Intégration de Données

Manipulation, récriture de requêtes

Source : Anhai DOAN *et al*,
Principles of Data Integration

Introduction

- Comment un système d'intégration décide quelles sources sont pertinentes pour une requête ? Lesquelles sont redondantes ? Comment les combiner pour répondre à une requête ?
- Réponse : en raisonnant sur le contenu des sources de données.
  - Les sources sont souvent décrites par des requêtes/vues.
- On va étudier les outils fondamentaux pour la manipulation d'expressions de requêtes.

SQL

Entretien

candidat, date, recruteur, décision, note

EvalEmploye

IDemp, nom, evalTrimestre, note, examinateur

```
select recruteur, candidat
from Entretien , EvalEmploye
where recruteur=nom AND
        note < 2.5
```

3

Requêtes Conjonctives

Q(X,T) :- Entretien(X,D,Y,H,F),
EvalEmploye(E,Y,T,W,Z),
W < 2.5.

Noter l’expression des jointures avec les occurrences de la même variable (ici Y)

```
select recruteur, candidat
from Entretien, EvalEmploye
where recruteur=nom AND
        note < 2.5
```

4

Requêtes Conjonctives (prédicats interprétés)

Q(X,T) :-

Entretien(X,D,Y,H,F), EvalEmploye(E,Y,T,W,Z),
W < 2.5.

Prédicats interprétés (comparaison) : on peut retrouver aussi des variables (w)

select recruteur, candidat
from **Entretien**, **EvalEmploye**
where recruteur=nom AND
note < 2.5

Requêtes Conjonctives (sous-but négatif)

Q(X,T) :-

Entretien(X,D,Y,H,F), EvalEmploye(E,Y,T,W,Z),
¬OffreEmploi(X, date).

Toute variable de tête doit apparaitre dans un sous-but positif.

Union de Requêtes Conjonctives

L'union est exprimée avec plusieurs règles avec le même prédicat de tête

Q(X,T) :-

Entretien(X,D,Y,H,F), EvalEmploye(E,Y,T,W,Z), W < 2.5.

Q(X,T) :-

Entretien(X,D,Y,H,F), EvalEmploye(E,Y,T,W,Z), Manager(y), W > 3.9.

7

Dépliage de requêtes

$$Q_1(X,Y) :- Vol(X,Z), Hub(Z), Vol(Z,Y)$$

$$Q_2(X,Y) :- Hub(Z), Vol(Z,X), Vol(X,Y)$$

$$Q_3(X,Z) :- Q_1(X,Y), Q_2(Y,Z)$$

Le dépliage de Q_3 est :

$$Q'_3(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Hub(W), Vol(W,Y), Vol(X,Z)$$

8

Dépliage de requêtes : Algorithme

- Trouver un sous-but $p(X_1, \ldots, X_n)$ tel que p est défini par une règle r.
- Unifier $p(X_1, \ldots, X_n)$ avec la tête de r.
- Remplacer $p(X_1, \ldots, X_n)$ par le résultat de l'application de l'unification à tous les sous-buts de r .
- Itérer jusqu'à épuiser les unifications.
- Si p est défini par une union de $r_1, \ldots, r_n$, créer n règles, une pour chaque règle r_i.

9

Inclusion de requêtes : Motivation (1)

Intuitivement, le dépliage de Q_3 est *équivalent* à Q_4:

$$Q'_3(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Hub(W), Vol(W,Y), Vol(X,Z)$$

$$Q_4(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Vol(X,Z)$$

Comment justifier formellement cette intuition ?

10

Inclusion de requêtes : Motivation (2)

De plus, Q_5 qui utilise 2 hubs est *incluse* dans Q'_3

$$Q_5(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Hub(Y), Vol(Y,Z)$$

$$Q'_3(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Vol(X,Z)$$

Besoin d'algorithmes pour détecter ces inclusions

11

Inclusion de requêtes et équivalence : définitions

Q_1 *est incluse dans* Q_2 si
pour **toute** base de données *D*
$Q_2(D) \supseteq Q_1(D)$

Q_1 est équivalente à Q_2 si
$Q_1(D) \supseteq Q_2(D)$ and $Q_2(D) \supseteq Q_1(D)$

Note: l'inclusion et l'équivalence sont des propriétés des requêtes et non de la base de données!

