EGC'2010, Hammamet

Découverte des Dépendances Fonctionnelles Conditionnelles Fréquentes

Thierno DIALLO
LIRIS Lyon

Noël NOVELLI
LIF Marseille

Contexte

- Qualité des données
- Nettoyage des données
- Dépendances entre les données

- Dépendances Fonctionnelles Exactes
- Dépendances Fonctionnelles Approximatives
- Règles d'Association

Contexte

- Dépendances Fonctionnelles Exactes

  La DF $X \rightarrow Y$ est valide dans r si

  $\forall\ t_1, t_2 \in r$, si $t_1[X] = t_2[X]$ alors $t_1[Y] = t_2[Y]$

- Dépendances Fonctionnelles Approximatives

  La DFA est valide en fonction d'une mesure d'erreur

  Exemple : *g3* compte le nombre minimum de tuples à supprimer pour qu'une DF soit valide

- Règles d'Association

  Ensemble d'items et ensemble de transactions d'items ➔ les items les plus présents

Contexte

- Dépendances Fonctionnelles Exactes

Trop strictes

- Dépendances Fonctionnelles Approximatives

Trop globales

- Règles d'Association

Trop détaillées et pas assez structurées

Contexte

DF

Trop strictes

Trop globales

RA

Trop détaillées

Pas assez structurées

Dépendances Fonctionnelles Conditionnelles

[Bohannon et al. ICDE'07]

Formalisation

[Bohannon et al. ICDE'07]

Soit R un schéma de relation et r une relation sur R

Une DFC θ sur R est une paire ($X \rightarrow Y$, Tp) où

- $X, Y \subseteq R$;
- $X \rightarrow Y$ est une DF ;
- Tp est un pattern tableau sur R.

Illustration

r	A	B	C	D
t_1 :	0	1	0	2
t_2 :	0	1	3	2
t_3 :	0	0	0	1
t_4 :	2	2	0	1
t_5 :	2	1	0	1

DF

AB → D

BD → A

…

DFA

g_3(A → B) = 2 {t_3 et (t_4 ou t_5)}

g_3(C → B) = 2 {t_3 et t_4}

g_3(C → D) = 1 {t_1}

…

RA : Items = { 0, 1, 2, 3 }

Freq(0) = 5, Freq(1) = 5,

Freq(2) = 4, Freq(3) = 1,

Freq(0, 1) = 5, Freq(0, 2) = 4,

Freq(0, 3) = 1, Freq(1, 2) = 4,

Freq(1, 3) = 1, Freq(2, 3) = 1,

…

DFC

(A → CD (2 || 0, 1))

(B → ACD (0 || 0, 0, 1))

(D → AB (2 || 0, 1))

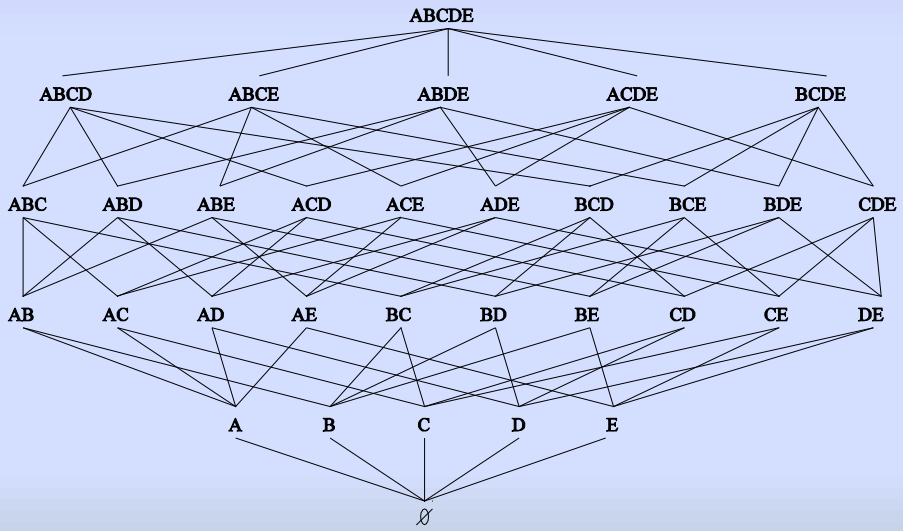
(AB → D (_ || _))

