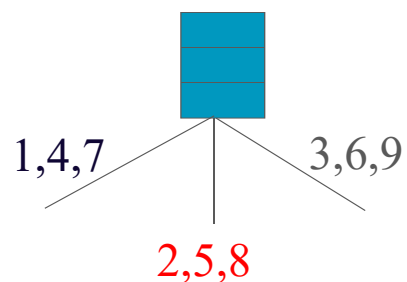


Candidatos a árboles hash

Hash sobre
1, 4 o 7

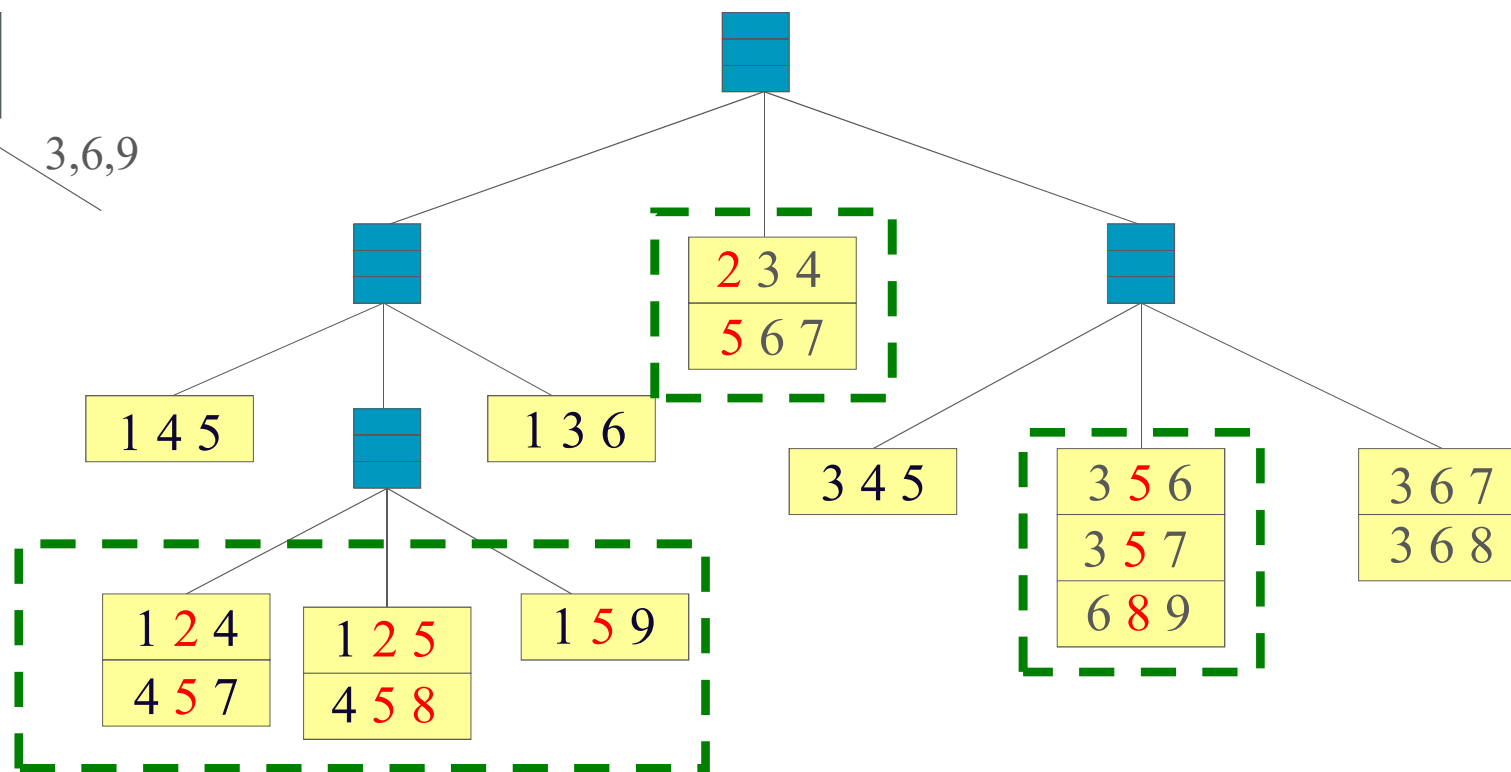


Función Hash

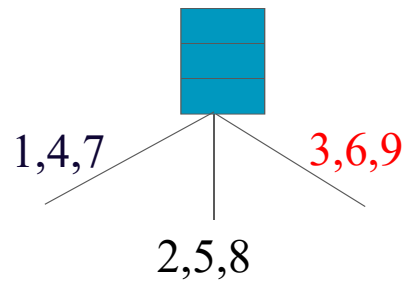


Hash sobre
2, 5 o 8

Candidatos a árboles hash

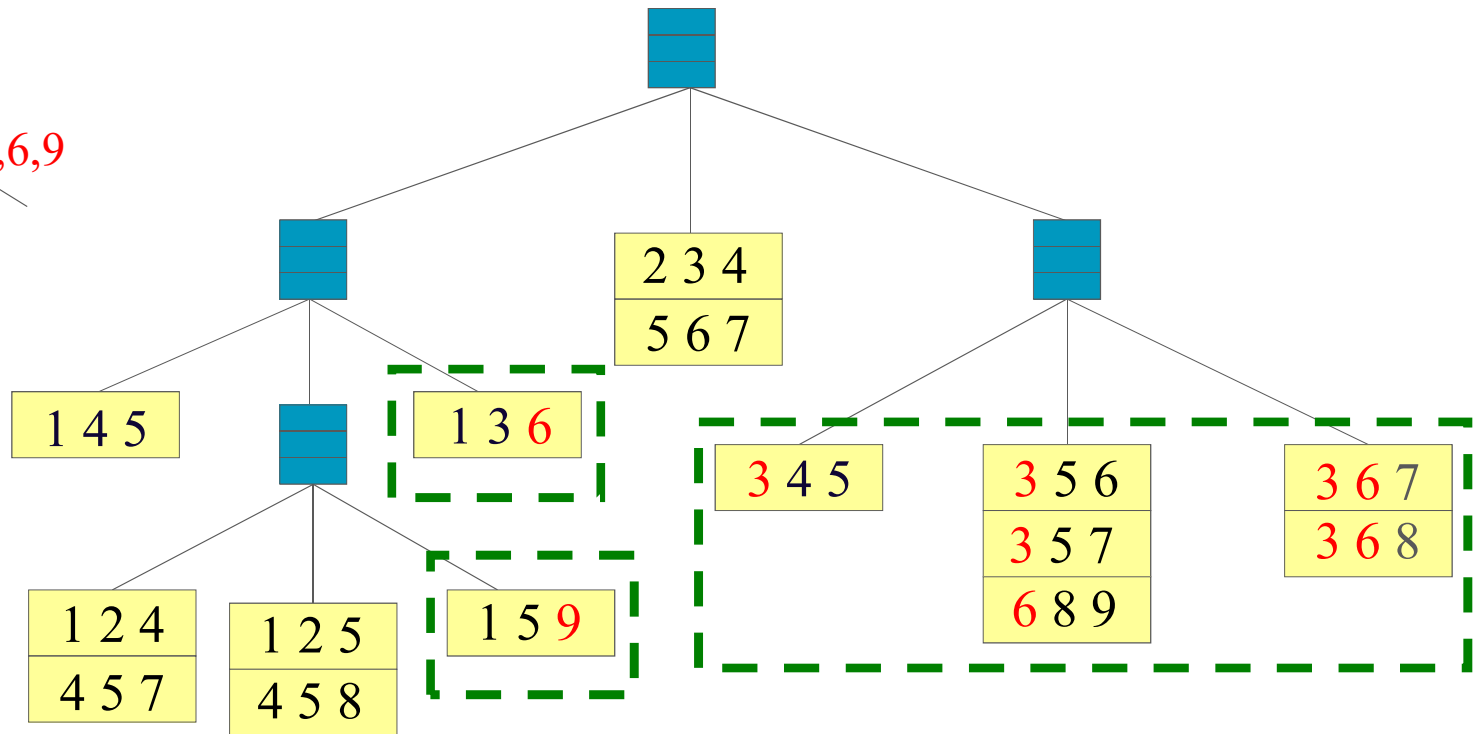


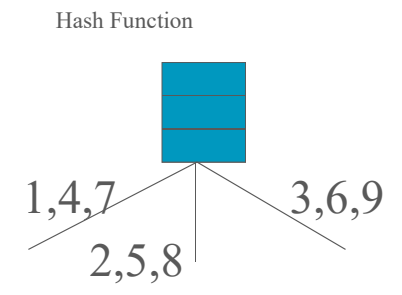
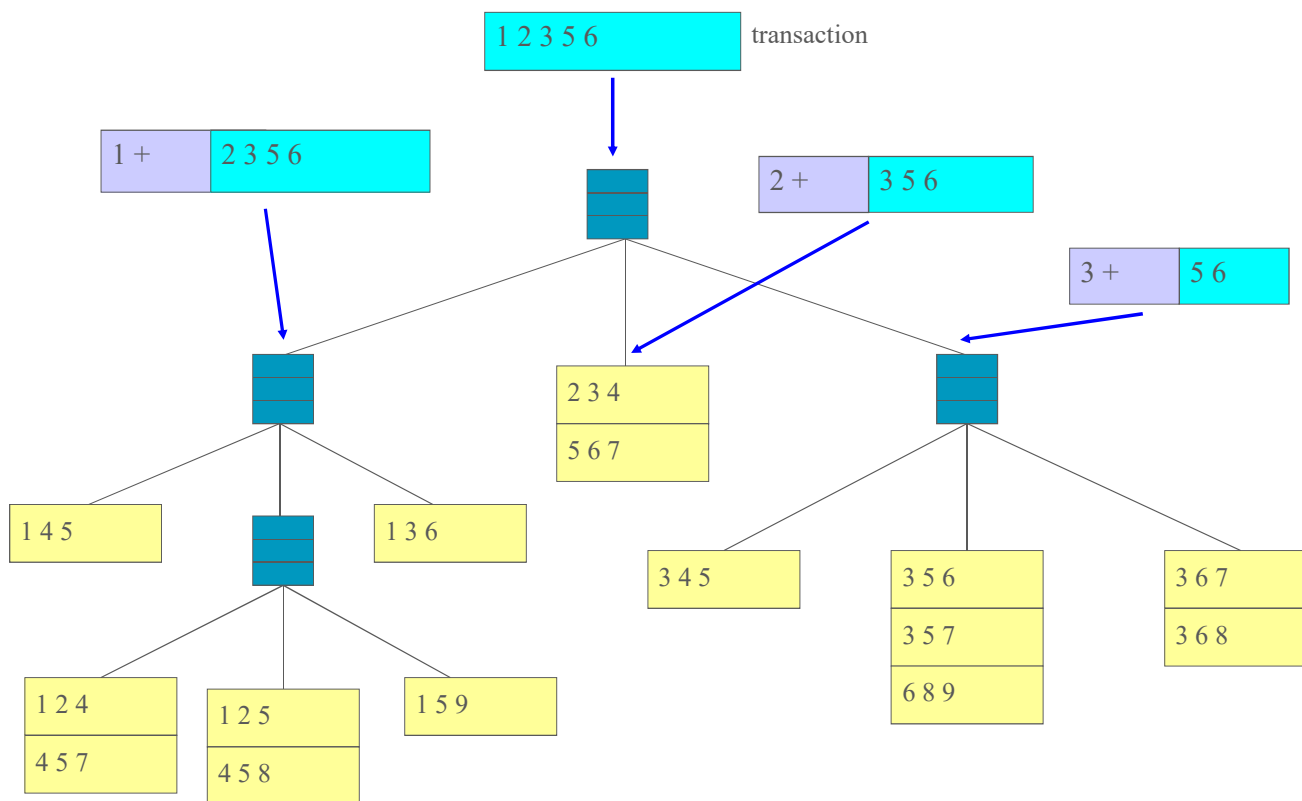
Función Hash

