

# Les Règles d'Association

**MARKET DATA ANALYSIS**  
OU  
L'analyse du panier de la ménagère

**Ricco RAKOTOMALALA**

# Données de transaction (I)

## Analyse des tickets de caisse

N°transaction (Caddie)	Contenu du caddie			
1	pastis	martini	chips	saucisson
2	martini	chips		
3	pain	beurre	pastis	
4	saucisson			
5	pain	lait	beurre	
6	chips	pain		
7	confiture			

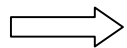
### Commentaires:

- » Une observation = Un caddie
- » Ne tenir compte que de la présence des produits (pas de leur quantité)
- » Nombre variable de produits dans un caddie
- » La liste des produits est immense !

### Objectifs :



- (1) Mettre en évidence les produits achetés ensemble
- (2) Transcrire la connaissance sous forme de règles d'association



Si antécédent Alors conséquent

listes de produits

Ex. Si pastis et martini Alors saucisson et chips

# Données de transaction (II)

Tableau de transactions → Tableau binaire 0/1

## Autre représentation des données de transactions

N° transaction (Caddie)	Contenu du caddie		
1	p1	p2	p3
2	p1		p3
3	p1	p2	p3
4	p1		p3
5		p2	p3
6			p4



Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1



Selon la granularité choisie, le nombre de colonnes peut être immense.  
(ex. détail par marques ou regroupement en familles → boîtes de cassoulet)

# Données de transaction (III)

Tableau individus x variables → Tableau binaire 0/1

## Codage disjonctif complet

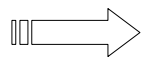
Observation	Taille	Corpulence
1	petit	mince
2	grand	enveloppé
3	grand	mince



Observation	Taille = petit	Taille = grand	Corpulence = mince	Corpulence = enveloppé
1	1	0	1	0
2	0	1	0	1
3	0	1	1	0



Dès que l'on peut se ramener à des données 0/1  
Il est possible de construire des règles d'association



Il s'agit de détecter les co-occurrences des modalités (attribut = valeur)  
Certaines associations sont impossibles par construction (ex. on ne peut pas être « petit » et « grand » en même temps)

# Critères d'évaluation des règles d'association

## Support et confiance

Soit la règle d'association  
R1 : Si p1 alors p2

### Données

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

SUPPORT : Un indicateur de « fiabilité » de la règle

en termes absolus  
 $\text{sup}(R1) = 2$  ou  $\text{sup}(R1) = 2/6 = 33\%$   
en termes relatifs

CONFIANCE : Un indicateur de « précision » de la règle

$$\begin{aligned} \text{conf}(R1) &= \frac{\text{sup}(R1)}{\text{sup}(\text{antécédent } R1)} \\ &= \frac{\text{sup}(p1 \rightarrow p2)}{\text{sup}(p1)} = \frac{2}{4} = 50\% \end{aligned}$$

⇒ « Bonne » règle = règle avec un support et une confiance élevée

# Extraction des règles d'association (I)

## Démarche

Paramètres : Fixer un degré d'exigence sur les règles à extraire

» Support min. (ex. 2 transactions)

» Confiance min. (ex. 75%)

→ L'idée est surtout de contrôler (limiter) le nombre de règles produites

Démarche : Construction en deux temps

» recherche des itemsets fréquents (support  $\geq$  support min.)

» à partir des itemsets fréquents, produire les règles (conf.  $\geq$  conf. min.)

Quelques définitions :

» item = produit

» itemset = ensemble de produits (ex. {p1,p3})

» sup(itemset) = nombre de transactions d'apparition simultanée des produits (ex. sup{p1,p3} = 4)

» card(itemset) = nombre de produits dans l'ensemble (ex. card{p1,p3} = 2)

# Extraction des Règles d'Association (II)

## Recherche des Itemsets Fréquents

Cas général :  $2^J - 1$

- » Nombre de calculs énormes
- » Chaque calcul impose de revenir scanner la base

