

02-04

Règles d'association

**NOUS ÉCLAIRONS.
VOUS BRILLENZ.**

FORMATION CONTINUE
ET SERVICES AUX ENTREPRISES



Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références

Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références

Tout commence dans la grande distribution ...



Modèle “Market-basket”

Problématique: comment identifier les articles (ou items en anglais) achetés ensemble par un nombre suffisamment grand de clients ?



Pains à hot dog

+



Saucisses à hot dog



Moutarde

Modèle “Market-basket”

- Tous les paniers d'épicerie sont enregistrés aux caisses (lecture des codes à barre)
- Pour une entreprise de grande distribution, traitement sur une grande échelle des données des différents points de ventes
- **Comment apprendre les associations les plus courantes ?**



Exemple d'une règle d'association

Ainsi, il est possible de découvrir des associations surprenantes !



Lait

+



Couches bébé



Bière

La preuve au IGA du coin ...



Bières !

Couches




En provenance du lait ...

Modèle “Market-basket”

- Un vaste ensemble d'**articles** (ou items)
Exemple: articles vendus dans une grande surface
- Un vaste ensemble de **paniers** (baskets)
Exemple: tous les articles achetés par un même client le même jour
- Nous voulons découvrir les **règles d'association** (association rules)
Exemple: les clients achetant $\{ x, y, z \}$ ont tendance à acheter $\{ v, w \}$

Modèle “Market-basket”

Transactions

TID	Paniers (liste d'articles)
1	Pain, Coca-Cola, Lait 
2	Bières, Pain 
3	Bière, Coca-Cola, Couches, Lait 
4	Bière, Pain, Couches, Lait 
5	Coca-Cola, Couches, Lait 

Règles d'association

{ Lait } → { Coca-Cola }

{ Couches, Lait } → { Bière }

Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références

Application: grande distribution

- **Articles / Items** = produits ou articles en vente
- **Paniers / Baskets** = ensemble des produits achetés par un client (panier d'épicerie)
- Les commerces conservent les données (articles achetés ensembles) liées aux transactions
 - Aide à comprendre le déplacement des clients au sein du magasin. Positionnement des rayons et des articles
 - Permet de découvrir les tie-in “tricks”, par exemple: promotion sur les couches, mais augmentation du prix des bières
 - Publicité sur les commerces en ligne
- Exemple d'Amazon: les clients ayant acheté **X** ont aussi acheté **Y**

Application: détection de plagiat

- **Items** = documents
- **Baskets** = phrases
- Les items (documents) apparaissant ensemble trop souvent peuvent indiquer du plagiat
- **Les items n'ont pas à être "inclus" (dans le sens des ensembles) dans les baskets**

Application: pharmacologie

- **Items** = médicaments et effets secondaires
- **Baskets** = patients ayant reçu un ou plusieurs médicaments
- Permet la détection d'effets secondaires induits par la combinaison de plusieurs médicaments
- Dans ce cas, il est également important de noter l'absence, comme la présence d'un item

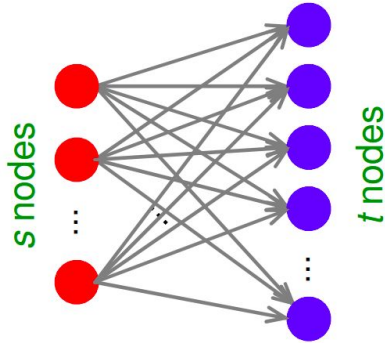
Application: trading algorithmique

- **Items** = valeur des actions
- **Baskets** = une période donnée
- Association de différents signaux à une tendance (bull/bear)
- Dans ce cas plus complexe, il faut tenir compte aussi de la chronologie (Temporal Association Rule Mining)



Application: analyse de réseaux sociaux

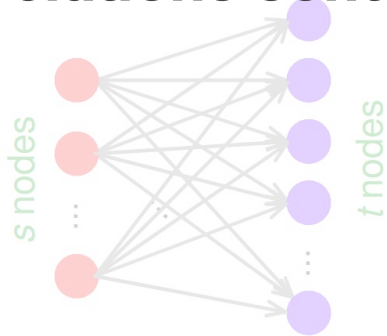
- **Items** = voisins
- **Baskets** = noeuds
- Recherche de sous-graphes bipartites $K_{s,t}$ d'un graphe



