

CND -Cube : Nouvelle représentation concise sans perte d'information d'un cube de données

Hanen Brahmi*, Tarek Hamrouni*
Riadh Ben Messaoud**, Sadok Ben Yahia*

* Département des Sciences de l'Informatique, Faculté des Sciences de Tunis
{tarek.hamrouni,sadok.benyahia}@fst.rnu.tn,

** Faculté des Sciences Économiques et de Gestion
riadh.benmessaoud@fsegn.rnu.tn

Résumé. Le calcul des cubes de données est excessivement coûteux aussi bien en temps d'exécution qu'en mémoire et son stockage sur disque peut s'avérer prohibitif. Plusieurs efforts ont été consacrés à ce problème à travers les cubes fermés, où les cellules préservant la sémantique d'agrégation sont réduites à une cellule, sans perte d'information. Dans cet article, nous introduisons le concept du *cube de données non-dérivable fermé*, nommé CND -Cube, qui généralise la notion des modèles non-dérivables fermés fréquents bidimensionnels à un contexte multidimensionnel. Nous proposons un nouvel algorithme pour extraire le CND -Cube à partir des bases de données multidimensionnelles en se basant sur trois contraintes anti-monotones, à savoir “être fréquent”, “être non dérivable” et “être un générateur minimal”. Les expériences montrent que notre proposition fournit la représentation la plus concise d'un cube de données et elle est ainsi la plus efficace pour réduire l'espace de stockage.

1 Introduction

Depuis les années 90, l'émergence des besoins en aide à la décision a conduit aux développements des *entrepôts de données*. Ces derniers ont apporté une solution adéquate et efficace au problème de stockage et de gestion des données. Un entrepôt est une base centralisée de grands volumes de données, historisées, organisées par sujet et consolidées à partir de diverses sources d'informations. Son contenu est analysé par les applications *Online Analytical Processing* (OLAP) qui fournissent aux utilisateurs des moyens pour naviguer dans les données multidimensionnelles afin d'y découvrir des connaissances interprétables, exploitables et utiles à la prise de décision.

Dans le but de répondre efficacement aux requêtes OLAP, le calcul des cubes de données est une solution fréquemment adoptée. Cependant, il est bien connu que le calcul des cubes de données est un problème combinatoire (Wang et al., 2002). Le volume des agrégats générés peut être incomparablement plus important que celui des données initiales, elles-mêmes déjà très volumineuses.

Par exemple, étant donné un contexte d'extraction R contenant n attributs, le nombre de tuples dans un cuboïde (Group-By) à k -attributs, tel que ($0 \leq k \leq n$), est le nombre de tuples