Chapitre 5

Explication dans les cubes de données par règles d'association

> " *Les petits faits inexpliqués contiennent toujours de quoi renverser toutes les explications des grands faits.* "
>
> Paul Valéry, "*Tel quel*"

5.1 Introduction

Avec les techniques développées pour la construction des cubes de données, les utilisateurs OLAP sont largement capables d'explorer les données multidimensionnelles, de naviguer dans les niveaux hiérarchiques des dimensions et d'en extraire des informations intéressantes selon plusieurs niveaux de granularité. Cependant, la technologie OLAP se limite à des tâches exploratoires et ne fournit pas d'outils automatiques pour expliquer les relations et les associations potentiellement existantes entre les données d'un cube.

Par exemple, un utilisateur peut noter, à partir d'un cube de données de ventes, que le niveau de vente des *sacs de couchage* est particulièrement élevé dans une ville donnée. En revanche, cette exploration ne permet par d'expliquer automatiquement les raisons de ce fait particulier. En effet, pour arriver à expliquer l'ordres de certains faits OLAP ou des phénomènes particuliers, un utilisateur est habituellement supposé explorer manuellement et observer l'ensemble des données selon plusieurs axes d'analyse. Par exemple, le niveau élevé des ventes des *sacs de couchage* peut s'expliquer par son association à une *saison estivale* et à une *clientèle relativement jeune*.

Dans les dernières années, beaucoup d'études ont abordé le problème de l'extraction des règles d'association à partir des cubes de données. Kamber *et al.* affirment

l'importance de l'exploration des cubes de données en employant les algorithmes de recherche des règles d'association [KHC97]. En plus, les auteurs considèrent que la structure multidimensionnelle des données, avec ses agrégats pré-calculés, représente un terrain favorable pour cette recherche. Imieliński *et al.* défendent le même point de vue et considère que l'OLAP est étroitement lié avec les règles d'association [IKA02]. Ils pensent également que l'extraction des connaissances est un objectif commun à la technologie OLAP et la recherche des règles d'association. Dans [GC98b], Goil et Choudhary avancent que, d'une manière générale, les techniques de fouille de données couplées avec l'OLAP peuvent rendre un système d'aide à la décision plus utile et plus facile à exploiter. La recherche de règles d'association est particulièrement intéressante dans les données multidimensionnelles. Elle peut également interagir avec les outils OLAP afin d'extraire automatiquement des connaissances à partir des cubes de données. En effet, les agrégats de comptage (COUNT) nécessaires pour la recherche des règles d'association sont déjà pré-calculés dans un cube de données. En plus, les hiérarchies des dimensions du cube peuvent être exploitées afin d'extraire des règles à plusieurs niveaux de granularité.

Dans ce chapitre, nous mettons en place une méthode pour l'*explication* des données multidimensionnelles en utilisant les règles d'association. Cette nouvelle proposition de couplage entre l'analyse en ligne et la fouille de données se base sur une approche qui adapte plutôt l'algorithme de la fouille aux données multidimensionnelles. Ainsi, nous introduisons un nouvel algorithme pour la recherche des règles d'association directement à partir des cubes de données sans transformation préalable de ce dernier. En effet, comme le montre l'aperçu de notre méthode dans la figure 6.5, la recherche des règles d'association se fait directement à partir du cube étudié et ne requiert pas de traitement sur les données de ce dernier.

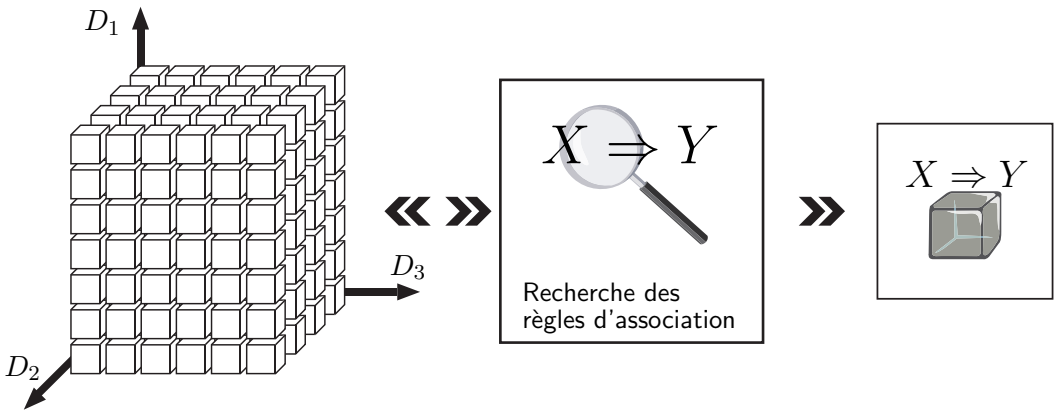

FIG. 5.1 – Étapes de l'explication dans les cubes de données par règles d'association

Nous établissons un cadre général pour la recherche de règles d'association à partir des cubes de données. Nous employons dans ce cadre le concept des *méta-règles inter-dimensionnelles* en vue d'offrir à l'utilisateur la possibilité de guider le processus de fouille vers des contextes d'analyse ciblés qui répondent à ses besoins d'explication et à partir desquels seront extraites les règles d'association.

Nous proposons également une redéfinition du support et de la confiance d'une règle d'association en y intégrant la mesure OLAP en vue de les adapter au contexte de l'analyse en ligne des données multidimensionnelles. Dans cette nouvelle définition, un support et une confiance d'une règle ne sont plus évalués selon la fréquence des faits qui les supportent, mais selon les unités de masse évaluées par une mesure OLAP. En plus du support et de la confiance, nous employons deux autres critères descriptifs (le *Lift* et l'indice de *Loevinger*) pour évaluer l'intérêt des règles d'association découvertes. Afin de valoriser les connaissances induites par les règles extraites, nous mettons en place un formalisme basé sur le principe de la *sémiologie graphique* de Bertin [Ber67] pour la visualisation des règles d'association. Ce formalisme permet d'intégrer des codages graphiques des règles extraites dans l'espace de représentation du cube étudié.

Ce chapitre est organisé comme suit. Dans la section 5.2, nous exposons une vue d'ensemble des travaux qui abordent le problème de découverte des règles d'association dans les données multidimensionnelles, suivie d'une étude comparative. Nous introduisons dans la section suivante, une formalisation des concepts que nous utilisons dans notre approche. La section 5.4 présente la notion de *méta-règles inter-dimensionnelles*. Nous revisitons la définition du support et de la confiance d'une règle d'association dans la section 5.5. La section suivante introduit le critère du *Lift* et l'indice de *Loevinger* pour une évaluation des règles extraites. Nous décrivons l'algorithme que nous proposons pour l'extraction des règles d'association inter-dimensionnelles dans la section 5.7. Le formalisme de la visualisation des règles extraites est présenté dans la section 5.8. La section suivante fait l'objet d'expérimentations empiriques afin de mesurer l'efficacité de notre proposition. Enfin, dans la section 5.10, nous concluons ce chapitre et proposons de nouvelles pistes de recherche pour notre proposition.

5.2 Extraction des règles d'association à partir des cubes de données

5.2.1 Historique des règles d'association

Le concept des règles d'association a été introduit la première fois par Agrawal *et al.* [AIS93]. Motivés par le problème de l'analyse du panier de la ménagère, les auteurs ont établi les premières bases d'un processus d'extraction de règles d'association. Ils sont aussi à l'origine de l'algorithme Apriori qui se base essentiellement sur *la propriété*

d'anti-monotonie, selon laquelle tout motif comprenant un sous-motif non fréquent est non fréquent. Depuis, les algorithmes d'extraction des règles d'association ont connu plusieurs évolutions. Ces évolutions couvrent diverses aspects parmi lesquels nous nous intéressons particulièrement à ceux liés aux types et aux structures des données.

La première génération des règles d'association d'Agrawal *et al.* [AIS93] concernait des données booléennes de transactions, où chaque produit (*item*) est codé selon sa présence ou son absence dans une transaction de vente. L'idée de base d'un algorithme d'extraction de règles, notamment Apriori, consiste à découvrir des relations intéressantes entre les produits qui s'achètent le plus souvent ensemble. Certaines références dans le domaine de la fouille de données parlent carrément de *règles d'association booléennes*. Un grand nombre de variantes de l'algorithme Apriori, travaillant toujours sur des données booléennes, ont été largement étudiées dans la littérature [AS94, MTV94, PCY95, SON95, Toi96].

L'extension des règles aux données quantitatives a été proposée pour la première fois par Srikant et Agrawal dans [SA96]. L'objectif de cette proposition consistait à extraire une nouvelle génération de *règles d'association quantitatives* à partir des tables d'une base de données relationnelle. Pour cela, les auteurs proposent une phase de pré-traitement qui discrétise les données quantitatives en variables qualitatives et les transforme ensuite en données booléennes selon un codage binaire. Suite à cette extension, beaucoup de travaux se sont basés sur les règles d'association quantitatives afin de les exploiter et de les étendre davantage pour couvrir des données de différentes natures liées à des domaines d'application spécifiques. On cite par exemple, l'étude des effets de causalité dans les données [BMS97, SBMU98], l'étude de phénomènes cycliques [ORS98, RMS98] ou de périodicités partielles [HDY99] dans des données temporelles. Pour un exposé plus complet sur les différents types de règles d'association quantitatives, nous renvoyons le lecteur à [Zhu98].

Toutes ces approches de règles d'association traitent des données se présentant selon des structures tabulaires. À notre connaissance, Kamber *et al.* [KHC97] sont les premiers à faire de la fouille de règles d'association dans les structures multidimensionnelles des cubes de données. Dans la suite, nous exposons en détail cette proposition ainsi que d'autres travaux qui traitent du sujet des règles d'association dans les structures multidimensionnelles des données.

5.2.2 Règles d'association dans les structures multidimensionnelles

Fouille guidée des règles d'association

Dans [KHC97], Kamber *et al.* ont introduit la fouille guidée des règles d'association dans les bases de données multidimensionnelles (*metarule-guided mining*). Cette

proposition consiste à utiliser une méta-règle qui va piloter le processus d'extraction pour la découverte de règles intéressantes répondant aux besoins d'analyse de l'utilisateur. Une méta-règle est un modèle général qui définit le contenu des règles d'association recherchées à partir d'un cube de données. Les auteurs définissent une méta-règle générale selon la forme :

$$P_1 \wedge P_2 \wedge \cdots \wedge P_m \Rightarrow Q_1 \wedge Q_2 \wedge \cdots \wedge Q_l$$

où P_i $(i = 1, \ldots, m)$ et Q_j $(j = 1, \ldots, l)$ sont des prédicats ou des instances de prédicats définit par l'utilisateur à partir des modalités du cube de données. Les auteurs affirment que la fouille guidée réduit l'espace de recherche dans le cube et permet de focaliser le processus d'extraction sur des régions de données ciblées par l'utilisateur. Ainsi, les règles d'association extraites répondent mieux aux attentes d'analyse de l'utilisateur. Quant à la structure multidimensionnelle des données, Kamber *et al.* confirment que la structuration des données dans un entrepôt et les agrégats pré-calculés d'un cube se prêtent au processus d'extraction de règles d'association.

Les auteurs proposent deux familles d'algorithmes d'extraction de règles à partir des cubes de données : (1) des algorithmes pour les cubes de données MOLAP matérialisés dont les agrégats sont tous pré-calculés (multi-D-slicing et n-D cube search) ; (2) des algorithmes pour les cubes de données ROLAP non matérialisés et dont les agrégats ne sont pas pré-calculés (abridged n-D cube construction et abridged multi-p-D cube construction). Tous ces algorithmes se basent sur la propriété d'anti-monotonie d'Apriori.

Analyse en ligne des règles d'association

Zhu distingue dans [Zhu98] trois types de règles d'association qui peuvent être extraites à partir d'un cube de données : les règles *inter-dimensionnelles*, les règles *intra-dimensionnelles* et les règles *hybrides*. À la différence de l'approche de Kamber *et al.* [KHC97] – où les règles sont extraites directement de la structure multidimensionnelle des données – Zhu aplatit le cube et le transforme selon une forme tabulaire appropriée, recherche les motifs fréquents en utilisant Apriori et génère ensuite les règles d'association.

Par exemple, supposons qu'un utilisateur souhaite découvrir des règles d'association inter-dimensionnelles dans un cube de *ventes* selon trois dimensions : *Lieu*, *Produit* et *Temps*. Dans ce cas, les faits du cube sont aplatis en fonction de ces trois dimensions comme le montre l'exemple du tableau 5.1.

Un motif inter-dimensionnel consiste en une conjonction de plusieurs modalités où chaque modalité provient d'une dimension distincte. Par exemple {*USA*, *DV-400*, *2002*} est un motif (*3-itemset*) inter-dimensionnel dans le tableau 5.1. Pour calculer le support de ce motif, Zhu prend en considération le nombre d'occurrences de ce dernier

Lieu	Produit	Temps	COUNT
Canada	iTwin	2002	30
Canada	iTwin	2003	10
Canada	aStar	2002	30
France	iPower	2005	20
France	DV-400	2005	85
France	DV-400	2004	25
France	EN-700	2006	25
France	EN-700	2003	20
USA	DV-400	2002	100
USA	iTwin	2005	20
USA	iTwin	2002	40
USA	aStar	2004	25
Japon	DV-400	2006	10
Japon	iTwin	2004	20
Japon	EN-700	2006	20

TAB. 5.1 – Aplatissement d'un cube de données pour l'extraction de règles inter-dimensionnelles [Zhu98]

fourni par l'agrégation COUNT. Si le motif est fréquent (son support est supérieur au support minimum), il peut ainsi générer les règles d'association inter-dimensionnelles suivantes :

$$\begin{array}{ll} \textit{USA} \wedge \textit{DV-400} \Rightarrow \textit{2002} & confiance = 1/1 = 100\% \\ \textit{USA} \wedge \textit{2002} \Rightarrow \textit{DV-400} & confiance = 1/2 = 50\% \\ \textit{DV-400} \wedge \textit{2002} \Rightarrow \textit{USA} & confiance = 1/1 = 100\% \end{array}$$

Un motif intra-dimensionnel est une conjonction de plusieurs modalités provenant d'une même dimension. Zhu considère qu'un processus d'extraction de règles d'association intra-dimensionnelles fait intervenir deux dimensions du cube : une première pour générer les modalités de la règle et une deuxième de regroupement, appelée *dimension de transaction*, dont les modalités sont considérées comme des identifiants de *transactions*. Dans le cube des ventes, on peut considérer par exemple la dimension *Produit* pour les éléments (items) des transactions regroupés selon les modalités de la dimension *Lieu*. Ainsi, l'auteur construit une table de transactions selon l'exemple du tableau 5.2 et cherche ensuite les motifs fréquents et les règles d'association intra-dimensionnelles à partir de cette table. Supposons que dans cet exemple, le motif $\{\textit{DV-400}, \textit{iTwin}, \textit{aStar}\}$ est un 3-itemset fréquent. À partir de ce motif, on peut obtenir les règles d'association intra-dimensionnelles suivantes :

ID transaction (Lieu)	Produit
Canada	iTwin, aStar
France	iPower, DV-400, EN-700
USA	DV-400, iTwin, aStar
Japon	DV-400, iTwin, EN-700

TAB. 5.2 – Aplatissement d'un cube de données pour l'extraction de règles intra-dimensionnelles [Zhu98]

$$\begin{array}{ll} DV\text{-}400 \wedge iTwin \Rightarrow aStar & confiance = 2/2 = 100\% \\ DV\text{-}400 \wedge aStar \Rightarrow iTwin & confiance = 2/2 = 100\% \\ iTwin \wedge aStar \Rightarrow DV\text{-}400 & confiance = 2/3 = 67\% \end{array}$$

Les règles d'association hybrides sont des combinaisons de règles inter et intra-dimensionnelles. Ainsi, une règle hybride consiste en un ensemble de modalités à la fois répétitives et provenantes de plusieurs dimensions. Dans ce cadre, un motif candidat L peut s'écrire d'une manière générale sous la forme d'une conjonction $L = \{L_{inter} \wedge L_{intra}\}$, où L_{inter} est un motif inter-dimensionnel et L_{intra} est un motif intra-dimensionnel. Pour trouver les motifs hybrides fréquents, l'auteur propose de chercher les motifs fréquents inter et intra-dimensionnels séparément, puis de fusionner les deux.

