Une extension de XQuery pour la recherche textuelle d'information dans des documents XML

Nicolas Faessel*, Jacques Le Maitre**

*LSIS (UMR CNRS 6168)
Université Paul CézanneDomaine Universitaire de Saint-Jérôme
Avenue Escadrille Normandie-Niemen
13397 Marseille Cedex 20
nicolas.faessel@lsis.org
**LSIS (UMR CNRS 6168)
Université du Sud Toulon-Var
BP 20132, 83957 La Garde
lemaitre@univ-tln.fr

Résumé. Nous présentons dans cet article une extension de XQuery que nous avons développée pour interroger le contenu et la structure de documents XML. Cette extension consiste à intégrer dans XQuery le langage NEXI, un sous-ensemble de XPath, défini dans le cadre de l'initiative INEX. Notre proposition est double : (i) équiper NEXI d'une sémantique floue, (ii) intégrer NEXI dans XQuery au moyen d'une métafonction appelée *nexi*, ayant une requête NEXI comme paramètre, et d'une extension de la clause *for* de l'opérateur FLWOR de XQuery. De plus, nous décrivons le prototype paramétrable que nous avons développé au dessus de deux moteurs XQuery classiques : Galax et Saxon.

1 Introduction

Il y a deux visions d'un document XML : une vision « centrée données » et une vision « centrée document ». Les documents XML « centrés données » sont constitués d'un ensemble d'éléments ayant une structure régulière : un ensemble de fiches bibliographiques, par exemple. Les documents XML « centrés document » décrivent des textes plus ou moins structurés : des livres scientifiques, par exemple. Pour interroger des documents XML « centrés données », le langage de requêtes XQuery (le SQL de XML), défini par le W3C (W3C, 2006b), est tout à fait bien adapté. Par contre, pour interroger des documents XML « centrés document » XQuery n'est pas suffisant lorsque l'interrogation est de nature sémantique, comme par exemple, la recherche des chapitres de livres qui concernent un certain sujet. De telles requêtes sont traitées traditionnellement par les systèmes de recherche d'information (Baeza-Yates et Ribeiro-Neto, 1999). Ce constat a conduit le W3C à proposer une extension de XQuery, XQuery Full-Text (W3C, 2006a), pourvue de fonctionnalités de recherche plein-texte. Le cœur de XQuery Full-Text est une fonction nommée *ftcontains* qui permet de tester si le contenu textuel d'un élément est conforme à une requête exprimée à l'aide d'opérateurs spécifiques : troncatures, connecteurs logiques, calcul de distance entre