

Fouille de Données

Recherche de motifs fréquents et de règles d'association

ENSTA 2010-2011

Antoine CORNUÉJOLS & Christine MARTIN
AgroParisTech
antoine.cornuejols@agroparistech.fr



Fouille de données : applications

Domaines d'applications

- 1 **Banques** : Quel prospect a le plus de chance d'être un client intéressant ?
- 2 **Vente** : Quel produit ce client sera-t-il tenté d'acheter ?
- 3 **Ministère des finances** : Quelle déclaration de revenu a-t-elle le plus de chance d'être malhonnête ?
- 4 **Sécurité nationale** : Quels événements doivent être considérés comme des indices de menaces pour la sécurité nationale ?

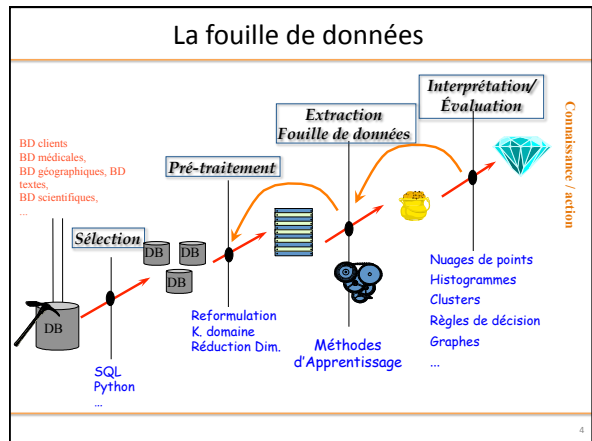
- Envoyer des promotions à certaines catégories de clients
- Regrouper des produits dans un même rayon
- Ne pas envoyer une promotion pour une tondeuse à gazon à quelqu'un qui habite en appartement
- ...

ENSHIE 2011 – Motifs fréquents et règles d'association 2

Fouille de données : définition

- Ensemble de techniques
 - d'aide à la **découverte de régularités** (règles, relations, corrélations, dépendances) pour **expliquer** et **prédire**
 - dans des **grandes bases de données** (traitements rapides)
 - **hétérogènes** (importance des prétraitements)
 - fondées sur des méthodes **statistiques**, d'apprentissage artificiel, de **visualisation**, etc.

Module Intégratif « Fouille de données et extraction de connaissances » 3



Extraction de Connaissances dans les Bases de Données (ECBD)

1. **Le nettoyage des données.** Afin d'éliminer le bruit et les données incohérentes.
2. **L'intégration de données.** Lorsque les données proviennent de sources différentes.
3. **La sélection de données.** Recherche des données pertinentes dans la base de données.
4. **La transformation des données.** Afin de les mettre dans une forme appropriée pour les opérations de recherche de régularités. Cela peut inclure des opérations de résumé ou d'agrégation de données.
5. **La fouille de données.** Extraction de régularités dans les données.
6. **L'évaluation des régularités.** Afin de ne garder que les régularités intéressantes.
7. **La représentation des connaissances extraites.** Des techniques de visualisation et de représentation sont utilisées pour transmettre le plus clairement possible les régularités extraites aux utilisateurs finaux.

Module Intégratif « Fouille de données et extraction de connaissances » 5

Critères de réussite

Pour bien mener un projet de DM, il faut :

- Identifier et énoncer clairement **les besoins**.
- Obtenir des **données de qualité** et représentatives du problème
- Identifier la ou **les méthodes adaptées**
 - Au type de **données**
 - Au type d'**hypothèses** que l'on souhaite
- Prétraiter les données de manière adéquate
- **Valider la pertinence** de la méthode et des résultats

