

# Les Règles d'Association

MARKET DATA ANALYSIS  
ou  
L'analyse du panier de la ménagère

Ricco RAKOTOMALALA

# Plan

1. Type de données traitées - Finalité de l'analyse
2. Recherche des « itemsets » fréquents
3. Construction des règles
4. Mesure d'évaluation des règles
5. Règles d'association et logiciels
6. Plus loin : recherche des motifs séquentiels

# Type de données traitées

## Finalité de l'extraction des règles d'association

# Données de transaction (I)

## Analyse des tickets de caisse

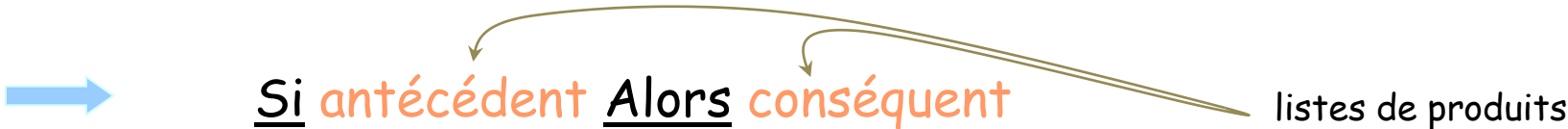
N° transaction (Caddie)		Contenu du caddie			
1	pastis	martini	chips	saucisson	
2	martini	chips			
3	pain	beurre	pastis		
4	saucisson				
5	pain	lait	beurre		
6	chips	pain			
7	confiture				

Commentaires: ou un ticket de caisse...

- » 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 règle d'association



Ex. Si pastis et martini Alors saucisson et chips

Rq : on ne traite (capte) pas les règles négatives. ex : s'il n'a pas acheté de vin alors il va acheter du soda...

# 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



ex. de regroupement en familles de produits : s'il achète un ouvre-boîte, alors il va acheter une boîte de conserve

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

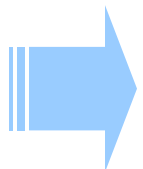
Observation	Taille	Corpulence
1	petit	mince
2	grand	enveloppé
3	grand	mince



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 cooccurrences des modalités (attribut = valeur)  
Certaines associations sont <sup>cependant</sup> 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

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

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

## Démarche globale

**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)
- » **itemset fréquent** = itemset dont le support est  $\geq$  à support min



# Recherche des itemsets fréquents

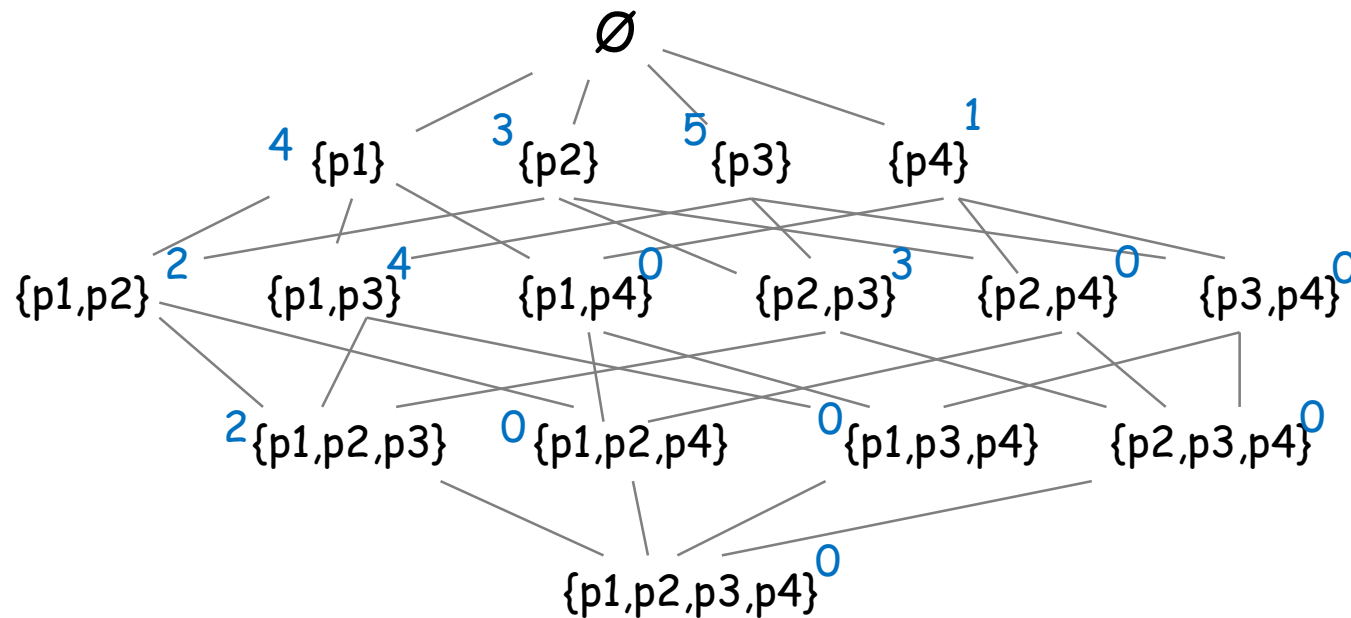
Idée : extraire tous les « itemsets fréquents » en minimisant les calculs, notamment le nombre d'accès à la base de données pour que le calcul soit réalisable sur de très grandes bases de données....