…

Problématique

Extraire une couverture de l'ensemble des DFC valides à partir d'une relation existante

- Pour une relation *r* de schéma *R = { A, B, C, D, E}*
  $\forall X, Y \subseteq R$, évaluée (X → Y (Tp))
  - *Ensemble des parties de R*
  - *Déterminer le Tp de chaque DFC => sous-ensemble de r*

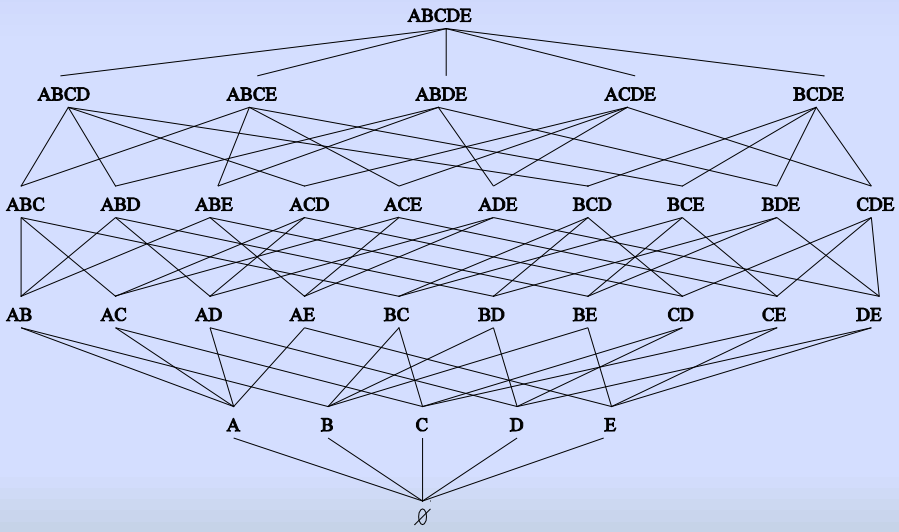


Problématique

Extraire une couverture de l'ensemble des DFC valides à partir d'une relation existante

- Pour une relation *r* de schéma *R = { A, B, C, D, E}*
  $\forall X \subseteq R$ *et* $A \in R$, évaluée (X \ A→ A (Tp))

➔ *Extraire uniquement les DFC minimales et non triviales dont la partie droite est réduite à un seul attribut.*



Autres approches

- Découverte des DFC fréquentes minimales et non triviales
- Adaptation d'approches d'extraction de DF pour les DFC [Fan et al. ICDE'09]
  - *TANE* ➔ *CTANE*
  - *FDMiner* ➔ *CFDMiner*
  - *FastFD* ➔ *FastCFD*

Cadre théorique

- Espace de recherche

  Soit *R* un schéma de relation et *r* une relation sur *R*. L'espace de recherche des DFC constantes pour *r* est défini comme suit:

  $\mathbf{ASP_{CFD}}\ (\boldsymbol{R}, \boldsymbol{r}) = \{\ (\boldsymbol{A}, \boldsymbol{a}) \mid \boldsymbol{A} \in \boldsymbol{R}, \boldsymbol{a} \in \mathbf{ADOM}(\boldsymbol{A}, \boldsymbol{r})\ \}$

r	A	B	C	D
t_1 :	0	1	0	2
t_2 :	0	1	3	2
t_3 :	0	0	0	1
t_4 :	2	2	0	1
t_5 :	2	1	0	1

$\mathbf{ASP_{CFD}}\ (\boldsymbol{ABCD}, \boldsymbol{r}) = \{\ (\boldsymbol{A,0}), (\boldsymbol{A,2}), (\boldsymbol{B,1}), (\boldsymbol{B,0}), (\boldsymbol{B,2}), (\boldsymbol{C,0}), (\boldsymbol{C,3}), (\boldsymbol{D,2}), (\boldsymbol{D,1})\ \}$