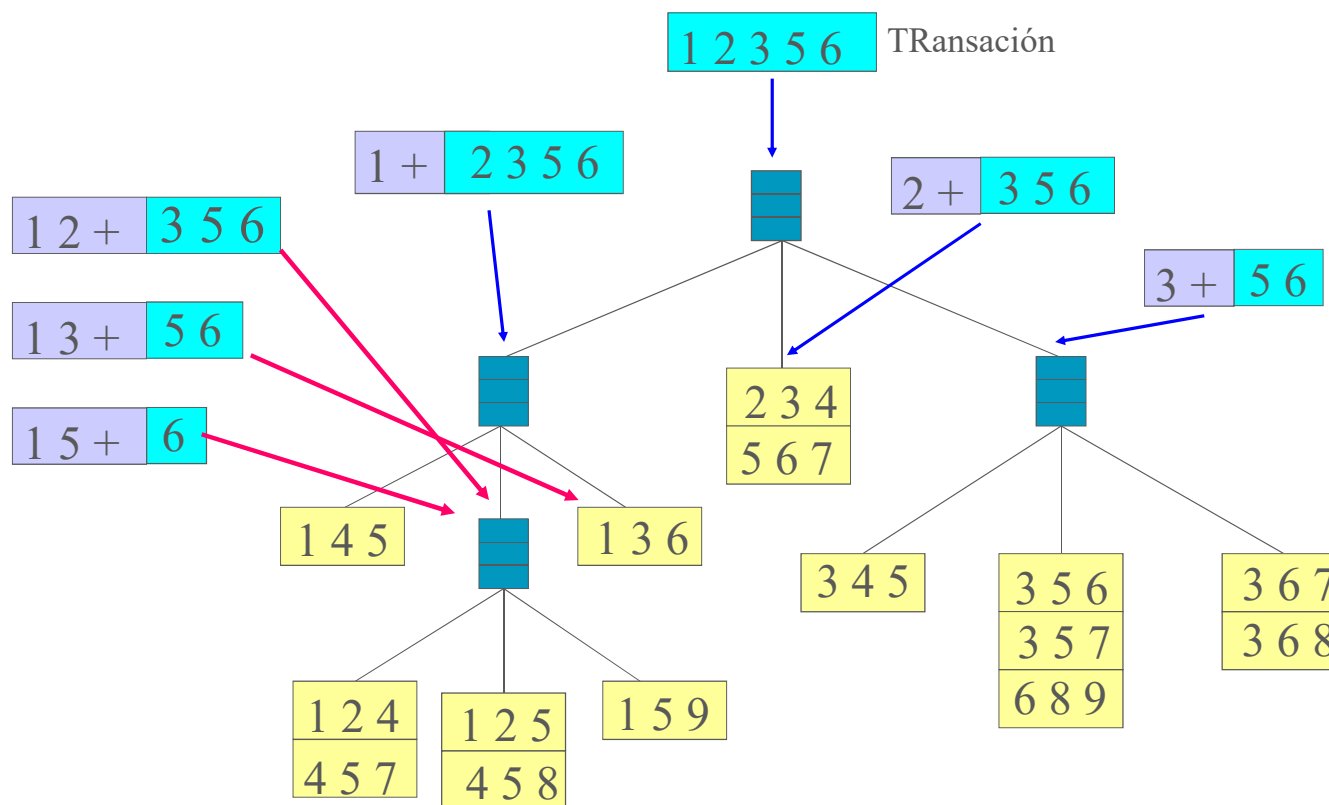


Hash sobre
3, 6 o 9

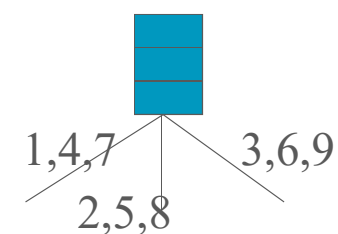
Candidatos a árboles hash

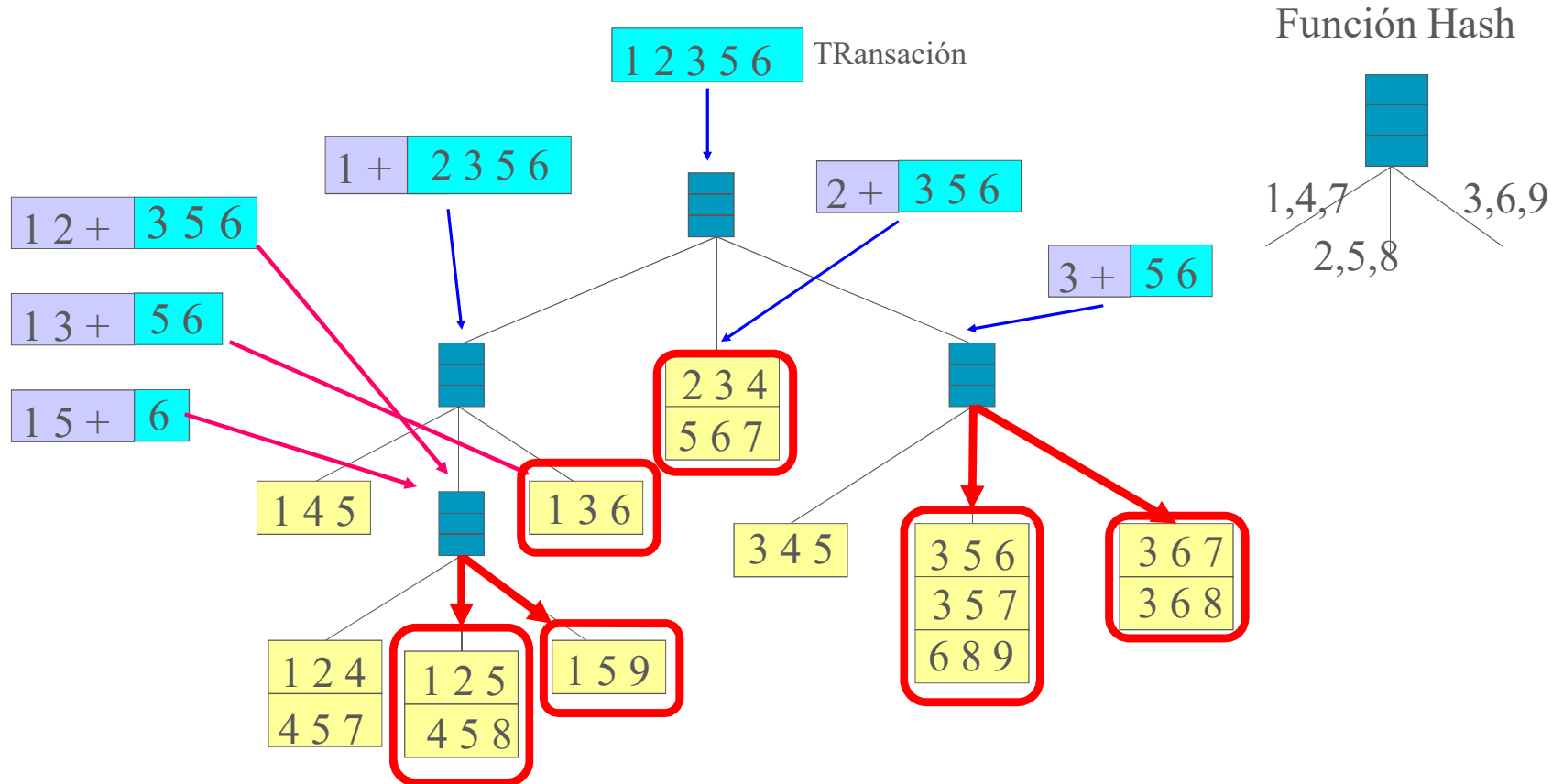






Función Hash





Factores que afectan la complejidad

Escoger un umbral mínimo de soporte

- Umbrales bajos de soporte resultan mayor numero de ítem frecuentes
- Esto puede aumentar el numero de candidatos y la máxima longitud de ítem frecuentes

Dimensional dad (Numero de items) del conjunto de datos

- Mas espacio se necesita para guardas la cuenta de soporte para cada ítem
- Si el numero de ítem frecuentes también se incrementa, los costos computacional y de I/O también se incrementa

Factores que afectan la complejidad

Tamaño de la base de datos

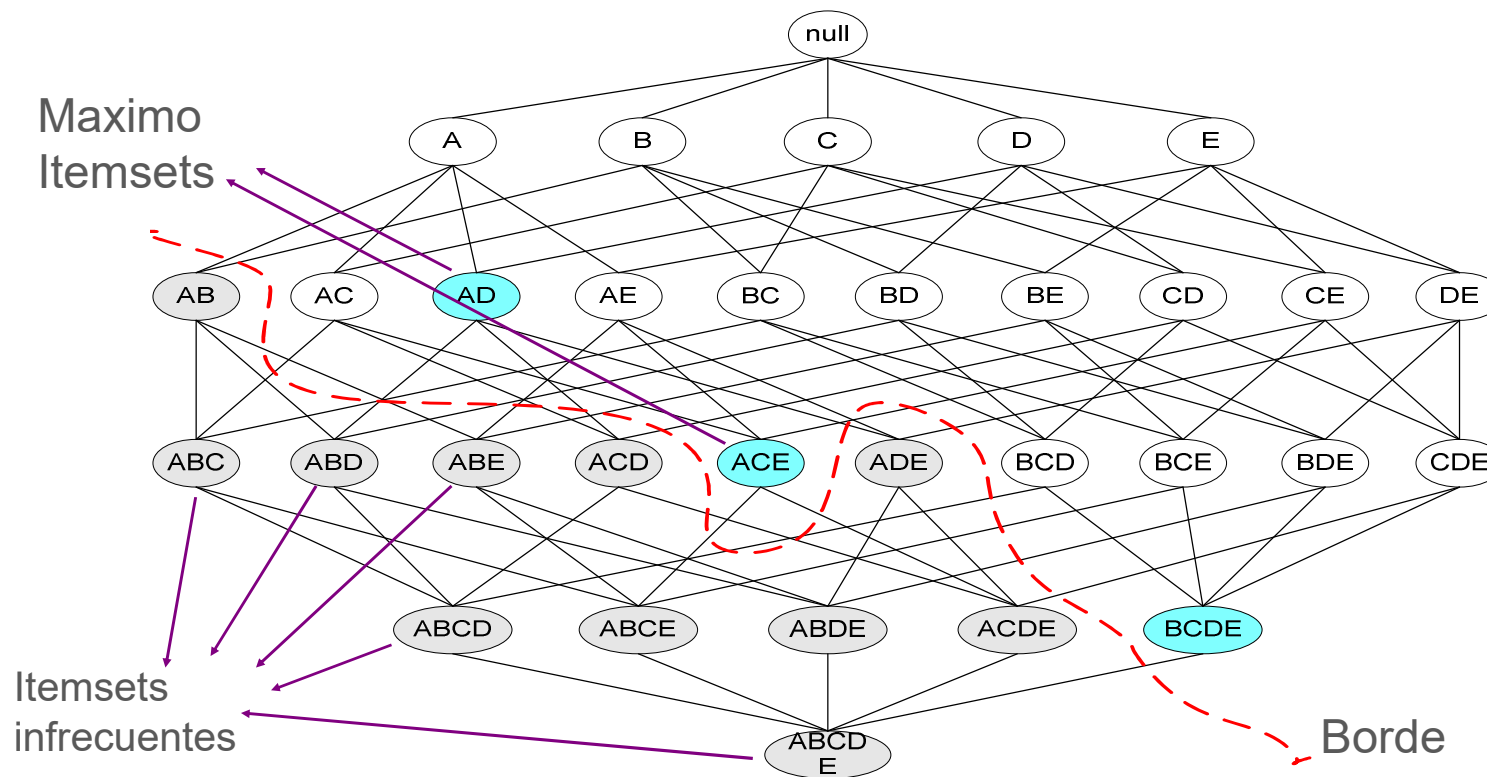
Amplitud promedio de la transacción

Algunos itemsets son redundantes dado que tiene un soporte identico

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	1	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	1	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	1	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	1	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	1	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	0	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	0	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	0	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	0	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	0	1	1	1	1	1	1	1	1	1

Número de itemset frecuentes $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$

