Utilisation de la théorie des sondages dans le cadre des OLAP

Sabine Goutier, Véronique Stéphan

Electricité de France Recherche et Développement
Département ICAME
1, av. du Général de Gaulle, 92 141 Clamart Cedex, France
{sabine.goutier, veronique.stephan}@edf.fr

Résumé. Dans le cadre de la théorie des sondages, le traitement de la non-réponse a donné lieu à différentes méthodologies reposant principalement sur la pondération ou sur l'imputation. L'objet de cet article est de montrer comment ce cadre formel statistique s'adapte naturellement au problème des valeurs manquantes dans le contexte des OLAP. Dans cette étude, nous nous limitons au cas des valeurs manquantes dans les dimensions, appelées alors dimensions creuses. La méthode d'ajustement est réalisée en intégrant un système de poids au sein du cube. La complexité algorithmique est fortement diminuée par la recherche d'un ensemble de systèmes de pondération minimum. Celle-ci, appelée méthode ROWN, est synthétisée et une validation expérimentale de l'évolution des estimations en fonction du support est présentée. Enfin, l'implémentation sous ORACLE EXPRESS est détaillée.
Mots-Clés : OLAP, cube de données, valeurs manquantes, redressement.

1. Introduction

Les technologies OLAP (OnLine Analytical Processing) permettent la consultation de grands volumes de données selon des directions multidimensionnelles (Codd 1993). Comme de nombreuses entreprises, Electricité De France (EDF) met en œuvre des OLAP pour la consultation de ses grandes bases de données. Au moyen des techniques OLAP, les marketeurs peuvent par exemple analyser les différents segments et tendances du marché, et répondre plus efficacement aux besoins opérationnels. Les données de cubes OLAP servent ainsi de support aux utilisateurs pour leur prise de décision. Il est donc indispensable de garantir la qualité des résultats fournis. En particulier, dans le cas de valeurs manquantes, un biais peut survenir par la seule prise en compte des valeurs renseignées pour le calcul des agrégats du cube.

S'agissant des données clients, l'existence dans les bases de valeurs manquantes est fréquente. Dans le contexte des bases EDF, les informations sont pour la plupart saisies lors de contacts avec le client. Hormis les informations tarifaires obligatoirement renseignées, les autres caractéristiques liées à un client (par exemple l'énergie du chauffage principal ou le type d'habitat) ne sont pas forcément (bien) remplies.

Le but de cet article est d'adapter les techniques de sondage pour améliorer la qualité des agrégats en présence de valeurs manquantes observées sur les dimensions du cube. Après avoir décrit le contexte de travail, nous présentons la repondération d'une dimension par

post-stratification. Puis, nous l'étendons au cas de plusieurs dimensions et nous testons l'influence du support sur les estimations. Nous présentons synthétiquement l'algorithme ROWN de réduction des poids (présenté dans Goutier et al. 2002). Dans cet article, nous détaillons son implémentation dans un système multidimensionnel : Oracle Express.

2. Contexte

Dans le modèle multidimensionnel, la structure logique de base est le cube qui contient un ensemble de valeurs numériques, définissant des mesures, observées dans un espace multidimensionnel constitué de plusieurs dimensions, appelées aussi axes d'étude. Parmi les multiples propositions de modélisation (Vassiliadis et al. 1999), nous reprenons la définition de (Gray et al. 1996). Pratiquement, un cube est construit à partir d'une table qui contient toutes les données détaillées sur lesquelles les variables peuvent être agrégées. Nous supposons la structure de table suivante :

$$\text{DETAIL} (\underline{\text{ID}}, D_1, D_2, \ldots, D_p, M_1, M_2, \ldots, M_q, W)$$

où:

- $D_1, D_2, \ldots, D_p$: attributs dimension qui forment les dimensions du cube,
- $M_1, M_2, \ldots, M_q$: attributs mesure qui forment les mesures du cube,
- *ID*: identifiant des enregistrements, aussi appelés individus, de la table DETAIL.
- W : poids de chaque individu (habituellement 1) qui sert pour pondérer les mesures. Cet attribut n'est habituellement pas considéré dans un cube.

Appliquée aux données clients, cette représentation s'écrit:

CLIENT(<u>IDCLI</u>, CONTRAT, HABITAT, CHAUFFAGE, FACTURE, POIDS)

Un exemple d'un cube CLIENT décrit pour les dimensions type de contrat, type d'habitat et énergie de chauffage, le nombre de clients[1] et la somme des factures.