12

Exemple 1

$$Q_1(X,Z) :- p(X,Y,Z)$$
$$Q_2(X,Z) :- p(X,X,Z)$$

$$Q_1 \supseteq Q_2$$

13

Exemple 2

$$Q_1(X,Y) :- p(X,Z), p(Z,Y)$$
$$Q_2(X,Y) :- p(X,Z), p(Z,Y), p(X,W)$$

$$Q_1 \supseteq Q_2$$

14

Inclusion : le pourquoi?

- Si les sources sont décrites par des vues, on utilise l'inclusion pour les comparer.
- Si on peut supprimer des sous-buts dans une requête, on peut l'évaluer plus efficacement.

15

Reprise de l'exemple

Relations: *Vol(source, destination)*
Hub(ville)

Vues:

$Q_1(X,Y)$:- $Vol(X,Z)$, $Hub(Z)$, $Vol(Z,Y)$
$Q_2(X,Y)$:- $Hub(Z)$, $Vol(Z,X)$, $Vol(X,Y)$

Requête:

$Q_3(X,Z)$:- $Q_1(X,Y)$, $Q_2(Y,Z)$

Dépliage:

$Q'_3(X,Z)$:- $Vol(X,U)$, $Hub(U)$, $Vol(U,Y)$, $Hub(W)$, $Vol(W,Y)$, $Vol(Y,Z)$

16

Supprimer les sous-buts redondants

sous-buts redondants?

Q'₃(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Hub(W), Vol(W,Y), Vol(Y,Z)

$\Rightarrow$

Q''₃(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y), Vol(Y,Z)

Est-ce que Q''_3 est équivalente à Q'_3?

Q'₃(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y) Hub(W), Vol(W,Y), Vol(Y,Z)

17

Inclusion : Requêtes Conjonctives

$$Q_1(\overline{X}) :- g_1(\overline{X_1}),\ldots,g_n(\overline{X_n})$$

Sans prédicats interprétés (≥,≠)
Ni négation.

Sémantique :

si φ mappe les sous-buts du corps de la requête sur des tuples dans *D*

alors, $\varphi(\overline{X})$ est une réponse.

18

Mappings d'inclusion

$$Q_1(\overline{X}) :- g_1(\overline{X_1}),\ldots,g_n(\overline{X_n})$$

$$Q_2(\overline{Y}) :- h_1(\overline{Y_1}),\ldots,h_m(\overline{Y_m})$$

φ: *Vars(Q_1) → Vars(Q_2)*
est un mapping d'inclusion si:

$$\varphi(g_i(\overline{X_i})) \in Corps(Q_2)$$

et $$\varphi(\overline{X}) = \overline{Y}$$

19

Exemple de mappings d'inclusion

Q'$_3$(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y),
Hub(W), Vol(W,Y), Vol(Y,Z)

Q''$_3$(X,Z) :- Vol(X,U), Hub(U), Vol(U,Y),
Vol(Y,Z)

mapping identité sur toutes les variables, sauf:

$$W \rightarrow U$$

20

Théorème
[Chandra and Merlin, 1977]

Q_1 inclut Q_2 ssi il existe un mapping d'inclusion de Q_1 vers Q_2.

21

Union de Requêtes Conjonctives

$$Q_1(X,Y) :- Vol(X,Z), Vol(Z,Y)$$
$$Q_1(X,Y) :- Vol(X,Z), Vol(Y,Z), Hub(Z)$$

$$Q_2(X,Y) :- Vol(X,Z), Vol(Z,Y), Hub(Z)$$

Théorème : une RC est incluse dans une union de RC ssi elle est incluse dans *une* des requêtes conjonctives.

