# Cubes Fermés / Quotients Émergents

Sébastien Nedjar, Alain Casali, Rosine Cicchetti & Lotfi Lakhal

Laboratoire d'Informatique Fondamentale de Marseille (LIF), CNRS UMR 6166
Aix-Marseille Université,
IUT d'Aix-en-Provence, Avenue Gaston Berger, 13625 Aix-en-Provence Cedex
prenom.nom@lif.univ-mrs.fr

**Résumé.** Le concept de Cube Émergent a été introduit afin de comparer deux data cubes. Dans cet article, nous introduisons deux nouvelles représentations réduites du Cube Émergent sans perte des mesures : le Cube Fermé Émergent et le Cube Quotient Émergent. La première représentation est basée sur le concept de fermeture cubique. C'est la plus petite représentation possible du cube de données émergent. À partir du Cube Fermé Émergent et donc en stockant le minimum d'informations, il est possible de répondre efficacement aux requêtes qui peuvent être exécutées sur le Cube Émergent lui-même. La seconde représentation s'appuie sur la structure du Cube Quotient qui a été proposé pour résumer un cube de données. Le Cube Quotient est revisité afin de le doter d'une sémantique basée sur la fermeture cubique et donc adapté au contexte du Cube Émergent. Le Cube Quotient Émergent résultant est moins réduit que le Cube Fermé Émergent mais il préserve la propriété de " spécialisation/généralisation " du data cube qui permet la navigation au sein du Cube Émergent. Nous établissons également le lien entre les deux représentations introduites et celle basée sur les bordures classiques en fouille de données. Des expérimentations effectuées sur divers jeux de données visent à comparer la taille des différentes représentations.

## 1 Introduction

Afin de capturer les renversements de tendances dans les bases de données OLAP, le concept de Cube Émergent (Nedjar et al., 2009) a été proposé. Ce concept résulte du couplage de deux structures intéressantes, le data cube (Gray et al., 1997) et les motifs émergents (Dong et Li, 2005), dans le même esprit que le Skycube (Pei et al., 2006) qui combine le concept de data cube et la dominance de Pareto (Skyline (Börzsönyi et al., 2001)). À partir des cubes de deux relations d'une base de données pourvues de dimensions (ou d'attributs catégories), le Cube Émergent rassemble tous les tuples satisfaisant une *double contrainte d'émergence* : la valeur de leur mesure est faible dans une relation (contrainte $C_1$) et significative dans l'autre (contrainte $C_2$).

Offrir aux décideurs des Cubes Émergents est loin d'être trivial car deux data cubes, probablement volumineux donc coûteux à obtenir, à stocker et à gérer, doivent être calculés pour ensuite être comparés. En abordant cette problématique, notre idée est d'étudier des représen-

tations réduites du Cube Émergent afin de minimiser chaque facette du problème. Des représentations réduites ont été largement étudiées en fouille de données et ont montré un véritable intérêt. Citons, par exemple, celles proposées pour les motifs fréquents (Pasquier et al., 1999; Zaki et Hsiao, 2002). Dans le contexte OLAP, différentes approches visent la réduction de la taille des cubes de données, par exemple en éliminant les redondances intrinsèques au sein des cubes (Lakshmanan et al., 2002; Casali et al., 2003; Morfonios et Ioannidis, 2006) ou en se focalisant sur les tendances les plus intéressantes à travers des cubes icebergs (Beyer et Ramakrishnan, 1999). Des représentations réduites pour les Cubes Émergents qui s'appuient sur les bordures (Nedjar et al., 2009) ont été proposées. Néanmoins, ces représentations ne permettent pas de retrouver les valeurs des mesures et donc ne peuvent pas être utilisées pour répondre à n'importe quelle requête OLAP.

Dans cet article, nous proposons deux représentations du Cubes Émergents qui éliminent les redondances, sans perte d'information et, bien sûr, évitent de calculer les deux cubes de données à comparer. La première, appelée Cube Fermé Émergent (cf. section 3), est la représentation la plus réduite. Elle est basée sur le concept de fermeture cubique (Casali et al., 2003). À partir de cette représentation, il est possible de retrouver les valeurs des mesures associées aux différentes tendances. Ainsi, le résultat de toute requête OLAP peut en être dérivé. Par exemple, nous pouvons répondre au type de question suivant : « Comment évoluent les ventes de tel produit entre 2007 et 2008 par ville et par saison ? ». Certaines applications OLAP nécessitent non seulement de traiter des requêtes mais aussi de disposer de capacités de navigation au sein des cubes. En effectuant une telle navigation, l'utilisateur observe les données à différents niveaux de granularité. Par exemple, si un renversement de tendances très significatif apparaît à un niveau très agrégé, l'utilisateur peut vouloir cerner les origines du phénomène et donc « forer » dans le cube pour obtenir les données détaillées sous-jacentes. Néanmoins, le Cube Fermé Émergent n'offre pas de telles capacité de navigation. C'est pourquoi nous proposons une seconde représentation qui, elle, les permet. Elle est basée sur la représentation par Cube Quotient (Lakshmanan et al., 2002) mais elle n'est pas une simple adaptation de la structure citée à notre problématique de capture des tendances émergentes car, pour la caractériser, nous devons établir le lien entre le concept de fermeture cubique et le quotient cube. La représentation proposée est évidemment appelée Cube Quotient Émergent (cf. section 4). De plus, nous établissons le lien existant entre les différentes représentations proposées qui est un lien d'inclusion (cf. section 4.3). Ce résultat analytique permet de guider le choix de la représentation la mieux adaptée au futur usage des Cubes Émergents.

