EXGRAF: Exploration et Fragmentation de Graphes au Service du Traitement Scalable de Requêtes RDF

Abdallah Khelil[1,3], Amin Mesmoudi[2], Jorge Galicia[1], Ladjel Bellatreche [1]

[1]LIAS, ISAE-ENSMA
(abdallah.khelil, bellatreche, jorge.galicia)@ensma.fr,
[2]Université de Poitiers
amin.mesmoudi@univ-poitiers.fr
[3]Université Oran 1, Algérie

Résumé. Les facilités de représentation de données offertes par RDF ont largement contribué à son succès et sa standardisation. Il est le langage incontournable pour le Web, la Biologie, etc. En allégeant la notion de schéma, il offre une représentation flexible des données. Son adoption rapide par les fournisseurs des données a contribué à la multiplication des masses de données RDF nécessitant à la fois d'un traitement efficace et scalable. Pour satisfaire ces besoins, plusieurs systèmes ont été proposés que nous divisons en deux catégories principales : (1) les systèmes orientés-mémoire comme gStore et (2) les systèmes orientés-disque comme RDF-3X et Virtuoso. Les systèmes de la 1ère catégorie sont très gourmands en mémoire surtout lorsqu'il s'agit de traiter une masse de données RDF. Ceux qui appartiennent à la 2ème catégorie utilisent la table traditionnelle en changeant automatiquement structure logique d'une donnée RDF. En conséquence, ils sont moins performants pour des requêtes complexes. Dans cet article, nous proposons une nouvelle approche orientée-disque (appelée EXGRAF) de traitement de requêtes RDF. Elle est basée sur l'EXploration logique de graphe RDF et la Fragmentation physique de triplets RDF. Nos expérimentations en utilisant des jeux de données réelles et synthétiques montrent le bon compromis entre l'efficacité et le passage à l'échelle de EXGRAF.

1 Introduction

Les systèmes traditionnels de stockage et de gestion des données ont été largement impactés par les nouveaux fournisseurs de données dites "graphes" comme le Web, le Web des données, les réseaux sociaux, la biologie. Leur approche consistant à définir un *schéma rigide (une structure)* dédié au stockage des données est rapidement devenue contraignante (Zou et al., 2011). Pour palier ce problème et satisfaire la structure graphe des données, de nouvelles représentations de données ont émergé. Resource Description Framework (RDF) est l'un des efforts menés par le World Wide Web Consortium (W3C) pour relier les données du Web. Il utilise la notion de triplets qui consiste à représenter chaque information sous forme d'un triplet $< sujet$, *prédicat*,$objet >$. Cette structure offre une flexibilité dans la collecte de données. Son adoption rapide par les fournisseurs des données a contribué à la multiplication des