Interrogation de données structurellement hétérogènes dans les bases de données orientées documents

Hamdi Ben Hamadou*, Faiza Ghozzi**
André Péninou*, Olivier Teste*

*IRIT, Université de Toulouse, UT3, UT2J, CNRS
118 Route de Narbonne - 31062 Toulouse, France
{hamdi.ben-hamadou, peninou, teste}@irit.fr
**Université de Sfax, ISIMS, MIRACL
Sakiet Ezzit 3021, Tunisie
faiza.ghozzi@isims.usf.tn

Résumé. Les systèmes orientés documents permettent de stocker tout document, quel que soit leur schéma. Cette flexibilité génère une potentielle hétérogénéité des documents qui complexifie leur interrogation car une même entité peut être décrite selon des schémas différents. Cet article présente une approche d'interrogation transparente des systèmes orientés documents. Pour cela, nous proposons de générer un dictionnaire de façon automatique lors de l'insertion des documents, et qui associe à chaque attribut tous les chemins permettant d'y accéder. Ce dictionnaire permet de réécrire la requête utilisateur à partir de disjonctions de chemins afin de retrouver tous les documents quelles que soient leurs structures. Nos expérimentations montrent des coûts d'exécution de la requête réécrite largement acceptables comparés au coût d'une requête sur schémas homogènes.

1 Introduction

Les systèmes de stockage « not-only SQL » (NoSQL) ont connu un important développement ces dernières années en raison de leur capacité à gérer de manière flexible et efficace d'importantes masses de données hétérogènes, Floratou et al. (2012); Stonebraker (2012). Les approches orientées documents sont couramment utilisées comme par exemple les systèmes MongoDB (Chodorow et Dirolf, 2010) ou CouchDB (Anderson et al., 2010). Ces systèmes reposent sur le principe de *« schemaless »* consistant à ne plus considérer un schéma unique pour un ensemble de données , appelé collection de documents (Chevalier et al., 2015). Cette flexibilité dans la structuration des données complexifie l'interrogation pour les utilisateurs qui doivent connaître les différents schémas des données manipulées (Chouder et al., 2017). Cet article traite de la problématique d'interrogation de données hétérogènes dans les systèmes NoSQL orientés documents.

Il existe différents types d'hétérogénéités (Shvaiko et Euzenat, 2005) : *L'hétérogénéité structurelle* désigne le problème de structures variables entre les documents. *L'hétérogénéité*