Cubes de données différentielles

Imieliński *et al.* proposent, dans un contexte OLAP, une approche de généralisation des règles d'association appelée Cubegrades [IKA02]. Un cubegrade est un formalisme qui calcule le différentiel d'une mesure agrégée d'un cube de données suite à des opérations de spécialisation (*drill-down*), de généralisation (*roll-up*) ou de changement de modalité dans une dimension (*switch*). Les auteurs reprochent aux règles d'association classiques de n'exploiter que les comptages – correspondant à la mesure COUNT dans un contexte OLAP – dans l'évaluation de l'implication existante entre l'antécédent et le conséquent d'une règle. Ils proposent d'exploiter dans les cubegrades d'autres agrégations de mesures. Formellement, un cubegrade est défini selon une implication de la forme générale :

$$\text{Cube source} \Rightarrow \text{Cube cible } [\textit{Mesures}, \textit{Valeurs}, \textit{Delta-valeurs}]$$

Cube source et le Cube cible représentent deux configurations de données du même cube où la deuxième configuration est obtenue à partir de la première suite à une des opérations sus-citées. Par exemple, comme le montre la figure 5.2, à partir d'une configuration source ($A = a_1$, $B = b_1$, $C = c_1$), le cube peut changer par :

(i) généralisation par agrégation de toute la dimension C ; on obtient alors le cube cible ($A = a_1$, $B = b_1$) ; (ii) spécialisation par rajout d'une nouvelle dimension D qui prend une modalité d_1 ; on obtient alors le cube cible ($A = a_1$, $B = b_1$, $C = c_1$, $D = d_1$) ; ou par (iii) mutation par changement de la modalité c_1 par c_2 dans la dimension C ; on obtient alors le cube cible ($A = a_1$, $B = b_1$, $C = c_2$).

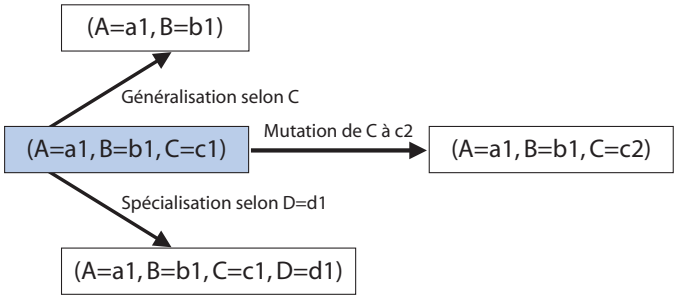

FIG. 5.2 – Opérations possibles dans un cubegrade [IKA02]

Mesures correspondent à un ensemble d'une ou de plusieurs mesure agrégées selon les fonctions SUM, AVG, MAX et MIN. Par exemple, à partir d'un cube de ventes, AVG(*Bénéfice*) permet d'agréger la mesure *Bénéfice* en calculant sa moyenne. *Valeurs* correspondent à l'ensemble des valeurs que prennent les mesures agrégées dans la configuration du cube source. *Delta-valeurs* mesurent les différentiels des valeurs des mesures agrégées entre le cube cible et le cube source. Pour résumer cette approche, considérons l'exemple du cubegrade suivant :

(Lieu=France) ⇒ (Lieu=France, Temps=2005)
[AVG(Bénéfice), AVG(Bénéfice) = $ 40 000, DeltaAVG(Bénéfice) = 80%]

Cet exemple signifie que la moyenne des bénéfices générés par les ventes en France, évalués à $ 40 000, enregistrent une baisse de 20% pendant l'année 2005.

Imieliński *et al.* affirment que les cubegrades sont une généralisation des règles d'association et des cubes de données. Nous pensons que cette approche généralise le concept d'une règle d'association et fait un rapprochement avec les cubes de données. Mais, nous pensons qu'elle ne généralise nullement le processus d'extraction des règles d'association à partir d'un cube de données. En effet, les auteurs ne proposent pas des algorithmes pour la découverte des cubegrades dans une base multidimensionnelle. Ils ne définissent pas non plus le calcul du support et de la confiance d'un cubegrade.

Règles inter-dimensionnelles basées sur les quantités

Guenzel *et al.* proposent un processus d'extraction de règles inter-dimensionnelles avec des prédicats non répétitifs à partir d'un environnement multidimensionnel des

données [GAL99]. Cette approche construit une règle d'association à partir d'un ensemble de modalités, appelé *éléments dimensionnels*, provenant de dimensions distinctes du cube. Chaque élément dimensionnel d'une règle d'association est pris à partir d'un seul niveau hiérarchique d'une dimension.

Les auteurs identifient chaque motif candidat d'une règle par une cellule ou un sous-cube dans le cube étudié. Le support et la confiance d'une règle sont ensuite exprimés en fonction des fréquences contenues dans ces cellules ou dans ces sous-cubes. Par exemple, soit la règle inter-dimensionnelles suivante :

$$Produit(iTwin) \Rightarrow Lieu(France) \wedge Temps(2004)$$

Le support de cette règle s'exprime selon la quantité du produit *iTwin* vendu en *France* pendant l'année *2004*. Par exemple, le support de cette règle peut être égal à 1200 unités vendues. La confiance de cette règle est calculée en divisant la quantité d'unités du produit *iTwin*, vendu en *France* pendant l'année *2004*, par la quantité d'unités totales vendues pour le produit *iTwin*. Cette approche de calcul du support et de la confiance rejoint le cas classique qui se base sur le comptage des faits supportés par la règle selon la mesure COUNT.

Règles intra-dimensionnelles contextualisées

Dans [CDH99, CDH00], Chen *et al.* proposent une plateforme OLAP pour la fouille dans les transactions relatives au commerce électronique (*distributed OLAP based infrastructure*). Selon les auteurs, cette plateforme inclut des outils d'entreposage, d'analyse en ligne et des techniques de fouille de données. Chen *et al.* introduisent dans cette plateforme un processus d'extraction de règles d'association intra-dimensionnelles. Une règle intra-dimensionnelle contient des modalités provenant du même niveau hiérarchique d'une même dimension, appelée *dimension de base*. Elle s'exprime selon un contexte de données en fonction d'autres dimensions du cube. Par exemple, considérons la règle suivante :

$$[x \in \mathsf{Client} : achète\_produit(x, A) \Rightarrow achète\_produit(x, B)]$$
$$| \ \textbf{Lieu} = France, \ \textbf{Temps} = 2005$$

Dans cet exemple, *Client* est la dimension de base, les produits sont les éléments (*item*) de la règle et *Lieu* et *Temps* sont les dimensions selon lesquelles l'utilisateur définit le contexte du cube d'où la règle est extraite.

Selon Chen *et al.*, le contexte d'une règle intra-dimensionnelle peut-être défini de différentes manières selon le niveau de granularité souhaité par l'utilisateur. Par exemple, la règle précédente peut également être exprimée dans des contextes différents :

$[x \in \textit{Client} : \textit{achète\_produit}(x, A) \Rightarrow \textit{achète\_produit}(x, B)]$
$|$ **Lieu** $=$ ***Lyon***, **Temps** $=$ *2005*

$[x \in \textit{Client} : \textit{achète\_produit}(x, A) \Rightarrow \textit{achète\_produit}(x, B)]$
$|$ **Lieu** $=$ *France*, **Temps** $=$ ***janvier 2005***

Règles d'association étendues

Dans [NJ03], Nestorov et Jukić introduisent un processus d'extraction de règles d'association étendues (*extended association rules*) à partir des entrepôts de données. Cette approche consiste à exploiter le langage de requête SQL fourni dans les systèmes de gestion des bases de données multidimensionnelles sans faire recours à des composantes extérieures de fouille de données.

Une règle d'association étendue est une règle intra-dimensionnelle avec prédicats répétitifs. Elle exprime une association entre les modalités d'une seule dimension (*item dimension*) et qui satisfont des conditions fixées par l'utilisateur dans d'autres dimensions (*non-item dimensions*).

Cependant, cette approche s'inscrit dans le problème d'analyse du panier de la ménagère. En effet, les éléments d'une règle d'association étendue désignent exclusivement des produits de ventes. Si un utilisateur cherche à découvrir les associations des produits vendus dans le sud de la France pendant la saison estivale, un exemple d'une règle d'association étendu peut-être :

Dans le *Sud* et pendant l'*Été* : **Tente** $\Rightarrow$ **Sac de couchage**
(*Support* $= 1\%$, *Confiance* $= 50\%$)

Pour obtenir une telle règle, l'utilisateur doit tout d'abord choisir la modalité *Sud* dans la dimension *Lieu* et la modalité *Été* dans la dimension *Temps*. L'utilisateur doit également fixer les seuils minimums du support et de la confiance. Le processus d'extraction des règles étendues utilise une séquence dynamique de requêtes SQL.

Règles d'association à partir d'un entrepôt de données

Tjioe et Taniar proposent une approche pour extraire des règles d'association inter-dimensionnelles à partir d'un entrepôt de données [TT05]. Cette approche consiste en un ensemble de procédures de pré-traitement des données afin de les préparer pour la phase de fouille. Ces procédures partent des dimensions choisies par l'utilisateur pour le processus de fouille. Les pré-traitements effectués ensuite sur les données de ces dimensions se basent essentiellement sur la fonction d'agrégation de la moyenne (AVG).

En effet, les auteurs proposent quatre algorithmes de pré-traitement : VAvg, HAvg, WMAvg et ModusFilter. Les trois premiers algorithmes consistent à calculer, dans

un premier temps, la valeur moyenne d'une mesure, sélectionnée par l'utilisateur. ModusFilter calcule le mode de la mesure, c'est-à-dire la valeur la plus fréquente de la mesure. Dans un second temps, ces algorithmes élaguent les faits OLAP ayant une mesure inférieure à la valeur moyenne. Les auteurs considèrent que les faits dont la mesure est en dessous de la valeur moyenne sont inintéressants pour le processus de fouille parce qu'ils ne peuvent pas générer de règles d'association.

L'algorithme VAvg calcule la moyenne verticale d'une mesure selon les dimensions choisies, alors que HAvg calcule plutôt la moyenne horizontale. WMAvg calcule la moyenne mobile pondérée verticalement dans les dimensions choisies. Par exemple, en partant du croisement des dimensions Temps et Produit, l'algorithme VAvg calcule la moyenne générale des bénéfices de chaque produit sur toutes les années. Ensuite, comme le résume le tableau 5.3, l'algorithme élimine pour chaque produit les faits dont les bénéfices annuels sont au-dessous de la moyenne générale. WMAvg fonctionne de la même manière que VAvg dans la phase d'élagage. En revanche, au lieu de calculer une simple moyenne d'un produit, WMAvg calcule plutôt une moyenne mobile pondérée par les quantités annuelles de ce produit. L'algorithme ModusFilter calcule pour chaque produit le mode, c'est-à-dire la valeur des bénéfices la plus fréquente dans le temps. Ensuite, pour chaque produit, il ne garde que les faits ayant une mesure égale au mode.

Temps	iTwin (Bénéfices)	...	aDream (Bénéfices)
2000	~~100~~		250
2001	~~120~~		~~125~~
2002	300		~~80~~
2003	~~200~~		~~110~~
2004	250		~~100~~
2005	270		150
2006	280		180
Vavg	**217,14**	...	**142,14**

TAB. 5.3 – Exemple de fonctionnement de l'algorithme Vavg [TT05]

Avec le même exemple de dimensions, l'algorithme HAvg calcule plutôt la moyenne générale des bénéfices de chaque année pour tout les produits. Comme le résume le tableau 5.4, pour chaque année, l'algorithme élimine ensuite les faits dont les bénéfices d'un produit sont en dessous de la moyenne générale.

Ces algorithmes de pré-traitement suivent tous une démarche relationnelle et emploient des requêtes SQL pour élaguer, dans la table des faits, les données jugées inutiles pour le processus de fouille. Les données filtrées sont aplaties selon un format tabulaire (*initialized table*). Les auteurs proposent ensuite trois algorithmes, de type Apriori, d'extraction de règles d'association inter-dimensionnelles à partir de

Temps	iTwin	DV-400	aStar	aDream	Havg
2000	~~100~~	200	150	~~125~~	**143**
2001	135	160	~~90~~	145	**132**
⋮					⋮
2006	125	~~50~~	175	150	**125**

TAB. 5.4 – Exemple de fonctionnement de l'algorithme Havg [TT05]

ces données filtrées : l'algorithme GenNLI pour les règles à prédicats non répétitifs et les algorithmes ComDims et GenHLI pour les règles à prédicats répétitifs.

5.2.3 Discussion et positionnement

Les approches proposées pour l'extraction des règles d'association à partir des cubes de données peuvent se décliner selon plusieurs aspects. Tout d'abord, nous remarquons que toutes ces approches sont validées avec des cubes de données de ventes, ce qui les rapproche de la problématique classique de l'analyse du panier de la ménagère.

Cependant, comme le montre le tableau 5.5, à l'exception de l'approche de Chen *et al.* [CDH99, CDH00] et de celle de Nestorov et Jukić [NJ03], nous pensons que les autres propositions peuvent aisément être étendues à d'autres domaines d'application. En effet, les règles contextualisées de Chen *et al.*, où chaque prédicat d'une règle concerne exclusivement l'achat d'un produit, sont restreintes au domaine des transactions de commerce électronique. Quant aux règles étendues de Nestorov et Jukić, l'extraction de ces dernières dépend essentiellement du contenu du cube de données.

Pour trouver les associations intra-dimensionnelles, les auteurs supposent que le cube de données des ventes contient l'information relative à la quantité vendue de deux, ou même de plusieurs, produits ensemble. En d'autres termes, les auteurs supposent qu'on peut croiser deux modalités d'une même dimension et qu'on peut observer la mesure de ce croisement. Cette information n'est pas toujours évidente à trouver dans un cube de données classique et elle va même à l'encontre de la modélisation multidimensionnelles d'un cube de données.