Enfin, nous effectuons des expérimentations (cf. section 5) afin de comparer la taille des représentations réduites et celle du Cube Émergent. Les résultats obtenus sont très positifs : pour les données réelles connues pour être fortement corrélées, les deux représentations apportent une importante réduction de taille.

## 2 Cubes Émergents

Considérons une relation $r$ dotée d'un ensemble d'attributs dimensions (notés $D_1$, $D_2$, ..., $D_n$) et d'une mesure (notée $M$). La caractérisation du Cube Émergent s'inscrit dans le contexte plus général du treillis cube de la relation $r$ : $CL(r)$ (Casali et al., 2003). Ce dernier est un espace de recherche bien adapté au calcul du cube de données de $r$. Il organise les tuples, solutions possibles du problème, selon un ordre de généralisation / spécialisation, noté

| Produit | Ville | Saison | Quantité |
|---|---|---|---|
| 1 | Marseille | Printemps | 100 |
| 1 | Marseille | Été | 100 |
| 2 | Paris | Été | 100 |
| 3 | Paris | Été | 100 |

TAB. 1: Relation exemple VENTE$_{07}$

| Produit | Ville | Saison | Quantité |
|---|---|---|---|
| 2 | Marseille | Printemps | 200 |
| 2 | Paris | Été | 100 |
| 1 | Marseille | Printemps | 100 |
| 3 | Paris | Été | 100 |
| 3 | Paris | Automne | 300 |

TAB. 2: Relation exemple VENTE$_{08}$

$\preceq_g$(Lakshmanan et al., 2002). Ces tuples sont structurés selon les attributs dimensions de $r$ et ces derniers peuvent prendre la valeur ALL (Gray et al., 1997). De plus, nous ajoutons à ces tuples un tuple virtuel ne contenant que des valeurs vides afin de fermer la structure. Tout tuple du treillis cube généralise le tuple de valeurs vides. Pour manipuler les tuples de $CL(r)$, l'opérateur + est défini. A partir d'un couple de tuples, il retourne le tuple le plus spécifique de $CL(r)$ qui généralise les deux opérandes.

**Exemple 2.1 -** Considérons la relation VENTE$_{07}$ (*cf.* Table 1) donnant les quantités de produit vendues par Type, Ville, et Saison. Dans $CL(\text{VENTE}_{07})$, considérons les ventes de produit 1 à Marseille, quelque soit la saison, $i.e$ le tuple (1, Marseille, ALL). Ce tuple est spécialisé par les deux tuples suivants de la relation : (1, Marseille, Été) et (1, Marseille, Printemps). De plus, (1, Marseille, ALL) $\preceq_g$ (1, Marseille, Eté) illustre l'ordre de généralisation entre tuples. Enfin, nous avons (1, Marseille, Été) + (1, Marseille, Printemps) = (1, Marseille, ALL).

Dans la suite de l'article, nous considérons seulement les fonctions agrégatives COUNT and SUM. De plus pour conserver la propriété d'anti-monotonie de SUM, nous supposons que les valeurs de la mesure sont strictement positives. Etant donné une fonction agrégative $f$ , $r$ une relation et $t$ un tuple de $CL(r)$. Nous notons $f_{val}(t, r)$ la valeur de la fonction agrégative $f$ associée au tuple $t$ dans $CL(r)$.

Un tuple émergent de $r_1$ vers $r_2$ (avec $r_1$ et $r_2$ deux relations uni-compatibles) a une valeur de mesure faible dans $r_1$ alors qu'elle est significative dans $r_2$.

**Définition 2.1 (Tuple Émergent) -** un tuple $t \in CL(r_1 \cup r_2)$ est dit émergent de $r_1$ vers $r_2$ si et seulement si il satisfait les deux contraintes $C_1$ et $C_2$ :

$$\begin{cases} f_{val}(t, r_1) < MinThreshold_1 \ (C_1) \\ f_{val}(t, r_2) \geq MinThreshold_2 \ (C_2) \end{cases}$$

où $MinThreshold_1$ et $MinThreshold_2$ sont des seuils minimaux positifs définis par l'utilisateur.

