Manipulation et fusion de données multidimensionnelles

Franck Ravat, Olivier Teste, Gilles Zurfluh
Institut de Recherche en Informatique de Toulouse / Equipe SIG-ED
118, Route de Narbonne 31062 TOULOUSE cedex 04
mél : {ravat, teste, zurfluh}@irit.fr

Résumé. Cet article définit une algèbre permettant de manipuler des tables dimensionnelles extraites d'une base de données multidimensionnelles. L'algèbre intègre un noyau minimum d'opérateurs unaires permettant d'effectuer les analyses décisionnelles par combinaison d'opérateurs. Cette algèbre intègre un opérateur binaire permettant la fusion de tables dimensionnelles facilitant les corrélations des sujets analysés.

1 Introduction

Nos travaux se situent dans le cadre des systèmes décisionnels intégrant des bases de données multidimensionnelles (BDM). Conceptuellement, ces BDM organisent les données en sujets appelés faits et axes d'analyses appelés dimensions (Kimball, 1996).

1.1 Contexte : notre modèle conceptuel

Definition : Un fait F_j est défini par (N_{Fj}, M_{Fj}, I_{Fj}, $IStar_{Fj}$) où
- N_{Fj} est le nom du fait,
- $M_{Fj} = \{m_1, m_2, \ldots, m_w\}$ est un ensemble de mesures (ou indicateurs d'analyse),
- $I_{Fj} = \{I_{F\_1}, I_{F\_2}, \ldots\}$ est l'ensemble des instances de F,
- $IStar_{Fj}$ est une fonction associant chaque instance de I_{Fj} à une instance de chaque dimension liée au fait.

Definition : Une dimension D_i est définie par (N_{Di}, A_{Di}, H_{Di}, I_{Di}) où
- N_{Di} est le nom de la dimension,
- $A_{Di} = \{a_{Di\_1}, a_{Di\_2}, \ldots, a_{Di\_u}\}$ est un ensemble d'attributs,
- $H_{Di} = \{h_{Di\_1}, h_{Di\_2}, \ldots, h_{Di\_y}\}$ est un ensemble de hiérarchies,
- $I_{Di} = \{I_{Di\_1}, I_{Di\_2}, \ldots\}$ est l'ensemble des instances de D_i.

Definition : Une hiérarchie représente une perspective d'analyse précisant les niveaux de granularité auxquels peuvent être manipulés les indicateurs d'analyse. Une hiérarchie h_{Di_x} définie sur la dimension D_i est un chemin élémentaire acyclique débutant par l'attribut de plus faible granularité et se terminant par un attribut de plus forte granularité. Elle est définie par (N_{Di_x}, $Param_{Di\_x}$, $Suppl_{Di\_x}$) où
- N_{Di_x} est le nom de la hiérarchie,
- $Param_{Di\_x} = \langle a_{Di\_k}, a_{Di\_l}, \ldots, a_{Di\_z} \rangle$ est un ensemble ordonné décrivant la hiérarchie des attributs (chaque attribut est appelé paramètre de la hiérarchie et correspond à un niveau de granularité d'analyse),
- $Suppl_{Di\_x}$: $Param_{Di\_x} \rightarrow 2^{(A_{Di} - Param_{Di\_x})}$ est une application spécifiant les attributs faibles qui complètent la sémantique des paramètres (chaque paramètre est associé à un ensemble d'attributs faibles).