La recherche des itemsets fréquents peut être une finalité en elle-même c.-à-d. détecter les produits qui sont achetés simultanément

recherche de co-occurrence simple, sans s'intéresser à la causalité

# Extraction des itemsets fréquents

Il s'agit de parcourir un treillis et de calculer les supports associés à chaque combinaison



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

$$C_4^1 = 4$$

# itemsets  
de card = 1

$$C_4^2 = 6$$

# itemsets  
de card = 2

$$C_4^3 = 4$$

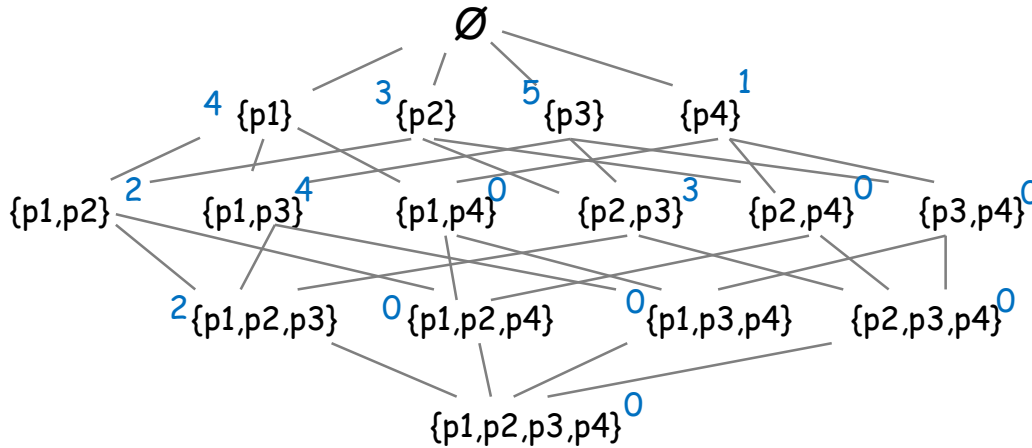
$$C_4^4 = 1$$

$$\Sigma = 15 = 2^4 - 1$$

Le nombre de configuration est très vite très élevé  
Chaque configuration nécessiterait un scan de la base de données  
→ Il faut s'appuyer sur le paramètre « support.min »  
→ Et les propriétés des itemsets  
Pour réduire le nombre de configuration à évaluer réellement

# Extraction des itemsets fréquents

## Quelques définitions



Sup.min = 2

**Itemset** : un ensemble d'items

**Superset** : B est un superset de A si  $\text{card}(A) < \text{card}(B)$  et  $A \subset B$   
→  $\text{sup}(B) \leq \text{sup}(A)$

**Itemset fréquent** :  
itemset dont le support  
est  $\geq$  à sup.min

Si un itemset n'est pas  
fréquent, tous ses supersets  
ne le seront pas non plus.



**Itemset fréquent fermé** :  
itemset fréquent dont  
aucun de ses supersets n'a  
un support identique (ex.  
{p1,p3} est fermé, {p1,p2}  
ne l'est pas)



**Itemset fréquent maximal** :  
itemset fréquent dont aucun  
de ses supersets n'est  
fréquent (ex. {p1,p2,p3} est  
maximal)

On prend usuellement comme  
point de départ les itemsets  
fréquents pour générer **toutes** les  
règles d'association.