Application: analyse de réseaux sociaux

- **Items** = voisins
- **Baskets** = noeuds
- D'une manière générale, on cherche à **apprendre** une relation **plusieurs-à-plusieurs** entre deux sortes de choses

Les associations sont recherchées entre items, et non entre baskets



Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références

Itemsets fréquents

- Les **itemsets fréquents** (frequent itemsets) sont les items apparaissant “fréquemment” ensemble dans les baskets
- Le **support** d'un itemset I est le nombre de baskets contenant tous les items de I
- Le support peut aussi être exprimé en pourcentage:
 - Rapport entre le nombre de baskets contenant tous les items de I et le nombre total de baskets

TID	Baskets (liste d'items)
1	Pain, Coca-Cola, Lait
2	Bières, Pain
3	Bière, Coca-Cola, Couches, Lait
4	Bière, Pain, Couches, Lait
5	Coca-Cola, Couches, Lait

Quel est le **support** de l'itemset {**Bière, Pain**} ?

Itemsets fréquents

- Les **itemsets fréquents** (frequent itemsets) sont les items apparaissant “fréquemment” ensemble dans les baskets

- Un itemset est **fréquent** si l'ensemble de ses items apparaît dans au moins s baskets

- Le support peut aussi être exprimé en pourcentage:

- s est appelé **seuil de support**
- Rapport entre le nombre de baskets contenant tous les items de I et le nombre total de baskets

TID	Baskets (liste d'items)
1	Pain, Coca-Cola, Lait
2	Bière, Pain
3	Bière, Coca-Cola, Couches, Lait
4	Bière, Pain, Couches, Lait
5	Coca-Cola, Couches, Lait

Quel est le **support** de l'itemset {**Bière, Pain**} ?

Exemple

- Items = {lait, coca-cola, pepsi, bière, jus}
- **Seuil de support = 3 (baskets)**

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

Quels sont les itemsets fréquents ?

Exemple

- Items = {lait, coca-cola, pepsi, bière, jus}
- **Seuil de support = 3 (baskets)**

$B_1 = \{\text{lait}, \text{coca-cola}, \text{bière}\}$	$B_2 = \{\text{lait}, \text{pepsi}, \text{jus}\}$
$B_3 = \{\text{lait}, \text{bière}\}$	$B_4 = \{\text{coca-cola}, \text{jus}\}$
$B_5 = \{\text{lait}, \text{pepsi}, \text{bière}\}$	$B_6 = \{\text{lait}, \text{coca-cola}, \text{bière}, \text{jus}\}$
$B_7 = \{\text{coca-cola}, \text{bière}, \text{jus}\}$	$B_8 = \{\text{bière}, \text{coca-cola}\}$

- Itemsets fréquents:
 - {lait}

Exemple

- Items = {lait, coca-cola, pepsi, bière, jus}
- **Seuil de support = 3 (baskets)**

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

- Itemsets fréquents:
 - {lait}, {coca-cola}

Exemple

- Items = {lait, coca-cola, pepsi, bière, jus}
- Seuil de support = 3 (baskets)

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

- Itemsets fréquents:
 - {lait}, {coca-cola}, {bière}

Exemple

- Items = {lait, coca-cola, pepsi, bière, jus}
- Seuil de support = 3 (baskets)

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

- Itemsets fréquents:
 - {lait}, {coca-cola}, {bière}, {jus}

D'autres possibilités ?

Exemple

- Items = {lait, coca-cola, pepsi, bière, jus}
- Seuil de support = 3 (baskets)

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

- Itemsets fréquents:
 - {lait}, {coca-cola}, {bière}, {jus}
 - {lait, bière}, {bière, coca-cola}, {coca-cola, jus}

Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références

Règles d'association

- Les règles d'association sont des règles de type **if-else**
- $\{i_1, i_2, \dots, i_k\} \rightarrow j$ signifie que si un basket contient tous les items i_1, i_2, \dots, i_k alors il est probable qu'il contienne aussi l'item j
- Il existe en pratique un nombre considérable de règles. L'objectif est de découvrir les plus significatives !
- L'**indice de confiance** (confidence) de la règle d'association $\{i_1, i_2, \dots, i_k\} \rightarrow j$ est la probabilité de j sachant $I = \{i_1, i_2, \dots, i_k\}$

$$\text{conf}(I \rightarrow j) = \frac{\text{support}(I \cup j)}{\text{support}(I)}$$