Se nesesita una representación compacta

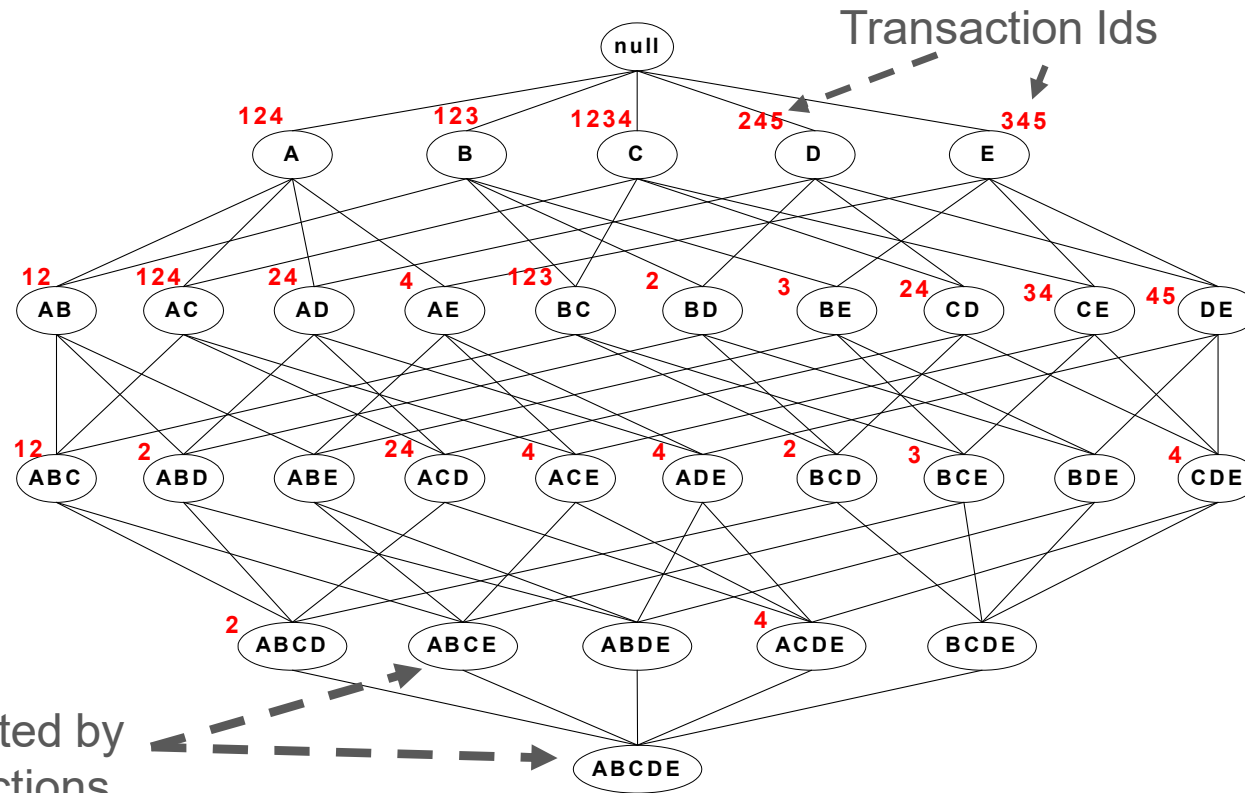


TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

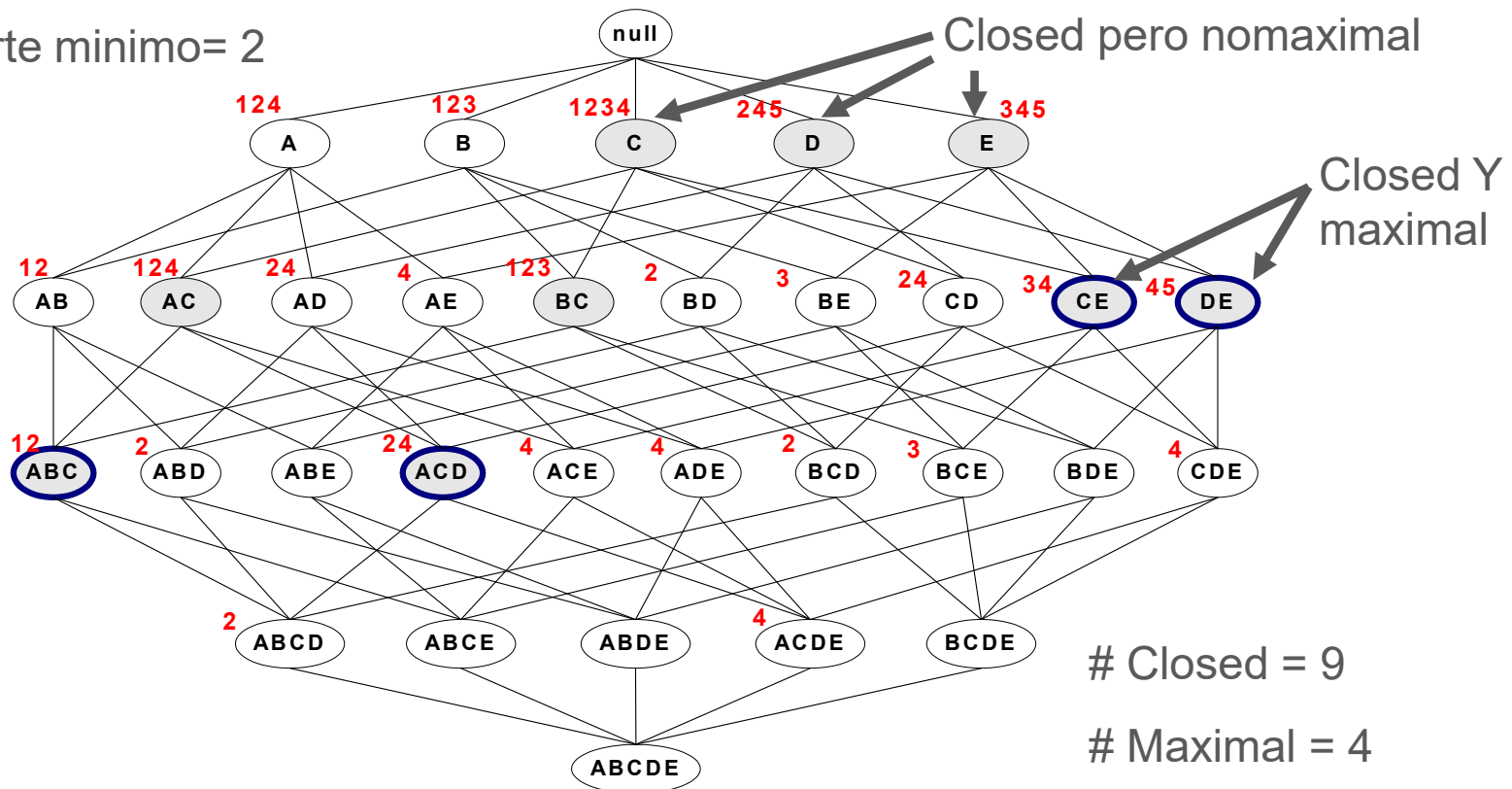
Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

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

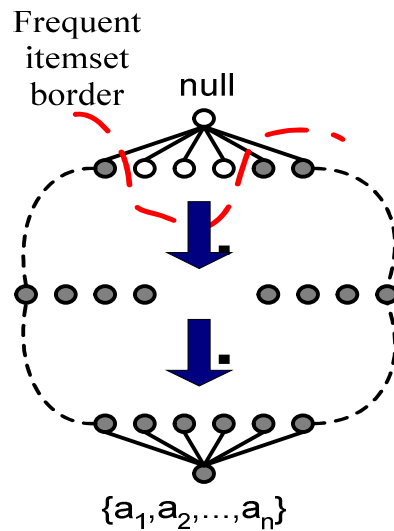
TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



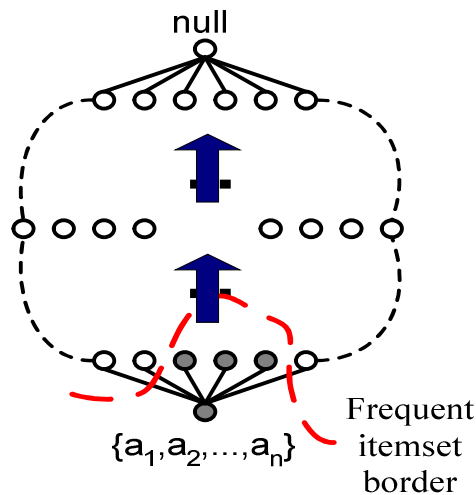
Soporte minimo= 2



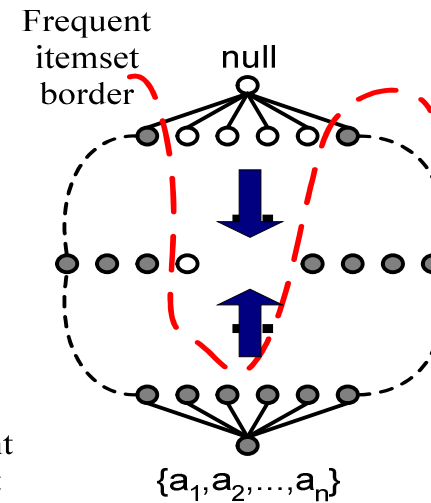
– General-to-specific vs Specific-to-general