22

RC avec prédicats de comparaison

Une vérification des mappings d'inclusions fournit une condition suffisante :

$$Q_1(\overline{X}) :- g_1(\overline{X_1}),\ldots,g_n(\overline{X_n}),C_1$$
$$Q_2(\overline{Y}) :- h_1(\overline{Y_1}),\ldots,h_m(\overline{Y_m}),C_2$$

φ: *Vars(Q₁) → Vars(Q₂):*

$$\varphi(g_i(\overline{X_i})) \in Corps(Q_2)$$
$$\varphi(\overline{X}) = \overline{Y}$$

et $C_2 \models \varphi(C_1)$

Exemple de mappings d'inclusion

$$Q_1(X,Y) :- Vol(X,Z), Vol(Z,Y),$$
$$Population(Z,P), P \leq 100{,}000$$

$$Q_2(U,V) :- Vol(U,W), Vol(W,V), Hub(W),$$
$$Population(W,S), S \leq 500{,}000$$

$$X \to U, Y \to V, Z \to W$$

$$P \leq 100{,}000 \models P \leq 500{,}000$$

Les mappings d'inclusion ne suffisent pas

$$Q_1(X,Y) :- R(X,Y), S(U,V), U \leq V$$

$$Q_2(X,Y) :- R(X,Y), S(U,V), S(V,U)$$

Pas de mappings d'inclusion, mais

$$Q_1 \supseteq Q_2$$

25

Raffinement de requête

$$Q_1(X,Y) :- R(X,Y), S(U,V), U \leq V$$

$$Q_2(X,Y) :- R(X,Y), S(U,V), S(V,U)$$

On considère les 2 raffinements de Q_2

$$Q_2(X,Y) :- R(X,Y), S(U,V), S(V,U), U \leq V$$
$$Q_2(X,Y) :- R(X,Y), S(U,V), S(V,U), V < U$$

Les mappings rouge s'appliquent au premier raffinement et les bleus au second.

26

Construction des raffinements de requêtes

- Considérer les regroupements complets de toutes les variables et constantes de la requête
- Pour chaque regroupement complet, créer une requête conjonctive.
- Résultat final = union de requêtes conjonctives.

27

Regroupement Complet

Soit

C une conjonction d'atomes interpétés sur
un ensemble de variables $X_1,\ldots,X_n$
un ensemble de constantes $a_1,\ldots,a_m$

C_T est un regroupement complet si: $C_T \models C$, *et*

$$\forall d_1,d_2 \in \{X_1,\ldots,X_n,a_1,\ldots,a_m\}$$

$$C_T \models d_1 < d_2 \quad or \quad C_T \models d_1 > d_2 \quad or \quad C_T \models d_1 = d_2$$

28

Raffinement de requête

$$Q_1(\overline{X}) :- g_1(\overline{X_1}),\ldots,g_n(\overline{X_n}),C_1$$

Soit C_T un regroupement complet de C_1

Alors:

$$Q_1(\overline{X}) :- g_1(\overline{X_1}),\ldots,g_n(\overline{X_n}),C_T$$

est un raffinement de Q_1

29

Théorème: [Requêtes avec prédicats interprétés]

[Klug, 88, van der Meyden, 92]

Q_1 inclut Q_2 ssi il existe un mapping d'inclusion de Q_1 vers tout raffinement de Q_2.

30

Requêtes avec négation

$$Q_1(\overline{X}) :- g_1(\overline{X_1}),\ldots,g_n(\overline{X_n}),$$
$$\neg h_1(\overline{Y_1}),\ldots,\neg h_k(\overline{Y_k})$$

Requêtes sûres :

toute variable de tête apparait dans un sous-but positif du corps.

Mappings d'inclusions :

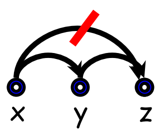
sous-buts négatifs de Q_1 sont mappés sur des sous-buts négatifs de Q_2.

➔ condition suffisante, mais pas nécessaire.

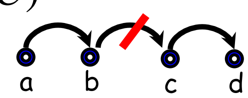
31

Inclusion de requêtes sans mappings d'inclusion

$$Q_2() :- a(X,Y), a(Y,Z), \neg a(X,Z)$$



$$Q_1() :- a(A,B), a(C,D), \neg a(B,C)$$



$$Q_1 \supseteq Q2$$

32

Théorème
[Requêtes avec négation]

Soit B le nombre total de variables et de constantes dans Q_2.

