

DATA MINING

LES RÈGLES ASSOCIATIVES

Mohamed Heny SELMI

medheny.selmi@esprit.tn

OBJECTIFS

- ✓ Rechercher les associations consiste à rechercher les règles de type :
« Si pour un individu, la variable $A = X_a$, la variable $B = X_b$, etc, alors, dans 80% des cas, la variable $Z = X_z$, cette configuration se rencontrant pour 30 % des individus »
- ✓ **Repérer** des règles liant les données avec un bon niveau de probabilité
 - découverte de relations fines entre attributs (ou variables)
 - généralisation des dépendances fonctionnelles
- ✓ Mettre en évidence les produits / des articles achetés ensemble
- ✓ **Transcrire la connaissance** sous forme de règles d'association
- ✓ Règles du style : $\langle \underline{\text{si}} [P(tid, X) := \text{prémisse}] \underline{\text{alors}} [P(tid, Y) := \text{conséquence}] \rangle$
- ✓ Différents types de règles
 - origine « panier de la ménagère »
 - étendues aux tables multiples et aux attributs continus

ANALYSE DES TICKETS DE CAISSE

N° Transaction (Caddie)	Contenu du caddie			
1	Poulet	Moutarde	Œufs	Pates
2	Moutarde	Œufs		
3	Pain	Beurre	Poulet	
4	Pates			
5	Pain	Lait	Beurre	
6	Œufs	Pain		
7	Confiture			



Une observation = un caddie

Ne tenir compte que de la présence des produits : peu importe leur quantité

Dans un caddie : le nombre de produits est variables

La liste des produits est immense et variable

TABLEAU DES TRANSACTIONS

- ✓ Mettre en évidence les produits / des articles achetés ensemble
- ✓ Transcrire la connaissance sous forme de règles d'association

si $[P(tid, X) := \text{prémisse}]$ alors $[P(tid, Y) := \text{conséquence}]$

N° Transaction (Caddie)	Contenu du caddie			
1	Poulet	Moutarde	Œufs	Pates

si Poulet et Moutarde alors Œufs et Pates

N° Transaction (Caddie)	Contenu du caddie	
6	Œufs	Pain

si Œufs alors Pain

TABLEAU BINAIRE

N° Transaction (Caddie)	Contenu du caddie			
1	Poulet	Moutarde	Œufs	Pates
2	Moutarde	Œufs		
3	Pain	Beurre	Poulet	
4	Pates			
5	Pain	Lait	Beurre	
6	Œufs	Pain		
7	Confiture			



	P1	P2	P3	P4	P5	P6	P7	P8
1	1	1	1	1	0	0	0	0
2	0	1	1	0	0	0	0	0
3	1	0	0	0	1	1	0	0
4	0	0	0	1	0	0	0	0
5	0	0	0	0	1	1	1	0
6	0	0	1	0	1	0	0	0
7	0	0	0	0	0	0	0	1

désignation

P1 = Poulet

P2 = Moutarde

P3 = Œufs

P4 = Pates

P5 = Pain

P6 = Beurre

P7 = Lait

P8 = Confiture

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

PASSAGE EN FORME DISJONCTIVE COMPLÈTE

- **Catégoriel, qualitatif, discret** : type marché, entreprises, taux, appartenance, ...

	Marché
c ₁	Part.
c ₂	Autre
c ₃	Part.
c ₄	Part.
c ₅	Entr.



	P	E	A
c ₁	1		
c ₂			1
c ₃	1		
c ₄	1		
c ₅		1	

- **Continu, quantitatif** : virement, âge, température, consommation, pourcentage, ...

	Dom
c ₁	1100
c ₂	0
c ₃	2200
c ₄	800
c ₅	3800



	d0	d1	d2	d3	d4
c ₁			1		
c ₂	1				
c ₃				1	
c ₄		1			
c ₅					1

CRITÈRES D'ÉVALUATION DES RÈGLES D'ASSOCIATION

SUPPORT

CONFIANCE

Règle d'association :
 $p1 \rightarrow p2$

- ✓ indicateur de « **fiabilité** »
- ✓ probabilité absolue :
 $P(X \cup Y)$
- ✓ $\frac{||X \cup Y||}{||BD||} = \%$
de transactions vérifiant la règle

- ✓ Indicateur de « **précision** »
- ✓ probabilité conditionnelle :
 $P(Y/X)$
- ✓ $\frac{||X \cup Y||}{||X||} = \%$
de transactions vérifiant l'implication

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

$Conf(R1) = sup(R1) / sup(\text{antécédant } R1)$
 $= sup(p1 \rightarrow p2) / sup(p1) = 2 / 4 = 50\%$

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



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

ANALYSE DES TICKETS

{ "crème" } \Rightarrow { "pain" }



ID	PRODUITS
1	pain, crème, eau
2	crème
3	pain, crème, vin
4	eau
5	crème, eau

Support = Prob. (crème et pain) :

$$\text{Sup} = \frac{\text{nom}(\text{tran. contenant crème et pain})}{\text{nom_total}(\text{tran.})} = \frac{2}{5} = 0.4$$

Confiance = Prob(crème et pain / crème) :

$$\text{Conf} = \frac{\text{nom}(\text{tran. contenant crème et pain})}{\text{nom}(\text{tran. contenant crème})} = \frac{2}{4} = 0.5 = \frac{\text{sup}(\text{crème et pain})}{\text{sup}(\text{crème})}$$

DÉMARCHE D'EXTRACTION DES RÈGLES D'ASSOCIATION

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

- ✓ Support min. (exp. 2 transactions)
- ✓ Confiance min. (exp. 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.)

