RAIRO. RECHERCHE OPÉRATIONNELLE

LOTFI LAKHAL
ROSINE CICCHETTI

STAR$^+$, un langage de manipulation de résumés statistiques structurés

RAIRO. Recherche opérationnelle, tome 24, n° 4 (1990), p. 365-432

<http://www.numdam.org/item?id=RO_1990__24_4_365_0>

Recherche opérationnelle/Operations Research
(vol. 24, n° 4, 1990, p. 365 à 432)

STAR$^+$, UN LANGAGE DE MANIPULATION DE RÉSUMÉS STATISTIQUES STRUCTURES (*)

par Lotfi LAKHAL ([1]) et Rosine CICCHETTI ([1, 2])

Résumé. – *Cet article présente le langage de résumés statistiques STAR$^+$ qui offre de nouvelles capacités de représentation, de création et de manipulation de résumés statistiques, calculés à partir de données détaillées ou d'autres résumés.*

Mots clés : Bases de données statistiques; langages de résumés statistiques; tables statistiques complexes.

Abstract. – *This paper presents the statistical summary language STAR$^+$, allowing new abilities to represent, create and manipulate statistical summaries, computed from raw data or other summaries.*

Keywords : Statistical databases; statistical summary languages; complex statistical tables.

1. INTRODUCTION ET MOTIVATIONS

1.1. Sémantique des résumés statistiques

Les *résumés statistiques* [ou macro-données statistiques (Wong [34])] peuvent être perçus comme des ensembles de données, proposant une vision synthétique d'un volume, généralement beaucoup plus important, de données factuelles, par opposition dites détaillées [ou micro-données statistiques (Wong [34])]. Au cours de leur *dérivation*, à partir de micro-données, deux opérations distinctes peuvent être mises en évidence : la classification des

(*) Reçu juin 1990.
([1]) Université de Nice-Sophia Antipolis, C.N.R.S., I3S, U.R.A., n° 1376, Lisan, Bât. n° 4, rue A.-Einstein, Sophia-Antipolis, 06560 Valbonne.
([2]) Centre d'Enseignement et de Recherche appliqués au Management, C.E.R.A.M., B.P. n° 120, Sophia-Antipolis, 06561 Valbonne Cedex.

Recherche opérationnelle/Operations Research, 0399-0559/90/04 365 68/$ 8.80

individus de la population sous-jacente puis le calcul de fonctions statistico-mathématiques sur la classification introduite (Malvestuto [22]). Une valeur résumée est ainsi obtenue pour chacune des classes établies, qui forment une partition de l'ensemble des individus concernés.

Conceptuellement, les résumés se caractérisent par deux types d'attributs, les *attributs catégories* et les *attributs résumés* (Shoshani *et al.* [30, 31], Rafanelli *et al.* [29], Fortunato *et al.* [10]). Les attributs catégories constituent les critères de classification. Un attribut catégorie provient généralement d'un attribut existant, *i.e.* d'un caractère considéré comme un critère lors de la classification des individus, mais il peut aussi être un nouvel attribut, dépendant fonctionnellement d'attributs décrivant la population étudiée (Malvestuto [22]). De par leur caractère descriptif ou qualitatif, les attributs catégories sont très fréquemment de type alphanumérique et ont généralement des ensembles de valeurs admissibles, appelées modalités, assez restreints (Turner *et al.* [33]). Les attributs résumés sont de type numérique puisque leurs valeurs sont obtenues par calcul de fonctions statistico-mathématiques. Une caractéristique importante des résumés est la densité de leurs données. Contrairement aux micro-données, qui typiquement contiennent un grand nombre de valeurs nulles, dont l'existence est souvent liée aux nombreuses manières de collecter les données brutes, les résumés véhiculent une information synthétique et pertinente. Aussi, l'idée largement répandue est d'assimiler la classification d'un résumé au *produit cartésien total* des sous-ensembles de valeurs de ses différents attributs catégories (Olken *et al.* [25], Shoshani *et al.* [31], Fortunato *et al.* [10]). Chaque combinaison de modalités détermine la valeur d'un attribut résumé et synthétise une classe de la population sous-jacente. Les résumés sont très fréquemment appelés *données multidimensionnelles* (Özsoyoglu *et al.* [27], Shoshani *et al.* [31], Olken *et al.* [25], Adam *et al.* [4]). Cette expression fait référence à leur représentation dans un espace de dimension n, où chaque critère de classification, *i.e.* chaque attribut catégorie, correspond à une dimension. La valeur résumée, associée à une combinaison de modalités, est assimilée à un point de cet espace.

Différents degrés de complexité peuvent être observés dans le contenu sémantique des résumés. Certains, que nous qualifions par la suite de *résumés élémentaires*, véhiculent une sémantique simple, correspondant à un unique attribut résumé (Özsoyoglu [28]). D'autres, via différents attributs résumés, ont un contenu sémantique complexe. Nous appelons *résumés complexes*, des ensembles de macro-données comportant différents attributs résumés,

structurés selon la même classification d'attributs catégories, et *résumés arbitrairement complexes*, ceux dont les valeurs des divers attributs résumés sont déterminées par les modalités d'attributs catégories différents. Un résumé complexe est donc doté d'une classification homogène organisant des valeurs résumées de sémantique différente, alors que de telles valeurs sont structurées, dans un résumé arbitrairement complexe, à travers des combinaisons de modalités hétérogènes.

Enfin, élément essentiel de la sémantique des ensembles de macro-données, la nature des attributs résumés rend compte du fait que les valeurs correspondantes traduisent des comptages, moyennes, écarts types, ... Cette notion, capitale lors de l'interprétation des résumés, apporte, dans le cadre de leur manipulation, une information fondamentale sur leur dérivabilité potentielle, *i.e.* leur capacité à produire, par agrégation ou calcul simple, de nouveaux résumés. En effet, suivant la nature de la fonction statistique : *additive* ou *calculable*, utilisée lors de leur création, les résumés pourront ou non être à nouveau dérivés. Les résumés dérivables sont, en fait, limités à ceux engendrés par des fonctions statistiques additives, *i.e.* : comptage et somme. Il s'avère que les principales autres fonctions statistiques peuvent être exprimées en termes de fonctions additives. Elles sont alors appelées fonctions statistiques calculables (Chen *et al.* [8]) et leur résultat est équivalent à celui d'un calcul arithmétique simple sur des résumés créés en utilisant leurs fonctions additives paramètres. Par exemple la moyenne peut être obtenue à partir de comptage et somme, la covariance admet les paramètres : comptage, somme de produits et somme de carrés, l'écart type : somme et somme de carrés; l'analyse de la régression est calculable à partir de comptage, somme, somme de carrés et somme de produits;...

Les fonctions statistiques : minimum et maximum extraient certaines valeurs particulières de caractères quantitatifs de la population étudiée ou d'attributs résumés. Elles peuvent être considérées comme des opérations de sélection.

Selon (Özsoyoglu *et al.* [27]), « une des fonctions essentielles des Systèmes de Gestion de Bases de Données Statistiques-SGBDS est *la création*, *la maintenance* et *la manipulation* des résumés, obtenus à partir de données détaillées ou d'autres résumés ». (Adam *et al.* [4]) ont même restreint la définition d'un SGBDS à cette unique fonctionnalité. Ces deux définitions résultent du fait que contrairement aux bases de données classiques, les requêtes statistiques visent non pas la recherche de données ponctuelles (telle que : « quels sont les noms des étudiants qui habitent à Nice? ») mais plutôt celle de données agrégées (telle que : « quel est le nombre des étudiants Niçois

par sexe, par discipline et par année d'étude?) (Q), ou encore « quel est le revenu moyen des Français par région, par sexe et par catégorie socio-professionnelle?, . . .) (Wong [34], Ghosh [11], Chen *et al.* [8], Adam *et al.* [4], Mcleish [24]). Les résumés résultats de ce type d'opération peuvent non seulement être calculés à partir de données détaillées, mais aussi, et ce de manière beaucoup plus performante, à partir d'autres résumés physiquement stockés dans la base (Ghosh [11], Hebrail [14], Chen [8]). Par exemple, si l'utilisateur s'intéresse au « nombre d'étudiants Niçois par sexe et par discipline » ou bien au « nombre d'étudiants Niçois de sexe masculin dans la discipline informatique », il semble paradoxal, comme c'est le cas dans la majorité des « packages » statistiques ou dans les SGBD relationnels commercialisés, d'effectuer une requête sur les données détaillées et non pas sur le résumé Q.

Les langages de bases de données existants (Jarke *et al.* [15]) sont définis en tenant compte des finalités de l'utilisation des applications de gestion, notamment la recherche de données détaillées. Celle-ci jouant un rôle mineur dans les applications statistiques, il convient d'expliciter les fonctionnalités attendues d'un langage manipulant des données synthétiques. En effet, nous pensons qu'une phase d'identification des besoins spécifiques doit précéder toute définition d'un tel langage.

Nos propositions font l'objet du paragraphe 1.3 et nous permettent d'introduire une présentation synthétique du langage STAR$^+$. Mais, au préalable, nous décrivons, de manière plutôt intuitive, notre approche de modélisation de bases de données statistiques.

1.2. Le modèle STAR

Le modèle STAR (STAtistical Relational database model) est basé sur une double structure, celle de relation (Delobel *et al.* [9]) et de Table Statistique Complexe-TSC (Lakhal *et al.* [18]). La structure de relation est utilisée pour la représentation des micro-données, celle de TSC pour les résumés statistiques à sémantique arbitrairement complexe. La structure de TSC peut être vue comme une combinaison des structures relationnelle et matricielle. En effet, elle s'inspire, en les adaptant, des concepts d'attribut, de domaine sémantique et de relation pour la représentation des attributs catégories et des classifications et utilise une structure matricielle pour les attributs résumés ou l'agencement de résumés élémentaires au sein d'un résumé complexe. Une des différences notables entre la structure des relations et celle des TSC est la notion d'ordre (par ailleurs inhérente aux matrices) intervenant au niveau de tous les éléments de notre structure. Cet ordre est directement induit par

la classification des critères dans un résumé. Ainsi, le concept de schéma d'attributs catégories, que nous introduisons, peut être vu, de manière intensive et extensive, comme une relation sur laquelle porteraient des contraintes d'ordre, aussi bien sur son schéma, *i.e.* sur ses attributs, que sur son extension, perçue comme un ensemble ordonné de tuples, similaire aux « tuple séquences » de (Abiteboul *et al.* [1]).

Les deux concepts de base structurels, que nous utilisons, sont ceux d'attribut catégorie et d'attribut résumé, qui peuvent réellement être perçus comme des unités d'information de type différent.

L'organisation de ces attributs, pour représenter des résumés arbitrairement complexes, fait appel aux constructeurs *ensemble*, *matrice*, *produit cartésien* et *concaténation*.

Les attributs catégories peuvent être organisés selon un schéma d'attributs catégories (ensemble ordonné d'attributs), dont l'extension, permettant de représenter une classification homogène de modalités, fait appel aux constructeurs *ensemble* et *produit cartésien*. De tels schémas interviennent dans la structuration des dimensions, ligne et colonne, d'une Table Statistique-TS. Une TS permet de représenter des résumés élémentaires, elle est dotée d'un unique attribut résumé, dont les valeurs sont agencées sous forme matricielle. Toute TS est caractérisée par la fonction statistique ayant permis d'en calculer les valeurs résumées. Les TS peuvent être concaténées pour obtenir des TSC, représentant des résumés arbitrairement complexes. Observé au niveau des éléments composants d'une TSC, un tel processus d'organisation met en jeu les constructeurs *ensemble* et *concaténation*, pour définir des multi-schémas d'attributs catégories, de manière intensive et extensive, à partir des schémas d'attributs catégories des TS, constituant la TSC, et de leur extension. A travers deux multi-schémas, ligne et colonne, il est possible de représenter la classification potentiellement complexe et hétérogène d'ensembles de macro-données. Les attributs résumés des différentes TS composants sont organisés sous forme matricielle, dans la TSC. La structuration des valeurs résumées de la TSC s'appuie sur les constructeurs *matrice* et *concaténation*. Les TSC sont non seulement dotées d'un schéma ou intension et d'une extension mais elles sont aussi caractérisées par les fonctions statistiques, ayant permis de calculer leurs attributs résumés, elles aussi organisées sous forme matricielle.

Le schéma d'une TS peut être perçu comme un tableau dans lequel les noms des attributs catégories apparaissent dans les entêtes lignes et colonnes, celui de l'unique attribut résumé correspondant à la cellule de la table. A cette TS est associée la fonction statistique utilisée pour calculer son attribut résumé. Les dimensions ligne et colonne d'une TS sont organisées selon un

schéma d'attributs catégories, défini comme un ensemble ordonné de noms d'attribut catégorie.

Exemple 1 : Les différents résumés, pris en exemple tout au long de cet article, sont supposés dérivés de micro-données décrivant une population estudiantine. Les individus de cette population sont notamment décrits par leur nom (NOM), prénom (PRENOM), sexe (SEXE), milieu socio-professionnel (MILIEU), cycle (CYCLE) et année d'étude dans le cycle (AN_ET), discipline étudiée (DISCIP), le fait qu'ils soient ou non boursier (caractère booléen BOURSIER) et le montant éventuel de leur bourse (BOURSE).

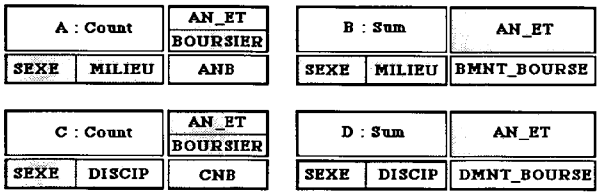

Figure 1. – Schéma des TS *A*, *B*, *C* et *D*.

Les TS A et C, dont les schémas sont illustrés par la figure 1, permettent de dénombrer les étudiants Niçois de deuxième cycle, respectivement :

– par sexe, milieu socio-professionnel, année d'étude et attribution ou non d'une bourse;

– par sexe, discipline, année d'étude et attribution ou non d'une bourse.

Les attributs catégories, dans ces TS, sont : SEXE défini sur l'ensemble $\{F, M\}$, MILIEU prenant les valeurs $\{$ Agriculteurs, Ouvriers, Cadres$\}$, DISCIP ayant pour modalités $\{$ Informatique, Mathématiques, Physique$\}$, AN_ET dont les valeurs sont : $\{1, 2\}$. Ces trois derniers attributs catégories admettent, dans ces TS, des valeurs qui constituent un sous-ensemble restreint de leur domaine initial. Les attributs résumés ANB et CNB de ces TS traduisent des comptages; ils sont calculés, à partir d'un sous-ensemble de la population initiale, en appliquant la fonction Count.

Les attributs résumés BMNT_BOURSE et DMNT_BOURSE des TS B et D sont issus de la sommation des valeurs de l'attribut BOURSE pour les individus considérés; ils donnent le montant total des bourses attribuées aux étudiants Niçois boursiers de deuxième cycle, respectivement :

– par sexe, milieu socio-professionnel et année d'étude;

– par sexe, discipline et année d'étude.

L'extension d'une TS est vue comme un tableau de valeurs des attributs catégories et de l'attribut résumé. Ces dernières sont structurées selon l'extension des schémas d'attributs catégories ligne et colonne définie par le produit cartésien total des ensembles de valeurs de leurs attributs.

Exemple 2 : L'extension de la TS A est présentée par la figure 2.

A : Count		1 O	1 N	2 O	2 N
F	Agriculteurs	15	14	9	12
F	Ouvriers	30	25	23	29
F	Cadres	2	5	1	5
M	Agriculteurs	10	9	3	5
M	Ouvriers	43	40	35	42
M	Cadres	7	8	5	8

Figure 2. – Extension de la TS *A*.

Une TSC permet de représenter des résumés arbitrairement complexes. Elle résulte de la concaténation, selon les lignes ou les colonnes, de différentes TS, organisées sous forme matricielle. La dimension ligne (ou colonne) d'une TSC est dite structurée par un multi-schéma d'attributs catégories (ensemble ordonné de schémas) ligne (ou colonne).

Exemple 3 : Si l'utilisateur désire travailler sur les valeurs des résumés *A*, *B*, *C* et *D*, il peut regrouper les TS, les représentant, au sein de la TSC Z1 dont le schéma est présenté par la figure 3.

Z1		AN_ET	AN_ET
A : Count	B : Sum	BOURSIER	
C : Count	D : Sum		
SEXE	MILIEU	ANB	BMNT_BOURSE
SEXE	DISCIP	CNB	DMNT_BOURSE

Figure 3. – Schéma de la TSC Z1.

L'extension d'une TSC est un tableau de valeurs d'attributs catégories et résumés, obtenu par concaténation, selon les lignes ou les colonnes, des extensions de ses TS composants.

Exemple 4 : L'extension de la TSC Z1 est illustrée par la figure 4.

La présentation tabulaire, utilisée dans les figures précédentes, propose la vision du schéma d'une TSC sous forme d'un tableau (exploitable pour une

Z1		1	1	2	2	1	2
A : Count	B : Sum						
C : Count	D : Sum	O	N	O	N		
F	Agriculteurs	15	14	9	12	23.625	11.700
F	Ouvriers	30	25	23	29	45.000	34.500
F	Cadres	2	5	1	5	1.600	1.000
M	Agriculteurs	10	9	3	5	13.220	4.560
M	Ouvriers	43	40	35	42	60.200	50.750
M	Cadres	7	8	5	8	7.700	5.500
F	Informatique	25	39	24	33	26.250	28.800
F	Mathématiques	18	21	15	20	27.000	17.900
F	Physique	27	14	20	18	37.125	28.200
M	Informatique	38	37	29	42	46.740	41.730
M	Mathématiques	17	29	19	28	23.630	28.880
M	Physique	25	23	25	25	27.250	29.250

Figure 4. – Extension de la TSC $Z1$.

interface conviviale) composé de quatre zones :

– la zone titre permettant l'affichage du nom de la TSC, ainsi que des identifiants des TS composants, sous forme matricielle. A chaque TS est associée la fonction statistique utilisée pour calculer son attribut résumé;

– la zone entête ligne (grisée) permettant de visualiser le multi-schéma d'attributs catégories ligne de la TSC, de telle manière que chaque ligne i décrive, de gauche à droite, le i-ième schéma du multi-schéma d'attributs catégories;

– la zone entête colonne (grisée) permettant, de manière analogue, de présenter le multi-schéma colonne;

– la zone cellule, regroupant les noms des attributs résumés, sous forme matricielle.

Dans la présentation choisie pour l'extension des TSC, les quatre zones du schéma des TSC, sont conservées. Chaque ligne (ou colonne) de la zone entête ligne (ou colonne) décrit, de gauche à droite, une combinaison de modalités, qui constitue une « instance » du multi-schéma d'attributs catégories (ligne ou colonne), les valeurs résumées apparaissant dans les cellules.

Ce format, largement utilisé par la suite pour illustrer les exemples, en raison de sa clarté et de sa concision, propose une vision duale d'une TSC, l'une en terme de TS, l'autre en terme d'attributs, catégories et résumés, clairement distingués.

Pour mieux expliciter les originalités du modèle STAR, il nous semble indispensable d'évoquer les différentes démarches de représentation de résumés, en examinant, dans un premier temps, celles modélisant les résumés au

travers de structures utilisées pour les données classiques puis les approches de structuration spécifiques.

Plusieurs recherches dans le domaine des bases de données statistiques utilisent la structure relationnelle pour la représentation ou le stockage des résumés (Malvestuo [21, 22], Ghosh [12, 13] Chen [8], Özsoyoglu *et al.* [28]). Or cette structure ne semble pas réellement adaptée à leur modélisation. En effet, le modèle relationnel n'établit aucune distinction entre attributs. Les attributs catégories et les attributs résumés, bien que fondamentalement différents, y sont donc représentés de la même manière. De plus, la nature des attributs résumés ne peut être prise en compte dans la définition relationnelle des résumés qui, en outre, ne permet que la représentation de résumés élémentaires ou complexes. Enfin, la dernière contrainte est relative au caractère statique de la structure d'une relation, alors que la modification dynamique de la structure des résumés est soulignée comme un besoin capital, au niveau de leur manipulation (*cf.* paragraphe 1.3).

Les résumés statistiques élémentaires, représentés au sein de la TSCZ1, nécessitent les quatre relations, en 3FN, suivantes :

A_EFFECTIF (SEXE, MILIEU, AN ET, BOURSIER, ANB)

C_EFFECTIF (SEXE, DISCIP, AN ET, BOURSIER, CNB)

B_MNTBOURSE (SEXE, MILIEU, AN ET, BMNT_BOURSE)

D_MNTBOURSE (SEXE, DISCIP, AN ET, DMNT_BOURSE)

Note : les clés primaires sont soulignées.

Nous illustrons, par la figure 5, l'extension de la relation A_EFFECTIF.

SEXE	MILIEU	AN_ET	BOURSIER	ANB
F	Agriculteurs	1	O	15
F	Agriculteurs	1	N	14
F	Agriculteurs	2	O	9
F	Agriculteurs	2	N	12
F	Ouvriers	1	O	30
F	Ouvriers	1	N	25
F	Ouvriers	2	O	23
F	Ouvriers	2	N	29
F	Cadres	1	O	2
F	Cadres	1	N	5
F	Cadres	2	O	1
F	Cadres	2	N	5
M	Agriculteurs	1	O	10
M	Agriculteurs	1	N	9
M	Agriculteurs	2	O	3
M	Agriculteurs	2	N	5
M	Ouvriers	1	O	43
M	Ouvriers	1	N	40
M	Ouvriers	2	O	35
M	Ouvriers	2	N	42
M	Cadres	1	O	7
M	Cadres	1	N	8
M	Cadres	2	O	5
M	Cadres	2	N	8

Figure 5. – Extension de la relation A_EFFECTIF.

Bien qu'aucune recherche sur la modélisation des résumés n'ait tiré partie des travaux sur les modèles relationnels sous non première forme normale-N1NF (Abiteboul *et al.* [2, 3]), il nous semble intéressant d'examiner si ces

nouvelles structures permettent la représentation de résumés statistiques à sémantique arbitrairement complexe. Nous le faisons, de manière intuitive, à travers l'exemple des résumés statistiques représentés par Z1 (*fig.* 4), en utilisant le modèle VERSO (Abiteboul *et al* [2]). Dans ce modèle, deux *V*-relations, seulement, sont nécessaires pour modéliser l'ensemble de macro-données Z1, avantage indéniable par rapport à la représentation relationnelle.

Leurs schémas sont les suivants :

AC_EFFECTIF =
SEXE AN_ET BOURSIER (MILIEU ANB)* (DISCIP CNB)*
BD_MNTBOURSE =
SEXE AN_ET (MILIEU BMNT_BOURSE)* (DISCIP DMNT_BOURSE)*

Nous nous contenons, par la figure 6, d'illustrer l'extension de la *V*-relation AC_EFFECTIF.

