Chapitre 7

Vers un cadre formel général

" N'importe quel objet peut être un objet d'art pour peu qu'on l'entoure d'un cadre. "

Boris Vian

7.1 Introduction

À la lumière de nos travaux, nous avons vu qu'il est possible d'enrichir l'analyse en ligne en la couplant avec la fouille de données. Nos différentes expériences nous amènent à réfléchir à la mise en place d'une base théorique pour le couplage entre la technologie OLAP et la fouille de données. Pour y parvenir, nous proposons d'utiliser les formalismes d'une algèbre OLAP et d'étendre ces derniers à des opérations OLAP basées sur des techniques de fouille de données.

Par analogie aux algèbres relationnelles, une algèbre OLAP consiste en une formalisation théorique d'un ensemble d'opérateurs dédiés à la manipulation des données multidimensionnelles. Ces opérateurs OLAP sont dotés de capacités de structuration et permettent l'organisation des données selon des axes d'analyse hiérarchisés et des mesures d'observation. Ils offrent aussi des opérations mécaniques capables de naviguer dans un cube et de visualiser ses données sous différents angles de vue afin d'extraire, par la suite, des informations pertinentes. Entre autres, la navigation repose sur des mécanismes d'agrégation des données pour résumer ces dernières à différents niveaux de granularité. Ces agrégations peuvent s'effectuer selon des critères portant sur certains faits ou certains descripteurs.

Dans ce chapitre, nous engageons des réflexions pour étendre les opérateurs OLAP classiques à une nouvelle génération d'opérateurs de *fouille de données en ligne*. L'objectif de cette extension est d'établir un cadre formel général, non seulement pour la structuration et la navigation, mais aussi pour la *description*, la *classification*, l'*explication* et la *prédiction* dans les données multidimensionnelles.

Dans une phase préliminaire, nous proposons un modèle de données multidimensionnelles supportant un *noyau minimal fermé* d'une algèbre OLAP. Ce noyau comporte un nombre minimal et suffisant d'opérateurs algébriques capables d'assurer toutes les manipulations possibles des données multidimensionnelles. Notre algèbre repose sur un ensemble d'opérations de *structuration* et de *navigation*. Ces dernières sont capables de construire un cube, de manipuler sa structure et d'explorer ses données selon plusieurs niveaux de granularité.

Dans une deuxième phase, en se basant sur nos approches de couplage entre l'analyse en ligne et la fouille de données, nous proposons une première tentative d'extension de notre algèbre. Nous formalisons alors deux opérateurs de fouille de données en ligne. Le premier, appelé OrCA, concerne la réorganisation des cubes de données par l'analyse des correspondances multiples. Le second, appelé OpAC, est un opérateur d'agrégation dans les cubes de données par la classification ascendante hiérarchique.

Ce chapitre est organisé de la façon suivante. Dans la section 7.2, nous exposons un état de l'art des travaux proposant des modèles multidimensionnels et des algèbres pour la manipulation des données. Nous fournissons aussi une étude comparative de ces travaux en vue de positionner notre contribution. La section 7.3 introduit notre modèle basé sur la notion d'*espaces de données multidimensionnelles*. Dans la section suivante, nous présentons notre noyau minimal fermé d'opérateur OLAP. Nous présentons la formalisation de nos deux opérateurs de fouille de données en ligne OrCA et OpAC dans la section 7.5. Dans la section 7.6, nous concluons ce chapitre et proposons de nouvelles directions de recherche.

7.2 Modèles multidimensionnels et algèbres OLAP

Malgré le développement des systèmes décisionnels, les bases de données multidimensionnelles et la technologie OLAP manquent d'une formalisation précise. Les concepts et les systèmes des modèles multidimensionnels existent sans fondement théorique stable [Mar98]. Par conséquent, aucun consensus n'a été admis pour les modèles de données multidimensionnelles. De plus, aucune terminologie commune n'a été établie pour ces derniers [VS99]. Le seul dispositif commun à ces modèles repose sur une structuration multidimensionnelle des informations [NHJ03].

En pratique, la modélisation multidimensionnelle se base souvent sur le schéma en étoile ou le schéma en flocons de neige [Inm96, Kim96, CD97]. Dans ces schémas, un fait OLAP représente le sujet d'analyse et il est associé à des dimensions représentant les axes de l'analyse suivant lesquels différentes graduations hiérarchisées peuvent êtres adoptées [RTZ06a]. Cependant, de tels schémas sont basés sur l'intuition plutôt que sur un formalisme précis [NHJ03].

En l'absence d'un formalisme standard pour les bases de données multidimensionnelles, plusieurs travaux de modélisation ont été publiés depuis la fin des années 90. Les études bibliographiques de Vassiliadis et Sellis [VS99], Pedersen *et al.* [PJD01] et Torlone [Tor03] permettent de distinguer deux grandes catégories dans ces travaux.

- Les travaux de la première catégorie se basent sur l'extension des modèles relationnels. Cette catégorie inclut les travaux de Gyssens *et al.* [GLS96], Li et Wang [LW96], Gyssens et Lakshmanan [GL97], Gray *et al.* [GCB+97] et Gingras et Lakshmanan [GL98]. Selon [Tor03], les schémas proposés dans ces travaux traitent les données sous forme de cube. Bien que les notions de fait, de dimension et de mesure sont plus au moins explicites, la hiérarchie entre les niveaux d'agrégation dans une dimension n'est pas explicitement exprimée dans ces schémas.
- La seconde catégorie comprend les travaux qui sont orientés vers la modélisation multidimensionnelle des cubes de données. Les modèles proposés dans cette seconde catégorie sont sémantiquement plus riche [RTZ06a]. Ils permettent d'expliciter précisément les hiérarchies dans les dimensions et offrent par conséquent une manipulation facile des cubes de données [Tor03]. Nous citons, par exemple, les travaux de Agrawal *et al.* [AGS96, AGS97], Cabibbo et Torlone [CT97, CT98], Vassiliadis [Vas98], Lehner *et al.* [LAW98, Leh98], Gebhardt *et al.* [GJJ97], Pedersen *et al.* [PJ99, PJD01], Thomas et Datta [TD01], Trujillo *et al.* [TLMS03], Franconi et Kamble [FK03, FK04], Abelló *et al.* [ASS03, ASS06] et Ravat *et al.* [RTZ06a, RTZ06b].

À l'image des bases de données classiques supportant l'algèbre relationnelle, un modèle multidimensionnel peut aussi supporter une algèbre de manipulation OLAP. Dans les travaux précédents, quelques propositions d'algèbres OLAP ont été présentées. D'une manière générale, les algèbres proposées dans le cadre de la première catégorie reposent sur une adaptation des opérateurs de l'algèbre relationnelle au contexte multidimensionnel. Quant aux algèbres de la seconde catégorie, elles proposent de nouvelles opérations propres à la manipulation des données multidimensionnelles. Parmi ces travaux, nous nous intéressons particulièrement à ceux ayant proposé des modèles pour les données multidimensionnelles supportant une algèbre OLAP. Dans la suite, nous relatons brièvement les formalismes proposés dans ces travaux.

7.2.1 Algèbre pour les matrices bi-dimensionnelles

Dans [GLS96], Gyssens *et al.* proposent un modèle multidimensionnel basé sur les données tabulaires. Ce modèle considère que les données sont rangées dans des matrices bi-dimensionnelles. Dans de telles matrices, le contenu d'une cellule appartient à un ensemble de *symboles*. Ce dernier contient des *noms*, des *valeurs numériques* et une *constante particulière* $\perp$ qui décrit les valeurs nulles (inapplicables

ou inconnues).

Le modèle de Gyssens *et al.* est associé à une algèbre, appelée *algèbre tabulaire* (*tabular algebra*), selon laquelle un programme algébrique est constitué d'instructions, d'itérations et d'affectations. Les affectations sont de la forme $T \leftarrow \langle ope \rangle \langle arg \rangle$, où T est le nom du tableau résultat, $\langle ope \rangle$ est un opérateur algébrique et $\langle arg \rangle$ sont les arguments nécessaires à l'opérateur. Les auteurs définissent quatre types d'opérateurs : les opérateurs *classiques*, les opérateurs de *restructuration*, un opérateur de *transposition* et un opérateur de *suppression des redondances*.

Les opérateurs classiques reposent sur une adaptation de l'algèbre relationnelle. Ils concernent les opérations de l'*union* (*union*), la *différence* (*difference*), le *produit cartésien* (*cartesian product*), le *renommage* (*renaming*), la *projection* (*projection*) et la *sélection* (*selection*). Les opérateurs de restructuration sont dédiés à la manipulation des matrices. La combinaison d'un opérateur de *groupement* (*grouping*) avec un autre opérateur de *fusion* (*merging*) permet de transformer une relation en un tableau croisé. Deux autres opérateurs permettent d'exprimer directement une opération d'*éclatement* (*split*) d'une matrice et son opération réciproque. Grâce à l'opérateur de *transposition* (*transposing*), il est possible d'exprimer pour toute opération concernant les lignes d'une matrice, une opération similaire concernant ses colonnes. L'opérateur de *suppression des redondances* (*clean-up*) élague dans une matrice les données à valeurs nulles $\perp$. Il permet ainsi de condenser la représentation de cette dernière et d'offrir une mise en forme plus pertinente des données.

7.2.2 Algèbre pour les cubes de données

Agrawal *et al.* ont introduit un modèle pour les cubes de données multidimensionnelles [AGS96, AGS97]. Les auteurs considèrent un cube de plusieurs dimensions où chaque dimension est associée à un domaine de valeur. Les modalités et les mesures sont prises parmi les domaines des dimensions. Une cellule dans un tel cube peut prendre (i) la valeur 0 dans le cas où le fait n'existe pas dans le cube ; (ii) la valeur 1 correspondant à la situation où le fait existe mais n'a pas de mesure ; ou (iii) un n-uplet $\langle X_1, \ldots, X_n \rangle$ dans le cas où le fait existe et a des valeurs de mesures.

En plus du modèle multidimensionnel, les auteurs définissent une algèbre de six opérateurs : la *conversion d'une dimension en mesure* (*push*), la *conversion d'une mesure en dimension* (*pull*), la *destruction d'une dimension* (*destroy dimension*), la *restriction* (*restriction*), la *jointure* (*join*) et la *fusion* (*merge*).

Les opérateurs *push* et *pull* convertissent une dimension en une mesure et *vice versa*. Ceci permet un traitement symétrique des dimensions et des mesures du cube. L'opérateur *destroy dimension* réduit la dimensionnalité d'un cube de données en éliminant une de ces dimensions. Les auteurs signalent que cette action est utile pour les dimensions à faible nombre d'attributs [AGS97]. La *restriction* élimine d'une dimension un ou plusieurs attributs qui ne satisfont pas un prédicat logique

P. La *jointure* relie deux cubes de données selon un certain nombre de dimensions communes. L'opérateur *merge* agit sur des attributs d'une dimension et fait appel à une fonction d'agrégation pour le calcul des mesures.

7.2.3 Algèbres pour les cubes de données relationnels

Li et Wang ont formalisé un cube de données basé sur le modèle relationnel [LW96]. Dans ce formalisme, deux algèbres sont proposées pour manipuler les relations et les cubes de données. La première, appelée *algèbre de groupement* (*grouping algebra*), est une extension de l'algèbre relationnelle à la manipulation des relations et des relations de groupement. La seconde, appelée *algèbre multidimensionnelle* (*multidimensional cube algebra*), vient en complément de la première et s'adresse aux cubes multidimensionnels.