Module Intégratif « Fouille de données et extraction de connaissances » 6

Nettoyage des données

Exemple

Data quality problems in a relational DB

ICDM Steering Committee

Name	Affiliation	City, State, Zip, Country	Phone
Platetsky-Shapiro G., PhD	U. of Massachusetts		617-264-9914
David J. Hand	Imperial College	London, UK	
Benjamin W. Wah	Univ. of Illinois	IL 61801, USA	(217) 333-6903
Hand D. J.			
Vipin Kumar	U. of Minnesota, MI, USA		
Xindong Wu	U. of Vermont	Burlington-4000 USA	
Philip S. Yu	U. of Illinois	Chicago IL, USA	999-999-9999
Osmar B. Zaiane	U. of Alberta	CA	111-111-1111
Laure EQUIL	Ramamohanarao Kolegiri, U. of Melbourne, Australia Heikki Mannila, U. of Helsinki, Finland Shusaku Tsumoto, Shimane Univ., Japan		

Annotations: Non-standard representation, Duplicates, Typos, Misfiled Value, Inconsistency, Obsolete Value, Missing Value, Incorrect Value, Incomplete Value.

3 records are missing!

Qualité des données

Qu'est-ce que c'est ?

Une combinaison subtile de propriétés :

- Précision.** EGC-2010 a eu lieu à Hammamet en Tunisie
- Cohérence.** Il y a une seule conférence EGC par an
- Complétude.** Chaque conférence EGC est localisée quelque part
- Fraîcheur.** Le lieu de la dernière conférence EGC était Brest en France
- Unicité :**
 - EGC est une conférence, pas une congrégation
 - EGC-2010 et *Extraction et Gestion des Connaissances 2010* font référence au même événement

ENSIE 2011 – Motifs fréquents et règles d'association 9

Qualité des données

La qualité des données est un problème fondamental

GIGO : Garbage In Garbage Out

- Omniprésent** dans toutes les applications
- Complexe.** Intrinsic dans toute BD, entrepôt de données ou système d'information.
La frontière entre bonnes données et mauvaises données est imprécise.
- Critique.** Coûte énormément

ENSIE 2011 – Motifs fréquents et règles d'association 10

Opérations

Intégration et transformation. Fusionner différentes sources de données.

- Identifier les entités, par ex. `client_id` et `cl_nr`.
- Uniformiser les données exprimées en unités différentes, par ex. DOLLAR et EURO
- Convertir les adresses en coordonnées
- Calculer la vente quotidienne à partir des ventes individuelles.
- Normaliser les variables entre 0 et 1.
- Remplacer toute valeur x d'un attribut A avec moyenne μ et variance σ^2 par $\frac{x-\mu}{\sigma}$ afin d'arriver à une moyenne de 0 et une variance de 1.

ENSIE 2011 – Motifs fréquents et règles d'association 11

Recherche de motifs fréquents

Recherche de motifs fréquents

- Motivé au départ par l'analyse de tickets de caisse
 - Achats fréquemment associés
 - Bière et couche bébé ?
- Objectif :**
 - découvrir des « motifs » communs à une partie « significative » des données
 - C'est ce que l'on appelle les **motifs fréquents** ou « Frequent ItemSets » (FIS).
- S'appuie sur des principes simples
- Etape algorithmiquement difficile **préalable à la formation de « règles »**

ENSIE 2011 – Motifs fréquents et règles d'association 13

Données

- Échantillon d'observations : $S = \{x_1, \dots, x_m\}$
- Ensemble des attributs : $A = \{a_1, \dots, a_d\}$
 - Descripteurs booléens des observations

Une observation x_i est un sous-ensemble de A
 $(x_i \subseteq A)$
 Constitué de toutes des conditions de A vérifiées par
 x_i

ENSIE 2011 – Motifs fréquents et règles d'association 14

Données

- Exemple concret**
 - Tickets de caisse :**
 - recensement d'achats faits par des individus
 - Les **attributs considérés** sont :
 - pain, chocolat, beurre, sucre
 - Les **observations :**
 - $X_1 = \{\text{pain, beurre, sucre}\}$
 - $X_2 = \{\text{pain, chocolat}\}$
 - $X_3 = \{\text{pain, beurre}\}$

ENSIE 2011 – Motifs fréquents et règles d'association 15

Données

- Les données peuvent être représentées sous forme d'un tableau booléen

X/A	a ₁	a ₂	a ₃	a ₄	a ₅
x ₁	1	0	1	1	0
x ₂	0	1	1	0	1
x ₃	1	1	1	0	1
x ₄	0	1	0	0	1
x ₅	1	1	1	0	1
x ₆	0	1	1	0	1