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 définitions :

- item = produit
- itemset = ensemble de produits (ex. $\{p1, p3\}$)
- $\text{sup}(\text{itemset})$ = nombre de transactions d'apparition simultanée des produits (ex. $\text{sup}\{p1, p3\} = 4$)
- $\text{card}(\text{itemset})$ = nombre de produits dans l'ensemble (ex. $\text{card}\{p1, p3\} = 2$)

ALGORITHME APRIORI [AGRAWAL94]

Première passe :

- ✓ recherche des 1-ensembles fréquents
- ✓ un compteur par produits

L'algorithme génère un candidat de taille k à partir de deux candidats de taille $k-1$ différents par le dernier élément

- ✓ procédure apriori-gen

Passe k :

- ✓ comptage des k -ensemble fréquents candidats
- ✓ sélection des bons candidats

APRIORI – FRÉQUENTS ITEMSETS

$L1 = \{ \text{frequent 1-ensemble} \} ;$

for ($k = 2 ; L_{k-1} \neq \emptyset ; k++$) do

{

$C_k = \text{apriori-gen}(L_{k-1}) ; // \text{Generate new candidates}$

foreach transactions $t \in \text{DB}$ do

{ // Counting

$C_t = \{ \text{subset}(C_k, t) \} ; // \text{get subsets of } t \text{ candidates}$

foreach $c \in C_t$ do $c.\text{count}++ ;$

}

$L_k = \{ c \in C_k \mid c.\text{count} \geq \text{minsup} \} ; // \text{Filter candidates}$

}

$F = \{L_k\} ;$

SIMULATION



On va prendre la valeur du Support minimal = **3**

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



1-Itemsets	Support
{p1}	4
{p2}	3
{p3}	5
{p4}	1

$L1 = \{ \{p1\}, \{p2\}, \{p3\} \}$

$C2 = \{ \{p1,p2\}, \{p1,p3\}, \{p2,p3\} \}$



SIMULATION



On va prendre le Support minimal = 3

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



2-Itemsets	Support
{p1,p2}	2
{p1,p3}	4
{p2,p3}	3

$$L_2 = \{ \{p1,p3\}, \{p2,p3\} \}$$

$$C_3 = \{ \{p1,p2,p3\} \}$$



SIMULATION



On va prendre le Support minimal = 3

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



3-Itemsets	Support
{p1,p2, p3}	2

$$L_3 = \emptyset$$

$$C_4 = \emptyset$$

$$F = \{ \{p1\}, \{p2\}, \{p3\}, \{p1,p3\}, \{p2,p3\} \}$$

SIMULATION



On va prendre le pourcentage de la confiance minimale = **65%**

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,p3}

p1→p3 : confiance = 4/4 = **100 %**

p3→p1 : confiance = 4/5 = **80 %**



{p2,p3}

p2→p3 : confiance = 3/3 = **100 %**

p3→p2 : confiance = 3/5 = **60 %**



INDICATEUR DE PERTINENCE DES RÈGLES

MESURE D'INTÉRÊT : *LIFT D'UNE RÈGLE*

- L'amélioration apportée par une règle, par rapport à une réponse au hasard est appelée « **lift** » et vaut :

$$\text{lift (règle)} = \text{confiance (règle)} / p(\text{résultat})$$

$$= p(\text{condition et résultat}) / [p(\text{condition}) \times p(\text{résultat})]$$

- Le lift est une bonne mesure de performance de la règle d'association.
- Le lift est la confiance de la règle divisée par la valeur espérée de la confiance.

Interprétation du lift :

- Un lift supérieur à 1 : Indique une corrélation positive
- Un lift de 1 indique une corrélation nulle
- Un lift inférieur à 1 : Indique une corrélation négative

T26	A	B	C	D	E
T1245	B	C	E	F	
T156	B	E			
T2356	A	B	D		
T145	C	D			

$$\text{lift}(C \rightarrow B) = 5/6 < 1 \quad \text{✗}$$

$$\text{lift}(B \rightarrow E) = 6/5 > 1 \quad \text{✓}$$

ETUDE DE CAS DE RECHERCHE D'ASSOCIATIONS INTÉRESSANTES

- ✓ Le principe de l'algorithme est de rechercher l'ensemble L_1 de tous les items apparaissant dans au moins $S_{\min} \times m$ transactions.
- ✓ Puis, parmi C_2 qui est le produit cartésien de L_1 avec lui-même, on construit l'ensemble L_2 de tous les couples d'items apparaissant dans au moins $S_{\min} \times m$ transactions.
- ✓ L'algorithme s'arrête quand L_k est vide.

UTILITÉ DES RÈGLES D'ASSOCIATION



La distribution réussit à mieux cibler ses mailings



La course à la fidélisation des clients



Réductions personnalisées à la caisse



Profil-client



Le test des nouveaux produits



Le panier moyen



Le parcours magasin



Cartes de fidélité