info710 : compléments de bases de données

Préservation des dépendances, troisième forme normale

Pierre Hyvernat
Laboratoire de mathématiques de l'université de Savoie
bâtiment Chablais, bureau 22
téléphone : 04 79 75 94 22
email : Pierre.Hyvernat@univ-savoie.fr

Ce TD est à faire à la maison et à rendre pour le lundi 4 décembre 2006. Il sera probablement noté sur 20 et comptera (avec un coefficient à déterminer) pour la note finale du cours. En plus du fond, la forme est partie intégrante du rapport : le correcteur (c.à.d. moi) doit prendre son pied en lisant votre document final !
Vous pouvez me rendre ce TD au format électronique (format pdf) par email ou en me le donnant en main propre. Vous pouvez également me le poser dans mon casier (batiment le Chablais, deuxième étage) en m'envoyant un email pour me prévenir.

Attention : *les rapports électroniques dans un autre format que pdf ne seront pas lus ! Utilisez Openoffice et sauvegardez au bon format. (Encore mieux, utilisez un membre de la famille de TEX, si vous connaissez...)*

Remarque : *les retards sont toléré, mais le paramètre* `exigeance` *du correcteur est une fonction exponentielle du retard. (Disons que jusqu'au jeudi 7 décembre, c'est OK...)*
Vous pouvez (mais ce n'est pas obligatoire) vous mettre par groupes de deux. Dans ce cas, vous ne rendez qu'un seul rapport qui précisera bien les deux noms.

On rappelle quelques définitions du cours :

Def. *Si $T(A_1,\dots,A_n)$ est un schéma de relation, et si F est un ensemble de dépendances fonctionnelles sur les attributs $A_1,\dots,A_n$, on dit que $\rho=(R_1,\dots,R_k)$ est une décomposition sans perte d'informations de T si l'égalité suivante est vraie, pour toutes les instances de T qui satisfont F :*

$$T \quad = \quad \bowtie_{i=1}^{k} \prod_{R_i} T$$

rq : $R_i \subseteq \{A_1,\dots,A_n\}$ pour $i=1,\dots,k$, et $\bigcup_{1\leq i\leq k} R_i = \{A_1,\dots,A_n\}$,

Def. *Un schéma de relation $T(A_1,\dots,A_n)$ muni d'un ensemble de dépendances F est en forme normale de Boyce-Codd si :*

$$X \to A_i \quad \in \quad F^+ \qquad \Rightarrow \qquad X \text{ est une clé} \quad ou \quad A_i \in X$$

Autrement dit, il n'y a pas de dépendances non-triviales...

Théorème. *Tout schéma de relation $T(A_1,\dots,A_n)$ avec un ensemble de dépendances F admet une décomposition sans perte d'information $\rho=(R_1,\dots,R_k)$ où chacune des tables $\prod_{R_i} T$ est en forme normale de Boyce-Codd par rapport à $F_i = \{X\to B \ \in F^+ \mid X\cup\{B\}\subseteq R_i\}$.*

Exercice 1 : mise en bouche

On considère une relation $L(T,A,G,R)$ d'une table dans une librairie ; cette relation contient
- le Titre d'un livre
- l'Auteur d'un livre
- le Genre d'un livre (SF, policier, cuisine, ...)

- le *R*ayon où est rangé le livre en question

On suppose qu'un auteur a pu écrire des livres dans plusieurs genres, et que dans un même genre, les rayons sont divisés en suivant l'ordre alphabétique. Tout ceci fait que l'on peut imposer les dépendances suivantes :

$$F \quad = \quad \{AG \to R \ , \ R \to G\}$$

Question 1 : quelles sont toutes les clés minimales possibles pour cette relation ?

Question 2 : est-ce que cette table est en forme normale de Boyce-Codd ?

Question 3 : est-ce que $\rho = (TA, AGR)$ est une décomposition sans perte d'information de L muni de F ? Prouvez-le ou donnez un contre-exemple en donnant une instance de la relation L.

Question 4 : est-ce que $\rho = (TAR, AG)$ est une décomposition sans perte d'information de L muni de F ? Prouvez-le ou donnez un contre-exemple en donnant une instance de la relation L.

Question 5 : en utilisant l'algorithme vu en cours, décomposez la relation L en relations qui sont toutes en forme normale de Boyce-Codd.

Question 6 : lors de l'application de l'algorithme dans la question précedente, il y a certain choix à faire (quelle dépendance choisir). Réappliquez l'algorithme en faisant des choix différents. Obtenez-vous le même résultat ? Si non, quel résultat trouvez-vous le plus intuitif ?

Exercice 2 : préservation des dépendance

Une décomposition sans perte d'information permet de conserver toute l'information d'une grosse relation dans plusieures petites relations. Ceci permet de séparer les attributs et permet d'eviter certains problèmes de redondance et de mélange de l'information.