La fouille guidée des règles d'association proposée par Kamber *et al.* [KHC97] est la seule approche qui prend en compte à la fois les cubes de données matérialisés MOLAP et les cubes de données non matérialisés ROLAP (voir tableau 5.5). Toutes les autres propositions considèrent que le cube est non matérialisé et que les données sont structurées dans une base relationnelle. Ceci permet à certaines approches [IKA02, GAL99, CDH99, CDH00, NJ03] d'exploiter le langage de requête

Proposition	Domaine d'application		Représentation des données		Structure des données	
	Général	Spécifique	ROLAP	MOLAP	Relationnelle	Multidimensionnelle
Kamber *et al.*	•		•	•	•	•
Zhu	•		•		•	
Imieliński *et al.*	•		•		•	
Guenzel *et al.*	•		•		•	
Chen *et al.*		•	•		•	
Nestorov et Jukić		•	•		•	
Tjioe et Taniar	•		•		•	
Notre proposition	•			•		•

TAB. 5.5 – Comparaison des propositions d'extraction de règles d'association selon le domaine d'application, la représentation et la structure des données

SQL du système de gestion de la base afin d'extraire les données nécessaires aux calculs et à la construction des règles d'association. D'autres approches [Zhu98, TT05] font plutôt recours à l'aplatissement des données du cube afin de les préparer à la phase de fouille.

L'approche que nous proposons est plutôt une approche générale qui ne dépend pas d'un domaine d'application particulier. Contrairement aux approches qui font recours à la structure relationnelle ROLAP pour employer le langage de requête SQL ou pour aplatir les données, la notre permet d'opérer directement sur des cubes de données multidimensionnels matérialisés MOLAP. Afin d'extraire des règles d'association à partir des cubes nous utilisons plutôt le langage de requêtes multidimensionnelles MDX (*Multi-Dimensional eXpression*).

Les règles d'association inter-dimensionnelles sont les plus exploitées dans l'ensemble des approches proposées. En général, une règle d'association inter-dimensionnelles est constituée de prédicats non répétitifs où chaque instance de prédicat provient d'une dimension distincte du cube. En revanche, une règle d'association intra-dimensionnelle est constituée de prédicats répétitifs dont les instances représentent des modalités provenant d'une même dimension du cube. Cependant, les travaux de Imieliński *et al.* [IKA02] et ceux de Tjioe et Taniar [TT05] dérogent à cette règle. Les cubegrades de Imieliński *et al.*, bien que nous pensons qu'ils ne peuvent être considérés comme de vraies règles d'association, font intervenir des prédicats de plusieurs dimensions

Proposition	Dimensions		Niveau d'abstraction		Prédicats		Mesures	
	Intra	Inter	Unique	Multiples	Répétitifs	Non répétitifs	COUNT	Autres mesures
Kamber *et al.*		•	•			•	•	
Zhu	•	•	•		•	•	•	
Imieliński *et al.*		•		•	•			•
Guenzel *et al.*		•		•		•		
Chen *et al.*	•			•	•		•	
Nestorov et Jukić	•			•	•		•	
Tjioe et Taniar		•		•	•	•	•	
Notre proposition		•		•		•		•

TAB. 5.6 – Comparaison des propositions d'extraction de règles d'association selon les dimensions, le niveau d'abstraction, les prédicats et les mesures

dont les instances peuvent être redondantes dans l'antécédent et le conséquent de l'implication. Quant à Tjioe et Taniar, ils sont les seuls à proposer des règles d'association inter-dimensionnelles qui peuvent avoir soit des prédicats répétitifs, soit des prédicats non répétitifs (voir tableau 5.6). Dans notre approche, les règles que nous découvrons permettent d'exprimer des associations entre les différentes dimensions du cube. Ainsi, nous nous plaçons dans le contexte des règles inter-dimensionnelles avec prédicats non répétitifs.

Il est à noter que, mises à part la proposition de Kamber *et al.* [KHC97] et celle de Zhu [Zhu98], toutes les autres approches sont capable d'exprimer les règles d'association à plusieurs niveaux d'abstraction profitant ainsi de l'aspect hiérarchique des dimensions d'un cube de données. Par exemple, un cubegrade [IKA02] peut s'exprimer par généralisation, en passant à un niveau de granularité plus haut, ou par spécialisation, en passant à un niveau de granularité plus bas. L'extraction des règles d'association contextualisées [CDH99, CDH00] peut aussi évoluer dans différents contextes de données selon les niveaux de granularité choisis par l'utilisateur. Nous exploitons dans notre proposition la notion des méta-règles afin que l'utilisateur OLAP puisse guider le processus de fouille et l'orienter selon ses besoins d'analyse. En plus, nous enrichissons les méta-règles en instaurant un contexte d'analyse qui peut se définir par l'utilisateur selon plusieurs niveaux d'abstraction en tirant profit des niveaux de granularité hiérarchiques des dimensions.

Comme le montre le tableau 5.6, globalement toutes les propositions se basent sur le comptage de la fréquence des données, en exploitant la mesure COUNT, afin de

calculer le support et la confiance des règles d'association. Il est vrai que Imieliński *et al.* [IKA02] intègrent des mesures agrégées outre que la mesure COUNT dans l'expression d'un cubegrade. Cependant, ces derniers ne calculent pas de support ou de confiance pour les cubegrades. Il est vrai aussi que Tjioe et Taniar [TT05] utilisent la fonction de la moyenne AVG pour élaguer les motifs inutiles dans le processus de découverte des règles d'association. Mais, il s'agit plutôt d'une phase de préparation des données. Dans la phase de fouille, les auteurs se basent sur les fréquences des données pour calculer le support et la confiance des règles. Dans notre approche, nous revisitons la définition du support et de la confiance des règles d'association et proposons plutôt de calculer ces derniers en fonction de l'agrégat SUM d'une mesure du cube qui répond au mieux aux attentes de l'utilisateur. Ainsi, contrairement à toutes les approches précédentes, une règle d'association n'est plus évaluée selon le nombre d'occurrences des faits qu'elle supporte. Elle est plutôt évaluée selon la mesure des faits qu'elle supporte. Cette nouvelle définition du support et de la confiance enrichit le sens des règles d'association extraites et les adapte le plus possible au contexte de l'analyse OLAP. De plus, pour contourner le problème du grand nombre de règles qui peuvent être découvertes et pour permettre à l'utilisateur une meilleure validation des règles extraites, nous mettons en œuvre deux indices, le *Lift* et l'indice de *Loevinger*, qui mesurent l'intérêt des associations des règles.

Proposition	Interaction avec l'utilisateur		Formalisation de la proposition		Représentation graphique des règles	
	Oui	Non	Oui	Non	Oui	Non
Kamber *et al.*	•			•		•
Zhu	•			•	•	
Imieliński *et al.*	•		•			•
Guenzel *et al.*	•			•		•
Chen *et al.*		•		•		•
Nestorov et Jukić	•			•		•
Tjioe et Taniar	•			•		•
Notre proposition	•		•		•	

TAB. 5.7 – Comparaison des propositions d'extraction de règles d'association selon l'interaction avec l'utilisateur, la formalisation de la proposition et la représentation graphique des règles extraites

Comme le montre le tableau 5.7, l'approche de Chen *et al.* [CDH99, CDH00] ne prévoit pas d'interaction possible avec les utilisateurs. En effet, il s'agit d'une plateforme intégrant l'analyse en ligne et la fouille de données totalement orientée pour des objectifs d'analyse sur des données transactionnelles. En revanche dans les autres approches, l'utilisateur se place au cœur du processus d'extraction des règles. Par exemple, dans [KHC97], à travers la définition des méta-règles, le support

minimum et la confiance minimale, l'utilisateur est capable d'orienter selon ses besoins le contenu et la qualité des connaissances qu'il cherche à découvrir. Cependant, nous remarquons que, exception faite pour l'approche des cubegrades [IKA02], la majorité des propositions manquent de formalisations capables d'établir un cadre théorique précis pour le processus d'extraction des règles d'association à partir des cubes de données.

De plus, dans toutes ces propositions, Zhu est le seul qui propose de valider les règles extraites par des représentations graphiques [Zhu98]. Cependant, ce dernier reprend des outils de visualisation prévus à la base pour les règles d'association classiques, tels que le graphe d'items (*Ball Graphical View*) [HDH+01] ou le graphe des règles en bâtonnets (*Bar Graphical View*) [Han98], qui n'ont pas de lien particulier avec le modèle des données multidimensionnelles. Dans notre approche, nous pallions ces limites et proposons un cadre formel pour l'extraction des règles d'association inter-dimensionnelles à partir des cubes de données. Nous proposons également une formalisation pour la visualisation des règles extraites. Cette visualisation se base sur un nouvel encodage graphique des règles d'association exclusivement adapté aux représentations multidimensionnelles des données. La représentation graphique des règles permet à l'utilisateur de cibler rapidement les règles les plus intéressantes et de mieux appréhender les connaissances véhiculées par ces dernières.

5.3 Définitions et formalisation

5.3.1 Sous-cube de données

Définition 5.3.1 (Sous-cube de données) *Soit $\mathcal{D}' \subseteq \mathcal{D}$ un sous-ensemble non vide de p dimensions $\{D_1, \ldots, D_p\}$ du cube de données $\mathcal{C}$ ($p \leq d$). Le p-uplet $(\Theta_1, \ldots, \Theta_p)$ est un sous-cube de données dans $\mathcal{C}$ selon $\mathcal{D}'$ si et seulement si $\forall i \in \{1, \ldots, p\}$, $\Theta_i \neq \emptyset$ et il existe un indice unique $j \geq 0$ tels que $\Theta_i \subseteq \mathcal{A}_{ij}$.*

Un sous-cube de données selon un ensemble de dimensions $\mathcal{D}'$ correspond à une portion du cube de données original $\mathcal{C}$. Il s'agit de fixer un niveau hiérarchique H_j^i dans chaque dimension de $D_i \in \mathcal{D}'$ et de sélectionner dans ce niveau un sous-ensemble Θ_i non vide de modalités appartenant à l'ensemble de toutes les modalités $\mathcal{A}_{ij}$ de H_j^i.

Par exemple, considérons le sous-ensemble des dimensions $\mathcal{D}' = \{D_1, D_2\}$ du cube $\mathcal{C}$ de la figure 5.3. Soient le sous-ensemble des modalités Θ_1=$\{$*Europe*$\}$ du niveau H_1^1 (*Continent*) de la dimension D_1 (*Lieu*) et le sous-ensemble des modalités Θ_2=$\{$*EN-700*, *aStar*, *aDream*$\}$ du niveau H_2^2 (*Produit*) de la dimension D_2 (*Produit*).

Dans ce cas, (Θ_1, Θ_2)=(*Europe*, $\{$*EN-700*, *aStar*, *aDream*$\}$) correspond au sous-cube grisé dans la figure 5.3 dans le cube $\mathcal{C}$ selon les dimensions $\mathcal{D}' = \{D_1, D_2\}$.

Il est à noter que, selon cette définition, un même sous-cube de données peut-être désigné par différentes notations :

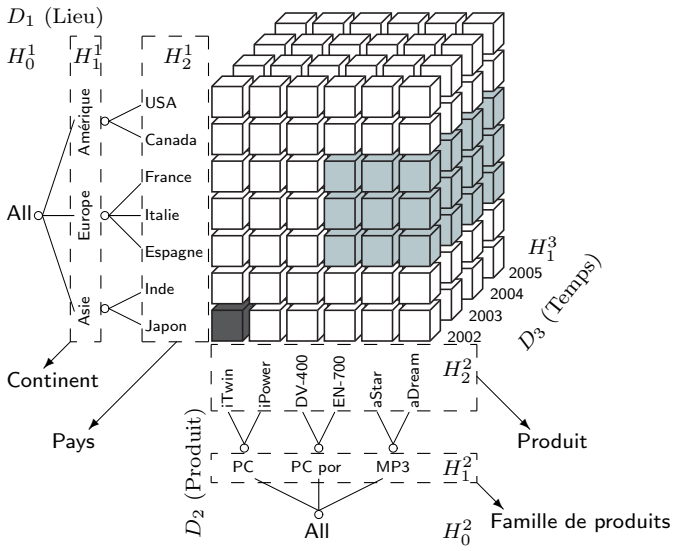

FIG. 5.3 – Exemple d'un sous-cube de données dans le cube des ventes

- en changeant le nombre des dimensions selon lesquelles est défini le sous-cube et en fixant à *All* les dimensions restantes. Par exemple, la portion grisée de la figure 5.3 peut aussi se définir comme le sous-cube de données (*Europe*, {*EN-700*, *aStar*, *aDream*}, *All*) selon l'ensemble des dimensions $\mathcal{D} = \{D_1, D_2, D_3\}$;
- en changeant, si possible, de niveau hiérarchique d'une des dimensions selon lesquelles est défini le sous-cube. Par exemple, la portion grisée de la figure 5.3 peut aussi se définir comme le sous-cube de données ({*France*, *Italie*, *Espagne*}, {*EN-700*, *aStar*, *aDream*}) selon l'ensemble des dimensions $\mathcal{D} = \{D_1, D_2\}$;
- en changeant, si possible, le nombre de dimensions selon lesquelles est défini le sous-cube et leurs niveaux hiérarchiques. Par exemple, la portion grisée de la figure 5.3 peut aussi se définir comme le sous-cube de données ({*France*, *Italie*, *Espagne*}, {*EN-700*, *aStar*, *aDream*}, *All*) selon l'ensemble des dimensions $\mathcal{D} = \{D_1, D_2, D_3\}$.

On note aussi qu'une cellule d'un cube de données $\mathcal{C}$ correspond au cas particulier d'un sous-cube de données défini selon l'ensemble entier des dimensions $\mathcal{D} = \{D_1, \ldots, D_d\}$ et tel que $\forall i \in \{1, \ldots, d\}$, Θ_i est un singleton contenant une seule modalité appartenant au niveau hiérarchique le plus fin de la dimension D_i. Par exemple, la cellule noire dans le cube de la figure 5.3 est exprimée selon le sous-cube (*Japon*, *iTwin*, *2002*) selon l'ensemble des dimensions $\mathcal{D} = \{D_1, D_2, D_3\}$.

5.3.2 Agrégation SUM d'un sous-cube de données

Chaque cellule du cube de données $\mathcal{C}$ représente un fait OLAP qui s'évalue dans $\mathbb{R}$ selon une mesure $M \in \mathcal{M}$. Dans notre proposition, on évalue un sous-cube de données selon l'agrégation SUM de la mesure M. Cette dernière est définie comme suit :

Définition 5.3.2 (Agrégation SUM d'un sous-cube de données) *Soient $(\Theta_1, \ldots, \Theta_p)$ un sous-cube de données dans $\mathcal{C}$ selon un sous-ensemble de dimensions $\mathcal{D}' \subseteq \mathcal{D}$ et une mesure $M \in \mathcal{M}$. L'agrégation SUM de la mesure M du sous-cube $(\Theta_1, \ldots, \Theta_p)$, notée $SUM_M(\Theta_1, \ldots, \Theta_p)$, est la somme de toutes les valeurs de la mesure M des faits présents dans le sous-cube.*

Par exemple, le *bénéfice des ventes* du sous-cube de données grisé dans la figure 5.3 peut être évalué selon l'agrégation $\mathsf{SUM}_{Bénéfice}(Europe, \{EN\text{-}700, aStar, aDream\})$ qui représente la somme des valeurs des bénéfices présentes dans toutes les cellules du sous-cube en question, c'est à dire les cellules grisées dans le cube des ventes.