Dans un cube multidimensionnel, les auteurs considèrent un ensemble de *noms de dimensions*, un ensemble de *noms d'attributs* et chaque attribut est associé à un *domaine de modalités*. Un ensemble de valeurs scalaires et une valeur nulle représentent le domaine de modalités d'une mesure. Une dimension du cube est définie par son nom et une relation qui lui est associée. Une référence d'une cellule dans ce cube correspond à la mise en relation d'un n-uplet, à partir des relations, avec l'ensemble des modalités des mesures.

L'algèbre multidimensionnelle de Li et Wang comprend six opérateurs : l'*ajout d'une dimension* (*add dimension*), le *transfert* (*transfer*), l'*union* (*union*), l'*agrégation* (*cube aggregation*), la *jointure* (*rc-join*) et la *construction* (*construct*). L'opérateur *add dimension* insère une nouvelle dimension dans le cube de données. Le *transfert* change l'emplacement d'un attribut d'une dimension dans une autre. Les auteurs affirment que cet opérateur est utile pour étudier le comportement de deux attributs en les mettant dans la même dimension [LW96]. L'*union* fusionne deux cubes de même structure selon une dimension de jointure. L'opération *cube aggregation* compresse les relations relatives aux dimensions du cube. Il permet plusieurs agrégations sur un cube de données sans passer par des relations de groupement. Le *rc-join* assure la jointure d'une dimension d'un cube avec une relation. Enfin, la *construction* génère un cube de données à partir d'une relation en spécifiant les attributs dimensionnels et les attributs métriques.

7.2.4 Algèbre pour les informations réparties en niveaux

Le modèle multidimensionnel, proposé par Cabibbo et Torlone, prend en compte l'aspect granulaire de l'information dans un cube de données [CT97, CT98]. Les cubes sont manipulés à l'aide d'un calcul paramétré par des fonctions interprétées.

Dans la définition d'une dimension, les auteurs introduisent un *ensemble de noms* pour ses niveaux hiérarchiques. Un ordre partiel permet l'ordonnancement

des niveaux de chaque hiérarchie. De plus, chaque niveau est associé à un ensemble fini de constantes présentant le domaine de ses modalités. Les auteurs définissent également un schéma d'un modèle multidimensionnel. Il consiste en un ensemble fini de dimensions et un ensemble fini de *tables de faits*. Le modèle comprend aussi un ensemble fini de *descriptions* des niveaux hiérarchiques des dimensions. Une référence d'une cellule, dans ce modèle, est une instance du schéma multidimensionnel. Une telle instance associe une mesure d'une table de faits à un ensemble de modalités prises dans les niveaux hiérarchiques des dimensions.

Bien qu'elle prenne en compte l'aspect granulaire de l'information souvent présent dans un cube de données, cette proposition n'établit pas un cadre algébrique complet pour les opérations OLAP. Les auteurs se limitent à la formalisation d'une famille de *fonctions de forage* (*roll-up functions*).

7.2.5 Algèbre pour les cubes vus comme des ensembles de relations

Dans le modèle proposé par Gyssens et Lakshmanan dans [GL97], le contenu d'un cube est dissocié de sa structure multidimensionnelle. Alors que son contenu correspond à celui d'un ensemble de tables relationnelles, la structure d'un cube est associée aux noms de ses dimensions, aux noms de ses attributs et à la nature de ces derniers (attributs de dimension ou attributs de mesure). Dans ce modèle, les auteurs distinguent entre deux *ensembles de symboles* : un *ensemble de noms* et un *ensemble de valeurs*. Ils introduisent aussi un *schéma multidimensionnel* de tables relationnelles (*n-dimensional table schema*) qui comprend un ensemble de dimensions, un ensemble d'attributs et une fonction pour associer à chaque dimension un sous-ensemble d'attributs. Une instance de ce schéma multidimensionnel correspond à un ensemble fini de $n + 1$ tables relationnelles composées de n tables de dimensions et une table de faits. Chaque dimension correspond à une table caractérisée par un identifiant unique pour chaque n-uplet. Cet identifiant permet de relier la table de la dimension avec la table de faits.

Le modèle de Gyssens et Lakshmanan supporte aussi une algèbre, fortement inspirée de l'algèbre relationnelle, pour la manipulation des cubes de données [GL97]. Elle se base sur la dissociation entre la structure du cube et les relations qui décrivent son contenu. Ainsi, l'algèbre se compose de deux ensembles d'opérations. Le premier repose sur des *opérateurs classiques* repris de l'algèbre relationnelle (les *opérateurs unaires*, l'*union*, l'*intersection*, la *différence*, et le *produit cartésien*). Le second ensemble propose deux opérateurs (*unfold* et *fold*) de restructuration qui concernent la manipulation de la structure d'un cube. L'opérateur *unfold* permet d'ajouter une nouvelle dimension dans le schéma relationnel d'un cube de données. Cet opérateur est équivalent à l'opérateur *add dimension* de [LW96]. Quant à l'opérateur *fold*, d'une manière équivalente à l'opérateur *destroy dimension* de [GLS96], il permet d'éliminer une dimension d'un cube de données.

7.2.6 Algèbre pour l'analyse en ligne OLAP

Thomas et Datta proposent un modèle de représentation des cubes de données et une algèbre dédiée à l'analyse en ligne [TD01]. Les auteurs reprochent aux modèles précédents le fait qu'ils ne considèrent pas la symétrie possible entre les dimensions et les mesures. Ils proposent un nouveau modèle de données multidimensionnelles qui prend en compte ce problème. Selon [TD01], un cube est composé par un *ensemble de caractéristiques*, un *ensemble d'attributs* et un *ensemble de cellules*. Une caractéristique peut représenter une dimension comme elle peut représenter une mesure du cube. Elle est aussi associée à un sous-ensemble d'attributs ordonnés selon un ordre hiérarchique. Une cellule du cube est identifiée par une *adresse* et un *contenu* dans l'espace de représentation du cube.

Ce modèle multidimensionnel supporte une algèbre de neuf opérateurs dédiés à la manipulation OLAP : la *restriction* (*restriction*), la *projection métrique* (*metric projection*), le *renommage* (*rename*), le *produit cubique* (*cubic product*), l'*union* (*union*), la *différence* (*difference*), l'*agrégation* (*aggregation*), la *conversion en mesure* (*force*) et la *conversion en dimension* (*extract*). L'opérateur de *restriction* permet de restreindre un ou plusieurs attributs à des valeurs particulières selon une condition exprimée sous forme de prédicats de premier ordre. La *projection métrique* réduit le nombre d'attributs de mesure pris en compte dans la représentation des données. Le *renommage* permet de renommer les éléments d'un ensemble d'attributs ou de caractéristiques. Le *produit cubique* combine deux cubes de données. Le produit de deux cubes de données est équivalent au produit cartésien classique de deux tables relationnelles. Alors que l'opérateur d'*union* consiste à unir deux cubes de données, la *différence* consiste plutôt à trouver la différence entre eux. L'*agrégation* permet de créer des agrégats à partir d'un ou de plusieurs attributs dimensionnels. Les mesures des cellules sont, par conséquent, calculées selon une fonction d'agrégation. L'opérateur *force* convertit une dimension en une mesure. Il est équivalent à l'opérateur *push* de [AGS96, AGS97]. Quant à l'opérateur *extract*, d'une manière similaire à l'opérateur *pull* [AGS96, AGS97], il effectue l'opération inverse en convertissant une mesure en une dimension.

7.2.7 Algèbre pour les tables multidimensionnelles

Ravat *et al.* proposent un modèle qui sert de support à une algèbre et un *langage graphique* pour l'analyse et la visualisation des données [RTZ06a, RTZ06b]. Il repose sur une *représentation multi-faits* où chaque fait est analysé en fonction d'axes d'analyse (dimensions) *multi-vues* (multi-hiérarchisées). Selon [RTZ06a], une dimension contient un certain nombre d'attributs, appelés *paramètres*, où chaque attribut représente une façon de graduer l'axe d'analyse. Les différents attributs d'une dimension sont organisés au sein d'une ou de plusieurs hiérarchies. Une hiérarchie est

un *chemin élémentaire acyclique* débutant par l'attribut de plus forte granularité et se terminant par celui de plus faible granularité.

Les auteurs représentent un cube de données par une structure de visualisation proche des arbres d'attributs [GMR98]. Ils se basent sur un tableau à double entrée hiérarchisée, appelé *table multidimensionnelle* (TM) [RTZ01]. L'algèbre proposée repose sur un opérateur de *construction*, produisant une TM à partir d'une base de données multidimensionnelles, et un ensemble d'*opérateurs fondamentaux* (*rotation*, *forage vers le bas*, *forage vers le haut*, *imbrication*, *sélection*, *classement*, *agrégation*, *conversion d'un paramètre*, *conversion d'une mesure*, *ajout/suppression de mesures*) portant sur les TM et facilitant la manipulation OLAP [RTZ06a].

L'opérateur de la *rotation* (*drotate*) permet, au sein d'une TM, soit de changer un axe d'analyse par un autre, soit de changer la hiérarchie sur un même axe. Les opérations de *forages vers le bas* ou *vers le haut* (*drilldown* ou *rollup*) permettent de changer le niveau de granularité des données observées dans la TM. Elles permettent de modifier les différents niveaux de graduation utilisés pour visualiser les données. L'*imbrication* (*nest*) permet d'intégrer, dans les dimensions d'une TM, les données provenant d'une ou de plusieurs dimensions. La *sélection* (*select*) restreint l'ensemble des valeurs affichées. La restriction peut bien porter sur les valeurs des attributs que sur les mesures du fait. L'opérateur de *classement* (*switch*) intervertit deux valeurs d'un attribut d'une dimension pour permettre l'ordonnancement des valeurs affichées. Le *calcul d'agrégats* (*aggregate*) permet d'ajouter dans une TM des calculs agrégeant ses lignes et/ou ses colonnes. D'une manière équivalente à l'opération *push* de [AGS96, AGS97], l'opération de *conversion d'un paramètre en mesure* (*push*) transforme un paramètre afin qu'il apparaisse dans la TM comme une mesure. Les valeurs du paramètre sont alors affichées dans les cellules. La *conversion d'une mesure en paramètre* (*pull*) transforme une mesure en paramètre de la TM. Les valeurs des mesures sont affichées au niveau des en-têtes en ligne ou en colonne. Cette opération est aussi équivalente à l'opération *push* de [AGS96, AGS97]. Les opérations d'*ajout* (*addm*) et de *suppression* (*delm*) *de mesures* permettent de modifier l'ensemble des mesures analysées.

Ravat *et al.* proposent aussi des opérations de *second niveau* construites par combinaison des opérations élémentaires suscitées. Le but de ces dernières est de répondre à des besoins d'analyse complexes. Ces opérations concernent la *rotation des hiérarchies* (*hrotate*), la *projection d'un paramètre* (*plot*), l'*ordonnancement des valeurs d'un paramètre* (*order*), la *rotation de faits* (*frotate*) et la *désélection* (*unselect*) [RTZ06a].

7.2.8 Discussion et positionnement

Plusieurs modèles pour les données multidimensionnelles ont été proposés. Bien qu'ils reposent tous sur une structuration multidimensionnelle des informations,

Proposition	Type du modèle		Hiérarchies des dimensions		Hiérarchies par dimension		Symétrie dimensions/mesures	
	Relationnel	MD	Oui	Non	Multiple	Unique	Oui	Non
Gyssens *et al.*	●			●				●
Agrawal *et al.*		●	○			●	●	
Li et Wang	●		○			●	○	
Cabibbo et Torlone		●	●		●		○	
Gyssens et Lakshmanan	●			●	○		●	
Thomas et Datta		●	●		○		●	
Ravat *et al.*		●	○		●		○	
Notre proposition		●	●		●		●	

● Caractéristique vraie

○ Caractéristique moins vraie

TAB. 7.1 – Comparaison des propositions pour les modèles de données multidimensionnelles

ces modèles adoptent des terminologies différentes, des notations diverses et des formalismes variés. D'une manière générale, ceci s'explique par le fait que les systèmes d'aide à la décision, reposant sur la technologie OLAP, ont existé avant la définition d'un fondement théorique formel standard et reconnu par la communauté des bases de données. Néanmoins, le mécanisme commun des modèles multidimensionnels se base sur une idée intuitive où les faits OLAP sont analysés selon plusieurs dimensions et sont observés selon un certain nombre de mesures.