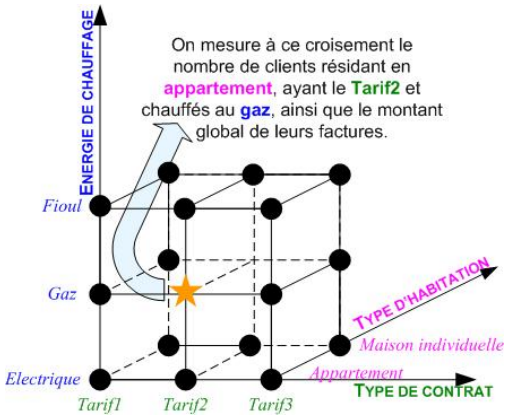


FIG. 1 – *Schéma d'un cube*

[1] obtenu par sommage de la variable POIDS égale dans ce cas à la valeur 1 sur tous les individus.

Nous traitons dans l'article, des valeurs manquantes dans les données de détail d'un cube. Cette problématique s'inscrit dans le cadre plus général de la qualité des données, qui a donné lieu à différentes approches selon les domaines d'étude.

La qualité d'un cube est intrinsèquement liée à la qualité de l'entrepôt de données. Le cas des valeurs manquantes s'y intègre naturellement dans le problème de complétude (Rahm et al. 2000). Le traitement des valeurs manquantes est décrit plus précisément dans le cadre du « nettoyage des données », qui regroupe un ensemble de techniques pour transformer et nettoyer les données lors de la construction de l'entrepôt de données (Galhardas et al. 2001). Notre approche diffère de celles proposées dans la mesure où nous ne nettoyons pas les données, mais où nous intégrons un système de poids pour tenir compte des valeurs manquantes.

En traitement d'enquêtes, la qualité des données intègre à la fois le traitement des valeurs erronées (Granquist 1997) et celle des valeurs manquantes (Little et al. 2002). Les deux approches majeures pour le traitement des valeurs manquantes, dites de non-réponse, sont celles de la repondération et de l'imputation (Tillé 2001). Dans l'article, nous utilisons des variables auxiliaires pour réaliser une repondération simple sans recours à des méthodes plus complexes (Deville 2002).

Notre travail s'inscrit également dans les problématiques d'imprécision dans les bases de données multidimensionnelles. En présence de valeurs manquantes, une première approche (Dyreson 1996) propose d'évaluer une requête alternative avec l'opérateur rollup. Dans (Pedersen et al. 2001), les auteurs adoptent une démarche générale pour la gestion des données imparfaites, c'est-à-dire manquantes ou imprécises par leur granularité. Ils proposent un traitement de l'imprécision sur les dimensions (resp. mesures) par repondération (resp. imputation). Contrairement à cette approche qui suppose la connaissance d'un modèle d'appartenance floue sur les données, notre travail intègre une étape de post-stratification sans recours à une connaissance experte.

Enfin, une approche (Laurent 2002) propose la détection dans un OLAP de cellules vides, qui peuvent être dues à des valeurs manquantes dans les données sources. Il s'agit d'identifier les cellules vides potentiellement intéressantes en analysant les cellules voisines du cube. Contrairement à cette approche, notre travail se situe en amont de la construction du cube. De plus, nous supposons que la population totale est connue avec éventuellement des valeurs manquantes sur les dimensions alors que l'hypothèse de l'auteur apparaît davantage comme celle de l'observation d'un échantillon (individus manquants).

3. Introduction de la repondération dans un cube OLAP

Le mode de traitement des valeurs manquantes dépend des hypothèses fixées sur les données. Dans le cas le plus simple, la probabilité d'observer une valeur manquante est supposée ne dépendre d'aucune variable : on parle alors de mécanisme de réponse uniforme. Cette hypothèse est cependant rarement réaliste. Pour cette raison, nous proposons d'intégrer une repondération pour le traitement des valeurs manquantes. La repondération vise à réduire le biais des estimations en supposant un mécanisme de réponse uniforme au sein de chaque strate de la population. Dans l'exemple précédent, cela reviendrait à diviser la population en plusieurs strates (par exemple les clients anciens et les nouveaux) et à considérer que la probabilité d'observer, dans chaque groupe, une valeur manquante est aléatoire.

