Recherche de sous-structures fréquentes pour l'intégration de schémas XML

Federico Del Razo López*, Anne Laurent*
Pascal Poncelet**, Maguelonne Teisseire*

* LIRMM - Université Montpellier II, 161 rue Ada 34392 Montpellier cedex 5
{delrazo,laurent,teisseire}@lirmm.fr
**EMA - LGI2P/Site EERIE, Parc Georges Besse 30035 Nîmes cedex 1
Pascal.Poncelet@ema.fr

Résumé. La recherche d'un schéma médiateur à partir d'un ensemble de schémas XML est une problématique actuelle où les résultats de recherche issus de la fouille de données arborescentes peuvent être adoptés. Dans ce contexte, plusieurs propositions ont été réalisées mais les méthodes de représentation des arborescences sont souvent trop coûteuses pour permettre un véritable passage à l'échelle. Dans cet article, nous proposons des algorithmes de recherche de sous-schémas fréquents basés sur une méthode originale de représentation de schémas XML. Nous décrivons brièvement la structure adoptée pour ensuite détailler les algorithmes de recherche de sous-arbres fréquents s'appuyant sur une telle structure. La représentation proposée et les algorithmes associés ont été évalués sur différentes bases synthétiques de schémas XML montrant ainsi l'intérêt de l'approche proposée.

1 Introduction

Étant donné l'explosion du volume de données disponibles sur Internet, il devient indispensable de proposer de nouvelles approches pour faciliter l'interrogation de ces grandes masses d'information afin de retrouver les informations souhaitées. L'une des conditions sine qua non pour permettre d'interroger des données hétérogènes est de disposer d'un (ou de plusieurs) "schéma général" que l'utilisateur pourra interroger et à partir duquel les données sources pourront être directement accédées. Malheureusement les utilisateurs ne disposent pas de moyen de connaître les modèles sous-jacents des données qu'ils souhaitent accéder et l'un des challenges dans ce contexte est donc de fournir des outils pour extraire, de manière automatique, ces schémas médiateurs. Un schéma médiateur est alors considéré comme une interface permettant à l'utilisateur l'interrogation des sources de données : l'utilisateur pose ses requêtes de manière transparente et n'a pas à tenir compte de l'hétérogénéité et de la répartition des données.

XML étant maintenant prépondérant sur Internet, la recherche de moyens d'intégration de tels schémas est un domaine de recherche actif. Si les recherches permettant l'accès aux données, quand un schéma d'interrogation est connu, sont maintenant bien avancées (Xylème, 2001), les recherches concernant la définition automatique d'un schéma médiateur restent incomplètes et non satisfaisantes (Tranier et al., 2004). Il est alors intéressant de considérer les