**Définition 2.2 (Taux d'Émergence) -** Soit $r_1$ et $r_2$ deux relations, $t \in CL(r_1 \cup r_2)$ un tuple et $f$ une des fonctions agrégatives citées. Le taux d'émergence de $t$ de $r_1$ vers $r_2$, noté $ER(t)$, est défini par :

$$ER(t) = \begin{cases} 0 \text{ si } f_{val}(t, r_1) = 0 \text{ et } f_{val}(t, r_2) = 0 \\ \infty \text{ si } f_{val}(t, r_1) = 0 \text{ et } f_{val}(t, r_2) \neq 0 \\ \dfrac{f_{val}(t, r_2)}{f_{val}(t, r_1)} \text{ sinon.} \end{cases}$$

| | |
|---|---|
| $U$ | (2, Marseille, Printemps) |
| | (3, Paris, Automne) |
| $L$ | (ALL, ALL, Printemps) |
| | (2, ALL, ALL) |
| | (3, ALL, ALL) |
| | ( ALL, ALL, Automne) |

FIG. 1: Bordures $[L; U]$ du Cube Émergent

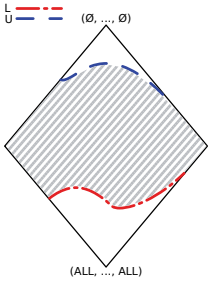


FIG. 2: Illustration des bordures

Nous observons que lorsque le taux d'émergence est supérieur à 1, il caractérise des tendances significatives dans $r_2$ mais insuffisamment marquées dans $r_1$. Au contraire, quand le taux d'émergence est inférieur à 1, il souligne les tendances immergentes, pertinentes dans $r_1$ mais pas dans $r_2$.

**Définition 2.3 (Cube Émergent) -** Nous appelons Cube Émergent l'ensemble des tuples de $CL(r_1 \cup r_2)$ émergents de $r_1$ vers $r_2$. Le Cube Émergent, noté $EC(r_1,\ r_2)$, est défini par :

$$EC(r_1,\ r_2) = \{t \in CL(r_1 \cup r_2) \mid C_1(t) \wedge C_2(t)\}$$

avec $C_1(t) = f_{val}(t, r_1) < MinThreshold_1$ et $C_2(t) = f_{val}(t, r_2) \geq MinThreshold_2$.

Le Cube Émergent est évidemment un cube convexe car il s'inscrit dans le contexte du treillis cube (Casali et al., 2003) et qu'il s'appuie sur une conjonction de contraintes monotone et antimonotone. Il peut donc être représenté par les bordures classiques $[L; U]$ (Nedjar et al., 2009).

**Définition 2.4 (Bordures $[L; U]$) -** Le Cube Émergent peut être représenté par les bordures : $U$ qui englobe les tuples émergents maximaux et $L$ qui contient tous les tuples émergents minimaux selon l'ordre de généralisation. $\preceq_g$.

$$\begin{cases} L = \min_{\preceq_g}(\{t \in CL(r_1 \cup r_2) \mid C_1(t) \wedge C_2(t)\}) \\ U = \max_{\preceq_g}(\{t \in CL(r_1 \cup r_2) \mid C_1(t) \wedge C_2(t)\}) \end{cases}$$

La figure 2 offre une illustration des bordures $[L; U]$ en les situant dans le treillis cube de $r_1 \cup r_2$. Bien sûr, la partie hachurée correspond au Cube Émergent.

**Exemple 2.2 -** Avec nos relations exemples $\text{VENTE}_{07}$ et $\text{VENTE}_{08}$, la figure 1 donne les bordures $[L; U]$ pour le Cube Émergent, en conservant une valeur similaire pour les seuils ($MinThreshold_1 = 200,\ MinThreshold_2 = 200$).

# 3 Structure du Cube Fermé Émergent

Dans ce paragraphe, nous introduisons la structure du Cube Fermé Émergent qui inclut à la fois *(i)* l'ensembles des tuples fermés émergents et *(ii)* la bordure $L$. Cette approche est

conçue dans le même esprit que celle proposée par (Bonchi et Lucchese, 2004) dans le contexte de bases de données de transactions et qui englobe les motifs fermés contraints et la *Lower* bordure ($L$).

Pour une raison de simplicité de notation, nous utilisons, à partir de maintenant $t$ au lieu de $(t, f_{val}(t, r))$ pour indiquer un tuple complet avec la valeur de sa mesure.

**Définition 3.1 (Fermeture Cubique) -** Soit $T \subseteq CL(r)$ un ensemble de tuples, l'opérateur de Fermeture Cubique $\mathbb{C} : CL(r) \rightarrow CL(r)$ selon $T$ peut être défini comme suit :

$$\mathbb{C}(t, T) = (\emptyset, \ldots, \emptyset) + \sum_{\substack{t' \in T, \\ t \preceq_g t'}} t'$$

où l'opérateur $\sum$ a la même sémantique que l'opérateur $+$ mais opère sur un ensemble et non un couple de tuples.

Considérons tous les tuples $t'$ dans $T$. Agrégeons les ensemble en utilisant l'opérateur $+$. Nous obtenons un nouveau tuple qui généralise tous les tuples $t'$ et qui est le plus spécifique. Ce nouveau tuple est la Fermeture Cubique de $t$.

**Exemple 3.1 -** Nous effectuons la Fermeture Cubique du tuple (1, ALL, ALL) dans la relation VENTE$_{07}$ en agrégeant tous les tuples qui le spécialisent grà¢ce à l'opérateur $+$. $\mathbb{C}$((1, ALL, ALL),VENTE$_{07}$)= $(\emptyset, \ldots, \emptyset)$ + (1, Marseille, Printemps) + (1, Marseille, Été) = (1, Marseille, ALL).

**Définition 3.2 (Tuple Fermé Émergent) -** Soit $t \in CL(r)$ un tuple, $t$ est un tuple fermé émergent si et seulement si : $t$ est un tuple émergent et $\mathbb{C}(t, r_1 \cup r_2) = t$.