6 observations sont notées x_i
5 attributs notés a_i

ENSIE 2011 – Motifs fréquents et règles d'association 16

Données

- Exemple concret**
 - Tickets de caisse : recensement d'achats faits par des individus

	Pain	Chocolat	Beurre	Sucre
x ₁	1	0	1	1
x ₂	1	1	0	0
x ₃	1	0	1	0

- L'individu x₁ a acheté du pain, du beurre et du sucre

ENSIE 2011 – Motifs fréquents et règles d'association 17

Notion de couverture

On dit qu'un ensemble d'attributs **I** **couvre** un exemple **x_i** si et seulement si

$I \subseteq x_i$

Cela signifie que x_i vérifie toutes les conditions booléennes exprimées par les attributs de I.

ENSIE 2011 – Motifs fréquents et règles d'association 18

Notion de couverture

Sur cet exemple :

X/L	a	b	c	d	e
x ₁	1	0	1	1	0
x ₂	0	1	1	0	1
x ₃	1	1	1	0	1
x ₄	0	1	0	0	1
x ₅	1	1	1	0	1
x ₆	0	1	1	0	1

- On peut dire que:
 - b couvre x₃
 - b ne couvre pas x₁
 - {c, e} couvre x₂
 - ...

ENSIE 2011 – Motifs fréquents et règles d'association 19

Base de données des monuments de Paris

X/L	BM	CP	LO	MO	ND	TE
x ₁	1		1	1	1	
x ₂	1		1	1	1	
x ₃	1		1	1	1	1
x ₄	1				1	1
x ₅	1		1	1	1	1
x ₆		1	1		1	
x ₇					1	
x ₈			1		1	
x ₉			1		1	
x ₁₀	1	1	1	1	1	

X Touristes
L lieux visités :

- Les bateaux mouches (BM),
- Le centre Pompidou (CP),
- musée du Louvre (LO),
- le musée d'Orsay (MO),
- la cathédrale Notre Dame (ND),
- la tour Eiffel (TE).

- x₁ = {BM, LO, MO, ND}
- {CP} couvre x₆ et x₁₀
- {BM, LO} couvre x₅
- {MO, TE} couvre x₃
- ...

ENSIE 2011 – Motifs fréquents et règles d'association 20

Motifs fréquents

Motif

- Un motif est un **ensemble d'attributs** (itemset) dont la valeur doit être vraie.
- Exemple : {BM, TE} est un motif (signifie : avoir visité BM et TE)

Taille d'un Motif

- La taille t d'un motif m_i est définie comme le cardinal de ses attributs, c'est-à-dire **t(m_i) = | m_i |**.
- Ainsi, la taille du motif {BM, TE} est égale à 2.

Couverture

- La couverture d'un motif est l'**ensemble des exemples couverts par ce motif**.

ENSIE 2011 – Motifs fréquents et règles d'association 21

Motifs fréquents

Support d'un motif

- Le support d'un motif est le **cardinal de la couverture** de ce motif.

Fréquence d'un motif

- La fréquence d'un motif est égale à son **support divisée par le nombre total d'exemples**.

Motif fréquent

- On dira qu'un motif est fréquent si **sa fréquence est supérieure** à un seuil défini a priori noté **f_{min}**.

ENSIE 2011 – Motifs fréquents et règles d'association 22

Motifs fréquents

• Monuments de Paris
• f_{min} = 40%

X/L	BM	CP	LO	MO	ND	TE
x ₁	1		1	1	1	
x ₂	1		1	1	1	
x ₃	1		1	1	1	1
x ₄	1				1	1
x ₅	1		1	1	1	1
x ₆		1	1		1	
x ₇					1	
x ₈			1		1	
x ₉			1		1	
x ₁₀	1	1	1	1	1	

Fréquence({BM, LO}) = 5/10 = 50 %
Donc le motif {BM, LO} est **fréquent**

Fréquence({MO, TE, ND}) = 2/10 = 20%
Donc le motif {MO, TE, ND} n'est **pas fréquent**

ENSIE 2011 – Motifs fréquents et règles d'association 23

Problème de fouille de données

Comment **trouver**
tous les motifs fréquents (ou « Frequent ItemSets » (FIS))
dans une base de donnée de grande taille,
c'est-à-dire tous les motifs dont **la fréquence est supérieure à f_{min}** ?