Par contre, une décomposition sans perte d'information ne permet pas de retrouver les dépendances originales de notre relation. Si l'on voit les dépendances fonctionnelles comme une contrainte qui permet de vérifier l'intégrité de la relation, la perte de contrainte peut être facheuse. Ceci justifie la définition suivante :

Def. *Si $\rho = (R_1, \ldots, R_n)$ est une décomposition de $T(A_1, \ldots, A_n)$ avec les dépendances F, on dit que ρ préserve les dépendances si F et $\bigcup_{1 \le i \le k} F_i$ sont équivalentes. (Où, comme précédemment, $F_i = \{X \to A \in F^+ \mid X \cup \{A\} \subseteq R_i\}$.)*

Rappel : deux ensembles de dépendances F et G sont équivalents si $F^+ = G^+$.
Remarque : on ne demande pas que ρ soit une décomposition sans perte d'information, mais juste que $R_i \subseteq \{A_1, \ldots, A_n\}$ pour tous les $i = 1, \ldots, k$ et $\bigcup_{1 \le i \le k} R_i = \{A_1, \ldots, A_n\}$.

Question 1 : est-ce que la décomposition de l'exercice précédent préserve les dépendances ?

Question 2 : en raisonnant de manière informelle, essayer de montrer que la dépendance $AG \to R$ ne sera jamais conservée par une décomposition en forme normale de Boyce-Codd. En déduire que le théorème est faux si l'on demande que les décompositions preservent les dépendances.

Exercice 3 : troisième forme normale

L'exercice précédent montre que, si la préservation des dépendances est importante, alors la forme normale de Boyce-Codd n'est pas très appropriée. On va la remplacer par une notion moins stricte qui permettra d'avoir un théorème de décomposition.

Def. *Étant donné une relation $T(A_1, \ldots, A_n)$ et un ensemble de dépendances F, un attribut B est dit premier s'il existe une clé minimale qui contient B. Si aucune clé minimale ne contient B, alors B est non-premier.*

Def. *Une relation $T(A_1, \ldots, A_n)$ avec un ensemble de dépendances F est en troisième forme normale si :*

$$X \to \{B\} \quad \in \quad F^+ \qquad \Rightarrow \qquad X \text{ est une clé, ou } B \in X, \text{ ou } B \text{ est premier}$$

Question 1 : quels sont les attributs premiers dans la relation $L(T, A, G, R)$ de l'exercice 1 ? Cette relation est-elle en troisième forme normale ?

Une autre relation pour la librairie est la relation des *F*ournisseurs suivante : $F(E, A, I, P)$:
- l'*E*diteur
- l'*A*dresse
- le numéro *I*SBN du livre ("International Standrad Book Number" : code unique de désignation d'un livre publié)
- le *P*rix du livre.

Les dépendances imposées sont

$$G \quad = \quad \{EI \to P \ , \ E \to A\}$$

Question 2 : Quelles sont les clés minimales de cette relation ? Quels sont les attributs premiers ? Cette table est-elle en troisième forme normale ?

L'algorithme pour décomposer une relation en relation en troisième forme normale est plus simple que celui pour décomposer en relation en forme normale de Boyce-Codd :
- **entrée :** un schéma de relation $T(A_1, \ldots, A_n)$ et un recouvrement minimal F des dépendances satisfaites par T
- **sortie :** une décomposition de T
- **algorithme :**
  `supprimer les attributs qui n'apparaissent dans aucune dépendance ;`
  si `une des dépendances de` F `utilise tous les attributs de la table`
  alors `renvoyer` $\rho = (A_1 \ldots A_n)$; (décomposition en un seul morceau)
  sinon `renvoyer` $(X_1 \cup \{B_1\}, \ldots, X_l \cup \{B_l\})$
  (où $X_j \to \{B_j\}$ sont toutes les dépendances de F)

La difficulté de ce calcul réside donc entièrement dans le calcul d'un recouvrement minimal.

Remarque : on peut également (et c'est souvent préferrable) créer une seule partie de la décomposition pour chaque ensemble d'attribut apparaissant à droite d'une dépendance : si $X \to C_1, \ldots, X \to C_p \ \in \ F$, alors on ne crée que la partie $X \cup \{C_1, \ldots, C_m\}$. (Au lieu des parties $X \cup \{C_1\}$, ..., $X \cup \{C_m\}$.)

Théorème. *Le résultat de l'algorithme précédent donne une décomposition en relations en troisième forme normale, et cette décomposition préserve les dépendances.*

Question 3 : faites la preuve de ce théorème.

> *indice : il faut démontrer que les dépendances sont préservées et que chaque relation* $\prod_{X_j \cup \{B_j\}}$ *est en troisième forme normale. Pour ce dernier point, raisonnez par contradiction et utilisez le fait que F est un recouvrement minimal.*

On regarde maintenant les décompositions sans perte d'information préservant les dépendances. On utilise l'algorithme suivant :

`utiliser le résultat` ρ `de l'algorithme précédent ;`
`choisir une clé minimale` X`, la rajouter à la décomposition` $\rho \cup X$ `;`

Si on veut obtenir une décomposition minimale, on peut ensuite enlever certaines parties de la décomposition en veillant à ne pas contredire les deux propriétés (décomposition sans perte d'information et préservation des dépendances).

Question 4 : utilisez l'algorithme décrit plus haut pour trouver une décomposition préservant les dépendances pour les relations L et F des exercices précédents.
Transformez ces décompositions en décompositions sans perte d'information préservant les dépendances en utilisant le deuxième algorithme.
Est-ce que ces décompositions sont optimales ? (Pouvez vous-supprimer certaines parties de la décomposition ?)