SEXE	AN_ET	BOURSIER	(MILIEU	ANB)*	(DISCIP	CNB)*
F	1	O	Agriculteurs	15	Informatique	25
			Ouvriers	30	Mathématiques	18
			Cadres	2	Physique	27
F	1	N	Agriculteurs	14	Informatique	39
			Ouvriers	25	Mathématiques	21
			Cadres	5	Physique	14
F	2	O	Agriculteurs	9	Informatique	24
			Ouvriers	23	Mathématiques	15
			Cadres	1	Physique	20
F	2	N	Agriculteurs	12	Informatique	33
			Ouvriers	29	Mathématiques	20
			Cadres	5	Physique	18
M	1	O	Agriculteurs	10	Informatique	38
			Ouvriers	43	Mathématiques	17
			Cadres	7	Physique	25
M	1	N	Agriculteurs	9	Informatique	37
			Ouvriers	40	Mathématiques	29
			Cadres	8	Physique	23
M	2	O	Agriculteurs	3	Informatique	29
			Ouvriers	35	Mathématiques	19
			Cadres	5	Physique	25
M	2	N	Agriculteurs	5	Informatique	42
			Ouvriers	42	Mathématiques	28
			Cadres	8	Physique	25

Figure 6. — Extension de la *V*-relation AC_EFFECTIF.

Malgré une représentation plus compacte et des possibilités de restructuration à travers l'opérateur original Struct (Abiteboul *et al.* [2]), permettant une réorganisation dynamique de la structure des *V*-relations, l'extension de la

V-relation AC_EFFECTIF montre que les inconvénients de la structuration relationnelle de résumés statistiques sont encore présents. De plus, si les *V*-relations permettent la modélisation de classifications de modalités potentiellement hétérogènes, celle-ci complexifie l'identification (déjà délicate dans la représentation relationnelle) des résumés élémentaires, constitutifs d'un résumé complexe.

Visant avant tout une meilleure adéquation de la structure logique proposée, aux résumés, certains travaux (Ghosh [12], Rafanelli *et al.* [29], Fortunato *et al.* [10]) ont opté pour des structures bi-dimensionnelles (ligne/colonne), similaires (du moins au niveau de leur présentation) aux TS précédemment introduites, appelées Table Relationnelle Statistique, Table Simple ou encore Table, dont les capacités de représentation sont cependant limitées. Or nous pensons que les statisticiens ne peuvent se contenter de manipuler des résumés élémentaires, dont ils ont besoin de rapprocher, comparer, ou combiner les données. Un exemple significatif, illustrant cette nécessité, est de considérer les données d'un résumé élémentaire et les totaux partiels ou globaux issus d'une agrégation de certains ou tous les critères de classification du résumé considéré (de tels totaux sont, pour les statisticiens, les marges du résumé). Les associer aux données dont ils sont dérivés, *i.e.* réunir au sein d'une même structure, un résumé et ses marges, afin de les manipuler comme un tout, nécessite impérativement la capacité de modélisation de résumés arbitrairement complexes. En effet, les valeurs globales sont structurées selon une classification plus grossière que celle du résumé initial, il faut donc se doter, notamment, des moyens de représentation de classifications hétèrogènes. De plus, étant donné qu'une des caractéristiques des bases de données statistiques est la prolifération des résumés (Shoshani [30]), une structure de données permettant la représentation de résumés arbitrairement complexes, peut s'avérer être, si les compositions et les décompositions requises sont proposées, un outil intéressant de regroupement des résumés élémentaires (dont l'interprétation globale et comparative est facilitée), permettant une réduction considérable de l'espace de stockage et une recherche plus aisée de résumés dans la base. A notre connaissance, le seul travail proposant une représentation de résumés à sémantique arbitrairement complexe, appelée Table Résumée-TR, est celui de (Özsoyoglu *et al.* [28]). Les TR proposent une représentation hiérarchique des attributs catégories. Il est important de noter que les TR ne sont pas considérées comme une structure de stockage, mais plutôt comme une interface offrant une présentation agréable des résumés.

Enfin, dans aucune des approches citées, la nature des attributs résumés, essentielle car indiquant leur dérivabilité potentielle, n'est prise en compte de manière structurelle.

1.3. Le langage STAR$^+$

Pour réellement tirer profit des bases de résumés statistiques, il convient de proposer des capacités de manipulation adaptées, aussi une phase d'identification des besoins spécifiques des statisticiens s'impose. Certains de ces besoins ont été intuitivement indiqués, dans (Wong [34], Shoshani *et al.* [31], Olken [26, 25]). Nous les retenons et les complétons à travers la classification générique suivante des différentes opérations nécessaires sur les résumés : Composition/Décomposition, Transposition, Dérivation, Affinement/Élargissement.

Cette classification, que nous détaillons ci-dessous, va nous servir comme support de base pour la définition des opérateurs du langage STAR$^+$ et comme référence pour la comparaison de notre contribution avec d'autres travaux.

Composition/Décomposition

Les possibilités de composition et de décomposition de résumés visent à la production de résumés dont la sémantique est de plus en plus complexe ou, au contraire, de plus en plus simple. Elles permettent notamment de regrouper différents résumés élémentaires calculés à partir de la même population, peuvent aussi représenter l'évolution de cette population dans le temps ou encore facilitent l'interprétation de la comparaison, au travers de résumés, de sous-ensembles distincts d'individus. De plus, elles ont un rôle important pour les calculs ultérieurs.

Transposition

Une des nécessités identifiées pour la manipulation de résumés est d'introduire un aspect dynamique dans la structure de données. Celle-ci, en effet, ne doit pas être figée, mais doit pourvoir évoluer, *i.e.* être réorganisée à la demande de l'utilisateur. Ces réorganisations, appelées opérations de transposition, ont un objectif de présentation, car elles permettent à l'utilisateur d'examiner les macro-données exactement sous l'angle désiré. Mais elles jouent un deuxième rôle essentiel dans le rapprochement de résumés, dont les classifications homogènes sont organisées différemment (un tel rapprochement peut être effectué dans un objectif de comparaison); elles facilitent aussi l'agrégation ou encore la composition des résumés.

Dérivation

La dérivation de nouvelles informations synthétisant des ensembles de macro-données est un besoin fondamental. Ce type de manipulation sous-entend des capacités d'agrégation et de calcul. Les opérations d'agrégation permettent de synthétiser les résumés, elles interviennent donc sur leur niveau de détail. La classification, moins fine que celle du résumé initial dérivable, permet, dans le résumé dérivé, de déterminer de nouvelles valeurs, calculées en utilisant la fonction additive somme. Il est important de souligner que de telles opérations ne « dénaturent » pas les attributs résumés, mais en proposent une vision moins détaillée. Les opérations de calcul, au contraire, visent à la création de nouveaux attributs résumés, à partir d'attributs, de même nature ou pas, mais associés à une même classification, conservée pour le résultat. Il s'agit alors de produire de nouvelles informations à un même niveau de détail. Les opérations de calcul permettent notamment de mener sur les résumés, les investigations statistiques équivalentes à celles envisagées sur les micro-données par utilisation de fonctions statistiques calculables (moyenne, variance, covariance, . . .).

Affinement/Élargissement

L'utilisateur doit aussi pouvoir choisir, dans un résumé, les données qu'il juge pertinentes. Ces manipulations d'affinement et d'élargissement opèrent sur les ensembles de valeurs des attributs et permettent de les restreindre ou de les étendre (à l'intérieur toujours de leur domaine de définition). Elles ont pour vocation l'interrogation des macro-données mais jouent également un rôle dans le rapprochement de résumés dotés de classifications homogènes mais cependant différentes.

Quelques langages de requêtes de bases de données statistiques ont été définis. Les plus significatifs sont QBSRT (Ghosh [11]), STAQUEL (Rafanelli *et al.* [29], Fortunato *et al.* [10]), STBE (Özsoyoglu *et al.* [28]) et STAR (Lakhal *et al.* [19, 20]) manipulant respectivement des Tables Relationnelles Statistiques, des Tables Simples, des Tables Résumées et des Tables Statistiques Complexes.

Le langage STAQUEL propose un ensemble d'opérateurs d'affinement/élargissement sur des Tables Simples (MACRO-SELECT, MACRO-UNION. RESTRICTION, COMPARAISON). QBRST offre un opérateur de sélection de modalités (PROJECT). Le langage STBE permet la manipulation de Tables Résumées, au travers d'une extension de l'algèbre relationnelle aux relations avec attributs multivalués et aux fonctions agrégatives, et grâce à deux opérateurs de conversion de structure DEC et REL. DEC extrait, à

partir d'une Table Résumée, une Table Résumée primitive (similaire à une TS), pouvant être convertie par REL en relation.

Les langages QBSRT, STAQUEL et STBE n'intègrent aucun opérateur de transposition, ni de composition/décomposition. Les dérivations proposées se limitent généralement à l'agrégation par fonctions statistiques additives et ne prennent en compte aucun contrôle relatif à la fiabilité des données produites.

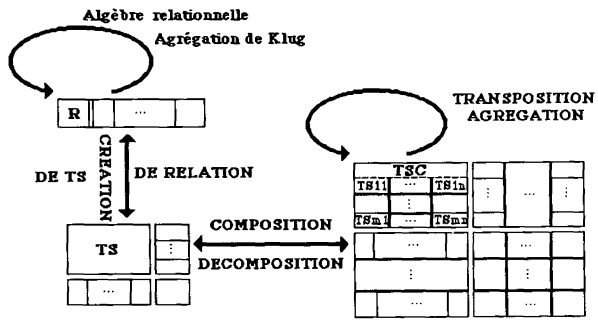

Figure 7. – Principe du langage STAR.

C': Count			BOURSIER
SEXE	DISCIP	AN_ET	C'NB

C' : Count			o	N
F	Informatique	1	25	39
F	Informatique	2	24	33
F	Mathématiques	1	18	21
F	Mathématique	2	15	20
F	Physique	1	27	14
F	Physique	2	20	18
M	Informatique	1	38	37
M	Informatique	2	20	42
M	Mathématiques	1	17	29
M	Mathématique	2	19	20
M	Physique	1	25	23
M	Physique	2	25	25

Figure 8. – Schéma et extension de la TS *C'*.

Le langage STAR, illustré par la figure 7, inclut des possibilités de création, de transposition, de composition/décomposition (limitées) et d'agrégation, par fonctions statistiques additives, de résumés statistiques et a recours à l'algèbre relationnelle pour les autres manipulations, comme dans STBE.

Nous insistons davantage sur la classe de transposition (reprise dans STAR+) afin d'en souligner l'intérêt. Cette classe comporte deux opérateurs de base : DISPLACE et ROTATE. DISPLACE a pour but de modifier le séquencement des attributs catégories dans un des multi-schémas ligne ou colonne d'une TSC. ROTATE modifie la dimension d'appartenance (ligne ou colonne) d'un attribut catégorie, en le faisant « basculer » dans l'autre dimension.

Comme nous l'avons signalé précédemment, les opérateurs de transposition n'ont pas seulement un objectif de présentation, leur intérêt est, en effet, indéniable dans un processus de regroupement de résumés, opéré en vue d'une meilleure interprétation des résumés résultats et d'une réduction considérable de l'espace de stockage, comme le montre l'exemple suivant.

Exemple 5 : Soit la TSC′ (dont la TSC est supposée être issue et qui lui est informationnellement équivalente). Son schéma et son extension sont donnés par la figure 8.

Les possibilités de transposition permettent d'envisager de regrouper les TSA et C′, en ayant, au préalable, transformé C′, grâce à l'opérateur ROTATE, afin d'obtenir la TSC. La TSC Z2 dont le schéma et l'extension sont donnés par la figure 9, est le résultat d'une telle opération.

Z2		AN_ET
A : Count		BOURSIER
C : Count		
SEXE	MILIEU	ANB
SEXE	DISCIP	CNB

Z2		1	1	2	2
A : Count					
C : Count		O	N	O	N
F	Agriculteurs	15	14	9	12
F	Ouvriers	30	25	23	29
F	Cadres	2	5	1	5
M	Agriculteurs	10	9	3	5
M	Ouvriers	43	40	35	42
M	Cadres	7	8	5	8
F	Informatique	25	39	24	33
F	Mathématiques	18	21	15	20
F	Physique	27	14	20	18
M	Informatique	38	37	29	42
M	Mathématiques	17	29	19	28
M	Physique	25	23	25	25

Figure 9. – Schéma et extension de la TSC Z2.

Le langage STAR+, schématisé par la figure 10, propose la manipulation de résumés statistiques à travers des opérateurs spécifiques sur les TSC, ce qui permet d'éviter de multiples opérations de conversion de structure entre relations et TS, inhérentes au principe du langage STAR, tout en contribuant à apporter une réponse adaptée à divers besoins de manipulation des résumés,

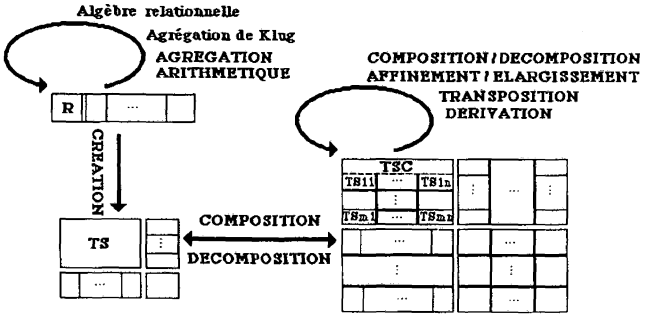

Figure 10. — Principe du langage STAR$^+$.

notamment la composition, l'interrogation ou encore les calculs arithmétiques et statistiques. La prise en compte de ces manipulations spécifiques a nécessité la re-définition de la structure des TSC par l'intégration, pour chaque résumé élémentaire, de la fonction statistique ayant permis de le calculer, ainsi qu'une étude de la compatibilité des TSC. Un complément à l'opérateur d'agrégation de Klug (Klug [17]) (utilisé dans STAR pour la création des résumés élémentaires à partir des micro-données) est apporté sous forme d'un opérateur d'agrégation arithmétique sur les relations, défini comme suit : soient une relation $R(U)$, X un sous-ensemble d'attributs de U, $t[X]$ la projection du n-uplet t sur X, f une fonction statistique additive, A et B deux attributs numériques, non nécessairement distincts, de U.

ARAGGREGATE $(R(X),\ f(\mathrm{op}(A,\ B)) = Y) = R'(X,\ Y) = \{\, t[X] \circ y / t \in R$ et $y = f(\mathrm{op}(t'[A],\ t'[B]))/t' \in R$ et $t'[X] = t[X]\,\}$; où op est l'un des opérateurs arithmétiques de base et $\circ$ symbolise la concaténation. Un tel opérateur, après un partitionnement préalable des tuples, effectue, pour chacun, un calcul arithmétique simple sur les valeurs de deux attributs numériques et somme les résultats ainsi obtenus pour chaque classe établie. Le rôle essentiel d'une telle agrégation est de permettre la création des résumés paramètres, nécessaires aux fonctions calculables (correspondant par exemple aux sommes de produits, somme de puissances, . . .).

Nous avons adopté le plan suivant, pour la présentation de la suite de cet article.

Le deuxième paragraphe propose un ensemble de définitions formelles, précisant la structure proposée et ses différents éléments composants. Nous y mettons en valeur la richesse sémantique des TSC et introduisons la notion de compatibilité entre TSC.

Les trois paragraphes suivants sont consacrés aux classes de manipulation : composition/décomposition, dérivation, affinement/élargissement. Pour chacune, nous proposons un ensemble d'opérateurs. Après en avoir précisé le rôle et la syntaxe, nous illustrons leur fonctionnement sur un exemple puis donnons l'expression formelle de leur résultat. Certains des opérateurs proposés sont dits complémentaires car ils peuvent être exprimés en fonction d'autres opérateurs, qualifiés eux de primitives. Ces expressions algébriques sont données en annexe.

2. CONCEPTS DE BASE ET DÉFINITIONS FORMELLES

Après avoir introduit un concept sémantique essentiel, celui de domaine catégorie, nous formalisons, dans ce paragraphe, la structure des TS et TSC, en définissant leur schéma, leur extension ainsi que la sémantique de leur(s) attribut(s) résumé(s). En correspondance avec sa perception intuitive duale, nous donnons une double définition du schéma d'une TSC, l'une en terme de TS, l'autre en terme d'attributs. Cette dernière est de plus homogène avec la définition d'une TS.

2.1. Domaine catégorie

Dans notre modèle, tout attribut, qu'il soit catégorie ou résumé, est défini sur un domaine, au sens relationnel (Delobel *et al.* [9]) comportant donc un double aspect, sémantique et syntaxique. Ce concept est enrichi lorsqu'il sert à spécifier l'ensemble des valeurs admissibles pour un attribut catégorie. Il est alors appelé domaine catégorie.

DÉFINITION 2.1 : *Domaine catégorie.* – Le domaine catégorie d'un attribut A, noté $\Delta(A)$, est défini comme un couple : $\Delta(A) = (D, \mathscr{R})$, où D est l'ensemble des modalités admissibles pour A et $\mathscr{R}$ une relation d'ordre total sur les éléments de D.

La notion d'ordre existant dans un domaine catégorie est essentielle dans notre modèle.

Cet ordre peut sembler inhabituel pour des domaines de type alphanumérique. Cependant, dans sa typologie des différents domaines de critères statistiques, (Steven [32]), identifie deux classes de domaines de ce type, dont une intègre une notion d'ordre. Il s'agit des domaines ordinaux, dont les éléments sont ordonnés (sans que la distance entre eux ne soit quantifiable).

L'autre classe regroupe des domaines nominaux, qui sont de simples ensembles de valeurs. Pour de tels domaines, la relation d'ordre est arbitrairement définie par spécification de la liste de leurs éléments. Le caractère arbitraire de cet ordre peut être perçu comme une contrainte, en fait l'usage fait qu'un ordre est implicitement adopté, ne serait-ce que pour faciliter la comparaison (y compris visuelle) de deux résumés (pouvant être des états de sortie).

Lorsque le domaine catégorie est de type numérique, (Steven [32]) identifie une relation d'ordre entre les différentes modalités du domaine et la présence ou non d'un zéro absolu. Dans ce cas, la relation d'ordre du domaine, dans notre modèle, est naturellement $\langle$ ou $\rangle$.

Exemple 2.1 : Considérons l'attribut DIPLOME, indiquant, pour les différents étudiants de la population analysée, le diplôme délivré à la fin de la formation suivie.

Le domaine de cet attribut est ordinal, ses valeurs sont indiquées par la liste suivante : {DEUG, Licence, Maîtrise, DEA, Doctorat}.

Le domaine de l'attribut SEXE est défini sous forme d'une liste, celui de AN_ET, comme un ensemble doté d'une relation d'ordre, présenté ici comme un intervalle.

$$\Delta(\text{SEXE}) = \{\text{F, M}\}; \qquad \Delta(\text{AN\_ET}) = [1 \ldots 7].$$

Ainsi, cette notion d'ordre confère au concept de domaine catégorie un aspect sémantique supplémentaire, en permettant une représentation fidèle des domaines ordinaux. Elle joue également un rôle organisationnel, exploité pour une manipulation aisée de l'extension des différents attributs catégories d'un résumé à sémantique arbitrairement complexe, comme nous le verrons par la suite.

2.2. Schémas des TS et TSC

Nous proposons ici la définition de l'intension des TS et TSC, en spécifiant au préalable quelques notions introduites de manière intuitive dans le premier paragraphe.

DÉFINITION 2.2 : *Schéma d'attributs catégories.* — Un schéma d'attributs catégories T est défini comme un ensemble ordonné de noms d'attribut catégorie. Il permet de structurer la dimension ligne d'une TS, il est alors noté $Tl = [A\,l_k / k \in [1 \ldots u]]$, ou sa dimension colonne : $Tc = [A\,c_k / k \in [1 \ldots w]]$. Un schéma peut être vide, il est noté : $\varnothing$.

DÉFINITION 2.3 : *Schéma de TS.* – Le schéma d'une TS P, noté SH(P), est défini par le triplet (Tl, Tc, R) où Tl est le schéma d'attributs catégories ligne de P, Tc est son schéma d'attributs catégories colonne et R est le nom de l'unique attribut résumé de P. De plus, pour que la TS ait un contenu sémantique valide, l'intersection de Tl et Tc doit être vide (un critère de classification n'est utilisé qu'une seule fois dans un résumé à sémantique simple).

Exemple 2.2 : Les schémas des TS A, B, C et D, illustrés par la figure 1, sont spécifiés par :

SH(A) = ([SEXE, MILIEU], [AN_ET, BOURSIER], ANB);

SH(B) = ([SEXE, MILIEU], [AN_ET], BMNT_BOURSE);

SH(C) = ([SEXE, DISCIP], [AN_ET, BOURSIER], CNB);

SH(D) = ([SEXE, DISCIP], [AN_ET], DMNT_BOURSE).

Les domaines catégories des attributs de ces TS sont les suivants :

Δ(SEXE) = {F, M}; Δ (MILIEU) = {Agriculteurs, Ouvriers, Employés, Commerçants, Cadres, Prof. libérales, Autres}; Δ(AN_ET) = [1 . . . 7]; Δ(BOURSIER) = {O, N};

Δ(DISCIP) = { . . ., Biologie, Géologie, Informatique, . . ., Mathématiques, Physique, . . . }.

Les attributs résumés ANB et BNB sont définis sur le même domaine D_NB : ENTIER. Le domaine commun aux attributs BMNT_BOURSE et DMNT_BOURSE est de type réel et appelé D_BOURSE.

DÉFINITION 2.4 : *Matrice des identifiants de TS d'une TSC.* – Le schéma d'une TSC, identifiée par son nom S et comportant N attributs résumés, est obtenu par regroupement des schémas des N TS : SH(P_{ij}) = (Tl_i, Tc_j, R_{ij}), selon des structures de ligne ou de colonne communes, en une matrice (m, n), (où $n \geqq 1$ et $m \geqq 1$) avec $n \times m = N$.

Celle-ci, notée MAT_TS(S), est appelée matrice des identifiants de TS de la TSC S. Chaque élément P_{ij} ($\forall i \in [1 . . . m]$, $\forall j \in [1 . . . n]$) de MAT_TS($S$) est l'identifiant de la TS constituant la i-ième ligne et j-ième colonne de la TSC S.

Cette définition correspond à la vision d'une TSC en terme de TS.

Exemple 2.3 : La matrice des identifiants de TS de la TSC Z1 est la suivante :

$$\text{MAT\_TS(Z1)} = [[A, B], [C, D]].$$

La particularité de MAT_TS(S) est que tous les éléments d'une ligne (resp. d'une colonne) dont l'indice est y, *i.e.*

$$\{P_{yj}/j \in [1 \ldots n]\} \text{ (resp. } \{P_{iy}/i \in [1 \ldots m]\})$$

ont le même schéma d'attributs catégories ligne Tl_y (resp. colonne Tc_y).

