Manipulation et fusion de données multidimensionnelles

Franck Ravat, Olivier Teste, Gilles Zurfluh
Institut de Recherche en Informatique de Toulouse / Equipe SIG-ED
118, Route de Narbonne 31062 TOULOUSE cedex 04
mél : {ravat, teste, zurfluh}@irit.fr

Résumé. Cet article définit une algèbre permettant de manipuler des tables dimensionnelles extraites d'une base de données multidimensionnelles. L'algèbre intègre un noyau minimum d'opérateurs unaires permettant d'effectuer les analyses décisionnelles par combinaison d'opérateurs. Cette algèbre intègre un opérateur binaire permettant la fusion de tables dimensionnelles facilitant les corrélations des sujets analysés.

1 Introduction

Nos travaux se situent dans le cadre des systèmes décisionnels intégrant des bases de données multidimensionnelles (BDM). Conceptuellement, ces BDM organisent les données en sujets appelés faits et axes d'analyses appelés dimensions (Kimball, 1996).

1.1 Contexte : notre modèle conceptuel

Definition : Un fait F_j est défini par (N_{Fj}, M_{Fj}, I_{Fj}, $IStar_{Fj}$) où
- N_{Fj} est le nom du fait,
- $M_{Fj} = \{m_1, m_2, \ldots, m_w\}$ est un ensemble de mesures (ou indicateurs d'analyse),
- $I_{Fj} = \{I_{F\_1}, I_{F\_2}, \ldots\}$ est l'ensemble des instances de F,
- $IStar_{Fj}$ est une fonction associant chaque instance de I_{Fj} à une instance de chaque dimension liée au fait.

Definition : Une dimension D_i est définie par (N_{Di}, A_{Di}, H_{Di}, I_{Di}) où
- N_{Di} est le nom de la dimension,
- $A_{Di} = \{a_{Di\_1}, a_{Di\_2}, \ldots, a_{Di\_u}\}$ est un ensemble d'attributs,
- $H_{Di} = \{h_{Di\_1}, h_{Di\_2}, \ldots, h_{Di\_y}\}$ est un ensemble de hiérarchies,
- $I_{Di} = \{I_{Di\_1}, I_{Di\_2}, \ldots\}$ est l'ensemble des instances de D_i.

Definition : Une hiérarchie représente une perspective d'analyse précisant les niveaux de granularité auxquels peuvent être manipulés les indicateurs d'analyse. Une hiérarchie h_{Di_x} définie sur la dimension D_i est un chemin élémentaire acyclique débutant par l'attribut de plus faible granularité et se terminant par un attribut de plus forte granularité. Elle est définie par (N_{Di_x}, $Param_{Di\_x}$, $Suppl_{Di\_x}$) où
- N_{Di_x} est le nom de la hiérarchie,
- $Param_{Di\_x} = \langle a_{Di\_k}, a_{Di\_l}, \ldots, a_{Di\_z} \rangle$ est un ensemble ordonné décrivant la hiérarchie des attributs (chaque attribut est appelé paramètre de la hiérarchie et correspond à un niveau de granularité d'analyse),
- $Suppl_{Di\_x}$: $Param_{Di\_x} \rightarrow 2^{(A_{Di} - Param_{Di\_x})}$ est une application spécifiant les attributs faibles qui complètent la sémantique des paramètres (chaque paramètre est associé à un ensemble d'attributs faibles).

Ces définitions textuelles servent de base à notre algèbre et sont complétées par une représentation graphique décrite dans la partie implantation (section 4).

1.2 Problématique et état de l'art

Dans le cadre des BDM, la manipulation la plus couramment utilisée s'effectue au travers d'une Table Dimensionnelle (TD) : tableau affichant les valeurs des mesures en fonction de deux dimensions. Différents travaux de recherche ont proposé une algèbre pour la définition et la manipulation des TD (Marcel 1999). Une proposition importante est l'opérateur cube (Gray *et al*., 1996). (Agrawal *et al*., 1995) propose différents opérateurs algébriques dans le contexte R-OLAP tandis que (Abello *et al*., 2003) (Cabibbo *et al*., 1998) définissent plusieurs opérateurs pour un langage algébrique et/ou graphique. A l'heure actuelle, il n'existe pas de consensus sur la définition d'un noyau minimum complet offrant une algèbre d'interrogation multidimensionnelle, à l'instar de l'algèbre relationnelle qui offre un support parfaitement défini et reconnu. Les algèbres proposées ne proposent pas d'opérateurs effectuant des analyses pour des BDM multi-faits avec une multi-hiérarchisations.

