Evolution de schéma par classification automatique pour les entrepôts de données

Ony RAKOTOARIVELO *, Fadila BENTAYEB **

* ERIC, Université Lumière Lyon2, 05 av. Pierre Mendès-France, 69676 BRON Cedex, France
ony.rakotoarivelo@eric.univ-lyon2.fr, bentayeb@eric.univ-lyon2.fr
http://eric.univ-lyon2.fr

Résumé. Les modèles et outils OLAP actuels gèrent les dimensions d'analyse d'un entrepôt de données de manière statique. Par conséquent, les axes d'analyse restent souvent figés malgré l'évolution des besoins et des données. Dans cet article, nous proposons une approche d'évolution de schéma basée sur une technique de classification automatique. Pour cela, nous cherchons le meilleur regroupement des instances d'un niveau d'analyse choisi par l'utilisateur en utilisant la méthode des k-means. Un nouvel axe d'analyse est ensuite construit à partir du résultat de cette classification. Pour choisir les descripteurs du niveau d'analyse à classifier, nous proposons deux solutions: la première utilise directement les attributs décrivant le niveau à classifier. Par contre, la deuxième solution décrit le niveau d'analyse par les mesures dans la table des faits. Pour valider notre approche, nous l'avons intégrée et testée à l'intérieur du SGBD *(Système de Gestion de Bases de Données)* Oracle 10g.

1 Introduction

Un entrepôt de données est une base de données multidimensionnelle dont l'utilisation principale est l'analyse en ligne. Ce type d'analyse permet à l'utilisateur de naviguer à l'intérieur des données et effectuer des comparaisons dans le temps. Ainsi, elle impose deux contraintes majeures sur l'entrepôt : les données entreposées doivent être non volatiles et historisées. Pour gérer ces contraintes, les modèles OLAP actuels préconisent de proscrire toute forme de modification sur l'entrepôt. En conséquence, il est difficile d'adapter le schéma de l'entrepôt par rapport à l'évolution des besoins d'analyse.

Quelques travaux de recherches se sont alors penchés sur ce problème. Hurtado et al. sont parmi les premiers à proposer une algèbre d'évolution de schéma pour les entrepôts de données. (Hurtado et al., 1999a,b). Pour cela, ils modélisent une dimension d'analyse par un graphe acyclique direct où les noeuds représentent les attributs de la dimension et les arêtes représentent les liens hiérarchiques entre ces attributs. Ils proposent par la suite des opérateurs qui permettent de modifier la structure du graphe tout en préservant les propriétés du graphe