Indexation de vues virtuelles dans un médiateur XML pour le traitement de XQuery Text

Clément Jamard*, Georges Gardarin*

Laboratoire PRiSM
Université de Versailles
78035, Versailles Cedex, France
prénom.nom@prism.uvsq.fr

Résumé: Intégrer le traitement de requêtes de recherche d'information dans un médiateur XML est un problème difficile. Ceci est notamment dû au fait que certaines sources de données ne permettent pas de recherche sur mot-clefs et distance ni de classer les résultats suivant leur pertinence. Dans cet article nous abordons l'intégration des fonctionnalités principales du standard XQuery Text dans XLive, un médiateur XML/XQuery. Pour cela nous avons choisi d'indexer des vues virtuelles de documents. Les documents virtuels sélectionnés sont transformés en objets des sources. L'opérateur de sélection du médiateur est étendu pour supporter des recherches d'information sur les documents de la vue. La recherche sur mots-clefs et le classement de résultat sont ainsi supportés. Notre formule de classement de résultats est adaptée au format de données semi-structurées, basé sur le nombre de mots-clefs dans les différents éléments et la distance entre les éléments d'un résultat.

1 Introduction

XQuery devenant le standard pour interroger XML, de nouveaux besoins apparaissent pour la recherche d'information. Buston et Rys (2003) spécifient des prédicats et fonctionnalités de recherche d'information à intégrer à XQuery, comme la recherche d'élément contenants des mots-clefs, le classement de résultats selon leur pertinence, la recherche basé sur des suffixes ou préfixes de mots. Un premier ensemble des fonctionnalités requises pour XQuery Text est défini par Buxton et Rys (2003). TexQuery, Amer-Yahia (2004), en est le langage précurseur.

Certaines des fonctionnalités citées précédemment, comme la simple recherche de mots-clefs, sont très communes et présentes dans la plupart des SGBD. Dans le cas de données distribuées, il faut d'abord recomposer les partitions avant de pouvoir effectuer une recherche sur le contenu ; d'importantes fonctionnalités souvent nécessaires aux applications ne sont pas faciles à implanter dans un système distribué. Le classement des résultats, les recherches conjonctives de mots-clefs, les recherches sur les racines de mots, leurs préfixes ou suffixes, sont difficilement réalisables car il faut auparavant recomposer les données dispersées.