De plus, les décideurs sont couramment confrontés à un problème peu étudié (Ravat, *et al*. 2002) (Benitez-Guerrero, *et al*. 2003), à savoir la fusion du contenu de tables dimensionnelles. Or, la fusion de deux TD répond à un besoin de corrélation nécessaire lors de la prise de décision. Par exemple, supposons qu'un décideur possède une table dimensionnelle TD1 contenant les ventes des produits pour 2002 et une table TD2 contenant les ventes des produits pour 2003. Une première fusion entre ces deux tables lui permettrait de calculer le montant total des ventes de produits en 2002 et 2003. Une seconde fusion pourrait calculer la différence des ventes et donc mettre rapidement en avant la variation des parts de marché. Ces calculs sont actuellement réalisés de manière empirique, et souvent fastidieuse, en effectuant des extractions et des recopies dans des tableurs. Une première solution (Franconi *et al*. 2004) permet d'effectuer l'union, l'intersection et la différence de deux cubes. Cette solution se limite à deux cubes ayant une structure strictement identique.

1.3 Contributions

Souhaitant proposer une algèbre répondant à l'ensemble des opérations qu'un décideur puisse appliquer à une ou plusieurs TD, nous proposons dans cet article, une algèbre permettant de manipuler et fusionner des TD issues d'une BDM intégrant plusieurs faits et plusieurs hiérarchies au sein d'une même dimension. Afin de faire abstraction des spécificités d'implantation d'une BDM, nous souhaitons que notre algèbre se situe au niveau conceptuel. Cette algèbre intégrera non seulement les opérations de manipulation multidimensionnelle classique (section 2) mais également les opérations de fusion de TD (section 3) tout en palliant les inconvénients de (Franconi et al. 2004). Cette algèbre fermée doit intégrer des opérateurs élémentaires à partir desquels les décideurs peuvent construire par combinaisons successives et incrémentales des analyses complexes. Le noyau minimal et complet d'opérateurs de cette algèbre doit servir de support formel au développement

- d'un langage textuel, basé sur une extension de SQL, plus particulièrement destiné aux administrateurs (Ravat et al., 2002),
- d'un langage graphique destiné aux décideurs intégrant une représentation du schéma de la BDM et de la TD (Tournier 2004)(section 4).

2 Opérations Unaires de Algèbre Multidimensionnelle

Notre algèbre de manipulation repose sur le concept de Table Dimensionnelle (TD).

Definition : Une table dimensionnelle TD_{Fk} est définie par (S, Ax, Pred) où

- $S=(F_k, MA_{Fk})$ avec F_k le sujet analysé, $MA_{Fk}=\{(a_i, f),\ldots\}$ ensemble des indicateurs (mesures et fonctions d'agrégation associées) - $MA_{Fk} \subseteq M_{Fk}$ -.
- $Ax=<(D_1, h_{D1}, Att_{D1}, pos_{D1}), (D_2, h_{D2}, A_{D2}, pos_{D2}),\ldots>$ est un ensemble de n-uplets représentant les axes de l'analyse ; D_1 est en colonnes et D_2 en lignes.
- Pred est un prédicat de sélection sur les mesures, paramètres et/ou attributs faibles.

Exemple : La TD ci-dessous visualise le total horaire et les montants annuels et trimestriels des cours effectués dans le bâtiment B1. TD est définie par (S, Ax, Pred) :

- S = (Assurer, {(NbHeures, SUM), (Montant, SUM)}, _),
- Ax = < (Cours, h_Ty, <TypeCours, IdCours>, PosCours), (Temps, h_Tps, <Année, Trimestre, Mois>) >, PosTemps), (Enseignant, _, _, _), (Salle, _, _, _) >,
- Pred = Salle.Idbatiment='B1'.