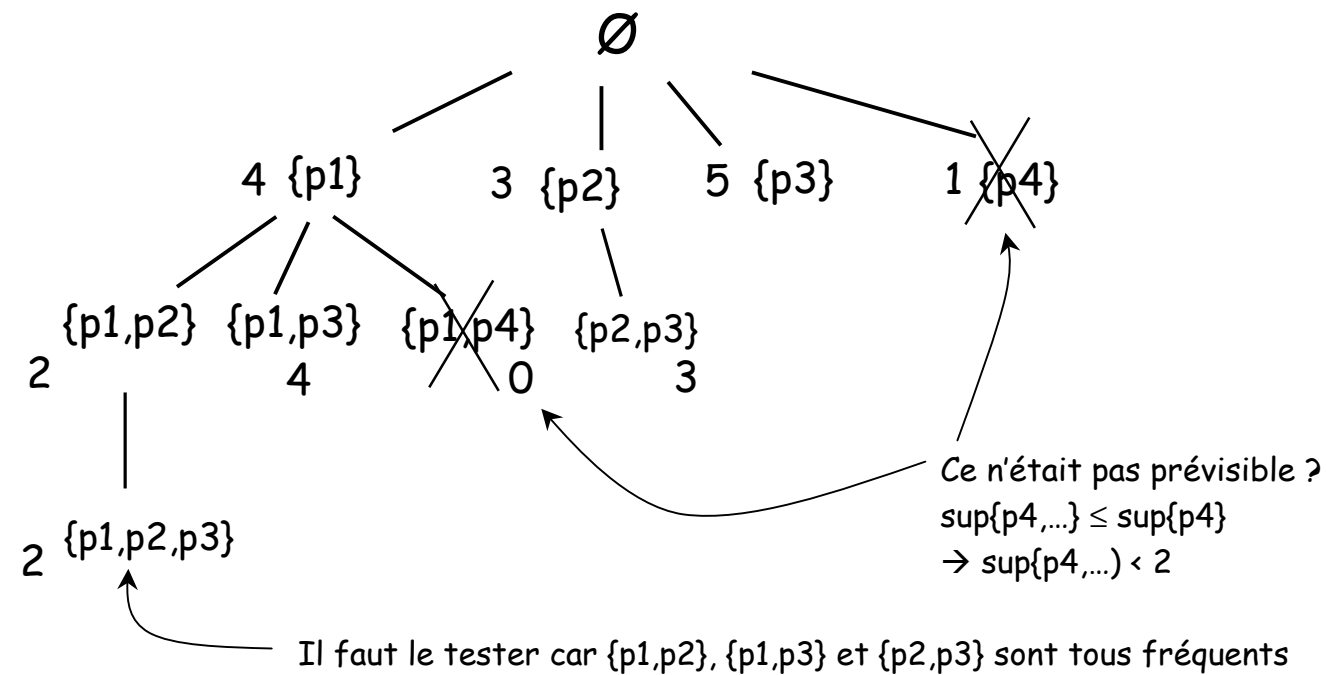
$$\begin{array}{r}
 C_4^1 = 4 \quad \leftarrow \text{Itemsets de card} = 1 \\
 C_4^2 = 6 \quad \leftarrow \text{Itemsets de card} = 2 \\
 C_4^3 = 4 \quad \leftarrow \text{Itemsets de card} = 3 \\
 C_4^4 = 1 \\
 \hline
 \Sigma = 15 = 2^4 - 1 \quad \dots
 \end{array}$$



Réduire l'exploration en éliminant d'emblée certaines pistes

### Données

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1



Que se passerait-il si nous avions sup. min. = 3 ?

# Extraction des Règles d'Association (III)

Recherche des règles pour les itemsets de card = 2



Il faut tester toutes les combinaisons : 2 tests par itemset

Données

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

{p1,p2}	$\left\{ \begin{array}{l} p1 \rightarrow p2 : \text{conf.} = 2/4 = 50\% \text{ (refusé)} \\ p2 \rightarrow p1 : \text{conf.} = 2/3 = 67\% \text{ (refusé)} \end{array} \right.$
{p1,p3}	$\left\{ \begin{array}{l} p1 \rightarrow p3 : \text{conf.} = 4/4 = 100\% \text{ (accepté)} \\ p3 \rightarrow p1 : \text{conf.} = 4/5 = 80\% \text{ (accepté)} \end{array} \right.$
{p2,p3}	$\left\{ \begin{array}{l} p2 \rightarrow p3 : \text{conf.} = 3/3 = 100\% \text{ (accepté)} \\ p3 \rightarrow p2 : \text{conf.} = 3/5 = 60\% \text{ (refusé)} \end{array} \right.$

