Fondements des bases de données

Normalisation (suite)

Équipe pédagogique BD

`https://perso.liris.cnrs.fr/marc.plantevit/doku/doku.php?id=lifbdw2_2020a`

Version du 9 octobre 2020

Les dépendances multivaluées

Comment normaliser une relation

Les dépendances multivaluées

Comment normaliser une relation

- Les dépendances fonctionnelles nous ont permis jusque-là de mettre en évidence une forme de redondance, que les formes normales cherchent à faire disparaître.
- Mais des cas de redondance ne sont pas capturés par les DF.

Exemple

Soit l'énoncé *les étudiants suivent des parcours, et sont inscrits dans des transversales indépendantes du parcours. Chaque étudiant peut être inscrit à plusieurs parcours* et la relation

$$R(\mathit{etudiant}, \mathit{parcours}, \mathit{transversale})$$

On ne peut dégager aucune DF dans ce schéma, donc la seule clé est la combinaison des trois attributs. Mais, il y a bien des redondances !

Dépendance multivaluée

Syntaxe

Une *dépendance multivaluée* (MVD) sur un schéma R est une expression de la forme $X \twoheadrightarrow Y$ avec $X \subseteq R$ et $Y \subseteq R$. On dit que X multidétermine Y.

Sémantique

Une relation r sur R satisfait la MVD $X \twoheadrightarrow Y$, noté $r \models X \twoheadrightarrow Y$, *ssi* pour toute paire de tuples t_1 et t_2 de r tels que $t_1[X] = t_2[X]$, ils existent des tuples t_3 et t_4 avec $t_3[X] = t_4[X] = t_1[X]$ tels que :

- $t_3[Y] = t_1[Y]$ et $t_3[R \setminus Y] = t_2[R \setminus Y]$
- $t_4[Y] = t_2[Y]$ et $t_4[R \setminus Y] = t_1[R \setminus Y]$

Informellement, pour chaque valeur de X fixée, les valeurs de Y et $Z = R \setminus XY$ sont indépendantes : on a toutes les combinaisons possibles de Y et Z.

Sur la relation *Livre*(*isbn*, *auteur*, *keyword*)

- Si un livre peut avoir plusieurs auteurs et plusieurs mots-clés (keyword), la relation livre possède la dépendance multivaluée

$$isbn \twoheadrightarrow auteur$$

- Autrement dit, à *isbn* fixé, on a toutes les combinaisons possibles d'auteurs et de mots-clés pour l'isbn en question :
  - Si (i, m_1, a_1) et (i, m_2, a_2) sont dans r
  - Alors (i, m_1, a_2) et (i, m_2, a_1) sont aussi dans r

	isbn	**keyword**	**auteur**
t_1	i	m_1	a_1
t_2	i	m_2	a_2
t_3	i	m_1	a_2
t_4	i	m_2	a_1

Dépendance multivaluée triviale

Une dépendance multivaluée $X \twoheadrightarrow Y$ d'une relation R est dite triviale si :

- soit $Y \subseteq X$: comme pour les DFs,
- soit $X \cup Y = R$: la MVD n'impose l'existence d'aucun tuple.

Propriétés des dépendances multivaluées

- les MVDs triviales sont toujours satisfaites.
- la définition de la satisfaction étant symétrique,
  si $r \models X \twoheadrightarrow Y$ alors $r \models X \twoheadrightarrow R \setminus XY$ également.
- si $r \models X \rightarrow Y$ alors $r \models X \twoheadrightarrow Y$ également
  toutes les DFs sont des MVD, mais la réciproque est fausse.

Axiomatisation des dépendances multivaluées

- Réflexivité

$$\frac{Y \subseteq X}{X \twoheadrightarrow Y}$$

- Augmentation

$$\frac{X \twoheadrightarrow Y}{WX \twoheadrightarrow WY}$$

- Transitivité

$$\frac{X \twoheadrightarrow Y \qquad Y \twoheadrightarrow Z}{X \twoheadrightarrow Z}$$

- Complémentation

$$\frac{X \twoheadrightarrow Y}{X \twoheadrightarrow R \setminus XY}$$

- Union

$$\frac{X \twoheadrightarrow Y \qquad X \twoheadrightarrow Z}{X \twoheadrightarrow YZ}$$

- Décomposition

$$\frac{X \twoheadrightarrow Y \qquad X \twoheadrightarrow Z}{X \twoheadrightarrow Y \cap Z, X \twoheadrightarrow Y \setminus Z, X \twoheadrightarrow Z \setminus Y}$$

- Sous-ensemble

$$\frac{X \twoheadrightarrow Y \qquad W \twoheadrightarrow Z \qquad Y \cap W = \emptyset}{X \twoheadrightarrow Y \cap Z, X \twoheadrightarrow Y \setminus Z}$$

Réflexivité, complémentation, augmentation et transitivité forment un système correct et complet pour l'inférence des MVD

Axiomatisation MVD et DF ensembles

- Généralisation

$$\frac{X \to Y}{X \twoheadrightarrow Y}$$

- Pseudo-transitivité mixée

$$\frac{X \twoheadrightarrow Y, Z \subseteq Y \qquad W \cap Y = \emptyset, W \to Z}{X \to Z}$$

Ajoutés à réflexivité, complémentation, augmentation et transitivité, ces règles forment un système correct et complet pour l'inférence des MVD et de DFs prises ensembles.

Quatrième forme normale (4FN)

Quatrième forme normale (4FN)

Une relation R est en 4FN, ssi pour chaque dépendance multivaluée $X \twoheadrightarrow Y$ non triviale, X est une super-clé de R.