Assurer NbHeures, Montant			COURS | h_Ty						
		TypeCours	TD			TP			CM
		IdCours	C1	C2	C3	C4	C5	C6	C7
Année	Trimestre								
TEMPS | h_Tps	2003	T1-03	(10, 100)	(11, 110)					(6, 75)
		T2-03				(5, 60)		(15, 170)	
		T3_03	(10, 100)		(6, 75)				
		T4_03	(15, 170)	(11, 110)			(10, 100)		(6, 75)
	2002	T1-02	(5, 60)		(5, 60)			(11, 110)	
		T2-02	(5, 62)						(15, 170)
ENSEIGNANT.All='all' Salle.IdBatiment='B1'									

FIG. 1 – *Représentation graphique d'une table dimensionnelle.*

Les opérateurs unaires de notre algèbre permettent de créer une TD (Display) et dans construire de nouvelles. Ces opérateurs sont présentés dans le tableau ci-dessous.

Opérateurs	Description
Display(C, F_k, MA_{Fk}, D_1, D_2, h_{D1}, h_{D2}, Att_{D1}, Att_{D2}) = TD_{Fk}	Construction d'une TD à partir d'un schéma C en constellation (multi-faits)
HRotate(TD_{Fk}, Di, h_{Di_1}, h_{Di_2}) = TD_{FkR}	Permutation de la hiérarchie h_{Di_1} et h_{Di_2}
DRotate(TD_{Fk}, D_1, D_2, h_{D2_x}) = TD_{FkR}	Permutation de la dimension D_1 et D_2
FRotate(TD_{Fk}, F_{k1}, F_{k2}) = TD_{FkR}	Permutation de la dimension F_{k1} et F_{k2} ayant des dimensions communes
DrillDown(TD_{Fk}, D_i, a_1) = TD_{FkR}	Forage descendant par ajout de paramètres de granularité inférieure
RollUp(TD_{Fk}, D_i, a_1) = TD_{FkR}	Forage ascendant par suppression de paramètres
AddM(TD_{Fk}, m_i, f) = TD_{FkR}	Ajout d'une mesure
DelM(TD_{Fk}, m_i, f) = TD_{FkR}	Suppression d'une mesure
Select(TD_{Fk}, $pred$) = TD_{FkR}	Restriction du domaine des valeurs visualisées
Switch(TD_{Fk}, D_i, a_{Di_x}, v_1, v_2) = TD_{FkR}	Permutation des valeurs d'un paramètre affiché
Nest(TD_{Fk}, D_i, a_{Di_x}, a_{Di_y}) = TD_{FkR}	Permutation des paramètres d'une dimension

TAB 1 – *Liste des opérateurs unaires.*

Exemple. Un décideur souhaite modifier la TD de la figure 1, intitulée TD1, afin d'analyser le nombre d'heures qu'effectue un enseignant (statut et code) pour un module (classé par code de formation et de module). Pour répondre à cette analyse, le décideur doit appliquer la combinaison d'opérations algébriques suivante :

Select(**DelM**(**DrillDown**(**DrillDown**(**HRotate**(**DRotate**(*TD1*, TEMPS, ENSEIGNANT, h_St), COURS, h_Ty, h_Fo), COURS, IdModule), ENSEIGNANT, CodeE), Montant, Sum), salle.All='all') = *TD2*

3 Opérations Binaires de Algèbre Multidimensionnelle

Nous proposons de spécifier les concepts de tables fortement, semi et faiblement compatibles. Deux tables TD1 et TD2 sont

- **fortement compatibles** si et seulement si les mesures des deux tables sont analysées en fonction des mêmes critères (paramètres, valeurs des paramètres, dimensions en ligne, en colonne et non développées identiques)
- **semi-compatibles** si et seulement si les paramètres sont identiques dans les deux tables mais les valeurs et/ou les mesures ne sont pas nécessairement communes. Les dimensions non développées sont positionnées sur les mêmes paramètres.
- **faiblement compatibles** si et seulement si les paramètres et/ou les mesures ne sont pas nécessairement communs. Les dimensions non développées sont positionnées sur les mêmes paramètres.

