Fouille de motifs

Brice Chardin

2022–2023

ISAE-ENSMA

INTRODUCTION

DATA MINING

Extraction de connaissances à partir de données (en anglais *Knowledge Discovery in Databases*, ou *KDD*)

Techniques et outils pour extraire de la connaissance à partir d'un volume important de données, autrement difficilement exploitables

Raisonnement sur les attributs, et non sur les instances

OBJECTIFS DE L'ANALYSE DE DONNÉES

Données (x, y) générées par un système [1]

$$x \rightarrow \begin{matrix}\text{système} \\ \text{(nature)}\end{matrix} \rightarrow y$$

Objectifs

- Extraire de l'information à partir des données, c'est-à-dire améliorer la connaissance du système

- Extraire des données expérimentales une approximation satisfaisante du comportement du système

1. Leo Breiman. *Statistical modeling : The two cultures*. Statistical science 16.3 (2001) : 199-231.

OBJECTIFS DE L'ANALYSE DE DONNÉES

Données (x, y) générées par un système [1]

$$x \rightarrow \begin{matrix}\text{système} \\ \text{(nature)}\end{matrix} \rightarrow y$$

Objectifs

- Extraire de l'information à partir des données, c'est-à-dire améliorer la connaissance du système

$\rightarrow$ *data mining*

- Extraire des données expérimentales une approximation satisfaisante du comportement du système

$\rightarrow$ *machine learning*

1. Leo Breiman. *Statistical modeling : The two cultures*. Statistical science 16.3 (2001) : 199-231.

PROCESSUS D'EXTRACTION DE CONNAISSANCES

1. Correction des données bruitées ou inconsistantes
2. Combinaison de données de sources multiples
3. Sélection des données pertinentes
4. Transformation des données au format approprié
5. Extraction de connaissance à partir des données (*data mining*)
6. Identification des motifs apportant de la connaissance
7. Présentation des résultats pour diffusion

PROCESSUS D'EXTRACTION DE CONNAISSANCES

1. Correction des données bruitées ou inconsistantes
2. Combinaison de données de sources multiples
3. Sélection des données pertinentes
4. Transformation des données au format approprié
5. Extraction de connaissance à partir des données (*data mining*)
6. Identification des motifs apportant de la connaissance
7. Présentation des résultats pour diffusion

Processus itératif

PROCESSUS D'EXTRACTION DE CONNAISSANCES

1. Correction des données bruitées ou inconsistantes
2. Combinaison de données de sources multiples
3. Sélection des données pertinentes
4. Transformation des données au format approprié
5. Extraction de connaissance à partir des données (*data mining*)
6. Identification des motifs apportant de la connaissance
7. Présentation des résultats pour diffusion

Processus itératif

Data mining : étape 5 ou ensemble du processus

PROCESSUS D'EXTRACTION DE CONNAISSANCES

0. Compréhension des données
1. Correction des données bruitées ou inconsistantes
2. Combinaison de données de sources multiples
3. Sélection des données pertinentes
4. Transformation des données au format approprié
5. Extraction de connaissance à partir des données (*data mining*)
6. Identification des motifs apportant de la connaissance
7. Présentation des résultats pour diffusion

Processus itératif

Data mining : étape 5 ou ensemble du processus

LIEN AVEC LA FORMATION

DOI – pré-traitement des données

Étapes 1 et 2 (nettoyage et combinaison)

LIEN AVEC LA FORMATION

DOI – pré-traitement des données

Étapes 1 et 2 (nettoyage et combinaison)

UED, IDM

Étapes 3, 4 et 7 (sélection, transformation et présentation)

LIEN AVEC LA FORMATION

DOI – pré-traitement des données

Étapes 1 et 2 (nettoyage et combinaison)

UED, IDM

Étapes 3, 4 et 7 (sélection, transformation et présentation)

CBD - entrepôts et cubes de données

Étapes 4, 5 et 6 (transformation, extraction et identification)

LIEN AVEC LA FORMATION

DOI – pré-traitement des données

Étapes 1 et 2 (nettoyage et combinaison)

UED, IDM

Étapes 3, 4 et 7 (sélection, transformation et présentation)

CBD - entrepôts et cubes de données

Étapes 4, 5 et 6 (transformation, extraction et identification)

DOI – fouille de motifs

Étapes 4, 5 et 6 (transformation, extraction et identification)

LIEN AVEC LA FORMATION

DOI – pré-traitement des données

Étapes 1 et 2 (nettoyage et combinaison)

UED, IDM

Étapes 3, 4 et 7 (sélection, transformation et présentation)

CBD - entrepôts et cubes de données

Étapes 4, 5 et 6 (transformation, extraction et identification)

DOI – fouille de motifs

Étapes 4, 5 et 6 (transformation, extraction et identification)

Formation ENSMA (dans son ensemble)

Étape 0 (pour les données en lien avec la formation)

- défaillances moteur, analyse d'image satellitaires, etc.

TYPES DE DONNÉES

Données structurées (relations)

Données semi-structurées (XML, json)

Données non structurées (texte, images, son, vidéos)

TYPES DE DONNÉES

Données structurées (relations) ← cadre de ce cours

Données semi-structurées (XML, json)

Données non structurées (texte, images, son, vidéos)

TYPES DE DONNÉES

Données structurées (relations) ← cadre de ce cours

Données semi-structurées (XML, json)

Données non structurées (texte, images, son, vidéos)

Algorithmes spécialisés

- biologie (e.g. séquences génétiques)
- séries temporelles
- données spatio-temporelles
- logs
- graphes
- etc.

EFFICACITÉ

Algorithmes applicables à de grands volumes de données

Processus interactif $\rightarrow$ temps de réponse court

Critères d'efficacité

- passage à l'échelle par rapport aux données
- capacité de parallélisation
- calcul incrémental

PERSPECTIVES DE RECHERCHE

Facteurs limitant l'usage des outils de fouille :

- interaction avec les utilisateurs
- efficacité
- généralité des algorithmes

→ Évolutions à prévoir pour les prochaines décennies

ITEMSETS FRÉQUENTS

ITEMSETS FRÉQUENTS

Cooccurrence d'éléments (*item*) dans un ensemble de transactions

Exemple typique (et historique) : le panier d'achats

no. transaction	café	sucre	lait	...	thé
1	1	1	0	...	0
2	1	1	0	...	0
3	1	0	1	...	1
4	0	1	0	...	1

ITEMSETS FRÉQUENTS

Cooccurrence d'éléments (*item*) dans un ensemble de transactions

Exemple typique (et historique) : le panier d'achats

no. transaction	café	sucre	lait	...	thé
1	1	1	0	...	0
2	1	1	0	...	0
3	1	0	1	...	1
4	0	1	0	...	1

Itemsets fréquents (seuil à 50% des transactions)

- {café} (75%)
- {sucre} (75%)
- {thé} (50%)
- {café, sucre} (50%)

EXEMPLE – OBJECTIFS POSSIBLES

Réorganisation du rayonnage

Gestion des stocks

Suggestions d'achats

Ajustement des prix

MESURE D'INTÉRÊT : LE SUPPORT

Support (aussi appelé *fréquence* ou *count*) de l'itemset

- relatif ou absolu

Calcul du support

no. transaction	café	sucre	lait	thé
1	1	1	0	0
2	1	1	0	0
3	1	0	1	1
4	0	1	0	1

- support({café}) = 3 (absolu), ou 75% (relatif)
- support({café, lait, thé}) = 1, ou 25%
- support({sucre, lait}) = 0

MESURE D'INTÉRÊT : LE SUPPORT

Support (aussi appelé *fréquence* ou *count*) de l'itemset

- relatif ou absolu

Calcul du support

no. transaction	café	sucre	lait	thé
1	1	1	0	0
2	1	1	0	0
3	1	0	1	1
4	0	1	0	1

- support({café}) = 3 (absolu), ou 75% (relatif)
- support({café, lait, thé}) = 1, ou 25%
- support({sucre, lait}) = 0