Règles d'association significatives

- Toutes les règles d'association ayant un indice de confiance élevé ne sont pas significatives ...
- **La règle** $X \rightarrow \text{lait}$ peut avoir un indice élevé pour beaucoup d'itemsets X , car le lait est une denrée souvent achetée indépendamment de X
- **L'intérêt** d'une règle d'association $I \rightarrow j$ est la différence entre son indice de confiance et la proportion de baskets contenant j

$$\text{Interest}(I \rightarrow j) = \text{conf}(I \rightarrow j) - P(j)$$

**Les règles d'association intéressantes ont un intérêt élevé
(généralement au dessus de 0.5)**

Exemple: indice de confiance et intérêt

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

Considérons la règle d'association **{lait, bière}** → **{coca-cola}** correspondant aux données ci-dessus.

- Quel est son indice de confiance ?
- Quel est son intérêt ?
- Cette règle est-elle intéressante ?

Exemple: indice de confiance et intérêt

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

Considérons la règle d'association **{lait, bière} → {coca-cola}** correspondant aux données ci-dessus.

- Quel est son indice de confiance ? **$2/4 = 0.5$**
- Quel est son intérêt ? **$0.5 - \frac{5}{8} = \frac{1}{8}$**
- Cette règle est-elle intéressante ? **Non, pas tellement ...**

Comment **trouver** les règles d'association ?

Trouver les règles d'association (1/2)

- **Problématique:** trouver toutes les règles d'association ayant un support s et un indice de confiance c
 - Remarque: le support d'une règle d'association est le support de l'itemset de la partie gauche
 - **Supp** $(\{ i_1, i_2, \dots, i_k \} \rightarrow j) = \text{Supp}(\{ i_1, i_2, \dots, i_k \})$
- **D'abord, il faut trouver les itemsets fréquents !**
 - Si $\{ i_1, i_2, \dots, i_k \} \rightarrow j$ possède un support et un indice de confiance élevé, alors les itemsets $\{ i_1, i_2, \dots, i_k \}$ et $\{ i_1, i_2, \dots, i_k, j \}$ sont fréquents

Trouver les règles d'association (2/2)

- **Étape 1:** apprendre tous les itemsets fréquents I (algorithme Apriori)
- **Étape 2:** générer les règles
 - Pour chaque sous ensemble A de I , générer la règle $A \rightarrow I \setminus A$
 - Puisque I est fréquent, A est également fréquent
 - Variante 1: Calcul de la confiance de la règle en une passe
 $\text{conf}(A, B \rightarrow C, D) = \text{support}(A, B, C, D) / \text{support}(A, B)$
 - Variante 2: Si $A, B, C \rightarrow D$ est en dessous d'une certaine confiance, alors $A, B \rightarrow C, D$ aussi. Génération de règles plus "grandes"
 - Ne garder que les règles au dessus d'un certain seuil de confiance

Exercice: trouver les règles d'association

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

$s = 3$

$c = 0.75$

Exercice: trouver les règles d'association

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

$s = 3$

$c = 0.75$

1. Itemsets fréquents

$\{\text{l}\}, \{\text{c}\}, \{\text{b}\}, \{\text{j}\}, \{\text{b,l}\}, \{\text{b,c}\}, \{\text{c,j}\}$

Exercice: trouver les règles d'association

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

$s = 3$

$c = 0.75$

1. Itemsets fréquents

$\{l\}, \{c\}, \{b\}, \{j\}, \{b,l\} \{b,c\} \{c,j\}$

2. Génération des règles

$b \rightarrow l: c=4/6; l \rightarrow b: c=4/5; b \rightarrow c: c=5/6; b,c \rightarrow l: c=3/5; b,l \rightarrow c: c=3/4; b \rightarrow c,l: c=3/6$

Exercice: trouver les règles d'association

$B_1 = \{\text{lait, coca-cola, bière}\}$	$B_2 = \{\text{lait, pepsi, jus}\}$
$B_3 = \{\text{lait, bière}\}$	$B_4 = \{\text{coca-cola, jus}\}$
$B_5 = \{\text{lait, pepsi, bière}\}$	$B_6 = \{\text{lait, coca-cola, bière, jus}\}$
$B_7 = \{\text{coca-cola, bière, jus}\}$	$B_8 = \{\text{bière, coca-cola}\}$

