Data Mining: Association Rules

Jef Wijsen

Université de Mons (UMONS)

La BD et les règles

1. Des items $\{A, B, C, D, \ldots\}$.
2. Tout ensemble fini d'items est appelé un itemset ou une transaction.
3. La BD est un ensemble (plutôt un *bag*) de transactions.

Par ex.,

$\{B, C, D, E\}$
$\{A, B, C\}$
$\{A, C, D\}$
$\{A, B, C, D\}$

Une règle d'association est une expression de la forme $X \Rightarrow Y$ où X et Y sont des itemsets disjoints. Par ex., $\{B, C\} \Rightarrow \{D\}$ ou $B, C \Rightarrow D$.

Le problème

- Le support d'une règle $X \Rightarrow Y$ est s si $s\%$ pourcent des transactions contiennent tous les items en $X \cup Y$.
- La confiance d'une règle $X \Rightarrow Y$ est c si $c\%$ des transactions qui contiennent X, contiennent aussi Y.
- Normalement, on utilise une fraction entre 0 et 1 au lieu d'un pourcentage.
- Dans l'exemple, la règle $B, C \Rightarrow D$ a un support de $2/4$ et une confiance de $2/3$.
- Soient donnés un seuil de support (souvent ≤ 0.1) et un seuil de confiance (plutôt proche de 1.0), trouver toutes les règles qui dépassent ces seuils.

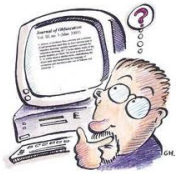
Fichier $\rightsquigarrow$ BD de transactions

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
	...			

$\rightsquigarrow$

{outlook=sunny,temperature=hot,humidity=high,windy=false,play=no}
{outlook=sunny,temperature=hot,humidity=high,windy=true,play=no}

- "outlook=sunny", "temperature=hot" jouent le rôle d'items.
- Il n'y a pas d'attribut cible !



Exercice

- Créer un fichier `paniers.arff` pour stocker le contenu du "panier de la ménagère" des clients qui passent à la caisse.
- Découvrir des règles d'association en `paniers.arff`.
- Comprendre les messages de type "`Size of set of large itemsets L(2): 26`"

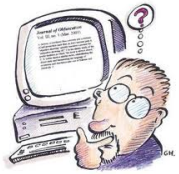
Attention : Confiance $\neq$ Corrélation

Consider "basket $\Rightarrow$ corn flakes" and suppose there are 5000 students as follows :

- 3000 students play basketball ;
- 3750 students eat corn flakes ;
- 2000 students both play basketball and eat corn flakes.

The rule "basket $\Rightarrow$ corn flakes" has a confidence of $\frac{2000}{3000} = 2/3$, but note :

$$\frac{Pr(\text{basket} \wedge \text{corn flakes})}{Pr(\text{basket})Pr(\text{corn flakes})} = 0.889 < 1 \quad .$$



Exercice

Pour l'ensemble

$$\{AB, ACDE, BCDF, ABCD, ABCF\},$$

trouvez tout sous-ensemble de $ABCDEF$ avec un *support count* ≥ 2.

- Le support count de X, dénoté $\sigma(X)$, est le nombre de transactions qui incluent X. Par exemple, $\sigma(BC) = 3$.
- Notez :
  - le support de $X \Rightarrow Y$ est égal à $\frac{\sigma(XY)}{N}$ avec N le nombre de transactions (à multiplier par 100 pour obtenir un pourcentage) ;
  - la confiance de $X \Rightarrow Y$ est égale à $\frac{\sigma(XY)}{\sigma(X)}$.