Prendre comme point de départ les  
itemsets fréquents fermés pour générer  
les règles permet de **réduire le nombre  
de règles redondantes**

Ex.  $(p1 \rightarrow p2)$  et  $(p1 \rightarrow p2, p3)$  auront la  
même confiance, la 1<sup>ère</sup> est redondante  
par rapport à la 2<sup>ème</sup>.

Donne la **représentation la plus  
compacte possible de la liste des  
itemsets**. Ex. si on sait que {p1,p2,p3}  
est fréquent, on sait que {p1,p2},  
{p1,p3} et {p2,p3} le sont également  
(mais on ne connaît pas leur support)

# Extraction des itemsets fréquents

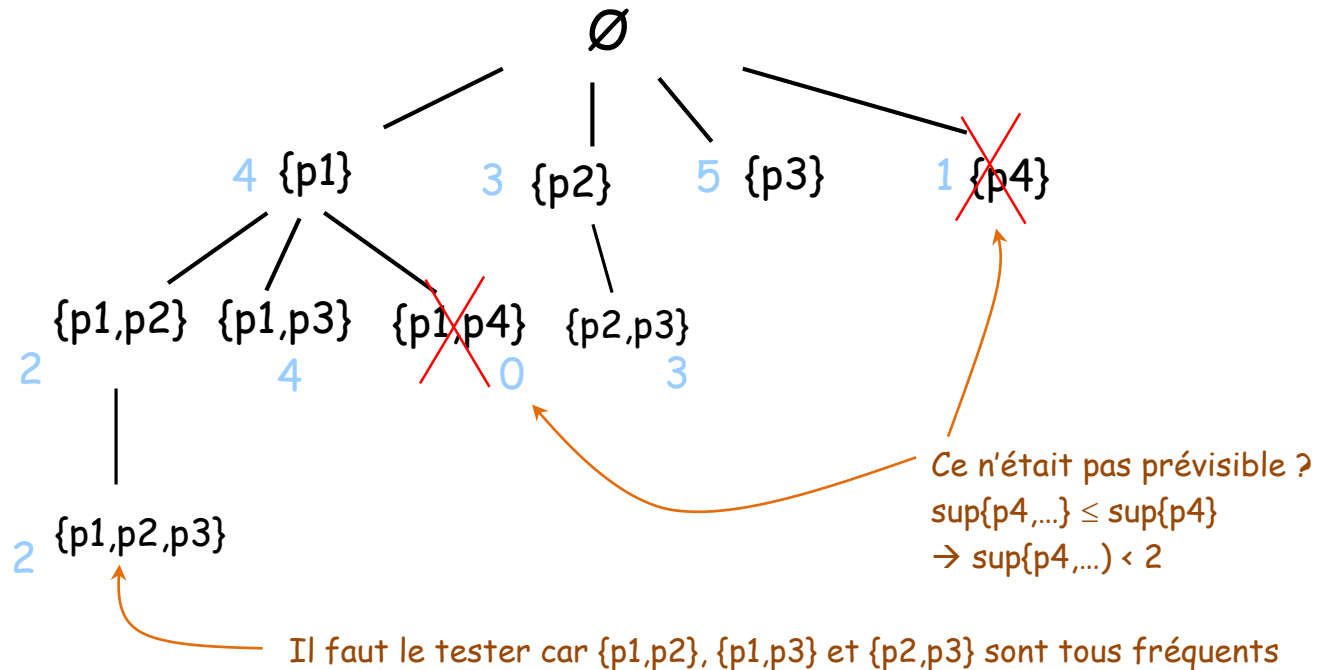
Une approche très simple

Réduire l'exploration en éliminant d'emblée certaines pistes via le support min (sup.min = 2 ici) et les propriétés des itemsets

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

## 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

Idée : déduire les règles à partir des itemsets fréquents

On limite la prolifération des règles en utilisant le critère confiance min.

éviter les règles redondantes....