ENSIE 2011 – Motifs fréquents et règles d'association 24

Motifs fréquents et représentation

Exemple :

- Observations : un échantillon d'humains adultes
- Descripteurs :
 - Numéro de sécurité sociale;
 - Sexe;
 - Âge;
 - Taille;
 - Couleur des cheveux;
 - ...

Extraire les motifs fréquents ?

↓

Problème : les descripteurs ne sont pas à valeurs booléennes

← **Transformer les données**

ENSIIE 2011 – Motifs fréquents et règles d'association 25

Motifs fréquents et représentation

Exemple :

- **Numéro de sécurité sociale**
 - Valeur individuelle → ne peut se répéter fréquemment dans la population → pas exploitable
- **Sexe et Couleur des cheveux**
 - Descripteurs qualitatifs → chaque modalité est considérée comme un attribut
 - Exemple : remplacement de Sexe par 2 attributs H et F;
- **Âge et Taille**
 - Descripteurs quantitatifs
 - Considérer des intervalles parlants
 - Exemple : âge → « moins de 15 ans », « 15-25 ans », « 25-40 ans », « 40-60 ans » et « Plus de 60 ans »

ENSIIE 2011 – Motifs fréquents et règles d'association 26

Extraction de motifs fréquents

- **Nombre total de motifs** pour un ensemble d'attributs de taille D :

$$2^D - 1$$
- Exemple :
 - pour 100 attributs → 1.27×10^{30} motifs à considérer

↓

Une stratégie d'élagage est nécessaire !!!

ENSIIE 2011 – Motifs fréquents et règles d'association 27

Propriété

Antimonotonie de la couverture et du support

- Si $m_1 \subseteq m_2$ alors :
 - $Couverture(m_2) \subseteq Couverture(m_1)$
 - $Support(m_2) \leq Support(m_1)$

↓

Si m_2 est fréquent (au niveau f_{min}) alors m_1 est fréquent également

Inversement si m_1 n'est pas fréquent n'est pas fréquent non plus !

ENSIIE 2011 – Motifs fréquents et règles d'association 28

Algorithme : principes

La propriété d'**antimonotonie** du support

↓

un motif ne peut être fréquent que si tous ses sous-motifs sont fréquents.

↓

Méthode de génération ascendante (« bottom-up »)

On peut **élaguer l'espace de recherche** en n'examinant pas les sur-motifs (sur-ensemble d'attributs) d'un motif non fréquent.

ENSIIE 2011 – Motifs fréquents et règles d'association 29

Motifs fréquents

- Monuments de Paris
- $f_{min} = 40\%$

X/L	BM	CP	LO	MO	ND	TE
x_1	1		1	1	1	
x_2	1		1	1	1	
x_3	1		1	1	1	1
x_4	1					1
x_5	1		1	1	1	1
x_6		1	1		1	
x_7						1
x_8			1		1	
x_9			1		1	
x_{10}	1	1	1	1	1	1

Frequency({MO, TE, ND}) = 20%

Donc le motif {MO, TE, ND} n'est pas fréquent

→ Tous les motifs contenant {MO, TE, ND} n'ont pas à être examinés !

ENSIIE 2011 – Motifs fréquents et règles d'association 30

Algorithme : principe

- 1 On cherche tous les **motifs de taille 1 fréquents**.
- 2
 - 2.1 On engendre les **motifs candidats de taille 2** :
 - ensembles de la forme $\{m_A, m_B\}$ où m_A et m_B sont de taille 1 et fréquents.
 - 2.2 On ne **retient que les motifs fréquents parmi eux**.
 - 3 On construit les **motifs candidats de taille 3**, puis on **sélectionne** parmi eux les motifs fréquents.

... Et ainsi de suite.

