Treillis des concepts SKYLINES : Analyse multidimensionnelle des SKYLINES fondée sur les ensembles en accord

Sébastien Nedjar, Fabien Pesci[1], Lotfi Lakhal & Rosine Cicchetti

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

Résumé. Le concept de SKYLINE a été introduit pour mettre en évidence les objets « les meilleurs » selon différents critères. Une généralisation multidimensionnelle du SKYLINE a été proposée à travers le SKYCUBE qui réunit tous les SKYLINES possibles selon toutes les combinaisons de critères et permet d'analyser les liens entre objets SKYLINES. Comme le data cube, le SKYCUBE s'avère extrêmement volumineux si bien que des approches de réduction sont incontournables. Dans cet article, nous définissons une approche de matérialisation partielle du SKYCUBE. L'idée sous-jacente est d'éliminer de la représentation les Skycuboïdes facilement re-calculables. Pour atteindre cet objectif de réduction, nous caractérisons un cadre formel : le treillis des concepts ACCORDS. Cette structure combine la notion d'ensemble en accord et le treillis des concepts. À partir de cette structure, nous dérivons le treillis des concepts SKYLINES qui en est une instance contrainte. Le point fort de notre approche est d'être orientée attribut ce qui permet de borner le nombre de nœuds du treillis et d'obtenir une navigation efficace à travers les Skycuboïdes.

1 Introduction

Dans un contexte décisionnel, certaines requêtes ne renvoient aucun résultat. Dans ces requêtes, l'utilisateur recherche les tuples pour lesquels les valeurs de certains critères sont optimales. C'est le caractère « multicritère » de ces interrogations qui les rend généralement infructueuses. En effet, tel tuple peut être optimal pour un critère mais pas pour un autre, il est alors éliminé du résultat alors qu'il aurait pu être pertinent pour l'utilisateur. Par exemple, si l'on considère une base de données immobilières, la recherche du logement « idéal » peut combiner des conditions sur le prix, le plus bas possible, la surface, la plus grande possible, et l'éloignement du lieu de travail, le plus réduit possible. Évidemment il est vraisemblable que ce logement idéal n'existe pas, d'où l'absence de réponse à ce type de requête. Pourtant certains logements pourraient s'avérer pertinents pour l'utilisateur parce que, situés dans une zone proche, mais non voisine, ils réunissent les critères de surface maximale et de prix minimal.

1. Fabien Pesci bénéficie d'une bourse doctorale co-financée par le conseil régional PACA et l'entreprise CA2I.

Afin d'apporter une réponse adéquate au type de requêtes décrites, l'opérateur SKYLINE (Börzsönyi et al., 2001) a été introduit. Il considère l'ensemble des critères de choix d'une recherche comme autant de préférences et extrait les tuples globalement optimaux pour cet ensemble de préférences. Ainsi plutôt que de rechercher une hypothétique solution idéale, il extrait les candidats les plus proches possibles des souhaits de l'utilisateur. Son principe général s'appuie sur la notion de dominance. Un objet ou un tuple est dit dominé par un autre si, pour tous les critères intéressant le décideur, il est moins optimal que cet autre. Un tel tuple est éliminé du résultat, non pas parce qu'il est non pertinent pour un des critères mais parce qu'il est non optimal selon la combinaison de tous les critères. En d'autres termes, il existe au moins une meilleure solution pour l'utilisateur qui, elle, sera retenue.

De la même manière que le cube de données permet de comprendre les liens existant entre plusieurs niveaux d'agrégation, une généralisation multidimensionnelle du SKYLINE a été proposée à travers le SKYCUBE (Yuan et al., 2005; Pei et al., 2005). Cette structure réunit tous les SKYLINES possibles suivant les différentes combinaisons de critères. Il est alors possible de rechercher efficacement des objets dominants selon différentes combinaisons de critères. De plus, grâce à cette structure, il devient possible d'observer le comportement des objets dominants à travers l'espace multidimensionnel et ainsi d'analyser et de comprendre les différents facteurs de dominance. Ce concept étant inspiré du cube de données, il souffre des mêmes inconvénients de coût de calcul et d'explosion de l'espace de stockage. Ainsi il est naturel, comme pour le data cube, d'essayer d'en proposer des représentations réduites et les algorithmes associés.

Dans cet article, nous proposons une approche basée sur le treillis des concepts qui permet de matérialiser partiellement les SKYCUBES et donc de réduire la taille de leur représentation tout en garantissant la reconstruction complète des résultats. Cette structure combine le concept d'ensemble en accord (Beeri et al., 1984; Lopes et al., 2002), issu de la théorie des bases de données, et celui du treillis des concepts, fondement de l'analyse de concepts formels. Contrairement à Pei et al. (2006), notre approche de réduction est orientée attribut ce qui lui confère les mêmes qualités pour la navigation que le SKYCUBE complet.

Le plan de l'article est le suivant. Au paragraphe 2, nous rappelons le contexte de notre travail : le SKYCUBE. Nous formalisons un cadre formel dans lequel s'inscrira notre travail au paragraphe 3. Puis nous présentons notre approche de matérialisation partielle des SKYCUBES.

2 SKYCUBE pour l'Analyse multidimensionnelle des SKYLINES

Dans cette section, nous présentons d'abord l'opérateur SKYLINE ainsi que la problématique à laquelle il répond. Dans un deuxième temps, nous présentons l'analyse multidimensionnelle des SKYLINES à travers le concept de SKYCUBE.