On note $\overline{A}$, le couple (A, v)

DFC Fréquentes

- Soit $\theta = (\overline{X} \rightarrow \overline{Y})$ une DFC constantes sur R et r une relation sur R. La fréquence de θ dans r est :

$$freq(\theta, r) = |\sigma_{\wedge_{(A,v)\in \overline{X}\cup\overline{Y}}(A=v)}(r)|$$

- Une DFC θ est dite fréquente par rapport à ε si **$freq(\theta, r) \geq \varepsilon$**

- $\overline{X}, \overline{Y} \subseteq ASP_{CFD}(R, r)$ tel que $\overline{X} \subseteq \overline{Y}$ et ε un seuil, nous avons : **$freq(Y, r) \geq \varepsilon$ => $freq(X, r) \geq \varepsilon$**

DFC Fréquentes

- Soit R un schéma de relation et r une relation sur R, $\overline{X}, \overline{Y} \subseteq \mathrm{ASP}_{\mathrm{CFD}}(R, r)$ et deux formules de sélection sur X et Y, on a:

$$r \models \overline{X} \rightarrow \overline{Y} \text{ ssi } |\sigma_{C_{\overline{X}}}(r)| = |\sigma_{C_{\overline{X}} \wedge C_{\overline{Y}}}(r)|$$

- Ensemble Conditionnel Libre

Soit $\overline{X} \subseteq ASP_{CFD}(R, r)$ un ensemble d'attributs conditionnels.
$\overline{X}$ est un ensemble conditionnel libre dans r ssi $\nexists \overline{X'} \subseteq \overline{X}$ tel que $|\sigma_{C_{\overline{X'}}}(r)| = |\sigma_{C_{\overline{X}}}(r)|$.

- $\overline{X}, \overline{Y} \subseteq \mathrm{ASP}_{\mathrm{CFD}}(R, r)$ tel que $\overline{X} \subseteq \overline{Y}$. Si $\overline{Y}$ est un ensemble conditionnel libre alors $\overline{X}$ l'est aussi.

DFC Fréquentes

- Fermeture d'un ensemble d'attributs conditionnels

$$\overline{X}^{*}_{\Sigma_r} = \overline{X} \cup \{\overline{A} \mid \overline{A}.att \in R - \overline{X}.att \wedge |\sigma_{C_{\overline{X}}}(r)| = |\sigma_{C_{\overline{X}} \wedge C_{\overline{A}}}(r)|\}$$

- Quasi-Fermeture d'un ensemble d'attributs conditionnels

$$\overline{X}^{\diamond}_{\Sigma_r} = \overline{X} \cup \bigcup_{\overline{A} \in \overline{X}} (\overline{X} - \overline{A})^{*}_{\Sigma_r}$$

- Fermeture d'un ensemble d'attributs conditionnels

$$\overline{X}^{*}_{\Sigma_r} = \overline{X}^{\diamond}_{\Sigma_r} \cup \{\overline{A} \mid \overline{A}.att \in R - \overline{X}^{\diamond}.att \wedge |\sigma_{C_{\overline{X}}}(r)| = |\sigma_{C_{\overline{X}} \wedge C_{\overline{A}}}(r)|\}$$

DFC Fréquentes

- L'ensemble de DFC fréquentes, minimales non triviales

$$\{\overline{X} \rightarrow \overline{A} \mid \overline{X} \in \mathcal{CFS}_r, freq(\overline{X}, r) \geq \epsilon \; et \; \overline{A} \in \overline{X}^{*}_{\Sigma_r} - \overline{X}^{\diamond}_{\Sigma_r}\}$$