Cet algorithme peut être vu comme le parcours avec élagage du treillis des parties de A (ensemble des attributs)

ENSIE 2011 – Motifs fréquents et règles d'association
31

Algorithme Apriori

- **Données** : S collection d'exemples décrits par les attributs A, f_{min} le seuil de fréquence
- **Résultat** : M_f l'ensemble des motifs fréquents avec leur fréquence associée

```

début
i = 1 ; Ci = {{A}} // Ci ensemble des candidats à l'étape i
tant que Ci ≠ ∅ ; faire
  pour m ∈ Ci faire // Etape 1 : Passe sur la base de données
    si fréquence(m) ≥ f_min alors // Tester si m est fréquent. Si oui, l'ajouter à Mf
      L_i ← L_i ∪ {m}
    fin si
  fin pour
  i ← i + 1 // Etape 2 : Engendrer les motifs candidats d'ordre supérieur
  C_i ← generer_candidats (L_{i-1}) // motifs de taille i dont tous les sous-motifs sont fréquents
  M_f ← M_f ∪ (L_i)
fin tant que
retourner M_f
fin
    
```

ENSIE 2011 – Motifs fréquents et règles d'association
32

Génération des candidats

- C_{k+1} : ensemble des candidats de taille k+1
- Pour tout m_A et m_B fréquents de taille k (inclus dans L_k)
 - Si m_A et m_B ont une intersection de taille k-1 notée m_{inter}
 - Et si toutes les parties de taille k-1 du motif :

$$m_C = m_{inter} \cup (m_A \setminus m_{inter}) \cup (m_B \setminus m_{inter})$$
 sont incluses dans L_k
 - Alors on ajoute le motif m_C à C_{k+1}
- Retourner C_{k+1}

ENSIE 2011 – Motifs fréquents et règles d'association
33

Algorithme Apriori : propriétés

Complexité :
fonction de deux étapes :

- **Génération des candidats** (Algorithme en « largeur-d'abord ») :
 - On cherche tous les $\{m_A, m_B\}$ appartenant à L_k dont l'union est de taille i + 1.
 - de l'ordre de $|L_k|^2$ ensembles à considérer au maximum.
- **Calcul des fréquences** dans la base de données.
 - l'opération est de complexité de l'ordre de $|L_k|$.
 - La complexité en **pire cas est donc en $O(|L_k|^3)$** .

En pratique, la complexité est souvent linéaire en $|L_k|$ car :

- **matrices très clairsemées** → peu d'intersections entre motifs.

ENSIE 2011 – Motifs fréquents et règles d'association
34

Recherche de motifs fréquents

	A	B	C	D
E1	1	0	1	0
E2	1	1	1	0
E3	0	1	1	1
E4	0	1	0	1
E5	1	1	1	0

Min_sup = 2

Algorithme par exploration en largeur d'abord et élagage

Motifs fréquents { ○ }

ENSIE 2011 – Motifs fréquents et règles d'association
35

Optimisation de l'algorithme


- Diminuer le nombre d'accès à la base
- Faire varier le seuil
 - Le seuil est fixé par l'expert
 - Plus le seuil est fort et plus de temps de calcul est faible

Trouver un compromis (temps de calcul raisonnable et nombre de motifs extrait satisfaisant)

ENSIE 2011 – Motifs fréquents et règles d'association
36

Optimisation

- Ne considérer que des motifs particuliers
 - Motifs maximaux (au sens de l'inclusion)
 - Motifs fermés
 -



Stratégie de croissance de frontière

ENSIE 2011 – Motifs fréquents et règles d'association 37

Recherche de Règles d'association

Règles d'association

Définition

- Les **règles d'association R** sont de la forme :

$$p_1 \rightarrow p_2$$
 Où p_1 et p_2 sont deux motifs.
- Une telle règle peut se lire :
 - Si un objet x possède p_1 alors il est plausible que x possède p_2

Exemple

- Si un client a acheté du pain et du beurre il est plausible qu'il ait aussi acheté de la confiture

ENSIE 2011 – Motifs fréquents et règles d'association 39

Les règles d'association et évaluation

- Les règles d'association sont l'une des **représentations** les plus **populaires** pour l'expression de **corrélations** locales en fouille de données.
- On définit 2 critères principaux de qualité d'une règle d'association :
 - **Support** : **utilisabilité, fiabilité** (E.g. dans $x\%$ des transactions on observe ...)
 - **Confiance** : **certitude des règles découvertes, précision** (E.g. $y\%$ des transactions vérifiant les prémisses de la règle, vérifient aussi les conclusions).