2.1 L'opérateur SKYLINE

Avant de définir de manière formelle l'opérateur SKYLINE, il est important de bien situer le contexte dans lequel la problématique se pose. En effet, il ne s'applique pas à n'importe quelle relation. Pour qu'il puisse effectuer les comparaisons nécessaires, il faut que les diffé-

RowId [1]	Propriétaire	...	Ville	Prix	Éloignement	Consommation	Voisins
1	Dupont		Marseille	220	15	275	5
2	Dupond		Paris	100	15	85	1
3	Martin		Marseille	220	7	180	1
4	Sanchez		Aubagne	340	7	85	3
5	Durand		Paris	100	7	180	1

TAB. 1 – *La relation* LOGEMENTS

rents domaines sur lesquels sont définis les attributs, critères de choix de l'utilisateur, soient totalement ordonnés (Börzsönyi et al., 2001; Pei et al., 2006). Par simplicité, les attributs ont systématiquement des valeurs numériques dans nos exemples.

Les tuples de nos relations pouvant être utilisées par l'opérateur SKYLINE sont de la forme $t = (d_1,d_2,\ldots,d_k,c_1,c_2,\ldots,c_l)$ où les d_i sont les dimensions non utilisées par l'opérateur SKYLINE alors que les c_i sont les critères sur lesquels l'utilisateur se fonde pour porter son choix.

Exemple 2.1 - La relation illustrée par la table 1 est typique pour l'utilisation du SKYLINE. Elle répertorie différents logements. Les attributs dimensions classiques sont ici *Propriétaire* et *Ville* et les critères de choix pour trouver le « meilleur logement » sont : le *Prix* de vente en milliers d'euros, l'*Éloignement* par rapport au lieu de travail en kilomètres, la *Consommation* Énergétique en kilowattheures par an et par mètre carré, le nombre de *Voisins*.

Définition 2.1 (Relation de dominance) - Soit $\mathcal{C} = \{c_1, c_2, \ldots, c_d\}$ l'ensemble des critères sur lesquels porte l'opérateur SKYLINE. Soit deux tuples t et t', la relation de dominance suivant l'ensemble de critères $\mathcal{C}$ est définie comme suit :

$$t \succeq_{\mathcal{C}} t' \;\Leftrightarrow\; t[c_1] \leq t'[c_1] \; et \; t[c_2] \leq t'[c_2] \; et \; \ldots \; et \; t[c_d] \leq t'[c_d]$$

Lorsque $t \succeq_{\mathcal{C}} t'$ et $\exists\, c_i \in \mathcal{C}$ tel que $t[c_i] < t'[c_i]$, la dominance est stricte, elle est notée $t \succ_{\mathcal{C}} t'$.

Lorsqu'un tuple t domine un tuple t' (*i.e.* $t \succeq_{\mathcal{C}} t'$), cela signifie que t est équivalent ou « meilleur » que le tuple t' pour tous les critères choisis. Comme nous considérons que les critères sont minimisés, les valeurs de t pour tous les critères sont inférieures ou égales à celles de t'. Ainsi dans le cadre d'une recherche multicritère les tuples dominés par d'autres (au moins un) ne sont pas pertinents et sont éliminés du résultat par l'opérateur SKYLINE.

Définition 2.2 (L'opérateur SKYLINE) - Soit r une relation, le SKYLINE de r suivant $\mathcal{C}$ est l'ensemble des tuples qui ne sont dominés par aucun autre, suivant l'ensemble de critères $\mathcal{C}$:

$$SKY_{\mathcal{C}}(r) = \{t \in r \mid \nexists\, t' \in r \setminus t,\; t' \succ_{\mathcal{C}} t\}$$

Exemple 2.2 - Avec notre relation exemple et les critères suivants $\mathcal{C} = \{$*Éloignement, Prix*$\}$ [2], $SKY_{\mathcal{C}}(\text{LOGEMENT}) = \{t_5\}$ car le tuple t_5 domine tous les autres. Il est donc le meilleur logement possible pour l'utilisateur.

1. *RowId* est un attribut implicite servant d'identifiant unique à chaque tuple. t_i est le tuple ayant i pour *RowId*.
2. Pour simplifier les notations, l'ensemble $\{A,B\}$ est noté AB.

2.2 SKYCUBE

L'opérateur SKYLINE est un outil fondamental pour l'analyse multicritère des bases de données. Nous pouvons calculer un SKYLINE suivant un ensemble de critères définis par l'utilisateur, mais lorsqu'il s'agit d'en calculer plusieurs sur les mêmes données, aucun d'eux ne sait exploiter à son avantage les liens qui peuvent exister entre les différents SKYLINES. C'est pour cela qu'une structure nommée SKYCUBE a été introduite (Yuan et al., 2005; Pei et al., 2005). On peut dire que le SKYCUBE est au SKYLINE ce que le cube de données est au GROUP-BY : une généralisation multidimensionnelle. Ainsi, dès lors que l'on souhaite répondre rapidement à toutes les requêtes posées sur un SKYCUBE, il vaut mieux opter pour une pré-matérialisation de ce cube.

Définition 2.3 (SKYCUBE) - Un SKYCUBE est l'ensemble de tous les SKYLINES dans tous les sous-espaces non vides possibles de $\mathcal{C}$:

$$SKYCUBE(r, \mathcal{C}) = \{(C, SKY_C(r)) \mid C \subseteq \mathcal{C}\}$$

$SKY_C(r)$ est appelé le cuboïde SKYLINE (ou Skycuboïde) du sous-espace C. Par convention le cuboïde selon l'ensemble de critère vide est vide (*i.e.* $SKY_\emptyset(r) = \emptyset$).

