Livraison 4 – normalisation II

Pascal Ostermann – pasccal@orange.fr

26 avril 2020

Au cours précédent, j'ai arrêté la normalisation du laboratoire de recherche sur la relation

R(num_emp, titre_art, date_art, nom_revue, mot_clef)

Les seules DFs vraies dans R sont triviales : par exemple, un article a plusieurs auteurs ; il lui est associé plusieurs mots-clefs, etc... R est donc en FNBC. Elle n'est pourtant pas satisfaisante, par exemple parce qu'un mot-clef associé à un article est répété autant de fois que cet article a de co-auteurs. En fait on désirerait la décomposer en les deux relations suivantes

Co-auteur(num_emp, titre_art, date_art, nom_revue)

Mot_clef(titre_art, date_art, nom_revue, mot_clef)

Mais pour justifier cette dernière décomposition, il me faut introduire un nouveau type de dépendances, avec un théorème de décomposition et une forme normale.

Dépendances multivaluées

> Soit X, Y et Z trois ensembles d'attributs d'une relation R, disjoints deux-à-deux. La dépendance multivaluée (DMV) X ↠ Y | Z est vraie dans R si et seulement si

$$\begin{aligned}\forall s,t \in R\,(s.X = t.X \supset \\ \exists u \in R\,(u.X = s.X \wedge \\ u.Y = s.Y \wedge \\ u.Z = t.Z))\end{aligned}$$

Propriétés

- si X et Y sont disjoints, alors X↠Y|∅ et X↠ ∅|Y (DMVs triviales)
- si X↠Y|Z alors X↠Z|Y
- si X, Y et Z sont disjoints, et si X→Y, alors X↠Y|Z
- Certains notent X↠Y (DMV globale) pour X↠Y|$\complement_R$XY
  mais je n'aime pas cette notation à cause du R implicite : X↠Y peut être vraie dans une relation R, et fausse dans une relation plus grande...

Exemple

Dans la relation R ci-dessus, la DMV suivante est vraie.

$$titre\_art, date\_art, nom\_revue \twoheadrightarrow num\_emp | mot\_clef$$

En effet, on a vu au cours précédent qu'un article était identifié par son titre, sa date et sa revue de publication. Donc si nous avons deux n-uplets concernant le même article, genre (je choisis un exemple plus parlant qu'un article scientifique)

Fantômas, Souvestre, crime
Fantômas, Allain, Fandor

Nous avons nécessairement également le n-uplet

Fantômas, Souvestre, Fandor

Puisque Souvestre est un des co-auteurs de Fantômas, il est également un des inventeurs du journaliste Jérôme Fandor, quand bien même il n'aurait jamais écrit une ligne sur ce personnage. On peut également dire que les mots_clefs associés à un article – qui correspondent grosso modo à son contenu – sont indépendants de ses auteurs : il ne change pas de contenu selon qu'on le considère comme écrit par untel ou par tel autre.

Théorème de décomposition

Soit X, Y, Z une *partition* de l'ensemble des attributs de R,

R se décompose SPI en $\prod_{XY}(R)$ et $\prod_{XZ}(R)$
si et seulement la DMV $X \twoheadrightarrow Y|Z$ est vraie dans R.

J'insiste sur le fait que X, Y, Z doivent former une partition : disjoints deux à deux comme dans toute DMV, et ils recouvrent R.

Quatrième forme normale (4FN)

Une relation R sera dite en quatrième forme normale si et seulement si toute DMV $X \twoheadrightarrow Y|Z$ (où X, Y, Z partition de R) est telle que

— Y ou Z est vide (DMV triviale) ou
— X est superclef de la relation R

Propriétés

— Le théorème de décomposition pour les DFs est un cas particulier de celui pour les DMVs.
— La 4FN est plus forte que la FNBC (elle-même plus forte que la 3FN, etc...) ; et si une relation n'admet pour DMVs que celles qu'on peut déduire des DFs, les deux formes normales sont équivalentes.
— Toute relation peut se décomposer SPI en des relations en 4FN.