3.1 Rappel de la repondération dans le contexte poststratifié

On suppose que la population est décrite par des variables renseignées sur tous les clients et des variables pour lesquelles il y a des valeurs manquantes. Parmi les premières, certaines d'entre-elles fournissent des informations sur les caractéristiques de la population. Ces dernières sont alors utilisées comme variables auxiliaires, au moyen d'une post-stratification, pour pondérer les individus renseignés, que l'on appelle par la suite échantillon. La repondération consiste à affecter à chaque unité renseignée sur la variable étudiée, un poids, calculé en fonction de la probabilité de réponse dans sa strate d'appartenance. On parle donc d'hypothèse de réponse uniforme à l'intérieur de chaque strate. En pratique, on cherche des strates ayant des effectifs sur l'échantillon suffisants (généralement, on fixe le seuil à 20). Les paramètres d'intérêt (comptage, total, moyenne, …) sont alors estimés par l'estimateur d'Horvitz-Thomson.

La première phase de stratification s'effectue par le croisement des variables auxiliaires qui détermine la partition de la population en strates. Notons H le nombre total de strates. Notons n_h le nombre d'individus de l'échantillon dans h et N_h le nombre d'individus de la population dans h. Nous notons n la taille de l'échantillon et N la taille de la population.

Soit M, une mesure d'intérêt observée sur la population. L'estimateur d'Horvitz-Thomson pour la moyenne de la mesure M est calculé par la formule suivante :

$$\overline{M} = \frac{1}{N} \times \sum_{i=1}^{n} w_i M_i$$

où w_i est le poids de l'élément i. Etant donnée la post-stratification, les poids des individus sont uniformes au sein de chaque strate.

Ainsi, chaque individu d'une strate h est affecté d'un poids égal à N_h / n_h qui dépend uniquement de la probabilité de réponse dans la strate.

On peut donc écrire :

$$\overline{M} = \frac{1}{N} \times \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left(\frac{N_h}{n_h} \right) \times M_{ih}$$

où M_{ih} est la valeur pour M de i appartenant à la strate h dans l'échantillon.

3.2 Application de la repondération à une dimension creuse

Nous considérons le cas où la table *DETAIL* contient l'ensemble de la population mais où, pour un ensemble de clients, les modalités d'un attribut noté D_1 ne sont pas connues. Cette dimension est une dimension *creuse* par opposition aux dimensions *denses* qui sont renseignées pour tous les clients de la population. L'ensemble des clients renseignés pour D_1 forme un échantillon de la table *DETAIL* que nous notons *DETAIL**. Ainsi *DETAIL* représente la population entière de taille N et *DETAIL** l'échantillon de taille n.

L'information issue de la post-stratification est supposée disponible dans la table *DETAIL* au moyen d'un attribut, noté S, résultant de ce croisement. Cette dimension est nécessairement une dimension dense. Ainsi, S possède H catégories (H étant le nombre de strates) notées 1, 2, … , h, … , H et nous pouvons écrire :

$$N_h = \#\left\{ t \in DETAIL \mid t[S] = h \right\}$$
$$n_h = \#\left\{ t \in DETAIL* \mid t[S] = h \right\}$$

Comme défini au pararaphe précédent, le poids associé à chacun des individus de la table $DETAIL^*$ (système de poids W) se calcule par la formule suivante :

$$\text{pour tout } t \in DETAIL^*, \text{ tel que } t[S] = h,\ t[W] = \frac{N_h}{n_h}$$

Dans ce cas, le cube contient les mesures suivantes :
- les mesures $M_1, M_2, \ldots, M_q$,
- W qui correspond à l'attribut poids précédemment défini,
- les mesures pondérées notées $WM_1, WM_2, \ldots, WM_q$.

Un système de poids est une mesure additive qui permet d'obtenir le comptage des individus par catégorie. Ainsi, pour chaque requête soumise, le comptage est calculé comme une somme du système de poids. Cette **propriété d'additivité** permet la consolidation[2] des données (opération de rollup en terminologie OLAP).

La moyenne d'une mesure s'obtient par la formule suivante:

$$\frac{\sum_{i \in DETAIL^*} W_i \times M_{ij}}{\sum_{i \in DETAIL^*} W_i} = \frac{1}{N} \sum_{i \in DETAIL^*} W_i \times M_{ij}$$

où M_{ij} est la valeur de la mesure M_j pour l'individu i de la table $DETAIL^*$.