A l'instar des opérations ensemblistes (union, intersection et différence) de l'algèbre relationnelle, l'opérateur de fusion que nous définissons a pour vocation de favoriser la combinaison de plusieurs TD, et donc faciliter les corrélations d'analyses.

Définition : La fusion combine deux tables fortement compatibles pour en construire une nouvelle de même structure où les valeurs des mesures sont calculées à l'aide d'une fonction.

Fusion(TD_{Fk1}, TD_{Fh2}, Fonct) = TD_{FkhR} où

- TD_{Fk1} et TD_{Fh2} sont deux tables dimensionnelles fortement compatibles,
- Fonct : fonction de calcul associée à la fusion (sum, min, max, diff,…)
- TD_{FkhR} est la table dimensionnelle résultat.

Exemple : Nous désirons fusionner les tables TD_A et TD_B faiblement compatibles.

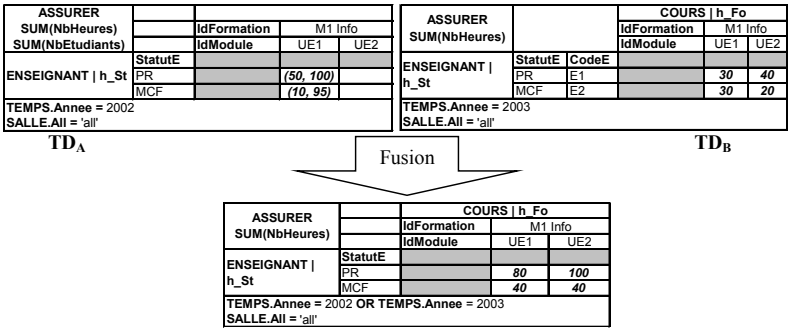

FIG. 2 – *Fusion de tables faiblement compatibles*

La fusion de tables semi ou faiblement compatibles s'effectue par l'application d'opérateurs unaires afin d'obtenir des tables fortement compatibles pour appliquer l'opération de fusion définie précédemment. Dans l'exemple précédent,

- soit on réduit les paramètres et les mesures aux valeurs communes (cf. figure 2)

Fusion(**DelM**(TD_A,NbEtudiant,Sum), **RollUp**(TD_B,ENSEIGNANT,StatutE))

- soit on étend les paramètres et les positions.

Fusion(**DrillDown**(TD_A,ENSEIGNANT,CodeE), **AddM**(TDB,NbEtudiant,Sum))

4 Implantation dans l'outil Graphic-OLAPSQL

Afin de valider nos propositions, nous développons au sein de notre équipe l'outil Graphic-OLAPSQL. A l'aide de cet outil, un décideur peut visualiser et interroger les constellations grâce à un langage graphique simple et incrémental (Tournier 2004). Le langage graphique permet d'exprimer des requêtes en manipulant directement le schéma de la BDM (basé sur une extension du formalisme de (Golfarelli et al. 1998)). Pour construire la requête, l'utilisateur sélectionne les éléments à l'aide de menus contextuels. La figure suivante présente les étapes de construction d'une requête graphique. L'utilisateur sélectionne un fait, deux hiérarchies distinctes, puis il soumet la requête à l'outil Graphic-OLAPSQL. Ce dernier transforme la requête graphique en un format interne algébrique (**Display**(AnalyseCours, Assurer, {(NbEtudiants,UM)}, Enseignant, Temps, h_Ens, h_Tps, <Statute,Codee>, <Année>)) puis retourne une TD visualisant les données de l'analyse.