Apriori : élagage

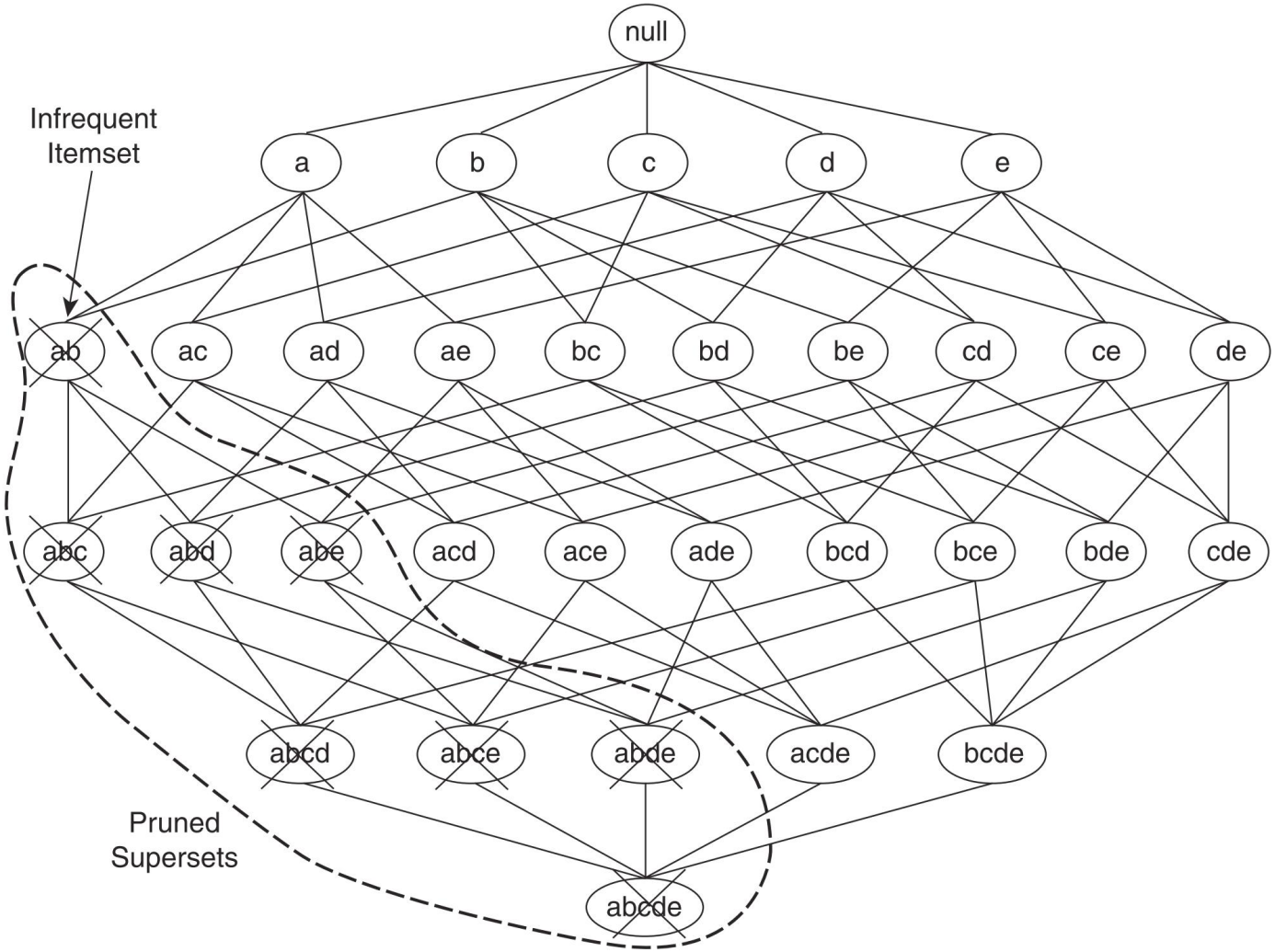


Figure 6.4. An illustration of support-based pruning. If $\{a, b\}$ is infrequent, then all supersets of $\{a, b\}$ are infrequent.

Apriori : k-itemsets inclus dans une transaction

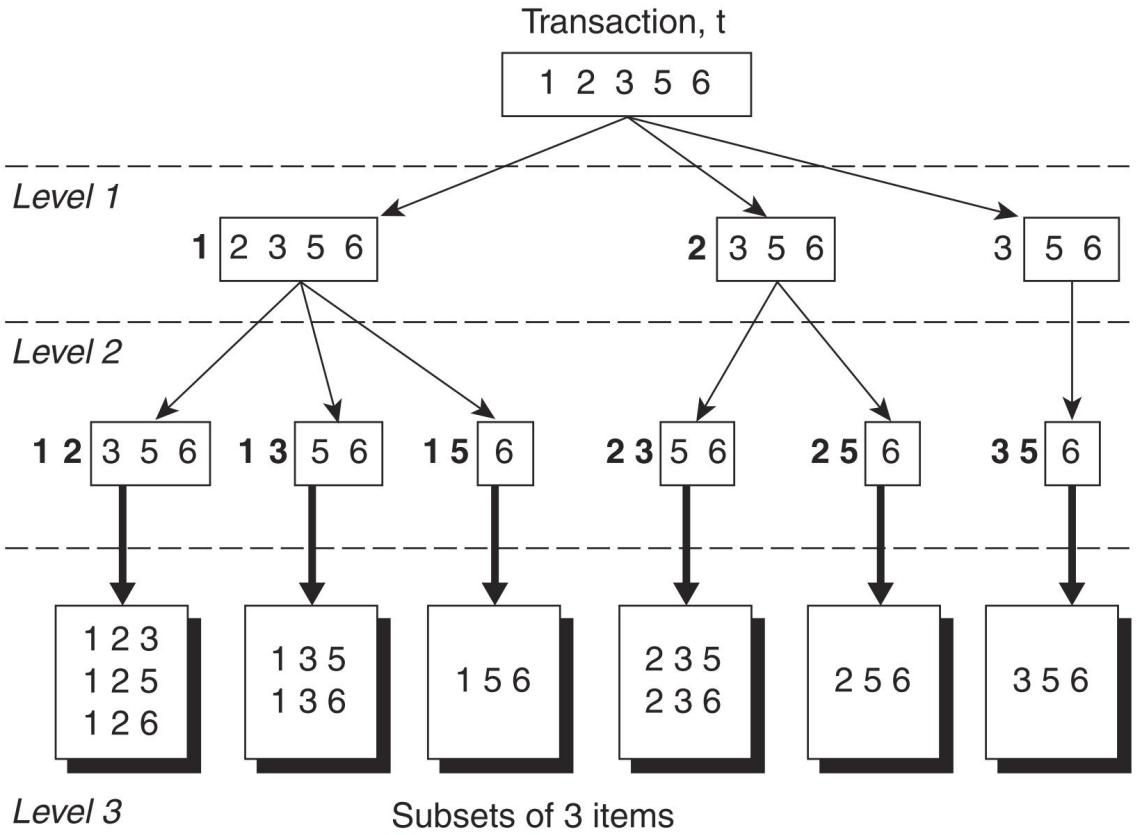


Figure 6.9. Enumerating subsets of three items from a transaction t.

Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

Apriori : candidats “hachés” et comptage

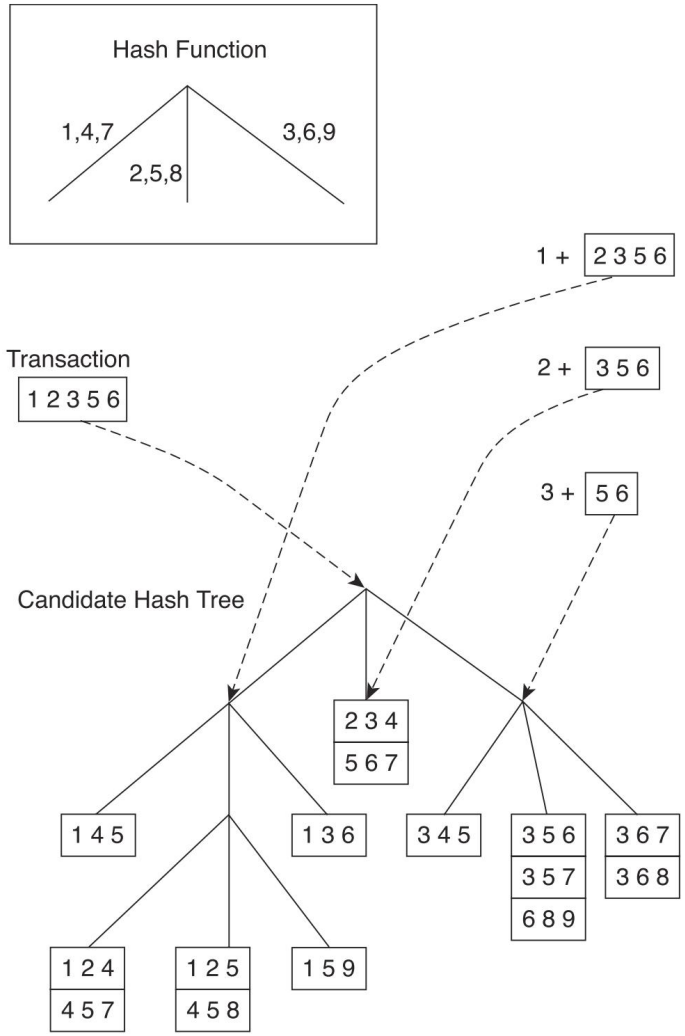


Figure 6.11. Hashing a transaction at the root node of a hash tree.

Apriori : candidats “hachés” et comptage

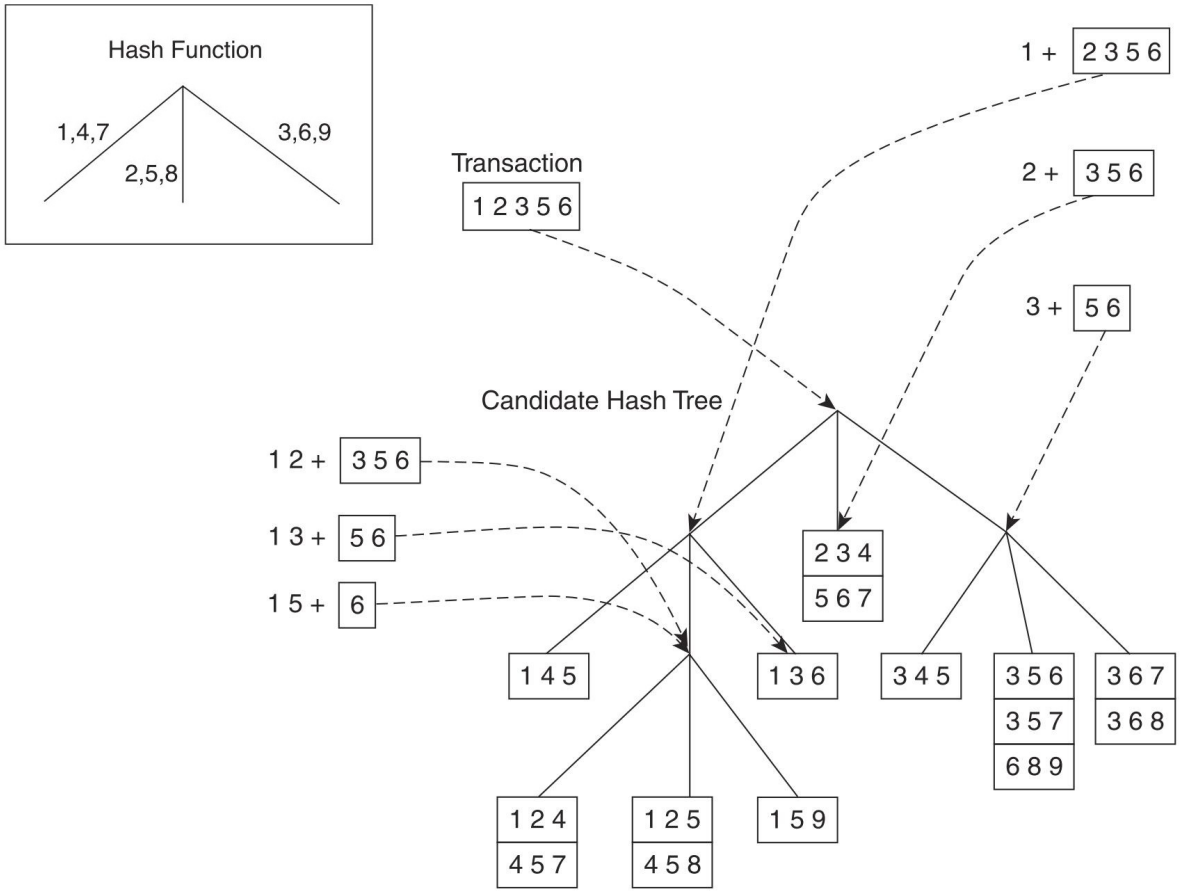
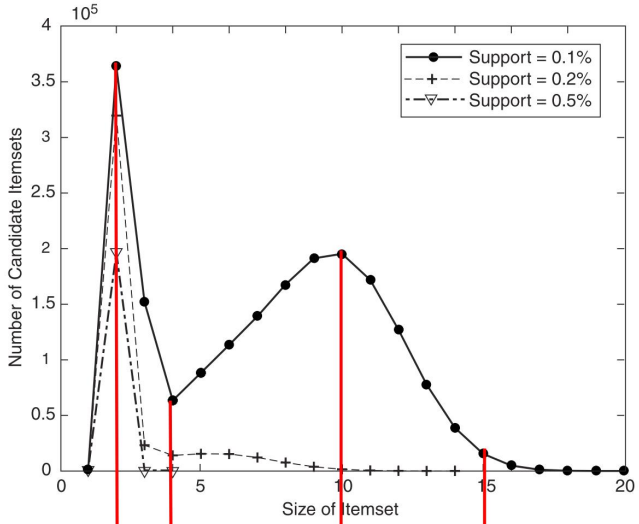