Intéressons nous à présent aux requêtes sur le cube. Deux types de requêtes peuvent être isolées: (Q1) requêtes n'effectuant pas de sélection sur la dimension D_1, (Q2) requêtes effectuant une sélection sur D_1.

Pour les requêtes de type (Q1), aucune repondération n'est nécessaire. Ainsi les réponses sont affichées en utilisant les mesures initiales $M_1, M_2, \ldots, M_q$ et le système de poids valant 1 pour tous les individus. Pour les requêtes de type (Q2), il est nécessaire d'utiliser le poids W et les mesures $WM_1, WM_2, \ldots, WM_q$, étant donné que la sélection ne retourne que les individus renseignés sur D_1.

Supposons que nous cherchons les proportions de clients par habitat et énergie de chauffage. Les résultats de cette requête de type (Q2) figurent dans le tableau 1. Chaque ligne de cette table (pour la colonne "pondéré") est obtenue en trois étapes :
- sélectionner dans la table *DETAIL* les tuples vérifiant la sélection (notons qu'ils appartiennent nécessairement à la table $DETAIL^*$),
- sommer l'attribut poids W sur les tuples sélectionnés,
- diviser le résultat par N (nombre de tuples dans la table *DETAIL*).

Ces proportions doivent être comparées à celles qui auraient été obtenues avec un cube standard. Dans ce cas, les valeurs manquantes sont remplacées par une valeur explicite supplémentaire 'MISS'. Si celles-ci sont occultées de l'analyse, l'interprétation des résultats se trouve erronée.

[2] Nous rappelons que la consolidation consiste au stockage d'agrégats précalculés à des niveaux spécifiés afin d'améliorer les temps de réponse dans le cas de très gros cubes.

Habitat	Chauffage	%clients (pondéré)	%clients (non pondéré)
Maison	Electricité	15.3%	17.3%
Maison	Gaz	22.6%	16.3%
Maison	Fioul	20.5%	13.5%
Maison	∅		12.5%
Appartement	Electricité	12.7%	15.7%
Appartement	Gaz	18.4%	12.6%
Appartement	Fioul	10.5%	3.7%
Appartement	∅		8.4%

TAB 1 – *Comparaison des résultats de la requête "Proportions de clients par habitat et énergie de chauffage" avec ou sans pondération*

3.3 Valeurs manquantes dans plusieurs dimensions

Il s'agit de généraliser la méthode au cas de plusieurs dimensions creuses. Considérons donc deux dimensions creuses D_1 et D_2. La table $DETAIL_1^*$ (resp. $DETAIL_2^*$) est la sélection des tuples de la table *DETAIL* qui sont renseignés sur la dimension D_1 (resp. D_2). Toute requête pouvant faire intervenir à la fois D_1 et D_2 (et éventuellement des dimensions denses), il est nécessaire de définir la table $DETAIL_{1,2}^*$ comme l'intersection de $DETAIL_1^*$ et de $DETAIL_2^*$ ($DETAIL_{1,2}^* = DETAIL_1^* \cap DETAIL_2^*$). En suivant l'approche exposée, nous sommes amenés à considérer **trois** systèmes de poids (notés W_1, W_2 et $W_{1,2}$) pour appliquer la méthode. Dans le cas de *m* dimensions creuses, le nombre de systèmes de poids s'élève alors à $2^m - 1$. Combiné au fait que chaque système de poids doit être répliqué pour chaque mesure (W_1M_1, ..., W_1M_p, W_2M_1, ..., W_2M_p, $W_{1,2}M_1$, ..., $W_{1,2}M_p$), cette solution n'est pas satisfaisante.

Si une dimension creuse est trop peu renseignée, elle ne doit pas être intégrée dans un OLAP. Cela implique également que tout croisement de cette dimension avec une autre n'est pas envisageable. Plus formellement, pour chaque dimension creuse *D*, définissons son support comme le rapport entre le nombre d'individus dans $DETAIL^*$ et le nombre d'individus dans *DETAIL* :

$$support(D) = \frac{\#(DETAIL^*)}{\#(DETAIL)}$$

En généralisant au cas de deux dimensions creuses D_1 et D_2, il est possible d'écrire :

$$support(D_1 \times D_2) = \frac{\#(DETAIL_1^* \cap DETAIL_2^*)}{\#(DETAIL)}$$

Une dimension creuse est *admissible* ssi son support est supérieur à une valeur fixée notée s_0. Cette condition n'est pas équivalente à la condition d'un nombre minimum d'individus dans les strates mais néanmoins étroitement liée. Dans nos expérimentations, s_0 est fixé à 10%.