Nous notons :

φl : l'application surjective associant à chaque ligne de MAT_TS(S) le schéma d'attributs catégories ligne correspondant :

$$\forall i \in [1 \ldots m], \quad \varphi l : \{P_{iy}/y \in [1 \ldots n]\} \to \{Tl_i/i \in [1 \ldots m]\} \quad \text{avec } \varphi l(P_{iy}) = Tl_i.$$

φc : l'application surjective associant à chaque colonne de MAT_TS(S) le schéma d'attributs catégories colonne correspondant :

$$\forall j \in [1 \ldots n], \quad \varphi c : \{P_{jy}/y \in [1 \ldots m]\} \to \{Tc_j/j \in [1 \ldots n]\} \quad \text{avec } \varphi c(P_{yj}) = Tc_j.$$

φr : la bijection associant à chaque élément de MAT_TS(S) son unique attribut résumé R_{ij} :

$$\varphi r : \quad \{P_{ij}/i \in [1 \ldots m], j \in [1 \ldots n]\} \to \{R_{ij}/i \in [1 \ldots m], j \in [1 \ldots n]\}$$

avec

$$\varphi r(P_{ij}) = R_{ij}.$$

DÉFINITION 2.5 : *Multi-schéma d'attributs catégories.* — Nous appelons un multi-schéma d'attributs catégories ligne (resp. colonne) d'une TSC S, le vecteur Fl (resp. Fc) dont les m (resp. n) éléments sont obtenus en appliquant φl (resp. φc) sur MAT_TS(S), d'où

$$Fl = [Tl_i/i \in [1 \ldots m]] \text{ (resp. } Fc = [Tc_j/j \in [1 \ldots n]]).$$

Un multi-schéma d'attributs catégories est dit vide, lorsque tous les schémas le composant sont vides.

DÉFINITION 2.6 : *Matrice d'attributs résumés.* — Nous appelons matrice d'attributs résumés d'une TSC S, la matrice $[R_{ij}/i \in [1 \ldots m], j \in [1 \ldots n]]$, notée RESUM, obtenue en appliquant φr sur MAT_TS(S).

DÉFINITION 2.7 : *Schéma de TSC.* — Le schéma d'une TSC S, noté SH(S), est définie par le triplet (Fl, Fc, RESUM), où Fl et Fc sont les multi-schémas d'attributs catégories ligne et colonne de S et RESUM est sa matrice d'attributs résumés.

Remarque 2.1 : A travers la définition précédente, une TSC est vue en terme d'attributs. Cette définition découle directement de la notion de matrice des identifiants de TS d'une TSC, où une TSC est vue en terme de TS. Ces deux perceptions sont équivalentes et facilitent non seulement la définition des manipulations de la structure, mais aussi l'interprétation même des données, en permettant une identification aisée des résumés élémentaires au sein d'un résumé complexe ou arbitrairement complexe et une distinction claire entre classification et valeurs résumées.

Exemple 2.4 : Le schéma de la TSC Z1, illustré par la figure 3, est spécifié par :

SH(Z1) = ([[SEXE, MILIEU], [SEXE, DISCIP]], [[AN_ET, BOURSIER], [AN_ET]], [[ANB, BMNT_BOURSE], [CNB, DMNT_BOURSE]]).

Son multi-schéma colonne est composé du schéma colonne des TS A et C : [AN_ET, BOURSIER] et de celui des TS B et D, réduit à un seul attribut : [AN_ET]. Sa matrice d'attributs résumés, notée ici comme un vecteur de vecteurs lignes, est carrée et regroupe les quatre attributs résumés des TS composants.

2.3. Extension des TS et TSC

Nous introduisons, dans ce paragraphe, les notions de cardinalité et/ou extension, pour les différents éléments structurels d'une TSC : attribut catégorie, schéma et multi-schéma d'attributs catégories, attribut résumé et matrice d'attributs résumés, TS et TSC.

DÉFINITION 2.8 : *Cardinalité et extension d'un attribut catégorie.* — Dans une TS P, nous appelons cardinalité d'un attribut catégorie A, défini sur le domaine $\Delta(A) = (D, \mathscr{R})$, le nombre de valeurs distinctes de D, prises par A dans P. Elle est notée Card (A, P).

L'extension de A dans P est définie comme l'ensemble ordonné des Card (A, P) valeurs distinctes de A dans P, et notée :

$$\mathrm{VAL}(A, P) = \{ a_p \in D / p \in [1 \ldots \mathrm{Card}(A, P)] \}$$

et

$$\forall p \in [1 \ldots \mathrm{Card}(A, P) - 1],\ a_p \mathscr{R} a_{p+1}.$$

DÉFINITION 2.9 : *Cardinalité et extension d'un schéma d'attributs catégories.* — La cardinalité d'un schéma d'attributs catégories ligne (ou colonne) $Tl = [A\, l_k / k \in [1 \ldots u]]$ dans une TS P, est définie comme le nombre de combinaisons possibles des valeurs de ses attributs catégories (u-uplets).

Elle est notée Card (Tl, P). Si $Tl = \varnothing$, Card $(Tl, P) = 1$, sinon :

$$\text{Card}(Tl, P) = \prod_{k=1}^{u} \text{Card}(Al_k, P).$$

L'extension d'un schéma d'attributs catégories ligne (ou colonne) Tl, dans une TS P, est l'ensemble ordonné des combinaisons totales des valeurs de ses attributs catégories (u-uplets), défini par :

$$\text{VAL}(Tl, P) = \bigotimes_{k=1}^{u} \text{VAL}(Al_k, P)$$

où $\otimes$ est le symbole de produit cartésien total d'ensembles ordonnés.

Si $Tl = \varnothing$, VAL $(Tl, P) = \{\text{NUL}\}$, où NUL représente la valeur nulle signifiant « non existant ».

DÉFINITION 2.10 : *Extension d'un attribut résumé.* – L'extension d'un attribut résumé R d'une TS P est définie comme une matrice de valeurs résumées, de dimension (Card (Tl, P), Card (Tc, P)), notée :

$$\text{VAR}(R, P) = [r_{sv}/s \in [1 \ldots \text{Card}(Tl, P)], v \in [1 \ldots \text{Card}(Tc, P)]],$$

dont chaque élément est la valeur de R structurée selon le s-ième tuple de VAL (Tl, P) et le v-ième tuple de VAL (Tc, P).

DÉFINITION 2.11 : *Extension de TS.* – L'extension d'une TS P, de schéma SH $(P) = (Tl, Tc, R)$, est notée VAL (P) et définie par le triplet : [VAL (Tl, P), VAL (Tc, P), VAL (RESUM, P)], où VAL (Tl, P) et VAL (Tc, P) sont respectivement les extensions des schémas d'attributs catégories ligne et colonne de P et VAL (R, P) est l'extension de son attribut résumé.

Remarque 2.2 : Lorsque les schémas d'attributs catégories ligne et colonne sont simultanément vides dans une TS (donc de cardinalité égale à un), l'extension de l'attribut résumé est réduite à une unique valeur. Le niveau de détail du résumé représenté est alors le plus grossier possible et la TS est qualifiée de constante.

Exemple 2.5 : L'extension de la TS A, illustrée par la figure 2, est donnée ci-dessous.

VAL $(A) = (\{\langle F, \text{Agriculteurs}\rangle, \langle F, \text{Ouvriers}\rangle, \langle F, \text{Cadres}\rangle, \langle M, \text{Agriculteurs}\rangle, \langle M, \text{Ouvriers}\rangle, \langle M, \text{Cadres}\rangle, [\langle 1, O\rangle, \langle 1, N\rangle, \langle 2, O\rangle, \langle 2, N\rangle], [[15, 14, 9, 12], [30, 25, 23, 29], [2, 5, 1, 5], [23, 30, 26, 25], [10, 9, 3, 5], [43, 40, 35, 42], [7, 8, 5, 8], [20, 32, 30, 40]]\})$.

Dans la TS A, l'attribut SEXE prend toutes les valeurs de son domaine catégorie, par contre, l'extension des attributs MILIEU et AN_ET correspond à un sous-ensemble des valeurs de leur domaine respectif, dont le séquencement respecte l'ordre du domaine.

VAL(MILIEU, A) = { Agriculteurs, Ouvriers, Cadres };

VAL(AN_ET, A) = { 1, 2 }.

DÉFINITION 2.12 : *Cardinalité et extension d'un multi-schéma d'attributs catégories.* — Soit le multi-schéma d'attributs catégories ligne (ou colonne) $Fl = [Tl_i / i \in [1 \ldots m]]$, dans le schéma d'une TSC S. Nous définissons la cardinalité de Fl, notée Card(Fl, S), comme le nombre total de tuples, dans l'extension des schémas Tl_i de Fl, dans S. Si tous les schémas de Fl sont vides, Card(Fl, S) = m, sinon chaque Tl_i permet de structurer la dimension ligne d'une TS P_{ij} composant de S. D'où, $\forall j \in [1 \ldots n]$;

$$\text{Card}(Fl, S) = \sum_{i=1}^{m} \text{Card}(Tl_i, P_{ij}).$$

L'extension d'un multi-schéma d'attributs catégories ligne (ou colonne) Fl dans une TSC S, est définie comme l'ensemble ordonné de l'extension de ses schémas, obtenu par : $\forall j \in [1 \ldots n]$,

$$\text{VAL}(Fl, S) = \mathop{©}_{i=1}^{m} \text{VAL}(Tl_i, P_{ij})$$

ou © est la concaténation d'ensembles ordonnés.

NOTATIONS : (i) L'extension d'un multi-schéma étant définie comme un ensemble ordonné de tuples τ, nous utilisons par la suite la notation indiqué ci-dessous, par exemple pour Fl :

$$\text{VAL}(Fl, S) = \left\{ \tau_s \in \mathop{©}_{i=1}^{m} \text{VAL}(Tl_i, P_{ij}) / s \in [1 \ldots \text{Card}(Fl, S)] \right\}.$$

(ii) Un attribut, appartenant à un schéma $Tl_i (i \in [1 \ldots m])$ du multi-schéma Fl d'une TSC S, est noté : Al_k^i où $k \in [1 \ldots u_i]$. Ses valeurs sont désignées par : al_p^{ik} où $p \in [1 \ldots \text{Card}(Al_k^i, S)]$.

DÉFINITION 2.13 : *Extension d'une matrice d'attributs résumés.* — Nous définissons l'extension d'une matrice d'attributs résumés RESUM dans une TSC S, notée VAL(RESUM, S), la matrice de valeurs résumées composée

des $(m \times n)$ matrices $\mathrm{VAL}(R_{ij}, P_{ij})$ concaténées suivant les lignes et les colonnes comme suit :

$$\mathrm{VAL}(\mathrm{RESUM}, S) = \overset{n}{\underset{j=1}{\copyright_c}} \overset{m}{\underset{i=1}{\copyright_l}} \mathrm{VAL}(R_{ij}, P_{ij})$$
$$= [r_{sv}/s \in [1 \ldots \mathrm{Card}(Fl, S)], \qquad v \in [1 \ldots \mathrm{Card}(Fc, S)]]$$

où $\copyright_c$ et $\copyright_l$ sont les symboles de concaténation de matrices selon les colonnes ou les lignes.

DÉFINITION 2.14 : *Extension de TSC.* — L'extension d'une TSC S, de schéma $\mathrm{SH}(S) = (Fl, Fc, \mathrm{RESUM})$, est notée $\mathrm{VAL}(S)$ et définie par le triplet suivant : $(\mathrm{VAL}(Fl, S), \mathrm{VAL}(Fc, S), \mathrm{VAL}(\mathrm{RESUM}, S))$, où $\mathrm{VAL}(Fl, S)$ et $\mathrm{VAL}(Fc, S)$ sont les extensions des multi-schémas d'attributs catégories ligne et colonne de S et $\mathrm{VAL}(\mathrm{RESUM}, S)$ est l'extension de sa matrice d'attributs résumés.

Exemple 2.6 : L'extension de la TSC Z1, illustrées par la figure 4, est définie par : VAL(Z1) = ({ ⟨ *F*, Agriculteurs ⟩, ⟨ *F*, Ouvriers ⟩, ⟨ *F*, Cadres ⟩, ⟨ *M*, Agriculteurs ⟩, ⟨ *M*, Ouvriers ⟩, ⟨ *M*, Cadres ⟩, ⟨ *F*, Informatique ⟩, ⟨ *F*, Mathématiques ⟩, ⟨ *F*, Physique ⟩, ⟨ *M*, Informatique ⟩, ⟨ *M*, Mathématiques ⟩, ⟨ *M*, Physique ⟩], [⟨ 1, *O* ⟩, ⟨ 1, *N* ⟩, ⟨ 2, *O* ⟩, ⟨ 2, *N* ⟩, ⟨ 1 ⟩, ⟨ 2 ⟩], [[15, 14, 9, 12, 23.635, 11.700], [30, 25, 23, 29, 45.000, 34.500], [2, 5, 1, 5, 1.600, 1.000], [23, 30, 26, 25, 23.150, 28.700], [10, 9, 3, 5, 13.220, 4.560], [43, 40, 35, 42, 60.200, 50.750], [7, 8, 5, 8, 7.700, 5.500], [20, 32, 30, 40, 16.500, 39.050], [25, 39, 24, 33, 26.250, 28.800], [18, 21, 15, 20, 27.000, 17.900], [27, 14, 20, 18, 37.125, 28.200], [38, 37, 29, 42, 46.740, 41.730], [17, 29, 19, 28, 23.630, 28.880], [25, 23, 25, 25, 27.250, 29.250] }).

2.4. Définition des TS et TSC

Après avoir, dans les paragraphes précédents, proposé une construction de résumé au travers de règles essentiellement organisationnelles, nous nous resituons, à présent, dans le contexte de gestion de macro-données statistiques où les opérations de dérivation sont fondamentales et cependant propices aux erreurs de la part de l'utilisateur.

Nous avons, en introduction, évoqué l'importance que revêt la connaissance de la dérivabilité des résumés. Aussi, nous proposons d'inclure la nature des attributs résumés dans la sémantique des ensembles de macro-données, sous forme de la fonction statistique ayant permis de les calculer.

C'est par cette information que nous complétons les définitions intensive et extensive d'une TS. Elle nous permet non seulement de connaître l'éventuelle dérivabilité de ses valeurs résumées et ainsi d'interdire toute opération invalide, mais elle joue aussi, de par son caractère sémantique, un rôle essentiel dans l'interprétation même de ces valeurs.

DÉFINITION 2.15 : *TS.* – Une TS, identifiée par son nom P, est définie par le triplet : $P(\mathrm{SH}(P), \mathrm{VAL}(P), f)$ où $\mathrm{SH}(P)$ est le schéma de la TS, $\mathrm{VAL}(P)$ est son extension et f est la fonction statistique ayant permis de calculer l'attribut résumé de P.

Exemple 2.7 : La TS A est définie par : $[\mathrm{SH}(A), \mathrm{VAL}(A), \mathrm{Count}]$, où $\mathrm{SH}(A)$ et $\mathrm{VAL}(A)$ sont donnés dans les exemples 2.2 et 2.5.

Nous notons φf, la bijection associant à chaque élément de $\mathrm{MAT\_TS}(S)$ la fonction f_{ij} ayant engendré son unique attribut résumé R_{ij} :

$$\varphi f : \{P_{ij}/i \in [1 \ldots m], j \in [1 \ldots n]\} \rightarrow \{f_{ij}/i \in [1 \ldots m], j \in [1 \ldots n]\} \quad \text{avec} \quad \varphi f(P_{ij}) = f_{ij}.$$

DÉFINITION 2.16 : *Matrice des fonctions statistiques de TSC.* – Nous appelons matrice des fonctions d'une TSC S, notée $\mathrm{MAT\_}f(S)$, la matrice obtenue en appliquant φf sur la matrice des identifiants de TS de la TSC.

DÉFINITION 2.17 : *TSC.* – Une TSC, identifiée par son nom S, est définie par le triplet : $S(\mathrm{SH}(S), \mathrm{VAL}(S), \mathrm{MAT\_}f(S))$ où $\mathrm{SH}(S)$ est le schéma de la TSC, $\mathrm{VAL}(S)$ est son extension et $\mathrm{MAT\_}f(S)$ est la matrice des fonctions statistiques de S.

Exemple 2.8 : La TSC Z1 est formellement définie par :

(SH (Z1), VAL (Z1), [[Count, Sum], [Count, Sum]]), où SH (Z1) et VAL (Z1) sont donnés dans les exemples 2.4 et 2.6 et illustrés par les figures 3 et 4.

2.5. Compatibilité de TSC

Les quelques définitions que nous donnons dans ce paragraphe sont relatives à la compatibilité entre TSC. Comme dans un contexte de gestion de données classiques, la compatibilité des attributs est essentielle, pour en assurer des comparaisons ou des rapprochements cohérents. Nous la définissons de manière analogue à E. Codd, dans le modèle relationnel. Cependant, l'organisation multidimensionnelle des TSC nécessite de prolonger cette prise en compte sémantique, en explicitant la compatibilité des schémas d'attributs catégories (que l'on peut rapprocher de l'uni-compatibilité de relations) et

des multi-schémas, que nous définissons en cohérence avec la précédente. La compatibilité de deux TSC se base sur celle d'un de leurs multi-schémas, ligne ou colonne. Elle assure la validité sémantique de leur rapprochement. Pourtant elle est insuffisante pour garantir la cohérence de certaines opérations. En effet, cette compatibilité entre multi-schémas peut être vue comme une « quasi-identité » de leur intension, en terme de similarité d'organisation et de cohérence sémantique des attributs, pris deux à deux. Or, les manipulations spécifiques, les calculs par excellence, imposent également des contraintes d'identité de classification des résumés, que nous précisons au travers de la définition de totale compatibilité de deux TSC, selon une de leurs dimensions.

DÉFINITION 2.18 : *Compatibilité d'attributs.* — Deux attributs sont dits compatibles s'ils sont définis sur le même domaine.

DÉFINITION 2.19 : *Compatibilité de schémas d'attributs catégories.* — Deux schémas d'attributs catégories $T1=[A1_k/k\in[1\ldots u]]$ et $T2=A2_k/k\in[1\ldots u]]$ sont dits compatibles si et seulement si : $\forall k\in[1\ldots u]$, $A1_k$ et $A2_k$ sont compatibles. Deux schémas vides sont compatibles.

DÉFINITION 2.20 : *Compatibilité de multi-schémas d'attributs catégories.* — Deux multi-schémas d'attributs catégories non vides $F1=[T1_i/i\in[1\ldots m]]$ et $F2=[T2_i/i\in[1\ldots m]]$ sont dits compatibles si et seulement si : $\forall i\in[1\ldots m]$, $T1_i$ et $T2_i$ sont compatibles.

Deux multi-schémas vides sont compatibles.

DÉFINITION 2.21 : *Compatibilité de TSC.* — Deux TSC $S1$ et $S2$, de schéma $\mathrm{SH}(S1)=(F1\,l, F1\,c, \mathrm{RESUM}\,1)$ et $\mathrm{SH}(S2)=(F2\,l, F2\,c, \mathrm{RESUM}\,2)$ sont dites compatibles, selon les lignes (resp. les colonnes) si et seulement si $F1\,l$ et $F2\,l$ sont compatibles (resp. $F1\,c$ et $F2\,c$).

DÉFINITION 2.22 : *Compatibilité totale de TSC.* — Deux TSC $S1$ et $S2$, de schéma : $\mathrm{SH}(S1)=(F1\,l,\ F1\,c,\ \mathrm{RESUM}\,1)$ et $\mathrm{SH}(S2)=(F2\,l,\ F2\,c,$ RESUM 2) sont dites totalement compatibles, selon les lignes (resp. les colonnes) si et seulement si elles sont compatibles selon les lignes (resp. les colonnes) et que

$$\mathrm{VAL}(F1\,l, S1)=\mathrm{VAL}(F2\,l, S2)\ [\text{resp. } \mathrm{VAL}(F1\,c, S1)=\mathrm{VAL}(F2\,c, S2)].$$

Ces deux dernières définitions expriment, en quelque sorte, une identité par schéma et par contenu d'une des dimensions de deux TSC (signalons que différents types d'identité ou d'égalité entre objets sont isolés par (Masunaga [23]), dans une approche orientée objet. Sans atteindre la variété

et la « subtilité » des notions qui y sont introduites, ce travail de recherche nous conforte dans l'idée qu'un double degré de finesse est nécessaire pour apprécier la compatibilité de résumés). De telles identités, par schéma et par contenu, peuvent non seulement être doubles, en ligne et en colonne, mais la compatibilité des attributs résumés, la similarité de leur organisation et l'identité des matrices de fonctions statistiques sont des contraintes complémentaires nous permettant de définir très précisément le cadre des manipulations valides opérant sur les TSC.

2.6. Notion complémentaire : Concept de blocs

Dans ce paragraphe, nous introduisons une nouvelle notion, celle de bloc de valeurs résumées. Les blocs sont des sous-matrices de l'extension d'un attribut résumé, dans une TS ou TSC. La « division » en blocs de valeurs résumées concorde avec l'établissement de classes regroupant des combinaisons de modalités. Ces blocs peuvent intuitivement être perçus au travers d'un « découpage » de cette extension, prolongeant, au niveau de l'attribut résumé, une certaine vision de la classification associée, parmi toutes celles la synthétisant à des degrés divers. Cette vision, caractérisée par un niveau de détail donné, met en évidence un séquencement particulier de modalités identiques, dans l'extension des multi-schémas, qui découle de l'ordre inhérent à la structure des TSC. L'utilisation de telles propriétés ainsi que la facilité de manipulation d'une matrice par blocs ou sous-matrices d'éléments nous permet d'exprimer, par la suite, le résultat des opérateurs d'agrégation et d'affinement/élargissement, avec une relative aisance.

Considérons une TS P, de schéma $\mathrm{SH}(P)=(Tl,\ Tc,\ R)$, avec

$$Tl=[A\,l_1,\ \ldots,\ A\,l_u].$$

L'extension de l'attribut résumé de P étant une matrice, elle peut être perçue comme un vecteur de vecteurs, ou plus précisément comme un vecteur de :

- $\mathrm{Card}(Tl,\ P)$ lignes, notées L_s, $s\in[1\ldots\mathrm{Card}(Tl,\ P)]$, définies par :

$$L_s=[r_{sv}\in\mathrm{VAL}(R,\ P)/v\in[1\ldots\mathrm{Card}(Tc,\ P)]],$$

ou encore;

- $\mathrm{Card}(Tc,\ P)$ colonnes, notées C_v, $v\in[1\ldots\mathrm{Card}(Tc,\ P)]$, définies par :

$$C_v=[r_{sv}\in\mathrm{VAL}(R,\ P)/s\in[1\ldots\mathrm{Card}(Tl,\ P)]].$$

DÉFINITION 2.23 : *Bloc de valeurs résumées selon un attribut catégorie.* – Nous définissons un bloc de lignes (resp. de colonnes), dans l'extension de l'attribut résumé R d'une TS P, selon l'attribut $A\,l_k$ (resp. $A\,c_k$), comme une sous-matrice de cette extension, constituée de lignes L_s (resp. de colonnes C_v) *consécutives* dans VAL(R, S), et *structurées par la même valeur de* $A\,l_k$ *(resp. de* $A\,c_k$*)*. Un tel bloc est noté : $B(R, A\,l_k)$.

Cet ensemble de lignes L_s est, en fait, structuré par la même modalité non seulement de $A\,l_k$, mais également de tous les attributs précédant $A\,l_k$ dans le schéma $T\,l$ et ce par définition de l'extension d'un schéma d'attributs catégories.