# Extraction des Règles d'Association

## Recherche des règles pour les itemsets de card = 2



Il faut tester toutes les combinaisons : 2 tests par itemset

Tous les supports sont dispos dans le treillis, pas besoin de scanner la base

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

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

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

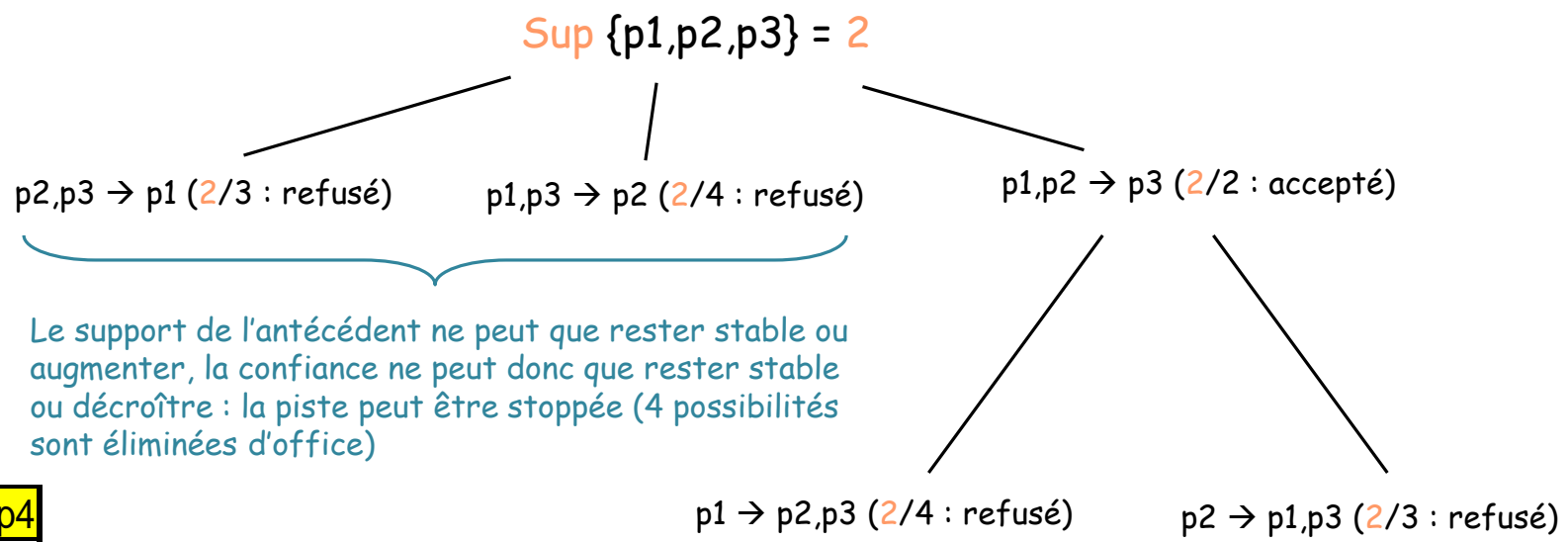
# Extraction des Règles d'Association

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



Réduire l'exploration en éliminant d'emblée certaines pistes  
Le support de la règle ne change jamais = support de l'itemset  
On peut jouer sur le support de l'antécédent

$C_3^1 = 3$  ← Règles avec conséquent de card = 1  
 $C_3^2 = 3$  ← Règles avec conséquent de card = 2



## 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 %



# Mesures d'évaluation des règles

Aller au-delà du support et de la confiance



$R : p_3 \rightarrow p_1$

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 (en  
termes relatifs)

$\text{sup}(R) = 4/6 = P(p_3 p_1)$

Proba. conjointe

Proba. conditionnelle

Confiance

$\text{conf}(R) = \text{sup}(\{p_1, p_3\}) / \text{sup}(\{p_3\}) = 4 / 5 = \frac{P(p_3 / p_1)}{P(p_1 / p_3)}$

Mais

Une règle peut avoir d'excellents supports et  
confiance sans être pour autant « intéressante »

Si Sexe = Masculin Alors Cerveau = présent

Support = 50%  
Confiance = 100%



Il faut un critère - une mesure d'intérêt - qui caractérise une forme de causalité c.-à-d. l'idée « la connaissance de l'antécédent amène de l'information (supplémentaire) sur la connaissance du conséquent »

# Un indicateur de pertinence des règles

Dépasser le support et la confiance avec le LIFT

R : Antécédent  $\rightarrow$  Conséquent

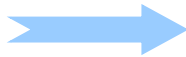
On sait que

$$\left\{ \begin{array}{l} \text{sup}(A) = P(A) \\ \text{sup}(C) = P(C) \\ \text{sup}(A \rightarrow C) = \text{sup}(AC) = P(AC) \\ \text{conf}(A \rightarrow C) = P(C / A) \end{array} \right.$$