La recherche de toutes les intersections valides peut être coûteuse et engendrer malgré tout un grand nombre de systèmes de poids à stocker dans le cube. La méthode développée (méthode ROWN pour *Reduction Of Weight Number*) permet de s'affranchir de ce problème. Elle est basée sur l'idée selon laquelle le système de poids résultant de l'intersection de deux dimensions creuses peut remplacer l'une ou l'autre (ou les deux). Ainsi $D_1 \times D_2$ peut être utilisé pour D_1 ou pour D_2.

Supposons que les supports de $D_1 \times D_2$, D_1 et D_2 sont dans la configuration indiquée sur la figure 2 (les rectangles représentent les échantillons des clients renseignés sur la dimension considérée). La perte d'informations acceptée pour remplacer D_1 par $D_1 \times D_2$ va être quantifiée à l'aide du niveau de confiance de $D_1 \times D_2$ par rapport à D_1 défini par :

$$conf((D_1 \times D_2)/D_1) = \frac{support((D_1 \times D_2))}{support(D_1)}$$

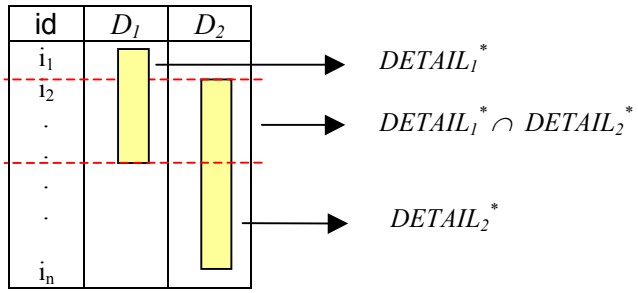


FIG. 2 – *Configuration des supports de D_1, D_2 et $D_1 \times D_2$*

Dans ce cas, D_1 pourra être remplacé par $D_1 \times D_2$ si le niveau de confiance est supérieur au seuil fixé, noté c_0 (qui vaut ici 75%). L'introduction du niveau de confiance permet de réaliser un compromis entre la précision des estimations et la complexité du nombre de systèmes de poids à calculer.

La méthode ROWN généralise l'approche dans le cas de m dimensions creuses : recherche d'un ensemble minimum de systèmes de poids à stocker dans le cube, en prenant en compte les conditions sur le support et le niveau de confiance.

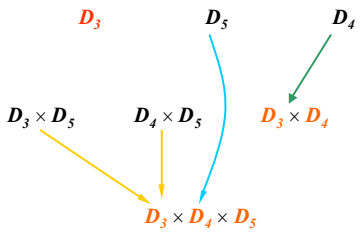
3.4 Réduire le nombre de poids : la méthode ROWN

Les notions de support et de confiance de ROWN sont très proches de celles définies dans l'algorithme *Apriori* de recherche des règles d'association (Agrawal et *al* 1994).

Nous ne détaillons pas notre algorithme (voir Goutier et *al.*, 2002). Il comporte trois grandes étapes, les deux premières adaptées de l'algorithme *Apriori* :

- sélection de tous les échantillons à support valide i.e. supérieur à s_0,
- pour chaque échantillon à support valide, recherche des sous-échantillons susceptibles de le remplacer avec un niveau de confiance supérieur à c_0,
- parmi les solutions, sélection de l'ensemble minimal de remplaçants.