Un bloc de valeurs de R dans P, selon un attribut catégorie $A\,l_k$, peut donc être exprimé ainsi :

$$B(R, A\,l_k) = [L_s / s \in [\inf \ldots \sup] \qquad \text{avec} \quad 1 \leqq \inf \leqq \sup \leqq \text{Card}(T\,l, P)$$

et

$$\forall s\,1, \quad s\,2 \in [\inf \ldots \sup], \; |\tau_{s1}|_k = |\tau_{s2}|_k]$$

où $|\tau_s|_k$ indique la restriction du s-ième tuple τ_s de VAL$(T\,l, P)$ à ses k premiers éléments (cette opération est assimilable à une projection relationnelle).

Considérons la « division » de VAL(R, P) en blocs selon l'attribut $A\,l_k$. Le nombre de lignes, noté $\eta(B, A\,l_k)$, dans les blocs ainsi obtenus, se détermine en fonction de la cardinalité des attributs suivants $A\,l_k$ dans le schéma $T\,l$ de P :

$$\eta(B, A\,l_k) = \prod_{y=k+1}^{u} \text{Card}(A\,l_y, P).$$

Le nombre de blocs selon l'attribut $A\,l_k$ dans l'extension de R, noté $\beta(A\,l_k, P)$, est donné par :

$$\beta(A\,l_k, P) = \text{Card}(T\,l, P) / \eta(B, A\,l_k) = \prod_{y=1}^{k} \text{Card}(A\,l_y, P).$$

Ainsi, l'extension d'un attribut résumé d'un TS peut être vue comme le résultat de la concaténation, selon les lignes, d'un ensemble ordonné de $\beta(A\,l_k, P)$ blocs $B_e(R, A\,l_k)$:

$$\text{VAL}(R, P) = \mathop{\textcircled{\scriptsize l}}_{e=1}^{\eta(B, A\,l_k)} B_e(R, A\,l_k).$$

Exemple 2.9 : Considérons la division de l'extension de la TS A, illustrée par la figure 2, en blocs de valeurs résumées selon l'attribut SEXE. Le nombre de ces blocs est déterminé, dans ce cas, par la cardinalité de l'attribut SEXE, *i.e.* 2, chacun comporte un nombre de lignes donné par la cardinalité de MILIEU : 4. D'où une vision de l'extension de l'attribut résumé ANB sous la forme suivante :

$$\text{VAL}(\text{ANB}, A) = B_1(\text{ANB}, \text{SEXE}) \copyright_1 B_2(\text{ANB}, \text{SEXE}) \text{ avec :}$$

B_1 (ANB, SEXE) = [[15, 14, 9, 12], [30, 25, 23, 29], [2, 1, 5, 5], [23, 30, 26, 25]] et B_2 (ANB, SEXE) = [[10, 9, 3, 5], [43, 40, 35, 42], [7, 8, 5, 8], [20, 32, 30, 40]].

3. OPÉRATEURS DE COMPOSITION/DÉCOMPOSITION

Les opérations de composition/décomposition des résumés permettent de rendre de plus en plus complexe ou de plus en plus simple la sémantique des résumés manipulés, sans toutefois modifier l'information véhiculée par les résumés élémentaires, ni leur dérivabilité potentielle. Ainsi, l'utilisateur peut créer de nouvelles TSC, par concaténation de deux TSC totalement compatibles, selon leur dimension, ligne ou colonne, commune, à travers l'opérateur de concaténation CONCATENATE, ou extraire différentes TS composants d'une TSC, grâce à PROJECT. Les opérateurs de jointure naturelle : JOIN et de jointure naturelle externe : OUTERJOIN sont deux alternatives à la concaténation, pour des TSC compatibles, mais dont l'extension du multi-schéma commun diffère. La TSC résultat est dotée d'un multi-schéma compatible avec celui des opérandes et dont l'extension correspond à leur intersection ou leur union extensive. Avec les quatre opérateurs de cette classe, l'utilisateur peut réaliser le rapprochement de différents résumés existants, selon un élément de structure commun, en ayant la possibilité au cours de cette opération de travailler assez finement sur l'extension du résultat.

CONCATENATE : Cet opérateur binaire permet de concaténer, selon les lignes ou les colonnes, deux TSC totalement compatibles, *i.e.* ayant un multi-schéma, ligne ou colonne, identique en intension et en extension. Il offre ainsi à l'utilisateur la possibilité de rapprocher différents résumés, pour faciliter l'interprétation de leur comparaison ou des opérations ultérieures. CONCATENATE n'a pas de répercussion sur les extensions de chacune des TSC opérandes, qui sont conservées dans la TSC résultat, ni sur leur matrice de fonctions statistiques.

La syntaxe de cet opérateur est la suivante : CONCATENATE (d, $S1$, $S2$), où d doit être remplacé par « l » ou « c » pour exprimer une concaténation des TSC $S1$ et $S2$ selon les lignes ou colonnes.

Exemple 3.1 : La TSC Z1 (illustrée par les figures 3 et 4) est obtenue à partir des TS A, C, B et D (*fig.* 1), totalement compatibles, selon les colonnes, deux à deux et dans l'ordre donné, par la séquence d'opérations de concaténation suivante : Z2 = CONCATENATE (c, A, C) puis Z3 = CONCATENATE (c, B, D) où les TSC résultats Z2 (*fig.* 9) et Z3 sont définies en terme de TS par : MAT_TS(Z2) = [A, C] et MAT_TS(Z3) = [B, D]. Z2 et Z3 étant totalement compatibles selon les lignes, nous avons : Z1 = CONCATENATE (l, Z2, Z3).

Formellement, considérons les deux TSC $S1$ et $S2$, dont les schémas sont les suivants :

SH($S1$) = ($F1\,l$, $F1\,c$, RESUM 1) et SH($S2$) = ($F2\,l$, $F2\,c$, RESUM 2), avec $F1\,l$ et $F2\,l$ deux multi-schémas compatibles et

$$\mathrm{VAL}(F1\,l,\ S1) = \mathrm{VAL}(F2\,l,\ S2).$$

La concaténation de $S1$ et $S2$, selon les lignes, produit la TSC S, définie comme suit :

CONCATENATE (l, $S1$, $S2$) = S, avec SH(S) = (Fl, Fc, RESUM) où :

- $Fl = F1\,l$ et $\mathrm{VAL}(Fl,\ S) = \mathrm{VAL}(F1\,l,\ S1) = \mathrm{VAL}(F2\,l,\ S2)$;
- $Fc = F1\,c \copyright F2\,c$ et $\mathrm{VAL}(Fc,\ S) = \mathrm{VAL}(F1\,c,\ S1) \copyright \mathrm{VAL}(F2\,c,\ S2)$

où © est le symbole de la concaténation d'ensembles ordonnés ;

- RESUM = RESUM 1 $\copyright_c$ RESUM 2 et

VAL(RESUM, S) = VAL(RESUM 1, $S1$) $\copyright_c$ VAL(RESUM 2, $S2$) où $\copyright_c$ représente la concaténation de matrices en colonne.

Enfin,

- MAT_f(S) = MAT_f($S1$) $\copyright_c$ MAT_f($S2$).

PROJECT : A travers cet opérateur unaire, l'utilisateur peut éliminer, dans un résumé, les différentes informations, qu'il juge non nécessaires, véhiculées par un ou plusieurs résumés élémentaires. Il peut donc extraire à partir d'une TSC, une de ses TS composants ou une nouvelle TSC, composée d'un sous-ensemble de TS, ayant en commun, deux à deux, un schéma d'attributs catégories ligne ou colonne. PROJECT n'a pas de répercussion sur les extensions des différentes TS extraites, ni sur la fonction statistique associée.

Sa syntaxe est la suivante : PROJECTS (S, $[P_{ij}]$), où $[P_{ij}]$ est une sous-matrice de la matrice des identifiants de TS de l'opérande, avec $i \in [1 \ldots m1]$, $j \in [1 \ldots n1]$, $1 \leqq m1 \leqq m$ et $1 \leqq n1 \leqq n$.

Exemple 3.2 : Considérons la TSC Z1, PROJECT (Z1, [A, C]) produit la TSC Z2, illustrée par la figure 9.

Formellement, soit une TSC S, de schéma : SH (S) = (Fl, Fc, RESUM). L'utilisateur désire opérer, à partir de S, l'extraction d'un ensemble de ($m1 \times n1$) TS, constituant une sous-matrice de MAT_TS (S), dont les ensembles d'indices ligne et colonne dans MAT_TS (S) sont respectivement notés : $\{i_1, \ldots, i_{m1}\}$ et $\{j_1, \ldots, j_{n1}\}$.

PROJECT (S, $[P_{ij}]$) = S' avec SH (S') = ($F'l$, $F'c$, RESUM'), où :

– seuls sont conservés, dans $F'l$, les schémas lignes des TS P_{ij} extraites, d'où :

$$F'l = [T'l_1, \ldots, T'l_{m1}] \quad \text{et} \quad \forall i' \in [1 \ldots m1], \quad T'l_{i'} = Tli_{i'}$$

et

$$\text{VAL}(T'l_{i'}, S') = \text{VAL}(Tli_{i'}, S);$$

– $F'c = [T'c_1, \ldots, T'c_{n1}]$ et $\forall j' \in [1 \ldots n1]$, $T'c_{j'} = Tcj_{j'}$ et $\text{VAL}(T'c_{j'}, S') = \text{VAL}(Tcj_{j'}, S)$;

– RESUM' est constituée uniquement des attributs résumés des TS extraites :

$$\forall i' \in [1 \ldots m1], \quad \forall j' \in [1 \ldots n1],$$

$$R'_{i'j'} = Ri_{i'}j_{j'} \quad \text{et} \quad \text{VAL}(\text{RESUM}'_{i'j'}, S') = \text{VAL}(\text{RESUM}\, i_{i'}j_{j'}, S).$$

Enfin,

– la matrice des fonctions statistiques de S' est une sous-matrice de celle de S, dont les éléments $f'_{i'j'}$ sont définis par :

$$\forall i' \in [1 \ldots m1], \quad \forall j' \in [1 \ldots n1], \quad f'_{i'j'} = fi_{i'}j_{j'}.$$

Les deux primitives CONCATENATE et PROJECT, définies ici, permettent par leur coopération une composition réversible de TSC (cette propriété est donnée en annexe).

JOIN : L'opérateur binaire JOIN travaille sur des TSC compatibles. Il opère une concaténation des deux opérandes selon leur dimension commune,

mais en ne considérant, pour chaque attribut catégorie, que les modalités qu'il prend à la fois dans les deux TSC opérandes, si bien que l'extension du multi-schéma ligne est constituée, dans le résultat, des tuples communs à celle des multi-schémas des opérandes. Seules les valeurs résumées, structurées par les combinaisons de modalités retenues, sont conservées pour les attributs résumés considérés. Le résultat de JOIN est compatible, selon la même dimension, avec les TSC opérandes.

Exemple 3.3 : Considérons le résumé complexe, représenté par la TSC Z4 (*fig.* 11), dénombrant tous les étudiants Niçois de deuxième cycle, par sexe,

Z4		Ø
E : Count		
F : Count		
SEXE	MILIEU	ENB
SEXE	DISCIP	FNB

Z4		Ø
E : Count		NUL
F : Count		
F	Agriculteurs	150
F	Ouvriers	320
F	Commerçants	405
M	Agriculteurs	102
M	Ouvriers	370
M	Commerçants	390
F	Informatique	294
F	Mathématiques	212
F	Physique	120
M	Informatique	350
M	Mathématiques	195
M	Physique	164

Figure 11. – Schéma et extension de la TSC $Z4$.

milieu socio-professionnel et par sexe, discipline. Dans Z4, les modalités de l'attribut MILIEU sont : {Agriculteurs, Ouvriers, Commerçants}.

Supposons que l'utilisateur cherche à analyser la représentativité des étudiants des deux premières années par rapport à tous ceux inscrits dans le même cycle, pour quelques disciplines scientifiques et certaines catégories socio-professionnelles. Il désire réaliser une composition préalable de Z1 et Z4. L'attribut MILIEU, étant, dans ces deux TSC, doté d'une extension différente, le rapprochement de Z1 et Z4, peut être réalisé, en éliminant les modalités non communes de cet attribut, par : JOIN (l, Z1, Z4). Le résultat de cette opération est la TSC Z5, illustrée par la figure 12.

Formellement, soient deux TSC $S1$ et $S2$, compatibles selon les lignes, de schéma :

$$\mathrm{SH}(S1)=(F1\,l,\ F1\,c,\ \mathrm{RESUM}\,1) \qquad \text{et} \qquad \mathrm{SH}(S2)=(F2\,l,\ F2\,c,\ \mathrm{RESUM}\,2)$$

Z5		
G: Count	H: Sum	I: Count
J: Count	K: Sum	L: Count

AN_ET	AN_ET	Ø
BOURSIER		

SEXE	MILIEU
SEXE	DISCIP

GNB	HMNT_BOURSE	INB
JNB	KMNT_BOURSE	LNB

Z5			1	1	2	2			
G: Count	H: Sum	I: Count							
J: Count	K: Sum	L: Count	O	N	O	N	1	2	NUL
F		Agriculteurs	15	14	9	12	23.625	11.700	150
F		Ouvriers	30	25	23	29	45.000	34.500	320
M		Agriculteurs	10	9	3	5	13.220	4.560	102
M		Ouvriers	43	40	35	42	60.200	50.750	370
F		Informatique	25	39	24	33	26.250	28.800	294
F		Mathématiques	18	21	15	20	27.000	17.900	212
F		Physique	27	14	20	18	37.125	28.200	120
M		Informatique	38	37	29	42	46.740	41.730	350
M		Mathématiques	17	29	19	28	23.630	28.880	195
M		Physique	25	23	25	25	27.250	29.250	164

Figure 12. – Schéma et extension de la TSC $Z5$.

avec :

$$\forall i \in [1 \ldots m], \quad \forall k \in [1 \ldots u_i],$$

les attributs $A1\,l_k^i$ et $A2\,l_k^i$ sont compatibles et $\mathrm{VAL}(A1\,l_k^i, S1) \cap \mathrm{VAL}(A2\,l_k^i, S2) \neq \varnothing$.

La vérification de cette dernière condition élimine le cas où deux attributs catégories compatibles de $S1$ et $S2$ ont des ensembles de modalités disjoints, dans l'extension de ces deux TSC.

JOIN $(l, S1, S2) = S$ avec $\mathrm{SH}(S) = (Fl, Fc, \mathrm{RESUM})$, où :

– $Fl = F1\,l$ et $\mathrm{VAL}(Fl, S) = \{\tau_s \in (\mathrm{VAL}(F1\,l, S1) \cap \mathrm{VAL}(F2\,l, S2))\}$, où $\cap$ respecte d'ordre sous-jacent des tuples dans l'extension des multi-schémas lignes de $S1$ et $S2$;

– $Fc = F1\,c \copyright F2\,c$ et $\mathrm{VAL}(Fc, S) = \mathrm{VAL}(F1\,c, S1) \copyright \mathrm{VAL}(F2\,c, S2)$;

– $\mathrm{RESUM} = \mathrm{RESUM}\,1 \copyright_c \mathrm{RESUM}\,2$.

Pour exprimer l'extension de RESUM, en fonction de celle de RESUM 1 et RESUM 2, nous introduisons les applications injectives $\psi 1$ et $\psi 2$, associant à chaque tuple de $\mathrm{VAL}(Fl, S)$ l'indice du tuple identique dans $\mathrm{VAL}(F1\,l, S1)$

et $\mathrm{VAL}(F2\,l,\ S2)$:

$$\psi 1 : \quad \mathrm{VAL}(F l,\ S) \to [1 \ldots \mathrm{Card}(F1\,l,\ S1)] \quad \text{et} \quad \psi 1(\tau_s) = s1/\tau_s = \tau 1_{s1}$$
$$\psi 2 : \quad \mathrm{VAL}(F l,\ S) \to [1 \ldots \mathrm{Card}(F2\,l,\ S2)] \quad \text{et} \quad \psi 2(\tau_s) = s2/\tau_s = \tau 2_{s2},$$

où

$$\tau 1_{s1} \in \mathrm{VAL}(F1\,l,\ S1) \quad \text{et} \quad \tau 2_{s2} \in \mathrm{VAL}(F2\,l,\ S2).$$

Alors :

$$\forall v \in [1 \ldots \mathrm{Card}(F1\,c,\ S1)],$$
$$r_{sv} = r\,1_{s1v}/r\,1_{s\,1\,v} \in \mathrm{VAL}(\mathrm{RESUM}\,1,\ S1) \quad \text{et} \quad s1 = \psi 1(\tau_s);$$
$$\forall v \in [\mathrm{Card}(F1\,c,\ S1)+1 \ldots \mathrm{Card}(F1\,c,\ S1)+\mathrm{Card}(F2\,c,\ S2)]$$
$$r_{sv} = r\,2_{s2v2}/r\,2_{s2v2} \in \mathrm{VAL}(F2\,c,\ S2);$$
$$v2 = v - \mathrm{Card}(F1\,c,\ S1) \quad \text{et} \quad s2 = \psi 2(\tau_s).$$

Enfin,

- $\mathrm{MAT}\_f(S) = \mathrm{MAT}\_f(S1)\ ©_c\ \mathrm{MAT}\_f(S2)$.

OUTERJOIN : L'opérateur binaire OUTERJOIN concatène deux TSC compatibles, selon leur dimension commune, par exemple les lignes. Il produit, une nouvelle TSC, compatible en ligne avec les opérandes, en considérant, pour chaque attribut catégorie, l'union des modalités dans l'extension des TSC opérandes. Pour chaque attribut résumé, les valeurs, correspondant aux mêmes combinaisons de modalités dans l'opérande considérée, sont retenues. Des valeurs résumées nulles sont introduites, pour les tuples rajoutés dans la classification. La sémantique de cette valeur nulle dépend de la fonction statistique associée aux TS. Elle est assimilée à *zéro* si la fonction est COUNT ou SUM, à *moins l'infini* si la fonction est MAX et à *plus l'infini* si la fonction est MIN.

Contrairement au précédent, cet opérateur permet de concaténer deux TSC non totalement compatibles, sans élimination d'aucune donnée existant dans les opérandes.

La syntaxe de l'opérateur est : OUTERJOIN $(d,\ S1,\ S2)$ où d indique la dimension selon laquelle les deux TSC opérandes sont compatibles.

Exemple 3.4 : Nous considérons la même opération de rapprochement entre Z1 et Z4 que dans l'exemple précédent, mais en réalisant l'union des modalités de l'attribut MILIEU.

Z6		
M: Count	N: Sum	O: Count
P: Count	Q: Sum	R: Count

AN_ET	AN_ET	Ø
BOURSIER		

SEXE	MILIEU	MNB	NMNT_BOURSE	ONB
SEXE	DISCIP	PNB	QMNT_BOURSE	RNB

Z6			1	1	2	2			
M: Count	N: Sum	O: Count							
P: Count	Q: Sum	R: Count	O	N	O	N	1	2	NUL
F		Agriculteurs	15	14	9	12	23.625	11.700	150
F		Ouvriers	30	25	23	29	45.000	34.500	320
F		Commerçants	NUL	NUL	NUL	NUL	NUL	NUL	405
F		Cadres	2	55	1	50	1.600	1.000	NUL
M		Agriculteurs	10	9	3	5	13.220	4.560	102
M		Ouvriers	43	40	35	42	60.200	50.750	370
M		Commerçants	NUL	NUL	NUL	NUL	NUL	NUL	390
M		Cadres	7	29	5	38	7.700	5.500	NUL
F		Informatique	25	39	24	33	26.250	28.800	294
F		Mathématiques	18	21	15	20	27.000	17.900	212
F		Physique	27	14	20	18	37.125	28.200	120
M		Informatique	38	37	29	42	46.740	41.730	350
M		Mathématiques	17	29	19	28	23.630	28.880	195
M		Physique	25	23	25	25	27.250	29.250	164

Figure 13. – Schéma et extension de $Z6$.

OUTERJOIN (l, Z1, Z4) produit comme résultat la TSC Z6, schématisée par la figure 13.

Formellement, soient deux TSC $S1$ et $S2$, compatibles selon les lignes, dont les schémas sont les suivants : SH ($S1$) = ($F1l$, $F1c$, RESUM 1) et SH ($s2$) = ($F2l$, $F2c$, RESUM 2). Nous avons donc : $\forall i \in [1 \ldots m]$, $\forall k \in [1 \ldots u_i]$, les attributs $A1l_k^i$ et $A2l_k^i$ sont compatibles (l'intersection de leur extension pouvant éventuellement être vide).

OUTERJOIN (l, $S1$, $S2$) = S, avec SH (S) = (Fl, Fc, RESUM) où :

– $Fl = F1l$. Contrairement à l'opérateur précédent, l'expression de l'extension de Fl, en fonction de celle de $F1l$ et $F2l$, nécessite la connaissance de l'ordre du domaine des différents attributs catégories. Pour cela, nous introduisons la fonction φ permettant d'ordonner l'union de deux ensembles ordonnés, en fonction de la relation d'ordre existant sur le domaine sous-jacent. Nous avons alors : $\forall i \in [1 \ldots m]$, $\forall k \in [1 \ldots u_i]$,

$$\mathrm{VAL}(F\,l, S) = \mathop{\copyright}_{i=1}^{m} \bigotimes_{k=1}^{u_i} \varphi(\mathrm{VAL}(A1\,l_k^i, S1) \cup \mathrm{VAL}(A2\,l_k^i, S2)).$$

– $Fc = F1\,c \,©\, F2\,c$ et $\mathrm{VAL}(Fc, S) = \mathrm{VAL}(F1\,c, S1) ©_c \mathrm{VAL}(F2\,c, S2)$;

– $\mathrm{RESUM} = \mathrm{RESUM}\,1 ©_c \mathrm{RESUM}\,2$.

Pour exprimer l'extension de RESUM, nous introduisons les applications injectives, notées $\varphi 1$ et $\varphi 2$, associant à chaque tuple respectivement de $\mathrm{VAL}(F1\,l, S1)$ et de $\mathrm{VAL}(F2\,l, S2)$, l'indice du tuple identique dans $\mathrm{VAL}(Fl, S)$:

$$\varphi 1 : \quad \mathrm{VAL}(F1\,l, S1) \to [1 \ldots \mathrm{Card}(Fl, S)] \quad \text{et} \quad \varphi 1(\tau 1_{s1}) = s/\tau 1_{s1} = \tau_s;$$

$$\varphi 2 : \quad \mathrm{VAL}(F2\,l, S2) \to [1 \ldots \mathrm{Card}(Fl, S)] \quad \text{et} \quad \varphi 2(\tau 2_{s2}) = s/\tau 2_{s2} = \tau_s,$$

où :

$$\tau 1_{s1} \in \mathrm{VAL}(F1\,l, S1); \; \tau 2_{s2} \in \mathrm{VAL}(F\,2\,l, S2)$$

$$\text{avec} \quad s1 \in [1 \ldots \mathrm{Card}(F1\,l, S1)]$$

et

$$s2 \in [1 \ldots \mathrm{Card}(F\,2\,l, S2)].$$

Chaque tuple τ_s, de l'extension de $F\,l$ dans S, permet de structurer un ensemble de valeurs résumées, constituant un vecteur dans VAL(RESUM, S), défini par :

$$[r_{sv}/v \in [1 \ldots \mathrm{Card}(F\,c, S)].$$

Toute valeur de ce vecteur est soit nulle, si elle est structurée par une combinaison de modalités ajoutée lors de la modification de multi-schéma ligne, soit égale à la valeur résumée de RESUM1 ou RESUM2 (suivant l'indice v), correspondant au même tuple de $F\,l$ dans $S1$ ou S2. Ainsi, $\forall v \in [1 \ldots \mathrm{Card}(F\,1\,c, S1)]$.