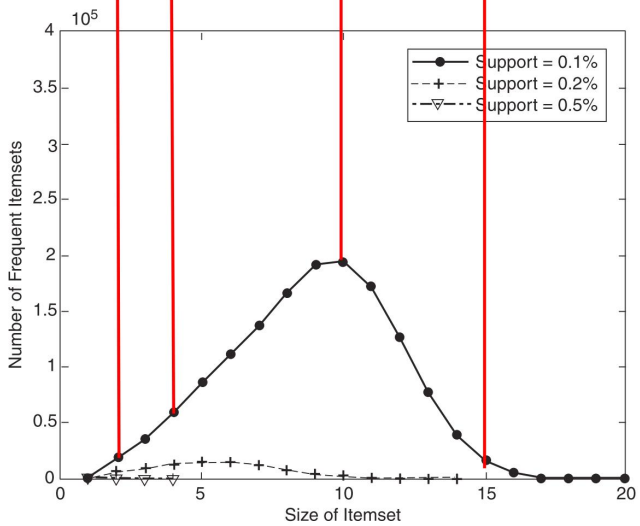
Figure 6.12. Subset operation on the leftmost subtree of the root of a candidate hash tree.

Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

Apriori : validation expérimentale



(a) Number of candidate itemsets.



Discuss this behavior.

Apriori : construire les règles

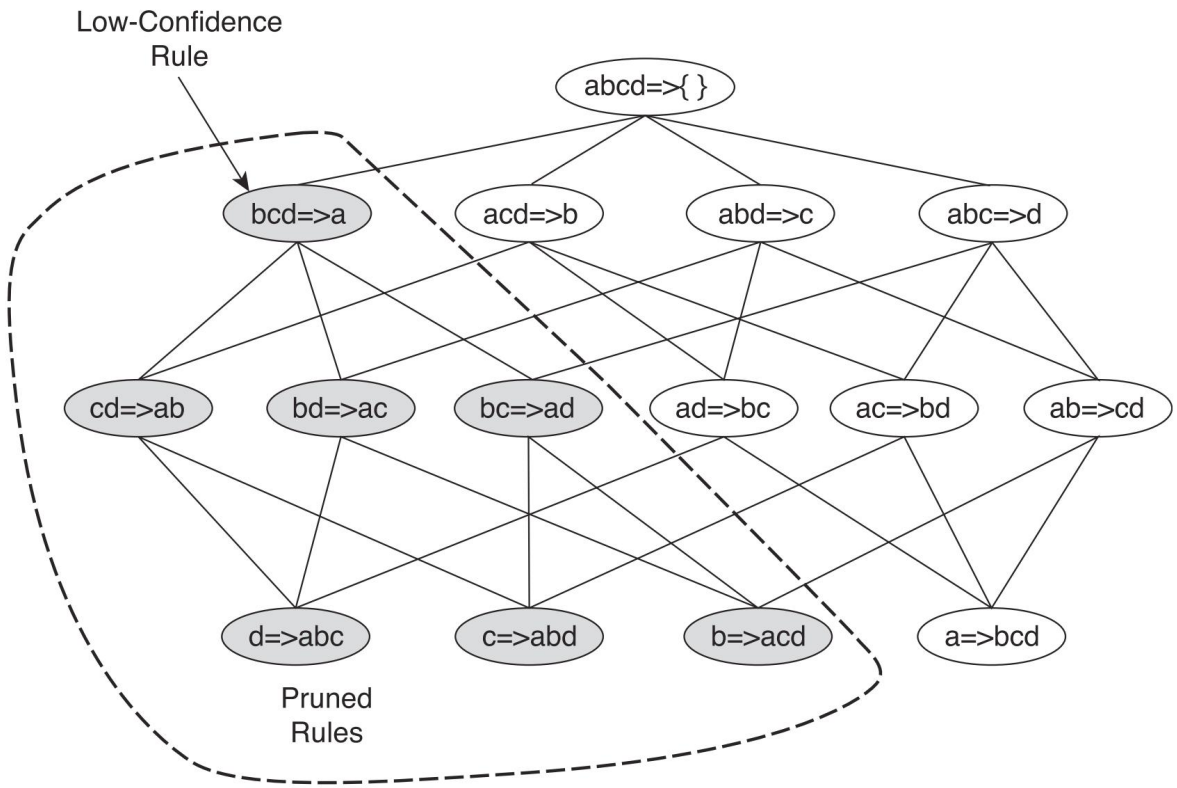


Figure 6.15. Pruning of association rules using the confidence measure.