Comme le montre le tableau 7.1, deux types de modèles existent. Le premier repose sur une extension des concepts du modèle relationnel [GLS96, LW96, GL97]. Le modèle conceptuel du schéma en étoile [Inm96, Kim96, CD97] et ses dérivés sont pris comme référence pour la modélisation multidimensionnelle des données. Dans le second type [AGS96, AGS97, CT97, CT98, TD01, RTZ06a, RTZ06b, RTZ01], les modèles proposés se détachent de la représentation relationnelle des données. Ils considèrent que les données sont déjà dans un espace de représentation multidimensionnel. Par exemple, le modèle de Thomas et Datta [TD01] est indépendant du contexte des données relationnelles. Les auteurs établissent un cadre de représentation multidimensionnel générique en considérant le cube de données à la fois comme source initiale des données et comme support pour l'analyse en ligne.

Le modèle que nous proposons s'insère dans la seconde catégorie. Il repose sur une modélisation multidimensionnelle dédiée à la manipulation des données pour l'analyse en ligne et la fouille de données. Dans notre formalisation, nous faisons abstraction des sources possibles de données.

Mises à part les propositions de Gyssens *et al.* [GLS96] et de Gyssens et Lakshmanan [GL97], les autres tentent d'inclure l'aspect hiérarchique des dimensions dans leurs modèles. Cabibbo et Torlone [CT97, CT98] définissent, pour chaque dimension, un ensemble de niveaux organisés selon un ordre partiel. Dans [TD01], les hiérarchies sont aussi mises en valeur selon la définition d'ordres partiels pour les attributs de chaque caractéristique (dimension) d'un cube de données. Cependant, la considération de plusieurs hiérarchies dans une dimension n'est pas prise en compte d'une manière très claire dans les modèles proposés. Ravat *et al.* [RTZ06a] considèrent une dimension comme un ensemble de chemins reliant les attributs selon un ordre hiérarchique. La plupart des formalismes proposés reposent sur une symétrie dans les traitements entre les dimensions et les mesures du modèle. Au même titre qu'une dimension, une mesure est souvent considérée comme un axe d'analyse à un seul niveau hiérarchique. Par exemple, dans [TD01], Thomas et Datta considèrent que les dimensions et les mesures d'un cube de données appartiennent toutes à un ensemble de caractéristiques propre au cube.

Notre modèle multidimensionnel repose sur une formalisation générale d'un *espace de données multidimensionnelles*. Ce dernier considère plusieurs hiérarchies par dimension. De plus nous ne faisons pas de distinction préalable entre une mesure et une dimension. Dans notre modèle, une mesure peut-être perçue comme une dimension dans un cube et une dimension peut aussi être perçue comme une mesure dans ce dernier.

Proposition	Granularité des données		Fonctions d'agrégation		Noyau minimal fermé		Combinaisons complexes	
	Oui	Non	Oui	Non	Oui	Non	Oui	Non
Gyssens *et al.*		●		●		●		●
Agrawal *et al.*	○		○			●		●
Li et Wang	●		●			●		●
Cabibbo et Torlone	○		○			●		●
Gyssens et Lakshmanan		●		●		●		●
Thomas et Datta	●		●		○		●	
Ravat *et al.*	●		●		●		●	
Notre proposition	●		●		●		●	

● Caractéristique vraie

○ Caractéristique moins vraie

Tab. 7.2 – Comparaison des propositions pour les algèbres OLAP

À l'image des formalismes qui étendent les concepts de la modélisation relationnelle, plusieurs algèbres OLAP proposées sont fortement inspirées de l'algèbre relationnelle. Dans l'algèbre tabulaire de Gyssens *et al.* [GLS96], à l'exception de

l'opérateur de l'*union*, les autres opérateurs fonctionnent soit exclusivement en ligne, soit exclusivement en colonne. Par exemple, la *sélection* permet de retenir les lignes d'un cube dont les données satisfont un certain nombre de critères. La *projection* permet de retenir plutôt un ensemble de colonnes. Le modèle multidimensionnel de Gyssens et Lakshmanan [GL97] est basé sur les tables relationnelles. Ce dernier dissocie entre le contenu et la structure d'un cube de données afin de simplifier la définition des opérateurs algébriques. L'algèbre proposée par Agrawal *et al.* [AGS96, AGS97] est aussi perçue comme une extension de l'algèbre relationnelle au contexte multidimensionnel. Nous considérons que la proposition de Li et Wang [LW96] est *à mi-chemin* entre le modèle relationnel est le modèle multidimensionnel des données. En employant deux types d'algèbres, les auteurs étendent le modèle relationnel au modèle multidimensionnel.

D'une manière générale, comme le montre le tableau 7.2, les travaux qui prennent en compte les hiérarchies des dimensions dans leurs modèles ont aussi mis en place un certain nombre d'opérateurs algébriques afin de manipuler les granularités des données. Dans l'algèbre de Agrawal *et al.* [AGS96, AGS97], les auteurs ont eu recours à des fonctions externes, notamment pour l'opération de *forage vers le haut* (*merge*), afin de spécifier la construction des groupements et des agrégations. Les deux algèbres proposées par Li et Wang [LW96] mettent aussi en valeur l'aspect granulaire de l'information. Ils définissent des relations de groupement dans un premier temps (dans l'algèbre de groupement). Dans un second temps, ils définissent un opérateur de *construction de cubes de données* à partir des relations et un opérateur d'*agrégation* (dans l'algèbre multidimensionnelle). Cependant, l'opérateur d'agrégation (*cube aggregation*) de l'algèbre multidimensionnelle n'exploite pas les relations de groupement dans la constitution des groupes. Le modèle de Cabibbo et Torlone [CT97, CT98] inclut dans sa construction les hiérarchies des dimensions d'un cube de données. Ceci permet de spécifier des opérations de *forage vers le haut et vers le bas* dans le modèle lui même. L'opération d'*agrégation* de Thomas et Datta [TD01] exploite également l'aspect hiérarchique des données dans les dimensions du cube. Ravat *et al.* [RTZ06a] sont les seuls qui exploitent les hiérarchies pour réaliser des *forages vers le haut* (*rollup*) *et vers le bas* (*drilldown*) dans la granularité des données.

Nous remarquons que tous les opérateurs d'*agrégation* et de *forage vers le haut* reposent sur des fonctions d'agrégation externes à leurs modèles respectifs. Agrawal *et al.* expliquent ce choix par le souci de conserver un nombre limité d'opérateurs pour des raisons de simplicité [AGS97]. De notre point de vue, nous pensons que la simplicité d'une algèbre doit aussi répondre à un critère de complétude. En effet, une algèbre doit permettre la plus large manipulation possible des données multidimensionnelles en se basant sur un *noyau minimal fermé* d'opérateurs. Ravat *et al.* ont clairement tenté de répondre à ce compromis entre le nombre d'opérateurs et la complétude de l'algèbre [RTZ06a]. De plus, afin de répondre à des besoins d'analyse en ligne avancés, Ravat *et al.* [RTZ06a] et Thomas et Datta [TD01] ont proposé des

opérateurs complexes construits par combinaisons à partir des opérateurs de base.

Dans notre proposition, nous considérons que les fonctions d'agrégation, au même titre que les dimensions, les hiérarchies et les mesures, font partie intégrante du modèle multidimensionnel d'un cube de données. Notre modèle n'est pas seulement perçu comme une structure pour la représentation des données, mais aussi comme une base pour la création d'une information agrégée et l'évaluation de cette dernière. Ainsi, un ensemble de fonctions d'agrégation représente une composante indissociable à notre modèle.

Par conséquent, dans l'algèbre que nous proposons, nous ne n'avons pas recours à des fonctions externes pour des opérations telles que l'agrégation ou le forage. De plus, grâce à la prise en compte de l'aspect hiérarchique des dimensions, notre algèbre manipule facilement les données à différents niveaux de granularité. Elle repose sur un noyau minimal fermé d'opérations algébriques. Nous distinguons dans ce noyaux deux familles d'opérations : les opérations de *structuration* et les opérations de *navigation*. La combinaison de plusieurs opérations est possible dans notre algèbre afin de répondre à des besoins d'analyse avancés et complexes. En plus de la formalisation des opérations de structuration et de navigation dédiées à l'analyse en ligne classique, nous essayons d'étendre notre algèbre en vue de fournir un cadre formel pour la *fouille de données en ligne*. L'objectif de ce cadre formel est d'établir un fondement théorique général pour une nouvelle génération d'opérateurs d'analyse basée sur le couplage entre l'analyse en ligne et la fouille de données.

7.3 Espace de données multidimensionnelles

Notre modèle multidimensionnel repose sur la notion d'un *espace de données multidimensionnelles*. Pour définir cette notion, nous nous inspirons du formalisme de Thomas et Datta [TD01]. Nous apportons des améliorations à ce dernier en y intégrant un espace de modalités et un ensemble de fonctions d'agrégation.

Définition 7.3.1 (Espace de données multidimensionnelles) *Un espace de données multidimensionnelles est un quintuplet ayant le schéma :*

$$\mathcal{E} = \langle \mathcal{C}, \mathcal{A}, \mathcal{M}, \mathcal{L}, \mathcal{F} \rangle$$

- $\mathcal{C} = \langle C, d \rangle$ *est un espace de caractéristiques ;*
- $\mathcal{A} = \langle A, f, O_A \rangle$ *est un espace d'attributs ;*
- $\mathcal{M} = \langle M, g, O_M, h \rangle$ *est un espace de modalités ;*
- $\mathcal{L}$ *est un ensemble fini et non vide de cellules ;*
- $\mathcal{F}$ *est un ensemble fini et non vide de fonctions d'agrégation.*

Dans la suite, nous définissons et détaillons les propriétés, avec des exemples à l'appui, les différentes composantes du quintuplet représentant l'espace de données

multidimensionnelles $\mathcal{E}$.

7.3.1 Espace de caractéristiques

Soit $\mathcal{C} = \langle C, d \rangle$ un *espace de caractéristiques* où :

- C est un ensemble non vide et fini de p *noms de caractéristiques* $\{c_1, \ldots, c_p\}$ $(p \geq 1)$;
- d est une fonction booléenne qui partitionne C en deux ensembles : un ensemble de *caractéristiques dimensionnelles* C_{dim} et un ensemble de *caractéristiques métriques* C_{mes}. On note $C = C_{dim} \otimes C_{mes}$, c'est-à-dire, $C = C_{dim} \cup C_{mes}$ et $C_{dim} \cap C_{mes} = \emptyset$. La fonction d est définie comme suit :

$$\forall c \in C, d(c) = \begin{cases} 1 & \text{si } c \in C_{dim} \\ 0 & \text{sinon} \end{cases}$$

Soient $p_{dim} = Card(C_{dim})$ le nombre de caractéristiques dimensionnelles et $p_{mes} = Card(C_{mes})$ le nombre de caractéristiques métriques $(p_{dim} + p_{mes} = p)$. Par exemple, C=$\{$*Lieu, Produit, Temps, Quantité*$\}$ est un ensemble de caractéristiques possible pour un cube de données. Dans le cube de la figure 3.3, selon la fonction booléenne d, ces caractéristiques sont réparties comme suit :

- C_{dim}=$\{$*Lieu, Produit, Temps*$\}$;
- C_{mes}=$\{$*Quantité*$\}$.