La structure du SKYCUBE peut être représentée par un treillis semblable à celui utilisé pour le cube de données. Les cuboïdes du SKYCUBE sont regroupés par niveau en fonction de leur nombre de critères.

Une requête SKYLINE multidimensionnelle retourne le sous-ensemble de tuples de la relation originelle formant le SKYLINE dans un sous-espace donné. Clairement, une fois le SKYCUBE calculé, il est possible de répondre à toute requête efficacement.

2.2.1 Problèmes associés à l'analyse multidimensionnelle des SKYLINES

En général, si un tuple t est dans les SKYLINES des sous-espaces C_1 et C_2 tels que $C_1 \subset C_2$, pouvons-nous affirmer que t appartiendra aussi au SKYLINE de n'importe quel sous-espace C situé entre C_1 et C_2 ($C_1 \subset C \subset C_2$) ? Une telle propriété serait fort attrayante puisqu'elle pourrait considérablement simplifier la détermination des SKYLINES multidimensionnels. Malheureusement elle n'est pas vérifiée dans le cas général.

Exemple 2.3 - Avec la relation LOGEMENT, en considérant les SKYLINES selon (*Prix, Éloignement, Voisin*) et (*Éloignement, Voisin*) on a : (*Prix, Éloignement, Voisin*) $\supseteq$ (*Éloignement, Voisin*) et $SKY_{PEV}(\text{LOGEMENT}) \subseteq SKY_{EV}(\text{LOGEMENT})$.

En revanche, si l'on regarde les SKYLINES suivants (*Prix, Éloignement, Consommation, Voisins*) et (*Prix, Éloignement, Consommation*) on a (*Prix, Éloignement, Consommation, Voisins*) $\supseteq$ (*Prix, Éloignement, Consommation*) et $SKY_{PECV}(\text{LOGEMENT}) \supseteq SKY_{PEC}(\text{LOGEMENT})$.

L'observation mise en évidence par cet exemple est la suivante : l'appartenance à un SKYLINE n'est pas monotone, c'est-à-dire qu'un tuple t appartenant à un cuboïde $SKY_{\mathcal{U}}(r)$ n'est pas automatiquement contenu dans les ancêtres de ce cuboïde.

Comme dans le cas du cube de données, le SKYCUBE peut contenir des informations superflues. C'est cette problématique qui a motivé la proposition de représentations réduites du SKYCUBE faite par Pei et al. (2006). Notre contribution s'intéresse à la même problématique

de réduction en combinant nous aussi l'analyse de concepts formels (Ganter et Wille, 1999) et le SKYLINE. Pour éviter le coût important de reconstruction des cuboïdes SKYLINES induit par le regroupement orienté valeur de Pei et al. (2006), notre méthode de réduction choisit une approche de regroupement orientée critère en se basant sur les ensembles en accord.

3 Treillis des Concepts ACCORDS d'une relation

Dans ce paragraphe, notre objectif est de définir un cadre de travail formel combinant le concept d'ensemble en accord et celui de treillis des concepts. Nous proposons une nouvelle structure, le treillis des concepts ACCORDS d'une relation, sur laquelle s'appuie notre matérialisation partielle du SKYCUBE. Après un rappel des notions d'ensemble en accord et des classes d'équivalence associées, nous caractérisons rigoureusement le treillis des concepts ACCORDS.

3.1 Ensembles en Accords

Le concept d'ensemble en accord (ainsi que le système de fermeture associé), introduit par Beeri et al. (1984) pour caractériser la relation d'Armstrong, a été utilisé avec succès pour l'extraction de dépendances fonctionnelles exactes et approximatives (Lopes et al., 2002). Deux tuples sont en accord sur un ensemble d'attributs X s'ils partagent la même valeur pour X.

Définition 3.1 (Ensembles en accord) - Soit t_i, t_j deux tuples de r et $X \subseteq \mathcal{C}$ un ensemble d'attributs (critères dans notre contexte). t_i, t_j sont en accord sur X si et seulement si $t_i[X] = t_j[X]$. L'ensemble des attributs en accord de t_i et t_j est défini comme suit :

$$\text{ACC}(t_i, t_j) = \{C \in \mathcal{C} \mid t_i[C] = t_j[C]\}$$

Cette définition peut être généralisée pour un ensemble de tuples $T \subseteq r$ composé d'au moins deux éléments :

$$\text{ACC}(T) = \{C \in \mathcal{C} \mid t[C] = t'[C], \forall t, t' \in T\}$$

Définition 3.2 (Ensembles en accord d'une relation) - L'ensemble des attributs en accord d'une relation r est défini comme suit :

$$\text{ACCORDS}(r) = \{\text{ACC}(t_i, t_j) \mid t_i, t_j \in r \text{ et } i \neq j\}$$

Cet ensemble peut être redéfini de manière équivalente :

$$\text{ACCORDS}(r) = \{\text{ACC}(T) \mid \forall T \subseteq r \text{ et } |T| \geq 2\}$$

Exemple 3.1 - Avec la relation LOGEMENT, $\text{ACC}(t_2, t_5) = PV$ car ces deux tuples partagent la même valeur sur les attributs *Prix*, *Voisins* et ont des valeurs différentes pour *Éloignement*, *Consommation*. De même, $\text{ACC}(\{t_3, t_4, t_5\}) = E$ car ces trois tuples ont la même valeur uniquement pour le critère *Éloignement*. L'ensemble des ensembles d'attributs en accord de la relation LOGEMENT est le suivant : $\text{ACCORDS}(\text{LOGEMENT}) = \{\emptyset, E, P, V, C, PV, ECV\}$.