(a) General-to-specific



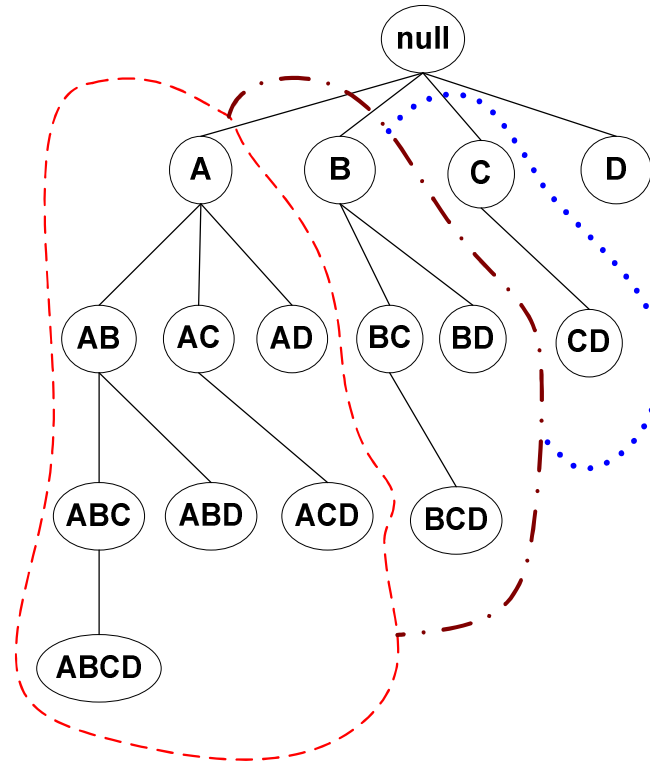
(b) Specific-to-general



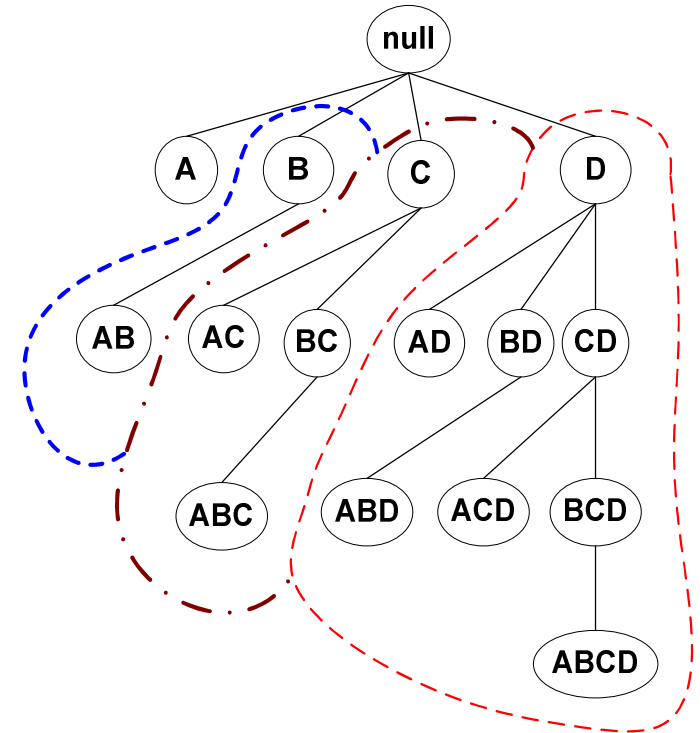
(c) Bidirectional

Traversal of Itemset Lattice

– Equivalent Classes

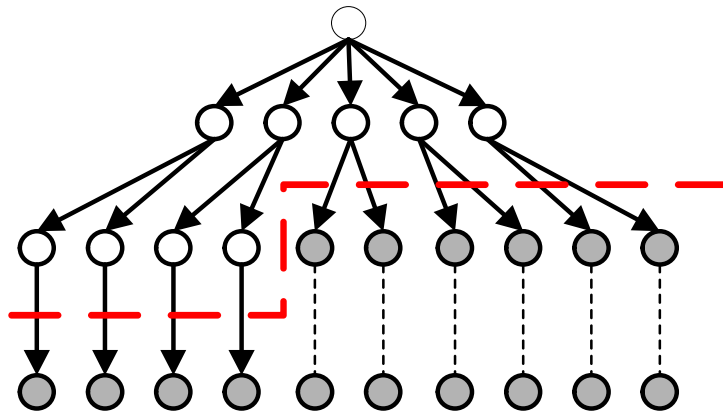


(a) Prefix tree

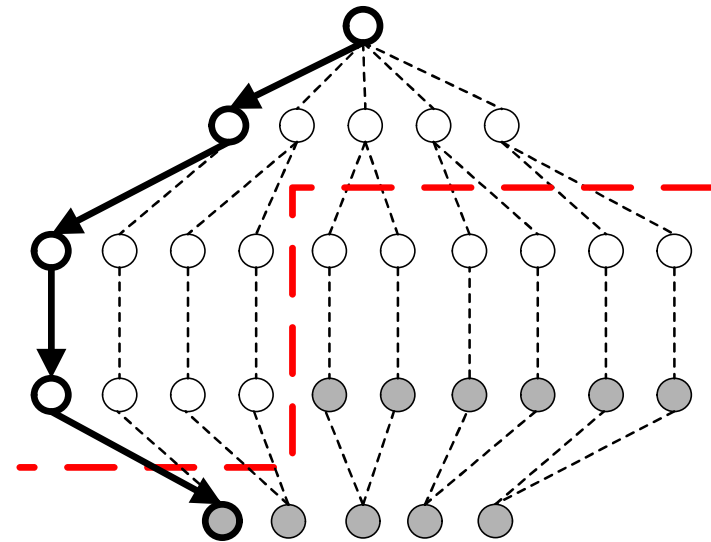


(b) Suffix tree

– Breadth-first vs Depth-first



(a) Breadth first



(b) Depth first

Representación de una base de datos

– Horizontal vs Vertical

Horizontal
Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Vertical Data Layout

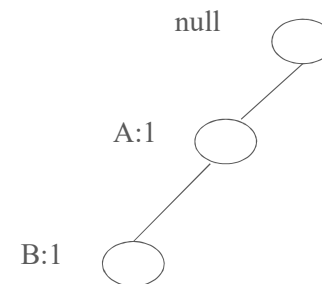
A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

FP-growth Algorithm

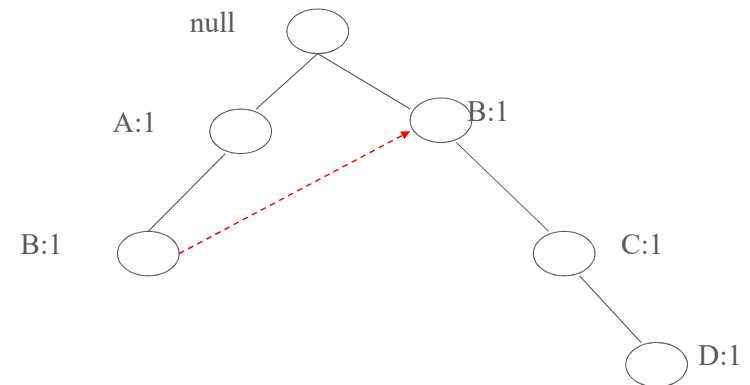
- Use a compressed representation of the database using an **FP-tree**
- Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

After reading TID=1:



After reading TID=2:

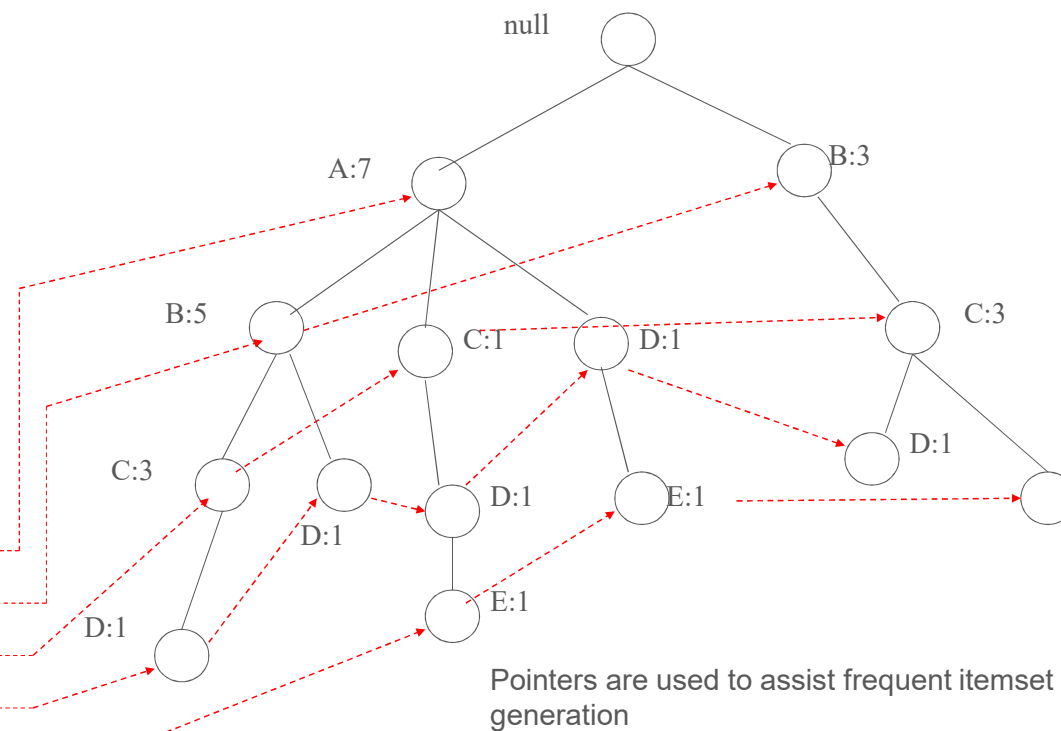


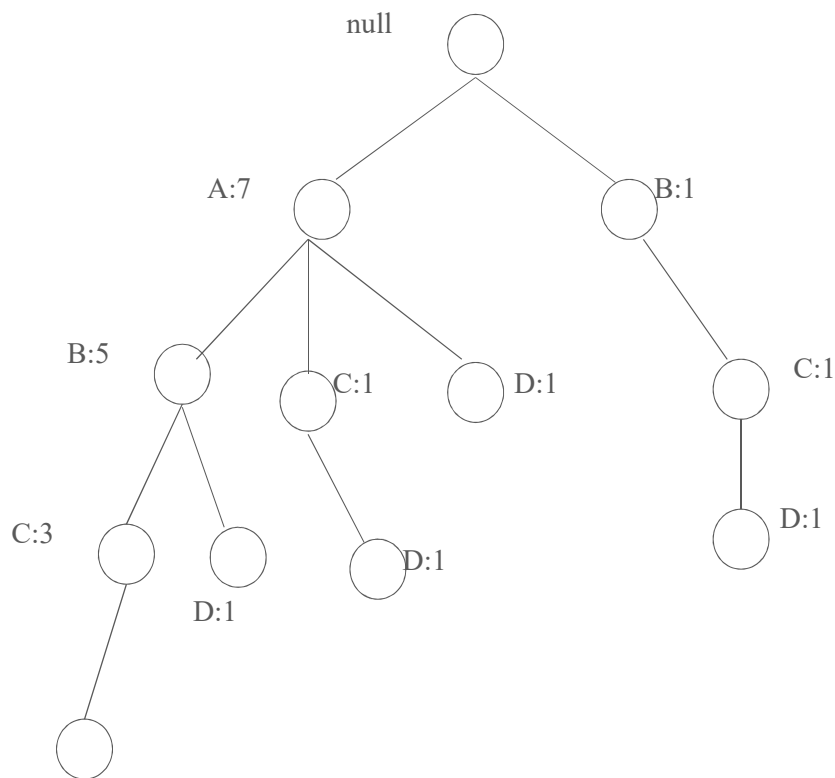
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Transaction Database