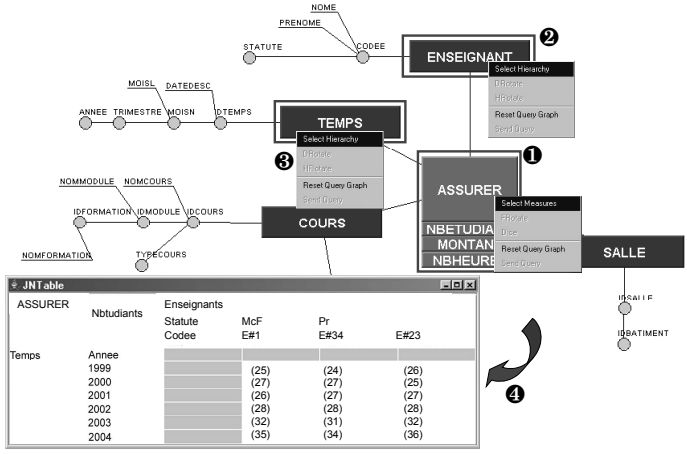

FIG. 3 – *Construction d'une requête graphique*

5 Conclusion et Perspectives

Dans cet article, nous avons proposé une algèbre manipulant les TD issues de BDM multi-faits analysés à l'aide de dimensions multi-hiérarchies. Les opérateurs unaires permettant de créer et de modifier une TD pour en construire une nouvelle (rotation, forage,

sélection, ordonnancement, ajout et suppression). Nous avons également proposé un opérateur de fusion de TD fortement compatibles afin de faciliter les corrélations inter-tables dimensionnelles inhérentes à un processus d'analyse. Pour les tables semi ou faiblement compatibles, il suffit de combiner les opérateurs unaires et l'opérateur de fusion. Le noyau minimal et complet d'opérateurs de cette algèbre fermée sert de support théorique aux futurs développements.

Nous souhaitons étendre le langage graphique destiné aux décideurs (Tournier 2004) afin de proposer une fusion graphique de TD. Nous envisageons également étendre le langage OLAP SQL (Ravat, et al. 2002) destiné aux concepteurs de BDM afin de construire des TD. De plus, à l'instar de l'algèbre relationnelle, nous souhaitons étudier l'optimisation de l'interrogation multidimensionnelle en développant des algorithmes d'ordonnancement des opérateurs algébriques du noyau minimal.

Références

A. Abello, J. Samos, and F. Saltor (2003), Implementing operations to navigate semantic star schemas, DOLAP, 2003.

R. Agrawal, A. Gupta, S. Sarawagi (1995), Modeling Multidimensional Databases, Research Report, IBM Almaden Research Center, San Jose (California), 1995

E. Benitez-Guerrero, C. Collet, M. Adiba (2003), Le système WHES pour l'évolution des entrepôts de données, BDA 2003, Lyon.

L. Cabibbo, R. Torlone (1998), From a Procedural to a Visual Query Language for OLAP, Proceedings of the 10th IEEE International Conference on Scientific and Statistical Database Management - SSDBM'98, 1998.

E. Franconi, A. Kamble (2004), The GMD Data Model and Algebra for Multidimensional Information, CAISE'04.

M. Golfarelli, D. Maio, S. Rizzi (1998), Conceptual design of data warehouses from E/R schemes, 31st Hawaii International Conference on System Sciences, 1998.

J. Gray, A. Bosworth, A. Layman, and H. Pirahesh (1996), Data cube: a relational aggregation operator generalizing group-by, cross-tabs and subtotals, Int'l Conf. on Data Engineering, 1996.

R. Kimball (1996), The data warehouse toolkit, John Wiley and Sons, 1996.

P. Marcel (1999), Modeling and querying multidimensional databases: An overview, Networking and Information Systems Journal. Volume 2, Number 5-6/1999, pp 515-548.

F. Ravat, O. Teste, G. Zurfluh (2002), Langage pour Bases Multidimensionnelles: OLAP-SQL, Revue ISI-NIS, Volume 7 - n°3/2002 - ISBN 2-7462-0579-3, p.11-38.

R. Tournier (2004), Bases de données multidimensionnelles : étude et implantation d'un langage graphique, rapport IRIT/2004-1-D, Juin 2004.

Summary

This article defines an algebra permitting to manipulate dimensional tables, which are issue from a multidimensional database. The algebra we define integrates unary operators permitting to decisional analysis by combining operators. Also, a binary operator permits to merge dimensional tables in order to facilitate OLAP analysis correlations.