Définition 3.3 (Classe d'équivalence d'un tuple) - Soit r une relation et $C \subseteq \mathcal{C}$ un ensemble de critères. La classe d'un tuple t selon C, notée $[t]_C$, est définie comme l'ensemble des identifiants i ($Rowid$) de tous les tuples $t_i \in r$ en accord avec t selon C (*i.e.* l'ensemble des identifiants des tuples t_i partageant avec t les mêmes valeurs pour C). Nous avons donc :

$$[t]_C = \{i \in Tid(r) \mid t_i[C] = t[C]\}$$

Exemple 3.2 - Avec la relation LOGEMENT, $[t_2]_P = \{2, 5\}$ car seuls les tuples t_2 et t_5 partagent la même valeur sur le critère *Prix*.

3.2 Concepts ACCORDS d'une relation

Notre objectif est de définir un treillis des concepts particulier se basant sur les ensembles en accord et les partitions (Spyratos, 1987). Pour atteindre ce but, nous caractérisons une instance de la connexion de Galois entre d'une part le treillis des parties de l'ensemble des critères et d'autre part le treillis des partitions de l'ensemble des identifiants de tuples. Cette connexion nous permet de dériver des opérateurs de fermeture duaux, de définir un concept ACCORD et de caractériser le treillis des concepts ACCORDS.

Définition 3.4 - Soit $Rowid : r \to \mathbb{N}$ une application qui associe à chaque tuple un unique entier naturel et $Tid(r) = \{Rowid(t) \mid t \in r\}$. Soit f, g deux applications entre les ensembles ordonnés $\langle \Pi(Tid(r)), \sqsubseteq \rangle$ [3] et $\langle \mathcal{P}(\mathcal{C}), \subseteq \rangle$ qui sont définies comme suit :

$$\begin{array}{rccl} f: & \langle \Pi(Tid(r)), \sqsubseteq \rangle & \longrightarrow & \langle \mathcal{P}(\mathcal{C}), \subseteq \rangle \\ & \pi & \longmapsto & \bigcap\limits_{[t] \in \pi} \text{ACC}(\{t_i \mid i \in [t]\}) \\ g: & \langle \mathcal{P}(\mathcal{C}), \subseteq \rangle & \longrightarrow & \langle \Pi(Tid(r)), \sqsubseteq \rangle \\ & C & \longmapsto & \{[t]_C \mid t \in r\} \end{array}$$

Pour un ensemble de critères C, l'ensemble des classes d'équivalence selon C forme une partition de $Tid(r)$. C'est la fonction g qui associe à C cette partition des identifiants. Cette dernière est notée π_C et définie comme suit : $\pi_C = g(C)$. L'ensemble de toutes les partitions π_C possibles est noté $\Pi_{\mathcal{P}(\mathcal{C})}$. La fonction f fait l'association inverse de g.

Exemple 3.3 - Avec la relation LOGEMENT, en prenant les ensembles de critères E, EC, ECV et PV, on a $g(E) = \{12,345\}$, $g(EC) = \{1,2,35,4\}$, $g(ECV) = \{1,2,35,4\}$ et $g(PV) = \{1,25,3,4\}$. Avec les partitions $\{1,2,35,4\}$ et $\{1,2,345\}$, on a $f(\{1,2,35,4\}) = ECV$ et $f(\{1,2,345\}) = E$.

Proposition 3.1 - Le couple d'applications $gc = (f, g)$ est une connexion de Galois entre le treillis des parties des critères $\mathcal{C}$ et le treillis des partitions de $Tid(r)$.

Définition 3.5 (Opérateurs de fermeture) - Le couple $gc = (f, g)$ étant un cas particulier de la connexion de Galois, les compositions $f \circ g$ et $g \circ f$ des deux applications sont des opérateurs

3. La définition du treillis des partitions $\Pi(E)$ d'un ensemble E est donnée dans la version étendue de ce papier (http://hal.archives-ouvertes.fr/hal-00528668/fr/).

de fermeture (Ganter et Wille, 1999) définis ci-dessous.

$$\begin{array}{lrcl} h: & \mathcal{P}(\mathcal{C}) & \longrightarrow & \mathcal{P}(\mathcal{C}) \\ & C & \longmapsto & f(g(C)) = \bigcap\limits_{\substack{C' \in \text{ACCORDS}(r) \\ C \subseteq C'}} C' \\ h': & \Pi(Tid(r)) & \longrightarrow & \Pi(Tid(r)) \\ & \pi & \longmapsto & g(f(\pi)) = \bullet\limits_{\substack{\pi' \in \Pi_{\mathcal{P}(\mathcal{C})} \\ \pi \sqsubseteq \pi'}} \pi' \end{array}$$

Exemple 3.4 - Avec la relation LOGEMENT, en prenant les ensembles de critères EC et ECV, d'après l'exemple précédent on a : $h(EC) = f(g(EC)) = f(\{1,2,35,4\}) = ECV$, $h(ECV) = f(g(ECV)) = f(\{1,2,35,4\}) = ECV$.
Avec les partitions $\{1,2,35,4\}$ et $\{1,2,345\}$, on a : $h'(\{1,2,35,4\}) = g(f(\{1,2,35,4\})) = g(ECV) = \{1,2,35,4\}$, $h'(\{1,2,345\}) = g(f(\{1,2,345\})) = g(E) = \{12,345\}$.