Support de l'antécédent

Support du conséquent

Le LIFT



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

Ex. LIFT(fumer  $\rightarrow$  cancer) = 3% / 1% = 3

Rapport de probabilité - Surcroît d'occurrence du conséquent quand l'antécédent est présent

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

~ fumer multiplie par 3 la survenue du cancer (~ proche de la notion de risque relatif)



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

Remarque :

Le LIFT peut se lire également comme un rapport de vraisemblance  $\rightarrow$  sous  $H_0$  : A et B sont indépendants

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

# Autres indicateurs de pertinence des règles

Données

$R : p3 \rightarrow p1$

Le point de départ est un tableau croisé

	Antécédent	Non(Antécédent)	Total
Conséquent	$n_{ac} = 4$ exemples		$n_c = 4$
Non(conséquent)	contre-exemples		
Total	$n_a = 5$		$n = 6$

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

Quelques mesures

Mesure	Formule
Support	$\frac{n_{ac}}{n}$
Confiance	$\frac{n_{ac}}{n_a}$
Lift	$\left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c}{n}\right)$
<u>Leverage</u>	$\frac{n_{ac}}{n} - \frac{n_a}{n} \times \frac{n_c}{n}$
Importance	$\ln \left[ \left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c - n_{ac}}{n - n_a}\right) \right]$
Conviction	$\frac{n_a \times (n - n_c)}{n \times (n_a - n_{ac})}$
Surprise	$\left(\frac{n_{ac}}{n} - \frac{n_a - n_{ac}}{n}\right) / \left(\frac{n_c}{n}\right)$

développé par  
Sergey Brin

- Il faut pouvoir les interpréter.
- Mis à part la confiance et le support, elles interviennent uniquement après coup pour trier les règles.
- Dans certaines configurations (ex. contre-exemples = 0), les mesures ne sont pas calculables. On peut s'en sortir avec l'estimation laplacienne des probabilités.
- Aucune n'est vraiment décisive parce que la notion de « règle intéressante » est difficile à situer.

# Les règles d'association dans les logiciels

Les logiciels s'appuient sur différents algorithmes  
A PRIORI, ECLAT, FP-GROWTH, etc.

Méthode présente dans tous les outils estampillés  
« Data Mining »

# Règles d'association dans la distribution SIPINA

L'outil d'extraction de règle - Interaction avec Excel

Le logiciel peut s'interfacer avec Excel

Association rule software - [Learning set editor]

Association rule software - [Association rule viewer (Beta = 1.00)]

Id	Antecedent	Consequent	Length	Support	Conf...	Recall	F-me...	Lift	Convi...
1	AA_Cell_Batteries & Eggs	2pct_Milk	3	0.0169	0.8519	0.1544	0.2614	7.7810	5.1617
2	Apple_Jelly & Wheat_Bread	2pct_Milk	3	0.0176	0.9231	0.1611	0.2743	8.4316	8.3062
3	Apples & Onions	2pct_Milk	3	0.0191	0.8667	0.1745	0.2905	7.9163	5.6957
4	Apples & Potato_Chips	2pct_Milk	3	0.0228	0.8611	0.2081	0.3351	7.8656	5.6363
5	Bananas & Onions	2pct_Milk	3	0.0206	0.8750	0.1879	0.3094	7.9924	6.0517
6	Bananas & Wheat_Bread	2pct_Milk	3	0.0220	0.8571	0.2013	0.3261	7.8293	5.4880
7	Cantaloupe & Pepperoni_Pizza_-Frozen	2pct_Milk	3	0.0169	0.8519	0.1544	0.2614	7.7810	5.1617

Les règles peuvent être facilement récupérées dans Excel pour être post-traitées (filtrer, trier selon les mesures...)

Tutoriel Tanagra, « [Associations dans la distribution SIPINA](#) », avril 2013.