5.3.3 Prédicat dimensionnel

Définition 5.3.3 (Prédicat dimensionnel) *Soit D_i une dimension d'un cube de données $\mathcal{C}$. Un prédicat dimensionnel dans D_i, noté α_i, est un prédicat de la forme $\langle a \in \mathcal{A}_{ij} \rangle$.*

Un prédicat dimensionnel est un prédicat qui prend la valeur d'une modalité de la dimension dans laquelle il est défini. Par exemple, dans la dimension D_1 de la figure 5.3, un prédicat dimensionnel possible peut prendre la forme $\alpha_1 = \langle a \in \mathcal{A}_{11} \rangle = \langle a \in \{Amérique, Europe, Asie\} \rangle$.

5.3.4 Prédicat inter-dimensionnels

Définition 5.3.4 (Prédicat inter-dimensionnels) *Soit $\mathcal{D}' \subseteq \mathcal{D}$ un sous-ensemble non vide de p dimensions $\{D_1, \ldots, D_p\}$ du cube de données $\mathcal{C}$ $(2 \leq p \leq d)$. $(\alpha_1 \wedge \cdots \wedge \alpha_p)$ est un prédicat inter-dimensionnels dans $\mathcal{D}'$ si et seulement si $\forall i \in \{1, \ldots, p\}$, α_i est un prédicat dimensionnel dans D_i.*

Par exemple, soit $\mathcal{D}' = \{D_1, D_2\}$ un sous-ensemble de dimensions du cube de données de la figure 5.3. Un prédicat inter-dimensionnels possible dans $\mathcal{D}'$ peut prendre la forme $(\langle a_1 \in \mathcal{A}_{11} \rangle \wedge \langle a_2 \in \mathcal{A}_{21} \rangle)$. Un prédicat inter-dimensionnels est une conjonction de prédicats dimensionnels non répétitifs. C'est-à-dire, chaque dimension de $\mathcal{D}'$ a un prédicat dimensionnel distinct dans l'expression du prédicat inter-dimensionnels.

5.4 Méta-règle inter-dimensionnelles

En s'inspirant du formalisme fourni par Plantevit *et al.* [PCL+05], nous établissons une partition dans les dimensions $\mathcal{D}$ du cube de données $\mathcal{C}$ selon trois sous-ensembles $\mathcal{D}_\mathcal{C}, \mathcal{D}_\mathcal{A}$ et $\mathcal{D}_\mathcal{I}$ tels que :

- $\mathcal{D}_\mathcal{C}$ est un sous-ensemble de p dimensions de *contexte*. Un sous-cube de données dans $\mathcal{C}$ selon $\mathcal{D}_\mathcal{C}$ est défini afin d'établir le contexte d'analyse à partir duquel les règles d'association seront extraites ;
- $\mathcal{D}_\mathcal{A}$ est un sous-ensemble de $(s+r)$ dimensions d'*analyse* à partir desquelles les prédicats d'une méta-règle inter-dimensionnelles sont choisis ;
- $\mathcal{D}_\mathcal{I}$ est le sous-ensemble des dimensions restantes. Ces dimensions sont fixées à l'agrégat total *All*. Il s'agit des dimensions *inutilisées* qui sont totalement agrégées et qui, par conséquent, n'interviennent ni dans la définition du contexte du processus d'extraction des règles d'association, ni dans la définition de la méta-règle.

Une méta-règle inter-dimensionnelles est un modèle de règles d'association défini par l'utilisateur selon un schéma général de la forme :

$$\mathcal{R} \left| \begin{array}{l} \text{Dans le contexte } (\Theta_1, \ldots, \Theta_p) \\ (\alpha_1 \wedge \cdots \wedge \alpha_s) \Rightarrow (\beta_1 \wedge \cdots \wedge \beta_r) \end{array} \right. \tag{5.4.1}$$

où $(\Theta_1, \ldots, \Theta_p)$ est un sous-cube de $\mathcal{C}$ défini selon le sous-ensemble des dimensions $\mathcal{D}_\mathcal{C}$. Ce sous-cube désigne la portion du cube de données dans laquelle sera conduit le processus d'extraction des règles d'association. À la différence du schéma de la méta-règle proposé par Kamber *et al.* dans [KHC97], notre méta-règle permet de cibler un contexte d'analyse précis dans le cube en définissant la population des faits qui se trouvent dans le sous-cube de données $(\Theta_1, \ldots, \Theta_p)$. Il est à remarquer que le cas où le sous-ensemble des dimensions de contexte est vide ($\mathcal{D}_\mathcal{C} = \emptyset$), correspond à un contexte d'analyse général qui couvre tous les faits du cube de données $\mathcal{C}$.

Il est à noter que $\forall k \in \{1, \ldots, s\}$ (respectivement $\forall k \in \{1, \ldots, r\}$), α_k (respectivement β_k) est un prédicat dimensionnel dans une dimension distincte de $\mathcal{D}_\mathcal{A}$. Par conséquent, la conjonction des prédicats $(\alpha_1 \wedge \cdots \wedge \alpha_s) \wedge (\beta_1 \wedge \cdots \wedge \beta_r)$ est un prédicat inter-dimensionnels dans $\mathcal{D}_\mathcal{A}$. Le nombre de prédicats $(s + r)$ dans la méta-règle est égal au nombre de dimensions dans $\mathcal{D}_\mathcal{A}$. Ainsi, notre méta-règle est un modèle qui définit des règles d'association inter-dimensionnelles avec des prédicats non répétitifs.

Par exemple, en plus des trois dimensions représentées dans la figure 5.3, supposons que le cube des ventes contient quatre autres dimensions : *Profil du consommateur* (D_4), *Profession du consommateur* (D_5), *Sexe* (D_6) et *Promotion* (D_7). Considérons alors la partition suivante des dimensions du cube des ventes :

- $\mathcal{D}_\mathcal{C} = \{D_5, D_6\} = \{$*Profession du consommateur*, *Sexe*$\}$;

– $\mathcal{D}_{\mathcal{A}} = \{D_1, D_2, D_3\}$ ={*Lieu*, *Produit*, *Temps*} ;
– $\mathcal{D}_{\mathcal{I}} = \{D_4, D_7\}$ ={*Profil du consommateur*, *Promotion*}.

Selon cette partition, un utilisateur peut souhaiter extraire des règles d'association répondant au modèle de la méta-règle inter-dimensionnelles suivante :

$$\left|\begin{array}{l} \text{Dans le contexte } (\mathit{Etudiant}, \mathit{Femme}) \\ \langle a_1 \in \mathit{Continent}\rangle \wedge \langle a_3 \in \mathit{Année}\rangle \Rightarrow \langle a_2 \in \mathit{Produit}\rangle \end{array}\right. \tag{5.4.2}$$

Selon cette méta-règle, les règles d'association inter-dimensionnelles sont extraites à partir du sous-cube de données (*Etudiant*, *Femme*) qui couvre les ventes concernant seulement la population des *étudiantes*. Les dimensions inutilisées (*Profil du consommateur*, *Promotion*) sont totalement agrégées et n'interviennent pas dans le processus d'extraction des règles d'association. En revanche, les dimensions d'analyse interviennent dans la découverte des règles. En effet, les prédicats des règles extraites proviennent des dimensions de $\mathcal{D}_{\mathcal{A}}$. Deux prédicats dimensionnels dans D_1 et D_3 sont prévus dans l'antécédent des règles, alors qu'un seul prédicat dimensionnel est prévu dans le conséquent des règles. Le premier prédicat dimensionnel de l'antécédent est fixé au niveau *Continent* de D_1. Le deuxième prédicat dimensionnel de l'antécédent est fixé au niveau *Année* de D_3. Quant au prédicat dimensionnel du conséquent, il est fixé au niveau *Produit* de D_2.

5.5 Support et confiance basés sur la mesure

5.5.1 Définition classique du support et de la confiance

Classiquement, le support (Supp) d'une règle d'association $X \Rightarrow Y$, dans une base de transactions $\mathcal{T}$, est la probabilité d'avoir dans la base $\mathcal{T}$ des transactions contenant à la fois les items X et Y. Cette probabilité correspond au rapport de la fréquence des transactions contentant X et Y dans $\mathcal{T}$ par la fréquence de toutes les transactions de $\mathcal{T}$. La confiance (Conf) de cette règle correspond à la probabilité conditionnelle d'avoir dans une transaction l'item Y sachant qu'elle contient déjà l'item X. Cette probabilité est le rapport de la fréquence des transactions contentant X et Y dans $\mathcal{T}$ par la fréquence des transactions contenant X dans $\mathcal{T}$. Une règle d'association est dite *fréquente* lorsque son support est plus grand ou égal à un support minimum (*minsupp*) fixé par l'utilisateur. De même, une règle est considérée intéressante lorsque sa confiance est supérieure ou égale à une confiance minimale (*minconf*) fixée par l'utilisateur.

Dans le contexte d'analyse en ligne, la structure multidimensionnelle d'un cube de données est différente de celle d'une base classique de transactions. En effet, un cube de données contient des agrégats pré-calculés. Ces derniers représentent les valeurs de la mesure des faits correspondant à tous les croisements possibles des modalités

provenant des différentes dimensions du cube. Dans un problème de recherche de règles d'association, les agrégats d'un cube de données s'avèrent d'une grande utilité dans le calcul du support et de la confiance d'une règle inter-dimensionnelles. Ces informations agrégées réduisent le temps de parcours des données de la base et permettent ainsi un calcul rapide qui ne dépend, dans ce cas, que du temps d'accès aux données du cube. En particulier, avec la mesure COUNT, on peut accéder directement aux fréquences des faits d'un cube de données $\mathcal{C}$. Ces fréquences nous permettent de calculer d'une manière relativement facile le support et la confiance classiques d'une règle d'association.

Supposons, par exemple, qu'un utilisateur éprouve le besoin de découvrir des règles d'association répondant à la méta-règle (5.4.2). Dans ce cas, R_1 est une règle possible qui pourrait être découverte à partir du cube $\mathcal{C}$:

$$R_1 \left| \begin{array}{l} \text{Dans le contexte } (\textit{Etudiant}, \textit{Femme}) \\ \textit{Amérique} \wedge \textit{2004} \Rightarrow \textit{MP3} \end{array} \right.$$

Le support de R_1, noté $\text{SUPP}(R_1)$, représente le rapport de la fréquence des ventes de lecteur MP3 pour les étudiantes du *continent américain* durant l'année *2004*, par la fréquence de toutes les ventes pour les *étudiantes*. En utilisant la mesure COUNT, $\text{SUPP}(R_1)$ s'exprime selon l'expression suivante :

$$\text{SUPP}(R_1) = \frac{\mathsf{COUNT}(\textit{Amérique}, \textit{MP3}, \textit{2004}, \textit{All}, \textit{Etudiant}, \textit{Femme}, \textit{All})}{\mathsf{COUNT}(\textit{All}, \textit{All}, \textit{All}, \textit{All}, \textit{Etudiant}, \textit{Femme}, \textit{All})}$$

La confiance de R_1, notée $\text{CONF}(R_1)$, représente le rapport de la fréquence des ventes de *lecteurs MP3* pour les *étudiantes* du *continent américain* durant l'année *2004*, par la fréquence de toutes les ventes pour les *étudiantes* du *continent américain* durant l'année *2004*. En utilisant la mesure COUNT, $\text{CONF}(R_1)$ s'exprime selon l'expression suivante :

$$\text{CONF}(R_1) = \frac{\mathsf{COUNT}(\textit{Amérique}, \textit{MP3}, \textit{2004}, \textit{All}, \textit{Etudiant}, \textit{Femme}, \textit{All})}{\mathsf{COUNT}(\textit{Amérique}, \textit{All}, \textit{2004}, \textit{All}, \textit{Etudiant}, \textit{Femme}, \textit{All})}$$

5.5.2 Nouvelle définition du support et de la confiance

Selon les expressions précédentes, le support et la confiance d'une règle d'association sont calculés en fonction des fréquences des faits OLAP en se basant sur la mesure COUNT. D'un point de vue statistique, il s'agit d'étudier la population des faits en fonction de leurs occurrences. Ainsi, pour le calcul du support, il s'agit de vérifier si une règle d'association est supportée par un nombre suffisant de faits afin de pouvoir affirmer qu'elle est fréquente. Cependant, du point de vue de l'analyse en ligne, les

faits OLAP sont le plus souvent observés selon des mesures plus intéressantes que leurs simples fréquences.

En effet, dans une analyse OLAP, un fait est rarement évalué par le nombre de ses occurrences. On s'intéresse le plus souvent à l'évaluer par une autre mesure qui répond aux besoins de l'expert. Par exemple, un directeur commercial d'une chaîne de magasins est plus intéressé d'observer les ventes d'un produit donné en fonction des bénéfices qu'il rapporte que le nombre de ventes de ce dernier. Il est naturellement indispensable, dans notre approche, de tenir compte de cette notion de mesure des faits dans la génération des règles d'association à partir des cubes de données.

Prenons, par exemple, des fragments de données prises dans le sous-cube de données précédent (*Etudiant*, *Femme*). La figure 5.4 (a) représente l'observation des faits de ventes selon leurs fréquences, alors que la figure 5.4 (b) représente l'observation des faits de ventes selon leurs niveaux de bénéfices.

	2004		2005	
	Amérique	Europe	Amérique	Europe
PC	1200	800	950	500
PC por	2500	2700	2800	3200
MP3	10600	5900	11400	9100

(a)

	2004		2005	
	Amérique	Europe	Amérique	Europe
PC	$ 60000	$ 33000	$ 28000	$ 10000
PC por	$ 500000	$ 567000	$ 420000	$ 544000
MP3	$ 116000	$ 118000	$ 57000	$ 91000

(b)

FIG. 5.4 – Fragment du cube des ventes selon (a) les fréquences et selon (b) la mesure des bénéfices

Dans l'exemple de la figure 5.4, pour un support minimum donné, on se rend compte que certains motifs qui sont fréquents, selon les fréquences des faits, ne le sont pas selon le niveau des bénéfices et *vice versa*. Par exemple, pour un support minimum $minsupp = 0,2$, les motifs ($\langle$ *Amérique* $\rangle$, $\langle$ *MP3* $\rangle$, $\langle$ *2004* $\rangle$) et ($\langle$ *Amérique* $\rangle$, $\langle$ *MP3* $\rangle$, $\langle$ *2005* $\rangle$) sont fréquents selon les fréquences des faits (cellules grisées dans le tableau de la figure 5.4 (a)). D'un autre côté, ces mêmes motifs ne sont pas fréquents selon les niveaux de bénéfices. Les motifs fréquents selon la mesure des bénéfices sont ($\langle$ *Europe* $\rangle$, $\langle$ *PC por* $\rangle$, $\langle$ *2004* $\rangle$) et ($\langle$ *Europe* $\rangle$, $\langle$ *PC por* $\rangle$, $\langle$ *2005* $\rangle$) (cellules grisées dans le tableau de la figure 5.4 (b)).

Dans le premier cas, les motifs sont évalués selon le nombre d'occurrence des ventes, mais ne mettent pas en valeur l'importance de ces ventes. En effet, pour des raisons conjoncturelles, sociales ou culturelles par exemple, un petit nombre de ventes d'un produit dans un contexte donné peut s'avérer plus conséquent en terme de bénéfice qu'un grand nombre de ventes de ce même produit dans un contexte différent. Dans le deuxième cas, les motifs ne dépendent pas du nombre d'occurrence des faits OLAP, mais plutôt de leurs mesures. Ces derniers motifs sont plus intéressants et répondent mieux au contexte d'analyse d'un expert.