MESURE D'INTÉRÊT : LE SUPPORT

Support (aussi appelé *fréquence* ou *count*) de l'itemset

- relatif ou absolu

Calcul du support

no. transaction	café	sucre	lait	thé
1	1	1	0	0
2	1	1	0	0
3	1	0	1	1
4	0	1	0	1

- support({café}) = 3 (absolu), ou 75% (relatif)
- support({café, lait, thé}) = 1, ou 25%
- support({sucre, lait}) = 0

MESURE D'INTÉRÊT : LE SUPPORT

Support (aussi appelé *fréquence* ou *count*) de l'itemset

- relatif ou absolu

Calcul du support

no. transaction	café	sucre	lait	thé
1	1	1	0	0
2	1	1	0	0
3	1	0	1	1
4	0	1	0	1

- support({café}) = 3 (absolu), ou 75% (relatif)
- support({café, lait, thé}) = 1, ou 25%
- support({sucre, lait}) = 0

MESURE D'INTÉRÊT : LE SUPPORT

Support (aussi appelé *fréquence* ou *count*) de l'itemset

- relatif ou absolu

Calcul du support

no. transaction	café	sucre	lait	thé
1	1	1	0	0
2	1	1	0	0
3	1	0	1	1
4	0	1	0	1

- support({café}) = 3 (absolu), ou 75% (relatif)
- support({café, lait, thé}) = 1, ou 25%
- support({sucre, lait}) = 0

En relatif, support(A) = P(A)

DÉFINITIONS

Définition (support absolu)

Pour une relation R, $\text{support}(X) = |\{t_i \in R \mid X \subseteq t_i\}|$

Définition (support relatif)

Pour une relation R, $\text{support}(X) = \dfrac{|\{t_i \in R \mid X \subseteq t_i\}|}{|R|}$

Définition (fréquence)

Un itemset X est fréquent pour un seuil s ssi $\text{support}(X) \geq s$

Propriété (monotonie)

$X \subseteq Y \implies \text{support}(X) \geq \text{support}(Y)$

Exemple (monotonie)

$\text{support}(\{\text{café}\}) \geq \text{support}(\{\text{café, sucre}\})$

CONSÉQUENCES DE LA MONOTONIE

Si un itemset X est fréquent, alors tout itemset $Y \subseteq X$ l'est aussi

Un itemset X contient $2^{|X|} - 2$ itemsets plus spécifiques (X et $\emptyset$ exclus)

CONSÉQUENCES DE LA MONOTONIE

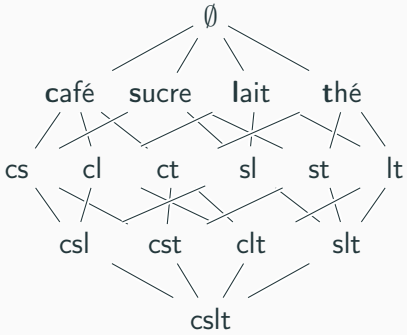
Si un itemset X est fréquent, alors tout itemset $Y \subseteq X$ l'est aussi

Un itemset X contient $2^{|X|} - 2$ itemsets plus spécifiques (X et $\emptyset$ exclus)

$\rightarrow$ Nombre potentiellement important d'itemsets fréquents

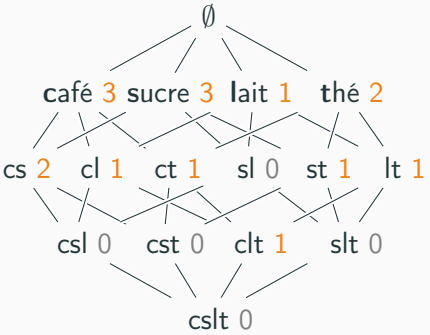
ITEMSETS FRÉQUENTS

Seuil de fréquence : 1



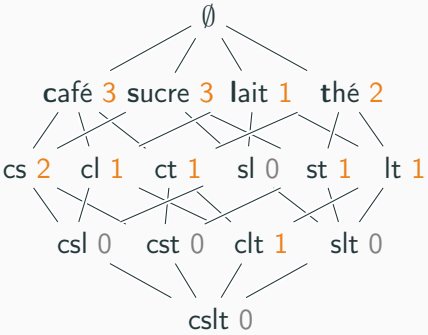
ITEMSETS FRÉQUENTS

Seuil de fréquence : 1



ITEMSETS FRÉQUENTS

Seuil de fréquence : 1 $\rightarrow$ 10 itemsets fréquents



FRÉQUENTS MAXIMAUX

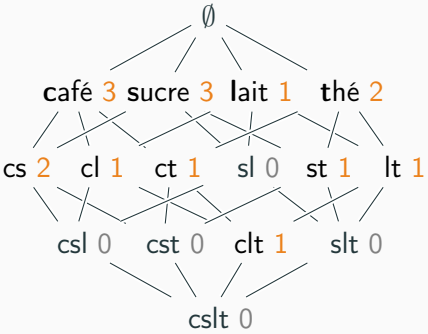
Définition (maximalité)

X est maximal ssi $\nexists Y \supset X \mid Y$ fréquent

I.e. il n'existe pas d'itemset contenant X (strictement) qui soit fréquent

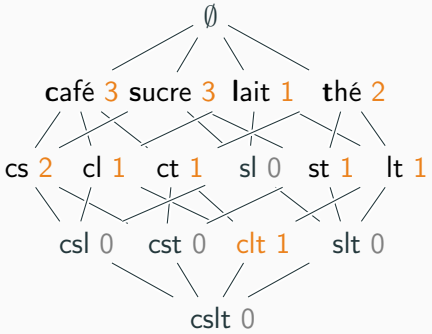
FRÉQUENTS MAXIMAUX

Seuil de fréquence : 1



FRÉQUENTS MAXIMAUX

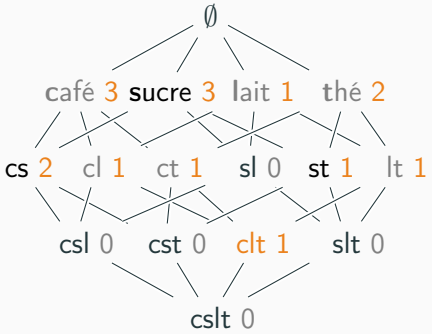
Seuil de fréquence : 1



- {café, lait, thé} (support = 1)

FRÉQUENTS MAXIMAUX

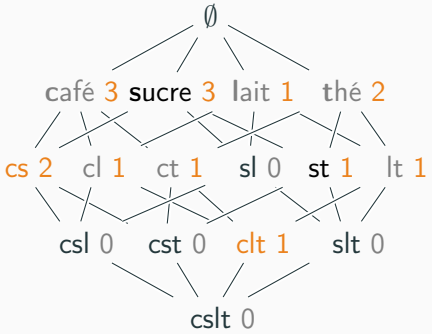
Seuil de fréquence : 1



- {café, lait, thé} (support = 1)

FRÉQUENTS MAXIMAUX

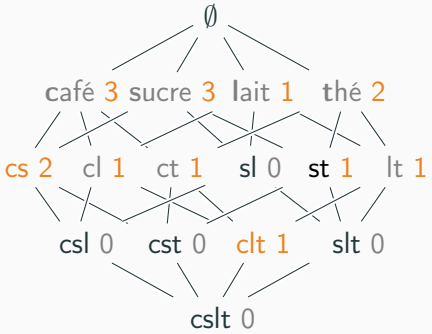
Seuil de fréquence : 1



- {café, lait, thé} (support = 1)
- {café, sucre} (support = 2)

FRÉQUENTS MAXIMAUX

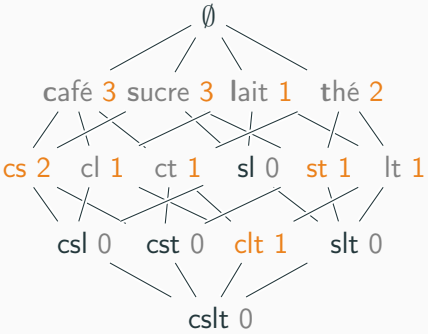
Seuil de fréquence : 1



- {café, lait, thé} (support = 1)
- {café, sucre} (support = 2)

FRÉQUENTS MAXIMAUX

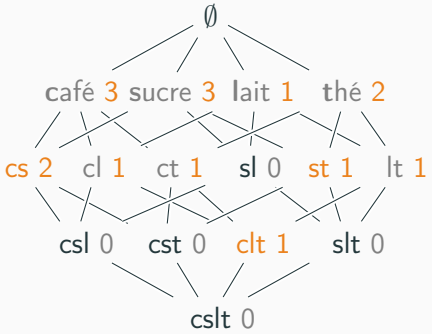
Seuil de fréquence : 1



- {café, lait, thé} (support = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)

FRÉQUENTS MAXIMAUX

Seuil de fréquence : $1 \rightarrow 3$ itemsets fréquents maximaux



- {café, lait, thé} (support = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)

FRÉQUENTS MAXIMAUX

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1

L'itemset {thé} est il fréquent ?