Définition 3.6 (Concepts ACCORDS) - Un concept ACCORD d'une relation r est un couple (C, π) associant un ensemble de critères à une partition des identifiants : $C \in \mathcal{P}(\mathcal{C})$ et $\pi \in \Pi(Tid(r))$. Les éléments de ce couple doivent être liés par les conditions suivantes : $C = f(\pi)$, $\pi = g(C) = \pi_C$. Soit $c_a = (C_{c_a}, \pi_{c_a})$ un concept ACCORD de r, nous appelons π_{c_a} *l'extension* de c_a (notée $ext(c_a)$) et C_{c_a} son *intension* (notée $int(c_a)$). L'ensemble de tous les concepts ACCORDS d'une relation r est noté CONCEPTSACCORDS(r).

Théorème 3.1 (Treillis des concepts ACCORDS) - Soit CONCEPTSACCORDS(r) l'ensemble des concepts ACCORDS d'une relation r. L'ensemble ordonné $\langle$CONCEPTSACCORDS$(r), \leq$ [4]$\rangle$ forme un treillis complet nommé treillis des concepts ACCORDS. Pour tout concept $P \subseteq$ CONCEPTSACCORDS(r), l'*infimum* ou borne inférieure ($\bigwedge$) et *supremum* ou borne supérieure ($\bigvee$) sont donnés ci-après :

$$\bigwedge P = (\bigcap_{c_a \in P} int(c_a), h'(\underset{c_a \in P}{+} ext(c_a))), \bigvee P = (h(\bigcup_{c_a \in P} int(c_a)), \bullet_{c_a \in P} ext(c_a))$$

Exemple 3.5 - La figure 1 donne le diagramme de Hasse du treillis des concepts ACCORDS de la relation LOGEMENT. Le couple $(ECV, \{1,2,35,4\})$ est un concept ACCORD car d'après les exemples précédents on a $g(ECV) = \{1,2,35,4\}$ et $f(\{1,2,35,4\}) = ECV$. À l'inverse le couple $(EC, \{1,2,35,4\})$ ne constitue pas un concept ACCORD car $f(\{1,2,35,4\}) \neq EC$. Soit $c_a = (ECV, \{1,2,35,4\})$ et $c_b = (PV, \{1, 25, 3, 4\})$ deux concepts ACCORDS, nous avons donc : $c_a \wedge c_b = (ECV \cap PV,\ h'(\{1, 2, 35, 4\} + \{1, 25, 3, 4\})) = (V,\ h'(\{1, 235, 4\})) = (V, \{1, 235, 4\})$ et $c_a \vee c_b = (h(ECV \cup PV), \{1, 2, 35, 4\} \bullet \{1, 25, 3, 4\}) = (h(PECV), \{1, 2, 3, 4, 5\}) = (PECV, \{1, 2, 3, 4, 5\})$.

Proposition 3.2 - Pour tout ensemble de critères $C \subseteq \mathcal{C}$, la partition associée π_C est identique à la partition de sa fermeture.

$$\forall C \subseteq \mathcal{C}, \pi_C = \pi_{h(C)}$$

Exemple 3.6 - Avec la relation LOGEMENT, en prenant comme ensemble de critères EC, d'après les exemples 3.3 et 3.4, on a : $\pi_{EC} = g(EC) = \{1,2,35,4\}$ et $\pi_{h(EC)} = g(h(EC)) = g(ECV) = \{1,2,35,4\}$.

La proposition précédente signifie que la fermeture d'un ensemble de critères C peut être vue comme le plus grand sur-ensemble de C ayant la même partition.

4. Soit $(C_1, \pi_1), (C_2, \pi_2) \in$ CONCEPTSACCORDS(r), $(C_1, \pi_1) \leq (C_2, \pi_2) \Leftrightarrow C_1 \subseteq C_2$ (ou $\pi_2 \sqsubseteq \pi_1$).

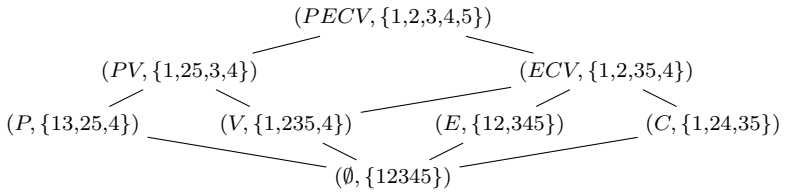

FIG. 1 – *Diagramme de Hasse du treillis des concepts* ACCORDS *de la relation* LOGEMENT

4 Treillis des Concepts SKYLINES pour la matérialisation partielle du SKYCUBE

Le treillis des concepts SKYLINES est un treillis des concepts ACCORDS contraints. Nous démontrons une propriété fondamentale de notre treillis nous permettant de ne matérialiser que partiellement le SKYCUBE en éliminant certains cuboïdes. Nous montrons ensuite comment de tels cuboïdes peuvent être reconstruits facilement.

