Inférence de dépendances (suite)

Marc Plantevit
marc.plantevit@liris.cnrs.fr

liris.cnrs.fr/marc.plantevit/doku/doku.php?id=lif10

Outline

Complétude du système d'Armstrong (1/4)

Preuve formelle

Une preuve (formelle) de f à partir de Σ dans le système d'Armstrong notée $\Sigma \vdash f$ est une *séquence* $\langle f_0, \ldots, f_n \rangle$ de DFs telles que $f_n = f$ et $\forall i \in [0..n]$:

- soit $f_i \in \Sigma$;
- soit f_i est la conséquence d'une règle dont toutes les prémisses $f_0 \ldots f_p$ apparaissent avant f_i dans la séquence.

Complétude : prouver que $\Sigma \models X \to Y \Rightarrow \Sigma \vdash X \to Y$

Il faut bien faire la différence entre

- la fermeture sémantique : $X^+ = \{A \mid \Sigma \models X \to A\}$
- la fermeture syntaxique : $X^* = \{A \mid \Sigma \vdash X \to A\}$

Complétude du système d'Armstrong (2/4)

Lemme : $\Sigma \vdash X \to Y \Leftrightarrow Y \subseteq X^\star$

($\Rightarrow$) On suppose sans perte de généralité que $Y = A_1 \dots A_n$. Pour chaque A_i, par réflexivité on a $Y \to A_i$ et on peut prolonger la preuve de $\Sigma \vdash X \to Y$ par transitivité pour obtenir une preuve $\Sigma \vdash X \to A_i$. Ainsi $\forall A_i \in Y$ on a bien $A_i \in X^\star$ et donc $Y \subseteq X^\star$.

($\Leftarrow$) Supposons que $Y \subseteq X^\star = \{A \mid \Sigma \vdash X \to A\}$, par définition de $X^\star$ on a une preuve de $\Sigma \vdash X \to A_i$ pour chaque A_i, dès lors, on peut concaténer toutes ces preuves et conclure par l'application de la règle de composition qu'on a prouvé valide.

Complétude du système d'Armstrong (3/4)

On va avoir le raisonnement logique suivant :

$$\Sigma \models X \to Y \Rightarrow \Sigma \vdash X \to Y$$

$$\equiv \Sigma \nvdash X \to Y \Rightarrow \Sigma \not\models X \to Y$$

$$\equiv \Sigma \nvdash X \to Y \Rightarrow \exists r.(r \models \Sigma \wedge r \not\models X \to Y)$$

L'astuce consiste à exhiber l'instance r,
avec $X^\star = X_1 \dots X_n$ et $Z_1 \dots Z_p = R \setminus X^\star$

r	X_1	$\dots$	X_n	Z_1	$\dots$	Z_p
s	x_1	$\dots$	x_n	z_1	$\dots$	z_p
t	x_1	$\dots$	x_n	y_1	$\dots$	y_p

Cette instance vérifie bien $r \models \Sigma$ et $r \not\models X \to Y$

Complétude du système d'Armstrong (4/4)

- Prouvons $r \models \Sigma$ par l'absurde. Supposons $V \rightarrow W \in \Sigma$ et $r \not\models V \rightarrow W$ pour aboutir à une contradiction. $r \not\models V \rightarrow W$, par définition de $\models$ on a $s[V] = t[V] \wedge s[W] \neq t[W]$, par construction de r cela implique $V \subseteq X^\star$ et $W \not\subseteq X^\star$. D'après le lemme $V \subseteq X^\star$ implique $\Sigma \vdash X \rightarrow V$ et par transitivité avec $V \rightarrow W \in \Sigma$ on obtient $\Sigma \vdash X \rightarrow W$. Or d'après le lemme (dans l'autre sens) on déduit que $W \subseteq X^\star$, une contradiction.
- Prouvons que $r \not\models X \rightarrow Y$. Pour cela, supposons, $r \models X \rightarrow Y$ par construction de r on a $s[Y] = t[Y]$ et donc $Y \subseteq X^\star$. D'après le lemme on a $\Sigma \vdash X \rightarrow Y$, une contradiction.