Dans notre approche, afin de répondre aux besoins d'analyse d'un utilisateur OLAP, nous redéfinissons les indices d'évaluation des règles d'association, notamment le support et la confiance, en se basant sur la mesure des faits. Notre définition consiste à étendre la notion de probabilité selon des rapports d'effectifs, représentant des unités statistiques, à la notion de concentration selon des rapports de mesures, représentant des unités de masse. Soit une règle d'association générale R qui répond à la méta-règle inter-dimensionnelles (5.4.1) :

$$R : \left| \begin{array}{l} \text{Dans le contexte } (\Theta_1, \ldots, \Theta_p) \\ (x_1 \wedge \cdots \wedge x_s) \Rightarrow (y_1 \wedge \cdots \wedge y_r) \end{array} \right.$$

Ainsi, nous définissons le support et la confiance d'une règle d'association inter-dimensionnelles selon une mesure $M \in \mathcal{M}$ du cube de données $\mathcal{C}$ par les expressions générales suivantes :

$$\textsc{Supp}(R) = \frac{\mathsf{SUM}_M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, \mathit{All}, \ldots, \mathit{All})}{\mathsf{SUM}_M(\mathit{All}, \ldots, \mathit{All}, \Theta_1, \ldots, \Theta_p, \mathit{All}, \ldots, \mathit{All})}$$

$$\textsc{Conf}(R) = \frac{\mathsf{SUM}_M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, \mathit{All}, \ldots, \mathit{All})}{\mathsf{SUM}_M(x_1, \ldots, x_s, \mathit{All}, \ldots, \mathit{All}, \Theta_1, \ldots, \Theta_p, \mathit{All}, \ldots, \mathit{All})}$$

Selon la définition 5.3.2, $\mathsf{SUM}_M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, \mathit{All}, \ldots, \mathit{All})$ correspond à l'agrégation SUM de la mesure M du sous-cube de données défini par l'expression $(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, \mathit{All}, \ldots, \mathit{All})$.

Dans ces nouvelles expressions du support et de la confiance, nous ramenons la population étudiée à la population des unités de masse mesurées. Il s'agit d'une définition plus générale que celle du cas classique concernant la population des unités des faits. Ceci-dit, le cas classique est un cas particulier qui correspond à la situation où la mesure M est égale à l'unité et où l'agrégation SUM de la mesure M est équivalente à la mesure COUNT. Dans la suite, et afin d'alléger les notations, nous utilisons délibérément ces nouvelles définitions et continuons à employer les termes de *support* et de *confiance*.

5.6 Critères de qualité pour les règles inter-dimensionnelles

5.6.1 Limites du support et de la confiance

Le support et la confiance sont les indices d'évaluation des règles d'association les plus utilisés. Ils constituent également des critères fondamentaux dans la génération des règles selon la famille des algorithmes basés sur Apriori [AIS93]. En effet, seules les règles ayant un support et une confiance plus grands que les seuils *minsupp* et *minconf* sont retenues par le processus de fouille. Cependant, le support et la confiance conduisent en général à la génération d'un grand nombre de règles dont la plupart peuvent s'avérer inintéressantes. À cause de ce grand nombre des règles, il est difficile pour un expert d'isoler les meilleures en fonction de ses préférences.

En vue de pallier cette limite, de nombreux indices d'évaluation[1] des règles d'association ont été proposés dans la littérature. Ces indices se caractérisent par des propriétés qui dépendent essentiellement des préférences de l'utilisateur et de la structure des données étudiées. Pour un exposé plus complet sur le sujet, nous renvoyons le lecteur à [Bla05, LVL05, LVL06].

Dans le cadre de notre approche, vu la grande masse de données dans un cube OLAP, nous sommes confrontés au même type de problème. Dès lors, nous pensons qu'il est nécessaire d'employer d'autres indices d'évaluation des règles inter-dimensionnelles.

5.6.2 Terminologie et notations

Soit une règle d'association R qui, dans le contexte $(\Theta_1, \ldots, \Theta_p)$, répond à la méta-règle inter-dimensionnelles (5.4.1). Considérons maintenant une expression générale de R de la forme $X \Rightarrow Y$, où $X = (x_1 \wedge \cdots \wedge x_s)$ et $Y = (y_1 \wedge \cdots \wedge y_r)$ sont des conjonctions de prédicats dimensionnels. Nous considérons aussi une mesure $M \in \mathcal{M}$ du cube de données $\mathcal{C}$. Dans la suite, nous adoptons les notations suivantes :

- P_X (respectivement P_Y) est la proportion de l'agrégation SUM de la mesure M des faits vérifiant X (respectivement Y) dans le sous-cube de données $(\Theta_1, \ldots, \Theta_p)$ défini selon les dimensions de contexte $\mathcal{D}_\mathcal{C}$. P_X et P_Y s'expriment selon les expressions :

[1] Dans la littérature, on parle plutôt de *mesures d'intérêt des règles d'association*. Nous avons délibérément choisi le terme *indices d'évaluation des règles d'association* afin d'éviter toute confusion avec les *mesures OLAP*.

$$P_X = \frac{\mathsf{SUM}_M(x_1, \ldots, x_s, All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{\mathsf{SUM}_M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

$$P_Y = \frac{\mathsf{SUM}_M(All, \ldots, All, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{\mathsf{SUM}_M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

- $P_{\overline{X}} = 1 - P_X$ (respectivement $P_{\overline{Y}} = 1 - P_Y$) est la proportion de l'agrégation SUM de la mesure M des fait ne vérifiant pas X (respectivement Y) dans le sous-cube de données $(\Theta_1, \ldots, \Theta_p)$ défini selon les dimensions de contexte $\mathcal{D}_\mathcal{C}$;
- P_{XY} est la proportion de l'agrégation SUM de la mesure M des faits vérifiant X et Y dans le sous-cube de données $(\Theta_1, \ldots, \Theta_p)$ défini selon les dimensions de contexte $\mathcal{D}_\mathcal{C}$. P_{XY} correspond au support de la règle R :

$$P_{XY} = \textsc{Supp}(R) = \frac{\mathsf{SUM}_M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{\mathsf{SUM}_M(All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

- $P_{Y/X}$ est la proportion de l'agrégation SUM de la mesure M des faits vérifiant X et Y dans le sous-cube de données $(\Theta_1, \ldots, \Theta_p)$ défini selon les dimensions de contexte $\mathcal{D}_\mathcal{C}$ et qui vérifie déjà X. $P_{Y/X}$ correspond à la confiance de la règle R :

$$P_{Y/X} = \textsc{Conf}(R) = \frac{\mathsf{SUM}_M(x_1, \ldots, x_s, y_1, \ldots, y_r, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}{\mathsf{SUM}_M(x_1, \ldots, x_s, All, \ldots, All, \Theta_1, \ldots, \Theta_p, All, \ldots, All)}$$

5.6.3 Choix des indices d'évaluation des règles d'association

Dans la littérature, on distingue deux grandes catégories d'indices d'évaluation des règles d'association : *les indices descriptifs* et *les indices statistiques*. En général, les indices statistiques dépendent fortement de la taille des données étudiées [LVL05]. En effet, quand le nombre d'exemples de la population étudiée est très important, un indice statistique perd son pourvoir discriminant et tend artificiellement vers 1. En plus, un indice statistique fait référence à un modèle probabiliste. Ceci suppose que l'on maîtrise la loi de probabilité des exemples étudiés. Ce qui n'est pas forcément le cas dans un contexte d'analyse en ligne où les utilisateurs, généralement, n'ont pas particulièrement une culture statistique, mais s'inscrivent plutôt dans une approche exploratoire des données. Ceci nous amène à écarter l'utilisation des indices statistiques qui sont difficiles à employer et à interpréter dans le cadre de notre approche.

Par conséquent, nous nous intéressons aux indices descriptifs qui sont plus simples à employer et plus abordables pour les utilisateurs OLAP. De plus, contrairement aux indices statistiques, le pouvoir discriminant d'un indice descriptif est indépendant du nombre d'exemples de la population étudiée. Cette propriété s'adapte bien avec notre contexte de données multidimensionnelles souvent caractérisées par un grand nombre de faits OLAP.

Dans le cadre de l'évaluation des règles d'association inter-dimensionnelles extraites à partir d'un cube de données, nous proposons d'utiliser, en plus du support et de la confiance, deux indices descriptifs, à savoir le *Lift* [BMS97] et l'indice de *Loevinger* [Loe47].

1. Le *Lift* représente l'écart à l'indépendance d'une règle d'association. Pour une règle $X \Rightarrow Y$, l'indépendance théorique entre l'antécédent X et le conséquent Y se mesure par le produit $P_X P_Y$. Le support de la règle, quant à lui, peut s'interpréter comme étant l'association brute réellement observée entre X et Y qui est égale au support de la règle P_{XY}. Ainsi, le *Lift* d'une règle $X \Rightarrow Y$ s'exprime selon le rapport de son support P_{XY} sur le produit $P_X P_Y$:

$$\textsc{Lift}(R) = \frac{P_{XY}}{P_X P_Y} = \frac{\textsc{Supp}(R)}{P_X P_Y} \tag{5.6.1}$$

   Le *Lift* s'interprète aussi comme le quotient du nombre d'exemples observés par celui attendus sous l'hypothèse de l'indépendance entre X et Y. Par exemple, une règle $X \Rightarrow Y$ ayant un *Lift* égal à 2 signifie que le nombre d'exemples de la règle est 2 fois plus grand que celui attendu sous l'hypothèse de l'indépendance, ce qui implique que le consommateur qui achète X a deux fois plus de chance d'acheter Y que le consommateur en général. Dans le cadre de notre approche, dans le contexte (*Etudiant*, *Femme*), si la règle inter-dimensionnelles *Amérique*$\wedge$ *2004* $\Rightarrow$ *MP3* a un *Lift* égal à n, ceci signifie que les ventes réalisées par les étudiantes en 2004 dans le continent américain ont n fois plus de chance d'être des ventes de lecteurs MP3 que dans le cas de ventes générales. Dans cet exemple précis, le *Lift* peut aussi s'interpréter comme le coefficient multiplicateur de la part du marché des lecteurs MP3 dû au fait qu'on se place dans le contexte d'une clientèle d'étudiantes du continent américain en 2004.

2. L'indice de *Loevinger*, avec le support et la confiance, est l'un des plus anciens indices répertoriés dans le domaine de l'évaluation de l'intérêt des règles d'association. Il consiste à améliorer l'interprétation de la confiance d'une règle par normalisation de cette dernière. Il normalise la confiance centrée de la règle par rapport aux objets ne vérifiant pas la conclusion. Ainsi, l'indice de *Loevinger* d'une règle $X \Rightarrow Y$ s'exprime selon le rapport de sa confiance centrée $(P_{Y/X} - P_Y)$ par $P_{\overline{Y}}$:

$$\textsc{Loev}(R) = \frac{P_{Y/X} - P_Y}{P_{\overline{Y}}} = \frac{\textsc{Conf}(R) - P_Y}{P_{\overline{Y}}} \tag{5.6.2}$$

L'indice de *Loevinger* mesure la puissance implicative d'une règle d'association. En effet, l'implication d'une règle d'association $X \Rightarrow Y$ ne dépend pas seulement du nombre d'exemples vérifiant Y, mais dépend aussi du nombre de contre-exemples ne vérifiant pas Y. Plus le nombre de contre-exemples P_Y est grand, plus la règle représente une faible implication.

5.7 Algorithme d'extraction des règles inter-dimensionnelles

5.7.1 Terminologie et notations

Dans cette section, nous reprenons les notations des sections précédentes auxquelles nous rajoutons les suivantes :

- k est un indice correspondant à l'itération en cours d'un algorithme itératif ;
- un k-itemset candidat, noté $C(k)$, est un motif de longueur k ($k \geq 1$) dont le support est inconnu. C'est-à-dire, $\forall A \in C(k)$, A est une conjonction de k prédicats (*items*) et A est susceptible d'être fréquent ;
- un k-itemset fréquent, noté $F(k)$, est un motif de longueur k ($k \geq 1$) dont le support est supérieur ou égal au support minimum (*minsupp*). C'est-à-dire, $\forall A \in F(k)$, F est une conjonction de k prédicats (*items*) et A est fréquent.

5.7.2 Approches de recherche des motifs fréquents

Classiquement, les algorithmes d'extraction des règles d'association reposent sur deux grandes étapes : la recherche des motifs fréquents ayant un support supérieur au support minimum (*minsupp*) ; et la génération, à partir de ces motifs fréquents, des règles d'association ayant une confiance supérieure à la confiance minimale (*minconf*).

La génération des règles d'association à partir d'un motif étant un problème qui se traite d'une manière identique dans tous les types d'algorithmes, la complexité d'un algorithme d'extraction de règles d'association dépend principalement de la première étape, à savoir la recherche des motifs fréquents. On retrouve dans la littérature un large éventail d'algorithmes considérés comme des variantes d'Apriori, permettant de générer tous les motifs fréquents à partir d'une base de données. Ces algorithmes reposent essentiellement sur la double propriété d'anti-monotonie : (i) tout sous-ensemble d'un motif fréquent est fréquent ; (ii) tout sur-ensemble d'un motif non fréquent est non fréquent. Selon cette double propriété, nous distinguons dans la littérature deux grandes classes d'algorithmes selon la stratégie adoptée dans la recherche des motifs fréquents :

Entrée $\mathcal{C}, \mathcal{D}_\mathcal{C}, \mathcal{D}_\mathcal{A}, \mathcal{D}_\mathcal{U}, \mathcal{R}, M, minsupp, minconf$
Sortie : $X \Rightarrow Y$, Supp, Conf, Lift, Loev

$C(1) \leftarrow \emptyset$
pour $k \leftarrow 1$ à $(s+r)$ **faire**
  $C(k) \leftarrow C(k) \cup \mathcal{A}_{kj}$
fin pour
$k \leftarrow 1$
tant que $C(k) \neq \emptyset$ **et** $k \leq (s+r)$ **faire**
  $F(k) \leftarrow \emptyset$
  pour tout $A \in C(k)$ **faire**
    si A est un prédicat inter-dimensionnels **alors**
      Supp $\leftarrow$ CalculSupport(A, M)
      si Supp $\geq$ *minsupp* **alors**
        $F(k) \leftarrow F(k) \cup \{A\}$
      fin si
    fin si
  fin pour
  pour tout $A \in F(k)$ **faire**
    pour tout $B \neq \emptyset$ **et** $B \in A$ **faire**
      si $A \backslash B \Rightarrow B$ répond à $\mathcal{R}$ **alors**
        Conf $\leftarrow$ CalculConfidence$(A \backslash B, B, M)$
        si Conf $\geq$ *minconf* **alors**
          $X \leftarrow A \backslash B$
          $Y \leftarrow B$
          Lift $\leftarrow$ CalculLift(X, Y, M)
          Loev $\leftarrow$ CalculLoevinger(X, Y, M)
          retourner $(X \Rightarrow Y$, Supp, Conf, Lift, Loev$)$
        fin si
      fin si
    fin pour
  fin pour
  $C(k+1) \leftarrow \emptyset$
  pour tout $A \in F(k)$ **faire**
    pour tout $B \in F(k)$ qui partage $k-1$ items avec A **faire**
      si Tout $Z \subset \{A \cup B\}$ ayant k items est un prédicat inter-dimensionnels **et** est fréquent **alors**
        $C(k+1) \leftarrow C(k+1) \cup \{A \cup B\}$
      fin si
    fin pour
  fin pour
  $k \leftarrow k+1$
fin tant que

Algorithme 3: Extraction des règles d'association inter-dimensionnelles à partir d'un cube de données

1. une recherche de type *descendante* (*Top-Down*) qui consiste à générer les motifs fréquents en partant des grands motifs vers les plus petits. À l'itération k, un algorithme descendant génère à partir d'un k-itemset fréquent les $(k-1)$-itemsets fréquents en se basant sur le fait que tout sous-ensemble d'un motif fréquent est fréquent ;
2. une recherche de type *ascendante* (*Bottom-Up*) qui consiste à générer les motifs fréquents en partant des petits motifs vers les motifs les plus grands. À l'itération k, un algorithme ascendant cherche dans les k-itemsets candidats ceux qui sont fréquents. En se basant sur le fait que tout sur-ensemble d'un motif non fréquent est non fréquent, seuls les k-itemsets fréquents sont utilisés pour générer les

$(k+1)$-itemsets candidats de la prochaine itération.