Algorithme et implémentation

- Algorithme par niveau

Algorithm CFUN

1 $L_0 := < \overline{\emptyset}, 1, \overline{\emptyset}, \overline{\emptyset} >$
2 $L_1 := \{ < \overline{A}, |\overline{A}|, \overline{A}, \overline{A} > | \overline{A} \in ASP_{CFD}(R, r) \wedge |\overline{A}.att| = 1 \}$
3 **for** ($k := 1$; $L_k \neq \emptyset$; $k := k + 1$) **do**
4 ComputeClosure(L_{k-1}, L_k)
5 ComputeQuasiClosure(L_k, L_{k-1})
6 DisplayCFD(L_{k-1})
7 PruneNonFreeSets(L_k, L_{k-1})
8 L_{k+1} := GenerateCandidate(L_k)
9 DisplayCFD(L_{k-1})

end CFUN

Algorithme et implémentation

- Génération des candidats
  - { (A, a_1), (A, a_2) } inutile
  - { (A, a_1), (B, b_1) } inutile si cette combinaison n'est pas présente dans la relation
- Utilisation de partitions pour les calculs
- Consommation mémoire ($k = |R|$, $n = |r|$)
  - Partitions : 2^k partitions de n éléments au maximum.
  - Quadruplets : 4 éléments.

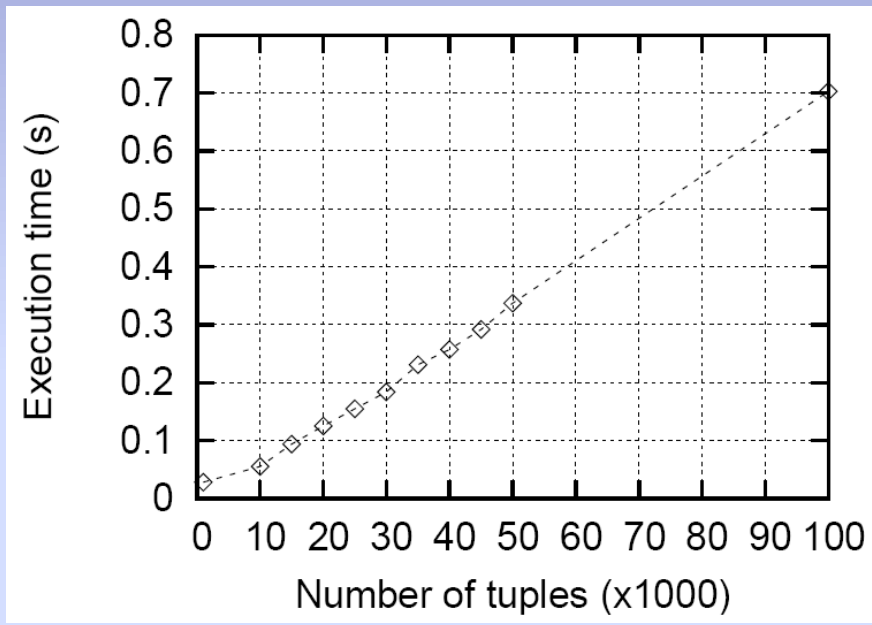
    ➔ $4 \times 4 \times n \times 2^k$
  - Algorithme par niveau donc seuls deux niveaux

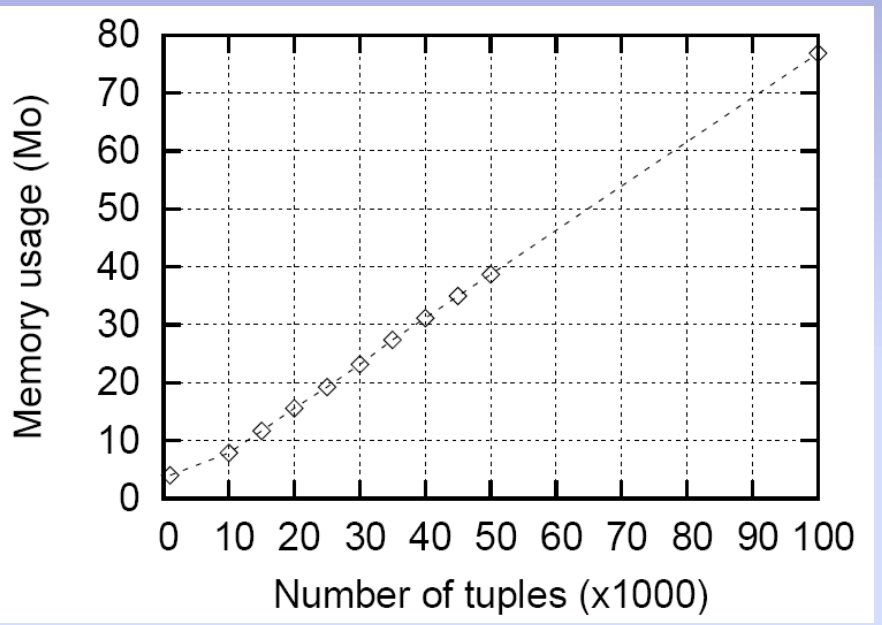
    ➔ $4 \times 4 \times n \times 2 \times binomial(\ k, k/2)$