FRÉQUENTS MAXIMAUX

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1

L'itemset {thé} est il fréquent ? → oui car, e.g., {sucre, thé} est fréquent

FRÉQUENTS MAXIMAUX

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1

L'itemset {thé} est il fréquent ? → oui car, e.g., {sucre, thé} est fréquent

Quel est son support ?

FRÉQUENTS MAXIMAUX

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1

L'itemset {thé} est il fréquent ? → oui car, e.g., {sucre, thé} est fréquent

Quel est son support ? → support({thé}) ≥ support({café, lait, thé}) = 1

support({thé}) ≥ support({sucre, thé}) = 1

FRÉQUENTS MAXIMAUX

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1

L'itemset {thé} est il fréquent ? → oui car, e.g., {sucre, thé} est fréquent

Quel est son support ? → support({thé}) ≥ support({café, lait, thé}) = 1

support({thé}) ≥ support({sucre, thé}) = 1

(en pratique support({thé}) = 2)

FRÉQUENTS MAXIMAUX

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1

L'itemset {thé} est il fréquent ? → oui car, e.g., {sucre, thé} est fréquent

Quel est son support ? → support({thé}) ≥ support({café, lait, thé}) = 1

support({thé}) ≥ support({sucre, thé}) = 1

(en pratique support({thé}) = 2)

→ Perte d'information sur la valeur exacte du support

ITEMSETS CLOS (*CLOSED*)

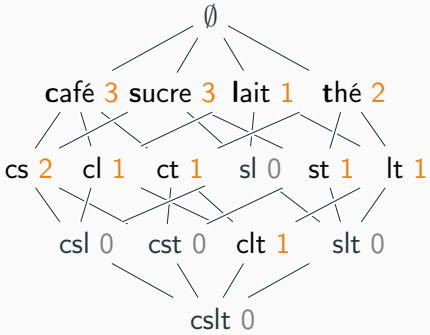
Définition (clôture)

X est clos ssi $\nexists Y \supset X \mid \text{support}(Y) = \text{support}(X)$

I.e. il n'existe par d'itemset contenant X (strictement) qui possède le même support

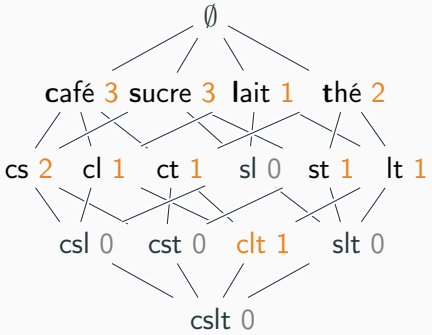
FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

Seuil de fréquence : 1



FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

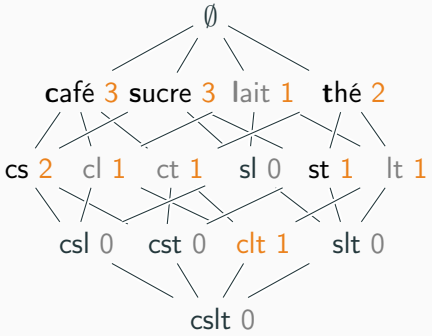
Seuil de fréquence : 1



- {café, lait, thé} (sup. = 1)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

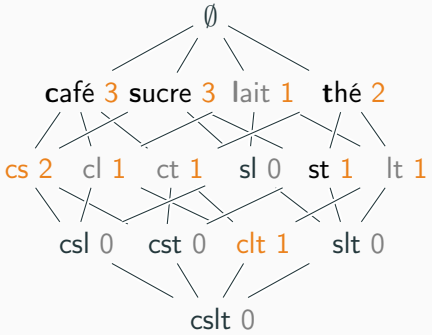
Seuil de fréquence : 1



- {café, lait, thé} (sup. = 1)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

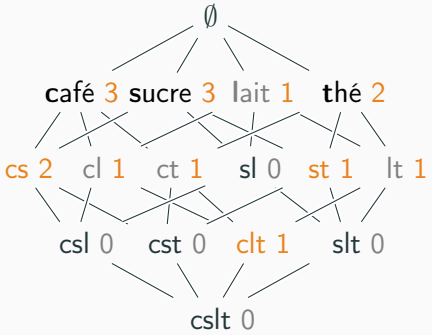
Seuil de fréquence : 1



- {café, lait, thé} (sup. = 1)
- {café, sucre} (support = 2)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

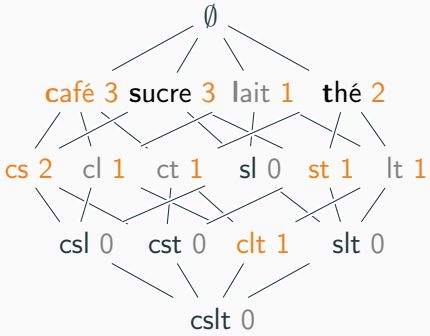
Seuil de fréquence : 1



- {café, lait, thé} (sup. = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

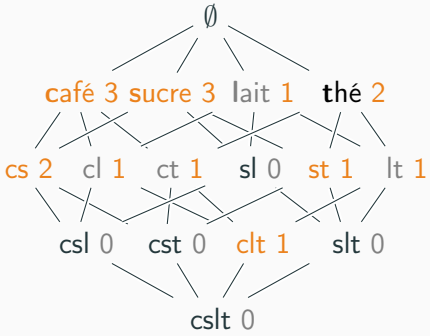
Seuil de fréquence : 1



- {café, lait, thé} (sup. = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)
- {café} (support = 3)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

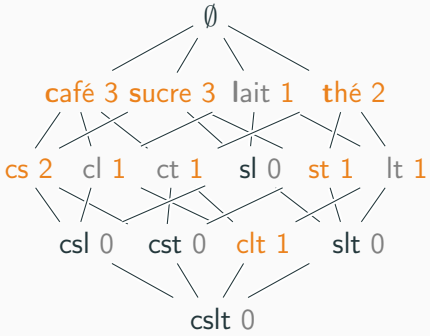
Seuil de fréquence : 1



- {café, lait, thé} (sup. = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)
- {café} (support = 3)
- {sucre} (support = 3)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

Seuil de fréquence : 1

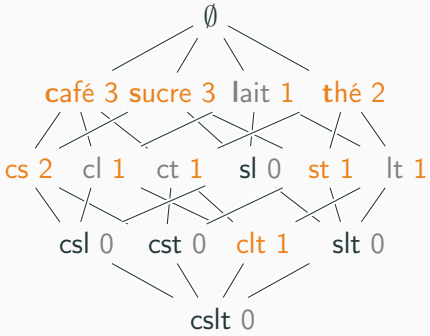


- {café, lait, thé} (sup. = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)

- {café} (support = 3)
- {sucre} (support = 3)
- {thé} (support = 2)

FRÉQUENTS CLOS (*CLOSED FREQUENT ITEMSETS*)

Seuil de fréquence : 1 $\rightarrow$ 6 itemsets fréquents clos



- {café, lait, thé} (sup. = 1)
- {café, sucre} (support = 2)
- {sucre, thé} (support = 1)

- {café} (support = 3)
- {sucre} (support = 3)
- {thé} (support = 2)

FRÉQUENTS CLOS

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1
- support({café}) = 3
- support({sucre}) = 3
- support({thé}) = 2

Soit C l'ensemble des fréquents clos :

$$\text{support}(X) = \max_{Y \in C | Y \supseteq X} \text{support}(Y)$$

FRÉQUENTS CLOS

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1
- support({café}) = 3
- support({sucre}) = 3
- support({thé}) = 2

Soit C l'ensemble des fréquents clos :

$$\text{support}(X) = \max_{Y \in C | Y \supseteq X} \text{support}(Y)$$

L'itemset {lait} est il fréquent ?