$s = 3$

$c = 0.75$

1. Itemsets fréquents

$\{l\}, \{c\}, \{b\}, \{j\}, \{b,l\} \{b,c\} \{c,j\}$

2. Génération des règles

~~$b \rightarrow l: c=4/6; l \rightarrow b: c=4/5; b \rightarrow c: c=5/6; b,e \rightarrow l: c=3/5; b,l \rightarrow c: c=3/4; b \rightarrow e, l: c=3/6$~~

Réduction du nombre de règles d'association

■ Maximal frequent itemsets

Si aucun superset immédiat n'est fréquent

■ Itemsets fermés

Si aucun superset immédiat n'a le même support

	Support	Maximal (s=3)	Fermé
A	4	Non	Non
B	5	Non	Oui
C	3	Non	Non
AB	4	Oui	Oui
AC	2	Non	Non
BC	3	Oui	Oui
ABC	2	Non	Oui

Réduction du nombre de règles d'association

■ Maximal frequent itemsets

Si aucun superset immédiat n'est fréquent

■ Itemsets fermés

Si aucun superset immédiat n'a le même support

	Support	Maximal (s=3)	Fermé
A	4	Non	Non
B	5	Non	Oui
C	3	Non	Non
AB	4	Oui	Oui
AC	2	Non	Non
BC	3	Oui	Oui
ABC	2	Non	Oui

Fréquent, mais superset BC aussi fréquent

Réduction du nombre de règles d'association

■ Maximal frequent itemsets

Si aucun superset immédiat n'est fréquent

■ Itemsets fermés

Si aucun superset immédiat n'a le même support

	Support	Maximal (s=3)	Fermé
A	4	Non	Non
B	5	Non	Oui
C	3	Non	Non
AB	4	Oui	Oui
AC	2	Non	Non
BC	3	Oui	Oui
ABC	2	Non	Oui

Fréquent, mais le seul superset ABC
n'est pas fréquent

Réduction du nombre de règles d'association

■ Maximal frequent itemsets

Si aucun superset immédiat n'est fréquent

■ Itemsets fermés

Si aucun superset immédiat n'a le même support

	Support	Maximal (s=3)	Fermé
A	4	Non	Non
B	5	Non	Oui
C	3	Non	Non
AB	4	Oui	Oui
AC	2	Non	Non
BC	3	Oui	Oui
ABC	2	Non	Oui

Le superset BC possède le même support

Réduction du nombre de règles d'association

■ Maximal frequent itemsets

Si aucun superset immédiat n'est fréquent

■ Itemsets fermés

Si aucun superset immédiat n'a le même support

Le seul superset ABC possède un support inférieur

	Support	Maximal (s=3)	Fermé
A	4	Non	Non
B	5	Non	Oui
C	3	Non	Non
AB	4	Oui	Oui
AC	2	Non	Non
BC	3	Oui	Oui
ABC	2	Non	Oui



Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références

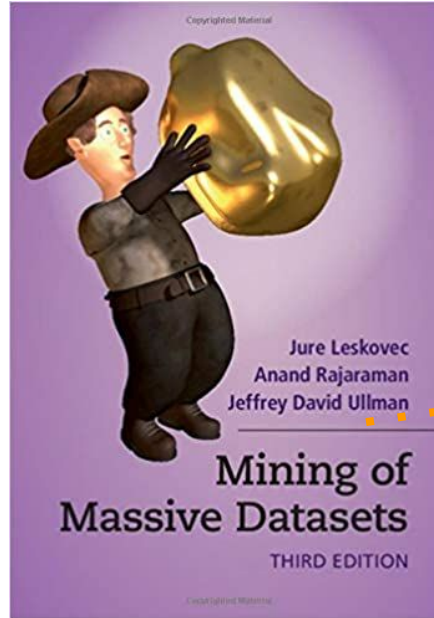


Pull de <https://github.com/mswawola-cegep/420-a58-sf-gr-12060.git>

02-04

Sommaire

1. Introduction au modèle “Market-basket”
2. Applications
3. Itemsets fréquents
4. Règles d’association
5. Ateliers
6. Lectures et références



6 Frequent Itemsets
p. 213-251

Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, **Mining of Massive Datasets,**
3rd edition

Références

- [1] [Mining of Massive Datasets, 3rd edition](#)
- [2] [Association Rule Mining via Apriori Algorithm in Python](#)