Header table

Item	Pointer
A	
B	
C	
D	
E	





Conditional Pattern base for D:

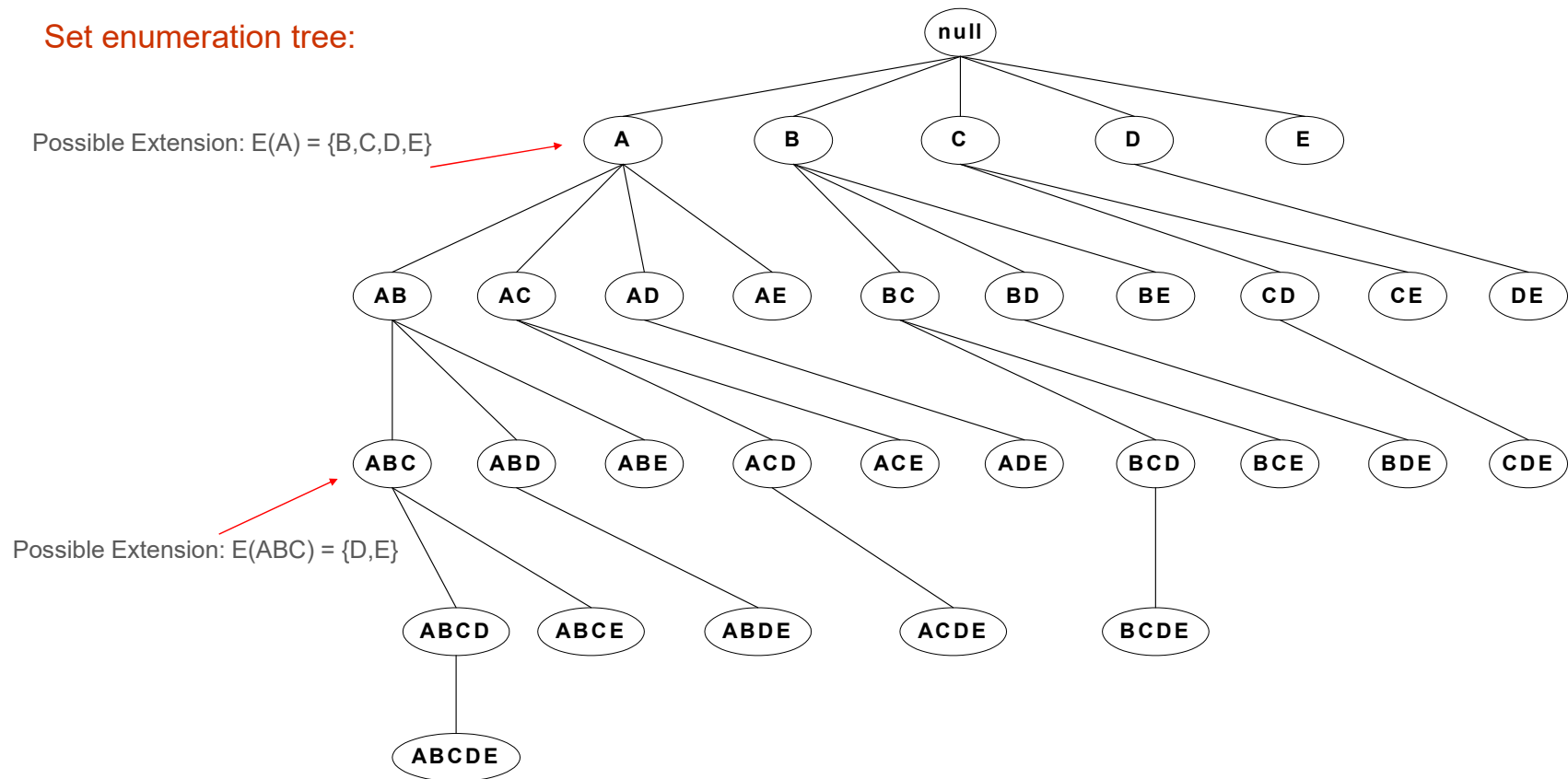
$$P = \{(A:1, B:1, C:1), \\ (A:1, B:1), \\ (A:1, C:1), \\ (A:1), \\ (B:1, C:1)\}$$

Recursively apply FP-growth on P

Frequent Itemsets found (with $\text{sup} > 1$):

AD, BD, CD, ACD, BCD

Set enumeration tree:





Items are listed in lexicographic order

Each node P stores the following information:

- Itemset for node P
- List of possible lexicographic extensions of P : $E(P)$
- Pointer to projected database of its ancestor node
- Bitvector containing information about which transactions in the projected database contain the itemset

Original Database:

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Projected Database for node A:

TID	Items
1	{B}
2	{}
3	{C,D,E}
4	{D,E}
5	{B,C}
6	{B,C,D}
7	{}
8	{B,C}
9	{B,D}
10	{}

For each transaction T, projected transaction at node A is $T \cap E(A)$

Generación de Reglas

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:

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

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

How to efficiently generate rules from frequent itemsets?

- In general, confidence does not have an anti-monotone property

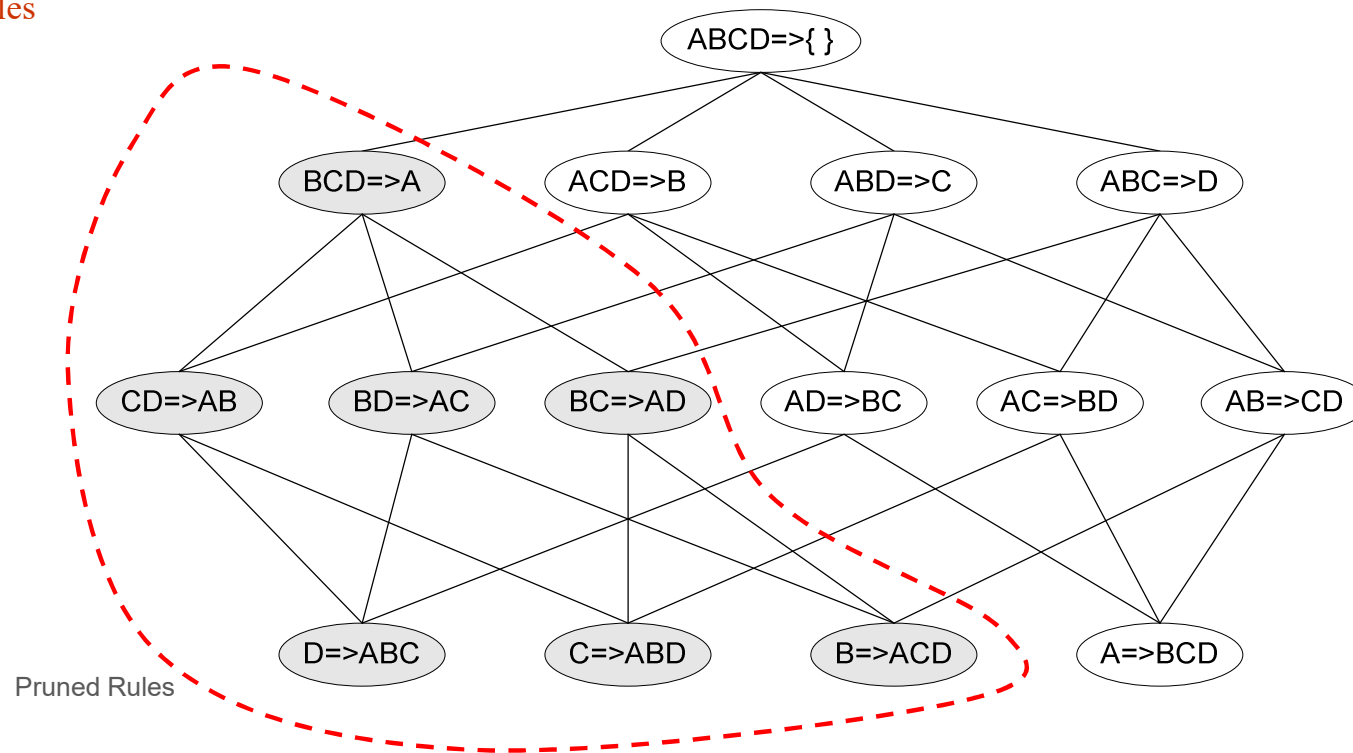
$c(ABC \rightarrow D)$ can be larger or smaller than $c(AB \rightarrow D)$