L'arbre de la figure 3 montre pour chacun des croisements, l'échantillon sur lequel sera calculé le poids de redressement. Chaque flèche indique pour chaque croisement, son remplaçant. Par exemple, D_5 sera redressée par le poids calculé sur l'échantillon engendré par les clients renseignés pour le croisement $D_3 \times D_4 \times D_5$.



initial	remplaçant
D_3	D_3
D_4	$D_3 \times D_4$
$D_3 \times D_4$	$D_3 \times D_4$
D_5	$D_3 \times D_4 \times D_5$
$D_3 \times D_5$	$D_3 \times D_4 \times D_5$
$D_4 \times D_5$	$D_3 \times D_4 \times D_5$
$D_3 \times D_4 \times D_5$	$D_3 \times D_4 \times D_5$

FIG. 3 – *Exemple d'application de la méthode au cas de trois dimensions creuses avec la table de substitution correpondante*

3.5 Expérimentation sur des bases de données EDF

La méthode proposée a été expérimentée sur plusieurs bases de données d'EDF, où chaque base stocke les données d'un ensemble de clients d'un département. Initialement, huit dimensions creuses avaient été choisies (par exemple l'énergie du chauffage principal, l'énergie de l'eau chaude sanitaire, le type d'habitat, …). Parmi ces huit dimensions, deux ne satisfaisaient pas la condition de support. La prise en compte de l'ensemble des croisements (concernant alors les six dimensions retenues) aurait donc nécessité le calcul de 2^6-1 (=63) poids sur ces données. De plus, il aurait été nécessaire d'effectuer 2^6-1 chargements de données distincts dans le cube multidimensionnel. La méthode ROWN nous a permis de réduire ce nombre prohibitif à 8 poids en moyenne. Bien sûr, certaines corrélations entre les dimensions creuses (en particulier la corrélation entre l'énergie de chauffage avec l'énergie de l'eau chaude sanitaire) expliquent une telle réduction. Une étude complémentaire devrait être menée pour déterminer l'impact de ces corrélations sur les échantillons retenus.

Nous avons testé également la validité des estimations à partir des échantillons retenus. Dans la mesure où les dimensions denses intervenaient comme variables de stratification (par exemple le type de tarif), nous avons simulé aléatoirement une dimension dense à trois modalités. La figure 4 indique l'évolution de l'estimation des proportions des différentes modalités pour cette dimension dense en fonction du support choisi. Les échantillons des différents supports ont été créés par croisement successif de plusieurs dimensions creuses. Le système de poids est calculé et permet alors de redresser l'échantillon pour estimer les proportions de la dimension dense étudiée. Nous constatons la convergence des estimations ainsi qu'une validation empirique du choix du seuil pour le support.

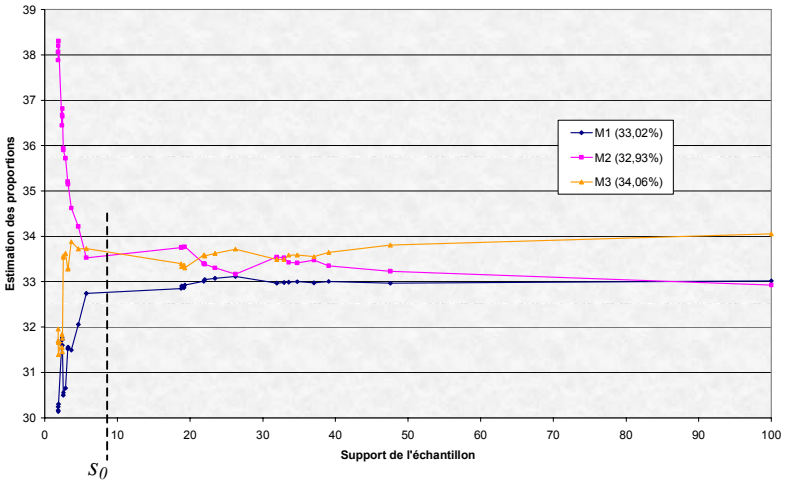


FIG. 4 – *Evolution de l'estimation de la proportion en fonction du support de l'échantillon*

4. Implémentation sous ORACLE EXPRESS

Considérons la mesure correspondant au calcul de la proportion de clients et cinq dimensions D_1, D_2, D_3, D_4 et D_5 telles que D_1, D_2 soient des dimensions *denses* et D_3, D_4, D_5 soient creuses. Les indices des dimensions sont imposés par l'ordre des dimensions dans la définition des mesures. Dans l'exemple suivant, les dimensions sont donc notées D_1, D_2, D_3, D_4 et D_5, alors que jusqu'à présent D_i représentait une dimension creuse et non pas, soit une dimension creuse, soit une dimension dense.

