

Extension de C-SPARQL pour l'échantillonnage de flux de graphes RDF

Amadou Fall Dia*, Zakia Kazi-Aoul*, Aliou Boly**, Yousra Chabchoub*

*ISEP, 10 rue de Vanves 92130 ISSY LES MOULINEAUX, France
prenom.nom@isep.fr,

**Université Cheikh Anta Diop de Dakar, BP 5005 Dakar-Fann, Sénégal
prenom.nom@ucad.edu.sn

Résumé. Les technologies du web sémantique sont de plus en plus utilisées pour la gestion de flux de données. Plusieurs systèmes de traitement de flux RDF ont été proposés : C-SPARQL, CQELS, SPARQL_{stream}, EP-SPARQL, SPARKWAVE, etc. Ces derniers étendent tous à la base, le langage d'interrogation sémantique SPARQL. Les données à l'entrée du système sont volumineuses et générées en continu à un rythme rapide et variable. De ce fait, le stockage et le traitement de la totalité du flux deviennent coûteux et le raisonnement presque impossible. Par conséquent, le recours à des techniques permettant de réduire la charge tout en conservant la sémantique des données, permet d'optimiser les traitements voire le raisonnement. Cependant, aucune des extensions de SPARQL n'inclut cette fonctionnalité. Ainsi, dans cet article, nous proposons d'étendre le système C-SPARQL pour générer des échantillons à la volée sur flux de graphes RDF. Nous ajoutons trois opérateurs d'échantillonnage (UNIFORM, RESERVOIR et CHAIN) à la syntaxe de C-SPARQL. Les expérimentations montrent la performance de notre extension en terme de temps d'exécution, et de la préservation de la sémantique des données.

1 Introduction

Notre usage quotidien des réseaux sociaux (Facebook, Twitter, Linkedin, etc.), des plateformes de contenus multimédias (YouTube, Flickr, iTunes, etc.), des réseaux de capteurs (observation, télé-relève, etc.), produit en continu des flux de données. Plusieurs groupes de recherches ont proposés des systèmes appliquant les technologies du web sémantique aux flux de données.

Étant donné la vitesse de génération des données et leur volume, le traitement de la totalité du contenu d'un flux est difficile et le raisonnement presque impossible. Ces systèmes nécessitent dès lors (i) des techniques d'allocation dynamique de ressources au système ou (ii) de réduction de la charge des données en entrée (résumé). Concernant cette dernière technique, aucune des extensions de SPARQL proposées (Streaming SPARQL Bolles et al. (2008), C-SPARQL Barbieri et al. (2010), CQELS Le-Phuoc et al. (2011), SPARQL_{stream} Calbimonte