FRÉQUENTS CLOS

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1
- support({café}) = 3
- support({sucre}) = 3
- support({thé}) = 2

Soit C l'ensemble des fréquents clos :

$$\text{support}(X) = \max_{Y \in C | Y \supseteq X} \text{support}(Y)$$

L'itemset {lait} est il fréquent ? $\rightarrow$ oui car {café, lait, thé} est fréquent

FRÉQUENTS CLOS

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1
- support({café}) = 3
- support({sucre}) = 3
- support({thé}) = 2

Soit C l'ensemble des fréquents clos :

$$\text{support}(X) = \max_{Y \in C | Y \supseteq X} \text{support}(Y)$$

L'itemset {lait} est il fréquent ? → oui car {café, lait, thé} est fréquent
Quel est son support ?

FRÉQUENTS CLOS

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1
- support({café}) = 3
- support({sucre}) = 3
- support({thé}) = 2

Soit C l'ensemble des fréquents clos :

$$\text{support}(X) = \max_{Y \in C | Y \supseteq X} \text{support}(Y)$$

L'itemset {lait} est il fréquent ? $\rightarrow$ oui car {café, lait, thé} est fréquent
Quel est son support ? $\rightarrow$ support({lait}) = max(support({café, lait, thé})) = 1

FRÉQUENTS CLOS

- support({café, lait, thé}) = 1
- support({café, sucre}) = 2
- support({sucre, thé}) = 1
- support({café}) = 3
- support({sucre}) = 3
- support({thé}) = 2

Soit C l'ensemble des fréquents clos :

$$\text{support}(X) = \max_{Y \in C | Y \supseteq X} \text{support}(Y)$$

L'itemset {lait} est il fréquent ? $\rightarrow$ oui car {café, lait, thé} est fréquent
Quel est son support ? $\rightarrow$ support({lait}) = max(support({café, lait, thé})) = 1

$\rightarrow$ Pas de perte d'information

APRIORI

Algorithme d'énumération des itemsets fréquents [2]

Principes du calcul

- Énumérer les itemsets fréquents de taille 1, puis 2, 3, etc.
- Tout sous-ensemble non vide d'un itemset fréquent doit être fréquent

2. Rakesh Agrawal and Ramakrishnan Srikant. *Fast algorithms for mining association rules*. Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.

APRIORI : ALGORITHME

1. Calculer le support de chaque 1-itemset
2. Sauvegarder les 1-itemset fréquents dans la liste L_1
3. Pour $i = 2\,;\ L_{i-1} \neq \emptyset\,;\ i{+}{+}$
   - Énumérer l'ensemble C_i des itemsets candidats par auto-jointure de L_{i-1} (i.e. union des paires de k-itemsets possédant $k-1$ éléments communs)
   - Retirer de C_i les sur-ensemble d'itemsets non fréquents de L_{i-1}
   - Calculer le support des itemsets de C_i
   - Lister les fréquents dans L_i
4. Les fréquents sont l'ensemble des L_i

APRIORI : EXEMPLE

Pour un seuil de fréquence de 2

Calculer le support de chaque 1-itemset

L_1

itemset	support
café	3
sucre	2
lait	1
thé	2

Pour un seuil de fréquence de 2

Sauvegarder les 1-itemset fréquents dans la liste L_1

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

Pour un seuil de fréquence de 2

Enumérer les itemsets candidats par auto-jointure de L_1

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

$\rightarrow$

L_2

itemset	support
café, sucre	
café, thé	
sucre, thé	

Pour un seuil de fréquence de 2

Retirer les sur-ensemble d'itemsets non fréquents de L_1

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

$\rightarrow$

L_2

itemset	support
café, sucre	
café, thé	
sucre, thé	

APRIORI : EXEMPLE

Pour un seuil de fréquence de 2

Calculer le support des itemsets candidats

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

$\rightarrow$

L_2

itemset	support
café, sucre	2
café, thé	1
sucre, thé	1

APRIORI : EXEMPLE

Pour un seuil de fréquence de 2

Lister les fréquents dans L_2

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

$\rightarrow$

L_2

itemset	support
café, sucre	2
~~café, thé~~	~~1~~
~~sucre, thé~~	~~1~~

Pour un seuil de fréquence de 2

Enumérer les itemsets candidats par auto-jointure de L_2

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

$\rightarrow$

L_2

itemset	support
café, sucre	2
~~café, thé~~	~~1~~
~~sucre, thé~~	~~1~~

$\rightarrow$ L_3

$\emptyset$

APRIORI : EXEMPLE

Pour un seuil de fréquence de 2

L_1

itemset	support
café	3
sucre	2
~~lait~~	~~1~~
thé	2

$\rightarrow$

L_2

itemset	support
café, sucre	2
~~café, thé~~	~~1~~
~~sucre, thé~~	~~1~~

$\rightarrow$ L_3 $\emptyset$

Itemsets fréquents : {café}, {sucre} et {thé}, {café, sucre}

ÉNUMÉRATION DES ITEMSETS FRÉQUENTS

Apriori n'est pas considéré comme l'algorithme le plus efficace

Solutions alternatives

- Linear time Closed itemset Miner (LCM) [3]
- FP-Growth [4]
- et beaucoup d'autres : Eclat, Relim, H-Mine, PrePost, etc.

3. Takeaki Uno, Tatsuya Asai, Yuzo Uchida and Hiroki Arimura. *An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases*. International Conference on Discovery Science. 2004.

4. Jiawei Han, Jian Pei, Yiwen Yin and Runying Mao. *Mining frequent patterns without candidate generation : A frequent-pattern tree approach*. Data Mining and Knowledge Discovery 8, no. 1 (2004) : 53-87.

RÈGLES D'ASSOCIATION

LIMITES DES ITEMSETS FRÉQUENTS

Milliers, millions d'itemsets énumérés

Itemsets fréquents

Au pire $2^n - 1$ itemsets, avec n le nombre d'items

Itemsets fréquents clos ou maximaux [5]

Au pire $\binom{n}{\lfloor n/2 \rfloor} \approx \sqrt{\frac{2}{\pi n}} \times 2^n$ itemsets

5. Par le théorème de Sperner pour la valeur du pire cas et la formule de Stirling pour l'approximation de la valeur du coefficient binomial.

LIMITES DES ITEMSETS FRÉQUENTS

Milliers, millions d'itemsets énumérés

Itemsets fréquents

Au pire $2^n - 1$ itemsets, avec n le nombre d'items

Itemsets fréquents clos ou maximaux [5]

Au pire $\binom{n}{\lfloor n/2 \rfloor} \approx \sqrt{\frac{2}{\pi n}} \times 2^n$ itemsets

Itemsets fréquents limités par le support comme seul indicateur
(plus éventuellement la taille de l'itemset)

5. Par le théorème de Sperner pour la valeur du pire cas et la formule de Stirling pour l'approximation de la valeur du coefficient binomial.

