Une représentation des arborescences pour la recherche de sous-structures fréquentes

Federico Del Razo Lopez, Anne Laurent, Maguelonne Teisseire

LIRMM - Université Montpellier II
161 rue Ada 34392 Montpellier cedex 5
{delrazo,laurent,teisseire}@lirmm.fr

Résumé. La recherche de structures fréquentes au sein de données arborescentes est une problématique actuellement très active qui trouve de nombreux intérêts dans le contexte de la fouille de données comme, par exemple, la construction automatique d'un schéma médiateur à partir de schémas XML. Dans ce contexte, de nombreuses propositions ont été réalisées mais les méthodes de représentation des arborescences sont très souvent trop coûteuses. Dans cet article, nous proposons donc une méthode originale de représentation de ces données. Les propriétés de cette représentation peuvent être avantageusement utilisées par les algorithmes de recherche de structures fréquentes (sous-arbres fréquents). La représentation proposée et les algorithmes associés ont été évalués sur des jeux de données synthétiques montrant ainsi l'interêt de l'approche proposée.

1 Introduction

L'explosion du volume de données disponible sur internet conduit aujourd'hui à réfléchir sur les moyens d'interroger les grosses masses d'information afin de retrouver les informations souhaitées. Les utilisateurs ne pouvant pas connaître les modèles sous-jacents des données qu'ils souhaitent accéder, il est donc nécessaire de leur fournir les outils automatiques de définition de schémas médiateurs. Un schéma médiateur fournit une interface permettant l'interrogation des sources de données par l'utilisateur au travers de requêtes. L'utilisateur pose alors ses requêtes de manière transparente à l'hétérogénéité et la répartition des données.

XML étant maintenant prépondérant sur internet, la recherche de moyens d'intégration de tels schémas est indispensable. Si les recherches permettant l'accès aux données quand un schéma d'interrogation est connu sont maintenant bien avancées (Xyleme, 2001), les recherches concernant la définition automatique d'un schéma médiateur restent incomplètes et sont donc non satisfaisantes (Tranier et al., 2004). Dans le but de proposer une approche permettant de répondre à cette dernière problématique, nous nous focalisons sur la recherche de sous-structures fréquentes au sein d'une base de données de schémas XML. Une sous-structure fréquente est un sous-arbre se trouvant dans *la plupart* des schémas XML considérés. Cette proportion est examinée au sens d'un *support* qui correspond à un nombre minimal d'arbres de la base dans lesquels doit se retrouver le sous-arbre pour être considéré comme *fréquent*. Une telle recherche