ENSIE 2011 – Motifs fréquents et règles d'association 40

Confiance et Support

Couverture ou support d'une règle d'association

On appelle *couverture* (ou *support*) de $a_1 \Rightarrow a_2$ la probabilité $P(a_1, a_2)$ que a_1 et a_2 soient $\forall x.a.i$ en même temps. Comme $P(a_1, a_2) = P(a_2, a_1)$, la couverture de $a_1 \Rightarrow a_2$ est la même que celle de $a_2 \Rightarrow a_1$.

On parle aussi de *lift* : $lift(X \rightarrow Y) = \frac{P(X \cap Y)}{P(X) \cdot P(Y)}$

Confiance d'une règle d'association

On appelle *confiance* de $a_1 \Rightarrow a_2$ la probabilité $P(a_2|a_1)$ que a_2 soit vérifiée quand a_1 l'est.

ENSIE 2011 – Motifs fréquents et règles d'association 41

Exemple

Monuments de Paris

- Quel est le support et la confiance de la règle :

X/L	BM	CP	LO	MO	ND	TE
X ₁	1		1	1	1	
X ₂	1		1	1	1	
X ₃	1		1	1	1	1
X ₄	1				1	1
X ₅	1		1	1	1	1
X ₆		1	1		1	
X ₇					1	
X ₈			1		1	
X ₉			1		1	
X ₁₀	1	1	1	1	1	

BM \wedge MO \rightarrow TE ?

Support (BM \wedge MO \rightarrow TE) = P (BM \wedge MO \wedge TE) = 2/10

Confiance (BM \wedge MO \rightarrow TE) = P(TE | BM \wedge MO) = 2/5

ENSIE 2011 – Motifs fréquents et règles d'association 42

Règles valides

- Une règle **R valide** est une règle qui vérifie à la fois :
 - Support (R) \geq support_{min}
 - Confiance (R) \geq confiance_{min}
 - Avec support_{min} et confiance_{min} deux seuils fixés
- On dit qu'une règle est **exacte** (\neq certaine) si sa **confiance vaut 1**
- Dans le cas contraire on la dit **approximative**

43

Motifs fréquents et recherche de règles

- **Problème** : Comment extraire toutes les règles valides ?
- La détermination des **motifs fréquents** permet de trouver un sur-ensemble de règles intéressantes
- De nouvelles mesures de « qualité » dont la **confiance** vont permettre de filtrer les règles dans ce sur-ensemble

44

Principe

Deux étapes principales :

- 1 La **recherche de motifs fréquents**. Par définition, ces motifs, ou itemsets, ont une fréquence au moins aussi grande que le seuil de support minimum : f_{min} .
- 2 La **recherche de règles d'association fortes à partir des motifs fréquents** qui vérifient à la fois un support et une confiance minimales.

Étape 2 : Recherche des règles d'association à partir d'un motif fréquent ℓ

- pour tout motif fréquent ℓ , engendre tous les sous-ensembles non vides de ℓ .
- pour tout sous-ensemble non vide s de ℓ , produire en sortie la règle « $s \Rightarrow (\ell \setminus s)$ » si $\frac{\text{support}(\ell)}{\text{support}(s)} \geq \text{min_confiance}$, où min_confiance est le seuil minimal de confiance exigé.