Le théorème est prouvé : $\Sigma \models X \rightarrow Y \Leftrightarrow \Sigma \vdash X \rightarrow Y$ et $X^\star = X^+$

Les relations d'Armstrong

Le coprs de la preuve de complétude

- La construction de r est un exemple d'une relation qui vérifie toutes les dépendances de Σ de la forme $X \rightarrow Y$ mais ne satisfait pas les autres avec X en partie gauche.
- On va répéter la construction de s, t dans la preuve pour chaque dépendance de Σ et avoir une instance qui concerne toutes les dépendances possibles.

Objectif et intérêt des relations d'Armstrong

- Représenter sur un exemple un ensemble de contraintes et uniquement celui-ci, i.e. toutes les autres contraintes sont fausses.
- Celle permet une représentation *par l'exemple* d'ensemble de dépendances : on manipule des valeurs, visualisation plus simple des éventuels conflits, d'incohérences, de mauvaise conception. . .

Construction de la relation

- Calculer les fermés de F : $Cl(F) = \{X^+ \mid X \subseteq R\}$
- Construire la relation d'Amstrong r correspondante :
  - Etape 0 : pour le fermé particulier R, construire le tuple $t_0 = \langle 0, \ldots, 0 \rangle$
  - Etape i : pour chaque $X \in Cl(F)$, ajouter un tuple t_i à r :
    - tel que $ti[A] = 0$ pour tout $A \in X^+$
    - tel que $t_i[B] = i$ pour tout $B \in (R \setminus X^+)$.

Instance canonique

Avec les mêmes arguments que dans la preuve de complétude :

- r vérifie chaque dépendance : $f \in \Sigma$ implique $r \models f$
- r ne vérifie aucune dépendance qui ne soit pas déductible : $\Sigma \not\models f$ implique $r \not\models f$

Exemple

Considérons :

- la relation $R = \{A, B, C\}$
- l'ensemble de DFs $\Sigma = \{A \rightarrow BC, B \rightarrow C\}$
- $Cl(F) = \{ABC, BC, C\}$

Relation d'Armstrong pour Σ

r	A	B	C
ABC	0	0	0
BC	1	0	0
C	2	2	0

Exercice

Soit Σ l'ensemble de dépendances

$AB \to C$ $\quad C \to A$ $\quad BC \to D$
$ACD \to B$ $\quad D \to E$ $\quad ABE \to C$
$C \to BD$ $\quad CE \to A$

1. Calculer l'ensemble des fermés de Σ ;
2. Construire la base à partir des fermés.

Outline

Système d'inférence de Casanova *et al.*

Soit I un ensemble de DI sur un schéma de base de données $\mathbf{R}$. Les règles d'inférence suivantes sont appelées système d'inférence de Casanova *et al.* pour les DI, dans lequel σ est une permutation d'un sous-ensemble de $\{1..n\}$:

- Réflexivité

$$R[X] \subseteq R[X]$$

- Permutation & projection

$$\frac{R[A_1 \ldots A_n] \subseteq S[B_1 \ldots B_n]}{S[A_{\sigma(1)} \ldots A_{\sigma(k)}] \subseteq S[B_{\sigma(1)} \ldots B_{\sigma(k)}]}$$

- Transitivité

$$\frac{R[X] \subseteq S[Y] \qquad S[Y] \subseteq T[Z]}{R[X] \subseteq T[Z]}$$

Propriétés du système

- Ce système d'inférence est lui aussi
  - correct
  - complet
- Contrairement aux DF, le problème d'inférence des DI est très difficile à traiter dans le cas général (PSPACE-complet) ;
- La notion I^+ s'applique aussi pour noter la fermeture d'un ensemble de DI.
- En revanche, la notion de fermeture d'un *ensemble* d'attributs par rapport à un ensemble de DI ne s'applique pas, car les DI manipulent des séquences et non des ensembles.

Combinaison DF et DI

- soit on inférait des DF à partir d'un ensemble de DF,
- soit on inférait des DI à partir d'un ensemble de DI.