Un schéma du cube sous ORACLE EXPRESS est indiqué sur la Fig. 5. La mesure $W_{1,2}$ (*W_12* sur le schéma du cube) est dimensionnée par D_1 et D_2. De plus, toutes les mesures sont dimensionnées par D_1 et D_2, étant donné que ces deux dimensions sont denses et ne pertubent pas les échantillons.

La modélisation des systèmes de poids dans ce système de gestion de bases multidimensionnelles se déroule en quatre étapes :

- ***définition des dimensions et chargement des positions sur celles-ci*** (une position correspond à une modalité de l'axe; par exemple, 'E' représentant Electricité est une position de la dimension *Chauffage principal*). Chacune des dimensions définies sont hiérarchisées. Le cas le plus simple est celui d'une hiérarchie à deux niveaux, un pour le niveau totalement agrégé, qui permet de ne pas ventiler les mesures selon les positions de cette dimension, et un pour les positions de la dimension. Trois cas se présentent alors:
  - Si la dimension présente des valeurs manquantes, deux stratégies sont envisageables. Soit les valeurs manquantes sont remplacées par une position fictive que nous notons 'MISS', soit aucun recodage n'est effectué.
  - Si la dimension ne présente pas de valeurs manquantes, aucun recodage n'est nécessaire.

- ***chargement des mesures***. Une mesure par système de poids est chargée. Notons que dans le cas où aucun recodage des valeurs manquantes n'est effectué, seules les tables $DETAIL^*$ sont chargées, i.e. les valeurs manquantes ne sont jamais chargées, ce qui entraîne une forte diminution de la volumétrie du cube. Dans l'exemple, quatre chargements sont donc effectués, un par système de poids retenus (voir Fig.3) et un chargement de $DETAIL^*_{1,2}$ (équivalente à $DETAIL_{1,2}$ étant données que D_1 et D_2 sont des dimensions denses) via la mesure de $W_{1,2}$ définie comme une variable dont les valeurs sont toutes égales à 1.
- ***opération d'agrégation (rollup)***. Les dimensions étant toutes hiérarchisées sur au-moins deux niveaux, il est alors nécessaire d'agréger les données. Comme nous l'avons rappelé plus haut dans l'article, les mesures sont additives, ce qui permet un rollup des systèmes de poids.
- ***gestion dynamique des systèmes de poids***. Pour un même croisement, plusieurs systèmes de poids peuvent convenir. En effet, si les systèmes $D_3 \times D_4$ et $D_3 \times D_4 \times D_5$ sont chargés, pour afficher la mesure sur le croisement $D_3 \times D_4$, deux choix sont possibles:
  - soit afficher directement la valeur obtenue par le système $D_3 \times D_4$,
  - soit afficher la valeur par le système $D_3 \times D_4 \times D_5$ en choisissant la position *Total* sur la dimension D_5.

  Pour lever toute ambiguïté sur le système de poids à afficher, un ensemble de formules (n'augmentant pas la volumétrie de la base) est défini pour permettre un affichage "transparent" pour chaque croisement du système de poids retenu.

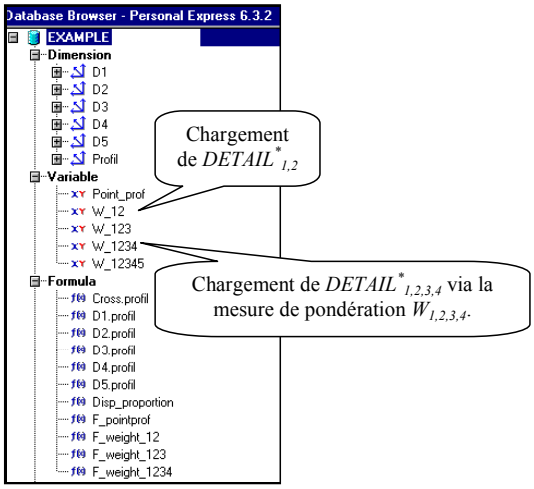


FIG. 5 – *Exemple d'un schéma Oracle Express*

Lors de l'affichage des résultats d'une requête multidimensionnelle, il est nécessaire, comme nous l'avons évoqué plus haut, de définir quelle **mesure redressée** doit être prise en compte en fonction du croisement étudié. Pour cela, un profil par dimension a été créé qui

permet de repérer quelles dimensions interviennent dans un croisement donné. Ce profil (formules *Di.profil* sur le schéma de la figure 5) indique si la dimension est active sur le croisement ou si les valeurs sont agrégées sur cette même dimension. L'ensemble des croisements ayant un support valide est stocké dans une dimension supplémentaire *Profil* et la table de substitution (un exemple est fourni à la figure 3) est définie dans une variable supplémentaire *Point_prof*. Pour finir, la formule *Disp_proportion* retourne les proportions ajustées utilisant le système de poids adéquat correspondant au croisement défini par la requête.