45

Illustration de l'étape 2

Soit la table de transactions :

Observations	Attributs				
	a1	a2	a3	a4	a5
x1	a1	a2			a5
x2		a2		a4	
x3		a2	a3	a4	
x4	a1	a2		a4	
x5	a1		a3		
x6		a2	a3		
x7	a1		a3		
x8	a1	a2	a3		a5
x9	a1	a2	a3		

Supposons que la table contient le motif fréquent $\ell = \{a1, a2, a5\}$. Quelles règles d'association peuvent être engendrées ?
 Les sous-ensembles non vides de $\{a1, a2, a5\}$ sont : $\{a1, a2\}$, $\{a1, a5\}$, $\{a2, a5\}$, $\{a1\}$, $\{a2\}$ et $\{a5\}$.
 Les règles d'association correspondantes sont les suivantes, avec leur confiance associée :

$a1 \wedge a2 \Rightarrow a5$	confiance = $2/4 = 50\%$
$a1 \wedge a5 \Rightarrow a2$	confiance = $2/2 = 100\%$
$a2 \wedge a5 \Rightarrow a1$	confiance = $2/2 = 100\%$
$a1 \Rightarrow a2 \wedge a5$	confiance = $2/6 = 33\%$
$a2 \Rightarrow a1 \wedge a5$	confiance = $2/7 = 29\%$
$a5 \Rightarrow a1 \wedge a2$	confiance = $2/2 = 100\%$

46

Autre exemple : recherche de motifs fréquents

Transaction ID	Liste d'items
T100	I1, I2, I5
T121	I2, I4
T124	I2, I3
T201	I1, I2, I4
T209	I1, I3
T287	I2, I3
T342	I1, I3
T378	I1, I2, I3, I5
T432	I1, I2, I3

47

Autre exemple : recherche de motifs fréquents

Transaction ID	Liste d'items
T100	I1, I2, I5
T121	I2, I4
T124	I2, I3
T201	I1, I2, I4
T209	I1, I3
T287	I2, I3
T342	I1, I3
T378	I1, I2, I3, I5
T432	I1, I2, I3

Parcourir C1 pour compter le nombre de chaque candidat

Itemset	Nombre
(I1)	6
(I2)	7
(I3)	6
(I4)	2
(I5)	2

Comparer à Min-sup

Itemset	Nombre
(I1)	6
(I2)	7
(I3)	6
(I4)	2
(I5)	2

Engendrer C2 avec les comptes

Itemset	Nombre
(I1, I2)	4
(I1, I3)	4
(I1, I4)	1
(I1, I5)	2
(I2, I3)	4
(I2, I4)	2
(I2, I5)	2
(I3, I4)	0
(I3, I5)	1
(I4, I5)	0

Comparer à Min-sup

Itemset	Nombre
(I1, I2)	4
(I1, I3)	4
(I1, I5)	2
(I2, I3)	4
(I2, I4)	2
(I2, I5)	2

48

Autre exemple : recherche de motifs fréquents

Engendrer C2
avec les comptes

Itemset	Nombre
{1,2}	4
{1,3}	4
{1,4}	1
{1,5}	2
{2,3}	4
{2,4}	2
{2,5}	2
{3,4}	0
{3,5}	1
{4,5}	0

Comparer à
Min-sup

Itemset	Nombre
{1,2}	4
{1,3}	4
{1,5}	2
{2,3}	4
{2,4}	2
{2,5}	2

Engendrer C3
avec les comptes

Itemset	Nombre
{1,2,3}	2
{1,2,5}	2

Comparer à
Min-sup

Itemset	Nombre
{1,2,3}	2
{1,2,5}	2

ENSIE 2011 – Motifs fréquents et règles d'association 49

Autre exemple : recherche de motifs fréquents

Soit la liste des motifs fréquents $\ell = \{I1, I2, I5\}$.

$I1 \wedge I2 \Rightarrow I5$	Confiance = $2/4 = 50\%$
$I1 \wedge I5 \Rightarrow I2$	Confiance = $2/2 = 100\%$
$I2 \wedge I5 \Rightarrow I1$	Confiance = $2/2 = 100\%$
$I1 \Rightarrow I2 \wedge I5$	Confiance = $2/6 \approx 33\%$
$I2 \Rightarrow I1 \wedge I5$	Confiance = $2/7 \approx 29\%$
$I5 \Rightarrow I1 \wedge I2$	Confiance = $2/2 = 100\%$

Si seuil = 70%, alors trois règles seulement sont produites.