**Définition 3.3 (Tuple Clé Émergent) -** Soit $t \in CL(r)$, $t$ est un tuple clé émergent si et seulement si : $t$ est un tuple émergent et $t \in \min(\{t' \in CL(r_1 \cup r_2)$ tel que $\mathbb{C}(t', r_1 \cup r_2) = \mathbb{C}(t)\})$.

**Exemple 3.2 -** Le tuple (2, Marseille, Printemps) est un tuple fermé émergent car : (2, Marseille, Printemps) est un tuple émergent . et $\mathbb{C}$((2, Marseille, Printemps), VENTE$_{07}$∪ VENTE$_{08}$) = (2, Marseille, Printemps). De plus, le tuple (2, ALL, ALL) est un tuple clé émergent car il appartient à la bordure $L$ (*cf.* Figure 1) et tous les tuples qui le généralisent ont une fermeture cubique différente.

L'ensemble des tuples fermés émergents n'est pas une représentation sans perte du Cube Émergent car pour certains tuples il est impossible de décider s'ils sont émergents ou pas. Ce sont tous les tuples plus généraux que les plus généraux tuples fermés émergents.

Afin d'obtenir une représentation sans perte, nous combinons d'une part l'ensemble des tuples fermés émergents à partir desquels les valeurs de la mesure peuvent être retrouvées et d'autre part les bordures qui délimitent l'espace des solutions. Cependant, la bordure $U$ est déjà incluse dans l'ensemble des tuples fermés car les éléments de $U$ sont les tuples émergents les plus détaillés (spécifiques). Ils sont donc obligatoirement des tuples fermés.

**Définition 3.4 (Cube Fermé Émergent) -** $ECC(r_1, r_2) = \{t \in CL(r_1 \cup r_2)$ tel que $t$ est un tuple fermé émergent$\} \cup L$.

**Exemple 3.3 -** Le Cube Fermé Émergent est représenté par la table 3 donnant l'ensemble des tuples fermés émergents et la figure 1 qui propose la bordure $L$.

| Tuple fermé émergent | ER |
|---|---|
| (ALL, Marseille, Printemps) | 3 |
| (2, ALL, ALL) | 3 |
| (3, Paris, ALL) | 3 |
| (2, Marseille, Printemps) | $\infty$ |
| (3, Paris, Automne) | $\infty$ |

TAB. 3: Ensemble des tuples fermés émergents

**Proposition 3.1 -** Pour tout tuple fermé émergent $t$, $\mathbb{C}(t, \mathit{ECC}(r_1, r_2)) = \mathbb{C}(t, r_1 \cup r_2)$

**Proposition 3.2 -** Soit $t$ et $u$ deux tuples de $EC(r_1, r_2)$, si $t$ et $u$ ont la même fermeture cubique sur $r_1 \cup r_2$, alors leur taux d'émergence est le même. $\forall t, u \in EC(r_1, r_2)$ tels que $\mathbb{C}(t, r_1 \cup r_2) = \mathbb{C}(u, r_1 \cup r_2)$, nous avons $ER(t) = ER(u)$.

Afin de faire du Cube Fermé Émergent une représentation sans perte, nous devons calculer la fermeture cubique sur $r_1 \cup r_2$ car deux tuples peuvent avoir les mêmes fermetures cubiques sur $r_1$ et sur $r_2$, mais des fermetures cubiques différentes sur $r_1 \cup r_2$. La proposition suivante assure que le Cube Fermé Émergent est une représentation sans perte du Cube Émergent.

**Proposition 3.3 -** Le Cube Fermé Émergent est une représentation sans perte du Cube Émergent : $\forall\, t \in CL(r_1 \cup r_2)$, $t$ est un tuple émergent si et seulement si $\mathbb{C}(t, \mathit{ECC}(r_1, r_2))$ est un tuple fermé émergent.

Par exemple, dérivons le taux d'émergence du tuple (ALL, Paris, Automne). Nous savons que ce tuple est émergent parce qu'il appartient à l'intervalle $[L; U]$ (*cf.* Figure 1). En calculant sa fermeture cubique sur $\mathit{ECC}(\text{VENTE}_{07}, \text{VENTE}_{08})$, nous obtenons le tuple (3, Paris, Automne). Puisque le taux d'émergence du tuple précédent est $\infty$, nous sommes sûrs que le taux d'émergence de (ALL, Paris, Automne) est $\infty$ et donc nous retrouvons le bon résultat.

# 4 Structure du Cube Quotient Émergent

## 4.1 Cubes Quotients et leur sémantique basée sur la fermeture cubique

Un Cube Quotient (Lakshmanan et al., 2002) offre un résumé d'un cube de données pour certaines fonctions agrégatives comme COUNT, SUM, ... De plus le Cube Quotient préserve la sémantique des opérateurs ROLL-UP/DRILL-DOWN sur le cube de données (Gray et al., 1997). Nous revisitons les définitions originales du Cube Quotient dans l'environnement du Treillis Cube. L'idée sous-tendant la représentation en question est d'éliminer les redondances en rassemblant ensemble les tuples véhiculant une information équivalente. Ceci résulte dans un ensemble de classes d'équivalence partitionnant les tuples du cube de données. Un tel partitionnement peut être effectué de plusieurs manières. Mais, afin de préserver les capacités de navigation, il est nécessaire de gérer des classes convexes.

