Fouille dans la structure de documents XML

Amandine Duffoux, Omar Boussaid,
Stéphane Lallich, Fadila Bentayeb

ERIC – Université Lumière Lyon 2
5 avenue Pierre Mendès-France – 69676 Bron Cedex – France
{ aduffoux | boussaid | lallich | bentayeb }@eric.univ-lyon2.fr

Résumé. La prolifération des documents XML appelle des techniques appropriées pour extraire et exploiter l'information contenue dans ces documents. On distingue deux approches de fouille : *XML Content Mining* portant sur le contenu et *XML Structure Mining* qui a trait à la structure des documents. Combiner ces deux approches est très intéressant. Les informations contenues dans la structure orientent la fouille sur le contenu. Nous présentons la première étape de cette démarche : une nouvelle méthode d'extraction des règles d'association à partir de la structure des documents XML qui permet de gérer les aspects hiérarchiques de ces documents tout en améliorant les mécanismes d'extraction grâce à la création d'une structure spéciale représentant la hiérarchie des balises rencontrées.

Mots-Clés : document XML, structure, règles d'association.

1 Introduction

La norme XML (eXtensible Markup Language) s'impose comme le nouveau standard de transport des données. Son succès est dû à sa capacité de décrire toutes sortes de données à travers les concepts de DTD (*Document Type Definition*) ou de schémas XML qui sont de véritables grammaires. Il devient donc essentiel de mettre en place des techniques appropriées pour extraire et exploiter les informations contenues dans ces documents. La fouille de données dans les documents XML [Garofalakis *et al.*, 1999] se divise en fouille sur le contenu (*XML Content Mining*) et en fouille à partir de la structure (*XML Structure Mining*). La fouille sur le contenu utilise habituellement les techniques de *text mining*. Certains auteurs [Braga *et al.*, 2002] se sont notamment intéressés à l'extraction de règles d'association. Néanmoins, ces travaux ont très peu pris en compte, voir pas du tout, l'aspect hiérarchique existant entre les balises. L'intérêt de la fouille à partir de la structure des documents XML (structure intra et inter-documents) [Nayak *et al.*, 2002] porte sur l'information que véhicule l'organisation hiérarchique des balises. Certains travaux [Moh *et al.*, 2000] ont porté sur l'extraction d'une DTD à partir d'un ensemble de documents XML de même structure. Ces deux approches, jusqu'à présent séparées, sont complémentaires. Il nous paraît donc judicieux d'orienter la fouille sur le contenu grâce aux connaissances extraites à partir de la structure. Pour cela, nous souhaitons dégager les liens existants entre les balises (qu'elles soient ou non imbriquées) et utiliser ces résultats pour étudier leur contenu.

Dans cet article, nous présentons la première étape de cette démarche, l'extraction de règles d'association à partir de la structure d'un ensemble de documents XML. Les règles d'association ont prouvé leur efficacité dans la découverte de relations intéressantes parmi une grande masse de données [Agrawal *et al.*, 1993]. La méthode que nous proposons permet non seulement de gérer les aspects hiérarchiques des documents