- But confidence of rules generated from the same itemset has an anti-monotone property
- e.g., $L = \{A, B, C, D\}$:

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

- Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

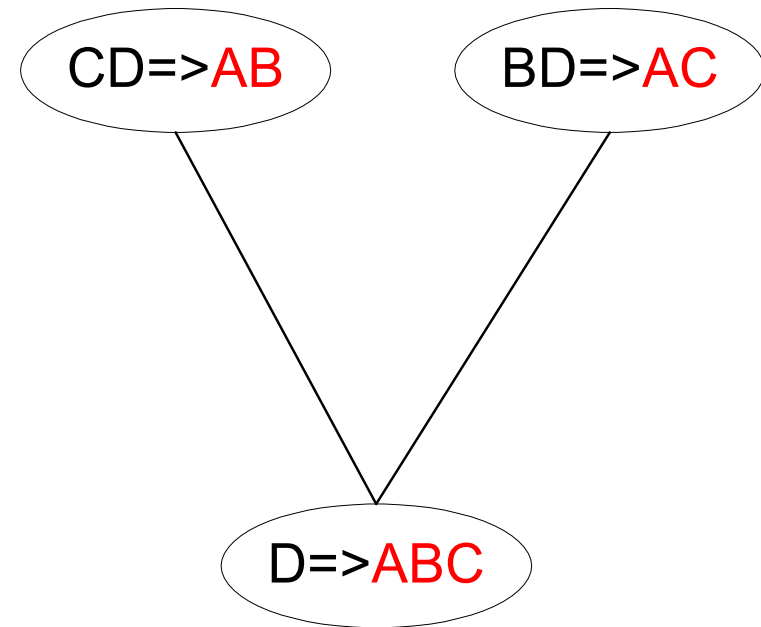
Lattice of rules



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

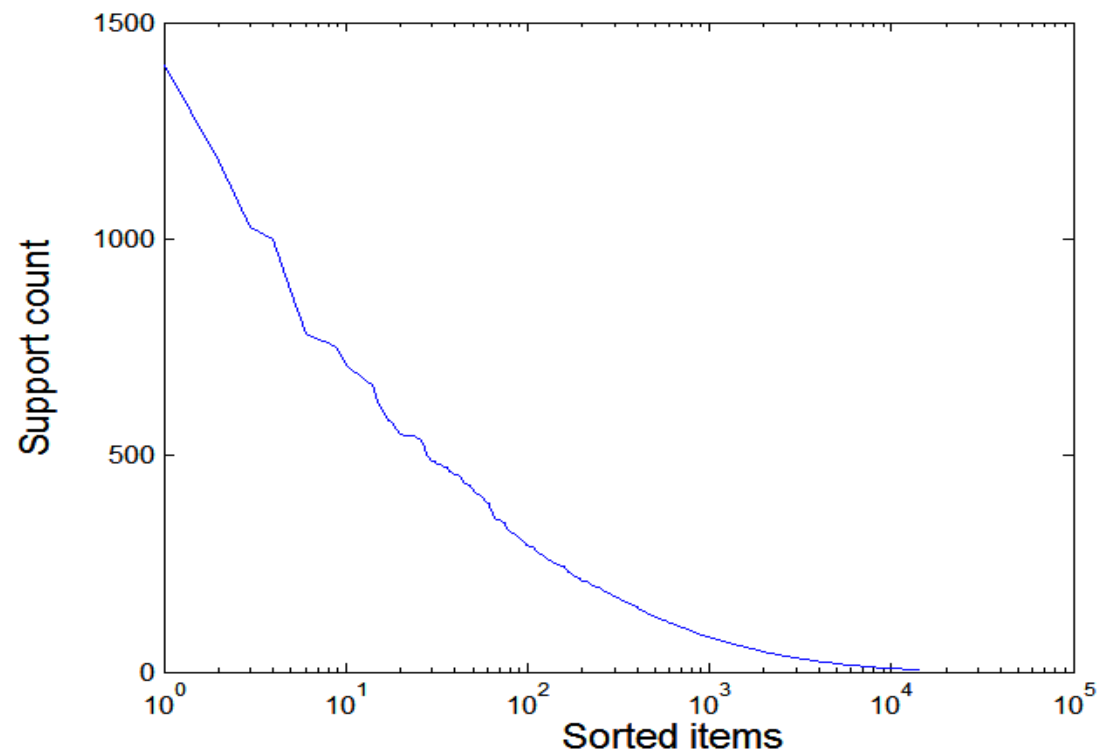
$\text{join}(\text{CD} \Rightarrow \text{AB}, \text{BD} \Rightarrow \text{AC})$
would produce the candidate rule $\text{D} \Rightarrow \text{ABC}$

Prune rule $\text{D} \Rightarrow \text{ABC}$ if its subset $\text{AD} \Rightarrow \text{BC}$ does not have high confidence



Many real data sets have skewed support distribution

Support
distribution of a
retail data set



How to set the appropriate *minsup* threshold?

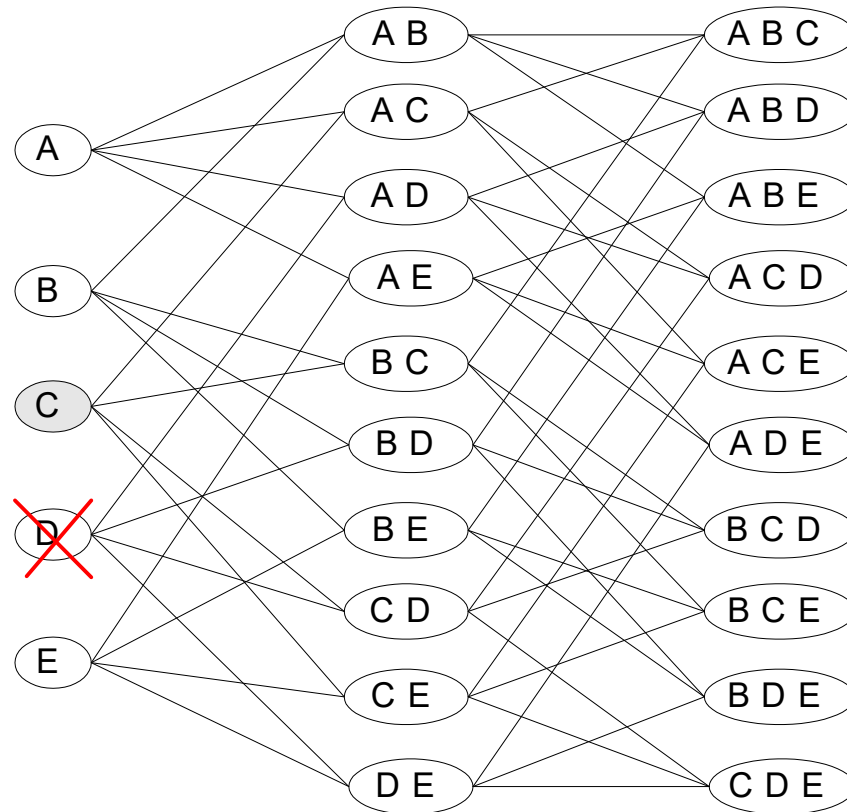
- If *minsup* is set too high, we could miss itemsets involving interesting rare items (e.g., expensive products)
- If *minsup* is set too low, it is computationally expensive and the number of itemsets is very large

Using a single minimum support threshold may not be effective

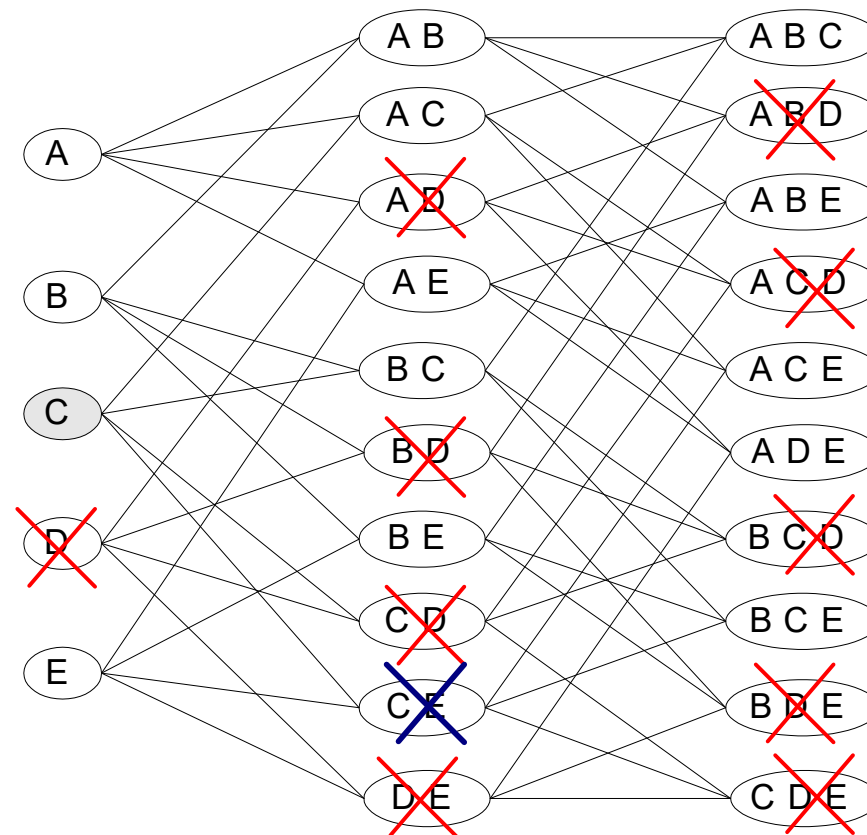
How to apply multiple minimum supports?