**Définition 4.1 (Classes d'équivalence convexes) -** Soit $\mathcal{C} \subseteq CL(r)$ une classe d'équivalence. Nous disons que $\mathcal{C}$ est convexe si et seulement si :

$$\forall t \in CL(r) \text{ si } \exists t', t'' \in \mathcal{C} \text{ tels que } t' \preceq_g t \preceq_g t'' \text{ alors } t \in \mathcal{C}.$$

Une partition $\mathcal{P}$ de $CL(r)$ qui comprend uniquement des classes d'équivalence convexes est appelée partition convexe.

La propriété de convexité rend possible la représentation de chaque classe d'équivalence à travers ses tuples minimaux et et son tuple maximal.

**Définition 4.2 (Relation d'équivalence quotient) -** Soit $f_{val}$ une fonction mesure. Nous définissons la relation d'équivalence $\equiv_f$ comme la fermeture réflexive transitive de la relation suivante $\tau$ : soit $t, t'$ deux tuples, $t \ \tau \ t'$ est vrai si et seulement si $(i)$ $f_{val}(t, r) = f_{val}(t', r)$ et $(ii)$ $t$ est soit un parent soit un enfant de $t'$.
La relation d'équivalence $\equiv_f$ est dite *relation d'équivalence quotient* si et seulement si elle satisfait la propriété de congruence faible : $\forall t, t', u, u' \in CL(r)$, si $t \equiv_f t', u \equiv_f u', t \preceq_g u$ et $u' \preceq_g t'$, alors $t \equiv_f u$.

Nous notons $[t]_{\equiv_f}$ la classe d'équivalence de $t$ ($[t]_{\equiv_f} = \{t' \in CL(r)$ tel que $t \equiv_f t'\}$). Alors le Cube Quotient est défini comme l'ensemble des classes d'équivalence, chacune d'entre elles étant pourvue de la valeur de la mesure.

**Définition 4.3 (Cube Quotient) -** Soit $CL(r)$ le treillis cube de la relation $r$ d'une base de données et $\equiv_f$ une relation d'équivalence quotient. Le Quotient Cube de $r$, noté $Quotient\text{-}Cube(r, \equiv_f)$, est défini comme suit :

$$QuotientCube(r, \equiv_f) = \{([t]_{\equiv_f}, f_{val}(t, r)) | t \in CL(r)\}.$$

Le Cube Quotient de $r$ est une partition convexe de $CL(r)$.

Pour deux classes d'équivalence $\mathcal{C}, \mathcal{C}' \in QuotientCube(r, \equiv_f), \mathcal{C} \preceq_{QC} \mathcal{C}'$ quand $\exists t \in \mathcal{C}$ et $\exists t' \in \mathcal{C}'$ tels que $t \preceq_g t'$.

La construction d'un Cube Quotient dépend de la relation d'équivalence quotient choisie. Par conséquent pour deux relations d'équivalence quotients, leur Cubes Quotients associés peuvent différer. De plus, la relation d'équivalence quotient la plus utile est la relation d'équivalence de couverture. La couverture de tout tuple $t$ est l'ensemble de tous les tuples agrégés ensemble pour obtenir $t$.

**Définition 4.4 (Couverture) -** Soit $t \in CL(r)$, La couverture de $t$ est un ensemble de tuples de $r$ qui sont généralisés par $t$ (*i.e.* $cov(t, r) = \{t' \in CL(r)$ tel que $t \preceq_g t'\}$).

Deux tuples $t, t' \in CL(r)$ sont dits équivalents selon leur couverture sur $r$, $t \equiv_{cov} t'$, s'ils ont la même couverture, *i.e.* $cov(t, r) = cov(t', r)$.

En utilisant la relation d'équivalence de couverture comme une instance de $\equiv_f$ dans la définition 4.3, nous pouvons définir le *cube quotient de couverture*.

Nous montrons maintenant que le Cube Quotient de Couverture est fortement relié à la fermeture cubique. Deux tuples $t, t' \in CL(r)$ sont équivalents selon la fermeture cubique, $t \equiv_{\mathbb{C}} t'$, si et seulement si $\mathbb{C}(t, r) = \mathbb{C}(t', r)$.

**Proposition 4.1 -** Soit $t, t' \in CL(r)$, $t$ équivalent selon la couverture à $t'$ sur $r$ si et seulement si $t$ est équivalent à $t'$ selon la fermeture cubique.

La proposition ci-dessus établit le lien entre le Cube Quotient et les concepts associés à la fermeture cubique présentés dans la section 3. De plus, elle montre qu'il est possible de définir un Cube Quotient de Couverture en utilisant toute fonction agrégative compatible avec la fermeture cubique.