CRITÈRES D'INTÉRÊT

Règle (ou motif) intéressante si :

- Compréhensible (par un humain)
- Valide par rapport aux données
  - y compris sur des mesures ultérieures
- Utile (actionnable)
- Nouvelle (contredit les aprioris)
  - ou connue pour valider une hypothèse

CRITÈRES D'INTÉRÊT

Règle (ou motif) intéressante si :

- Compréhensible (par un humain)
- Valide par rapport aux données
  - y compris sur des mesures ultérieures
- Utile (actionnable)
- Nouvelle (contredit les aprioris)
  - ou connue pour valider une hypothèse

Critères en partie subjectifs

RÈGLES D'ASSOCIATION

Règle d'association : $X \rightarrow Y$

X est l'antécédent (ou *LHS*, left-hand side)

Y est le conséquent (ou *RHS*, right-hand side)

Intuitivement, permet d'étudier l'influence de X sur Y

RÈGLES D'ASSOCIATION

Règle d'association : $X \rightarrow Y$

X est l'antécédent (ou *LHS*, left-hand side)

Y est le conséquent (ou *RHS*, right-hand side)

Intuitivement, permet d'étudier l'influence de X sur Y

Exemples

- {café} $\rightarrow$ {sucre}
- {café, thé} $\rightarrow$ {sucre, lait}

Définition (support)

$\text{support}(X \to Y) = \text{support}(X \cup Y) = P(X \wedge Y)$

Proportion des faits pour lesquels la règle est vérifiée positivement (i.e. où X et Y sont présents)

Définition (support)

$\text{support}(X \to Y) = \text{support}(X \cup Y) = P(X \wedge Y)$

Proportion des faits pour lesquels la règle est vérifiée positivement (i.e. où X et Y sont présents)

Définition (confiance)

$$\text{confidence}(X \to Y) = \frac{\text{support}(X \to Y)}{\text{support}(X)} = \frac{P(X \wedge Y)}{P(X)} = P(Y \mid X)$$

Mesure de validité de la règle

Définition (support)

$\text{support}(X \to Y) = \text{support}(X \cup Y) = P(X \wedge Y)$

Proportion des faits pour lesquels la règle est vérifiée positivement (i.e. où X et Y sont présents)

Définition (confiance)

$$\text{confidence}(X \to Y) = \frac{\text{support}(X \to Y)}{\text{support}(X)} = \frac{P(X \wedge Y)}{P(X)} = P(Y \mid X)$$

Mesure de validité de la règle

$$\text{confidence}(X \to Y) = 1 \equiv X \implies Y$$

DÉCOMPOSITION DES RÈGLES

Propriétés

$\forall Z : \text{support}(X \rightarrow Y) \geq \text{support}(X \rightarrow Y \cup Z)$

$\text{confidence}(X \rightarrow Y) \geq \text{confidence}(X \rightarrow Y \cup Z)$

Si $X \rightarrow Y \cup Z$ est une règle intéressante alors $X \rightarrow Y$ et $X \rightarrow Z$ le sont aussi

Tendance à décomposer en $X \rightarrow A$, avec $|A| = 1$

ÉNUMÉRATION DES RÈGLES D'ASSOCIATION AVEC SEUILS MINIMAUX DE SUPPORT ET DE CONFIANCE

Propriété

$\text{confidence}(X \to Y) \geq \text{support}(X \to Y)$

Mesures nécessaires : $\text{support}(X \cup Y)$ et $\text{support}(X)$

1. énumérer les itemsets fréquents avec le minimum $\min(\tau_c, \tau_s)$ des seuils de confiance τ_c et de support τ_s
2. générer les règles d'associations pour chaque itemset fréquent X
   - pour tout sous ensemble non vide $Y \subset X$, générer la règle $Y \to (X - Y)$ si $\frac{\text{support}(X)}{\text{support}(Y)} \geq \tau_c$
3. calculer les mesures d'intérêt complémentaires

Exemples

- $\text{support}(\{\text{café, lait, thé}\}) = 1$
- $\text{support}(\{\text{café, lait}\}) = 1$
- $\text{support}(\{\text{lait}\}) = 1$
- $\text{support}(\{\text{thé}\}) = 2$

$$\text{confidence}(\text{café, lait} \rightarrow \text{thé}) = \frac{\text{support}(\{\text{café, lait, thé}\})}{\text{support}(\{\text{café, lait}\})} = 1$$

$$\text{confidence}(\text{lait} \rightarrow \text{café, thé}) = \frac{\text{support}(\{\text{café, lait, thé}\})}{\text{support}(\{\text{lait}\})} = 1$$

$$\text{confidence}(\text{thé} \rightarrow \text{café, lait}) = \frac{\text{support}(\{\text{café, lait, thé}\})}{\text{support}(\{\text{thé}\})} = 1/2$$

$$\text{confidence}(\text{thé} \rightarrow \text{lait}) = \frac{\text{support}(\{\text{thé, lait}\})}{\text{support}(\{\text{thé}\})} = 1/2$$

$$\text{confidence}(\text{lait} \rightarrow \text{thé}) = \frac{\text{support}(\{\text{thé, lait}\})}{\text{support}(\{\text{lait}\})} = 1$$

Exemple de règle trompeuse

Sur 100 élèves :

- 90 réussissent LOS (logiciels sûrs)
- 75 réussissent APM (applications mobiles)
- 65 réussissent les deux

$\text{support}(\text{APM} \rightarrow \text{LOS}) = 65/100 = 65\%$

$\text{confidence}(\text{APM} \rightarrow \text{LOS}) = 65/75 = 86.7\%$

Exemple de règle trompeuse

Sur 100 élèves :

- 90 réussissent LOS (logiciels sûrs)
- 75 réussissent APM (applications mobiles)
- 65 réussissent les deux

$\text{support}(\text{APM} \rightarrow \text{LOS}) = 65/100 = 65\%$

$\text{confidence}(\text{APM} \rightarrow \text{LOS}) = 65/75 = 86.7\%$

$P(\text{LOS} \mid \text{APM}) = 0.867$

$P(\text{LOS}) = 0.9$

$\rightarrow$ Réussir APM fait décroitre la probabilité de réussir LOS

Exemple de règle trompeuse

Sur 100 élèves :

- 90 réussissent LOS (logiciels sûrs)
- 75 réussissent APM (applications mobiles)
- 65 réussissent les deux

$\text{support}(\text{APM} \rightarrow \text{LOS}) = 65/100 = 65\%$

$\text{confidence}(\text{APM} \rightarrow \text{LOS}) = 65/75 = 86.7\%$

$P(\text{LOS} \mid \text{APM}) = 0.867$

$P(\text{LOS}) = 0.9$

$\rightarrow$ Réussir APM fait décroitre la probabilité de réussir LOS

$\rightarrow$ Mesures complémentaires d'intérêt

LE LIFT

$$\begin{aligned}\text{lift}(X \to Y) &= \frac{P(X \wedge Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)} \\ &= \frac{\text{confidence}(X \to Y)}{\text{support}(Y)} \\ &= \frac{\text{support}(X \cup Y)}{\text{support}(Y)\,\text{support}(X)}\end{aligned}$$

LE LIFT

$$\begin{aligned}\text{lift}(X \to Y) &= \frac{P(X \wedge Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)} \\ &= \frac{\text{confidence}(X \to Y)}{\text{support}(Y)} \\ &= \frac{\text{support}(X \cup Y)}{\text{support}(Y)\,\text{support}(X)}\end{aligned}$$

$$\text{lift}(X \to Y) = \text{lift}(Y \to X) = \text{lift}(X, Y)$$

Ratio entre la fréquence d'apparition de $X \cup Y$ et ce qu'on obtiendrait si X et Y étaient indépendants

Intuitivement, l'apparition d'un item élève (*lifts*) la fréquence d'apparition de l'autre

LE LIFT

Si $\text{lift}(X, Y) < 1$, alors X et Y sont corrélés négativement

Si $\text{lift}(X, Y) = 1$, alors X et Y sont indépendants

Si $\text{lift}(X, Y) > 1$, alors X et Y sont corrélés positivement

LOS et APM

$$\text{lift}(\text{APM} \rightarrow \text{LOS}) = \frac{0.867}{0.90} = 0.963$$
$$= \frac{0.65}{0.75 \times 0.90}$$

LE PROBLÈME DU LIFT, ILLUSTRÉ

Interlude : exercice sur les métriques (questions 1 et 2)

MESURES COMPLÉMENTAIRES

leverage$(X \rightarrow Y) = P(X \wedge Y) - P(X)P(Y)$

Différence entre la fréquence d'apparition de $X \cup Y$ et ce qu'on obtiendrait si X et Y étaient indépendants

MESURES COMPLÉMENTAIRES

$$\text{leverage}(X \to Y) = P(X \wedge Y) - P(X)P(Y)$$

Différence entre la fréquence d'apparition de $X \cup Y$ et ce qu'on obtiendrait si X et Y étaient indépendants

$$\begin{aligned}\text{conviction}(X \to Y) &= \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)} = \frac{1}{\text{lift}(X \to \neg Y)} \\ &= \frac{1 - P(Y)}{1 - P(Y \mid X)}\end{aligned}$$

MESURES COMPLÉMENTAIRES

$$\text{leverage}(X \to Y) = P(X \wedge Y) - P(X)P(Y)$$

Différence entre la fréquence d'apparition de $X \cup Y$ et ce qu'on obtiendrait si X et Y étaient indépendants

$$\begin{aligned}\text{conviction}(X \to Y) &= \frac{P(X)P(\neg Y)}{P(X \wedge \neg Y)} = \frac{1}{\text{lift}(X \to \neg Y)} \\ &= \frac{1 - P(Y)}{1 - P(Y \mid X)}\end{aligned}$$

Une bonne mesure devrait être invariante par rapport aux transactions non pertinentes [6]

6. Selon Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining : Concepts and Techniques*, Third Edition. The Morgan Kaufmann Series in Data Management Systems (2011) : 269-271.