On suppose que pour $1 \leq i \leq p_{dim}$, $c_i \in C_{dim}$ désigne la $i^{ième}$ caractéristique dimensionnelle dans l'espace $\mathcal{C}$. On suppose également que pour $1 \leq k \leq p_{mes}$, $c_k \in C_{mes}$ désigne la $k^{ième}$ caractéristique métrique dans l'espace $\mathcal{C}$.

7.3.2 Espace d'attributs

Soit $\mathcal{A} = \langle A, f, O_A \rangle$ un *espace d'attributs* où :

- A est un ensemble non vide et fini de t *noms d'attributs* notés $\{a_1, \ldots, a_t\}$ $(t \geq 1)$. Par exemple, A =$\{$*Pays*, *Continent*, *Produit*, *Famille de produits*, *Année*, *Nombre d'articles en stock*, *Nombre d'articles vendus*$\}$ est l'ensemble des attributs du cube de la figure 3.3 ;
- $f : C \longrightarrow 2^A$ est une bijection de l'ensemble des caractéristiques C dans les sous-ensembles des attributs de A. f fait correspondre à chaque caractéristique $c \in C$ un sous-ensemble distinct et non vide d'attributs dans A. Il est à préciser que pour deux caractéristiques différentes, la bijection f fait correspondre deux sous-ensembles d'attributs disjoints. C'est-à-dire, $\forall c, c' \in C, c \neq c', f(c) \cap f(c') = \emptyset$. Inversement, pour chaque attribut a de A, il existe une et une seule caractéristique c de C tel que $a \in f(c)$. En d'autres termes, f crée une partition de p sous-ensembles dans l'ensemble des attributs A où chaque sous-ensemble

correspond à une caractéristique dans C. Par exemple, la partition engendrée par f dans l'ensemble des attributs du cube de la figure 3.3 est :
- $f(Lieu) = \{Pays, Continent\}$;
- $f(Produit) = \{Produit, Familles\ de\ produit\}$;
- $f(Temps) = \{Année\}$;
- $f(Quantité) = \{Nombre\ d'articles\ en\ stock, Nombre\ d'articles\ vendus\}$.

– O_A est un ensemble non vide et fini de p *ordres partiels* $\{o^A_{c_1}, \ldots, o^A_{c_p}\}$. Un ordre partiel de O_A exprime un *ordre hiérarchique* entre des attributs de A. Par exemple, soient a et a' deux attributs de A, on dit que a' est plus petit que a selon un ordre $o^A \in O_A$ si l'attribut a' possède une granularité d'information plus fine que celle de a. On écrit cette relation d'ordre hiérarchique entre a et a' selon la notation :

$$a' \preceq^A_c a$$

Chaque sous-ensemble d'attributs $f(c)$ d'une caractéristique c de C est associé à une relation d'ordre hiérarchique o^A_c dans O_A. $f(c)$ est un ensemble *bien ordonné* d'attributs, c'est-à-dire, chaque partie non vide de $f(c)$ possède un plus petit élément au sens de l'ordre partiel o^A_c. Ce dernier permet ainsi de construire des hiérarchies dans $f(c)$.

Par exemple, supposons que la caractéristique dimensionnelle c ="*Produit*" correspond à l'ensemble des attributs $f(Produit)$ ={*Nom du produit, Catégorie de produits, Nom commercial du produit, Marque commerciale, Producteur*}. La relation d'ordre partiel o^A_c associée à $f(c)$ permet d'établir les deux ordres hiérarchiques suivants :

$$Nom\ du\ produit \preceq^A_{Produit} Catégorie\ de\ produits$$

$$Nom\ commercial\ du\ produit \preceq^A_{Produit} Marque\ commerciale \preceq^A_{Produit} Producteur$$

Définition 7.3.2 (Plus petits attributs d'une caractéristique) *Pour une caractéristique $c \in C$, l'ensemble des plus petits attributs, noté $f(c)^\star$, est un ensemble d'attributs de $f(c)$ où $\forall a \in f(c)^\star$ il n'existe pas d'attribut a' dans $f(c)$ tel que $a' \preceq^A_c a$ selon l'ordre partiel o^A_c.*

Selon cette définition, on identifie pour chaque caractéristique c l'ensemble des attributs ayant les granularités d'information les plus fines dans leurs hiérarchies respectives. Dans l'exemple précédent, $f(Produit)^\star$ ={*Nom du produit, Nom commercial du produit*}.

Pour le cas particulier d'une caractéristique admettant un seul attribut (un singleton), on identifie son ensemble des plus petits attributs au singleton comprenant cet attribut. C'est-à-dire, pour une caractéristique $c \in C$, si $f(c) = \{a\}$ $(a \in A)$, alors $f(c)^\star = \{a\}$.

Définition 7.3.3 (Plus petits attributs de A) *On désigne par $A^\star$ l'ensemble des plus petits attributs dans A de toutes les caractéristiques de C.*

$$A^\star = \{f(c)^\star \subset A, \forall c \in C, \text{ où } f(c)^\star \text{ sont les plus petits attributs de } f(c)\}$$

Par exemple dans le cube de la figure 3.3, l'ensemble des plus petits attributs de A est :

$A^\star$ ={*Pays*, *Produit*, *Année*, *Nombre d'articles en stock*, *Nombre d'articles vendus*}

7.3.3 Espace de modalités

Soit $\mathcal{M} = \langle M, g, O_M, h\rangle$ un *espace de modalités* où :

- M est un ensemble non vide de *modalités* notées $\{b_1, b_2, \ldots\}$;
- $g : A \longrightarrow 2^M$ est une fonction de l'ensemble des attributs A à dans les sous-ensembles des modalités de M. g associe à chaque attribut $a \in A$ un sous-ensemble non vide de modalités dans M. g permet donc d'identifier pour chaque attributs de A son domaine de modalités. Dans l'exemple de la figure 3.3, les domaines des attributs sont :
  - g(*Continent*)={*Amérique, Europe, Asie*} ;
  - g(*Pays*)={*USA, Canada, France, Italie, Espagne, Inde, Japon*} ;
  - g(*Famille de produits*)={*PC, PC por, MP3*} ;
  - g(*Produit*)={*iTwin, iPower, DV-400, EN-700, aStar, aDream*} ;
  - g(*Année*)={*2002, 2003, 2004, 2005*} ;
  - g(*Nombre d'articles en stock*)={*20, 24, 30, 40, 54, 240, 304, 400, ...*} $\subseteq \mathbb{R}$;
  - g(*Nombre d'articles vendus*)={*4, 5, 10, 21, 55, 63, 104, 232, ...*} $\subseteq \mathbb{R}$.
- O_M est un ensemble non vide et fini de t *ordres totaux* $\{o^M_{a_1}, \ldots, o^M_{a_p}\}$. Chaque $o^M_a \in O_M$ établit un ordre total dans le sous-ensemble des attributs $g(a)$, où a est un attribut dans A. Cet ordre correspond à un ordonnancement lexical des modalités d'un niveau de granularité d'une caractéristique. En général, cet ordre correspond à l'ordre alphabétique des noms des modalités. Ceci-dit, dans des cas particuliers, o^M_a peut correspondre à un ordre spécifique différent de l'ordre alphabétique. Par exemple, pour un attribut temporel, l'ordre de ses modalités peut correspondre à l'ordre chronologique des dates. Pour un attribut numérique, l'ordre de ses modalités peut correspondre à un ordre arithmétique croissant ou décroissant. Soient b et b' deux modalités d'un attribut a ($\{b, b'\} \subseteq g(a)$), on dit que b' est plus petit que b si la modalité b' précède la modalité b selon l'ordre total o^M_a. On écrit cette relation d'ordre entre b et b' selon la notation :

$$b' \prec^M_a b$$

Par exemple, dans le cube de la figure 3.3, les modalités de l'attributs $a =$"*Année*" sont ordonnées selon la relation d'ordre chronologique :

$$2002 \prec^{M}_{Année} 2003 \prec^{M}_{Année} 2004 \prec^{M}_{Année} 2005$$

– h est une *fonction d'appartenance* qui associe à une modalité un ensemble de modalités qui lui appartiennent selon un *ordre sémantique*. Avant de définir la fonction h, nous introduisons dans la suite la notion d'ordre sémantique de modalités.

Définition 7.3.4 (Ordre sémantique des modalités) *Soit b et b' ($b \neq b'$) deux modalités dans M. On dit que b' appartient à b selon un ordre sémantique, et on note $b' \sqsubseteq b$, si et seulement si les quatre conditions suivantes sont vérifiées :*

1. $\exists a \in A$ *tel que* $b \in g(a)$;
2. $\exists a' \in A$ *tel que* $b' \in g(a')$;
3. $\exists c \in C$ *tel que* $\{a, a'\} \subseteq f(c)$;
4. $a' \preceq^{A}_{c} a$ *selon l'ordre hiérarchique* o^{A}_{c}.

Il est à remarquer que, d'après la quatrième condition de cette définition, l'attribut a – auquel appartient la modalité b – ne peut pas être dans l'ensemble $f(c)^{\star}$ des plus petits attributs de l'ensemble des attributs $f(c)$ (puisqu'il existe forcément un autre attribut a' qui lui est plus petit). Par conséquent, la fonction d'appartenance h exclut de son ensemble de départ les modalités les plus fines qui correspondent aux plus petits attributs de A.

Définition 7.3.5 (Modalités les plus fines) *On désigne par $M^{\star}$ l'ensemble des modalités les plus fines dans M correspondant à l'ensemble des plus petits attributs $A^{\star}$ selon la bijection g :*

$$M^{\star} = \{b \in M \text{ tel que } \exists a \in A^{\star} \text{ et } b \in g(a)\}$$

Nous définissons la fonction d'appartenance $h : M \backslash M^{\star} \longrightarrow 2^{M}$, de l'ensemble des modalités M privé des modalités les plus fines $M^{\star}$ dans les sous-ensembles de M. Cette fonction associe à chaque modalité b de l'ensemble de départ un sous-ensemble de modalités $h(b)$ qui appartiennent à b selon un ordre sémantique. Formellement, on écrit :

$$\forall b \in M \backslash M^{\star}, h(b) = \{b' \in M \text{ tel que } b' \sqsubseteq b\}$$

Par exemple, dans le cube de la figure 3.3, les appartenances sémantiques des modalités se résument par :

- $h(Amérique)$=$\{USA,\ Canada\}$;
- $h(Europe)$=$\{France,\ Italie,\ Espagne\}$;
- $h(Asie)$=$\{Inde,\ Japon\}$;
- $h(PC)$=$\{iTwin,\ iPower\}$;
- $h(PC\ por)$=$\{DV\text{-}400,\ EN\text{-}700\}$;
- $h(MP3)$=$\{aStar,\ aDream\}$.

7.3.4 Ensemble de cellules

Soit $\mathcal{L}$ un ensemble fini et non vide de *cellules*.

Définition 7.3.6 (Cellule) *Une cellule exprime un fait OLAP. Elle est définie par le couple $\langle adresse, contenu\rangle$ vérifiant les deux conditions suivantes :*

1. *l'adresse est un vecteur à p_{dim} composantes $\langle\beta_1, \ldots, \beta_i, \ldots, \beta_{p_{dim}}\rangle$, où $\forall i \in \{1, \ldots, p_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in C_{dim}$ et $\beta_i \in g(a)$;*
2. *le contenu est un vecteur à p_{mes} composantes $\langle\gamma_1, \ldots, \gamma_k, \ldots, \gamma_{p_{mes}}\rangle$, où $\forall k \in \{1, \ldots, p_{mes}\}$, il existe $a \in f(c_k)$ tels que $c_k \in C_{mes}$ et $\gamma_k \in g(a) \cup \{\perp\}$. $\perp$ étant une valeur nulle désignant une modalité inapplicable ou inconnue.*

Dans l'adresse d'une cellule, une composante β_i représente une position dans l'axe de la $i^{ième}$ caractéristique dimensionnelle c_i. Cette position β_i s'identifie par une modalité qui appartient à un attribut a dans la caractéristique dimensionnelle c_i. C'est-à-dire, la $i^{ième}$ composante de l'adresse d'une cellule appartient à l'ensemble des modalités $g(a)$, où a est un attribut qui appartient à son tour à l'ensemble $f(c_i)$ des attributs de la $i^{ième}$ caractéristique dimensionnelle c_i.