Si $\tau_s \notin \mathrm{VAL}(F\,1\,l, S1)$ alors $r_{sv} = \mathrm{NUL}$, sinon $r_{sv} = r\,1_{s1v}$ avec $s = \varphi 1(\tau 1_{s1})$.

$$\forall v \in [\mathrm{Card}(F\,1\,c, S1) + 1 \ldots \mathrm{Card}(F1\,c, S1) + \mathrm{Card}(F2\,c, S2)]$$

Si $\tau_s \notin \mathrm{VAL}(F\,2\,l, S2)$ alors $r_{sv} = \mathrm{NUL}$, sinon $r_{sv} = r\,2_{s2v2}$ avec :

$$s = \varphi 2(\tau 2_{s2}) \qquad \text{et} \qquad v = v2 - \mathrm{Card}(F\,1\,c, S1).$$

Enfin,

– $\mathrm{MAT}\_f(S) = \mathrm{MAT}\_f(S1) ©_c \mathrm{MAT}\_f(S2)$.

4. OPÉRATEURS DE DÉRIVATION

Nous englobons, dans la classe de dérivation, des possibilités d'agrégation et de calcul. Les premières permettent de synthétiser la classification d'ensembles de macro-données, alors que les secondes ne modifient pas le niveau de détail des données manipulées, mais toutes incluent une phase de calcul de nouvelles valeurs résumées. Mettant en jeu la fonction somme, l'opérateur AGGREGATE permet, appliqué sur des résumés obtenus par l'utilisation de fonction additive, d'en tirer la même information à un niveau moindre de détail. Pour cela, il procède à l'élimination d'un attribut catégorie dans le schéma d'une TSC et à la sommation des valeurs résumées concernées. L'opérateur complémentaire SAGGREGATE réalise la même opération mais sur tous les attributs catégories d'un schéma d'attributs ligne ou colonne, dans une TSC. Parmi les opérateurs de calcul proposés, nous en distinguons deux types : les opérateurs de calcul arithmétique et ceux de calcul statistique. A travers des calculs très simples sur les valeurs résumées, les premiers permettent des conversions d'unité, des consolidations ou des cumuls, par exemple pour déterminer le résumé représentant deux sous-ensembles distincts mais comparables de la population sous-jacente, à partir de résumés les synthétisant respectivement. Les opérateurs de calcul statistique fournissent les résultats équivalents à ceux de traitements sur les micro-données, faisant appel aux fonctions calculables. Ils s'appliquent sur les différents résumés, issus du calcul des fonctions additives paramètres de la fonction calculable, et effectuent alors les opérations arithmétiques requises sur les valeurs résumées.

AGGREGATE : AGGREGATE est un opérateur unaire, permettant la production de résumé dont le niveau de détail est plus grossier que celui de l'opérande. Ce dernier doit obéir à la contrainte de dérivabilité suivante : tous les résumés élémentaires de la TSC, concernés par l'aggrégation, doivent avoir été calculés par application d'une fonction additive.

AGGREGATE effectue une synthèse de toute ou une partie de la classification d'un résumé, en éliminant le dernier attribut catégorie d'un des schémas de la TSC le représentant, si bien que les nouvelles classes introduites correspondent à un regroupement des classes initiales. Les valeurs des différents attributs résumés, concernés par la manipulation, sont sommées, selon des blocs de valeurs correspondant à la nouvelle classification (*cf.* paragraphe 2.6). L'opérateur AGGREGATE ne dénature pas les attributs résumés sur lesquels il s'applique. La matrice des fonctions statistiques du résumé résultat est donc analogue à celle de l'opérande.

La syntaxe utilisée est : AGGREGATE (d, S, P_{ij}), où d est la dimension, ligne ou colonne, de l'attribut catégorie à supprimer. Celui-ci appartient au schéma, ligne ou colonne, d'une ou plusieurs TS, dont l'identifiant ou un des identifiants (P_{ij}) doit être précisé comme argument. La fonction somme est appliquée aux valeurs résumées de tous les attributs R_{ij}($i \in [1 \ldots m]$ ou $j \in [1 \ldots n]$), au préalable regroupées ensemble si elles sont structurées par les mêmes tuples dans le schéma concerné Tl_i ou Tc_j. De plus, $\forall i \in [1 \ldots m]$ ou $\forall j \in [1 \ldots n]$, f_{ij} doit être une fonction additive.

Exemple 4.1 : La TSC Z7, dont le schéma et l'extension sont illustrés par la figure 14, est obtenue à partir de la TSC Z1, en utilisant différentes

Z7			BOURSIER	Ø
S : Count	T : Sum			
U : Count	V : Sum			
SEXE	MILIEU	AN_ET	SNB	TMNT_BOURSE
SEXE	DISCIP	AN_ET	UNB	VMNT_BOURSE

Z7			O	N	NUL
S : Count		T : Sum			
U : Count		V : Sum			
F	Agriculteurs	1	15	14	23.625
F	Agriculteurs	2	9	12	11.700
F	Ouvriers	1	30	25	45.000
F	Ouvriers	2	23	29	34.500
F	Cadres	1	2	55	1.600
F	Cadres	2	1	50	1.000
M	Agriculteurs	1	10	9	13.220
M	Agriculteurs	2	3	5	4.560
M	Ouvriers	1	43	40	60.200
M	Ouvriers	2	35	42	50.750
M	Cadres	1	7	29	7.700
M	Cadres	2	5	38	5.500
F	Informatique	1	25	39	26.250
F	Informatique	2	24	33	28.800
F	Mathématiques	1	18	21	27.000
F	Mathématiques	2	15	20	17.900
F	Physique	1	27	14	37.125
F	Physique	2	20	18	28.200
M	Informatique	1	38	37	46.740
M	Informatique	2	20	42	41.730
M	Mathématiques	1	17	29	23.630
M	Mathématiques	2	19	20	28.880
M	Physique	1	25	23	27.250
M	Physique	2	25	25	29.250

Figure 14. – Schéma et extension de $Z7$.

opérations de transposition. Elle lui est équivalente informationnellement mais est organisée différemment.

L'utilisateur cherche à connaître le nombre d'étudiants, objets de l'examen, par sexe, milieu socio-professionnel et attribution ou non d'une bourse et

par sexe, discipline et attribution ou non d'une bourse. Il désire aussi obtenir le montant total des bourses attribuées, par sexe, milieu et par sexe, discipline. Or, le résumé Z7 contient l'information souhaitée mais à un niveau de détail plus fin. Tous les éléments de MAT_f (Z7) étant des fonctions additives, deux opérations d'agrégation successives permettent alors à l'utilisateur d'obtenir le résultat escompté, correspondant à la TSC Z8 (schématisée par la figure 15) :

AGGREGATE (l, Z7, S) = Z7', puis AGGREGATE (l, Z7', U) = Z8.

Formellement, considérons une TSC S, de schéma SH(S) = ($F\,l$, $F\,c$, RESUM), où $F\,l$ est un multi-schéma ligne non vide, contenant le schéma

Z8		BOURSIER	Ø
W : Count	X : Sum		
Y : Count	Z : Sum		
SEXE	MILIEU	WNB	XMNT_BOURSE
SEXE	DISCIP	YNB	ZMNT_BOURSE

Z8				
W : Count	X : Sum	O	N	NUL
Y : Count	Z : Sum			
F	Agriculteurs	24	26	35.325
F	Ouvriers	53	54	79.500
F	Cadres	3	105	2.600
M	Agriculteurs	13	14	17.780
M	Ouvriers	78	82	110.950
M	Cadres	12	67	13.200
F	Informatique	49	72	55.050
F	Mathématiques	33	41	44.900
F	Physique	47	32	65.325
M	Informatique	58	79	88.470
M	Mathématiques	36	49	52.510
M	Physique	50	48	56.500

Figure 15. – Schéma et extension de la TSC $Z8$.

$T\,l_d = [A\,l_k^d / k \in [1 \ldots u]]$, avec $u \geq 1$. De plus, tous les éléments f_{dj} ($j \in [1 \ldots n]$) de la matrice des fonctions statistiques de S sont des fonctions additives.

AGGREGATE (l, S, P_{dj}) = S' avec SH(S') = ($F'\,l$, $F'\,c$, RESUM') où :

– $F'\,l = [T'\,l_i / i \in [1 \ldots m]]$ et $\forall\, i \in [1 \ldots m] / i \neq d$, $T'\,l_i = T\,l_i$ et

$$\mathrm{VAL}(T'\,l_i, S') = \mathrm{VAL}(T\,l_i, S);$$

si $u = 1$, $T'\,l_d = \varnothing$ et si $u > 1$, $T'\,l_d = [A\,l_k^d / k \in [1 \ldots u-1]]$ avec :

$$\mathrm{VAL}(T'\,l_d, S') = \bigotimes_{k=1}^{u-1} \mathrm{VAL}(A\,l_k^d, S);$$

- $F'c = Fc$ et $\mathrm{VAL}(F'c, S) = \mathrm{VAL}(Fc, S)$.
- $\mathrm{RESUM}' = \mathrm{RESUM}$, et

chaque valeur $r'_{s'v}$ ($s' \in [1 \ldots \mathrm{Card}(F'1, S')]$, $v \in [1 \ldots \mathrm{Card}(F'c, S')]$) de l'extension de RESUM' est calculée en fonction des valeurs de VAL (RESUM, S) par :

$$\forall s' \in \left[1 \ldots \sum_{i=1}^{d-1} \mathrm{Card}(T' l_i, S')\right], \qquad r'_{s'v} = r_{s'v}$$

$$\forall s' \in \left[\sum_{i=1}^{d} \mathrm{Card}(T' l_i, S') + 1 \ldots \mathrm{Card}(F' l, S')\right], \; r'_{s'v} = r_{sv}$$

où

$$s = s' + (\mathrm{Card}(T l_d, S) - \mathrm{Card}(T' l_d, S'))$$

$$= s' + \mathrm{Card}(T l_d, S)(1 - 1/\mathrm{Card}(A l_u^d, S)),$$

$$\forall s' \in \left[\sum_{i=1}^{d-1} \mathrm{Card}(T' l_i, S') + 1 \ldots \sum_{i=1}^{d} \mathrm{Card}(T' l_i, S')\right],$$

$$r'_{s'v} = \sum_{s = (\omega+1) + (\mathrm{Card}(A l_u^d, S)(s' - \omega - 1))}^{(\omega+1)\,\mathrm{Card}(A l_u^d, S)(s' - \omega)} r_{sv} \qquad \text{où} \qquad \omega = \sum_{i=1}^{d-1} \mathrm{Card}(T' l_i, S').$$

Enfin,

- $\mathrm{MAT\_}f(S') = \mathrm{MAT\_}f(S)$.

Cas particulier : Si l'extension de l'attribut $A l_k^d$ est réduite à une unique valeur dans la TSC opérande (par exemple à la suite d'opérations d'affinement), l'opérateur AGGREGATE permet l'élimination de cet attribut; la phase de calcul des valeurs résumées est, dans ce cas, réduite à une simple affectation.

SAGGREGATE : Cet opérateur complémentaire réalise la même opération que la primitive AGGREGATE, mais sur tous les attributs d'un schéma ligne ou colonne, dans une TSC. Il impose la même contrainte de dérivabilité des résumés élémentaires concernés par l'agrégation et ne modifie pas les fonctions statistiques associées. Cet opérateur est particulièrement bien adapté au calcul de marge d'un tableau statistique (*i. e.* les totaux en ligne et en colonne), très utile aux statisticiens, comme nous l'illustrons à travers l'exemple suivant.

La syntaxe à utiliser est : SAGGREGATE (d, S, P_{ij}) où d est la dimension du schéma d'attributs à agréger, lequel organise notamment la TS P_{ij} dans la TSC S.

Exemple 4.2 : Supposons que l'utilisateur désire effectuer le calcul des marges pour le résumé élémentaire représenté par la TS A (*fig.* 1 et 2). Pour cela, il réalise les opérations suivantes :

$$AB = SAGGREGATE(c, A, A);\ AC = SAGGREGATE(l, A, A)$$

et

$$AD = SAGGREGATE(l, AB, AB) = SAGGREGATE(c, AC, AC).$$

Enfin, la concaténation des différentes TS obtenues est effectuée par :

$$A' = CONCATENATE(l, A, AB);\ A'' = CONCATENATE(l, AC, AD)$$

puis :
$Z9 = CONCATENATE(c, A', A'')$. Le résultat de cette série d'opérations est un résumé arbitrairement complexe, modélisé par la TSC Z9, dont le schéma et l'extension sont illustrés par la figure 16.

Z9		AN_ET	Ø
A : Count	AB : Count	BOURSIER	
AC : Count	AD : Count		
SEXE	MILIEU	ANB	ABNB
Ø		ACNB	ADNB

Z9		1	1	2	2	NUL
A : Count	AB : Count					
AC : Count	AD : Count	O	N	O	N	
F	Agriculteurs	15	14	9	12	50
F	Ouvriers	30	25	23	29	107
F	Cadres	2	5	1	5	13
M	Agriculteurs	10	9	3	5	27
M	Ouvriers	43	40	35	42	160
M	Cadres	7	8	5	8	28
	NUL	107	101	76	101	385

Figure 16. – Schéma et extension de la TSC Z9.

Formellement, considérons une TSC S, de schéma SH (S) = (Fl, Fc, RESUM), où Fl est un multi-schéma ligne non vide, contenant le schéma Tl_d. De plus, tous les éléments f_{dj} ($j \in [1 \ldots n]$) de la matrice des fonctions statistiques de S sont des fonctions additives.

SAGGREGATE $(l, S, P_{dj}) = S'$ avec $\mathrm{SH}(S') = (F'\,l,\ F'\,c, \mathrm{RESUM}')$ où :

– F' $[T'\ l_1, \ldots, T'\,l_m]$ et $\forall i \in [1 \ldots m]/i \neq d$, $T'\,l_i = T\,l_i$ et $\mathrm{VAL}(T'\,l_i, S') = \mathrm{VAL}(T\,l_i, S)$; et $T'\,l_d = \varnothing$;

– $F'\,c = F\,c$ et $\mathrm{VAL}(F'\,c, S') = \mathrm{VAL}(F\,c, S)$;

– $\mathrm{RESUM}' = \mathrm{RESUM}$.

Chaque valeur $r'_{s'v}(s' \in [1 \ldots Card(F'\,l, S')]$, $v \in [1 \ldots \mathrm{Card}(\mathrm{F}'\,c, S')]$ de l'extension de RESUM' est obtenue en fonction des valeurs de $\mathrm{VAL}(\mathrm{RESUM}, S)$ par :

$$\forall s' \in \left[1 \ldots \sum_{i=1}^{d-1} \mathrm{Card}(T'\,l_i, \mathrm{S}') \right], \qquad r'_{s'v} = r_{s'v}$$

$$\forall s' \in \left[\sum_{i=1}^{d+1} \mathrm{Card}(T'\,l_i, S') + 1 \ldots \mathrm{Card}(F'\,l, S') \right], \qquad r'_{s'v'} = r_{sv'}$$

où

$$s = s' - (\mathrm{Card}(T\,l_d, S) - 1).$$

Si

$$s' = \sum_{i=1}^{d-1} \mathrm{Card}(T'\,l_i, S') + 1,$$

$$r'_{s'v} = \sum_{s=\omega+1}^{(\omega+1)+\mathrm{Card}(T\,l_d, S)} \qquad \text{où} \quad \omega = \sum_{i=1}^{d-1} \mathrm{Card}(T'\,l_i, S').$$

Enfin,

– $\mathrm{MAT\_}f(S') = \mathrm{MAT\_}f(S)$.

Opérateurs de calcul arithmétique

Les quatre opérateurs arithmétiques +, −, *, / sont, dans leur fonctionnement, analogues. Cependant, pour contrôler la cohérence du résultat et en connaître la sémantique, l'utilisation directe des deux derniers est volontairement limitée.

Les primitives + et − peuvent être appliquées sur deux TSC totalement compatibles à la fois selon les lignes et selon les colonnes, et dont les attributs résumés sont également compatibles deux à deux. De plus, la matrice des fonctions statistiques des opérandes doit être identique. La TSC résultat d'une telle opération est totalement compatible, en ligne et en colonne,

avec les TSC initiales. Ses valeurs résumées sont calculées en sommant ou soustrayant, deux à deux, celles des opérandes. Il est également possible de fournir comme argument l'identifiant d'une TSC et une valeur constante. Ce n'est que dans le cadre de ce cas particulier que peuvent être utilisés les opérateurs * et /.

La syntaxe de l'opérateur + est donnée à titre d'exemple : +(S1, S2) ou +(S1, C). où S1 et S2 sont deux TSC dont les contraintes de compatibilité sont indiquées ci-dessus et C est une constante numérique, différente de zéro.

Exemple 4.3 : Supposons qu'il existe, dans la base exemple, un résumé Z10 fournissant exactement les mêmes comptages que Z9 (*fig.* 17), selon la même classification, mais pour les étudiants de l'Université de Toulon.

Z10		AN_ET	Ø
AE: Count	AF: Count		
AG: Count	AH: Count	BOURSIER	
SEXE	MILIEU	AENB	AFNB
Ø		AGNB	AHNB

Z10		1	1	2	2	NUL
AE : Count	AF : Count					
AG : Count	AH : Count	O	N	O	N	
F	Agriculteurs	2	0	1	3	6
F	Ouvriers	11	5	11	9	36
F	Cadres	1	20	3	16	40
M	Agriculteurs	5	3	4	9	21
M	Ouvriers	5	13	17	21	56
M	Cadres	0	10	8	9	27
	NUL	24	51	44	67	186

Figure 17. — Schéma et extension de la TSC *Z* 10.

Si l'utilisateur cherche à dénombrer les étudiants du sud-est, il peut déterminer les comptages globaux, *i. e.* pour les Universités de Nice et Toulon, selon la même classification, en opérant l'addition des deux TSC Z9 et Z10. Le résultat de l'opération +(Z9, Z10) est la TSC Z11, présentée par la figure 18.

Nous présentons, formellement, la TSC résultat de l'opérateur +, dans les deux cas d'utilisation possible, *i. e.* avec comme opérandes soit deux TSC, soit une TSC et une constante.

Considérons, d'abord, les deux TSC S1 et S2, dont les schémas sont les suivants :

SH(S1)=($F1l$, $F1c$, RESUM1) et SH($S2$)=($F2l$, $F2c$, RESUM2). $S1$ et $S2$ doivent vérifier les contraintes suivantes :

– VAL($F1l$, $S1$)=VAL($F2l$, $S2$) et VAL($F1c$, $S1$)=VAL($F2c$, $S2$);

Z11		AN_ET	Ø
AI : Count	AJ : Count		
AK : Count	AL : Count	BOURSIER	
SEXE	MILIEU	AINB	AJNB
Ø		AKNB	ALNB

Z11		1	1	2	2	NUL
AI : Count	AJ : Count					
AK : Count	AL : Count	O	N	O	N	
F	Agriculteurs	17	14	10	15	56
F	Ouvriers	41	30	34	38	143
F	Cadres	3	25	4	21	53
M	Agriculteurs	15	12	7	14	48
M	Ouvriers	48	53	52	63	216
M	Cadres	7	18	13	17	55
	NUL	131	152	120	168	571

Figure 18. – Schéma et extension de la TSC Z11.

– les attributs résumés de $S1$ et $S2$ sont compatibles deux à deux :

$\forall i \in [1 \ldots m]$ et $\forall j \in [1 \ldots n]$,

$R1_{ij}$ et $R2_{ij}$ sont définis sur le même domaine $D(R_{ij})$;

– les éléments de MAT_f ($S1$) et MAT_f($S2$) sont identiques deux à deux :

$$\forall i \in [1 \ldots m] \quad \text{et} \quad \forall j \in [1 \ldots n], \qquad f1_{ij} = f2_{ij}.$$

L'addition de ces TSC s'exprime par :

$+$**(S1, S2)** = **S** avec SH(S) = ($F\,l$, $F\,c$, RESUM) où :

– $F\,l = F\,1\,l$ et VAL($F\,l$, S) = VAL($F\,1\,l$, $S1$) = VAL($F\,2\,l$, $S2$);

– $F\,c = F\,1\,c$ et VAL($F\,c$, S) = VAL($F\,1\,c$, $S1$) = VAL($F\,2\,c$, $S2$);

– RESUM = RESUM1 et

$$\forall s \in [1 \ldots \text{Card}(F\,l, S)], \quad \forall v \in [1 \ldots \text{Card}(F\,c, S)],$$

si $r1_{sv}$ = NUL ou $r2_{sv}$ = NUL alors r_{sv} = NUL, sinon $r_{sv} = r1_{sv} + r2_{sv}$.

Enfin,

– MAT_f (S) = MAT_f ($S1$) = MAT_f ($S2$).

Pour le deuxième cas, nous utilisons la TSC S et une constante numérique C spécifiée par l'utilisateur.

$+(\mathbf{S}, \mathbf{C}) = \mathbf{S}'$ avec $\mathrm{SH}(S') = \mathrm{SH}(S)$ et :

- $\mathrm{VAL}(F'\,l, S') = \mathrm{VAL}(F\,l, S)$;
- $\mathrm{VAL}(F'\,c, S') = \mathrm{VAL}(F\,c, S)$;
- $\forall s \in [1 \ldots \mathrm{Card}(F'\,l, S')]$, $\forall v \in [1 \ldots \mathrm{Card}(F'\,c, S')]$

les valeurs résumées r'_{sv} de VAL(RESUM, S) s'expriment par :

si $r_{sv} = \mathrm{NUL}$ alors $r'_{sv} = \mathrm{NUL}$ sinon $r'_{sv} = r_{sv} + C$.

Enfin,

- $\mathrm{MAT\_}f(S') = \mathrm{MAT\_}f(S)$.

L'opérateur − est strictement similaire à +, mais il effectue la soustraction de valeurs résumées. Quant aux primitives * et /, utilisables uniquement avec une constante comme argument, elles sont analogue au deuxième cas d'utilisation de + et mettent en jeu l'opérateur arithmétique correspondant.

Opérateurs de calcul statistique

Pour être à même d'effectuer les calculs statistiques usuels, tout en contrôlant, voire guidant, l'utilisation, notre démarche se base sur les travaux de paramétrisation des fonctions statistiques calculables et propose une famille d'opérateurs, chacun étant dédié à un calcul particulier. Tous les résumés opérandes dans ces calculs doivent être dérivables (les paramètres des fonctions calculables étant des fonctions additives). De plus, il est possible d'affiner les contraintes portant sur les opérandes, en fonction de chaque opérateur.

Quel que soit le calcul envisagé, il n'existe pas de différence fondamentale dans la définition de l'opérateur correspondant. Aussi, nous nous contentons de deux exemples, AVERAGE et COVARIANCE, pour expliciter formellement la TSC résultat par rapport aux opérandes.