MESURES COMPLÉMENTAIRES

$$\text{all\_confidence}(X \to Y) = \min(P(X \mid Y), P(Y \mid X))$$

$$\text{max\_confidence}(X \to Y) = \max(P(X \mid Y), P(Y \mid X))$$

MESURES COMPLÉMENTAIRES

$$\text{all\_confidence}(X \to Y) = \min(P(X \mid Y), P(Y \mid X))$$

$$\text{max\_confidence}(X \to Y) = \max(P(X \mid Y), P(Y \mid X))$$

Mesure de Kulczynski

$$\text{Kulc}(X \to Y) = \frac{P(X \mid Y) + P(Y \mid X)}{2}$$

MESURES COMPLÉMENTAIRES

$$\text{all\_confidence}(X \to Y) = \min(P(X \mid Y), P(Y \mid X))$$

$$\text{max\_confidence}(X \to Y) = \max(P(X \mid Y), P(Y \mid X))$$

Mesure de Kulczynski

$$\text{Kulc}(X \to Y) = \frac{P(X \mid Y) + P(Y \mid X)}{2}$$

Mesure cosinus

$$\text{cosine}(X \to Y) = \frac{P(X \wedge Y)}{\sqrt{P(X)P(Y)}}$$

MESURES COMPLÉMENTAIRES

$$\text{all\_confidence}(X \to Y) = \min(P(X \mid Y), P(Y \mid X))$$

$$\text{max\_confidence}(X \to Y) = \max(P(X \mid Y), P(Y \mid X))$$

Mesure de Kulczynski

$$\text{Kulc}(X \to Y) = \frac{P(X \mid Y) + P(Y \mid X)}{2}$$

Mesure cosinus

$$\text{cosine}(X \to Y) = \frac{P(X \wedge Y)}{\sqrt{P(X)P(Y)}}$$

Déséquilibre (imbalance ratio)

$$\text{IR}(X \to Y) = \frac{|\text{support}(X) - \text{support}(Y)|}{\text{support}(X) + \text{support}(Y) - \text{support}(X \cup Y)}$$

MESURES COMPLÉMENTAIRES

Toutes les mesures ne dépendent que de :

- support(X)
- support(Y)
- support($X \cup Y$)

Mesures non influencées par le nombre de transactions non pertinentes :

- all_confidence
- max_confidence
- Kulc
- cosine
- IR

IR pas considéré comme une mesure de qualité, mais une caractérisation

INTERPRÉTATION DES MESURES

Intervalles de valeurs (valeur élevée toujours préférable) :

- support : [0, 1]
- confidence : [0, 1]
- lift : [0, $+\infty$] $\rightarrow$ 1 représente l'indépendance
- leverage : [-1, 1] $\rightarrow$ 0 représente l'indépendance
- conviction : [0, $+\infty$] $\rightarrow$ 1 représente l'indépendance
- all_confidence : [0, 1]
- max_confidence : [0, 1]
- Kulc : [0, 1]
- cosine : [0, 1]
- IR : [0, 1]

Pour deux itemsets parfaitement corrélés :

- confidence, all_confidence, max_confidence, Kulc, cosine valent 1
- conviction vaut $+\infty$

CALCUL DES MESURES COMPLÉMENTAIRES

Interlude : exercice sur les métriques (questions 3 à 9)

CALCUL DES MESURES COMPLÉMENTAIRES

Interlude : exercice sur les métriques (questions 3 à 9)

Pas de consensus sur les meilleures mesures d'intérêt pour les règles d'association

GÉNÉRALISATIONS ET AUTRES RÈGLES

DONNÉES NON BOOLÉENNES

age	gender	hair_length
35	male	0

$(\text{age} \in [30, 39]) \wedge (\text{gender} = \text{male}) \Rightarrow (\text{hair\_length} = 0)$

DONNÉES NON BOOLÉENNES

age	gender	hair_length
35	male	0

$(\text{age} \in [30, 39]) \wedge (\text{gender} = \text{male}) \Rightarrow (\text{hair\_length} = 0)$

Catégorisation et encodage 1 parmi n (*one-hot*) des données

twenty	thirty	male	female	bald
0	1	1	0	1

thirty, male $\rightarrow$ bald

Règles d'association

$$X \rightarrow Y \equiv \Big(\forall t \in R : (\forall A \in X : t.A = 1) \Rightarrow (\forall A \in Y : t.A = 1)\Big)$$

FAMILLE DES IMPLICATIONS

Règles d'association

$$X \to Y \equiv \Big(\forall t \in R : (\forall A \in X : t.A = 1) \Rightarrow (\forall A \in Y : t.A = 1)\Big)$$

Dépendances fonctionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

FAMILLE DES IMPLICATIONS

Règles d'association

$$X \to Y \equiv \Big(\forall t \in R : (\forall A \in X : t.A = 1) \Rightarrow (\forall A \in Y : t.A = 1)\Big)$$

Dépendances fonctionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

id	city	department
1	Poitiers	Vienne
2	Châtellerault	Vienne
3	Poitiers	Vienne
4	Niort	Deux-Sèvres

FAMILLE DES IMPLICATIONS

Règles d'association

$$X \to Y \equiv \Big(\forall t \in R : (\forall A \in X : t.A = 1) \Rightarrow (\forall A \in Y : t.A = 1)\Big)$$

Dépendances fonctionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

id	city	department
1	Poitiers	Vienne
2	Châtellerault	Vienne
3	Poitiers	Vienne
4	Niort	Deux-Sèvres

- city $\to$ department

FAMILLE DES IMPLICATIONS

Règles d'association

$$X \to Y \equiv \Big(\forall t \in R : (\forall A \in X : t.A = 1) \Rightarrow (\forall A \in Y : t.A = 1)\Big)$$

Dépendances fonctionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

id	city	department
1	Poitiers	Vienne
2	Châtellerault	Vienne
3	Poitiers	Vienne
4	Niort	Deux-Sèvres

- city $\to$ department
- id $\to$ city, department

FAMILLE DES IMPLICATIONS

Règles d'association

$$X \to Y \equiv \Big(\forall t \in R : (\forall A \in X : t.A = 1) \Rightarrow (\forall A \in Y : t.A = 1)\Big)$$

Dépendances fonctionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

id	city	department
1	Poitiers	Vienne
2	Châtellerault	Vienne
3	Poitiers	Vienne
4	Niort	Deux-Sèvres

- city $\to$ department
- id $\to$ city, department

Dépendances fonctionnelles conditionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in \sigma(R) : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

$\sigma(R)$: sélection sur R (tests d'égalité avec des constantes)

FAMILLE DES IMPLICATIONS

Dépendances fonctionnelles conditionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in \sigma(R) : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

$\sigma(R)$: sélection sur R (tests d'égalité avec des constantes)

Dépendances fonctionnelles approximatives

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A \approx t_2.A) \Rightarrow (\forall A \in Y : t_1.A \approx t_2.A)\Big)$$