## 4.2 Cubes Quotients Émergents

Dans le paragraphe précédent, nous avons revisité la structure du Cube Quotient qui a été originellement proposé comme une représentation concise des cubes de données (Lakshmanan et al., 2002) préservant les opérateurs de navigation (ROLL-UP / DRILL-DOWN). Motivés par les propriétés pertinentes du Cube Quotient, nous voulons tirer profit d'une telle représentation pour condenser les Cubes Émergents. Comme le taux d'émergence n'est pas une fonction monotone, l'adaptation nécessaire est difficile à exprimer en utilisant les concepts originaux. C'est pourquoi nous établissons le lien entre le Cube Quotient et les concepts associés à la fermeture cubique. Il est possible que deux tuples liés par l'ordre de généralisation aient tous deux un taux d'émergence infini. Néanmoins ces deux tuples peuvent avoir une fermeture différente. Ainsi pour définir le Cube Quotient Émergent, il n'est pas possible d'utiliser le taux d'émergence comme fonction mesure. À la place, nous utilisons le couple $(f_{val}(t, r_1), f_{val}(t, r_2))$ car il est composé de deux fonctions qui elles mêmes sont compatibles avec la fermeture cubique.

**Définition 4.5 (Cube Quotient Émergent) -** Nous appelons Cube Quotient Émergent l'ensemble des classes d'équivalence de $CL(r_1 \cup r_2)$ émergeant de $r_1$ vers $r_2$ noté $EQC(r_1, r_2)$ : $EQC(r_1, r_2) = \{([t]_{\equiv_f}, f_{val}(t, r_1), f_{val}(t, r_2))$ tel que $[t]_{\equiv_f} \in QuotientCube(r_1 \cup r_2, \equiv_f)$ et $t$ est émergent de $r_1$ vers $r_2\}$.

Chaque classe d'équivalence du Cube Quotient Émergent est représentée par son élément maximal (selon la généralisation) qui est un tuple fermé émergent et ses éléments minimaux qui sont les tuples clés associés aux tuples fermés cités.

**Proposition 4.2 -** Les bordures classiques sont incluses dans le Cube Quotient Émergent. La caractérisation de ces bordures basée sur le Cube Quotient Émergents est la suivante :

1. $U = \max_{\preceq_g}(\{\max_{\preceq_{QC}}(\{[t]_{\equiv_f}\})\})|([t]_{\equiv_f}, f_{val}(t, r_1), f_{val}(t, r_2)) \in EQC(r_1, r_2)$.
2. $L = \min_{\preceq_g}(\{\min_{\preceq_{QC}}(\{[t]_{\equiv_f}\})\})|([t]_{\equiv_f}, f_{val}(t, r_1), f_{val}(t, r_2)) \in EQC(r_1, r_2)$.

La proposition suivante prouve que la représentation ci-dessus est correcte.

**Proposition 4.3 -** Le Cube Quotient Émergent est un résumé du Cube Émergent : $\forall\ t \in CL(r_1 \cup r_2)$, $t$ est émergent si et seulement si $([t]_{\equiv_f}, f_{val}(t, r_1), f_{val}(t, r_2))$ appartient au Cube Quotient Émergent.

**Exemple 4.1 -** Avec les deux relations exemples $\text{VENTES}_{07}$ et $\text{VENTES}_{08}$, la table 4 donne le Cube Quotient Émergent.

| Tuple Maximal | Tuples Minimaux | |
|---|---|---|
| (ALL, Marseille, Printemps) | (ALL, ALL, Printemps) | (3,1) |
| (2, ALL, ALL) | (2, ALL, ALL) | (3,1) |
| (3, Paris, ALL) | (3, ALL, ALL) | (3,1) |
| (2, Marseille, Printemps) | (2, ALL, Printemps)<br>(2, Marseille, Printemps) | (2,0) |
| (3, Paris, Automne) | (ALL, ALL, Automne) | (2,0) |

TAB. 4: Cube Quotient Émergent

### 4.3 Liens entre et utilisations des différentes structures

Nous ajoutons à la représentation réduite classique du Cube Émergent, les bordures $L$ et $U$, deux nouvelles structures sans perte d'information : le Cube Fermé Émergent et le Cube Quotient Émergent. Chaque structure a des usages particuliers pour la fouille de bases de données OLAP. Disposant simplement des bordures, l'utilisateur peut savoir si un tuple est émergent ou pas. Comme les motifs émergents qui offrent des classifieurs précis dans les bases de données de transactions (Dong et Li, 2005), les bordures citées peuvent être utilisées pour des taches de classification dans les bases de données OLAP.

Néanmoins, à partir des bordures, il est impossible d'obtenir le taux d'émergence des tuples. Afin d'éviter cet inconvénient, nous proposons le Cube Fermé Émergent pour répondre à toutes les requêtes OLAP qu'il est possible d'exprimer sur le Cube Émergent (sans avoir besoin de le calculer). Enfin, afin d'offrir aux utilisateurs des outils pour naviguer au sein des Cubes Émergents, nous caractérisons le Cube Quotient Émergent. À travers le théorème suivant, nous établissons les liens d'inclusion entre les différentes représentations.

**Theoréme 4.1** - Soit $[L;U]$, $ECC$ and $EQC$ les différentes représentations pour le Cube Émergent ($EC$) de deux relations $r_1$ et $r_2$. Nous avons alors :

$$[L;U] \subseteq ECC \subseteq EQC \subseteq EC$$

Toutes les représentations proposées sont réduites comparées au Cube Émergent lui-même excepté dans les deux cas extrêmes : *(i)* Quand il n'existe aucun tuple émergent et *(ii)* quand tous les tuples émergents sont fermés et donc que le Cube Émergent ne contient aucune redondance.