La syntaxe adoptée pour ces opérateurs est la suivante : CALCUL_STAT($S\,l, \ldots, S\,x$), où CALCUL_STAT est un terme générique à substituer par le calcul voulu et $S\,l, \ldots, S\,x$ sont les résumés dérivables paramètres.

Exemple 4.4 : Supposons que l'utilisateur cherche à déterminer la moyenne des bourses attribuées par sexe, milieu socio-professionnel et par sexe, discipline, pour les étudiants Niçois boursiers des deux premières années de deuxième cycle. Cette information peut être tirée de la TSC Z8 (*fig.* 15), qui donne, pour la même classification, le montant total des bourses attribuées

et également le nombre d'étudiants boursiers. Mais, au préalable, les manipulations suivantes doivent être réalisées :

– une opération d'affinement doit être effectuée, afin de ne conserver pour l'attribut catégorie BOURSIER, que la modalité « O ». Une telle opération permet d'éliminer les comptages relatifs aux étudiants non boursiers, et fait appel à la primitive SELECT (*cf.* paragraphe 5);

– une opération d'agrégation est alors nécessaire pour supprimer l'attribut catégorie BOURSIER. Elle correspond au cas particulier où la fonction Somme est appliquée sur des classes réduites à une unique valeur. Cette opération ne modifie donc pas les extensions des attributs résumés WNB et YNB de Z8, qui conservent les valeurs retenues dans l'étape précédente. Son objectif est simplement d'éliminer l'attribut BOURSIER de la TSC, dont le multi-schéma colonne est alors vide;

– deux opérations d'extraction permettent ensuite d'isoler les paramètres nécessaires au calcul, sous forme de deux TSC dont la sémantique correspond respectivement au nombre d'étudiants boursiers et au montant total des bourses attribuées pour la classification suivante : sexe, milieu socio-professionnel et sexe, discipline.

Cette séquence de manipulations produit donc comme résultat deux TSC, totalement compatibles en ligne et en colonne, Z12 et Z13 illustrées par la figure 19.

Ces deux TSC peuvent alors être utilisées comme arguments de l'opérateur de calcul AVERAGE, afin d'obtenir les valeurs moyennes des bourses attribuées, demandées par l'utilisateur, de la manière suivante : AVERAGE(Z13, Z12). Le résumé, résultat de la requête, est représenté par la TSC Z14, dont le schéma et l'extension sont donnés par la figure 20.

Pour obtenir ces valeurs moyennes, par sexe, ou pour calculer la moyenne générale des bourses pour les étudiants examinés, le même opérateur peut être appliqué sur les résumés opérandes, au préalable agrégés pour obtenir le niveau de détail requis.

Nous nous intéressons, dans un premier temps, à l'opérateur AVERAGE, en définissant la TSC résultat en fonction des TSC opérandes, en ayant au préalable indiqué les contraintes que doivent vérifier ces dernières. Nous procédons, ensuite, de manière analogue pour l'opérateur COVARIANCE.

Considérons, d'abord, les deux TSC S1 et S2, dont les schémas sont les suivants : SH($S1$)=($F1l$, $F1c$, RESUM1) et SH($S2$)=($F2l$, $F2c$,

Z12		
AM : Count		Ø
AN : Count		
SEXE	MILIEU	AMNB
SEXE	DISCIP	ANNB

Z13		
AO : Sum		Ø
AP : Sum		
SEXE	MILIEU	AOMNT_BOURSE
SEXE	DISCIP	APMNT_BOURSE

Z12		
AM : Count		NUL
AN : Count		
F	Agriculteurs	24
F	Ouvriers	53
F	Cadres	3
M	Agriculteurs	13
M	Ouvriers	78
M	Cadres	12
F	Informatique	49
F	Mathématiques	33
F	Physique	47
M	Informatique	58
M	Mathématiques	36
M	Physique	50

Z13		
AO : Sum		NUL
AP : Sum		
F	Agriculteurs	35.325
F	Ouvriers	79.500
F	Cadres	2.600
M	Agriculteurs	17.780
M	Ouvriers	110.950
M	Cadres	13.200
F	Informatique	55.050
F	Mathématiques	44.900
F	Physique	65.325
M	Informatique	88.470
M	Mathématiques	52.510
M	Physique	56.500

Figure 19. – Schéma et extension de la TSC $Z12$ et $Z13$.

Z14		
AQ : Avg		Ø
AR : Avg		
SEXE	MILIEU	AVG(MNT_BOURSE)
SEXE	DISCIP	AVG(MNT_BOURSE)

Z14		
AQ : Avg		NUL
AR : Avg		
F	Agriculteurs	1472
F	Ouvriers	1500
F	Cadres	867
M	Agriculteurs	1368
M	Ouvriers	1422
M	Cadres	1100
F	Informatique	1123
F	Mathématiques	1361
F	Physique	1390
M	Informatique	1525
M	Mathématiques	1459
M	Physique	1130

Figure 20. – Schéma et extension de la TSC $Z14$.

RESUM2), où $S1$ et $S2$ vérifient les containtes suivantes :

- $\text{VAL}(F1\,l, S1) = \text{VAL}(F2\,l, S2)$ et $\text{VAL}(F1\,c, S1) = \text{VAL}(F2\,c, S2)$;
- $\forall i \in [1 \ldots m]$ et $\forall j \in [1 \ldots n]$, $f1_{ij} = \text{Sum}$ et $f2_{ij} = \text{Count}$.

AVERAGE (S1, S2) = S avec $\text{SH}(S) = (F\,l, F\,c, \text{RESUM})$ où :

- $F\,l = F1\,l$ et $\text{VAL}(F\,l, S) = \text{VAL}(F1\,l, S1) = \text{VAL}(F2\,l, S2)$;
- $F\,c = F1\,c$ et $\text{VAL}(F\,c, S) = \text{VAL}(F1\,c, S1) = \text{VAL}(F2\,c, S2)$;
- RESUM = AVG (RESUM1)

et

$$\forall s \in [1 \ldots \text{Card}(F\,l, S)], \forall v \in [1 \ldots \text{Card}(F\,c, S)],$$

si $r1_{sv} \neq \text{NUL}$, $r2_{sv} \neq \text{NUL}$ et $r2_{sv} \neq 0$ alors $r_{sv} = r1_{sv}/r2_{sv}$, sinon $r_{sv} = \text{NUL}$.

Enfin,

– $\forall i \in [1 \ldots m]$ et $\forall j \in [1 \ldots n]$, $f_{ij} = \text{A}vg$.

Considérons, à présent, la TS $S1$ et les TSC compatibles en ligne et colonne $S2$, $S3$ et $S4$ dont les schémas sont : $\text{SH}(S1) = (T1\,l, T1\,c, \text{R1})$, $\text{SH}(S2) = (F\,2l, F\,2c, \text{RESUM2})$, $\text{SH}(S3) = (F\,3l, F\,3c, \text{RESUM3})$ et $\text{SH}(S4) = (F\,4l, F\,4c, \text{RESUM4})$, et qui vérifient les contraintes suivantes :

– $T1\,l = \varnothing$; $T1\,c = \varnothing$; l'unique valeur rl de VAL(R1, $S1$), est différente de zéro et de la valeur nulle et la fonction statistique associée à $S1$ est : $f1 = \text{Count}$. $S1$ est donc une TS constante ;

– $\text{VAL}(F\,2l, S2) = \text{VAL}(F\,3l, S3) = \text{VAL}(F\,4l, S4)$;

– $\text{VAL}(F\,2c, S2) = \text{VAL}(F\,3c, S3) = \text{VAL}(F\,4c, S4)$;

– $\forall i \in [1 \ldots m]$ et $\forall j \in [1 \ldots n]$, $f\,2_{ij} = \text{Sum (Prod)}$; $f\,3_{ij} = \text{Sum}$ et $f\,4_{ij} = \text{Sum}$.

COVARIANCE (S1, S2, S3, S4) = S avec $\text{SH}(S) = (F\,l, F\,c, \text{RESUM})$ où :

– $F\,l = F\,2l$ et

$$\text{VAL}(F\,l, S) = \text{VAL}(F\,2l, S2) = \text{VAL}(F\,3l, S3) = \text{VAL}(F\,4l, S4) ;$$

– $F\,c = F\,2c$ et

$$\text{VAL}(F\,c, S) = \text{VAL}(F\,2c, S2) = \text{VAL}(F\,3c, S3) = \text{VAL}(F\,4c, S4) ;$$

– $\text{RESUM} = \text{COV (RESUM3, RESUM4)}$ et

$$\forall s \in [1 \ldots \text{Card}(F\,l, S)], \quad \forall v \in [1 \ldots \text{Card}(F\,c, S)], \quad \text{si} \quad r2_{sv} \neq \text{NUL}, \quad r3_{sv} \neq \text{NUL} \quad \text{et} \quad r4_{sv} \neq \text{NUL}$$

alors $r_{sv} = (1/(r1 - 1))\,(r2_{sv} - (1/r1)\,r3_{sv}\,r4_{sv})$ sinon $r_{sv} = \text{NUL}$;

Enfin,

– $\forall i \in [1 \ldots m]$ et $\forall j \in [1 \ldots n]$, $f_{ij} = \text{Cov}$.

5. OPÉRATEURS D'AFFINEMENT/ÉLARGISSEMENT

Les manipulations d'affinement et d'élargissement permettent à l'utilisateur de travailler très finement sur les valeurs d'un résumé et en particulier d'intervenir sur sa classification. En affinant ou élargissant, d'où leur nom, l'ensemble des valeurs d'un attribut catégorie, elles offrent à l'utilisateur la possibilité de restreindre ou d'étendre le champ de son investigation statistique et ainsi d'obtenir un résumé dont l'ensemble des données lui semble le plus pertinent pour l'étude menée. Ces manipulations ont également un rôle essentiel dans un processus de composition de résumés statistiques ou encore de calcul, car elles permettent de rendre totalement compatibles deux résumés dotés de classifications homogènes mais cependant différentes. Pour cette classe de manipulation, nous définissons les opérateurs UNION, INTERSECTION, DIFFERENCE, NRESTRICT, CSELECT et RSELECT. Les opérateurs de calcul ou la primitive CONCATENATE, définis sur les TSC, imposent la contrainte forte de totale compatibilité des résumés opérandes. Agencer des résumés par exemple, au sein d'une nouvelle TSC, ne peut se faire que s'ils ont une dimension commune, en intension et en extension, par définition même de cette structure. Pour éviter les limitations de manipulation découlant de cette contrainte, nous nous sommes intéressés, à la mise en cohérence totale de deux TSC compatibles. Dans cet objectif, nous définissons les opérateurs binaires UNION et INTERSECTION. Ils créent une nouvelle TSC, à partir de la première opérande, en réalisant un « complément », par union des combinaisons de modalités, ou un « croisement », par intersection, de sa classification par rapport à celle de la deuxième TSC opérande. Ces deux primitives permettent, avec CONCATENATE, l'expression des opérateurs complémentaires JOIN et OUTERJOIN. Dans le même esprit, DIFFERENCE effectue la différence ensembliste des combinaisons de modalités de deux TSC. NRESTRICT permet, notamment au cours d'un processus de décomposition de résumés, de retrouver l'ensemble des données significatives d'une TSC, composant de ce résumé, en éliminant les éventuelles modifications apportées sur la TSC, lors d'une composition ou plus généralement d'une comparaison avec d'autres. Il s'agit de supprimer les combinaisons manquantes existant dans la classification d'un résumé, tout en respectant la définition extensive d'un schéma d'attributs catégories. Par l'opérateur CSELECT, l'utilisateur restreint l'ensemble des modalités d'un attribut catégorie, aux seules valeurs vérifiant une qualification donnée. Il peut, de la même manière, éliminer des valeurs résumées jugées non pertinentes, ou encore effectuer, parmi elles, la recherche du minimum ou du maximum, grâce à RSELECT.

UNION : Cet opérateur de base admet comme arguments deux TSC compatibles, selon une dimension, et engendre une nouvelle TSC, compatible avec les précédentes, selon la même dimension. Utilisé pour compléter une TSC par rapport à une autre, et inversement, UNION engendrent deux TSC totalement compatibles, dont le multi-schéma commun peut être perçu, de manière extensive, comme l'union de ses extensions initiales dans les TSC opérandes. En effet, UNION procède en ajoutant, à l'extension des attributs catégories de la première TSC, des modalités existant pour les attributs compatibles dans la seconde, d'où une insertion de nouveaux tuples dans l'extension des schémas et du multi-schéma concernés. Cet opérateur agit également sur l'extension de la matrice des attributs résumés, par un ajout de valeurs nulles correspondant aux nouvelles combinaisons de modalités introduites.

D'un point de vue syntaxe, UNION admet trois arguments : l'identifiant de la TSC à « compléter » et celui de la TSC par rapport à laquelle s'effectue ce « complément », ainsi que la dimension ligne ou colonne de leur multi-schéma commun : UNION(d, $S1$, $S2$).

Exemple 5.1 : Nous reprenons l'exemple du rapprochement des TSC Z1 (*fig.* 3 et 4) et Z4 (*fig.* 11). Ces deux TSC ont un multi-schéma ligne compatible, mais l'extension de l'attribut catégorie MILIEU diffère d'une TSC à l'autre. Pour les rendre totalement compatibles, l'utilisateur peut effectuer une double opération UNION. La première : UNION(l, Z1, Z4) engendre la TSC Z15, illustrée par la figure 21. Z15 est dotée d'une matrice de TS, qui est une sous-matrice de celle de la TSC Z6, introduite dans l'exemple 3.4, en utilisant l'opérateur de composition OUTERJOIN sur les mêmes TSC opérandes. Une deuxième opération UNION(l, Z4, Z1) permet la création d'une TSC totalement compatible en ligne avec Z15 et dont la matrice des identifiants de TS est : [[O], [R]] (*cf.* exemple 3.4, *fig.* 13).

Formellement, considérons les deux TSC $S1$ et $S2$, compatibles selon les lignes, de schéma :

SH($S1$)=($F1l$, $F1c$, RESUM1) et SH($S2$)=($F2l$, $F2c$, RESUM2), où :

VAL($F1l$, $S1$)≠VAL($F2l$, $S2$).

UNION(l, $S1$, $S2$)=$S1'$, avec SH($S1'$)=SH($S1$), et :

— l'extension de $F1l$ dans $S1'$ s'exprime en fonction de celle de $F1l$ dans $S1$ et $F2l$ dans $S2$ en utilisant la fonction φ (introduite pour l'opérateur OUTERJOIN), permettant d'ordonner l'union de deux ensembles ordonnés

Z15		AN_ET	AN_ET
M : Count	N : Sum		
P : Count	Q : Sum	BOURSIER	
SEXE	MILIEU	MNB	NMNT_BOURSE
SEXE	DISCIP	PNB	QMNT_BOURSE

Z15		1	1	2	2		
M : Count	N : Sum						
P : Count	Q : Sum	O	N	O	N	1	2
F	Agriculteurs	15	14	9	12	23.625	11.700
F	Ouvriers	30	25	23	29	45.000	34.500
F	Commerçants	NUL	NUL	NUL	NUL	NUL	NUL
F	Cadres	2	55	1	50	1.600	1.000
M	Agriculteurs	10	9	3	5	13.220	4.560
M	Ouvriers	43	40	35	42	60.200	50.750
M	Commerçants	NUL	NUL	NUL	NUL	NUL	NUL
M	Cadres	7	29	5	38	7.700	5.500
F	Informatique	25	39	24	33	26.250	28.800
F	Mathématiques	18	21	15	20	27.000	17.900
F	Physique	27	14	20	18	37.125	28.200
M	Informatique	38	37	29	42	46.740	41.730
M	Mathématiques	17	29	19	28	23.630	28.880
M	Physique	25	23	25	25	27.250	29.250

Figure 21. – Schéma et extension de la TSC $Z15$.

de modalités, en fonction de la relation d'ordre existant sur le domaine sous-jacent. Nous avons alors :

$$\mathrm{VAL}(F\,1\,l,\,S1') = \bigcirc_{i=1}^{m} \otimes_{k=1}^{u_i} \varphi\,\mathrm{VAL}(\mathrm{A}\,1\,l_k^i,\,S1) \cup \mathrm{VAL}(\mathrm{A}\,2l_k^i,\,S2)).$$

– $\mathrm{VAL}(F\,1\,c,\,S1') = \mathrm{VAL}(F\,1\,c,\,S1)$;

– l'extension de RESUM1 dans $S1'$ regroupe les valeurs résumées de RESUM1 dans $S1$ et des valeurs nulles. Pour l'exprimer, nous utilisons l'application injective $\varphi\,1$, introduite pour OUTERJOIN, afin d'associer à chaque tuple de $\mathrm{VAL}(F\,1\,l,\,S1)$, l'indice du tuple identique dans $\mathrm{VAL}(F\,1\,l,\,S1')$.

Chaque tuple $\tau'_{s'}$ de $\mathrm{VAL}(F\,1\,l,\,S1')$ structure l'ensemble de valeurs résumées : $[r'_{s'v}/v \in [1 \ldots \mathrm{Card}(F\,1\,c,\,S1')]]$.

Toute valeur résumée de ce vecteur est soit nulle, si elle est structurée par une combinaison de modalités ajoutée lors de la modification du multi-schéma ligne, soit égale à la valeur résumée de RESUM1, correspondant à la même combinaison de modalités dans l'extension de $S1$.

Si $\tau'_{s'} \notin \mathrm{VAL}(F\,1\,l,\,S1)$ alors $\forall v \in [1 \ldots \mathrm{Card}(F\,1\,c,\,S1')]$, $r'_{s'v} = \mathrm{NUL}$,

sinon $\forall v \in [1 \ldots \text{Card}(F\,1\,c, S1')]$, $r'_{s'v} = r_{sv}/r_{sv} \in \text{VAL}(\text{RESUM1}, S1)$ et $s' = \varphi\,1\,(\tau_s)$.

Enfin,

– $\text{MAT\_}f\,(S1') = \text{MAT\_}f\,(S1)$.

INTERSECTION : L'opérateur binaire INTERSECTION est analogue à UNION, mais il réalise l'intersection des ensembles de combinaisons de modalités définissant la classification de deux résumés et est donc un opérateur d'affinement. Les deux opérandes de INTERSECTION doivent être des TSC compatibles selon une dimension.

La syntaxe de l'opérateur est la suivante : INTERSECTION (d, $S1$, $S2$) où d est la dimension du multi-schéma compatible de $S1$ et $S2$.

Exemple 5.2 : Nous reprenons l'exemple du rapprochement des TSC Z1 et Z4, traité par l'opérateur JOIN dans l'exemple 3.3, dont le résultat est la TSC Z5, illustrée par la figure 12.

Z5 peut être obtenue par la séquence d'opérations suivantes : INTERSECTION (l, Z1, Z4) = Z1′ puis INTERSECTION (l, Z4, Z1) = Z4′. Les TSC obtenues peuvent alors être rapprochées par : CONCATENATE (l, Z1′, Z4′) pour produire la TSC Z5.

Soient deux TSC $S1$ et $S2$ de schéma $\text{SH}(S1) = (F\,1\,l, F\,1\,c, \text{RESUM1})$ et $\text{SH}(S2) = (F\,2\,l, F\,2\,c, \text{RESUM2})$, compatibles selon les lignes, avec : $\text{VAL}(F\,1\,l, S1) \neq \text{VAL}(F\,2\,l, S2)$.

INTERSECTION (l, S1, S2) = S1′, avec $\text{SH}(S1') = \text{SH}(S1)$ et :

– Si $\forall i \in [1 \ldots m]$

et

$$\forall k \in [1 \ldots u_i], \text{VAL}(\text{A}\,1\,l^i_k, S1) \cap \text{VAL}(\text{A}\,2\,l^i_k, S2) \neq \varnothing,$$

alors l'extension du multi-schéma $F\,1\,l$ est définie par :

$$\text{VAL}(F\,1\,l, S1') = \mathop{\copyright}_{i=1}^{m} \bigotimes_{k=1}^{u_i} (\text{VAL}(\text{A}\,1\,l^i_k, S1) \cap \text{VAL}(\text{A}\,2\,l^i_k, S2))$$

où $\cap$ respecte l'ordre sous-jacent dans l'extension des multi-schémas de $S1$ et $S2$.

S'il n'existe aucune modalité commune dans l'extension des attributs $\text{A}\,1\,l^i_k$ et $\text{A}\,2\,l^i_k$, la construction de l'extension du schéma d'attributs concerné, par

produit cartésien, conduit à un ensemble vide, ce qui implique, pour la ou les TS dotées de ce schéma, l'absence d'extension.

Pour éviter ce problème, si l'intersection des extensions de $A1\,l_k^i$ et $A1\,l_k^i$ est vide, l'extension du schéma Tl_i est définie par une unique combinaison de modalités, dont les éléments al^{ik} sont des valeurs nulles. Une telle combinaison est notée COMB_NUL.

– $\mathrm{VAL}(F\,1\,c, S1') = \mathrm{VAL}(F\,1\,c, S1)$;

– L'extension de RESUM1 dans $S1'$ s'exprime en fonction de celle de RESUM1 dans $S1$, par : $\forall v \in [1 \ldots \mathrm{Card}(F\,1\,c, S1')]$,
si $\tau'_{s'} \neq$ COMB_NUL où $\tau'_{s'} \in \mathrm{VAL}(F\,l1, S1')$,
$r'_{s'v} = r1_{sv} \in \mathrm{VAL}(\mathrm{RESUM1}, S1)$ et $s = \psi 1(\tau'_{s'})$, où $\psi 1$ associe à chaque tuple de l'extension de $F\,1\,l$ dans $S1'$, l'indice du tuple identique dans $\mathrm{VAL}(F\,1\,l, S1)$.

Si $\tau'_{s'} =$ COMB_NUL, deux cas doivent être envisagée :

• Si $\tau_s \neq$ COMB_NUL où $\tau_s \in \mathrm{VAL}(F\,l1), S1)$ et $s = \psi 1(\tau'_{s'})$,
alors la combinaison considérée a été engendrée par une intersection vide de l'extension d'attributs catégories, d'où :

$$\forall v \in [1 \ldots \mathrm{Card}(F\,1\,c, S')], \qquad r'_{s'v} = \mathrm{NUL},$$

• Si $\tau_s =$ COMB_NUL alors l'expression des valeurs résumées est identique au premier cas traité (où $\tau'_{s'} \neq$ COMB_NUL).

Enfin,

– $\mathrm{MAT\_}f(S1') = \mathrm{MAT\_}f(S1)$.

DIFFÉRENCE : L'opérateur binaire DIFFERENCE travaille sur des TSC compatibles selon une de leurs dimensions. Il produit, à partir de la première opérande, une nouvelle TSC, mais en ne considérant, pour chaque attribut catégorie, que les modalités ne faisant pas partie de l'extension de l'attribut compatible correspondant, dans la seconde opérande. Seules les valeurs résumées, ainsi structurées, sont conservées pour les attributs résumés considérés.

Syntaxiquement, DIFFERENCE admet comme arguments les identifiants de deux TSC et la dimension selon laquelle elles sont compatibles : DIFFERENCE(d, $S1$, $S2$).