{a,b,c,d} is fixed.

Rules are ordered by set inclusion on right-hand sides.

Itemsets fréquents maximaux

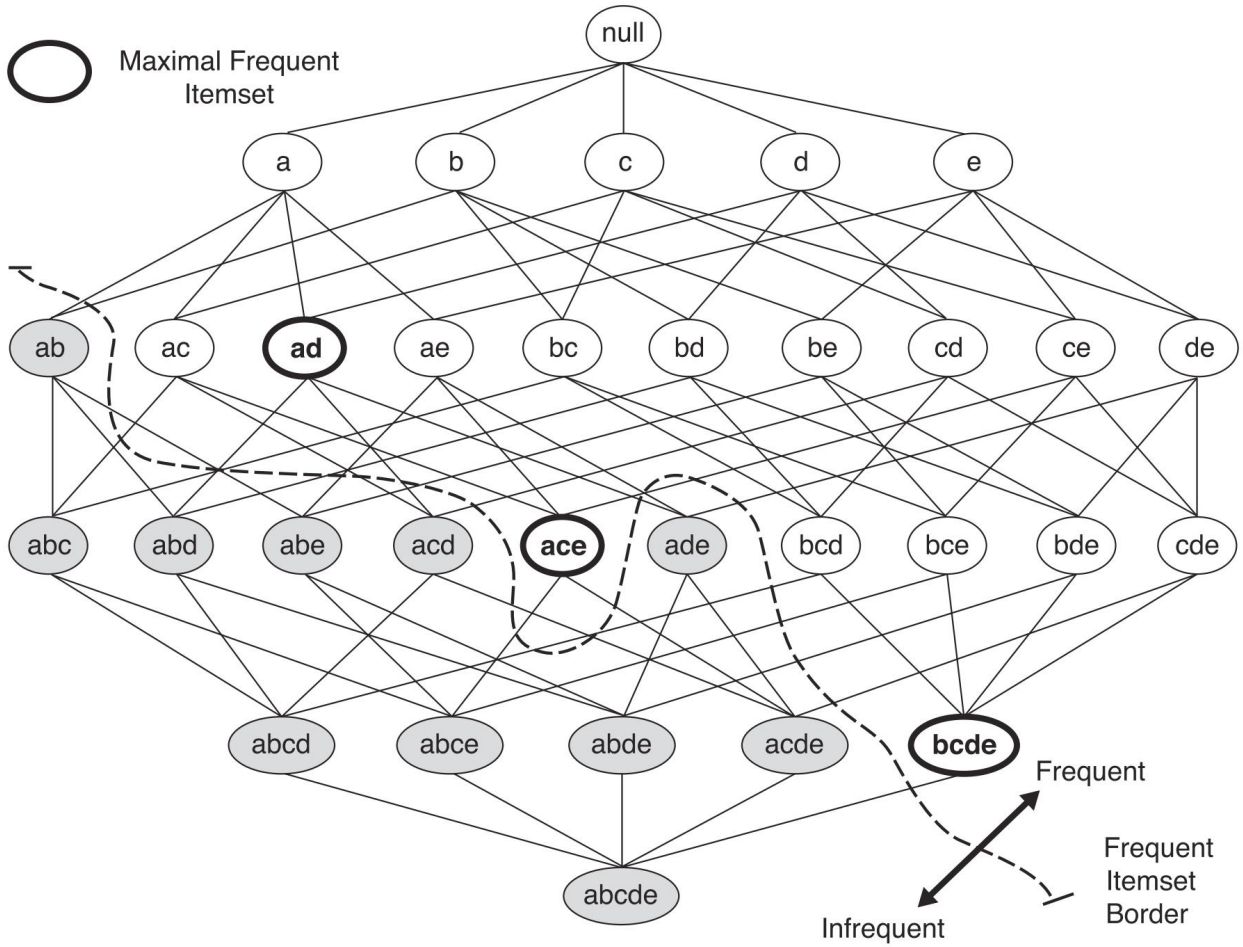


Figure 6.16. Maximal frequent itemset.