4.1 Treillis des concepts SKYLINES

Soit π_C une partition de r sur C. Par définition, les tuples d'une même classe d'équivalence $[t]_C \in \pi_C$ sont indistinguables sur C (ils partagent tous la même projection sur C). Si t est non dominé sur C, il en sera de même pour tous les autres tuples de sa classe (et réciproquement). Il suffit donc de tester la dominance d'un seul tuple d'une classe pour savoir si l'ensemble des tuples de la classe appartient ou non au Skycuboïde selon C. Ainsi, pour optimiser le calcul d'un Skycuboïde selon C à partir de sa partition π_C, on conserve uniquement un seul représentant de chaque classe d'équivalence. Cet ensemble est noté $reps(\pi_C)$. On diminue alors la taille de l'entrée en éliminant les tuples qui auraient conduit à un grand nombre de comparaisons inutiles. Pour prendre en compte les particularités du calcul de l'opérateur SKYLINE sur une partition, nous introduisons l'opérateur suivant.

Définition 4.1 - Soit C un ensemble de critères et π une partition de r. On définit un nouvel opérateur Π-SKY de la manière suivante :

$$\begin{aligned}\Pi\text{-SKY}_C(\pi_C) &= \{[t_i] \in \pi_C \mid \forall t_j \in r \text{ on a } t_j \not\succ_C t_i\} \\ &= \{[t_i] \in \pi_C \mid t_i \in \text{SKY}_C(r)\}\end{aligned}$$

Définition 4.2 (Concepts SKYLINES) - Soit $c_a = (C, \pi) \in \text{CONCEPTSACCORDS}(r)$ un concept ACCORD d'une relation r. Le concept SKYLINE associé c_s est défini de la manière suivante :

$$c_s = (C, \Pi\text{-SKY}_C(\pi))$$

Il y a exactement autant de concepts SKYLINES que de concepts ACCORDS. L'ensemble des concepts SKYLINES associés aux concepts ACCORDS de r est noté $\text{CONCEPTSSKYLINES}(r)$.

Les concepts SKYLINES sont des concepts ACCORDS où les partitions ont été contraintes. Ainsi, ce type de concept n'est plus forcément ordonné par la relation $\sqsubseteq$ entre partitions. La

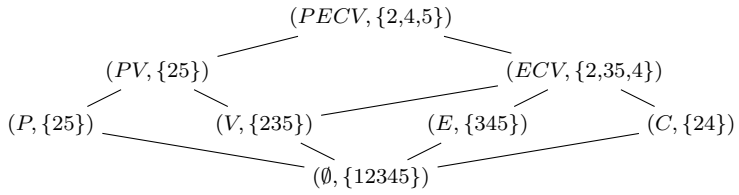

FIG. 2 – *Diagramme de Hasse du treillis des concepts* SKYLINES *de la relation* LOGEMENT

relation d'ordre $\leq$ entre concepts SKYLINES s'exprime donc de la manière suivante : Pour tout $(C_1, \Pi\text{-SKY}_{C_1}(\pi_1))$, $(C_2, \Pi\text{-SKY}_{C_2}(\pi_2)) \in \text{CONCEPTSSKYLINES}(r)$, nous avons alors $(C_1, \Pi\text{-SKY}_{C_1}(\pi_1)) \leq (C_2, \Pi\text{-SKY}_{C_2}(\pi_2)) \Leftrightarrow C_1 \subseteq C_2$.

Définition 4.3 (Treillis des concepts SKYLINES) -
L'ensemble ordonné $\langle \text{CONCEPTSSKYLINES}(r), \leq \rangle$ forme un treillis complet nommé treillis des concepts SKYLINES. Il est isomorphe au treillis des concepts ACCORDS.

Exemple 4.1 - D'après l'exemple 3.5, le couple $c_a = (ECV, \{1,2,35,4\})$ est un concept ACCORD. Le concept SKYLINE c_s associé est $c_s = (ECV, \Pi\text{-SKY}_{ECV}(\{1, 2, 35, 4\})) = (ECV, \{2, 35, 4\})$. L'identifiant 1 est éliminé de l'extension par l'opérateur Π-SKY car le tuple t_1 est dominé par t_2. La figure 2 donne le diagramme de Hasse du treillis des concepts SKYLINES de la relation LOGEMENT.

4.2 Matérialisation partielle du SKYCUBE

Ce paragraphe propose le treillis des concepts SKYLINES comme une matérialisation partielle des SKYCUBES. L'idée sous-jacente, afin d'obtenir une représentation réduite, est d'éliminer les cuboïdes les plus efficacement reconstructibles.

Définition 4.4 (Condition de non accord) - Soit r une relation et C un ensemble de critères. La condition $\text{CNA}_C(r)$ est vérifiée quand :

$$\nexists t_i, t_j \in r \text{ tels que } t_i[C] = t_j[C] \text{ avec } i \neq j, C \subseteq \mathcal{C}$$

Lorsque $\text{CNA}_C(r)$ est vérifiée, $\text{CNA}_X(r)$ l'est aussi avec X un sur-ensemble de C.

Cette condition de non accord est une version « affaiblie » de la condition de valeurs distinctes (Yuan et al., 2005) puisqu'elle s'applique sur les projections et non pas sur les valeurs individuelles de chacun des critères.