Exemple 5.3 : Nous reprenons les TSC Z1 et Z4 et réalisons la différence de leur classification par : DIFFERENCE(l, Z4, Z1). Le résultat est la TSC Z16 (illustrée par la figure 22), compatible en ligne avec les deux opérandes.

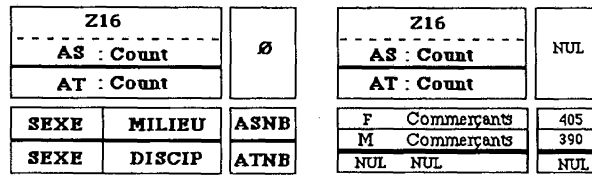

Figure 22. – Schéma et extension de la TSC $Z16$.

Soient deux TSC $S1$ et $S2$ compatibles selon les lignes mais non totalement, de schéma :

$$SH(S1) = (F1l, F1c, \text{RESUM1})$$

et

$$SH(S2) = (F2l, F2c, \text{RESUM2}).$$

DIFFERENCE (I, S1, S2) = S1', avec $SH(S1') = SH(S1)$ et :

– Si $\forall i \in [1 \ldots m]$ et $\forall k \in [1 \ldots u_i]$,

$$VAL(A1l_k^i, S1) - VAL(A2l_k^i, S2) \neq \varnothing,$$

où – est la différence ensembliste, alors l'extension du multi-schéma $F1l$ est définie par :

$$VAL(F1l, S1') = \mathop{\textcircled{c}}_{i=1}^{m} \mathop{\otimes}_{k=1}^{u_i} (VAL(A1l_k^i, S1) - VAL(A2l_k^i, S2))$$

où – respecte l'ordre sous-jacent dans l'extension des multi-schémas de $S1$ et $S2$.

Si $\exists i \in [1 \ldots m]$ et $\exists k \in [1 \ldots u_i] / VAL(A11_k^i, S1) - VAL(A21_k^i, S2) = \varnothing$, l'extension du schéma Tl_i est définie par une unique combinaison de modalités, dont les éléments al^{ik} sont des valeurs nulles :

$$VAL(Tl_i, S) = \text{COMB\_NUL}.$$

– $VAl(F1c, S1') = VAL(F1c, S1)$;

– L'extension de RESUM1 dans $S1'$ s'exprime en fonction de celle de RESUM1 dans $S1$, de manière totalement analogue à celle donnée pour l'opérateur INTERSECTION.

Enfin,

– $\mathrm{MAT\_}f(S1') = \mathrm{MAT\_}f(S1)$.

NRESTRICT: Cet opérateur de base unaire procède à l'élimination des modalités permettant de structurer un ensemble de valeurs résumées qui sont toutes des valeurs nulles, dans l'extension des attributs catégories correspondants. NRESTRICT élimine ainsi les tuples éventuellement rajoutés dans l'extension des multi-schémas par l'opérateur UNION ou un opérateur de composition, ainsi que les valeurs résumées nulles correspondantes. Il permet ainsi de ne conserver que les données significatives en « compactant » l'extension de la TSC opérande.

Syntaxiquement, le seul argument à fournir est l'identifiant de la TSC : NRESTRICT (S).

Exemple 5.4*:* Considérons la TS M, composant de la TSC $Z15$ (*fig.* 21). M est obtenue à partir de la TS A (*fig.* 1 et 2), en complétant l'extension de l'attribut MILIEU, par rapport aux valeurs qu'il prend dans la TS E de la TSC $Z4$ (*fig.* 11). Nous avons alors : NRESTRICT (M) = A.

Formellement, soit une TSC S, de schéma $\mathrm{SH}(S) = (\mathrm{F}l, \mathrm{F}c, \mathrm{RESUM})$.

NRESTRICT (S) = S', avec $\mathrm{SH}(S') = \mathrm{SH}(S)$, et :

– Lorsqu'une modalité d'un attribut catégorie permet de structurer des valeurs résumées, d'un ou plusieurs attributs résumés, qui sont des valeurs nulles, elle est éliminée par NRESTRICT, d'où :

$$\mathrm{VAL}(\mathrm{T}'l_i, S') \subseteq \mathrm{VAL}(Tl_i, S).$$

Une modalité al_q^{ik} de Al_k^i est conservée dans l'extension de l'attribut correspondant dans S', si et seulement s'il existe au moins une valeur résumée non nulle dans l'extension de la TSC résultat, structurée par la modalité considérée. L'ensemble des tuples de Tl_i, comprenant une modalités al_q^{ik}, peut être déterminé en fonction des cardinalités des différents attributs du schéma et de la position q de la modalité dans l'extension de l'attribut catégorie considéré.

Si la modalité al_q^{ik} est le q-ième, mais non le dernier, composant de tuple dans l'extension de Tl_i, l'indice de ces tuples, dans $\mathrm{VAL}(Tl_i, S)$, appartient aux intervalles :

$$\left[\alpha\left(\left(\sum_{y=k+1}^{u_i} \mathrm{Card}(Al_y^i, S)(q-1)\right)+1\right)\ldots\alpha\left(\left(\prod_{y=k+1}^{u_i} \mathrm{Card}(Al_y^i, S)\,q\right)+1\right)\right]$$

où :

$$\alpha = \beta\left(\prod_{y=k}^{u_i} \text{Card}\,(A\,l_y^i,\ S)\right)$$

avec β un coefficient entier variant dans l'intervalle :

$$\left[0 \ldots \prod_{y=1}^{k-1} \text{Card}\,(A\,l_y^i,\ S) - 1\right].$$

Si la modalité al_q^{ik} est la dernière dans l'extension de l'attribut $A\,l_k^i$, *i.e.* $q = \text{Card}\,(A\,l_k^i,\ S)$, alors les intervalles précédents sont définis par : $[\alpha q \ldots \alpha(q+1)]$.

Nous notons $\mathbb{T}_S(al_q^{ik}) = \{\tau_s / s \in [1 \ldots \text{NB}]\}$ l'ensemble des tuples de Fl dans S comprenant la modalité al_q^{ik} avec :

$$\text{NB} = \left(\prod_{y=1}^{k-1} \text{Card}\,(A\,l_y^i,\ S)\right)\left(\prod_{y=k+1}^{u_i} \text{Card}\,(A\,l_y^i,\ S)\right).$$

Alors $\mathbb{T}_S(\text{al}_q^{ik}) \in F'\,l$ si et seulement si,

$$\exists\, \tau_s \in \mathbb{T}_S(\text{al}_q^{ik}) / \exists\, r_{sv} \neq \text{NUL}$$

$$\text{où} \quad r_{sv} \in \text{VAL}(\text{RESUM},\ S) \qquad \text{et} \qquad v \in [1 \ldots \text{Card}\,(F'c,\ S')].$$

– L'expression de l'extension de $F'c$ dans S' par rapport à celle de Fc dans S est strictement analogue à celle du multi-schéma ligne.

– Pour exprimer l'extension de la matrice des attributs résumés de S' en fonction de celle de S, nous introduisons les applications injectives $\delta 1$ et $\delta 2$, associant à chaque tuple respectivement de $\text{VAL}\,(F'\,l,\ S')$ et de $\text{VAL}\,(F'c,\ S')$, leur indice dans $\text{VAL}\,(Fl,\ S)$ et $\text{VAL}\,(Fc,\ S)$.

$$\text{VAL}\,(F'\,l,\ S') \to [1 \ldots \text{Card}\,(Fl,\ S)] \text{ et } \delta 1\,(\tau_{s'}) = s;$$

$$\delta 2 : \quad \text{VAL}\,(F'c,\ S') \to [1 \ldots \text{Card}\,(Fc,\ S)] \qquad \text{et} \qquad \delta 2\,(\tau_{v'}) = v.$$

Alors, $\forall\, s' \in [1 \ldots \text{Card}\,(F'\,l,\ S')]$ et $\forall\, v' \in [1 \ldots \text{Card}\,(F'c,\ S')]$,

$$r_{s'v'} = r_{sv} \qquad \text{où} \quad r_{sv} \in \text{VAL}\,(\text{RESUM},\ \text{S}) \quad \text{et} \quad \delta 1\,(\tau_{s'}) = s; \qquad \delta 2\,(\tau_{v'}) = v.$$

Enfin,

– $\text{MAT\_}f\,(S') = \text{MAT\_}f\,(S)$.

CSELECT. – L'opérateur unaire CSELECT permet à l'utilisateur d'exprimer une qualification sur un attribut catégorie d'une TSC et élimine les valeurs ne la vérifiant pas, dans l'extension de l'attribut considéré. Ainsi, l'extension du multi-schéma concerné est, dans le résultat, un sous-ensemble des combinaisons de modalités initiales. Les valeurs résumées correspondant aux tuples éliminés sont supprimées.

La syntaxe de l'opérateur est : CSELECT (S, $\zeta(P_{dj}.A\,l_u^d)$) où S est l'identifiant de la TSC sur laquelle opérer et ζ est une condition simple exprimée sur le dernier attribut catégorie $A\,l_u^d$ de Tl_d dans S. L'ambiguïté possible sur l'identification de l'attribut catégorie est levée en le préfixant par le nom d'une des TS dont le schéma inclut $A\,l_u^d$.

Exemple 5.5 : L'utilisateur désire travailler sur la TSC $Z1$ (*fig.* 3 *et* 4) mais seules les disciplines Informatique et Mathématiques l'intéressent. Une sélection lui permet alors de restreindre l'extension de l'attribut DISCIP aux valeurs voulues :

CSELECT ($Z1$, C.DISCIP ≠ « Physique »). Le résultat de cette opération est la TSC $Z17$, présentée intensivement et extensivement par la figure 23.

Z17		AN_ET	AN_ET
A : Count	B : Sum		
AU : Count	AV : Sum	BOURSIER	
SEXE	MILIEU	ANB	BMNT_BOURSE
SEXE	DISCIP	AUNB	AVMNT_BOURSE

Z17		1	1	2	2		
A : Count	B : Sum						
AU : Count	AV : Count	O	N	O	N	1	2
F	Agriculteurs	15	14	9	12	23.625	11.700
F	Ouvriers	30	25	23	29	45.000	34.500
F	Cadres	2	5	1	5	1.600	1.000
M	Agriculteurs	10	9	3	5	13.220	4.560
M	Ouvriers	43	40	35	42	60.200	50.750
M	Cadres	7	8	5	8	7.700	5.500
F	Informatique	25	39	24	33	26.250	28.800
F	Mathématiques	18	21	15	20	27.000	17.900
M	Informatique	38	37	29	42	46.740	41.730
M	Mathématiques	17	29	19	28	23.630	28.880

Figure 23. – Schéma et extension de la TSC $Z17$.

La sélection de modalités s'effectue sur le dernier attribut catégorie dans un des schémas du multi-schéma ligne ou colonne de la TSC opérande (DISPLACE peut être utilisé si ce n'est pas le cas).

Soit une TSC S de schéma $\mathrm{SH}(S)=(Fl, Fc, \mathrm{RESUM})$, avec Tl_d un schéma non vide de $Fl : Tl_d=[A\,l_k^d/k\in[1\ldots u]]$.

CSELECT $(\mathbf{S}, \zeta(\mathbf{P_{dj}}.\mathbf{A\,l_u^d}))=\mathbf{S'}$ avec $\mathrm{SH}(S')=\mathrm{SH}(S)$, et

- $\forall i\in[1\ldots m]/i\neq d$, $\mathrm{VAL}(Tl_i, S')=\mathrm{VAL}(Tl_i, S)$.

Tous les attributs catégorie de Tl_d conservent leur cardinalité et extension dans $T'l_d$, sauf $A\,l_u^d$, dont certaines modalités peuvent être éliminées.

Nous notons : $\alpha=\{al_q^{du}/q\in[1\ldots\mathrm{Card}(A\,l_u^d)]$ et $\zeta(al_q^{du})$ VRAI$\}$ l'ensemble des valeurs de $A\,l_u^d$ répondant à la condition exprimée par l'utilisateur et qui constituent donc l'extension de cet attribut dans le résultat. Si $\alpha\neq\varnothing$, nous avons alors :

Si $u>1$, alors

$$\mathrm{VAL}(Tl_d, S')=\{|\tau_s|_{(u-1)}/\tau_s\in\mathrm{VAL}(Fl, S)\}\otimes\alpha$$

où :

$$s\in\left[\sum_{i=1}^{d-1}\mathrm{Card}(Tl_i, S)+1\ldots\sum_{i=1}^{d}\mathrm{Card}(Tl_i, S)+1\right].$$

Si $u=1$, alors $\mathrm{VAL}(T'l_d, S')=\alpha$.

Le cas particulier où $\alpha=\varnothing$ correspond à l'élimination de toutes les modalités d'un attribut catégorie.

Deux interprétations en sont possibles : une erreur dans la formulation de la condition, de la part de l'utilisateur, ou sa volonté d'éliminer l'attribut catégorie considéré. C'est cette deuxième hypothèse que nous retenons. Dans ce cas, la définition extensive de Tl_d dans S', donnée ci-dessus, produit en résultat un ensemble vide (puisque, si E et F sont deux ensembles : $E\otimes F=\varnothing\Leftrightarrow E=\varnothing$ OU $F=\varnothing$).

Or, nous considérons que l'utilisateur cherche à éliminer un attribut catégorie et non le schéma complet dans lequel intervient cet attribut. Pour produire un résultat sémantiquement valide, cette suppression doit se réaliser parallèlement à un calcul de nouvelles valeurs résumées, car elle correspond très exactement à une opération d'agrégation.

Ainsi, lorsque $\alpha=\varnothing$, l'opérateur CSELECT indiquée ci-dessus est strictement équivalente à : $\mathrm{AGGREGATE}(l, S, P_{dj})$.

- $\mathrm{VAL}(F'c, S')=\mathrm{VAL}(Fc, S)$;
- $\forall j\in[1\ldots n]$, $\forall i\in[1\ldots m]$ et $i\neq d$, $\mathrm{VAL}(R'_{ij}, S')=\mathrm{VAL}(R_{ij}, S)$.

Les attributs résumés R'_{dj} et R_{dj} $(j\in[1\ldots n])$ diffèrent par leur extension en raison de la modification apportée sur Tl_d par CSELECT, en effet, les valeurs

résumées structurées par des tuples supprimés dans VAL$(T l_d, S)$, ne sont pas conservées. Nous nous plaçons ici dans le cas général où $\alpha \neq \varnothing$, et utilisons la fonction ψl, introduite pour l'opérateur JOIN, associant à chaque tuple de l'extension de Fl dans S', l'indice du tuple identique dans VAL$(F l, S)$. Alors,

$$\forall s' \in [1 \ldots \text{Card}(F' l, S')],$$

$$\forall v \in [1 \ldots \text{Card}(Fc, S')] : \quad r'_{s'v} = r_{sv} \quad \text{avec } \psi 1(\tau'_{s'}) = s.$$

Enfin,

– MAT_$f(S') =$ MAT_$f(S)$.

RSELECT. Cet opérateur unaire joue un rôle similaire au précédent, mais il permet à l'utilisateur d'exprimer une condition de sélection sur un attribut résumé. Il n'intervient pas au niveau de la classification du résumé qui reste identique, mais il élimine les valeurs résumées visées en leur substituant des valeurs nulles, de manière à conserver la cardinalité des multi-schémas. Les valeurs résumés répondant à la qualification donnée, associées au tuples de modalités les structurant en ligne et en colonne, dont donc assimilées par cet opérateur à des combinaisons manquantes. La primitive NRESTRICT permet d'assurer qu'une élimination, même partielle, de ces combinaisons est possible, avant de la réaliser.

Pour cet opérateur, nous entendons « condition de sélection » au sens large, puisqu'elle peut, en effet, correspondre aux qualifications classiques (attribut, comparateur, constante), mais également à l'application des fonctions statistiques de recherche, comme minimum ou maximum. Dans ce dernier cas, l'extraction de valeurs particulières peut être considérée comme sous plusieurs aspects : une recherche globale rendant, pour l'attribut résumé concerné, une unique valeur, mais aussi des recherches plus détaillées, qui permettent d'isoler plusieurs valeurs différentes, pour des combinaisons de critères spécifiés. En effet, la recherche de valeurs minimales ou maximales peut s'effectuer à l'intérieur de blocs de valeurs résumées (*cf.* paragraphe 2.6), dont la taille peut atteindre les dimensions de la matrice définissant l'extension de l'attribut résumé (recherche globale) ou être fonction des cardinalités d'une partie des attributs catégories (recherche détaillée).

La syntaxe utilisée est la suivante : RSELECT$(S, \zeta(P_{db}.R_{db}), [\{A l_k^d\}])$ où ζ symbolise une condition de sélection, au sens large, sur les valeurs de l'attribut R_{db} de S. Lorsqu'elle inclut une fonction de recherche, le dernier argument (optionnel) permet à l'utilisateur de donner une liste des critères

(dont l'ordre est cohérent avec celui des schémas ligne et/ou colonne concernés), qui indique les combinaisons de modalités pour lesquelles les valeurs minimales ou maximales sont recherchées.

Exemple 5.6 : L'utilisateur désire identifier les différentes combinaisons des valeurs de sexe et milieu socio-professionnel, pour lesquelles les effectifs

Z18	
AW: Count	B : Sum
AU: Count	AV: Sum
SEXE	MILIEU
SEXE	DISCIP

AN_ET	AN_ET
BOURSIER	
AWNB	BMNT_BOURSE
AUNB	AVMNT_BOURSE

Z18		1	1	2	2		
AW : Count	B : Sum						
AU : Count	AV : Sum	O	N	O	N	1	2
F	Agriculteurs	NUL	NUL	9	NUL	23.625	11.700
F	Ouvriers	NUL	NUL	23	NUL	45.000	34.500
F	Cadres	NUL	NUL	1	NUL	1.600	1.000
M	Agriculteurs	NUL	NUL	3	NUL	13.220	4.560
M	Ouvriers	NUL	40	NUL	NUL	60.200	50.750
M	Cadres	NUL	NUL	5	NUL	7.700	5.500
F	Informatique	25	39	24	33	26.250	28.800
F	Mathématiques	18	21	15	20	27.000	17.900
M	Informatique	38	37	29	42	46.740	41.730
M	Mathématiques	17	29	19	28	23.630	28.880

Figure 24. – Schéma et extension de $Z18$.

sont les plus bas, dans la TSC $Z17$ (*fig.* 23), quelque soit l'année d'étude et pour les boursiers et non boursiers. Une réponse à sa requête peut être obtenue par : RSELECT(Z17, Min(A.ANB), { SEXE, MILIEU }), qui donne en résultat la TSC $Z18$.

Formellement, considérons la TSC S, de schéma SH(S) = ($F1$, Fc, RESUM).

RSELECT(S, ζ ($\mathbf{P_{db}} \cdot \mathbf{R_{db}}[\{ \mathbf{Al_k^d} \}]$) = S', avec SH($S'$) = SH($S$).

La seule modification du résultat S' par rapport à S concerne l'extension de la matrice d'attributs résumés et plus précisément de l'attribut R_{db}. Les valeurs de l'extension de RESUM dans S' sont ainsi déterminées :

$$\forall s \notin \left[\sum_{i=1}^{d-1} \text{Card}(Tl_i, S) \ldots \sum_{i=1}^{d} \text{Card}(Tl_i, S) \right];$$

$$\forall v \notin \left[\sum_{j=1}^{b-1} \text{Card}(Tc_j, S) \ldots \sum_{j=1}^{b} \text{Card}(Tc_j, S) \right], \qquad r'_{sv} = r_{sv};$$

Si ζ est une condition classique

$$\forall v \in \left[\sum_{i=1}^{d-1} \text{Card}\,(Tl_i, S) \ldots \sum_{i=1}^{d} \text{Card}\,(Tl_i, S) \right]$$

$$\forall v \in \left[\sum_{j=1}^{b-1} \text{Card}\,(Tc_j, S) \ldots \sum_{j=1}^{b} \text{Card}\,(Tc_j, S) \right]$$

Si ($\zeta(r_{sv})$ VRAI), alors $r'_{sv} = r_{sv}$, sinon $r'_{sv} = \text{NUL}$.

Si ζ inclut une fonction de recherche, les deux cas suivants doivent être considérés :

– aucun attribut catégorie n'est spécifié (recherche globale), alors :

$$\text{si } r_{sv} = \zeta(R_{db}), \quad \text{alors } r'_{sv} = r_{sv}, \quad \text{sinon } r'_{sv} = \text{NUL}.$$

– si une liste d'attributs catégories est indiquée (recherche détaillée) :

en notant $\{ A\, l^d_y / y \in [k1 \ldots ku] \}$ et $\{ A\, c^b_y / y \in [h1 \ldots hv] \}$ les critères spécifiés, l'extraction des valeurs $r_{sv} = \zeta(R_{db})$ est effectuée à l'intérieur de blocs dont les dimensions sont :

$$\left(\prod_{y=k1}^{ku} \text{Card}\,(A\, l^d_y, S), \prod_{y=h1}^{hv} \text{Card}\,(A\, c^b_y, S) \right).$$

Pour toutes les valeurs r_{sv} ne vérifiant pas $\zeta(R_{db})$, $r'_{sv} = \text{NUL}$.

6. CONCLUSION

Nous avons présenté, dans cet article, un langage de requête algébrique permettant la manipulation des résumés statistiques initialement créés à partir de relations et modélisés au travers d'une structure originale, celle des Tables Statistiques Complexes.

Un langage de manipulation de résumés statistiques doit impérativement mettre l'accent sur les possibilités de dérivation. Si de telles opérations constituent « le cœur » du langage STAR+, elles ne sont pas les seules. Nous leur adjoignons, en effet, la classe de composition/décomposition, tout aussi importante dès lors que la structure de données proposée vise à la représentation de résumés arbitrairement complexes. Bien que les opérateurs de transposition, d'affinement/élargissement aient une fin en eux-mêmes (présentation des résumés pour les premiers, interrogation pour les autres), ils jouent un rôle encore plus important lorsqu'ils sont combinés aux manipulations

« vitales », que sont les dérivations et les compositions, car ils élargissent leur portée de manière très significative. En effet, sans possibilité de transposition, l'organisation des Tables Statistiques Complexes deviendrait vite un « carcan » pour l'utilisateur, qui ne pourrait plus rapprocher deux résumés parfaitement comparables mais dont les attributs catégories seraient ordonnés différemment ou appartiendrait à des dimensions opposées. C'est, notamment, à ce type de contraintes que se heurtent les langages QBSRT, STAQUEL et STBE évoqués en introduction. En jouant un rôle équivalent à celui des transpositions, mais au niveau de l'extension des résumés, les manipulations d'affinement/élargissement permettent aussi d'amplifier le champ d'application des compositions et des calculs arithmétiques et statistiques, car elles permettent de rendre identiques des classifications homogènes.

