Un formalisme pour l'intégration de données hétérogènes

Sana Hamdoun*, Faouzi Boufares*

*Institut Galilée, Université Paris Nord Av. J. B. Clément 93430 Villetaneuse France
Sh@lipn.univ-paris13.fr
Boufares@lipn.univ-paris13.fr

Résumé. Dans ce papier nous proposons, un formalisme d'intégration de données hétérogènes. Nous définissons, d'une manière générale, une source de données comme un ensemble de composants muni des relations et fonctions qui relient ces composants, et un environnement d'intégration comme un ensemble de sources associé à un ensemble de ″liens d'intégration″ entre ces dernières.
L'approche générale d'intégration que nous proposons s'inscrit dans le cadre de la construction d'entrepôts de données hétérogènes basés sur des sources de catégories différentes, structurées, semi-structurées et non structurées telles que relationnelles, objet-relationnelles et XML. Le processus d'intégration est composé de trois étapes : le filtrage de l'ensemble des composants de l'entrepôt, la génération de son schéma global et la construction des vues qui le composent.

1 Introduction

Très généralement parlant, une base de données (BD) est une grande quantité d'informations stockées sur support informatique de telle sorte que son exploitation (mise-à-jour, recherche, extraction d'information, analyse et fouille de données (Benabdeslem et al., 2001) soit facilitée.

Dans une BD les ″éléments″ d'information sont reliés logiquement et physiquement. Ces éléments et liens sont organisés selon un modèle de données. Une *donnée* est une sorte de *composition* des éléments d'information via les liens envisagés, et dans ce cas *une BD est un ensemble de données*.

Dans ce papier, nous faisons l'abstraction des concepts de données et de bases de données, aussi bien au niveau logique qu'au niveau sémantique, et ceci indépendamment de tout processus de modélisation.

Une BD sera vue comme un ensemble de ″composants″ muni des liens pour les relier.

Sémantiquement, les données et leurs liens seront interprétés dans un *système de types*. Bien que ce dernier puisse dépendre du modèle envisagé, nous faisons aussi l'abstraction d'un tel système pour BDs, indépendamment du modèle.

Gérer plusieurs BDs revient donc à gérer plusieurs ensembles et plusieurs relations (au sens général du terme) sur ces derniers.

L'*intégration* des données est l'ensemble des transformations que subissent certaines données et leurs organisations afin d'être fusionnées et présentées éventuellement sous une nouvelle forme (entre-