Dans le contenu d'une cellule, une composante γ_k représente la valeur de la $k^{ième}$ caractéristique métrique c_k. Cette position γ_k s'identifie par une modalité qui appartient à un attribut a dans la caractéristique métrique c_k. Si cette modalité métrique n'existe pas, alors γ_k est identifiée par la valeur nulle $\perp$. C'est-à-dire, la $k^{ième}$ composante du contenu d'une cellule appartient à l'ensemble des modalités $g(a) \cup \{\perp\}$, où a est un attribut qui appartient à son tour à l'ensemble $f(c_k)$ des attributs de la $k^{ième}$ caractéristique métrique c_k.

Par exemple, dans le cube de la figure 3.3 :

- $\langle\langle Italie,\ DV\text{-}400,\ 2002\rangle, \langle 240,\ 63\rangle\rangle$ est une cellule qui définit le fait OLAP des ventes de l'article *DV-400* en *Italie* pendant l'année *2002*. Ce fait est mesuré par *240* articles en stock et *63* articles vendus ;
- $\langle\langle France,\ aDream,\ 2002\rangle, \langle\perp, \perp\rangle\rangle$ est une cellule qui définit le fait OLAP des ventes de l'article *aDream* en *France* pendant l'année *2002*. Ce fait étant inexistant ou inconnu, le contenu de sa cellule est vide.

7.3.5 Ensemble de fonctions d'agrégation

Soit $\mathcal{F}$ un ensemble fini et non vide de *fonctions d'agrégation*.

Définition 7.3.7 (Fonction d'agrégation) *Soit* $E = \{e_1, \ldots, e_j, \ldots, e_n\}$ *un ensemble non vide de* n *cellules dans* $\mathcal{L}$*, tel que :*

$$\forall j \in \{1, \ldots, n\},\ e_j = \langle\langle \beta_1^j, \ldots, \beta_i^j, \ldots, \beta_{p_{dim}}^j \rangle, \langle \gamma_1^j, \ldots, \gamma_k^j, \ldots, \gamma_{p_{mes}}^j \rangle\rangle.$$

On dit que ξ *est une fonction d'agrégation sur l'ensemble des cellules* E*, si les trois conditions suivantes sont vérifiées :*

1. $\xi(E) = \langle \Gamma_1, \ldots, \Gamma_k, \ldots, \Gamma_{p_{mes}} \rangle$;
2. $\forall k \in \{1, \ldots, p_{mes}\},\ \Gamma_k = \bigodot_{j=1}^{n}(\gamma_k^j)$;
3. $\bigodot$ *est une fonction de* $\mathbb{R}^n$ *dans* $\mathbb{R}$.

Une fonction d'agrégation permet d'évaluer un ensemble de cellules (faits OLAP) par des mesures agrégées. La fonction $\bigodot$ détermine la nature des mesures agrégées. Par exemple, $\bigodot$ peut s'identifier à :

- une opération de calcul de la *somme* de n éléments dans $\mathbb{R}$. Dans ce cas, $\forall k \in \{1, \ldots, p_{mes}\}$, $\Gamma_k = \sum_{j=1}^{n}(\gamma_k^j)$. On dit que la fonction ξ correspond à l'agrégation SUM ;
- une opération de calcul de la *moyenne* de n éléments dans $\mathbb{R}$. Dans ce cas, $\forall k \in \{1, \ldots, p_{mes}\}$, $\Gamma_k = \frac{1}{n}\sum_{j=1}^{n}(\gamma_k^j)$. On dit que la fonction ξ correspond à l'agrégation AVG ;
- une opération de calcul du *maximum* parmi n éléments dans $\mathbb{R}$. Dans ce cas, $\forall k \in \{1, \ldots, p_{mes}\}$, $\Gamma_k = \max_{j=1}^{n}(\gamma_k^j)$. On dit que la fonction ξ correspond à l'agrégation MAX.

Il est à noter que, selon la définition précédente, une fonction d'agrégation prend en compte les faits OLAP mesurés par des modalités numériques dans l'ensemble des réels $\mathbb{R}$. Dans la terminologie OLAP, on parle de *mesures additives*.

Les valeurs nulles $\perp$ des faits OLAP sont considérées comme les *éléments neutres* de la fonction $\bigodot$. En d'autres termes, ces valeurs ne sont pas prises en compte dans la fonction d'agrégation. Par exemple, soient trois faits OLAP, du cube de la figure 3.3, représentés par les cellules :

- $e_1 = \langle\langle$*Italie, DV-400, 2002*$\rangle, \langle$*240, 63*$\rangle\rangle$
- $e_2 = \langle\langle$*France, aDream, 2002*$\rangle, \langle \perp, \perp \rangle\rangle$
- $e_3 = \langle\langle$*Inde, DV-400, 2002*$\rangle, \langle 30, \perp \rangle\rangle$

La fonction d'agrégation sur $\{e_1, e_2, e_3\}$ selon la somme est égale à $\xi(\{e_1, e_2, e_3\}) = \langle 240 + \perp + 30, 63 + \perp + \perp \rangle = \langle 270, 63 \rangle$

7.4 Noyau minimal fermé d'une algèbre multidimensionnelle

Dans cette section, nous élaborons un ensemble d'opérateurs algébriques basés sur notre espace de données multidimensionnelles. Ces opérateurs constituent un noyau minimal et fermé d'une algèbre OLAP complète. Nous distinguons dans ce noyau deux familles d'opérateurs : les *opérateurs de structuration* et les *opérateurs de navigation*.

7.4.1 Opérateurs de structuration

Cette première famille d'opérateurs permet de construire un cube de données, à partir d'un espace de données multidimensionnelles, et de manipuler sa structure. Elle fournit à l'utilisateur des outils pour créer une représentation multidimensionnelle associée à ses objectifs d'analyse. Ces opérateurs servent à la gestion de l'ensemble des caractéristiques (dimensions et mesures) d'un espace de données multidimensionnelles afin de mettre en place la structure d'un cube de données qui reflète au mieux le contexte d'analyse souhaité par l'utilisateur. Nous avons défini six opérateurs de structuration : la *construction d'un cube de données* (ConstructCube), l'*ajout d'une caractéristique* (AddCharacteristic), la *suppression d'une caractéristique* (DltCharacteristic), l'*imbrication d'un attribut* (Nest), la *conversion d'une dimension* (Push) et la *conversion d'une mesure* (Pull). Dans la suite, nous présentons le rôle et le formalisme de chacun de ces opérateurs.

Construction d'un cube de données

La construction d'un cube de données (ConstructCube) permet de créer un cube $\mathcal{U}_0$ à partir d'un espace de données multidimensionnelles $\mathcal{E} = \langle \mathcal{C}, \mathcal{A}, \mathcal{M}, \mathcal{L}, \mathcal{F} \rangle$. Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_0 &= \textsc{ConstructCube}(\mathcal{E}, \mathcal{C}_0, \mathcal{F}_0) \\ &= \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_0, \mathcal{F}_0 \rangle \end{aligned}$$

avec :

- $\mathcal{C}_0 = \langle C_0, d \rangle \subseteq \mathcal{C}$ est un sous-espace de caractéristiques dans $\mathcal{C} = \langle C, d \rangle$, où la fonction logique d permet de distinguer entre les caractéristiques dimensionnelles $C_{0dim} \subseteq C_{dim}$ et les caractéristiques métriques $C_{0mes} \subseteq C_{mes}$. On suppose que $n_{dim} = Card(C_{0dim})$ ($n_{dim} \leq p_{dim}$), $n_{mes} = Card(C_{0mes})$ ($n_{mes} \leq p_{mes}$) et que $n = n_{dim} + n_{mes}$ est le nombre de caractéristiques du cube à construire ;
- $\mathcal{A}_0 = \langle A_0, f, O_{A_0} \rangle \subseteq \mathcal{A}$ est le sous-espace des attributs engendré par l'ensemble des caractéristiques C_0 selon la bijection f dans $\mathcal{E}$;
- $\mathcal{M}_0 = \langle M_0, g, O_{M_0}, h \rangle \subseteq \mathcal{M}$ est le sous-espace des modalité engendré par l'ensemble des attributs A_0 selon la fonction g dans $\mathcal{E}$;
- $\mathcal{L}_0 \subseteq \mathcal{L}$ est un sous-ensemble fini et non vide de cellules dans $\mathcal{L}$ engendrées par l'ensemble des caractéristiques C_0 ;

– $\mathcal{F}_0 \subseteq \mathcal{F}$ est un sous-ensemble fini et non vide de fonctions d'agrégation dans $\mathcal{F}$.

Dans la suite, nous supposons que $\mathcal{U}_0$ est le cube de départ sur lequel sont appliqués les opérateurs de notre algèbre OLAP.

Ajout d'une caractéristique

Cet opérateur (AddCharacteristic) consiste à rajouter une nouvelle caractéristique $c_{new} \in C \backslash C_0$ dans le cube $\mathcal{U}_0$. Il est aussi valable pour l'ajout d'une dimension ($d(c_{new}) = 1$) que pour l'ajout d'une mesure ($d(c_{new}) = 0$). Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \textsc{AddCharacteristic}(\mathcal{U}_0, c_{new}) \\ &= \langle \mathcal{C}_1, \mathcal{A}_1, \mathcal{M}_1, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{aligned}$$

avec :

– $\mathcal{C}_1 = \langle C_1, d \rangle$, où $C_1 = C_0 \cup \{c_{new}\}$;
– $\mathcal{A}_1 = \langle A_1, f, O_{A_1} \rangle$, où $A_1 = A_0 \cup \{f(c_{new})\}$;
– $\mathcal{M}_1 = \langle M_1, g, O_{M_1}, h \rangle$, où $M_1 = M_0 \cup M_{new}$, avec $M_{new} = \bigcup_{a \in f(c_{new})} g(a)$. M_{new} est l'ensemble des modalités engendrées par la nouvelle caractéristique c_{new} ;
– $\mathcal{L}_1$ est l'ensemble des cellules engendrées par l'ensemble des caractéristiques C_1.

Suppression d'une caractéristique

Cet opérateur (DltCharacteristic) est l'opérateur inverse de l'ajout d'une caractéristique. Il consiste à supprimer une dimension ou une mesure $c_{old} \in C_0$ du cube $\mathcal{U}_0$. Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \textsc{DltCharacteristic}(\mathcal{U}_0, c_{old}) \\ &= \langle \mathcal{C}_1, \mathcal{A}_1, \mathcal{M}_1, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{aligned}$$

avec :

– $\mathcal{C}_1 = \langle C_1, d \rangle$, où $C_1 = C_0 \backslash \{c_{old}\}$;
– $\mathcal{A}_1 = \langle A_1, f, O_{A_1} \rangle$, où $A_1 = A_0 \backslash \{f(c_{old})\}$;
– $\mathcal{M}_1 = \langle M_1, g, O_{M_1}, h \rangle$, où $M_1 = M_0 \backslash M_{old}$, avec $M_{old} = \bigcup_{a \in f(c_{old})} g(a)$. M_{old} est l'ensemble des modalités engendrées par la caractéristique c_{old} ;
– $\mathcal{L}_1$ est l'ensemble des cellules engendrées par l'ensemble des caractéristiques C_1.