# Règles d'association avec TANAGRA (I)

## Interaction avec Excel

Plusieurs mesures sont disponibles, un tutoriel spécifique y est consacré

The screenshot shows the TANAGRA 1.4.50 software interface. On the left, a Microsoft Excel window (bpress\_discrete.xlsx) is open, displaying a dataset with columns A and B. A red dashed arrow points from the 'Execute Tanagra' button in the Excel ribbon to the TANAGRA window. The TANAGRA window has a menu bar (File, Diagram, Component, Window, Help) and a toolbar. The 'Analysis' pane on the left shows a tree structure with 'Dataset (tanC6.txt)', 'Define status 1', and 'A priori MR 1'. The main 'Rules evaluation' table displays the following data:

N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise
1	"BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"AGNEAUSD=non" - "taille=infreqMed"	360	132	174	119	0.33056	0.90152	1.86520	0.15333	1.31833	5.24615	0.60920
2	"BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"taille=infreqMed"	360	132	183	124	0.34444	0.93939	1.84799	0.15806	1.28929	8.11250	0.63388
3	"AGNEAUSD=non" - "BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"taille=infreqMed"	360	127	183	119	0.33056	0.93701	1.84329	0.15123	1.22709	7.80521	0.60656
4	"BEURRE=FREQUENTLY" - "AGNEAUSD=non" - "SEXE=FEMALE"	"taille=infreqMed"	360	152	153	119	0.33056	0.78289	1.84211	0.15111	1.56642	2.64848	0.56209
5	"AGNEAUSD=non" - "BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"taille=infreqMed"	360	153	152	119	0.33056	0.77778	1.84211	0.15111	1.58490	2.60000	0.55921

Below the table, the 'Components' section shows various analysis options: Data visualization, Statistics, Nonparametric statistics, Instance selection, Feature construction, Feature selection, Regression, Factorial analysis, PLS, Clustering, Spv learning, Meta-spv learning, Spv learning assessment, Scoring, and Association. The 'Association' component is selected. At the bottom, a row of icons represents different analysis methods: A priori, A priori MR, A priori PT, Assoc Outlier, Frequent Itemsets, Spv Assoc Rule, and Spv Assoc Tree.

Les règles sont facilement récupérables dans Excel pour être filtrées et triées

## Tutoriels

- « Règles d'association - Orange, Tanagra et Weka », avril 2008.
- « Règles d'association - Comparaison de logiciels », novembre 2008.
- « Règles d'association avec APRIORI PT », avril 2008.

# Règles d'association avec TANAGRA (II)

## Règles d'association supervisée

TANAGRA 1.4.50 - [Dataset (tanC6.txt)]

File Diagram Component Window Help

Analysis

- Dataset (tanC6.txt)
- Define status 1
- Spv Assoc Tree 1

Download information

Association tree spv

Parameters

Support : 0.05

Confidence : 0.6

Max card itemsets : 4

Lift : 1

Learning set ratio : 1

Class value : HIGH

OK Cancel Help

Définir une modalité cible

Define attribute statuses

Parameters

Attributes :

- TYPELAIT
- FRITURES
- BOEUFSD
- PORCSD
- VOLAILLES
- POISSONS
- AGNEAUSD
- AUTREVIANDES
- FROMAGES
- EOUFSD
- REPASVIANDES
- SELALIMENT
- SFI CONSO

Target

RISQUEATTAQUE

Clear all Clear selected

OK Cancel Help

Définir une variable cible

Spv Assoc Tree 1

Rules

"RISQUEATTAQUE" is "HIGH" -- IF ...