Définition *alternative*

R est en 4FN ssi

- R est en 3FN
- Les seules dépendances multivaluées sont du type $X \twoheadrightarrow R \setminus X$.

Remarque

- R ne doit pas être décomposable en deux relations sans perte de jointure.
- La 4FN implique la FNBC puisqu'une dépendance fonctionnelle est un cas particulier de dépendance multivaluée.

Exemple

$$\{isbn \twoheadrightarrow auteur;\ isbn \twoheadrightarrow mot\_cle\}$$

isbn	*auteur*	*mot_cle*
2-7117-8645-5	Abiteboul	BD
2-7117-8645-5	Hull	BD
2-7117-8645-5	Vianu	BD
2-7117-8645-5	Abiteboul	Relationnel
2-7117-8645-5	Hull	Relationnel
2-7117-8645-5	Vianu	Relationnel

Pas en 4FN

isbn	*auteur*
2-7117-8645-5	Abiteboul
2-7117-8645-5	Hull
2-7117-8645-5	Vianu

isbn	*mot_cle*
2-7117-8645-5	BD
2-7117-8645-5	Relationnel

Cette décomposition est sans perte et préserve les dépendances.

Comment normaliser une relation

Algorithme de normalisation : principe

Entrée : L'ensemble U de tous les attributs du problème

Entrée : Un ensemble de contraintes sur U (DF, DI, MVD).

Sortie : Un schéma de BD normalisé (3FN, BCNF ou 4FN).

Deux grandes catégories d'algorithmes de normalisation

- les algorithmes *de décomposition*,
- les algorithmes *de synthèse*.

Dans les deux cas on va s'appuyer sur un ensemble canonique de dépendances et utiliser le théorème de décomposition.

Décomposition SPI

$R(XYZ)$ est décomposable Sans Perte d'Information (SPI) sur $R_1 = XY$ et $R_2 = XZ$ ssi

$$R = \pi_{XY}(R) \bowtie \pi_{XZ}(R)$$

Théorème de décomposition (Heath)

Soit $R(XYZ)$ (X, Y et Z disjoints) une relation. Si R vérifie $X \rightarrow Y$ alors la décomposition en $R_1(XY)$ et $R_2(XZ)$ est SPI.

Calcul d'une couverture canonique

Pour décomposer selon F, on va utiliser un ensemble F' qui soit :

- Couverture de F : $F^+ = F'^+$,
- Minimal : on ne peut pas retirer de DF en préservant toujours la couverture,
- Sans attributs redondants, ni à droite ni à gauche,
- Regroupé : il n'y a pas deux DF avec la même partie gauche.

On a vu des algorithmes qui permettent de produire une telle couverture. Ces étapes sont nécessaires pour assurer que les algorithmes vont bien produire un bon schéma !

Algorithme de décomposition

- R la relation (universelle) à décomposer
- F un *ensemble minimal* de dépendances sur R
- $S = \{R\}$ le schéma de base de données

Algorithme de décomposition en 4FN/FNBC

Tant qu'il existe dans $R \in S$ qui n'est pas en 4FN/FNBC :

- On cherche dans D une dépendance non triviale $X \twoheadrightarrow Y$ [1] telle que $R(X, Y, Z)$ et X n'est pas une clé de R.
- On ajoute à Y l'ensemble Z' des attributs de Z fonctionnellement déterminés par X, produisant la dépendance $X \twoheadrightarrow YZ'$.
- On remplace R dans S par les deux relations
  - $R_1(X, Y \cup Z')$
  - $R_2(X, Z \setminus Z')$

Propriété : cet algorithme est sans perte d'information mais pas toujours sans perte de dépendances

1. une DMV est un cas particulier de DF

Algorithme de synthèse

- R la relation (universelle) à décomposer
- F un *ensemble* de dépendances sur R

Principe général

- Construire une couverture canonique de F.
- Générer une relation XY pour chaque DF $X \to Y$;
- Générer une relation XY' pour chaque DMV $X \twoheadrightarrow Y$ avec $F \models Y' \to Y$;
- On supprime les schémas de relation qui ne sont pas maximaux par inclusion.
- S'il y a perte de jointure, alors on rajoute une relation composée d'une clé de F.

Propriété : l'algorithme est sans perte d'information et donne un schéma en 3FN ou en FNBC quand c'est possible sans perte de dépendance.

Exemple

On gère une liste de **projets**. *Chaque* **projet** *a un* **responsable**, *un* **ensemble d'employés**, *et utilise certains* **produits** *en une* **quantité** *donnée. Pour un* **produit** *et un* **projet**, *plusieurs* **fournisseurs** *à des* **cout** *différents sont concernés. Un* **fournisseur** *a plusieurs* **adresses**. *Finalement, un* **employé** *a une* **date d'embauche** *et un* **salaire**.

- Les DFs sont :
  - $projet, produit \rightarrow quantite$
  - $projet, fournisseur, produit \rightarrow cout$
  - $projet \rightarrow responsable$
  - $employe \rightarrow salaire, dateE$
- Les MVDs sont :
  - $fournisseur \twoheadrightarrow adresse$
  - $projet \twoheadrightarrow employe, salaire, dateE$

Bilan

	3FN	FNBC	4FN
Elimination de la redondance due aux DF	Pas toutes	Oui	Oui
Elimination de la redondance due aux DMV	Non	Non	Oui
Préservation des DF	Oui	Pas toujours	Pas toujours
Préservation des DMV	Pas toujours	Pas toujours	Pas toujours

Fin