Imbrication d'un attribut

L'imbrication d'un attribut (Nest) permet d'intégrer dans une dimension $c \in C_{0dim}$, du cube $\mathcal{U}_0$, un attribut $a_{new} \in f(c_{new})$ provenant d'une nouvelle dimension $c_{new} \in C \backslash C_0$. Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \textsc{Nest}(\mathcal{U}_0, a_{new}, c) \\ &= \langle \mathcal{C}_1, \mathcal{A}_1, \mathcal{M}_1, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{aligned}$$

avec :

- $\mathcal{C}_1 = \langle C_1, d \rangle$, où $C_1 = C_0 \cup \{c_{new}\}$;
- $\mathcal{A}_1 = \langle A_1, f, O_{A_1} \rangle$, où $A_1 = A_0 \cup \{a_{new}\}$;
- $\mathcal{M}_1 = \langle M_1, g, O_{M_1}, h \rangle$, où $M_1 = M_0 \cup g(a_{new})$;
- une cellule $e = \langle adresse, contenu \rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :
  1. l'adresse de e est un vecteur à $n_{dim}+1$ composantes $\langle \beta_1, \ldots, \beta_i, \ldots, \beta_{n_{dim}+1} \rangle$, où $\forall i \in \{1, \ldots, n_{dim}+1\}$, il existe $a \in (f(c_i) \cup \{a_{new}\})$ tels que $c_i \in C_{0dim}$ et $\beta_i \in g(a)$;
  2. le contenu est un vecteur à n_{mes} composantes $\langle \gamma_1, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}} \rangle$, où $\forall k \in \{1, \ldots, n_{mes}\}$, il existe $a \in f(c_k)$ tels que $c_k \in C_{mes}$ et $\gamma_k \in g(a) \cup \{\perp\}$.

Conversion d'une dimension

La conversion d'une dimension (PUSH) permet de transformer un attribut $a \in f(c)$ d'une dimension $c \in C_{0dim}$ en une mesure dans le cube $\mathcal{U}_0$. Formellement, nous écrivons :

$$\begin{array}{rcl} \mathcal{U}_1 & = & \textsc{Push}(\mathcal{U}_0, a) \\ & = & \langle \mathcal{C}_0, \mathcal{A}_1, \mathcal{M}_1, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{array}$$

avec :

- $\mathcal{C}_0 = \langle C_0, d \rangle$ est l'espace des caractéristiques de $\mathcal{U}_1$ où $c \in C_{0mes}$ $(d(c) = 0)$;
- $\mathcal{A}_1 = \langle A_1, f, O_{A_1} \rangle$, où $A_1 = (A_0 \backslash f(c)) \cup \{a\}$;
- $\mathcal{M}_1 = \langle M_1, g, O_{M_1}, h \rangle$, où $M_1 = M_0 \backslash (\bigcup_{a' \in (f(c) \backslash a)} g(a'))$;
- une cellule $e = \langle adresse, contenu \rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :
  1. l'adresse de e est un vecteur à $n_{dim}-1$ composantes $\langle \beta_1, \ldots, \beta_i, \ldots, \beta_{n_{dim}-1} \rangle$, où $\forall i \in \{1, \ldots, n_{dim}-1\}$, il existe $a' \neq a$ et $a' \in f(c_i)$ $(c_i \neq c)$ tels que $c_i \in C_{0dim}$ et $\beta_i \in g(a')$;
  2. le contenu de e est un vecteur à $n_{mes}+1$ composantes $\langle \gamma_1, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}+1} \rangle$, où $\forall k \in \{1, \ldots, n_{mes}+1\}$, il existe $a' \in (f(c_k) \cup \{a\})$ tels que $c_k \in C_{0mes}$ et $\gamma_k \in g(a') \cup \{\perp\}$.

Conversion d'une mesure

La conversion d'une mesure (PULL) est l'opérateur inverse du PUSH. Il permet de transformer un attribut $a \in f(c)$ d'une mesure $c \in C_{0mes}$ en une dimension dans le cube $\mathcal{U}_0$. Formellement, nous écrivons :

$$\begin{array}{rcl} \mathcal{U}_1 & = & \textsc{Pull}(\mathcal{U}_0, a) \\ & = & \langle \mathcal{C}_0, \mathcal{A}_1, \mathcal{M}_1, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{array}$$

avec :

- $\mathcal{C}_0 = \langle C_0, d\rangle$ est l'espace des caractéristiques de $\mathcal{U}_1$ où $c \in C_{0dim}$ $(d(c) = 1)$;
- $\mathcal{A}_1 = \langle A_1, f, O_{A_1}\rangle$, où $A_1 = A_0 \cup f(c)$;
- $\mathcal{M}_1 = \langle M_1, g, O_{M_1}, h\rangle$, où $M_1 = M_0 \cup (\bigcup_{a' \in f(c)} g(a'))$;
- une cellule $e = \langle adresse, contenu\rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :
  1. l'adresse de e est un vecteur à $n_{dim}+1$ composantes $\langle \beta_1, \ldots, \beta_i, \ldots, \beta_{n_{dim}+1}\rangle$, où $\forall i \in \{1, \ldots, n_{dim}+1\}$, il existe $a' \in (f(c_i) \cup f(c))$ tels que $c_i \in C_{0dim}$ et $\beta_i \in g(a')$;
  2. le contenu de e est un vecteur à $n_{mes}-1$ composantes $\langle \gamma_1, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}-1}\rangle$, où $\forall k \in \{1, \ldots, n_{mes}-1\}$, il existe $a' \in (f(c_k)\backslash\{a\})$ tels que $c_k \in C_{0mes}$ et $\gamma_k \in g(a') \cup \{\perp\}$.

7.4.2 Opérateurs de navigation

Cette seconde famille d'opérateurs est dédiée à la navigation dans un cube de données. Elle offre à l'utilisateur des outils pour la manipulation du contenu d'un cube afin d'en extraire des informations intéressantes. Ces opérateurs permettent particulièrement d'explorer les niveaux de granularités des données, d'observer les modalités des dimensions et de calculer des agrégats dans un cube de données. Nous avons défini cinq opérateurs de navigation : le *forage vers haut* (RollUp), le *forage vers le bas* (DrillDown), la *permutation de modalités* (Switch), la *sélection de modalités* (Select) et le *calcul d'agrégats* (Aggregate). Nous présentons, dans la suite, le rôles et le formalisme de chacun de ces opérateurs.

Forage vers le haut

Dans une dimension $c \in C_{0dim}$ du cube $\mathcal{U}_0$, le forage vers le haut (RollUp) permet de passer du niveau hiérarchique en *cours* $a_{old} \in f(c)$ à un niveau *supérieur* $a_{new} \in f(c) \cup \{All\}$ (*All* est l'agrégat total de la caractéristique c), tel que $a_{old} \preceq_c^{A_0} a_{new}$. C'est-à-dire, a_{old} et a_{new} appartiennent à la même hiérarchie dans la dimension c et a_{old} est plus petit que a_{new} selon l'ordre hiérarchique $o_c^{A_0}$. Une fonction d'agrégation $\xi \in \mathcal{F}_0$ permet également de calculer les nouvelles valeurs des mesures. Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \textsc{RollUp}(\mathcal{U}_0, c, a_{old}, a_{new}, \xi) \\ &= \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_1, \mathcal{F}_0\rangle \end{aligned}$$

En supposant que la caractéristique c correspond à la première caractéristique dimensionnelle $(c = c_1)$ dans le cube $\mathcal{U}_0$, une cellule $e = \langle adresse, contenu\rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :

1. l'adresse de e est un vecteur à n_{dim} composantes $\langle \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}}\rangle$, où (i) $\beta_1 \in g(a_{new})$ et (ii) $\forall i \in \{2, \ldots, n_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in (C_{0dim}\backslash\{c\})$ et $\beta_i \in g(a)$;

2. le contenu de e est un vecteur à n_{mes} composantes $\langle \Gamma_1, \ldots, \Gamma_k, \ldots, \Gamma_{n_{mes}} \rangle = \xi(E)$, où E est un ensemble de cellules dans $\mathcal{L}_0$ dont les adresses, dans le cube $\mathcal{U}_0$, sont de la forme $\langle b, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$ tels que $b \in g(a_{old})$ et $b \sqsubseteq \beta_1$. C'est-à-dire, le contenu d'une nouvelle cellule e dans $\mathcal{L}_1$ est le résultat de la fonction d'agrégation ξ sur un ensemble de cellules E. Ces cellules correspondent à celles dans $\mathcal{L}_0$ qui, pour la dimension c concernée par le forage, *avaient* des modalités b appartenant à la modalité β_1 (selon un ordre sémantique) et les mêmes modalités que e pour les autres dimensions.

Forage vers le bas

Inversement à l'opération de forage vers le haut, le forage vers le bas (DrillDown) permet de passer, pour une dimension $c \in C_{0dim}$ du cube $\mathcal{U}_0$, du niveau hiérarchique en *cours* a_{old} à un niveau *inférieur* a_{new}, tel que $a_{new} \preceq_c^{A_0} a_{old}$ et $a_{old} \in (f(c) \cup \{All\}) \backslash f(c)^\star$ (All est l'agrégat total de la caractéristique c). C'est-à-dire, a_{new} et a_{old} appartiennent à la même hiérarchie dans la dimension c, a_{old} n'est pas un plus petit attribut de c et a_{new} est plus petit que a_{old} selon l'ordre hiérarchique $o_c^{A_0}$. Les nouvelles mesures des cellules sont restituées à des valeurs mesurant des faits ayant une granularité d'information plus fine. Formellement, nous écrivons :

$$\begin{array}{rcl} \mathcal{U}_1 & = & \textsc{DrillDown}(\mathcal{U}_0, c, a_{old}, a_{new}) \\ & = & \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{array}$$

En supposant que la caractéristique c correspond à la première caractéristique dimensionnelle ($c = c_1$) dans le cube $\mathcal{U}_0$, une cellule $e = \langle adresse, contenu \rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :

1. l'adresse de e est un vecteur à n_{dim} composantes $\langle \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$, où (i) $\beta_1 \in g(a_{new})$ et (ii) $\forall i \in \{2, \ldots, n_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in (C_{0dim} \backslash \{c\})$ et $\beta_i \in g(a)$;
2. le contenu de e est un vecteur à n_{mes} composantes $\langle \gamma_1, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}} \rangle$. Soit un ensemble E de cellules ayant les mêmes propriétés de la cellule e dans $\mathcal{L}_1$. C'est-à-dire, leurs adresses est de la forme $\langle \beta, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$, où $\beta \in g(a_{new})$. Il existe donc une fonction d'agrégation $\xi \in \mathcal{F}_0$ selon une fonction $\bigodot$ et une modalité $b \in g(a_{old})$ tel que $\beta \sqsubseteq b$ tel que la cellule de $\mathcal{L}_0$ dont l'adresse est $\langle b, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$ est associée à un contenu $\langle \Gamma_1, \ldots, \Gamma_k, \ldots, \Gamma_{n_{mes}} \rangle = \xi(E)$.

Permutation de modalités

Le classement des modalités (Switch) permet d'intervertir deux modalités b et b' d'un attribut $a \in f(c)$ d'une dimension $c \in C_{0dim}$ du cube $\mathcal{U}_0$, tel que $b' \prec_a^{M_0} b$. C'est-à-dire, la modalité b' précède la modalité b selon l'ordre total $o_a^{M_0}$. Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \textsc{Switch}(\mathcal{U}_0, a, b, b') \\ &= \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_0, \mathcal{F}_0 \rangle \end{aligned}$$

Dans le nouveau cube $\mathcal{U}_1$, l'ordonnancement de ces modalités selon $o_a^{M_0}$ est changé et on a plutôt $b \prec_a^{M_0} b'$.