FAMILLE DES IMPLICATIONS

Dépendances fonctionnelles conditionnelles

$$X \to Y \equiv \Big(\forall t_1, t_2 \in \sigma(R) : (\forall A \in X : t_1.A = t_2.A) \Rightarrow (\forall A \in Y : t_1.A = t_2.A)\Big)$$

$\sigma(R)$: sélection sur R (tests d'égalité avec des constantes)

Dépendances fonctionnelles approximatives

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A \approx t_2.A) \Rightarrow (\forall A \in Y : t_1.A \approx t_2.A)\Big)$$

Dépendances d'ordre

$$X \to Y \equiv \Big(\forall t_1, t_2 \in R : (\forall A \in X : t_1.A < t_2.A) \Rightarrow (\forall A \in Y : t_1.A < t_2.A)\Big)$$

Nombreux outils pour l'énumération des dépendances [7]

7. Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. *Data profiling with metanome*. Proc. VLDB Endow. 8, 12 (2015), 1860–1863.
https://hpi.de/naumann/projects/data-profiling-and-analytics/metanome-data-profiling/algorithms.html

EXEMPLE DE DÉPENDANCES

Empno	Lastname	Workdept	Job	Educlevel	Gender	Sal	Bonus	Comm	Mgrno
10	SPEN	C01	FINANCE	18	F	52750	500	4220	20
20	THOMP	null	MANAGER	18	M	41250	800	3300	null
30	KWAN	null	FINANCE	20	F	38250	500	3060	10
50	GEYER	null	MANAGER	16	M	40175	700	3214	20
60	STERN	D21	SALE	14	M	32250	500	2580	30
70	PULASKI	D21	SALE	16	F	36170	700	2893	100
90	HENDER	D21	SALE	17	F	29750	500	2380	10
100	SPEN	C01	FINANCE	18	M	26150	800	2092	20

- Workdep $\rightarrow$ Job (dépendance fonctionnelle)
- $[\text{Educlevel} = 18] \wedge \text{Gender} \rightarrow \text{Bonus}$ (dépendance fonctionnelle conditionnelle)
- $[\text{Educlevel} > 16] \wedge \text{Gender} \rightarrow \text{Bonus}$ (généralisation)
- Sal $\rightarrow$ Com (dépendance fonctionnelle approximative : $A \approx B$ ssi $\frac{2|A-B|}{A+B} < 0.1$)
- Sal $\rightarrow$ Com (dépendance d'ordre)

RQL : RULE QUERY LANGUAGE

Généralisation des dépendances précédentes [8]

$$X \rightarrow Y \equiv \Big(\forall t_1, t_2, \ldots, t_n \in R : \\ (\forall A \in X : \delta(A, t_1, t_2, \ldots, t_n)) \Rightarrow (\forall A \in Y : \delta(A, t_1, t_2, \ldots, t_n))\Big)$$

δ : conditions sur les variables de tuples t_i et la variable d'attribut A

8. Brice Chardin, Emmanuel Coquery, Marie Pailloux, Jean-Marc Petit. *RQL : A Query Language for Rule Discovery in Databases.* Theoretical Computer Science, Elsevier, 2017, 658, pp.357-374.

RQL : RULE QUERY LANGUAGE

Généralisation des dépendances précédentes [8]

$$X \rightarrow Y \equiv \Big(\forall t_1, t_2, \ldots, t_n \in R : \\ (\forall A \in X : \delta(A, t_1, t_2, \ldots, t_n)) \Rightarrow (\forall A \in Y : \delta(A, t_1, t_2, \ldots, t_n))\Big)$$

δ : conditions sur les variables de tuples t_i et la variable d'attribut A

Exemple

$\delta(A, t_1, t_2) = (t_1.A = t_2.A)$ (dépendances fonctionnelles)

8. Brice Chardin, Emmanuel Coquery, Marie Pailloux, Jean-Marc Petit. *RQL : A Query Language for Rule Discovery in Databases.* Theoretical Computer Science, Elsevier, 2017, 658, pp.357-374.

RQL : SYNTAXE

```
FINDRULES
OVER <attributs>
SCOPE <variables de tuple>
[WHERE <relation entre les variables de tuple>]
CONDITION ON <variable d'attribut>
          IS <condition δ sur les variables>
```

RQL : EXEMPLES

Dépendances fonctionnelles

```
FINDRULES
OVER Empno, Lastname, Workdept, Job, Gender, Bonus, Mgrno
SCOPE t1, t2 Emp
CONDITION ON A IS t1.A = t2.A
```

RQL : EXEMPLES

Dépendances fonctionnelles

```
FINDRULES
OVER Empno, Lastname, Workdept, Job, Gender, Bonus, Mgrno
SCOPE t1, t2 Emp
CONDITION ON A IS t1.A = t2.A
```

Dépendances fonctionnelles conditionnelles

```
FINDRULES
OVER Empno, Lastname, Workdept, Job, Gender, Bonus, Mgrno
SCOPE t1, t2 (SELECT * FROM Emp WHERE Educlevel > 16)
CONDITION ON A IS t1.A = t2.A
```

RQL : EXEMPLES

Dépendances fonctionnelles approximatives

```
FINDRULES
OVER Educlevel, Sal, Bonus, Comm
SCOPE t1, t2 Emp
CONDITION ON A IS 2*ABS(t1.A-t2.A)/(t1.A+t2.A) < 0.1
```

RQL : EXEMPLES

Dépendances fonctionnelles approximatives

```
FINDRULES
OVER Educlevel, Sal, Bonus, Comm
SCOPE t1, t2 Emp
CONDITION ON A IS 2*ABS(t1.A-t2.A)/(t1.A+t2.A) < 0.1
```

Règles d'association

```
FINDRULES
OVER Coffee, Sugar, Milk, Tea
SCOPE t Transactions
CONDITION ON A IS t.A = 1
```

RQL : EXEMPLES

Dépendances fonctionnelles approximatives

```
FINDRULES
OVER Educlevel, Sal, Bonus, Comm
SCOPE t1, t2 Emp
CONDITION ON A IS 2*ABS(t1.A-t2.A)/(t1.A+t2.A) < 0.1
```

Règles d'association

```
FINDRULES
OVER Coffee, Sugar, Milk, Tea
SCOPE t Transactions
CONDITION ON A IS t.A = 1
```

RQL : énumération des règles avec une confiance de 100%

DENIAL CONSTRAINTS

Contraintes d'intégrité et motifs [9]

$\forall t_1, t_2 \in R : \neg(p_1 \wedge p_2 \wedge \ldots \wedge p_n)$

Dépendances fonctionnelles

$$X \rightarrow A \equiv \forall t_1, t_2 \in R : \neg\left(\bigwedge_{x \in X} (t_1.x = t_2.x) \wedge t_1.A \neq t2.A\right)$$

9. Xu Chu, Ihab F. Ilyas and Paolo Papotti. *Discovering denial constraints*. Proceedings of the VLDB Endowment 6.13 (2013) : 1498-1509.

DENIAL CONSTRAINTS

Contraintes d'intégrité et motifs [9]

$\forall t_1, t_2 \in R : \neg(p_1 \wedge p_2 \wedge \ldots \wedge p_n)$

Dépendances fonctionnelles

$$X \rightarrow A \equiv \forall t_1, t_2 \in R : \neg\left(\bigwedge_{x \in X} (t_1.x = t_2.x) \wedge t_1.A \neq t2.A\right)$$

Énumération en fixant une liste d'opérateurs de comparaison $\{\phi_i\}$ close par la négation

Prédicats p_i de la forme $(t_j.A\phi_i t_k.B)$ ou $(t_j.A\phi_i \text{constante})$

9. Xu Chu, Ihab F. Ilyas and Paolo Papotti. *Discovering denial constraints*. Proceedings of the VLDB Endowment 6.13 (2013) : 1498-1509.