5. Conclusions et perspectives

Dans cet article, nous montrons comment adapter simplement le traitement de la non-réponse aux valeurs manquantes dans les dimensions du cube, appelées alors dimensions creuses. Cette approche consiste à introduire un système de poids associé à toute mesure portant une dimension creuse. Considérant plusieurs dimensions creuses, le problème de complexité se pose. Nous présentons la méthode ROWN qui permet de s'affranchir du nombre exponentiel de systèmes de poids dû à l'ensemble des croisements à évaluer. Un aspect intéressant de notre solution est qu'elle préserve la propriété d'additivité. Enfin, le principe d'implémentation sous Oracle Express est fourni.

Parmi les différentes pistes, deux nous apparaissent essentielles :

- Restituer à l'utilisateur la précision des résultats dans le cube sous la forme d'intervalle de confiance, par exemple.
- Traiter des valeurs manquantes dans les mesures et également introduire les dimensions hiérarchiques.

6. Bibliographie

Agrawal R. et Srikant R. (1994), Fast Algorithms for Mining Association Rules, Proceedings of the 20th Int'l Conference on Very Large Databases, Santiago CHILE, September 1994.

Codd E.F. (1993), Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate, Technical report, E.F. Codd and Associates, 1993.

Deville J.C. (2000), Generalized calibration and application to weighting for non-response, Proceedings of COMPSTAT'2000, Utrecht, Juillet 2000.

Dyreson C.E. (1996), Information retrieval from an incomplete data cube, Proceedings of the 22nd VLDB Conference, Bombay India, pp 532-543, 1996.

Galhardas H. et al. (2001), Declarative Data Cleaning: Language, Model, and Algorithms, Proceedings of VLDB, Roma Italy, September 2001.

Goutier S., Hébrail G. et Stéphan V. (2002), Getting right answers from incomplete multidimensional databases, Revue Ingénierie des systèmes d'information, Editions Hermès, 7(3), pp 67-88, 2002.

Gray J., Bosworth A., Layman A. et Pirahesh H. (1996), Data Cube: A Relational Aggregation Operator Generalizing Group-By Cross-Tabs, and Sub-Totals, Proceedings of the ICDE'96, New-Orleans USA, 1996.

Granquist L. (1997), The New View on Editing, International Statistical Review, 65(3), pp 381-387, 1997.
Laurent A. (2002), Bases de données multidimensionnelles floues et leur utilisation pour la fouille de données, Thèse de Doctorat de l'Université Paris 6, Septembre 2002.
Little R. et Rubin D. (2002), Statistical analysis with missing data, 2nd edition, John Wiley & Sons, Canada, 2002.
Pedersen T.-B., Jensen C. et Dyreson C. (2001), A Foundation for Capturing and Querying Complex Multidimensional Data, Information Systems - Special Issue: Data Warehousing -, 26(5), pp 383-423, 2001.
Rahm E. et Do H. (2000), Data Cleaning: Problems and Current Approaches, IEEE Bulletin on Data Engineering, 23(4), 2000.
Tillé Y. (2001), Théorie des sondages, Editions DUNOD, Paris, 2001.
Vassiliadis P. et Sellis T. (1999), A survey on Logical Models for OLAP Databases, SIGMOD Record 28(4), pp 64-69, 1999.

Summary

In the context of sampling theory, the non response problem has been handled by weighting or imputation techniques. This paper suggests how this formal statistical context can be adapted to missing value problem within OLAP data. In this article, we consider only the case of missing values occurring in dimension attributes, also called sparse dimensions. Integration of these concepts within the OLAP data cube model is solved, by adjusting the data cube measures with a well-chosen weighting system. This method, called ROWN method, is briefly described together with an experimental validation of the quality estimates according to the support level. As a conclusion, we detail the implementation of he weighting method on the ORACLE EXPRESS system.
Keywords: OLAP, data cube, missing values, weighting system.