Découverte de régularités pour l'intégration de données semi structurées

Pierre-Alain Laur, Xavier Baril

LIRMM – 161, Rue Ada
34392 Montpellier Cedex 5
{laur, baril}@lirmm.fr
http://www.lirmm.fr

Résumé. Cet article présente l'utilisation d'une technique de fouille de données pour aider à la spécification de vues sur des sources XML. Notre langage de vues permet d'intégrer des données XML provenant de sources hétérogènes. Cependant, la définition de motifs sur les sources permettant de spécifier les données à extraire est souvent difficile, car la structure des données n'est pas toujours connue. Nous proposons d'extraire les structures fréquentes dans les données des sources pour spécifier des motifs pertinents à utiliser dans la spécification des vues.

1 Introduction

L'objectif d'un système d'intégration de données est de fournir un accès unifié à différentes sources hétérogènes. Pour cela, on utilise généralement un mécanisme de vues et un langage commun (Halevy, 2003). Le choix d'XML est principalement motivé par deux raisons : (1) tout d'abord c'est un langage flexible qui permet de représenter des données provenant de différents modèles, (2) et ensuite de nombreuses applications actuelles permettent d'exporter leurs données en XML.

Pour intégrer des données XML, nous avons proposé VIMIX : *View Model for Integration of Xml sources* (Baril, 2003). Cependant, la définition de vues XML est souvent difficile car la structure des données des sources n'est pas toujours connue à l'avance. Cet article présente l'utilisation d'un algorithme de découverte de régularités pour aider à la spécification de vues XML. Nous avions présenté dans (Baril et al., 2001) une interface graphique pour la spécification de vues et un mécanisme d'aide qui ne tenait pas compte de la fréquence d'apparition des motifs dans les sources.

La recherche de régularités dans les bases de données a fait l'objet de nombreux travaux ces dernières années. La plupart des approches proposées s'intéressent à des structures plates ou fortement structurées. Cependant, contrairement aux bases de données traditionnelles où l'on décrit d'abord la structure des données (i.e. le type ou le schéma), les données XML peuvent ne pas être validées par une DTD ou un *XML-Schema*.

Un des objectifs de notre proposition est donc de découvrir les régularités structurelles existantes au sein d'un ensemble de documents semi structurés. Un exemple d'un tel ensemble est une source de données hétérogènes en XML. Nous considérons par la suite un arbre comme un graphe connecté acyclique et une forêt comme une collection d'arbres où