Définition 4.5 (Dominance sous CNA) - Soit $C \subseteq \mathcal{C}$ un ensemble de critères et r une relation. La relation de dominance $t \succ_C t$ sous l'hypothèse $\text{CNA}_C(r)$ s'exprime plus simplement :

Soit C tel que $\text{CNA}_C(r)$, $\forall t_i, t_j \in r$ on a $t_j \succ_C t_i$ ssi $\forall c \in C, t_j[c] \leq t_i[c]$ avec $i \neq j$

Lemme 4.1 - Soit r une relation. Pour tout ensemble de critères $C \subseteq \mathcal{C}$ vérifiant l'hypothèse de non accord $\text{CNA}_C(r)$, on a $\text{SKY}_C(r) \subseteq \text{SKY}_{C \cup c_0}(r)$ avec $c_0 \in \mathcal{C} \setminus C$.

Le contre exemple suivant montre que la propriété réciproque n'est pas vérifiée.

Exemple 4.2 - Soit $r = \{t_1, t_2\}$ une relation (avec $t_1 = (0, 1)$ et $t_2 = (1, 0)$) et $\mathcal{C} = \{A, B\}$ l'ensemble de ses critères. La condition $CNA_A(r)$ est bien vérifiée. On a $t_2 \notin \text{SKY}_A(r)$ car $t_1 \succ_A t_2$ alors que t_2 appartient bien à $\text{SKY}_{A \cup B}(r)$.

Corollaire 4.1 - Soit C tel que $\text{CNA}_C(r)$ est vérifiée. Par définition, pour tout X ($C \subseteq X$) vérifiant la condition $\text{CNA}_X(r)$, on a $\text{SKY}_C(r) \subseteq \text{SKY}_X(r)$.

Cette propriété est satisfaite pour tout sur-ensemble de C jusqu'à l'ensemble de tous les critères $\mathcal{C}$.

Théorème 4.1 (Théorème fondamental) - Soit r une relation, C un ensemble de critères et $h(C)$ sa fermeture. Alors :

$$\forall C \subseteq \mathcal{C} \text{ on a } \text{SKY}_C(r) \subseteq \text{SKY}_{h(C)}(r)$$

Exemple 4.3 - Dans la relation LOGEMENT, on a $\text{SKY}_{EC}(\text{LOGEMENT}) = \{t_4\}$. De plus, $h(EC) = ECV$ et $\text{SKY}_{ECV}(\text{LOGEMENT}) = \{t_2, t_3, t_5, t_4\}$. L'inclusion $\text{SKY}_{EC}(\text{LOGEMENT}) \subseteq \text{SKY}_{h(EC)}(\text{LOGEMENT})$ est donc bien vérifiée. Aussi, on a $\text{SKY}_E(\text{LOGEMENT}) = \{t_3, t_4, t_5\}$ avec $h(E) = E$. On note donc $\text{SKY}_E(\text{LOGEMENT}) \nsubseteq \text{SKY}_{EC}(\text{LOGEMENT})$.

Le théorème précédent montre une inclusion entre certains cuboïdes. Plus exactement, pour tout ensemble de critères, il y a une chaîne d'inclusions allant de ses générateurs minimaux jusqu'à sa fermeture. Cela implique que l'on peut calculer un cuboïde à partir d'un autre qui le contient. Ainsi au lieu d'utiliser la relation complète, on ne considère qu'un sous-ensemble restreint de celle-ci. Les concepts SKYLINES représentent les plus grands cuboïdes (selon l'inclusion) permettant d'en calculer d'autres. Ainsi en ne conservant que les concepts SKYLINES (*i.e.* non matérialisation des cuboïdes non fermés), on peut reconstruire rapidement les cuboïdes manquants simplement en trouvant le fermé à partir duquel ils peuvent être recalculés.

De plus grâce à la notion de classe d'équivalence des concepts SKYLINES, on évite un grand nombre de comparaisons inutiles. Les éléments indistinguables n'étant considérés que par groupes, la complexité du calcul ne dépend plus du nombre de tuples mais du nombre de groupes (classes d'équivalence).

Le treillis des concepts SKYLINES constitue donc une matérialisation partielle du SKYCUBE, où l'information non matérialisée est efficacement recalculable.

5 Comparaison avec les travaux antérieurs

L'opérateur SKYLINE (Börzsönyi et al., 2001) s'intéresse à trouver les éléments les plus intéressants dans un contexte base de données. Il tire ses origines du *maximal vector problem* (Bentley et al., 1978). Cependant les particularités du contexte font que des algorithmes spécifiques ont été développés. Les plus connus d'entre eux sont BNL (Börzsönyi et al., 2001), SFS (Chomicki et al., 2003), LESS (Godfrey et al., 2007). D'autres algorithmes plus efficaces tels que BBS (Papadias et al., 2005) ont été proposés mais se basent sur des structures d'index coûteuses à maintenir.

Le SKYLINE étant un outil d'aide à la décision, l'utilisateur va généralement calculer plusieurs SKYLINES avant de trouver celui qui l'intéresse vraiment. Pour répondre à cette problématique, le concept de SKYCUBE (Yuan et al., 2005; Pei et al., 2005) a été proposé. L'idée

générale étant de pré-calculer tous les SKYLINES, il est impératif de restreindre autant que possible le coût de stockage du résultat.

Une première approche de réduction a été proposée par Pei et al. (2006). Son objectif est de résoudre le problème de l'appartenance d'un objet donné à différents Skycuboïdes. Sa solution consiste à proposer une représentation du SKYCUBE fondée sur l'analyse formelle des concepts où chaque nœud correspond à un couple composé d'une classe d'équivalence (ensemble d'objets) et des sous-espaces dans lesquels elle est SKYLINE. Le principal défaut de cette approche orientée valeur est que le nombre de nœuds est borné par la cardinalité du treillis des parties des tuples ($|\mathcal{P}(Tid(r))|$). Les objets SKYLINES sont considérés avant tout, ce qui fait que pour reconstruire un Skycuboïde il faut parcourir un grand nombre de nœuds du treillis. Dans les approches orientées attribut comme la nôtre, le nombre de nœuds est bien plus restreint car il est borné par $|\mathcal{P}(\mathcal{C})|$.