## 5 Évaluations experimentales

Pour réaliser les évaluations présentées, nous utilisons les mêmes relations de bases de données que celles exploitées dans (Xin et al., 2007). Les expérimentations sont menées sur des données provenant d'un éventail large et varié de domaines. Il est bien connu que les données synthétiques sont faiblement corrélées alors que dans de nombreuses bases réelles ou statistiques les données sont fortement corrélées (Pasquier et al., 1999). Pour les données syn-

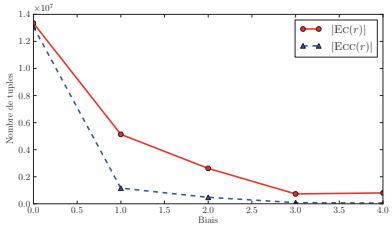


FIG. 3: Taille du Cube Fermé Émergent avec $\mathcal{D}$ = 10, $\mathcal{C}$ = 100, $\mathcal{T}$ = 1000K

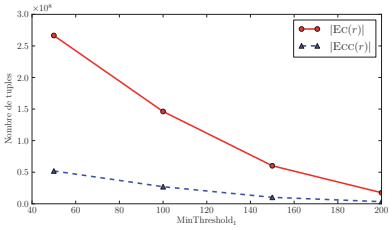


FIG. 4: Taille du Cube Fermé Émergent pour les relations de données météorologiques

thétiques [1], nous utilisons les notations suivantes pour décrire les relations : $\mathcal{D}$ le nombre de dimensions, $\mathcal{C}$ la cardinalité de chaque dimension, $\mathcal{T}$ le nombre de tuples de la relation, $\mathcal{M}_1$ (respectivement $\mathcal{M}_2$) le seuil correspondant à la contrainte d'émergence $C_1$ (respectivement $C_2$), et $\mathcal{S}$ le biais ou zipf des données. Quand $\mathcal{S}$ est égal à 0, les données sont uniformes. Quand $\mathcal{S}$ croît, les données sont de plus en plus biaisées. $\mathcal{S}$ est appliqué sur toutes les relations dans chacune des bases de données. Pour les données réelles, nous utilisons les relations de données météorologiques SEP83L.DAT et SEP85L.DAT utilisées par (Xin et al., 2007), qui ont 1.002.752 tuples avec 8 dimensions sélectionnées. Les attributs dimensions (avec leur cardinalité) sont les suivantes : année mois jour heure (238), latitude (5260), longitude (6187), numéro de station (6515), météo actuelle (100), code de changement (110), altitude solaire (1535) et luminance lunaire relative (155).

Dans les figures 3 et 4, nous reportons les résultats obtenus en comparant les tailles du Cube Fermé Émergent et du Cube Émergent pour nos deux jeux de données. Comme attendu, en faisant varier le biais des données, nous observons que, comparé au Cube Émergent, le Cube Fermé Émergent a une taille de plus en plus réduite au fur et à mesure que le biais des données augmente. Le facteur de réduction varie de 7 à 11. Le phénomène est illustré par la figure 3. De plus, nous utilisons des données réelles, connues pour être fortement corrélées, pour comparer le Cube Émergent et le Cube Fermé Émergent. Le paramètre qui varie est le seuil minimal appliqué à la relation $r_2$. Nous observons dans la figure 4 que plus ce seuil s'accroit plus la taille du Cube Émergent et du Cube Fermé Émergent décroit. Effectivement quand le seuil minimal est élevé, il est logique que le nombre de tuples émergents soit moindre. Cependant,le Cube Fermé Émergent est toujours plus réduit que le Cube Émergent avec un gain appréciable. Lors de la comparaison des tailles du Cube Quotient Émergent et du Cube Émergent, nous considérons exactement les mêmes cas d'expérimentation que pour le Cube Fermé Émergent et, bien sûr, obtenons des résultats similaires. Pour les données synthétiques, en augmentant le facteur biais des données, la figure 5 montre que le Cube Quotient Émergent permet une réduction effective. Pour les données réelles, les résultats sont donnés dans la figure 6.

1. Le générateur de données synthétiques est disponible à l'adresse : `http://illimine.cs.uiuc.edu/`

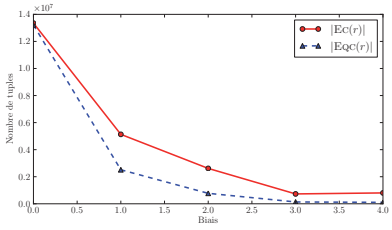


FIG. 5: Taille du Cube Quotient Émergent avec $\mathcal{D} = 10$, $\mathcal{C} = 100$, $\mathcal{T} = 1000$K

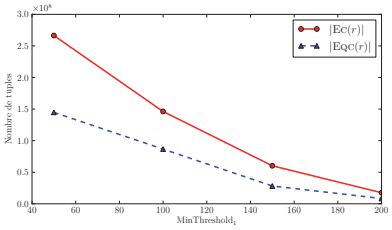


FIG. 6: Taille du Cube Quotient Émergent pour les relations de données météorologiques

# 6 Conclusion

Le concept de Cube Émergent a été proposé afin de capturer les renversements de tendances entre les data cubes de deux relations comparables. Dans cet article, nous introduisons deux nouvelles représentations sans perte des mesures pour les Cubes Émergents qui sont basées sur le concept de fermeture cubique. Pour valider nos structures, nous effectuons des expérimentations. Nous avons évalué la taille des représentations introduites avec des résultats très intéressants pour les données fortement corrélées. Les deux représentations proposées sont originales. Elles peuvent être directement appliquées pour manipuler une représentation réduite des motifs émergents ou plus généralement de motifs contraints Bonchi et Lucchese (2004).