Dans la pratique, l'efficacité d'un algorithme ascendant de recherche de motifs fréquents, tel que Apriori, dépend naturellement du support minimum utilisé. Elle dépend aussi de la nature des données étudiées [Bla05]. Quand les données étudiées sont creuses (éparses) la deuxième propriété d'anti-monotonie devient plus efficace et réduit considérablement l'espace de recherche.

En effet, les données éparses se caractérisent naturellement par des motifs peu fréquents. Ceci revient à dire que les motifs non fréquents ont plus de chance d'être plus nombreux que les motifs fréquents dans des données éparses. Dans ce cas, il est préférable de s'attaquer au problème de recherche des motifs fréquents selon une approche ascendante qui élimine dès le départ les petits motifs non fréquents (qui sont plus nombreux) et de ne retenir que ceux qui sont fréquents (qui sont moins nombreux) et enfin de construire à partir de ces derniers les grands motifs fréquents.

Dans le cadre d'un contexte OLAP, un cube de données se caractérise en général par l'éparsité de ses données. Il convient alors, dans notre approche, d'adopter une stratégie ascendante. Nous utilisons, en particulier l'algorithme Apriori et l'adaptons aux données multidimensionnelles ainsi qu'au formalisme que nous avons établi pour l'extraction des règles selon les méta-règles inter-dimensionnelles.

5.7.3 Adaptation d'Apriori aux données multidimensionnelles

Notre algorithme adopte une stratégie itérative ascendante pour la recherche des motifs fréquents. L'algorithme 3 résume notre démarche générale de génération des règles d'association à partir d'un cube de données $\mathcal{C}$.

L'extraction des motifs fréquents est effectuée niveau par niveau. Cependant, une étape d'initialisation est nécessaire. Cette étape consiste à capturer les 1-itemsets candidats $C(1)$ à partir des dimensions d'analyse $\mathcal{D}_{\mathcal{A}}$ définies par l'utilisateur dans le cube de données $\mathcal{C}$ (lignes 1 à 4 dans l'algorithme 3). Les éléments de $C(1)$ correspondent aux modalités des ensembles $\mathcal{A}_{ij}$, où $\forall D_i \in \mathcal{D}_{\mathcal{A}}$, $\mathcal{A}_{ij}$ est l'ensemble des modalités du $j^{ième}$ niveau hiérarchique H_j^i de la dimension D_i. Les niveaux hiérarchiques des dimensions de $\mathcal{D}_{\mathcal{A}}$ sont implicitement sélectionnés par l'utilisateur lors de la définition des prédicats dimensionnels de la méta-règle inter-dimensionnelles $\mathcal{R}$. Par exemple, reprenons la méta-règle (5.4.2) définie sur le cube de données de la figure 5.3 :

> Dans le contexte $(Etudiant, Femme)$
> $\langle a_1 \in Continent\rangle \wedge \langle a_3 \in Année\rangle \Rightarrow \langle a_2 \in Produit\rangle$

Les 1-itemsets candidats, dans ce cas, correspondent à l'ensemble des motifs suivants $C(1)$= $\{\{Amérique\}, \{Europe\}, \{Asie\}, \{iTwin\}, \{iPower\}, \{DV\text{-}400\}, \{EN\text{-}700\}, \{aStar\}, \{aDream\}, \{2002\}, \{2003\}, \{2004\}, \{2005\}\}$.

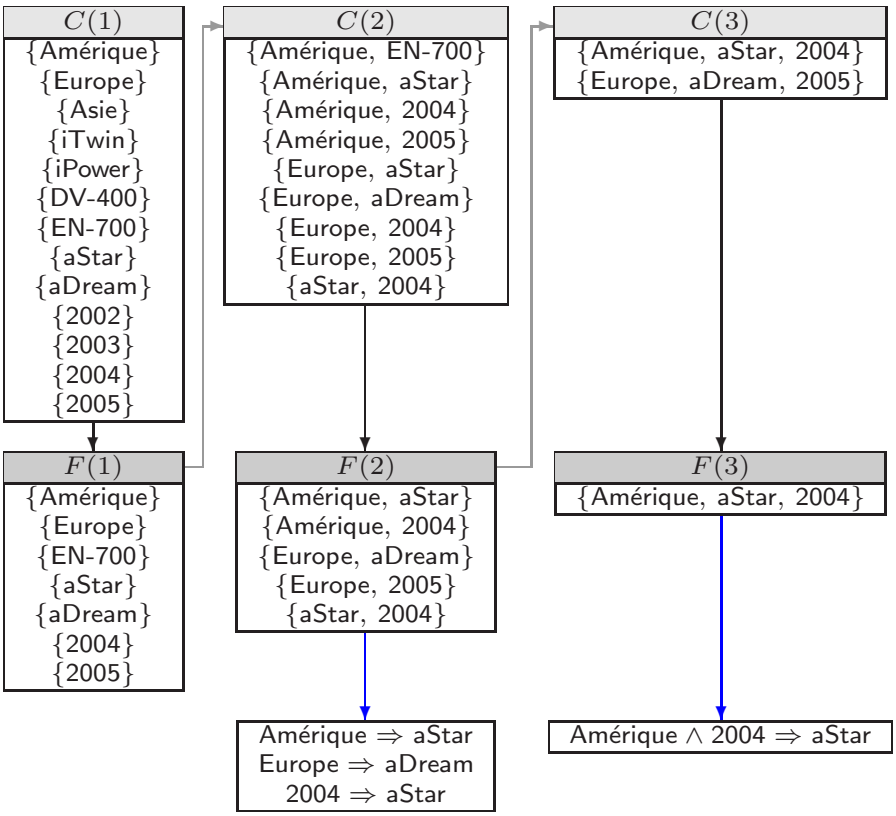

FIG. 5.5 – Exemple d'une recherche ascendante de règles d'association

Suite à l'étape d'initialisation, l'algorithme effectue un traitement itératif. Chaque itération k de l'algorithme consiste en trois étapes : (i) l'algorithme détermine l'ensemble des k-itemsets fréquents $F(k)$ à partir des k-itemsets candidats $C(k)$; (ii) génère les règles d'association selon la méta-règle inter-dimensionnelles $\mathcal{R}$; et (iii) prépare, à partir des k-itemsets fréquents $F(k)$, les $(k+1)$-itemsets candidats pour la prochaine itération $k+1$. En se référant à l'exemple de la figure 5.5 de génération de règles d'association selon la méta-règle (5.4.2) défini sur le cube de données de la figure 5.3, nous détaillons dans la suite les différentes étapes de notre algorithme :

La première étape consiste à chercher dans les k-itemsets candidats de $C(k)$ ceux qui vont appartenir à l'ensemble des k-itemsets fréquents $F(k)$ (lignes 7 à 15 dans l'algorithme 3). Pour faire partie de $F(k)$, un itemset $A \in C(k)$ doit vérifier deux

conditions :

1. A doit être une instance d'un prédicat inter-dimensionnels selon un sous-ensemble de k ($k \leq (s+r)$) dimensions de $\mathcal{D}_\mathcal{A}$. C'est-à-dire, A est une conjonction de k instances de prédicats dimensionnels de k dimensions distinctes de $\mathcal{D}_\mathcal{A}$. Par exemple, dans le processus d'extraction de la figure 5.5, le 2-itemset {*Europe*, *Amérique*} ne peut pas appartenir à $F(2)$ puisque les deux modalités *Europe* et *Amérique* sont des instances du même prédicat dimensionnel $\langle a_1 \in \mathit{Continent} \rangle$;
2. A doit être fréquent. C'est-à-dire, le support de A est supérieur ou égal au support minimum *minsupp* défini par l'utilisateur. Rappelons que, dans le cadre de notre approche, le support est calculé en fonction de l'agrégation SUM d'une mesure $M \in \mathcal{M}$ sélectionnée par l'utilisateur. Par exemple, en tenant compte de la méta-règle (5.4.2), le support du 2-itemset {*Amérique*, *aStar*} se calcule selon l'expression :

$$\text{SUPP}(\mathit{Amérique} \wedge \mathit{aStar}) = \frac{\mathsf{SUM}_M(\mathit{Amérique}, \mathit{aStar}, \mathit{All}, \mathit{All}, \mathit{Etudiant}, \mathit{Femme}, \mathit{All})}{\mathsf{SUM}_M(\mathit{All}, \mathit{All}, \mathit{All}, \mathit{All}, \mathit{Etudiant}, \mathit{Femme}, \mathit{All})}$$

La deuxième étape consiste à extraire les règles d'association à partir des k-itemsets fréquents $F(k)$ (lignes 16 à 29 dans l'algorithme 3). Soit A un itemset fréquent dans $F(k)$. Afin de générer les règles d'association candidates à partir de A, l'algorithme cherche dans A tous les sous-itemsets B non vides de A et considère, pour chaque sous-itemset B, la règle $A \backslash B \Rightarrow B$. Dans le cadre de notre approche, B est une conjonction de l instances de prédicats dimensionnels de l dimensions distinctes de $\mathcal{D}_\mathcal{A}$, où $1 \leq l < k$. Une règle $A \backslash B \Rightarrow B$ n'est retournée par l'algorithme que lorsqu'elle est jugée intéressante. Pour cela, la règle doit satisfaire deux conditions :

1. la règle $A \backslash B \Rightarrow B$ doit répondre au schéma de la méta-règle inter-dimensionnelles $\mathcal{R}$ définie selon les besoins d'analyse de l'utilisateur. C'est-à-dire, l'itemset $A \backslash B$ (respectivement, B) doit être composé d'instances de prédicats dimensionnels définis dans l'antécédent (respectivement, dans le conséquent) de la méta-règle $\mathcal{R}$. Par exemple, dans le processus d'extraction de la figure 5.5, à partir du 2-itemset {*aStar*, *2004*} de $F(2)$, l'algorithme n'autorise pas la génération de la règle *aStar* $\Rightarrow$ *2004*. En effet, selon la méta-règle (5.4.2), *aStar* est une instance du prédicat dimensionnel $\langle a_2 \in \mathit{Produit} \rangle$, qui est prévu dans le conséquent, et non pas dans l'antécédent. De même, *2004* est une instance du prédicat $\langle a_3 \in \mathit{Année} \rangle$, qui est prévu dans l'antécédent, et non pas dans le conséquent des règles à générer. En revanche, la règle *2004* $\Rightarrow$ *aStar* est acceptée vue que l'emplacement de ses éléments répondent bien à celui imposé par la méta-règle (5.4.2) ;
2. la règle $A \backslash B \Rightarrow B$ doit avoir une confiance supérieure ou égale à la confiance minimale *minconf* définie par l'utilisateur. La confiance d'une règle est également

calculée en fonction de l'agrégation SUM d'une mesure $M \in \mathcal{M}$ sélectionnée par l'utilisateur. Par exemple, en tenant compte de la méta-règle (5.4.2), la confiance de la règle *2004* $\Rightarrow$ *aStar* se calcule selon l'expression :

$$\textsc{Conf}(\textit{2004} \Rightarrow \textit{aStar}) = \frac{\mathsf{SUM}_M(\textit{All}, \textit{aStar}, \textit{2004}, \textit{All}, \textit{Etudiant}, \textit{Femme}, \textit{All})}{\mathsf{SUM}_M(\textit{All}, \textit{All}, \textit{2004}, \textit{All}, \textit{Etudiant}, \textit{Femme}, \textit{All})}$$

Dans le cas où les deux conditions précédentes sont vérifiées, notre algorithme calcule le *Lift* et l'indice de *Loevinger*. La règle $A \backslash B \Rightarrow B$ est ainsi retournée avec les valeurs de son support, sa confiance, son *Lift* et son indice de *Loevinger*. Dans le cas contraire, la règle est rejetée.

La troisième étape consiste à générer à partir des k-itemsets fréquents $F(k)$ un nouvel ensemble de $(k+1)$-itemsets candidats $C(k+1)$ (lignes 30 à 37 dans l'algorithme 3). Un $(k+1)$-itemset candidat est généré selon l'union de deux k-itemsets fréquents A et B de $F(k)$, où A et B vérifient les trois conditions suivantes :

1. A et B doivent partager $(i-1)$ éléments communs. Par exemple, dans le processus d'extraction de la figure 5.5, les deux 2-itemsets fréquents {*Amérique*, *aStar*} et {*Europe*, *2005*} de $F(2)$ ne répondent pas à cette condition puisqu'ils n'ont pas un élément commun. Ce qui signifie que ces deux 2-itemsets ne peuvent pas générer de 3-itemsets candidats. En revanche, les deux 2-itemsets fréquents {*Amérique*, *aStar*} et {*Amérique*, *2004*} de $F(2)$ répondent bien à cette condition et, par conséquent, le 3-itemset {*Amérique*, *aStar*, *2004*} peut appartenir à $C(3)$;
2. tout sous-itemset non vide de $A \cup B$ doit être une instance d'un prédicat inter-dimensionnels dans $\mathcal{D}_{\mathcal{A}}$. C'est-à-dire, chaque sous-itemset non vide de $A \cup B$ doit correspondre à des instances de prédicats dimensionnels dans des dimensions distinctes de $\mathcal{D}_{\mathcal{A}}$;
3. tout sous-itemset non vide de $A \cup B$ doit être fréquent.

Enfin, l'algorithme s'**arrête** (ligne 6 dans l'algorithme 3) quand l'une des deux conditions suivantes est vérifiée :

1. quand l'algorithme aurait fait $(s+r)$ itérations, où $(s+r)$ correspond au nombre de dimensions d'analyse $\mathcal{D}_{\mathcal{A}}$;
2. quand il n'y a plus de motifs candidats. C'est-à-dire, l'algorithme peut s'arrêter à une itération $k < (s+r)$ quand il n'arrive plus à générer de $(k+1)$-itemsets candidats à partir des k-itemsets fréquents $F(k)$.