Si le langage STAR$^+$ trouve sa vocation en particulier dans le domaine des bases de données statistiques, son champ d'application n'y est pas limité. Des domaines voisins comme ceux des bases de données classiques, des bases de données historiques ou encore des systèmes d'information et d'aide à la décision peuvent tirer profit de notre contribution. En effet, par l'interrogation de résumés, il est possible d'apporter une réponse rapide à un certain type de requête, demandant des informations agrégées et nécessitant traditionnellement le balayage d'une grande partie, voire de la totalité, de la base de données. Dans certaines applications, où la constitution d'un historique circonstancié n'est pas nécessaire, l'archivage d'informations synthétiques juducieuses s'avèrent beaucoup moins coûteux que le stockage d'importants volumes de données détaillées, tout en étant plus facilement exploitable. Enfin, il existe toute une classe d'utilisateurs, impliqués dans un processus de prise de décidion, qui ne trouvent pas, à travers les langages relationnels, une réponse à leurs besoins. Ces utilisateurs ne manipulent, généralement, les données que sous une forme agrégée. Le langage STAR$^+$ leur donne la possibilité de rechercher l'information « dense » et pertinente pour étayer, au mieux, leur choix, et leur permet de la manipuler de manière adaptée.

REMERCIEMENTS

Ce travail a été en partie financé par le Ministère de l'Éducation nationale, Direction de l'évaluation et de la prospective.

Nous tenons à exprimer tous nos remerciements à M. Jean-Paul DISPAGNE, Chef du Centre de l'Informatique Statistique et d'Aide à la Décision-C.I.S.A.D. du Ministère de l'Éducation nationale et M. Serge MIRANDA, Professeur à l'Université de Nice, pour leur aide amicale et leurs encouragements constants.

BIBLIOGRAPHIE

1. S. ABITEBOUL et S. GINSBURG, Tuple Sequences and Lexicographic Indexes, *Journal of the ACM*, 1986, *33*, p. 409-422.
2. S. ABITEBOUL et N. BIDOIT, Non First Normal Form Relation : an Algebra Allowing Data Restructuring, *Journal of Computer and System Sciences*, 1986, *33*, p. 361-393.
3. S. ABITEBOUL, P. C. FISCHER et H. J. SCHEK (Eds), Nested Relations and Complex Objects in Databases, *Lecture Notes in Computer Science*, 361, Springer-Verlag, 1989.
4. R. ADAM et J. C. WORTMANN, Security-Control Methods for Statistical Databases : A Comparative Study, *ACM Computing Survey*, 1989, *21*, p. 514-556.
5. M. ADIBA et C. COLLET, Management of Complex Objects as Dynamic Forms, *Proceedings of the International Conference on Very Large Databases*, 1988, p. 134-147.
6. F. BRY et G. THAURONT, Gestion interne de données statistiques, *Revue-AFCET, Modèles et Base de Données*, 1986, *4*, p. 25-38.
7. R. CICCHETTI, L. LAKHAL, N. LE THANH et S. MIRANDA, A Logical Summary-Data Model for Macro Statistical Databases, *Proceedings of the International Symposium on Database Systems for Advanced Applications*, 1989, p. 43-51.
8. M. C. CHEN et L. MCNAMEE, On the Data Model and Access Method of Summary Data Management, *I.E.E.E. Transaction on Knowledge and Data Engineering*, 1989, *1*, p. 519-529.
9. C. DELOBEL et M. ADIBA, Bases de données et systèmes relationnels, *Dunod*, Paris, 1982.
10. E. FORTUNATO, M. RAFANELLI, F. L. RICCI et A. SEBASTIO, An Algebra for Statistical Data, *Proceedings of the International Conference on Statistical and Scientific Database Management*, 1986, p. 122-134.
11. S. P. GHOSH, Statistical Relational Tables for Statistical Database Management, *I.E.E.E. Transaction on Software Engineering*, 1986, *12*, p. 1106-1116.
12. S. P. GHOSH, Statistical relational model, *Statistical and Scientific Database Management, Lecture Notes in Computer Science*, *339*, Springer-Verlag, 1988, p. 338-355.
13. S. P. GHOSH, Statisticians and statistical Database Management, I.B.M. Research Report n° RJ6975, San Jose, 1989.
14. G. HEBRAIL, Définition de Résumés et Incertitude dans les Grandes Bases de Données, *Thèse de Doctorat*, Université de Paris-Sud, 1987.
15. M. JARKE et Y. VASSILIOU, A Framework for Choosing a Database Query Language, *ACM Computing Surveys*, 1985, *17*, p. 313-340.
16. G. JOMIER, O. KEZOUIT et H. RALAMBONDRAINY, Data Analysis for Relational Databases : The Pepin-Sicla System, *Proceedings of the International Conference on Statistical and Scientific Database Management*, 1986, p. 211-218.
17. A. KLUG, Equivalence of Relational Algebra and Relational Calculus Query Languages Having Aggregate Functions, *ACM Transaction On Database Systems*, 1982, *29*, p. 699-717.
18. L. LAKHAL, R. CICCHETTI et S. MIRANDA, Complex-Statistical-Table Structure and Operators for Macro Statistical Databases, *Foundations of Data Organization and Algorithms. Lecture Notes in Computer Science*, Springer-Verlag, 1989, *367*, p. 421-438.

19. L. LAKHAL, R. CICCHETTI et S. MIRANDA, Relation and Table Language for Statistical Databases, *Mathematical Foundamentals of Database Systems, Lecture Notes in Computer Science*, Springer-Verlag, 1989, *364*, p. 285-300.

20. L. LAKHAL, R. CICCHETTI et S. MIRANDA, STAR, un langage de manipulation de bases de données statistiques, *Revue-AFCET, Modèles et Bases de données*, 1989, *12*, p. 3-20.

21. F. M. MALVESTUTO et C. ZUFFADA, Modelling Large Bases of Categorical Data with Acyclic Schemes, *Database Theory, Lecture Notes in Computer Science*, Springer-Verlag, 1986, *243*, p. 323-340.

22. F. M. MALVESTUTO, The Classification Problem with Semantically Heterogeneous Data, *Statistical and Scientific Database Management, Lecture Notes in Computer Science*, Springer-Verlag, 1989, *339*, p. 155-176.

23. Y. MASUNAGA, Object Identity, Equality and Relational Concept, *Proceeding of the International Conference on Deductive and Object Oriented Databases*, 1989, p. 170-187.

24. M. MCLEISH, Further Results on the Security of Partioned Dynamic Statistical Databases, *ACM Transaction On Database Systems*, 1989, *14*, p. 98-113.

25. F. OLKEN, D. ROTEM, A. SHOSHANI et H. K. T. WONG, Scientific and Statistical Data Management Research at LBL. *Proceedings of the International Conference on Statistical and Scientific Database Management*, 1986, pp. 1-20.

26. F. OLKEN, Physical Database Support for Scientific and Statistical Database Management, *Proceedings of the International Conference on Statistical and Scientific Database Management*, 1986, p. 44-60.

27. G. ÖZSOYOGLU et M. Z. M. ÖZSOYOGLU, Statistical Database Query Languages, *I.E.E.E. Transaction on Software Engineering*, 1985, *11*, p. 1071-1081.

28. G. ÖZSOYOGLU, V. MATOS et Z. M. ÖZSOYOGLU, Query Processing Techniques in the Summary-Table-By-Example Database Query Language, *ACM Transaction On Database Systems*, 1989, *14*, p. 526-573.

29. M. RAFANELLI et F. L. RICCI, STAQUEL : a quercy language for statistical macro database management systems, *Proceedings of the Convention Informatique Latine*, 1985, p. 16-19.

30. A. SHOSHANI, Statistical Databases : Characteristics, Problems and some Solutions, *Proceedings of the International Conference on Very Large Databases*, 1982, p. 147-160.

31. A. SHOSHANI et H. K. T. WONG, Statistical and Scientific Database Issues, *I.E.E.E Transaction on Software Engineering*, 1985, *11*, p. 1071-1081.

32. S. S. STEVEN, On the theory of scales measurement, *Science*, 1946, *103*, p. 677-680.

33. M. TURNER, R. HAMMOND et F. COTTOB, A DBMS for Large Statistical Databases, *Proceedings of the International Conference on Very Large Databases*, 1979, p. 319-327.

34. H. K. T. WONG, Micro and Macro Statistical/Scientific Database Management. *Proceedings of the I.E.E.E. International Conference on Data Engeneering*, 1984, p. 104-110.

ANNEXE

ÉQUIVALENCE ALGÉBRIQUE

Nous examinons, dans cette annexe, l'expression, en terme de primitive de manipulation, des opérateurs complémentaires définis sur les TSC *i.e.* : JOIN, OUTERJOIN, SAGGREGATE et INTERSECTION. Enfin, nous nous intéressons au processus de décomposition réversible d'une TSC en TS. L'aspect intensif de cette décomposition met en jeu les opérateurs CONCATENATE et PROJECT, pour son aspect extensif, nous avons recours aux primitives UNION et NRESTRICT.

$E1$: OUTERJOIN$(d, S1, S2)=$
CONCATENATE$(d,$ UNION$(d, S1,$ S2), UNION$(d, S2, S1)$).

$E2$: JOIN$(d, S1, S2)=$
CONCATENATE$(d,$ INTERSECTION$(d, S1, S2)$, INTERSECTION$(d, S2, S1)$).

$E3$: L'opérateur SAGGREGATE s'exprime grâce à la primitive de dérivation AGGREGATE.

$E4$: INTERSECTION$(d, S1, S2)=$DIFFERENCE$(d, S1,$ DIFFERENCE$(d, S1, S2)$).

Preuve de $E1$: Soit S la TSC résultat de la partie gauche de l'égalité $E1$, S' la TSC résultat de sa partie droite avec $S1'=$UNION$(d, S1, S2)$ et $S2'=$UNION$(d, S2, S1)$.

Par definition de l'opérateur OUTERJOIN, nous avons :

$$\mathrm{SH}(S')=(Fl, Fc, \mathrm{RESUM}) \quad \text{où} \quad Fl=F1\,l, \quad Fc=F1\,c \copyright \mathrm{F2c}$$

et

$$\mathrm{RESUM}=\mathrm{RESUM}\,1 \copyright_c \mathrm{RESUM}\,2.$$

Par définition des opérateurs UNION et CONCATENATE, nous avons :

$$\mathrm{SH}(S1')=\mathrm{SH}(S1) \quad \text{et} \quad \mathrm{SH}(S2')=\mathrm{SH}(S2) \quad \text{et} \quad \mathrm{SH}(S')=(F'l, F'c, \mathrm{RESUM}')$$

avec :

$$F'1=F1\,l, \qquad F'c=F1\,c \copyright F2\,c \quad \text{et} \quad \mathrm{RESUM}'=\mathrm{RESUM}\,1 \copyright_c \mathrm{RESUM}\,2,$$

d'où : $\mathrm{SH}(S)=\mathrm{SH}(S')$.

En poursuivant ce raisonnement au niveau extensif, nous avons, par définition des opérateurs mis en jeu :

$$\mathrm{VAL}(Fl, S)=\mathop{\copyright}_{i=1}^{m}\ \mathop{\copyright}_{k=1}^{u_i} \varphi(\mathrm{VAL}(A\,1\,l_k^i, S1) \cup \mathrm{VAL}(A\,2\,l_k^i, S2)).$$

$$\mathrm{VAL}(Fc, S)=\mathrm{VAL}(Fc\,1, S1) \copyright \mathrm{VAL}(Fc\,2, S2)$$

et

$$\mathrm{VAL}(\mathrm{RESUM}, S)=\mathrm{VAL}(\mathrm{RESUM}\,1, S1) \copyright_c \mathrm{VAL}(\mathrm{RESUM}\,2, S2).$$

$$\mathrm{VAL}(Fl, S)=\mathrm{VAL}(F'1, S'1)$$

$$=\mathop{\copyright}_{i=1}^{m}\ \mathop{\otimes}_{k=1}^{u_i} \varphi(\mathrm{VAL}(A\,1\,l_k^i, S1) \cup \mathrm{VAL}(A\,2\,l_k^i, S2)).$$

$$\mathrm{VAL}(Fc, S)=\mathrm{VAL}(F1\,c, S1) \copyright_c \mathrm{VAL}(F2\,c, S2)$$

et

$$\text{VAL}(\text{RESUM}, S) = \text{VAL}(\text{RESUM}\,1, S1) \copyright_c \text{VAL}(\text{RESUM}\,2, S2).$$

D'où $\text{VAL}(S) = \text{VAL}(S')$.

Comme

$$\text{MAT\_}f(S) = \text{MAT\_}f(S1) \copyright_c \text{MAT\_}f(S2) = \text{MAT\_}f(S'),$$

nous avons : $S = S'$.

Preuve de E2 : Soit S la TSC résultat de la partie gauche de l'égalité $E2$, S' la TSC résultat de sa partie droite avec

$$S1' = \text{INTERSECTION}(d, S1, S2) \quad \text{et} \quad S2' = \text{INTERSECTION}(d, S2, S1).$$

Par définition de l'opérateur JOIN, nous avons :

$$\text{SH}(S') = (F1, Fc, \text{RESUM}) \quad \text{où} \quad Fl = F1\,l,\ Fc = F1\,c \copyright \text{F2c}$$

et

$$\text{RESUM} = \text{RESUM}\,1 \copyright_c \text{RESUM}\,2.$$

Par définition des opérateurs INTERSECTION ET CONCATENATE, nous avons :

$$\text{SH}(S1') = \text{SH}(S1) \quad \text{et} \quad \text{SH}(S2') = \text{SH}(S2)$$

et

$$\text{SH}(S') = (F'l, F'c, \text{RESUM}') \quad \text{avec} \quad F'l = F1\,l,$$
$$F'c = F1\,c \copyright \text{F2c} \quad \text{et} \quad \text{RESUM}' = \text{RESUM}\,1 \copyright_c \text{RESUM}\,2.$$

D'où : $\text{SH}(S) = \text{SH}(S')$.

En poursuivant ce raisonnement au niveau extensif, nous avons par définition des opérateurs mis en jeu :

$$\text{VAL}(Fl, S) = \mathop{\copyright}_{i=1}^{m} \bigotimes_{k=1}^{u_i} (\text{VAL}(A1\,l_k^i, S1) \cap \text{VAL}(A2\,l_k^i, S2)).$$

$$\text{VAL}(Fc, S) = \text{VAL}(Fc1, S1) \copyright_c \text{VAL}(Fc2, S2)$$

et

$$\text{VAL}(\text{RESUM}, S) = \text{VAL}(\text{RESUM}\,1, S1) \copyright_c \text{VAL}(\text{RESUM}\,2, S2).$$

$$\text{VAL}(Fl, S) = \text{VAL}(F'l, S'1) = \mathop{\copyright}_{i=1}^{m} \bigotimes_{k=1}^{u_i} (\text{VAL}(A1\,l_k^i, S1) \cap \text{VAL}(A2\,l_k^i, S2)).$$

$$\text{VAL}(Fc, S) = \text{VAL}(Fc1, S1) \copyright_c \text{VAL}(Fc2, S2)$$

et

$$\text{VAL}(\text{RESUM}, S) = \text{VAL}(\text{RESUM}\,1, S1) \copyright_c \text{VAL}(\text{RESUM}\,2, S2).$$

D'où $\text{VAL}(S) = \text{VAL}(S')$.

Comme

$$\text{MAT\_}f(S) = \text{MAT\_}f(S1) \copyright_c \text{MAT\_}f(S2) = \text{MAT\_}f(S'),$$

nous avons : $S = S''$.

Preuve de E3 : Considérons l'opération suivante : SAGGREGATE$(l.\ S,\ P_{il,j}) = S'$ où :

$$\mathrm{SH}(S) = (Fl,\ Fc,\ \mathrm{RESUM}) \quad \text{et} \quad \mathrm{SH}(S') = (F'l,\ F'c,\ \mathrm{RESUM}')$$

avec

$$F'l = [T'l_i / i \in [1 \ldots m]] \quad \text{et} \quad \forall i \in [1 \ldots m],$$
$$i \neq il, \quad T'l_i = Tl_i; \quad R'_{ij} = R_{ij} \quad \text{et} \quad T'l_{il} = \varnothing; \quad R'_{ij} = R_{ij};$$

Considérons à présent la dérivation, effectuant une agrégation successive de tous les attributs $A\,l_k^i\,(k \in [1 \ldots ui])$ de Tl_i, définie de manière récurrente par :

AGGREGATE$(l,\ S,\ P_{ij} \cdot A\,l_{ui}^i) = \mathrm{RESULT}_1$,

AGGREGATE$(l,\ \mathrm{RESULT}_1,\ \mathrm{P}_{ij} \cdot \mathrm{A}\,l_{ui-1}^i) = \mathrm{RESULT}_2$,

$k[2 \ldots ui-1]$ AGGREGATE$(l,\ \mathrm{RESULT}_k,\ \mathrm{P}_{ij} \cdot \mathrm{A}\,l_{ui-k}^i) = \mathrm{RESULT}_{k+1}$.

Nous notons S'' la TSC résultat RESULT_{ui}, son schéma est :

$$\mathrm{SH}(S'') = (F''l,\ Fc,\ \mathrm{RESUM}'') \quad \text{avec } F''l = [T''l_i / i \in [1 \ldots m]];\ \forall i \in [1 \ldots m],$$
$$i \neq i1,\ T''l_i = Tl_i; \qquad R''_{ij} = R_{ij} \qquad \text{et} \qquad T''l_{i1} = \varnothing; \qquad R''_{ij} = R_{ij};$$

Nous avons alors : $\mathrm{SH}(S') = \mathrm{SH}(S'')$, $\mathrm{VAL}(S') = \mathrm{VAL}(S'')$ et $\mathrm{MAT\_}f(S') = \mathrm{MAT\_}f(S'')$.

D'où $S' = S''$.

Preuve de E4 : Soit S le résultat de la partie gauche de l'égalité E4, S' celui de sa partie droite et $S'1$ la TSC obtenue par : DIFFERENCE $(d,\ S1,\ S2)$.

Par définition des opérateurs, nous avons : $\mathrm{SH}(S) = \mathrm{SH}(S1)$ et $\mathrm{SH}(S') = \mathrm{SH}(S1)$. S et S' sont donc dotées d'un même schéma.

Au niveau extensif, en considérant la dimension d comme étant celle des lignes, l'extension du multi-schéma dans S et S' s'exprime par :

$$\mathrm{VAL}(Fl,\ S) = \{\tau_s / \tau_s \in (\mathrm{VAL}(F1\,l,\ S1) \cap \mathrm{VAL}(F2\,l,\ S2))\}.$$
$$\mathrm{VAL}(F'1\,l,\ S'1) = \{\tau_s / \tau_s \in (\mathrm{VAL}(F1\,l,\ S1) - \mathrm{VAL}(F2\,l,\ S2))\}.$$
$$\mathrm{VAL}(F'l,\ S') = \{\tau_s / \tau_s \in (\mathrm{VAL}(F1\,l,\ S1) - \mathrm{VAL}(Fl1,\ S'1))\}$$
$$= \{\tau_s / \tau_s \in (\mathrm{VAL}(F1\,l,\ S1) - (\mathrm{VAL}(F1\,l,\ S1) - \mathrm{VAL}(Fl2,\ S2)))\}$$

Or, l'intersection ensembliste s'exprime en fonction de la différence, d'où :

$$\mathrm{VAL}(F'l,\ S') = \{\tau_s / \tau_s \in (\mathrm{VAL}(F1\,l,\ S1) \cap \mathrm{VAL}(F2\,l,\ S2))\}.$$

Comme $\mathrm{MAT\_}f(S') = \mathrm{MAT\_}f(S1) = \mathrm{MAT\_}f(S)$, nous avons $S = S'$.

Nous nous intéressons maintenant au processus de décomposition d'une TSC en l'ensemble ordonné de ses TS composants, grâce à PROJECT, puis de composition de ces dernières, par CONCATENATE, permettant de retrouver la TSC initiale. Enfin, il est également important de montrer que l'extraction d'une TS, à partir d'un résumé initialement obtenu en composant cette TS avec d'autres, avec au préalable la mise en totale cohérence des TS concernées par UNION, permet, en ayant recours à l'opérateur NRESTRICT de retrouver la TS initiale.

Décomposition et composition de TSC

Considérons une TSC S, de schéma $\mathrm{SH}(S) = (Fl,\ Fc,\ \mathrm{RESUM})$

L'obtention des différentes TS est réalisée par $(m \times n)$ opérations PROJECT, qui produisent les TS : P_{ij} de schéma $\mathrm{SH}(P_{ij}) = (Tl_i,\ Tc_j,\ R_{ij})$.

Pour reconstruire une TSC, à partir de ces TS, nous fixons j et appliquons $(m-1)$ fois l'opérateur CONCATENATE sur toutes les TS P_{ij}, i variant de 1 à m. Le résultat de cette

séquence est une TSC, appelée S'_j, dont la matrice des identifiants de TS correspond à la j-ième colonne de celle de S.

L'obtention de S'_j est définie de manière récurrente comme suit :

$$\text{CONCATENATE}(c, P_{1j}, P_{2j}) = \text{RESULT}_1$$

$$\text{CONCATENATE}(c, \text{RESULT}_1, P_{3j}) = \text{RESULT}_2$$

$$\forall i \in [2 \ldots (m-1)], \quad \text{CONCATENATE}(c, \text{RESULT}_{i-1}, P_{(i+1)j}) = \text{RESULT}_i.$$

avec $\text{RESULT}_m = S'_j$.

De plus, par définition de l'opérateur CONCATENATE, la matrice des identifiants de TS de S'_j est réduite au vecteur : $\text{MAT\_TS}(S'_j) = [P_{ij},\ i \in [1 \ldots m]]$.

Les n S'_j, $j \in [1 \ldots n]$, sont alors concaténées selon les lignes, en suivant le même processus :

$$\text{CONCATENATE}(1, S'_1, S'_2) = \text{RESULT}'_1$$

$$\text{CONCATENATE}(1, \text{RESULT}'_1, S_3) = \text{RESULT}'_2$$

$$\forall j \in [2 \ldots (n-1)], \quad \text{CONCATENATE}(1, \text{RESULT}'_{j-1}, S_{j+1}) = \text{RESULT}'_j$$

En notant S' le résultat obtenu RESULT'_n, nous avons :

$$\text{MAT\_TS}(S') = [P_{ij} / i \in [1 \ldots m] \text{ et } j \in [1 \ldots n]],$$

et donc : $\text{MAT\_TS}(S') = \text{MAT\_TS}(S)$, d'où $S' = S$.

Complément et réduction de TS

D'un point de vue extensif, nous supposons qu'il existe une TS P_{ij} telle que :

$\text{VAL}(Tl_i, S) \neq \text{VAL}(Tl_i, P_{ij})$. Cela signifie que lors de la concaténation de P_{ij} avec les autres TS composants de S, il y a eu ajout de combinaisons manquantes. Autrement dit :

$$\forall \tau_s \in \text{VAL}(Tl_i, S) \quad \text{si} \quad \tau_s \notin \text{VAL}(Tl_i, P_{ij}),$$

alors :

$$\forall v \left[\sum_{y=1}^{j-1} \text{Card}(Tc_y, S) + 1 \ldots \sum_{y=1}^{i} \text{Card}(Tc_y, S) \right], \qquad r_{sv} = \text{NUL}.$$

Or, par définition de l'opérateur NRESTRICT, de tels tuples sont éliminés de l'extension de Tl_i et les valeurs résumées nulles, dans l'extension de l'attribut résumé correspondant sont également supprimées. Nous avons donc :

$$\text{NRESTRICT}(\text{PROJECT}(S, P_{ij})) = P_{ij}.$$