- $MS(i)$: minimum support for item i
- e.g.: $MS(\text{Milk})=5\%$, $MS(\text{Coke}) = 3\%$,
 $MS(\text{Broccoli})=0.1\%$, $MS(\text{Salmon})=0.5\%$
- $MS(\{\text{Milk}, \text{Broccoli}\}) = \min (MS(\text{Milk}), MS(\text{Broccoli}))$
 $= 0.1\%$
- Challenge: Support is no longer anti-monotone
 - Suppose: $\text{Support}(\text{Milk}, \text{Coke}) = 1.5\%$ and
 $\text{Support}(\text{Milk}, \text{Coke}, \text{Broccoli}) = 0.5\%$
 - $\{\text{Milk}, \text{Coke}\}$ is infrequent but $\{\text{Milk}, \text{Coke}, \text{Broccoli}\}$ is frequent

Item	MS (I)	Sup (I)
A	0.10 %	0.25 %
B	0.20 %	0.26 %
C	0.30 %	0.29 %
D	0.50 %	0.05 %
E	3 %	4.20 %



Item	MS (I)	Sup (I)
A	0.10 %	0.25 %
B	0.20 %	0.26 %
C	0.30 %	0.29 %
D	0.50 %	0.05 %
E	3 %	4.20 %



Multiple Minimum Support (Liu 1999)

Order the items according to their minimum support (in ascending order)

- e.g.: $MS(\text{Milk})=5\%$, $MS(\text{Coke}) = 3\%$,
 $MS(\text{Broccoli})=0.1\%$, $MS(\text{Salmon})=0.5\%$
- Ordering: Broccoli, Salmon, Coke, Milk

Need to modify Apriori such that:

- L_1 : set of frequent items
- F_1 : set of items whose support is $\geq MS(1)$
where $MS(1)$ is $\min_i(MS(i))$
- C_2 : candidate itemsets of size 2 is generated from F_1
instead of L_1

Modifications to Apriori:

- In traditional Apriori,
 - A candidate $(k+1)$ -itemset is generated by merging two frequent itemsets of size k
 - The candidate is pruned if it contains any infrequent subsets of size k
- Pruning step has to be modified:
 - Prune only if subset contains the first item
 - e.g.: Candidate={Broccoli, Coke, Milk} (ordered according to minimum support)
 - {Broccoli, Coke} and {Broccoli, Milk} are frequent but {Coke, Milk} is infrequent
 - Candidate is not pruned because {Coke, Milk} does not contain the first item, i.e., Broccoli.

Evaluación de Reglas de asociación

Los algoritmos de asociación tienden a producir muchas reglas

- Muchas de ellas redundantes o poco interesantes
- Redundantes si $\{A,B,C\} \rightarrow \{D\}$ y $\{A,B\} \rightarrow \{D\}$ tienen le mismo soporte y confianza

Las medidas de interés sirven para podar o ranquiar los patrones derivados

Computar medidas interesantes

Dada la regla $X \rightarrow Y$, La información necesaria para computar las medidas de interés está en la tabla de contingencia

Tabla de contingencia para $X \rightarrow Y$

	Y	\bar{Y}	
X	f_{11}	f_{10}	f_{1+}
\bar{X}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	$ T $

f_{11} : Soporte de X y Y

f_{10} : Soporte de X y \bar{Y}

f_{01} : Soporte de \bar{X} y Y

f_{00} : Soporte de \bar{X} y \bar{Y}

Usada para varios indicadores

◆ Soporte, Confianza, lift, Gini, J-measure, etc.

Inconveniente de a confianza

	Ron	Ron	
Agua	15	5	20
Agua	75	5	80
	90	10	100

Regla de asociacion: **Agua** → **Ron**

Confianza= $P(\text{Ron}|\text{Agua}) = 0.75$

Pero $P(\text{Ron}) = 0.9$

⇒ Aunque a confianza es alta, la regla es engañosa

⇒ $P(\text{Ron}|\text{Agua}) = 0.9375$

Independencia estadística

Oblation de 1000 Estudiantes

- 600 Estudiantes saben nadar (S)
- 700 Estudiantes saben montar en bici (B)
- 420 Estudiantes saben nadar y montar en bici(S,B)
- $P(S \cap B) = 420/1000 = 0.42$
- $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$
- $P(S \cap B) = P(S) \times P(B) \Rightarrow$ Independencia estadística
- $P(S \cap B) > P(S) \times P(B) \Rightarrow$ Positivamente correlacionados
- $P(S \cap B) < P(S) \times P(B) \Rightarrow$ Negativamente correlacionados

Medidas basadas en estadística

Measures that take into account statistical dependence

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

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

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

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

Ejemplo: Lift/Interest

	Ron	Ron	
Agua	15	5	20
Agua	75	5	80
	90	10	100

Regla de asociacion: Agua \rightarrow Ron

Confianza= $P(\text{Ron}|\text{Agua}) = 0.75$

Pero $P(\text{Ron}) = 0.9$

$\Rightarrow \text{Lift} = 0.75/0.9 = 0.8333 (< 1, \text{ Así estan negativamente asociados})$

Fallos de Lift & Interest

	Y	Y	
X	10	0	10
X	0	90	90
	10	90	100

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

	Y	Y	
X	90	0	90
X	0	10	10
	90	10	100

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

Independencia estadística:

If $P(X,Y)=P(X)P(Y) \Rightarrow Lift = 1$

#	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)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\bar{A}\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A}\bar{B}) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\bar{A}\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A}\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
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)	$\max \left(P(A, B) \log \left(\frac{P(B A)}{P(B)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{B} \bar{A})}{P(\bar{B})} \right), \right. \\ \left. P(A, B) \log \left(\frac{P(A B)}{P(A)} \right) + P(\bar{A}\bar{B}) \log \left(\frac{P(\bar{A} \bar{B})}{P(\bar{A})} \right) \right)$
9	Gini index (G)	$\max \left(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] \right. \\ \left. - P(B)^2 - P(\bar{B})^2, \right. \\ \left. P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] \right. \\ \left. - P(A)^2 - P(\bar{A})^2 \right)$
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)	$\max \left(\frac{P(A)P(B)}{P(\bar{A}\bar{B})}, \frac{P(B)P(A)}{P(\bar{B}\bar{A})} \right)$
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)	$\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A) + P(B) - P(A,B)}$
21	Klogsen (K)	$\sqrt{P(A,B) \max(P(B A) - P(B), P(A B) - P(A))}$

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?

Propiedades de una Buena medida

Piatetsky-Shapiro:

3 properties a good measure M must satisfy:

- $M(A,B) = 0$ if A and B are statistically independent
- $M(A,B)$ increase monotonically with $P(A,B)$ when $P(A)$ and $P(B)$ remain unchanged
- $M(A,B)$ decreases monotonically with $P(A)$ [or $P(B)$] when $P(A,B)$ and $P(B)$ [or $P(A)$] remain unchanged

Comparación de diferentes medidas

10 examples of contingency tables:

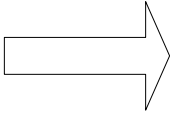
Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

Rankings of contingency tables using various measures:

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Propiedades sobre permutación de variables

	B	$\overline{\mathbf{B}}$
A	p	q
$\overline{\mathbf{A}}$	r	s



	A	$\overline{\mathbf{A}}$
B	p	r
$\overline{\mathbf{B}}$	q	s

Does $M(A,B) = M(B,A)$?

Symmetric measures:

- ♦ support, lift, collective strength, cosine, Jaccard, etc

Asymmetric measures:

- ♦ confidence, conviction, Laplace, J-measure, etc

The slide features a white background with abstract blue geometric shapes in the corners. In the top right, there is a large, light blue triangle pointing towards the center. In the bottom left, there is a smaller, darker blue triangle pointing towards the center. The text is centered in the middle of the slide.

THANK YOU!

ANY QUESTIONS?