Sélection de modalités

Cet opérateur (SELECT) consiste à sélectionner, dans le cube $\mathcal{U}_0$, l'ensemble des modalités d'une caractéristique $c \in C_0$ qui satisfont une condition P formée de conjonctions de prédicats logiques sur un certain nombre d'attributs de la caractéristique c. Notons que la sélection est valable aussi bien pour les modalités d'une caractéristique dimensionnelle ($d(c) = 1$) que pour les modalités d'une caractéristique métrique ($d(c) = 0$). Formellement, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \textsc{Select}(\mathcal{U}_0, c, P) \\ &= \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{aligned}$$

Dans le cas où c est une caractéristique dimensionnelle, supposons que cette dernière est la première caractéristique dans C_{0dim}. Une cellule $e = \langle adresse, contenu \rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :

1. l'adresse de e est un vecteur à n_{dim} composantes $\langle \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$, où (i) β_1 satisfait la condition P et (ii) $\forall i \in \{2, \ldots, n_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in C_{0dim}$ et $\beta_i \in g(a)$;
2. le contenu est un vecteur à n_{mes} composantes $\langle \gamma_1, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}} \rangle$, où $\forall k \in \{1, \ldots, n_{mes}\}$, il existe $a \in f(c_k)$ tels que $c_k \in C_{0mes}$ et $\gamma_k \in g(a) \cup \{\perp\}$.

Dans le cas où c est une caractéristique métrique, supposons que cette dernière est la première caractéristique dans C_{0mes}. Une cellule $e = \langle adresse, contenu \rangle$ de $\mathcal{L}_1$ vérifie les deux propriétés suivantes :

1. l'adresse de e est un vecteur à n_{dim} composantes $\langle \beta_1, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$, où $\forall i \in \{1, \ldots, n_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in C_{0dim}$ et $\beta_i \in g(a)$;
2. le contenu est un vecteur à n_{mes} composantes $\langle \gamma_1, \gamma_2, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}} \rangle$, où (i) γ_1 satisfait la condition P et (ii) $\forall k \in \{2, \ldots, n_{mes}\}$, il existe $a \in f(c_k)$ tels que $c_k \in C_{0mes}$ et $\gamma_k \in g(a) \cup \{\perp\}$.

Calcul d'agrégats

Le calcul d'agrégat (AGGREGATE) permet de construire de nouveaux agrégats dans le cube $\mathcal{U}_0$. Pour un attribut a d'une dimension c de $\mathcal{U}_0$, on considère un un sous-ensemble de modalités $B \subset g(a)$. L'opérateur AGGREGATE consiste à agréger

des faits en regroupant les modalités du sous-ensemble B dans une nouvelle modalité *artificielle*, notée agg_B. Les nouvelles mesures agrégées sont calculées selon une fonction d'agrégation $\xi \in \mathcal{F}_0$. Formellement, nous écrivons :

$$\begin{array}{rcl} \mathcal{U}_1 & = & \textsc{Aggregate}(\mathcal{U}_0, c, B, \xi) \\ & = & \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_1, \mathcal{L}_1, \mathcal{F}_0 \rangle \end{array}$$

avec :

- $\mathcal{M}_1 = \langle M_1, g, O_{M_1}, h \rangle$, où $M_1 = M_0 \cup \{agg_B\}$;
- en supposant que la caractéristique c correspond à la première caractéristique dimensionnelle ($c = c_1$) dans le cube $\mathcal{U}_0$, une cellule $e = \langle adresse, contenu \rangle$ de $\mathcal{L}_1$ vérifie les propriétés suivantes :
  1. l'adresse de e est un vecteur à n_{dim} composantes $\langle \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$, où (i) $\beta_1 \in (f(c) \backslash B) \cup \{agg_B\}$ et (ii) $\forall i \in \{2, \ldots, n_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in (C_{0dim} \backslash \{c\})$ et $\beta_i \in g(a)$;
  2. si $\beta_1 \in (f(c) \backslash B)$, alors le contenu de e est un vecteur à n_{mes} composantes $\langle \gamma_1, \ldots, \gamma_k, \ldots, \gamma_{n_{mes}} \rangle$, où $\forall k \in \{1, \ldots, n_{mes}\}$, il existe $a \in f(c_k)$ tels que $c_k \in C_{0mes}$ et $\gamma_k \in g(a) \cup \{\perp\}$;
  3. si $\beta_1 = agg_B$, alors le contenu de e est un vecteur à n_{mes} composantes $\langle \Gamma_1, \ldots, \Gamma_k, \ldots, \Gamma_{n_{mes}} \rangle = \xi(E)$, où E est un ensemble de cellules dans $\mathcal{L}_0$ dont les adresses, dans le cube $\mathcal{U}_0$, sont de la forme $\langle b, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}} \rangle$ tel que $b \in B$.

7.5 Vers une extension à la fouille de données en ligne

Malgré l'absence d'un dispositif théorique commun, nous avons vu que plusieurs algèbres OLAP existent. Ces dernières formalisent, tant bien que mal, les mécanismes nécessaires pour la manipulation des données multidimensionnelles.

Nous avons également vu que le modèle multidimensionnel et l'algèbre que nous proposons se distinguent par rapport aux autres propositions par un certain nombre de propriétés intéressantes. Notamment, ils incluent plusieurs hiérarchies par dimension, ils prennent en compte les granularités des données, ils manipulent d'une manière symétrique les dimensions et les mesures, ils intègrent des fonctions d'agrégation, etc.

Cependant, jusqu'à maintenant, notre algèbre établit *un cadre formel de plus* pour l'analyse en ligne en proposant un noyau minimal d'opérateurs OLAP afin de répondre à des besoins classiques de structuration et de navigation dans les cubes de données.

Rappelons que, dans le cadre de notre problématique de base, notre objectif n'est pas de formaliser une nouvelle algèbre OLAP. Nous cherchons plutôt à établir un cadre théorique général pour le couplage entre l'analyse en ligne et la fouille de données.

D'après nos différentes expériences, nous avons prouvé qu'il est possible d'enrichir l'analyse en ligne en l'associant à la fouille de données. En effet, classiquement, l'analyse en ligne se limite à des tâches de structuration, de visualisation et de navigation dans les données. Avec des techniques de fouille de données, nous avons apporté des extensions à l'analyse en ligne portant sur des capacités de *description*, de *classification* et d'*explication*. Ces extensions nous ont aussi permis d'aborder le problème de l'analyse des données complexes.

À l'image de ces extensions, nous pensons que, sur un plan formel, il est aussi possible d'étendre une algèbre OLAP en vue de formaliser une nouvelle famille d'opérateurs basés sur le couplage entre l'analyse en ligne et la fouille de données. Ainsi, notre modèle multidimensionnel et notre algèbre OLAP sont plutôt perçus comme une base d'outils théoriques qui serviront à la mise en place d'un *cadre formel général* pour des opérateurs de *fouille de données en ligne*.

À ce stade de nos travaux, nous proposons une première tentative de formalisation théorique concernant nos approches de couplage entre l'analyse en ligne et la fouille de données. Pour y parvenir, nous nous basons sur notre modèle multidimensionnel et nous exploitons des combinaisons d'opérateurs de notre algèbre. Dans la suite, nous exposons la formalisation de deux opérateurs :

1. le premier opérateur, appelé OrCA (**O**perator for Reo**r**ganization by Multiple **C**orrespondence **A**nalysis), est dédié à la réorganisation d'un cube de données par une analyse des correspondances multiples (ACM). Il correspond à notre première approche de couplage développée dans le chapitre 3 ;
2. le deuxième opérateur, appelé OpAC (**Op**erator for **A**ggregation by **C**lustering), concerne l'agrégation par une classification ascendante hiérarchique (CAH) dans un cube de données. Il correspond à notre deuxième approche de couplage présentée dans le chapitre 4.

7.5.1 Réorganisation par une ACM (OrCA)

Principe

L'opérateur OrCA permet d'améliorer la représentation des faits d'un cube de données. Il repose sur l'arrangement des modalités des dimensions du cube en fonction de résultats fournis par une ACM. Comme nous l'avons formulé dans le chapitre 3, cet arrangement établit, pour chaque attribut, un ordonnancement de ses modalités selon un nouvel ordre fourni par les résultats de l'ACM.

Afin de formaliser notre opérateur OrCA, nous introduisons, en préliminaire, un

nouvel opérateur d'*ordonnancement de modalités*.

Ordonnancement de modalités

Cet opérateur (ManySwitches) permet d'arranger, selon un ordre total donné, un sous-ensemble de modalités associées à un attribut dimensionnel. Il s'agit d'une combinaison de plusieurs permutations de modalités (Switch). Soient :

- $\mathcal{U}_0 = \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_0, \mathcal{F}_0 \rangle$ un cube de données ;
- $c \in C_{0dim}$ une caractéristique dimensionnelle du cube $\mathcal{U}_0$ $(d(c) = 1)$;
- $a \in f(c)$ un attribut dimensionnel appartenant à l'ensemble des attributs de la caractéristique c ;
- $B \subseteq g(a)$ un sous-ensemble de modalités de l'attribut a ;
- o^B un ordre total dans l'ensemble des modalité B différent de l'ordre total $o_a^{M_0}$ initialement établi dans l'ensemble des modalités $g(a)$.

L'opérateur ManySwitches fournit un nouvel ordonnancement des modalités du sous-ensemble B dans la représentation de l'attribut a. Ainsi, dans le nouveau cube $\mathcal{U}_1$, l'ordonnancement des modalités de B selon $o_a^{M_0}$ sera remplacé par un nouvel ordonnancement selon o^B. Formellement, nous écrivons :

$$\mathcal{U}_1 \quad = \quad \textsc{ManySwitches}(\mathcal{U}_0, a, B, o^B)$$

Formalisation

Pour formaliser l'opérateur de réorganisation par une ACM, nous considérons les éléments suivants :

- $\mathcal{U}_0 = \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_0, \mathcal{F}_0 \rangle$ un cube de données ;
- $\forall i \in \{1, \ldots, n_{dim}\}$, on considère un attribut $a_i \in f(c_i)$, où les c_i sont les caractéristiques dimensionnelles du cube $\mathcal{U}_0$. Les modalités $g(a_i)$ de chaque attribut a_i sont initialement organisées selon l'ordre total $o_{a_i}^{M_0}$ dans le cube $\mathcal{U}_0$;
- X un tableau disjonctif complet extrait à partir du cube $\mathcal{U}_0$. Comme nous l'avons défini dans la section 3.5, selon un codage binaire, X représente les faits (cellules) $\mathcal{L}_0$ en fonction des modalités $g(a_i)$ de tous les attributs a_i $(i \in \{1, \ldots, n_{dim}\})$. Une ACM est appliquée sur le tableau X ;
- $\forall i \in \{1, \ldots, n_{dim}\}$, $o_{a_i}^{ACM}$ est le nouvel ordre total fourni par l'ACM pour l'ensemble des modalités $g(a_i)$. Cet ordre est établi soit selon les projections des modalités (voir section 3.6.1), soit selon leurs valeurs-test (voir section 3.6.2) ;
- $O_{ACM} = \{o_{a_1}^{ACM}, \ldots, o_{a_{n_{dim}}}^{ACM}\}$ est l'ensemble fini et non vide de n_{dim} ordres totaux fournis par l'ACM pour les attributs a_i.