Mais il existe des interactions entre DF et DI

Proposition

Si $|X| = |T|$, les propriétés suivantes sont vérifiées

1. $\{R[XY] \subseteq S[TU], S : T \rightarrow U\} \models R : X \rightarrow Y$
2. $\{R[XY] \subseteq S[TU], R[XZ] \subseteq S[TV], S : T \rightarrow U\} \models R[XYZ] \subseteq S[TUV]$
3. $\Sigma = \{R[XY] \subseteq S[TU], R[XZ] \subseteq S[TV], S : T \rightarrow U\}$, si $r \models \Sigma$, $s \models \Sigma$ et $u \in r$ alors $u[Y] = u[Z]$

Outline

La notion de couverture est une relation d'équivalence entre des ensembles de contraintes.

Couverture d'un ensemble de DFs

Soit Σ et Γ deux ensembles de DFs, Γ est une couverture de Σ *ssi*

$$\Gamma^+ = \Sigma^+$$

- Une couverture d'un ensemble de DF est donc une représentation alternative
- Mais qui possède exactement la même sémantique.
- C'est exactement le même ensemble de DF qui est implicite.
- On a intérêt à choisir de bons représentants au sein des classes.

Des critères pour de bon ensembles équivalents

Propriétés des couvertures

- un ensemble F de DF est dit non redondant s'il n'existe pas de couverture G de F telle que $G \subseteq F$ avec $G \neq F$.
- un ensemble F de DF est dit minimum s'il n'existe pas de couverture G de F tel que $|G| \leq |F|$.
- F est dit optimal s'il n'existe pas de couverture G de F avec moins d'attributs que dans F.

Propriétés immédiates

- une couverture minimum est non redondante ;
- une couverture optimum est minimum.

Algorithme : couverture minimum

Data: F un ensemble de DF
Result: G une couverture minimum de F
$G := \emptyset$
for $X \to Y \in F$ **do**
  $G := G \cup \{X \to X^+\}$
end
for $X \to X^+ \in G$ **do**
  if $G - \{X \to X^+\} \vdash X \to X^+$ **then**
    $G := G - \{X \to X^+\}$
  end
end
return G

- Cet algorithme est polynomial dans le nombre de DF dans F et le nombre d'attributs dans F.
- La couverture minimum calculée par l'algorithme n'est pas forcément unique : d'autres couvertures peuvent avoir le même nombre de DF, mais être différentes.
- Parmi celles-ci, certaines sont optimum ; malheureusement, leur calcul est un problème difficile dans le cas général (NP-Complet).

Exemple

$$F = \{ \begin{array}{lll} AB \to C & C \to A & BC \to D \\ ACD \to B & D \to EF & ABE \to C \\ CF \to BD & CE \to AF & \end{array} \}$$

1. A partir des df $X \to Y$ de F, construire G l'ensemble des règles de la forme $X \to X^+$.
   - $AB \to AB^+ = ABCDEF$
   - ...
2. Supprimer de G les règles inutiles ;
3. Retourner G.

Exercice

Calculer les couvertures des ensembles suivants

F_1

$AB \to D$ $\quad C \to A$ $\quad BC \to D$
$D \to EF$ $\quad BE \to C$ $\quad CF \to B$
$CE \to A$ $\quad CE \to G$

F_2

$AB \to C$ $\quad C \to A$ $\quad BC \to D$
$D \to EF$ $\quad BE \to C$ $\quad CF \to B$
$CE \to F$

Conclusion : ce qu'il faut retenir

Dualité syntaxe et sémantique

- syntaxe des DFs (et des DI)
  - preuve formelle $\Sigma \vdash X \rightarrow Y$
  - fermeture syntaxique $X^\star = \{A \mid \Sigma \vdash X \rightarrow Y\}$
- sémantique l'évaluation de la vérité
  - définition de la sémantique $r \models X \rightarrow Y$
  - fermeture sémantique $X^+ = \{A \mid \Sigma \models X \rightarrow Y\}$

Le système d'Armstrong

- Les règles d'inférence, la preuve formelle
  - le système *pur* : réflexivité, augmentation, transitivité
  - les *raccourcis* : composition, décomposition, pseudo-transitivité
- Le système d'Armstrong est bon
  - correct $\Sigma \vdash X \rightarrow Y \Rightarrow \Sigma \models X \rightarrow Y$
  - complet $\Sigma \models X \rightarrow Y \Rightarrow \Sigma \vdash X \rightarrow Y$
- Construction des relations d'Armstrong

Fin du quatrième cours.