ENSIE 2011 – Motifs fréquents et règles d'association 50

Règles valides

- Dans le pratique les mesures de support et de confiance fournissent
 - un grand nombre de règles
 - Pas toujours significatives

Besoin d'autres critères de sélection

ENSIE 2011 – Motifs fréquents et règles d'association 51

Principe de la recherche de règles

- Dans la pratique
 - On y rajoute des critères supplémentaires pour
 - limiter le nombre de règles produites
 - augmenter leur significativité
 - Préférence pour des règles à antécédent minimal et conclusion maximale
 - Néanmoins le passage à l'échelle reste difficile
- Comment choisir le support et la confiance de référence ou autres seuils?

ENSIE 2011 – Motifs fréquents et règles d'association 52

Évaluation

Les notions de *support* et de *confiance* sont **intelligibles**, ce qui est très séduisant pour l'utilisateur non expert. Leur utilisation exclusive présente cependant certains **inconvenients** et **limites**.

- Tendence à engendrer **beaucoup de règles**
- Si on veut des **règles exceptionnelles**, il faut baisser le support, mais explosion du nombre de règles
- Engendre aussi des règles du type : $a \Rightarrow b$ avec $P(b|a) = P(b)$, ce qui signifie que a et b sont **indépendants**.

⇒ recours à d'**autres mesures d'intérêt**.

ENSIE 2011 – Motifs fréquents et règles d'association 53

Mesures d'intérêt


$X \Rightarrow Y$

- **Confiance.** $\frac{|X \cap Y|}{|X|}$
- **Lift.** $\frac{|X \cap Y|}{|X| \cdot |Y|}$
- **Leverage.** The proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other.
- **Conviction.** $\frac{Pr(X) \cdot Pr(\neg Y)}{Pr(X \wedge \neg Y)}$

ENSIE 2011 – Motifs fréquents et règles d'association 54

Autres mesures de « qualité »

- On parle aussi de lift : $\text{lift}(X \rightarrow Y) = P(Y|X)/P(Y)$



Lift >1 → la règle est importante
Lift <=1 → la règle ne sert strictement à rien !

- Lift (fumer → cancer) = 3 signifie qu'on a 3 fois plus de chances d'avoir un cancer lorsqu'on est fumeur !
- En revanche, ne peut être calculé qu'après coup pour filtrer les règles mais ne peut servir à guider la recherche

ENSIIE 2011 – Motifs fréquents et règles d'association 55


Autres mesures de « qualité »

- Pourcentage de réussite
 - Taux d'individus concernés par la règle
 - Réussite $(X \rightarrow Y) = \text{confiance}(X \rightarrow Y) * \text{support}(X \rightarrow Y)$
 - Réussite $(X \rightarrow Y) = P(X, Y)$

ENSIIE 2011 – Motifs fréquents et règles d'association 56

Autres mesures de « qualité »

- Capacité de déploiement
 - Pourcentage d'individus qui vérifient les conditions mais pas les résultats
 - Capacité $(X \rightarrow Y) = P(X, \text{non } Y)$



Plus la capacité de déploiement est faible et plus la règle est intéressante

ENSIIE 2011 – Motifs fréquents et règles d'association 57

Autres mesures de « qualité »

- Progression brute :
 - $PB(X \rightarrow Y) = \text{confiance}(X \rightarrow Y) - P(Y)$
 - Une progression brute intéressante est > 0

ENSIIE 2011 – Motifs fréquents et règles d'association 58

Autres mesures de « qualité »

- ...
- Liste non exhaustive !
- De nombreux travaux sur le sujet
 - Déterminer des mesures
 - plus discriminantes
 - plus pertinentes
 - adaptées à des domaines précis
 -

ENSIIE 2011 – Motifs fréquents et règles d'association 59

Outils

- Commerciaux :
 - Clementine (SPSS)
 - Entreprise Miner (SAS)
 - Intelligent Miner (IBM)
- Libres
 - Weka (universitaire)
 - Orange
 - Tanagra

Voir <http://www.kdnuggets.com/>

Module Intégratif « Fouille de données et extraction de connaissances » 60