N°	Antecedent	Length	Support	Confidence	Lift
1	AGNEAUSD=non - POISSONS=non - AGE=63-72	3	0.056 ( 0.00 )	0.625 ( 0.00 )	2.922 ( 0.00 )
2	HEURESOMMEIL=sup8h - taille=infeqMed - AGE=63-72	3	0.094 ( 0.00 )	0.607 ( 0.00 )	2.839 ( 0.00 )
3	HEURESOMMEIL=sup8h - SEXE=FEMALE - AGE=63-72	3	0.078 ( 0.00 )	0.636 ( 0.00 )	2.975 ( 0.00 )
4	HEURESOMMEIL=sup8h - POISSONS=non - AGE=63-72	3	0.050 ( 0.00 )	0.621 ( 0.00 )	2.902 ( 0.00 )
5	HEURESOMMEIL=sup8h - HAB_BOISSON=NEVER - AGE=63-72	3	0.058 ( 0.00 )	0.600 ( 0.00 )	2.805 ( 0.00 )
6	BOEUFSD=moderement - POISSONS=non - AGE=63-72	3	0.053 ( 0.00 )	0.613 ( 0.00 )	2.866 ( 0.00 )
7	BEURRE=FREQUENTLY - ACTIVITESPORT=NEVER - AGE=63-72	3	0.075 ( 0.00 )	0.614 ( 0.00 )	2.869 ( 0.00 )
8	BEURRE=FREQUENTLY - POISSONS=non - AGE=63-72	3	0.056 ( 0.00 )	0.690 ( 0.00 )	3.224 ( 0.00 )
9	VOLAILLES=moderement - SEXE=FEMALE - AGE=63-72	3	0.081 ( 0.00 )	0.604 ( 0.00 )	2.825 ( 0.00 )
10	nbanneescol=infq2 - taille=infeqMed - AGE=63-72	3	0.069 ( 0.00 )	0.625 ( 0.00 )	2.922 ( 0.00 )
11	nbanneescol=infq2 - SEXE=FEMALE - AGE=63-72	3	0.056 ( 0.00 )	0.645 ( 0.00 )	3.016 ( 0.00 )
12	AUTREVIANDES=non - taille=infeqMed - AGE=63-72	3	0.086 ( 0.00 )	0.608 ( 0.00 )	2.842 ( 0.00 )

Peut être utilisé pour caractériser un groupe issu d'un clustering par ex.

Tutoriels

- « Règles d'association prédictives », février 2009.
- « Règles d'association "supervisées" », avril 2008.



# Règles d'association avec R

Le package « [arules](#) »

(l'équivalent existe sous Python -

<http://tutoriels-data-mining.blogspot.com/2019/02/regles-d-assoc-iation-sous-python.html>)

Extraction de  
différents types  
d'itemsets  
fréquents

```
itemset_mining.r
1 #clear the memory
2 rm(list=ls())
3 #importing the dataset
4 library(xlsReadWrite)
5 dataset <- read.xls(file="itemset_mining.xls", colNames=T, sheet=1)
6 print(dataset)
7 #loading arule library
8 library(arules)
9 #extracting the frequent itemsets
10 params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="frequent itemsets")
11 result <- apriori(as.matrix(dataset), parameter = params)
12 inspect(result)
13 #extracting the closed itemsets
14 params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="closed frequent itemsets")
15 result <- apriori(as.matrix(dataset), parameter = params)
16 inspect(result)
17 #extracting the maximally itemsets
18 params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="maximally frequent itemsets")
19 result <- apriori(as.matrix(dataset), parameter = params)
20 inspect(result)
```

Extraction et  
visualisation des  
règles

```
assoc rule on german.r
1 #charger le package
2 library(arules)
3 #charger le fichier de données
4 german <- read.table(file="credit-german.txt", header=T, dec=".", sep="\t")
5 summary(german)
6 #transformer les données attributs-variables
7 #en données transactionnelles
8 german.trans <- as(german, "transactions")
9 summary(german.trans)
10 #extraction des règles
11 german.regles <- apriori(german.trans, parameter=
12     list(supp=0.25, conf=0.75, minlen=2, maxlen=10, target="rules"))
13 summary(german.regles)
14 #afficher les 10 premières règles trouvées
15 inspect(german.regles[1:10])
16 #afficher les 5 règles avec le lift le + élevé
17 regles.triees <- sort(german.regles, by="lift")
18 inspect(regles.triees[1:5])
```