5.7.4 Calcul des indices des règles inter-dimensionnelles

Il est à noter que notre algorithme d'extraction des règles inter-dimensionnelles fait appel à des fonctions externes pour le calcul du support (CalculSupport), de la confiance (CalculConfiance), du *Lift* (CalculLift) et de l'indice de *Loevinger* (CalculLoevinger) (lignes 10, 19, 23 et 24 dans l'algorithme 3). Ces fonctions emploient, d'une manière dynamique, des requêtes MDX afin d'extraire les valeurs des agrégats nécessaires pour le calcul de chaque indice. MDX est un langage de requêtes adapté à la structure multidimensionnelle des cubes de données. Il permet de lancer des requêtes sur un cube de données et de retourner des jeux de cellules multidimensionnelles contenant les données du cube.

Reprenons l'exemple de l'extraction des règles d'association de la figure 5.5 à partir du cube des ventes la figure 5.3, selon la méta-règle (5.4.2) et la mesure des bénéfices des ventes. Soit R_1 : *Amérique* $\wedge$ *2004* $\Rightarrow$ *aStar* une règle d'association extraite dans le cadre de cet exemple. À titre indicatif, nous exposons dans la suite le mécanisme des requêtes MDX employé pour le calcul du support et de la confiance de la règle R_1. Afin d'alléger les notations, nous reprenons celle de la section 5.6 et considérons que $X =$ *Amérique* $\wedge$ *2004* et $Y =$ *aStar*.

Calcul du support

Selon le formalisme développé dans la section 5.5, le support de R_1 s'écrit selon l'expression :

$$\text{Supp}(R_1) = \frac{\text{SUM}_{Bénéfice}(Amérique, aStar, 2004, All, Etudiant, Femme, All)}{\text{SUM}_{Bénéfice}(All, All, All, All, Etudiant, Femme, All)}$$

Pour calculer la valeur de cette expression, notre algorithme fait appel à la fonction CalculSupport($X \wedge Y$, *Bénéfice*) qui interroge dynamiquement le cube de données via les deux requêtes MDX suivantes. La première requête retourne la valeur de l'agrégation SUM des bénéfices du numérateur et la deuxième retourne celle de l'agrégation SUM des bénéfices du dénominateur de l'expression de Supp(R_1).

$\text{SUM}_{Bénéfice}(Amérique, aStar, 2004, All, Etudiant, Femme, All)$

```
SELECT
NON EMPTY {[Lieu].[Continent].[Amérique]} ON AXIS(0),
NON EMPTY {[Temps].[Année].[2004]} ON AXIS(1),
NON EMPTY {[Produit].[Produit].[aStar]} ON AXIS(2)
FROM Ventes
WHERE ([Measures].[Bénéfice],
       [Profession].[Catégorie de profession].[Etudiant],
       [Sexe].[Sexe].[Femme])
```

$\mathsf{SUM}_{Bénéfice}(All, All, All, All, Etudiant, Femme, All)$

```
SELECT
NON EMPTY {[Profession].[Catégorie de profession].[Etudiant]} ON AXIS(0),
NON EMPTY {[Sexe].[Sexe].[Femme]} ON AXIS(1)
FROM Ventes
WHERE ([Measures].[Bénéfice])
```

Calcul de la confiance

Selon le formalisme développé dans la section 5.5, la confiance de R_1 s'écrit selon l'expression :

$$\textsc{Conf}(R_1) = \frac{\mathsf{SUM}_{Bénéfice}(Amérique, aStar, 2004, All, Etudiant, Femme, All)}{\mathsf{SUM}_{Bénéfice}(Amérique, All, 2004, All, Etudiant, Femme, All)}$$

Pour calculer la valeur de cette expression, notre algorithme fait appel à la fonction $\textsc{CalculConfiance}(X, Y, Bénéfice)$. Le numérateur de l'expression $\textsc{Conf}(R_1)$ étant déjà calculé dans l'expression de $\textsc{Supp}(R_1)$, la fonction CalculConfiance ne fait appel qu'à la requête MDX suivante pour retourner la valeur de l'agrégation SUM des bénéfices du dénominateur de $\textsc{Conf}(R_1)$.

$\mathsf{SUM}_{Bénéfice}(Amérique, All, 2004, All, Etudiant, Femme, All)$

```
SELECT
NON EMPTY {[Lieu].[Continent].[Amérique]} ON AXIS(0),
NON EMPTY {[Temps].[Année].[2004]} ON AXIS(1)
FROM Ventes
WHERE ([Measures].[Bénéfice],
       [Profession].[Catégorie de profession].[Etudiant],
       [Sexe].[Sexe].[Femme])
```

5.8 Visualisation des règles inter-dimensionnelles

Dans cette section, nous proposons une approche pour la visualisation des règles d'association inter-dimensionnelles découvertes à partir d'un cube de données. Cette approche a pour objectif de pallier les problèmes liés aux grandes quantités de règles d'association générées. La visualisation des règles d'association est un moyen pour valoriser les connaissances induites par ces dernières. Elle permet de rendre plus exploitable les règles et contribue à l'accélération du processus d'analyse.

Dans le cadre du contexte de l'analyse en ligne, notre approche intègre la visualisation des règles d'association dans l'environnement multidimensionnel. En

effet, par analogie au principe de la navigation OLAP dans les données, notre approche tente d'établir un cadre de navigation et d'exploration interactive des règles d'association en se basant sur des principes de visualisation simples, efficaces et surtout compréhensibles par les utilisateurs OLAP. Pour cela, nous utilisons les principes de la *sémiologie graphique* de *Bertin* [Ber67].

5.8.1 Principes de visualisation de *Bertin*

Les travaux de *Bertin* [Ber67], issus de la cartographie, sont comptés parmi les plus anciennes propositions dans le domaine de l'encodage graphique de l'information. Ces travaux sont toujours d'actualité et font référence dans le domaine de la visualisation d'information sur ordinateur [Bla05].

Les principes de sémiologie graphique de *Bertin*, proposés dans [Ber67], consistent à organiser des éléments visuels d'une ou de plusieurs informations selon des variables graphiques. Les variables graphiques de *Bertin* incluent : la *position*, la *taille*, la *luminosité*, la *texture*, la *couleur*, l'*orientation* et la *forme*.

La position est une variable particulièrement importante puisqu'elle exprime une information visuelle perceptivement dominante dans une représentation graphique. La répartition des éléments graphiques dans l'espace de représentation visuel permet à la rétention humaine d'établir implicitement un ordre géométrique des informations portées par ces éléments. Quant aux autres variables, elles sont appelées *variables rétiniennes* [Ber67] car il est possible de percevoir leurs variations indépendamment de la position de leurs éléments associés sans mettre à contribution les muscles de l'œil humain.

Il est à noter que la variable de la taille concerne plus les surfaces que les longueurs. *Bertin* affirme que l'œil humain est plus sensible aux variations de la surface d'une forme qu'aux variations de ses longueurs. Ainsi, les encodages graphiques basés sur les surfaces sont plus pertinents que ceux basés sur les longueurs.

5.8.2 Codage graphique d'une règle d'association

En se basant sur les principes de visualisation de *Bertin*, nous proposons de représenter une règle d'association selon un encodage graphique qui tient compte de ses différents indices, à savoir : son support, sa confiance, son *Lift* et son indice de *Loevinger*. Notre encodage fait intervenir les variables graphiques de la forme, la taille, la luminosité et la couleur.

Soit une règle d'association de forme générale $X \Rightarrow Y$. Nous représentons une telle règle selon les critères suivants :

- l'itemset $\{X, Y\}$ est codé par un carré de couleur bleu ;
- le support de la règle est codé par la surface du carré ;

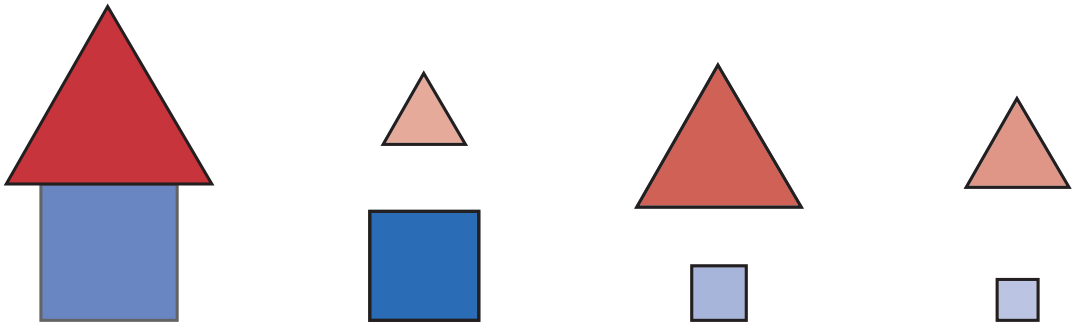

Fig. 5.6 – Exemples de représentations graphiques de règles d'association

- le *Lift* de la règle est codé par la luminosité du carré bleu ;
- l'implication $X \Rightarrow Y$ est codée par un triangle isocèle de couleur rouge ;
- la confiance de la règle est codée par la surface du triangle ;
- l'indice de *Loevinger* est codé par la luminosité du triangle rouge.

Pour une règle d'association, nous utilisons deux formes différentes pour distinguer entre l'itemset de la règle et l'implication de cette dernière. En plus, nous choisissons d'attribuer une couleur différente pour chaque forme pour mieux distinguer entre les deux notions. Nous verrons plus loin que les deux formes (carré et triangle) seront aussi associées au critère graphique de la *position*.

Nous utilisons la surface pour représenter les ordres de grandeur. Une forme ayant une grande surface représente naturellement une entitée associée à une grande valeur. Plus le support (respectivement, la confiance) d'une règle est grand, plus la surface du carré (respectivement, du triangle) est grande. Par exemple, comme le montre la figure 5.6, la règle d'association (a) ayant un support de 100% est représentée par un carré de surface plus grande que celui de la règle (d) qui a un support de 30% seulement. La surface étant une variable graphique pertinente à la perception humaine, il est plus convenable de représenter les indices des règles d'association les plus employés, à savoir le support et la confiance, avec cette variable.

Le *Lift* et l'indice de *Loevinger* sont plutôt codés selon l'intensité de la luminosité de leur forme respective. Une luminosité intense d'une forme correspond à une couleur claire de cette dernière. En revanche, une faible luminosité d'une forme correspond à une couleur plutôt foncée de cette dernière. Ainsi, nous associons les grandes valeurs du *Lift* (respectivement, de l'indice de *Loevinger*) à des faibles luminosités du carré

bleu (respectivement, du triangle rouge). D'un autre côté, nous associons les petites valeurs du *Lift* (respectivement, de l'indice de *Loevinger*) à des luminosités intenses du carré bleu (respectivement, du triangle rouge). Par exemple, la règle d'association (b) de la figure 5.6 a un indice de *Loevinger* plus petit que celui de la règle (a), ce qui explique la différence de l'intensité de luminosité de leurs triangles rouges.

5.8.3 Visualisation des règles d'association

Nous proposons d'utiliser le codage graphique des règles d'association dans notre approche de visualisation. Cette dernière a pour objectif de mieux exploiter via des outils graphiques les connaissances véhiculées par les règles d'association. L'idée clé consiste à intégrer les règles d'association dans l'espace de représentation du cube de données qui a servi pour leur génération. Nous offrons ainsi à l'utilisateur OLAP la possibilité de naviguer d'une manière interactive dans le cube pour explorer aussi bien les faits de ce cube que les associations existantes entre ses données.

Classiquement, un utilisateur OLAP observe les valeurs des faits d'un cube de données selon des axes d'analyse. Les valeurs des faits correspondent aux mesures du cube et les axes d'analyse correspondent aux dimensions d'analyse. Ainsi, l'utilisateur choisit une mesure $M \in \mathcal{M}$ et un ensemble de dimensions d'analyse $\mathcal{D}_{\mathcal{A}} \subseteq \mathcal{D}$ et visualise dans un éditeur graphique OLAP le croisement de ces dimensions.

Ce croisement produit un ensemble de cellules, où chaque cellule représente un fait OLAP qui est la conjonction de modalités distinctes provenant chacune d'une dimension de $\mathcal{D}_{\mathcal{A}}$. Une cellule contient également la valeur du fait qu'elle représente selon la mesure M. L'ensemble des ces cellules constitue ainsi un espace de représentation *numérique* du cube de données.

En général, dans la plupart des éditeurs OLAP, l'affichage d'un tel espace de représentation se fait selon une visualisation plane où les modalités de plusieurs dimensions peuvent être emboîtées en lignes et/ou en colonnes d'un *tableau croisé dynamique*. À titre d'exemple, la figure 5.7 représente l'affichage d'une vue d'un cube de données selon un tableau croisé dynamique dans MicroSoft SQL Server 7.0.

Dans notre approche de visualisation des règles d'association, nous partons de ce même principe d'affichage des vues d'un cube de données. Il est à noter que, mise à part son contenu, la position d'une cellule dans cet espace de représentation correspond à un prédicat inter-dimensionnels dans l'ensemble des dimensions d'analyse retenues pour l'affichage de la vue du cube. C'est-à-dire, une conjonction d'instances de prédicats dimensionnels distincts. En d'autres termes, selon notre formalisme des règles d'association inter-dimensionnelles, chaque cellule dans l'espace de représentation du cube est identifiée par un motif qui peut correspondre à un des cas de figure suivants :

– le motif n'est pas fréquent ;

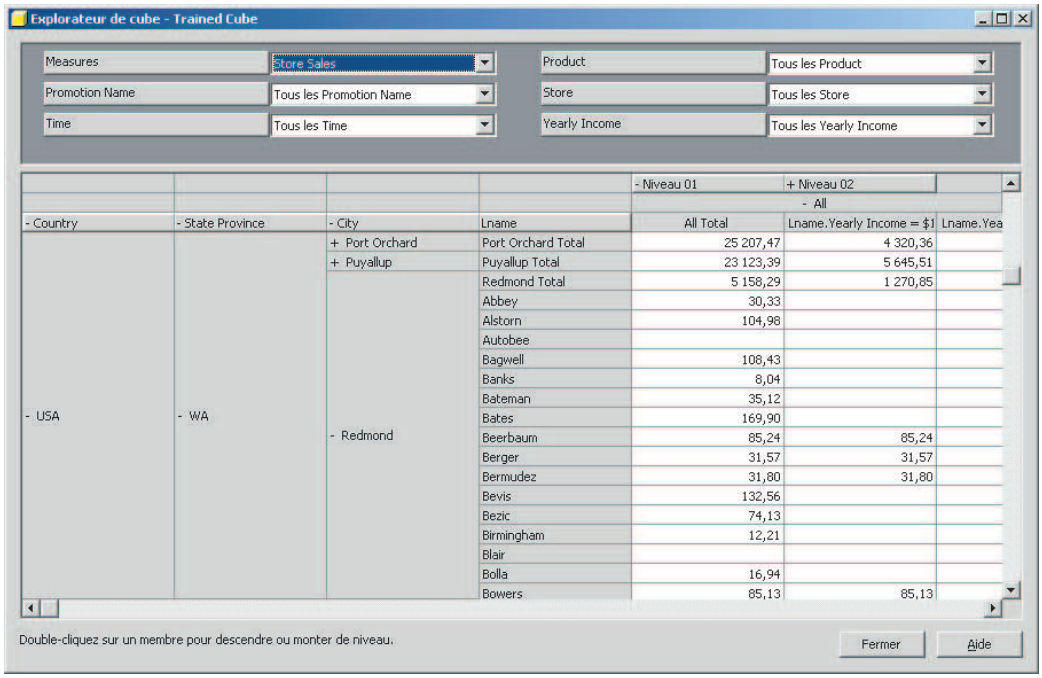

FIG. 5.7 – Tableau croisé dynamique dans MicroSoft SQL Server 7.0

– le motif est fréquent, mais ne génère pas de règles d'association ;
– le motif est fréquent et génère une seule règle d'association ;
– le motif est fréquent et génère plusieurs règles d'association.