Expérimentations

- Implémenté en C++ (Windows et Linux)
- Pentium Centrino 2GHz, 2Go de RAM
- Données synthétiques
  - Variation du nombre d'attributs, de tuples et de la corrélation entre les données

Expérimentations



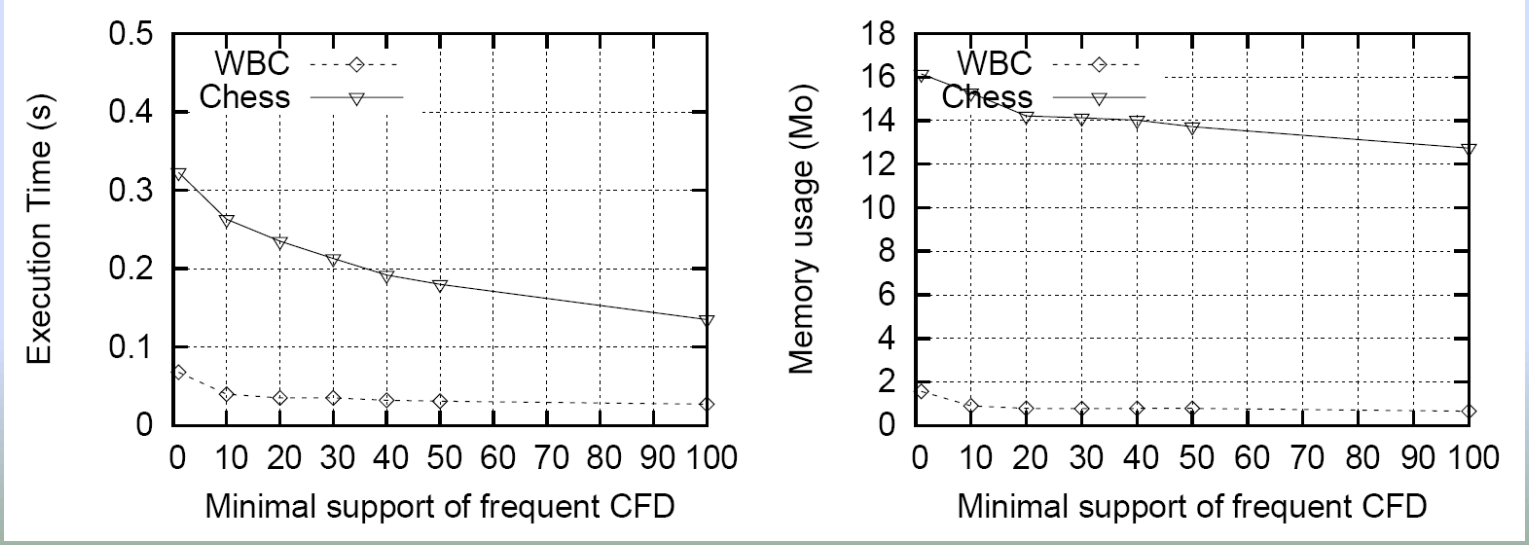


Expérimentations

- Données réelles

Datasets	#Attributs	#Tuples	Taille (Ko)
Wirsconsin Breast Cancer	11	699	19 917
Chess	7	28 056	531 820

Même comportement que les approches antérieures



Conclusion et perspectives

- Nouvelles caractérisations pour les DFC
- Nouvel algorithme par niveau
- Passage à l'échelle

- Amélioration de l'extraction de DFC en utilisant les DF
- Passage aux DFC générales