Dépendances de jointure et cinquième forme normale

À lire le théorème de décomposition pour les DMVs et son « si et seulement », on pourrait penser que la quatrième forme normale est aussi la forme terminale. Ce n'est pas tout à fait vrai. Soit par exemple la relation suivante, tirée de Ullman [1] :

$$\text{Buveurs}(\underline{\text{buveur, bar, bière}})$$

On considère que cette relation signifie qu'un buveur fréquente un bar, que ce bar sert une bière, et que cette bière est aimée par le buveur. En d'autres termes, elle est par construction équivalente à

$$\text{Fréquente}(\underline{\text{buveur, bar}}) \bowtie \text{Sert}(\underline{\text{bar, bière}}) \bowtie \text{Aime}(\underline{\text{buveur, bière}})$$

Il n'y a pas de DMV dans la relation Buveurs, sans quoi elle serait décomposable en seulement deux relations. On dit qu'il y a ici une **dépendance de jointure** (DJ). Je n'en donne pas de formalisme, chacun des ouvrages que j'ai consultés en donnant une version différente, mais on peut aisément étendre les définitions liées aux DMVs par une cinquième forme normale : les DJs y seraient triviales ou déductibles des clefs. Et énoncer un théorème de décomposition. Mais c'est là que le bât blesse, le dit théorème signifiant qu'on peut décomposer ssi il y a DJ, et la DJ étant définie comme la possibilité de décomposer : on se retrouve à dire « on peut décomposer si et seulement si on peut décomposer. »

En fait, on peut éviter de faire apparaître des DJs avec un petit peu d'attention lors de l'écriture de la liste des attributs. Lorsqu'on définit le troisième attribut (disons que c'est la bière), on découvre qu'il a deux significations, et on décide de le dupliquer, en obtenant cette liste d'attributs :

$$\text{buveur, bar, bière\_servie, bière\_aimée}$$

après quoi le résultat désiré s'obtient via les seules DMVs

1. buveur $\twoheadrightarrow$ bar, bière_servie | bière_aimée
2. bar $\twoheadrightarrow$ buveur | bière_servie

DMVs à partie gauche vide

Une autre maladresse que l'on peut commettre en établissant la liste des attributs est d'ajouter un ensemble d'attributs X qui n'ont aucun lien avec les autres Y. Cela peut notamment se produire lorsqu'on duplique des attributs comme ci-dessus, mais inutilement. Dans ce cas la relation globale pourra quand même se décomposer via le produit cartésien :

$$R = \prod_X R \times \prod_Y R$$

En d'autres termes, on a ici la DMV $\emptyset \twoheadrightarrow X|Y$.

1. Principles of database systems, à mon avis l'ouvrage de référence sur le sujet.

Conclusion : ce qu'il faut retenir

La méthode de normalisation consiste donc à

— établir une liste d'attributs – Cette liste doit être plate, sans structuration, afin de donner à chaque attribut un nom sans ambiguïté ; ces attributs doivent vraiment en être : des VALEURS ATOMIQUES, dont vous devez pouvoir signifier le domaine ; par ailleurs, il est plus prudent d'éviter les attributs qui n'ont pas de lien avec les précédents (afin d'éviter les DMVs de partie gauche vide), ou ceux qui auraient plusieurs liens avec ceux-là (afin d'éviter les DJs).

— établir la liste des dépendances – Elles devront être les plus fortes et les plus concises qu'il est possible : par exemple pas de $XY \to Z$ si $X \to Z$ est vrai ; pas non plus de $X \to Z$ si $X \to Y$ et $Y \to Z$ sont vraies. [2]

— Décomposer la relation globale constituée de la liste des attributs à l'aide des dépendances, pour obtenir un schéma normalisé – Les seules formes normales à retenir devraient être les troisième, Boyce-Codd et quatrième.

Exemple

Pour finir, le traitement complet de l'exemple du centre de recherche

Liste des attributs

nom_bât, num_salle, nom_labo, num_tél, num_emp, nom_emp, prénom_emp, tél_perso, titre_art, date_art, nom_revue, texte_art, mot_clef

Dépendances

1. nom_bât, num_salle → nom_labo, num_tél
2. nom_labo → nom_bât
3. num_emp → nom_emp, prénom_emp, tél_perso, nom_bât, num_salle
4. titre_art, date_art, nom_revue → texte_art
5. titre_art, date_art, nom_revue ↠ num_emp | mot_clef

Résultat

Salle(<u>nom_bât, num_salle</u>, nom_labo, num_tél) – en 3FN
Emp(<u>num</u>, nom, prénom, tél_perso, bât_bur, num_bur)
Article(<u>titre_art, date_art, nom_revue</u>, texte_art)
Co-auteur(<u>num_emp, titre_art, date_art, nom_revue</u>)
Mot_clef(<u>titre_art, date_art, nom_revue, mot_clef</u>)
Sauf Salle, toutes les relations sont en 4FN.

2. Dans des cours plus théoriques et peut-être plus rigoureux – en tous cas plus longs –, on parle ici de la *couverture minimale* d'un ensemble de DFs, en on propose des méthodes pour la calculer, mais... aux dernières nouvelles, cela ne marchait pas vraiment pour les DMVs ; et la couverture minimale est généralement celle qui vient le plus naturellement à l'esprit.