Itemsets fréquents fermés

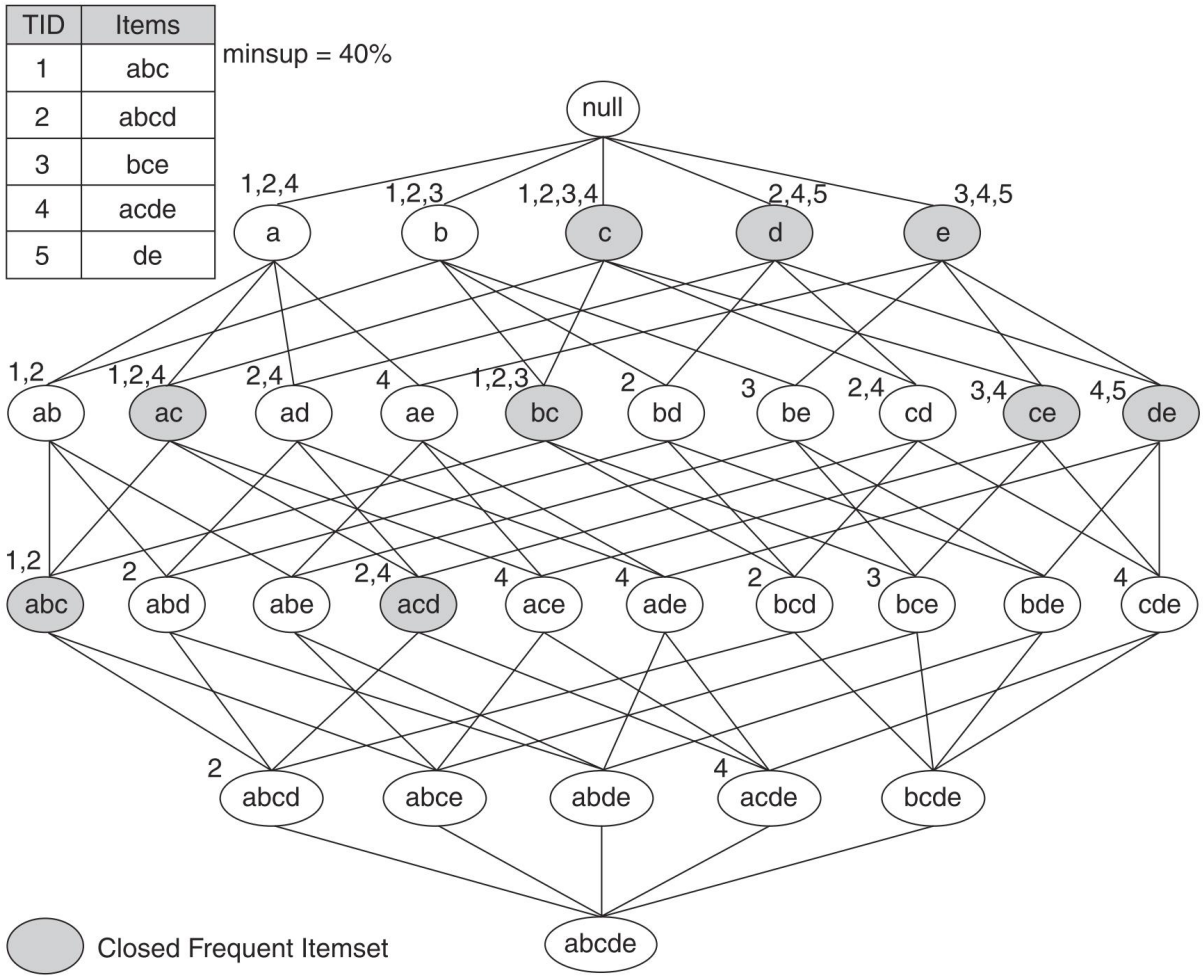
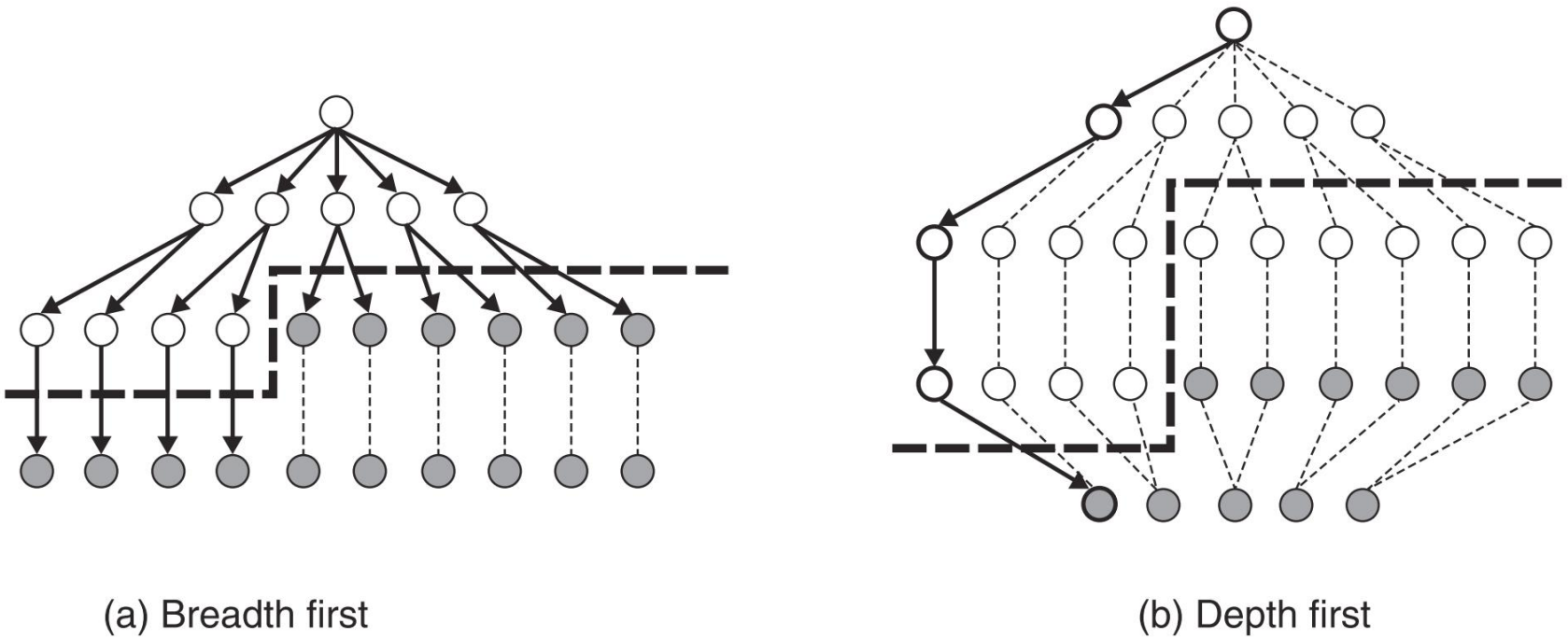