L'opérateur OrCA consiste à fournir, à partir de $\mathcal{U}_0$, un cube de données avec une nouvelle représentation. Dans cette représentation, les modalités $g(a_i)$, de chaque

attribut a_i ($i \in \{1, \ldots, n_{dim}\}$), ne sont plus arrangées selon l'ordre $o_{a_i}^{M_0}$, mais plutôt selon l'ordre $o_{a_i}^{ACM}$ fourni par l'ACM. On dit que le nouveau cube est arrangé selon l'ensemble des ordres totaux O_{ACM}.

D'une manière concrète, notre opérateur OrCA est une combinaison de n_{dim} opérations d'ordonnancement de modalités (ManySwitches) successives. Chacune de ces dernières arrange, dans le cube, l'ensemble des modalités $g(a_i)$ d'un attribut a_i selon l'ordre total $o_{a_i}^{ACM}$.

Par exemple, dans le cas d'un cube de données $\mathcal{U}_0$ à deux dimensions ($n_{dim} = 2$), l'opération de réorganisation du cube $\mathcal{U}_0$ par l'ACM est une combinaison de deux opérations d'ordonnancement de modalités. On peut alors écrire :

$$\begin{aligned} \mathcal{U}_2 &= \text{OrCA}(\mathcal{U}_0, O_{ACM}) \\ &= \text{ManySwitches}(\text{ManySwitches}(\mathcal{U}_0, a_1, g(a_1), o_{a_1}^{ACM}), a_2, g(a_2), o_{a_2}^{ACM}) \end{aligned}$$

Formellement, dans le cas général d'un cube à n_{dim} dimensions, nous écrivons :

$$\begin{aligned} \mathcal{U}_1 &= \text{ManySwitches}(\mathcal{U}_0, a_1, g(a_1), o_{a_1}^{ACM}) \\ \mathcal{U}_2 &= \text{ManySwitches}(\mathcal{U}_1, a_2, g(a_2), o_{a_2}^{ACM}) \\ \mathcal{U}_3 &= \text{ManySwitches}(\mathcal{U}_2, a_3, g(a_3), o_{a_3}^{ACM}) \\ \vdots & \quad \vdots \\ \mathcal{U}_{n_{dim}} &= \text{ManySwitches}(\mathcal{U}_{n_{dim}-1}, a_{n_{dim}}, g(a_{n_{dim}}), o_{a_{n_{dim}}}^{ACM}) \end{aligned}$$

$$\begin{aligned} \mathcal{U}_{n_{dim}} &= \text{OrCA}(\mathcal{U}_0, O_{ACM}) \\ &= \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_{n_{dim}}, \mathcal{L}_0, \mathcal{F}_0 \rangle \end{aligned}$$

Notons que le nouvel espace des modalités, noté $\mathcal{M}_{n_{dim}} = \langle M, g, O_{ACM}, h \rangle$, n'est plus associé à l'ensemble des ordres totaux O_{M_0}, mais plutôt à ceux de O_{ACM} fournis par l'ACM.

7.5.2 Agrégation par une CAH (OpAC)

Principe

L'opérateur OpAC consiste à créer des agrégats dans un cube en regroupant les modalités d'un niveau hiérarchique d'une dimension. Comme nous l'avons exposé dans le chapitre 4, les groupes des modalités correspondent à des classes fournies par une CAH.

Formalisation

Pour formaliser l'opérateur d'agrégation par une CAH, nous considérons les

éléments suivants :

- $\mathcal{U}_0 = \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_0, \mathcal{L}_0, \mathcal{F}_0 \rangle$ un cube de données ;
- $c \in C_{0dim}$ une caractéristique dimensionnelle du cube $\mathcal{U}_0$ $(d(c) = 1)$;
- $a \in f(c)$ un attribut dimensionnel appartenant à l'ensemble des attributs de la caractéristique c ;
- $\Omega = g(a)$ l'ensemble des modalités de l'attribut a représentant les individus de la classification ;
- Σ un ensemble fini et non vide comprenant les variables de la classification extraites à partir du cube $\mathcal{U}_0$ selon le formalisme que nous avons développé dans la section 4.3 ;
- une CAH est appliquée sur l'ensemble des individus Ω selon les variables Σ ;
- $\mathcal{P}$ une partition de x classes $\{agg_1, \ldots, agg_q, \ldots, agg_x\}$ dans l'ensemble des modalité $g(a)$ retenue à partir des résultats de la CAH ;
- $\xi \in \mathcal{F}_0$ est une fonction d'agrégation.

Notre opérateur d'agrégation par la CAH permet de générer des faits agrégés en regroupant les modalités de l'attribut a selon la partition $\mathcal{P}$. En d'autres termes, l'opérateur OpAC crée, pour chaque classe de modalités $agg_q \in \mathcal{P}$ $(q \in \{1, \ldots, x\})$, une nouvelle modalité *artificielle* dans la dimension c. Ainsi, l'agrégation par la CAH est une combinaison de x opérations d'agrégation (Aggregate) successives.

Par exemple, pour une partition $\mathcal{P}$ à deux classes $\{agg_1, agg_2\}$ $(x = 2)$, l'opération d'agrégation par la CAH des modalités $g(a)$, selon la partition $\mathcal{P}$, est une combinaison de deux opérations d'agrégation. La première concerne la création de l'agrégat agg_1 et la seconde concerne la création de l'agrégat agg_2. Ces deux opérations engendrent un nouvel ensemble de cellules dont les mesures sont calculées selon la fonction d'agrégation ξ. On peut alors écrire :

$$\begin{array}{rcl} \mathcal{U}_2 & = & \textsc{OpAC}(\mathcal{U}_0, c, \mathcal{P}, \xi) \\ & = & \textsc{Aggregate}(\textsc{Aggregate}(\mathcal{U}_0, c, agg_1, \xi), c, agg_2, \xi) \\ & = & \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_2, \mathcal{L}_2, \mathcal{F}_0 \rangle \end{array}$$

Dans le cas général, pour une partition $\mathcal{P}$ à x classes, nous écrivons :

$$\begin{array}{rcl} \mathcal{U}_1 & = & \textsc{Aggregate}(\mathcal{U}_0, c, agg_1, \xi) \\ \mathcal{U}_2 & = & \textsc{Aggregate}(\mathcal{U}_1, c, agg_2, \xi) \\ \mathcal{U}_3 & = & \textsc{Aggregate}(\mathcal{U}_2, c, agg_3, \xi) \\ \vdots & & \vdots \\ \mathcal{U}_x & = & \textsc{Aggregate}(\mathcal{U}_{x-1}, c, agg_x, \xi) \end{array}$$

$$\begin{array}{rcl} \mathcal{U}_x & = & \textsc{OpAC}(\mathcal{U}_0, c, \mathcal{P}, \xi) \\ & = & \langle \mathcal{C}_0, \mathcal{A}_0, \mathcal{M}_x, \mathcal{L}_x, \mathcal{F}_0 \rangle \end{array}$$

Le nouveau cube $\mathcal{U}_x$, obtenu suite à l'opération OPAC, est caractérisé par un nouvel espace de modalités $\mathcal{M}_x = \langle M_x, g, O_{M_x}, h\rangle$, où $M_x = M_0 \cup \{agg_1, \ldots, agg_x\}$.

Le cube $\mathcal{U}_x$ est aussi caractérisé par un nouvel ensemble de cellules $\mathcal{L}_x$. En supposant que la caractéristique c correspond à la première caractéristique dimensionnelle ($c = c_1$) dans le cube $\mathcal{U}_0$, une cellule $e = \langle adresse, contenu\rangle$ de $\mathcal{L}_x$ vérifie les deux propriétés suivantes :

1. l'adresse de e est un vecteur à n_{dim} composantes $\langle \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}}\rangle$, où (i) $\beta_1 \in \{agg_1, \ldots, agg_x\}$ et (ii) $\forall i \in \{2, \ldots, n_{dim}\}$, il existe $a \in f(c_i)$ tels que $c_i \in (C_{0dim} \backslash \{c\})$ et $\beta_i \in g(a)$;
2. le contenu de e est un vecteur à n_{mes} composantes $\langle \Gamma_1, \ldots, \Gamma_k, \ldots, \Gamma_{n_{mes}}\rangle = \xi(E)$, où E est un ensemble de cellules dans $\mathcal{L}_0$ dont les adresses, dans le cube $\mathcal{U}_0$, sont de la forme $\langle b, \beta_2, \ldots, \beta_i, \ldots, \beta_{n_{dim}}\rangle$ tel que $b \in agg_q$ ($q \in \{1, \ldots, x\}$).

7.6 Conclusion et perspectives

Suite à l'ensemble de nos contributions, nous sommes amenés à mettre en place des fondements théoriques capables de définir un cadre formel général pour le couplage entre l'analyse en ligne et la fouille de données. Pour cela, nous pensons qu'il est possible d'étendre les opérateurs classiques d'une algèbre OLAP à une nouvelle génération d'opérateurs d'analyse en ligne basés sur la fouille de données.

Dans ce chapitre, nous avons mis en place les premières bases d'un tel cadre formel. Pour cela, nous avons défini un modèle de données multidimensionnelles et une algèbre OLAP. Ces derniers manipulent d'une manière symétrique les dimensions et les mesures d'un cube, prennent en compte l'aspect hiérarchiques des données, intègrent des fonctions d'agrégation et considèrent les niveaux granulaires des données. En plus, notre algèbre repose sur un noyau minimal fermé d'opérateurs OLAP dédiés à la structuration et la navigation. Ces opérateurs permettent de créer un cube de données, de manipuler sa structure et d'explorer son contenu selon plusieurs niveaux de granularité afin de répondre à toute sorte de besoins d'analyse en ligne possibles.

En se basant sur nos différentes expériences de couplage entre l'analyse en ligne et la fouille de données, nous avons proposé une première tentative pour étendre notre algèbre à des opérateurs de fouille de données en ligne. Cette extension a fait l'objet d'une formalisation de deux nouveaux opérateurs : ORCA et OPAC. Le premier est un opérateur de réorganisation d'un cube de données par une ACM et le second est un opérateur d'agrégation dans un cube de données par une CAH. Ces deux opérateurs reposent sur la formalisation des résultats des techniques de fouille (ACM et CAH) et la combinaison d'opérateurs classiques de notre algèbre OLAP.

Rappelons que ces deux opérateurs représentent les premiers résultats de nos travaux actuels concernant la mise en place d'un cadre formel général pour le couplage

de l'analyse en ligne et de la fouille de données. Dans la suite de ces travaux, plusieurs perspectives méritent d'être étudiées.

Tout d'abord, en complément de nos résultats actuels, nous avons dore et déjà engagé des réflexions pour la formalisation d'un troisième opérateur, appelé AROx (**A**ssociation **R**ules **O**perator for E**x**plication), dédié à l'explication dans un cube de données. Ce dernier se base sur notre approche d'extraction des règles d'association à partir des cubes de données.

À l'image de nos trois propositions de couplage entre l'analyse en ligne et la fouille de données, nous pensons que, dans le cadre formel que nous allons établir, il serait nécessaire de considérer trois familles d'opérateurs de fouille de données en ligne : (i) des opérateurs pour la visualisation et la description, tel que OrCA ; (ii) des opérateurs pour la structuration et la classification, tel que OpAC et (iii) des opérateurs pour l'explication et la prédiction, tel que AROx. Nous souhaitons aussi généraliser les formalismes de chaque famille d'opérateurs de fouille de données en ligne. Ainsi, notre cadre formel peut jouer le rôle d'une base théorique générale pour toute sorte d'analyse combinant l'OLAP avec la fouille de données.

Enfin, nous projetons la validation de ce cadre formel par une implémentation à l'aide d'un langage approprié aux algèbres OLAP et à la fouille de données. Cette implémentation pourrait faire l'objet d'un nouveau module dans notre plateforme MiningCubes.