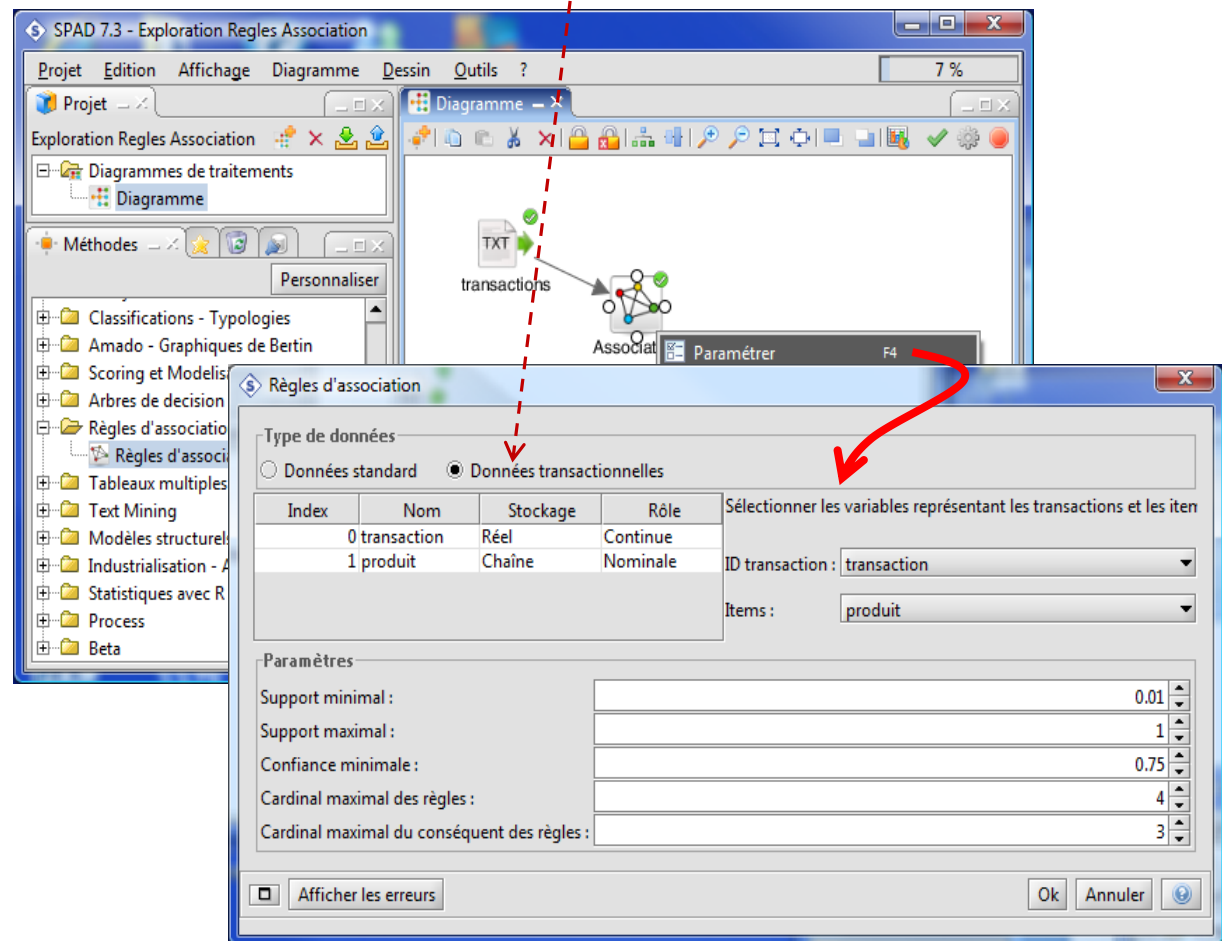
# Règles d'association avec SPAD

Peut traiter indifféremment les bases « individus x variables » et les bases transactionnelles

Spécification du format de données pour le traitement

## Format « transactions »

transaction	produit
1	B
1	E
1	H
2	A
2	B
2	E
2	F
3	B
3	C
3	F
3	H



Un outil interactif permet de filtrer et trier les règles

# Extraction des motifs séquentiels

Intégrer des contraintes temporelles (succession)  
dans la recherche des règles

# 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

Datées (au moins succession d'achats)

Données de transactions

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 < (1, 3) (2) (6, 2) > = 3 (ou  $\frac{3}{4} = 75\%$ )  
Si (1, 3) Alors (2) (6, 2) → confiance =  $\frac{3}{4} = 75\%$   
Si (1, 3) (2) Alors (6, 2) → confiance =  $3/3 = 100\%$



libres  
Calculs très complexes, très peu de logiciels proposent cette approche  
<http://himalaya-tools.sourceforge.net/Spam/>

# Références

Wan Aezwani Wab, « [Apriori and Eclat Algorithms in Association Rule Mining](#) », Slideshare, Avril 2014.

P. Tan, M. Steinbach, V. Kumar, « [Association Analysis: Basic Concepts and Algorithms](#) », chapitre 6 de l'ouvrage « Introduction to Data Mining », 2005.

Tutoriels Tanagra consacrés aux [Règles d'Association](#) (mise en œuvre, les différentes études possibles, comparaison des logiciels).