Par exemple, soit une cellule c dans l'espace de représentation d'une vue plane d'un cube de données $\mathcal{C}$. La position de c dans cette représentation correspond au croisement de X en ligne et de Y en colonne. X et Y sont des conjonctions de modalités provenant chacune d'une dimension distincte de l'ensemble des dimensions d'analyse retenues pour l'affichage. En d'autres termes, X et Y sont des instances de prédicats inter-dimensionnels distincts dans les dimensions d'analyse retenues pour l'affichage. Par conséquent, la cellule c correspond au motif $\{X, Y\}$. Selon les propriétés du motif $\{X, Y\}$, nous proposons de représenter dans la cellule c un codage graphique approprié :

– si $\{X, Y\}$ n'est pas fréquent, seule la valeur de la mesure M, si elle existe, est affichée dans la cellule c (voir l'exemple de figure 5.8 (a)) ;
– si $\{X, Y\}$ est fréquent et ne génère pas de règles d'association, un carré blanc est affiché dans la cellule c. La surface du carré renseigne sur le support du motif. Plus le support est grand, plus la surface du carré est grande. Il est à noter que la couleur bleue du carré, dont la luminosité code la valeur du *Lift*, est éliminée.

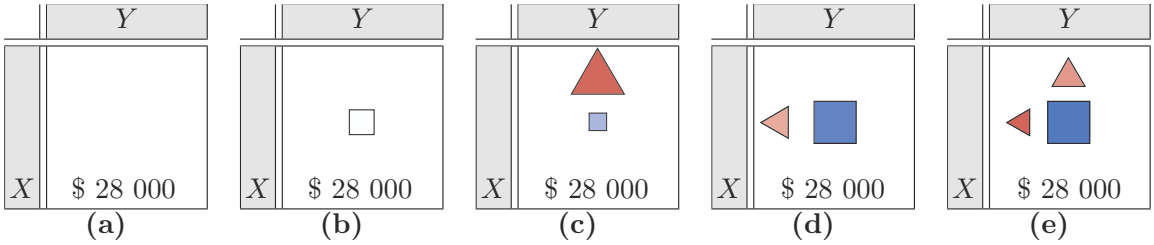

FIG. 5.8 – Exemples de visualisation d'une cellule d'un cube de données

En effet, dans ce cas de figure, on ne peut pas parler de *Lift* puisqu'il ne s'agit pas d'une règle d'association (voir l'exemple de figure 5.8 (b)) ;
- si $\{X, Y\}$ est fréquent et génère la règle d'association $X \Rightarrow Y$, un carré bleu et un triangle isocèle rouge sont affichés dans la cellule c. Le triangle pointe vers Y selon le sens de l'implication de la règle (voir l'exemple de figure 5.8 (c)) ;
- si $\{X, Y\}$ est fréquent et génère la règle d'association $Y \Rightarrow X$, un carré bleu et un triangle isocèle rouge sont affichés dans la cellule c. Le triangle pointe vers X selon le sens de l'implication de la règle (voir l'exemple de figure 5.8 (d)) ;
- si $\{X, Y\}$ est fréquent et génère les règles d'association $X \Rightarrow Y$ et $Y \Rightarrow X$, un carré bleu et deux triangles isocèles rouges sont affichés dans la cellule c. Le premier triangle pointe vers Y selon le sens de l'implication de la règle $X \Rightarrow Y$ et le second triangle pointe vers X selon le sens de l'implication de la règle $Y \Rightarrow X$ (voir l'exemple de figure 5.8 (e)).

5.9 Expérimentations et performances

Afin d'évaluer les performances de notre algorithme d'extraction de règles d'association inter-dimensionnelles, nous avons mené un ensemble d'expérimentations. Celles-ci ont été réalisées sous un environnement Windows XP sur une machine de 480MB de mémoire vive, un processeur Intel Pentium 4 avec une fréquence de 1,60GHz.

La figure 5.9 montre le temps de d'exécution de notre algorithme en fonction du support minimum selon différents seuils de confiances minimales. On remarque que, en général, le temps d'exécution de l'algorithme décroît en fonction du support minimum. Plus le support minimum est grand, plus l'algorithme devient rapide. Ceci s'explique par la propriété d'anti-monotonie qui, pour des seuils élevés du support minimum, élague d'une manière significative les motifs non fréquents dès les premières itérations de l'algorithme. On note aussi que, quand il s'agit de seuils élevés de la confiance minimale, le temps d'exécution de l'algorithme baisse globalement.

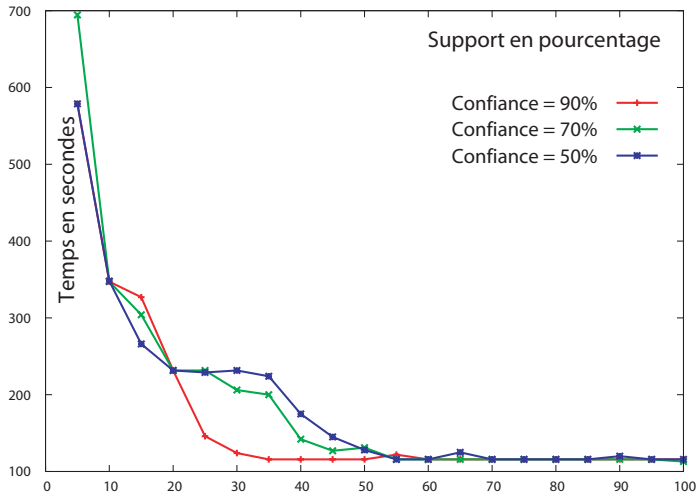

FIG. 5.9 – Temps d'exécution de l'algorithme en fonction du support minimum selon différentes confiances minimales

Cependant, contrairement au support minimum, la confiance minimale n'influence pas d'une manière sensible la rapidité de l'algorithme.

La figure 5.10 résume des tests de performance de l'algorithme pour des cubes de données de différents volumes en fonction du support minimum. Chaque cube est caractérisé par le nombre de faits qu'il contient. On remarque que pour les petites valeurs du support minimum (moins de 40%), le nombre de faits étudiés est un élément déterminant dans la rapidité de l'algorithme. En revanche, pour les grandes valeurs du support minimum, le nombre de faits n'a pratiquement aucune influence sur le temps d'exécution de l'algorithme. En effet, quelque soit le nombre de faits, l'algorithme garde globalement le même temps d'exécution.

La figure 5.11 confirme en partie le précédent constat. En effet, pour un support et une confiance minimums égaux à 5%, on note que la performance de l'algorithme dépend fortement du nombre de motifs fréquents et de règles d'association générés. Le temps d'exécution de l'algorithme croît d'une manière remarquable en fonction du nombre de motifs fréquents et de règles d'association. Cependant, on note également que la génération des règles d'association à partir des motifs fréquents consomme plus de temps que la génération des motifs fréquents. En effet, particulièrement pour les données éparses, la propriété d'anti-monotonie est capable de réduire considérablement la complexité de la recherche ascendante des motifs fréquents. Cependant, cette propriété n'a aucun pouvoir sur la génération des règles à partir des motifs. Pour chaque motif fréquent, l'algorithme doit générer toutes les règles possibles et chercher celles qui répondent au schéma de la méta-règle et qui ont une

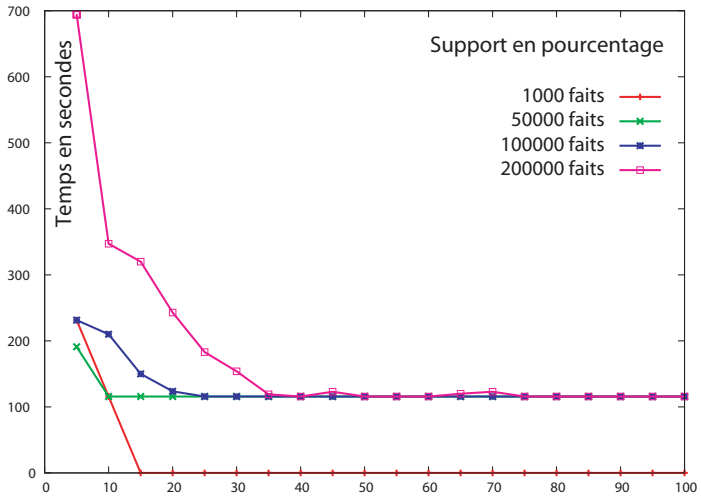

FIG. 5.10 – Temps d'exécution de l'algorithme en fonction du support minimum selon différents nombres de faits

confiance supérieure ou égale à la confiance minimale.

5.10 Conclusion et perspectives

Dans ce chapitre, nous avons proposé une troisième approche de couplage entre l'analyse en ligne et la fouille de données. Cette dernière établi un cadre général pour l'extraction des règles d'association inter-dimensionnelles pour l'explication dans les cubes de données. Cette nouvelle approche couple les règles d'association avec la technologie OLAP en adaptant l'algorithme de recherche des règles au contexte des données multidimensionnelles. Selon cette approche, aucun pré-traitement préalable est nécessaire sur les cubes de données. L'algorithme que nous proposons est une adaptation d'Apriori aux données multidimensionnelles. Il repose sur une recherche ascendante des motifs fréquents qui exploite la propriété d'anti-monotonie particulièrement adaptée aux données éparses.

Nous avons employé les méta-règles inter-dimensionnelles afin de piloter le processus de recherche des règles dans un cube de données. Ainsi, un utilisateur peut cibler un contexte d'analyse spécifique défini par une portion particulière dans le cube étudié.

Nous avons également revisité les principes classiques du support et de la confiance d'une règle d'association. Nous proposons un formalisme qui redéfinit ces derniers en offrant la possibilité de les calculer en fonction des unités de masse d'une mesure

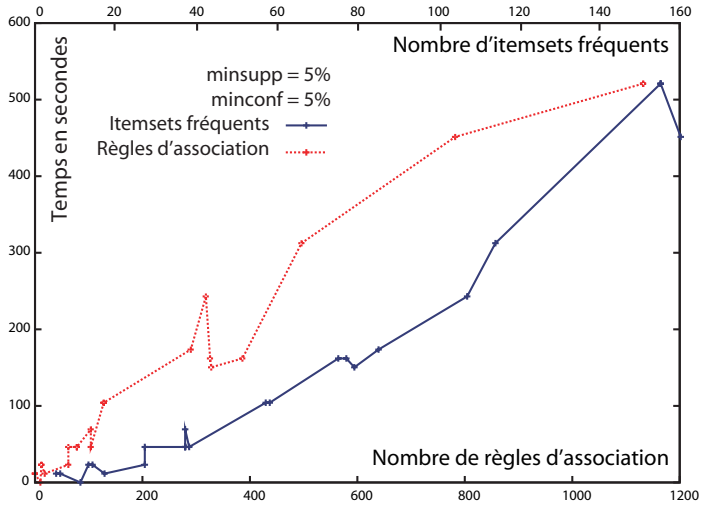

FIG. 5.11 – Temps d'exécution de l'algorithme en fonction du nombre d'itemsets fréquents et en fonction des règles d'association

choisie par l'utilisateur. Nous avons montré que cette nouvelle façon d'évaluer une règle d'association est plus pertinente au sens d'une analyse en ligne.

En général, le support et la confiance entraînent la génération d'un grand nombre de règles d'association qui sont inexploitables dans la plupart des cas. Pour cela, nous proposons de filtrer les règles extraites en ne gardant que celles les plus intéressantes aux sens du critère du *Lift* et de l'indice de *Loevinger*.

Afin de valoriser les règles d'association extraites, nous avons proposé un codage graphique de ses dernières selon la *sémiologie graphique* de *Bertin* [Ber67]. Ce codage prend en compte l'ordre d'importance de chaque règle en fonction des valeurs de ses critères d'évaluation. Nous utilisons également ce codage dans le cadre d'une nouvelle approche de visualisation des règles d'association dans un espace de représentation du cube de données étudié.

Suite à ce travail, des améliorations possibles et de nouvelles pistes de recherche méritent d'être étudiées. Tout d'abord, nous pensons qu'il est aussi intéressant d'intégrer la valeur de la mesure dans l'expression de la règle inter-dimensionnelles. La mesure peut aussi faire l'objet d'un codage graphique intégré dans celui de la règle. Ainsi, nous pouvons offrir à l'utilisateur une visualisation complète de l'espace de représentation du cube de données incluant les mesures des faits OLAP et les liens entre ces faits par les règles d'association.

Vu le grand nombre de travaux sur les règles d'association dans les cubes de données, nous pensons qu'il est nécessaire d'élaborer une étude comparative afin de

positionner notre approche, en terme de performance, par rapport aux approches existantes.

Enfin, nous projetons aussi la généralisation de notre approche et son extension aux règles d'association inter-dimensionnelles avec prédicats répétitifs et aux règles d'association intra-dimensionnelles. Une autre amélioration possible consisterait à mieux profiter de l'aspect hiérarchique des dimensions du cube de données étudié afin d'en extraire des règles d'association avec des prédicats appartenant à plusieurs niveaux de granularité.

Implémentation et cas d'application aux données complexes

Résumé

Dans ce chapitre, nous présentons une plateforme logicielle que nous avons mis en place pour concrétiser, sur un plan technique, nos contributions théoriques sur le couplage entre l'analyse en ligne et la fouille de données. Nous proposons aussi un cas d'application aux données complexes avec l'agrégation par classification. Nous utilisons un jeu de données médicales relatives au domaine du dépistage du cancer du sein.

Dans une étape préliminaire, nous préparons et traitons ce jeu de données à partir duquel nous construisons un cube de données complexes. Pour cela, nous mettons en œuvre une méthodologie générale d'entreposage de données complexes basée sur le formalisme XML.

Sommaire

Publications

[BMCA06a] Boussaid O., Messaoud R.B., Choquet R., Anthoard S., « Conception et construction d'entrepôts XML », in $2^{ème}$ *journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'2006)*, *Revue des Nouvelles Technologies de l'Information*, pp. 3–21, Versailles, France : Cépaduès Editions. Juin 2006.

[BMCA06b] Boussaid O., Messaoud R.B., Choquet R., Anthoard S., « X-Warehousing : an XML-Based Approach for Warehousing Complex Data », in *Proceedings of the* 10^{th} *East-European Conference on Advances in Databases and Information Systems (ADBIS'2006)*, pp. 39–54, Thessaloniki, Greece : Springer-Verlag. September 2006.

[MBR06a] Messaoud R.B., Boussaid O., Rabaséda S.L., « A Data Mining-Based OLAP Aggregation of Complex Data : Application on XML Documents », *International Journal of Data Warehousing and Mining*, 2(4) :1–26. 2006.