Une perspective intéressante est le calcul d'une estimation de la taille du Cube Émergent afin de l'utiliser comme un guide non seulement pour calibrer les seuils d'émergence mais aussi pour choisir l'algorithme de calcul de cube le mieux adapté. En effet, l'efficacité de ces algorithmes est fortement liée à la nature même des données (fortement ou faiblement corrélées) et la taille des futurs résultats est un paramètre crucial. D'autres perspectives seraient d'une part de prendre en compte les hiérarchies de dimensions Hurtado et Mendelzon (2002) dans le même esprit que Cure for Cubes Morfonios et Ioannidis (2006) et d'autre part de raffiner l'estimation de la taille des représentations.

# Références

Beyer, K. S. et R. Ramakrishnan (1999). Bottom-up computation of sparse and iceberg cubes. In A. Delis, C. Faloutsos, et S. Ghandeharizadeh (Eds.), *SIGMOD Conference*, pp. 359–370. ACM Press.

Bonchi, F. et C. Lucchese (2004). On closed constrained frequent pattern mining. In K. Morik et R. Rastogi (Eds.), *ICDM*, pp. 35–42. IEEE Computer Society.

Börzsönyi, S., D. Kossmann, et K. Stocker (2001). The skyline operator. In *ICDE*, pp. 421–430. IEEE Computer Society.

Casali, A., R. Cicchetti, et L. Lakhal (2003). Extracting semantics from data cubes using cube transversals and closures. In L. Getoor, T. E. Senator, P. Domingos, et C. Faloutsos (Eds.), *KDD*, pp. 69–78. ACM.

Dong, G. et J. Li (2005). Mining border descriptions of emerging patterns from dataset pairs. *Knowl. Inf. Syst. 8*(2), 178–202.

Gray, J., S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, et H. Pirahesh (1997). Data cube : A relational aggregation operator generalizing group-by, cross-tab, and sub totals. *Data Min. Knowl. Discov. 1*(1), 29–53.

Hurtado, C. A. et A. O. Mendelzon (2002). Olap dimension constraints. In L. Popa (Ed.), *PODS*, pp. 169–179. ACM.

Lakshmanan, L. V. S., J. Pei, et J. Han (2002). Quotient cube : How to summarize the semantics of a data cube. In F. H. Lochovsky et W. Shan (Eds.), *VLDB*, pp. 778–789. Morgan Kaufmann.

Morfonios, K. et Y. E. Ioannidis (2006). Cure for cubes : Cubing using a rolap engine. In U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, et Y.-K. Kim (Eds.), *VLDB*, pp. 379–390. ACM.

Nedjar, S., A. Casali, R. Cicchetti, et L. Lakhal (2009). Emerging cubes : Borders, size estimations and lossless reductions. *Information Systems 34*(6), 536–550.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems 24*(1), 25–46.

Pei, J., Y. Yuan, X. Lin, W. Jin, M. Ester, Q. Liu, W. Wang, Y. Tao, J. X. Yu, et Q. Zhang (2006). Towards multidimensional subspace skyline analysis. *ACM Trans. Database Syst. 31*(4), 1335–1381.

Xin, D., J. Han, X. Li, Z. Shao, et B. W. Wah (2007). Computing iceberg cubes by top-down and bottom-up integration : The starcubing approach. *IEEE Trans. Knowl. Data Eng. 19*(1), 111–126.

Zaki, M. J. et C.-J. Hsiao (2002). Charm : An efficient algorithm for closed itemset mining. In R. L. Grossman, J. Han, V. Kumar, H. Mannila, et R. Motwani (Eds.), *SDM*. SIAM.

## Summary

The concept of Emerging Cube has been introduced in order to compare two data cubes. In this paper, we introduce two new and reduced representations without measure loss: the Emerging Closed Cube and Emerging Quotient Cube. The former representation is based on the concept of cube closure. It is the smallest possible representation of cubes. Provided with the Emerging Closed Cube and thus by storing the minimal information, it is possible to answer efficiently queries which can be answered from the Emerging Cube itself. The latter representation is supported by the structure of the Quotient Cube which was proposed to summarize data cubes. The Quotient Cube is revisited in order to provide it with a closure-based semantics and thus adapt it to the context of Emerging Cube. The resulting Emerging Quotient Cube is less reduced than the Emerging Closed Cube but it preserves the "specialization / generalization" property of the data cube which makes it possible to navigate within the Emerging Cube. We also state the relationship between the two introduced representations and the one based on the borders. Experiments performed on various data sets are intended to measure the size of the three representations.