LIEN ENTRE IMPLICATIONS ET DENIAL CONSTRAINTS

Règles d'association

$$X \to A \equiv \Big(\forall t \in R : (\forall x \in X : t.x = 1) \Rightarrow t.A = 1\Big)$$

LIEN ENTRE IMPLICATIONS ET DENIAL CONSTRAINTS

Règles d'association

$$X \to A \equiv \Big(\forall t \in R : (\forall x \in X : t.x = 1) \Rightarrow t.A = 1\Big)$$

$$\begin{aligned} p \Rightarrow q &\equiv \neg p \vee q \\ &\equiv \neg(p \wedge \neg q) \end{aligned}$$

LIEN ENTRE IMPLICATIONS ET DENIAL CONSTRAINTS

Règles d'association

$$
\begin{aligned}
X \to A &\equiv \Big(\forall t \in R : (\forall x \in X : t.x = 1) \Rightarrow t.A = 1\Big) \\
&\equiv \Big(\forall t \in R : \neg\Big((\forall x \in X : t.x = 1) \wedge t.A \neq 1\Big)\Big) \\
&\equiv \Big(\forall t \in R : \neg\Big(\bigwedge_{x \in X} (t.x = 1) \wedge t.A \neq 1\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
p \Rightarrow q &\equiv \neg p \vee q \\
&\equiv \neg(p \wedge \neg q)
\end{aligned}
$$

LIEN ENTRE IMPLICATIONS ET DENIAL CONSTRAINTS

Règles d'association

$$X \rightarrow A \equiv \Big(\forall t \in R : (\forall x \in X : t.x = 1) \Rightarrow t.A = 1\Big)$$
$$\equiv \Big(\forall t \in R : \neg\Big((\forall x \in X : t.x = 1) \wedge t.A \neq 1\Big)\Big)$$
$$\equiv \Big(\forall t \in R : \neg\Big(\bigwedge_{x \in X} (t.x = 1) \wedge t.A \neq 1\Big)\Big)$$

$$p \Rightarrow q \equiv \neg p \vee q$$
$$\equiv \neg(p \wedge \neg q)$$

Même démarche pour les autres types de dépendances

Ne fonctionne que pour un attribut unique en conséquent

EXEMPLE DE DENIAL CONSTRAINTS

Name	Gender	Area	Phone	City	State	Zip	Salary	Tax_rate	Tax_exemption
Ballin	M	304	232-7667	Anthony	WV	25813	5000	3	2000
Black	M	719	154-4816	Denver	CO	80290	60000	4.63	0
Puerta	F	501	378-7304	West Crossett	AR	72045	85000	7.22	0
Landram	M	319	150-3642	Giffort	IA	52404	15000	2.48	40
Murro	M	970	190-3324	Denver	CO	80251	60000	4.63	0
Billinghurst	F	501	154-4816	Kremlin	AR	72045	70000	7	0

EXEMPLE DE DENIAL CONSTRAINTS

Name	Gender	Area	Phone	City	State	Zip	Salary	Tax_rate	Tax_exemption
Ballin	M	304	232-7667	Anthony	WV	25813	5000	3	2000
Black	M	719	154-4816	Denver	CO	80290	60000	4.63	0
Puerta	F	501	378-7304	West Crossett	AR	72045	85000	7.22	0
Landram	M	319	150-3642	Giffort	IA	52404	15000	2.48	40
Murro	M	970	190-3324	Denver	CO	80251	60000	4.63	0
Billinghurst	F	501	154-4816	Kremlin	AR	72045	70000	7	0

Clé (Area, Phone) : $\forall t_1, t_2 \in R, \neg(t_1.Area = t_2.Area \wedge t_1.Phone = t_2.Phone)$

EXEMPLE DE DENIAL CONSTRAINTS

Name	Gender	Area	Phone	City	State	Zip	Salary	Tax_rate	Tax_exemption
Ballin	M	304	232-7667	Anthony	WV	25813	5000	3	2000
Black	M	719	154-4816	Denver	CO	80290	60000	4.63	0
Puerta	F	501	378-7304	West Crossett	AR	72045	85000	7.22	0
Landram	M	319	150-3642	Giffort	IA	52404	15000	2.48	40
Murro	M	970	190-3324	Denver	CO	80251	60000	4.63	0
Billinghurst	F	501	154-4816	Kremlin	AR	72045	70000	7	0

Clé (Area, Phone) : $\forall t_1, t_2 \in R, \neg(t_1.Area = t_2.Area \wedge t_1.Phone = t_2.Phone)$

Zip $\rightarrow$ State : $\forall t_1, t_2 \in R, \neg(t_1.Zip = t_2.Zip \wedge t_1.State \neq t_2.State)$

EXEMPLE DE DENIAL CONSTRAINTS

Name	Gender	Area	Phone	City	State	Zip	Salary	Tax_rate	Tax_exemption
Ballin	M	304	232-7667	Anthony	WV	25813	5000	3	2000
Black	M	719	154-4816	Denver	CO	80290	60000	4.63	0
Puerta	F	501	378-7304	West Crossett	AR	72045	85000	7.22	0
Landram	M	319	150-3642	Giffort	IA	52404	15000	2.48	40
Murro	M	970	190-3324	Denver	CO	80251	60000	4.63	0
Billinghurst	F	501	154-4816	Kremlin	AR	72045	70000	7	0

Clé (Area, Phone) : $\forall t_1, t_2 \in R, \neg(t_1.Area = t_2.Area \wedge t_1.Phone = t_2.Phone)$

Zip $\rightarrow$ **State** : $\forall t_1, t_2 \in R, \neg(t_1.Zip = t_2.Zip \wedge t_1.State \neq t_2.State)$

Tax_exemption $\leq$ **Salary** : $\forall t \in R, \neg(t.Tax\_exemption > t.Salary)$

EXEMPLE DE DENIAL CONSTRAINTS

Name	Gender	Area	Phone	City	State	Zip	Salary	Tax_rate	Tax_exemption
Ballin	M	304	232-7667	Anthony	WV	25813	5000	3	2000
Black	M	719	154-4816	Denver	CO	80290	60000	4.63	0
Puerta	F	501	378-7304	West Crossett	AR	72045	85000	7.22	0
Landram	M	319	150-3642	Giffort	IA	52404	15000	2.48	40
Murro	M	970	190-3324	Denver	CO	80251	60000	4.63	0
Billinghurst	F	501	154-4816	Kremlin	AR	72045	70000	7	0

Clé (Area, Phone) : $\forall t_1, t_2 \in R, \neg(t_1.Area = t_2.Area \wedge t_1.Phone = t_2.Phone)$

Zip $\rightarrow$ **State** : $\forall t_1, t_2 \in R, \neg(t_1.Zip = t_2.Zip \wedge t_1.State \neq t_2.State)$

Tax_exemption $\leq$ **Salary** : $\forall t \in R, \neg(t.Tax\_exemption > t.Salary)$

Denver dans le Colorado : $\forall t \in R, \neg(t.City = \text{'Denver'} \wedge t.State \neq \text{'CO'})$

EXEMPLE DE DENIAL CONSTRAINTS

Name	Gender	Area	Phone	City	State	Zip	Salary	Tax_rate	Tax_exemption
Ballin	M	304	232-7667	Anthony	WV	25813	5000	3	2000
Black	M	719	154-4816	Denver	CO	80290	60000	4.63	0
Puerta	F	501	378-7304	West Crossett	AR	72045	85000	7.22	0
Landram	M	319	150-3642	Giffort	IA	52404	15000	2.48	40
Murro	M	970	190-3324	Denver	CO	80251	60000	4.63	0
Billinghurst	F	501	154-4816	Kremlin	AR	72045	70000	7	0

Clé (Area, Phone) : $\forall t_1, t_2 \in R, \neg(t_1.Area = t_2.Area \wedge t_1.Phone = t_2.Phone)$

Zip $\rightarrow$ **State** : $\forall t_1, t_2 \in R, \neg(t_1.Zip = t_2.Zip \wedge t_1.State \neq t_2.State)$

Tax_exemption $\leq$ **Salary** : $\forall t \in R, \neg(t.Tax\_exemption > t.Salary)$

Denver dans le Colorado : $\forall t \in R, \neg(t.City = \text{'Denver'} \wedge t.State \neq \text{'CO'})$

Pour chaque état, Salary $\propto$ **Tax_rate** :
$\forall t_1, t_2 \in R, \neg(t_1.State = t_2.State \wedge t_1.Salary < t_2.Salary \wedge t_1.Tax\_rate > t_2.Tax\_rate)$