Que se passerait-il si nous avions conf. min. = 55 %

# Extraction des Règles d'Association (IV)

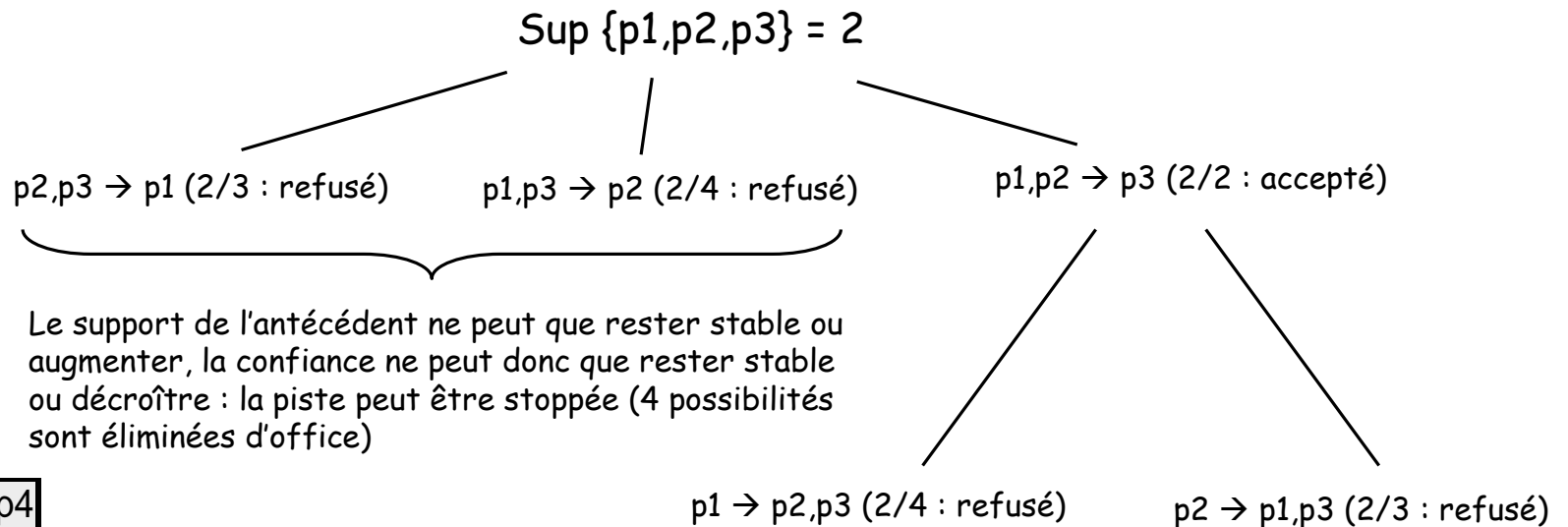
Recherche des règles pour les itemsets de card = 3 et plus...

$C_3^1 = 3$  ← Règles avec conséquent de card = 1

$C_3^2 = 3$  ← Règles avec conséquent de card = 2



Réduire l'exploration en éliminant d'emblée certaines pistes



## Données

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

Que se passerait-il si nous avions conf. min. = 55 %

# Un indicateur de pertinence des règles

## Dépasser le support et la confiance avec le LIFT

Support très élevé  
Confiance = 100%

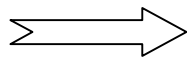
Que faut-il penser de la règle ?

Si cheveux = brun Alors cerveau = présent

La confiance en termes probabilistes

$$\begin{aligned} \text{conf}(A \rightarrow C) &= \frac{\text{sup}(AC)}{\text{sup}(A)} \\ &= \frac{P(AC)}{P(A)} \\ &= P(C / A) \end{aligned}$$

Le LIFT



$$\text{lift}(A \rightarrow C) = \frac{P(C / A)}{P(C)}$$

Support du conséquent en termes relatifs

Rapport de probabilité

S'interprète comme un odd-ratio (une cote)

S'interprète comme une « vraisemblance » de la règle

Lift  $\leq 1$   $\rightarrow$  La règle ne sert absolument à rien...

Interprétation :  $\text{LIFT}(\text{fumer} \rightarrow \text{cancer}) = 3\% / 1\% = 3$

En fumant, on a 3 fois plus de chances d'avoir un cancer.



Le LIFT ne peut être calculé qu'après coup pour filtrer les règles  
Nous ne pouvons pas l'utiliser pour guider l'apprentissage

# Des règles d'association aux motifs séquentiels

Introduire la date des transactions (ou du moins tenir compte de leur succession)

Peut-on produire des règles du type ?

Si « destruction véhicule » et « remboursement intégral » Alors « achat nouveau véhicule »

Étape 1

Étape 2

Étape 3

Données de transactions

Datées (au moins succession d'achats)

Clients	Achat 1	Achat 2	Achat 3	Achat 4
C1	(1, 2, 3)	(4, 2, 5)	(1, 6, 2)	(4, 1)
C2	(1, 3, 2)	(1, 2, 3)	(6, 3, 2)	
C3	(4, 8)	(1, 3, 7)	(5, 8)	(1, 4)
C4	(5, 2, 3)	(1, 2, 3)	(1, 2, 8)	(1, 6, 2)

Itemset et règles

Support  $\langle (1, 3) (2) (6, 2) \rangle = 3$  (ou  $\frac{3}{4} = 75\%$ )

Si  $(1, 3)$  Alors  $(2) (6, 2) \rightarrow$  confiance  $= \frac{3}{4} = 75\%$

Si  $(1, 3) (2)$  Alors  $(6, 2) \rightarrow$  confiance  $= \frac{3}{3} = 100\%$



Calculs très complexes, très peu de logiciels proposent cette approche

# Références

Très peu d'ouvrages consacrés spécifiquement à cette approche. Descriptifs succincts souvent dans des livres plus généraux (Lefebure & Venturi, etc.)

D'autres références sur internet (thèses de doctorat) et supports de cours. Les principales sont les articles de ARAWAL et MANILLA.

Quelques tutoriels (ORANGE, TANAGRA, WEKA) en ligne

<http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>

(section Règles d'Association et Comparaison de logiciels)