Figure 6.17. An example of the closed frequent itemsets (with minimum support count equal to 40%).

Parcours en largeur ou en profondeur



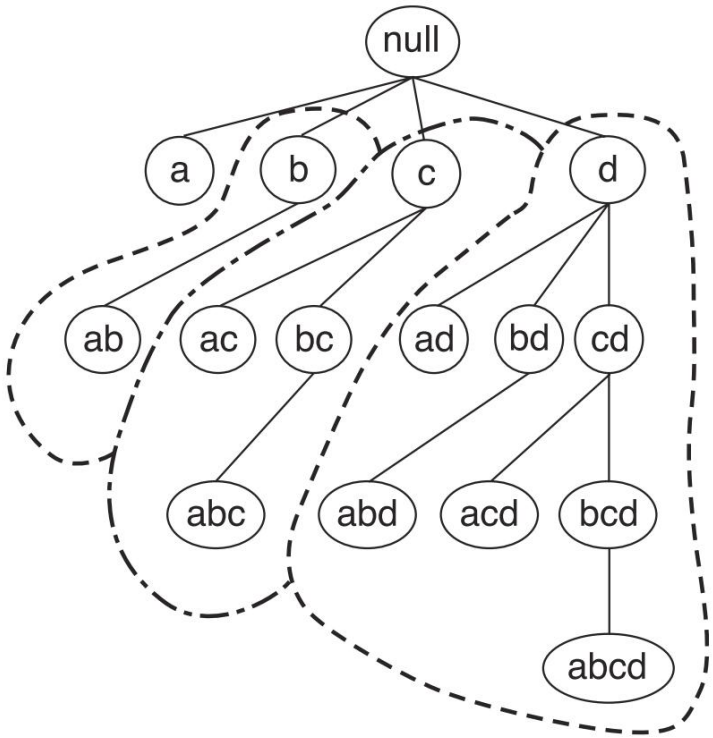
(a) Breadth first

(b) Depth first

Figure 6.21. Breadth-first and depth-first traversals.

Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

Parcours en profondeur



(b) Suffix tree.

Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

Représentations dites “horizontale” et “verticale”

Transaction Data Set

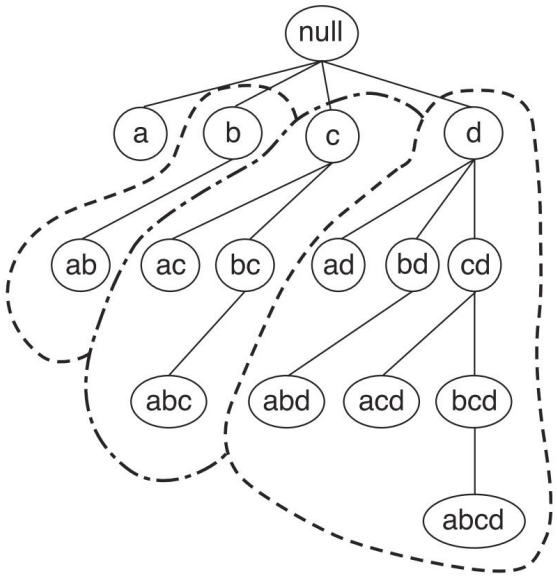
TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

$$
\begin{aligned}
\text{cover}(a) &= \{\ 1,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9\ \} \\
\text{cover}(b) &= \{\ 1,\ 2,\ 5,\ 6,\ 8,\ 9,\ 10\ \} \\
\text{cover}(c) &= \{\ 2,\ 3,\ 5,\ 6,\ 8,\ 10\ \} \\
\text{cover}(d) &= \{\ 2,\ 3,\ 4,\ 6,\ 9\ \} \\
\text{cover}(e) &= \{\ 3,\ 4,\ 10\ \}
\end{aligned}
$$

Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

Calcul récursif de cover

Pour chaque enfant $x \cdot s$ de s, cover$(x \cdot s)$ est l'intersection de cover(s) avec le cover du preceding-sibling de s qui contient x.