Q_1 inclut Q_2 ssi
$Q_1(D) \supseteq Q_2(D)$ pour tout base de données D ayant au plus B constantes.

33

Exemple (1)

Film(ID,titre,année,genre)
réalisateur(ID,réalisateur)
acteur(ID, acteur)

$$Q(T,Y,D) :- Film(I,T,Y,G), Y \geq 1950, G = "comédie"$$
$$Réalisateur(I,D), Acteur(I,D)$$

$$V_1(T,Y,D) :- Film(I,T,Y,G), Y \geq 1940, G = "comédie"$$
$$Réalisateur(I,D), Acteur(I,D)$$

$$V_1 \supseteq Q \qquad \Rightarrow \qquad Q'(T,Y,D) :- V_1(T,Y,D), Y \geq 1950$$

L'inclusion suffit pour montrer que V_1 peut servir pour répondre à Q.

34

Exemple (2)

$$Q(T,Y,D):-Film(I,T,Y,G),Y\geq 1950,G="comédie"$$
$$Réalisateur(I,D),Acteur(I,D)$$

$$V_2(I,T,Y):-Film(I,T,Y,G),Y\geq 1950,G="comédie"$$
$$V_3(I,D):-Réalisateur(I,D),Acteur(ID,D)$$

Pas d'inclusion mais intutivement, V_2 et V_3 servent pour répondre à Q.

$$Q''(T,Y,D):-V_2(I,T,Y),V_3(I,D)$$

Comment exprimer cette intuition ?

Récriture de requêtes avec des vues ! 35

Récriture : formalisation

Input: Requête Q

Des vues: $V_1,\ldots,V_n$

Récriture = une requête Q' composée des vues et des prédicats interprétés

Récriture équivalente de Q avec $V_1,\ldots,V_n$= récriture Q', tel que $Q' \Leftrightarrow Q$.

36

Exemple (3)

Film(ID,titre,année,genre)
réalisateur(ID,réalisateur)
acteur(ID, acteur)

$$Q(T,Y,D) :- Film(I,T,Y,G), Y \geq 1950, G = "comédie"$$
$$Réalisateur(I,D), Acteur(I,D)$$

$$V_4(I,T,Y) :- Film(I,T,Y,G), \underline{Y \geq 1960}, G = "comédie"$$
$$V_3(I,D) :- Réalisateur(I,D), Acteur(I,D)$$
$$Q'''(T,Y,D) :- \underline{V_4(I,T,Y)}, V_3(I,D)$$

Récriture maximalement incluse

37

Récriture maximalement incluse

Input: Requête Q
L langage de (récriture de) requête
Vues $V_1,\ldots,V_n$

Q' = récriture maximalement incluse de Q si:

1. *Q' $\in$ **L**,*
2. *Q' $\subseteq$ Q, and*
3. *Il n'existe pas de Q'' dans **L** tel que Q'' $\subseteq$ Q et Q' $\subset$ Q''*

38

Justification (pratique)

- Integration LAV (Local-as-View)
  - Besoin de récritures maximalement incluses
- Optimisation de requêtes
  - Besoin de récritures maximalement incluses
  - Implanté dans la plupart des SGBD industriels
- Conception physique de la base de données
  - Description des structures de stockage comme des vues

Exercice : quelles vues peuvent servir à la récriture de Q ?

$$Q(T,Y,D):-Film(I,T,Y,G),Y \geq 1950,G = "comédie"$$
$$Directeur(I,D), Acteur(I,D)$$

$$V_2(I,T,Y):-Film(I,T,Y,G),Y \geq 1950,G = "comédie"$$

$$V_3(I,D):-Directeur(I,D), Acteur(I,D)$$

$$V_6(T,Y):-Film(I,T,Y,G),Y \geq 1950,G = "comédie"$$

$$V_7(I,T,Y):-Film(I,T,Y,G),Y \geq 1950,$$
$$G = "comédie", Prix(I,W)$$

$$V_8(I,T):-Film(I,T,Y,G),Y \geq 1940,G = "comédie"$$