Une autre approche intéressante est celle de Xia et Zhang (2006). Elle est orientée cuboïde (orientée attribut). L'objectif ici est de proposer une réduction très importante du SKYCUBE en éliminant les éléments considérés comme redondants des différents cuboïdes. Son inconvénient est que la reconstruction d'un Skycuboïde est délicate et relativement coûteuse car il faut parcourir une structure de données pour retrouver les éléments redondants. De plus, contrairement à l'approche de Pei et al. (2006, 2007) et à la nôtre, celle-ci n'est pas fondée sur un support théorique aussi solide que l'analyse des concepts formels. Cependant, l'un des principaux intérêts de cette approche en plus de la réduction importante du coût de stockage est l'efficacité de la mise à jour des données.

Notre approche ayant pour objectif de reconstruire les Skycuboïdes à moindre coût, elle se comporte idéalement dans ce cas-là. Malgré cette orientation ciblée, elle représente un bon compromis que ce soit pour la mise à jour ou pour la réduction de l'espace de stockage.

6 Conclusion

Dans ce papier, nous avons proposé le treillis des concepts SKYLINES qui est une instance contrainte d'un cadre formel plus général : le treillis des concepts ACCORDS. Cette nouvelle structure permet non seulement la matérialisation partielle sans perte d'information mais aussi la reconstruction efficace des cuboïdes manquants. Nous pouvons facilement étendre notre matérialisation au cas des ϵ-SKYLINES ou SKYLINES approximatifs (Papadias et al., 2005), en généralisant la définition d'ensemble en accord en remplaçant l'égalité stricte par une égalité à ϵ près. Une telle extension permet à l'utilisateur de relâcher la contrainte de dominance lorsque le nombre de résultats n'est pas suffisant.

Nous travaillons actuellement sur les aspects algorithmiques aussi bien pour le calcul du treillis des concepts SKYLINES que pour la reconstruction des résultats. Ce travail a pour objectif d'évaluer en pratique la contribution théorique décrite dans cet article.

Références

Beeri, C., M. Dowd, R. Fagin, et R. Statman (1984). On the structure of armstrong relations for functional dependencies. *J. ACM 31*(1), 30–46.

Bentley, J. L., H. T. Kung, M. Schkolnick, et C. D. Thompson (1978). On the average number of maxima in a set of vectors and applications. *J. ACM 25*(4), 536–543.

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

Chomicki, J., P. Godfrey, J. Gryz, et D. Liang (2003). Skyline with presorting. In U. Dayal, K. Ramamritham, et T. M. Vijayaraman (Eds.), *ICDE*, pp. 717–816. IEEE Computer Society.

Ganter, B. et R. Wille (1999). *Formal Concept Analysis : Mathematical Foundations*. Springer.

Godfrey, P., R. Shipley, et J. Gryz (2007). Algorithms and analyses for maximal vector computation. *VLDB J. 16*(1), 5–28.

Lopes, S., J.-M. Petit, et L. Lakhal (2002). Functional and approximate dependency mining : database and fca points of view. *J. Exp. Theor. Artif. Intell. 14*(2-3), 93–114.

Papadias, D., Y. Tao, G. Fu, et B. Seeger (2005). Progressive skyline computation in database systems. *ACM Trans. Database Syst. 30*(1), 41–82.

Pei, J., A. W.-C. Fu, X. Lin, et H. Wang (2007). Computing compressed multidimensional skyline cubes efficiently. In A. Dogac, T. Ozsu, et T. Sellis (Eds.), *ICDE*, pp. 96–105. IEEE.

Pei, J., W. Jin, M. Ester, et Y. Tao (2005). Catching the best views of skyline : A semantic approach based on decisive subspaces. In *VLDB*, pp. 253–264.

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.

Spyratos, N. (1987). The partition model : A deductive database model. *ACM Trans. Database Syst. 12*(1), 1–37.

Xia, T. et D. Zhang (2006). Refreshing the sky : the compressed skycube with efficient support for frequent updates. In S. Chaudhuri, V. Hristidis, et N. Polyzotis (Eds.), *SIGMOD Conference*, pp. 491–502. ACM.

Yuan, Y., X. Lin, Q. Liu, W. Wang, J. X. Yu, et Q. Zhang (2005). Efficient computation of the skyline cube. In *VLDB*, pp. 241–252.

Summary

The SKYLINE concept has been introduced in order to exhibit the best objects according to all the criterion combinations and makes it possible to analyse the relationships between SKYLINE objets. Like the data cube, the SKYCUBE is so voluminous that reduction approaches are really necessary. In this paper, we define an approach which partially materializes the SKYCUBE. The underlying idea is to discard from the representation the skycuboids which can be computed again easily. To meet this reduction objective, we characterize a formal framework: the AGREE concept lattice. From this structure, we derive the SKYLINE concept lattice which is one of its constrained instances. The strong points of our approach are: (i) it is attribute oriented; (ii) it provides a boundary for the number of lattice nodes; (iii) it facilitates the navigation within the Skycuboids.