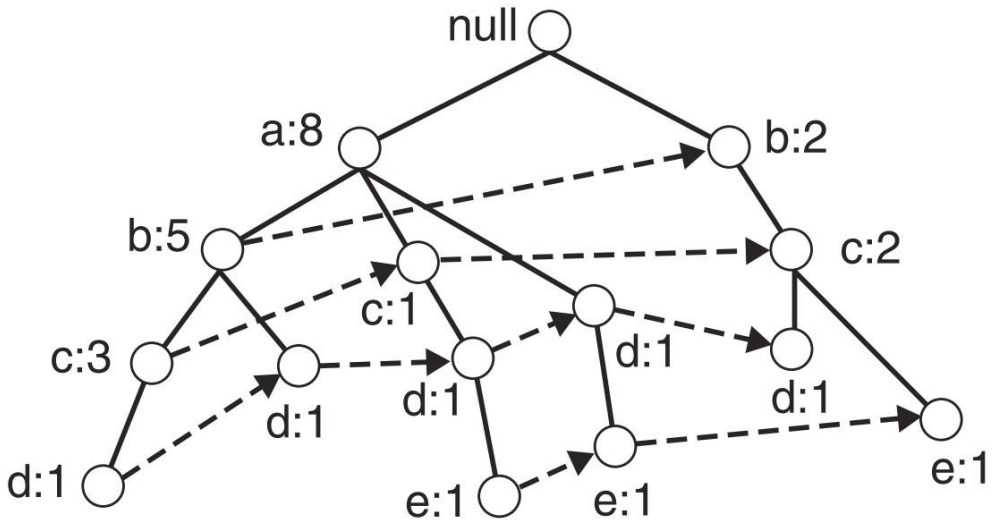
(b) Suffix tree.

$$\text{cover}(a \cdot d) = \text{cover}(d) \cap \text{cover}(a)$$
$$\text{cover}(b \cdot d) = \text{cover}(d) \cap \text{cover}(b)$$
$$\text{cover}(c \cdot d) = \text{cover}(d) \cap \text{cover}(c)$$

$$\text{cover}(a \cdot bd) = \text{cover}(b \cdot d) \cap \text{cover}(a \cdot d)$$
$$\text{cover}(a \cdot cd) = \text{cover}(c \cdot d) \cap \text{cover}(a \cdot d)$$
$$\text{cover}(b \cdot cd) = \text{cover}(c \cdot d) \cap \text{cover}(b \cdot d)$$

$$\text{cover}(a \cdot bcd) = \text{cover}(b \cdot cd) \cap \text{cover}(a \cdot cd)$$

FP-tree : une structure de données pour calculer les intersections en RAM (Random Access Memory)

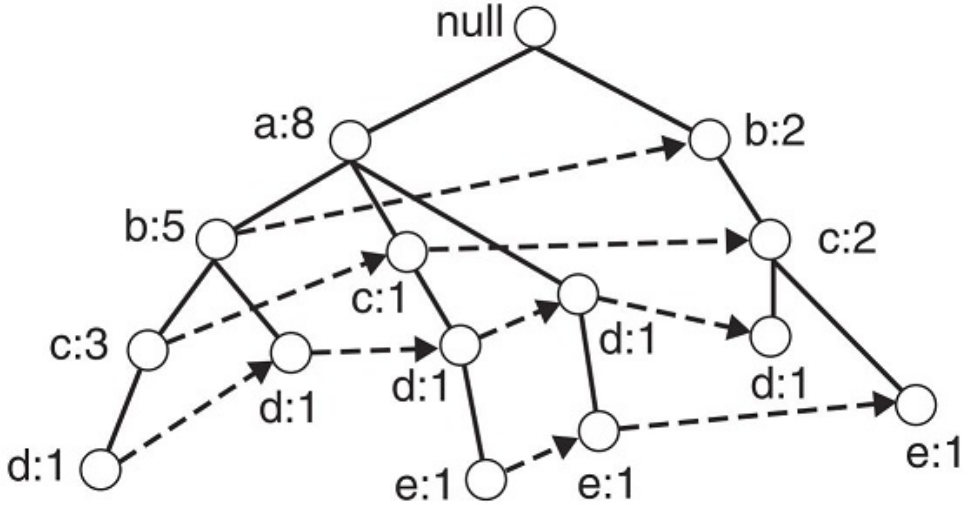


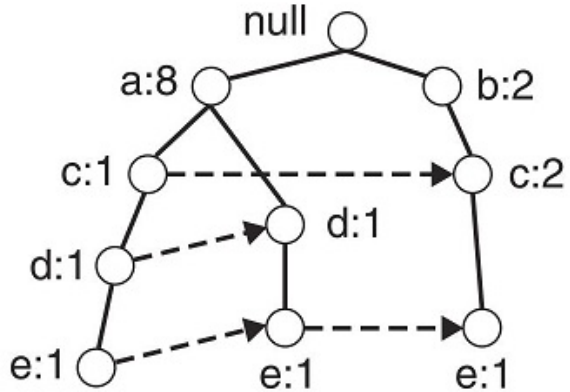
Transaction Data Set

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	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

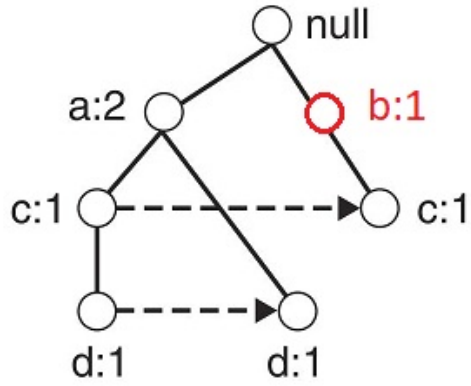
Parmi les 8 transactions qui contiennent *a*, il y en a 5 qui contiennent aussi *b*. Parmi ces 5 transactions qui incluent *ab*, il y en a 3 qui contiennent aussi *c*. Etc.

Exemple : les transactions avec suffixe *e*





(a) Prefix paths ending in e



(b) Conditional FP-tree for e