Échantillonnage de flux de données sémantiques : Une approche orientée graphe

Fethi Belghaouti*, Amel Bouzeghoub*, Zakia Kazi-aoul**, Raja Chiky**

*Institut Mines-Télécom, Télécom SudParis, 9 rue Charles Fourier 91011 Evry Cedex France
prenom.nom@telecom-sudparis.eu,
**ISEP, 28 Rue Notre-Dame des Champs 75006 Paris France
prenom.nom@isep.fr

1 Introduction

Ces dernières années, nous assistons à la sémantisation des données statiques et dynamiques (flux de données). Toutefois, vu la spécificité de ces derniers ni les technologies du web sémantique ni celles des Systèmes de Gestions de Flux de Données (SGFD) ne peuvent les traiter. Pour ce faire, les chercheurs proposent aujourd'hui de nouveaux systèmes tels que C-SPARQL (Barbieri et al., 2010), CQELS (Phuoc) et SPARQL Stream (Calbimonte et al.). Lorsque le débit du flux en entrée de ces systèmes dépasse les seuils supportés, deux solutions existent : 1- Allouer au système autant de ressources que nécessaires (Hoeksema et Kotoulas, 2011) et (Phuoc et al.) ; 2- Réduire sa charge en entrée en se délestant d'une partie des données (Jain et al., 2013) et (Gao et al., 2014). Néanmoins, le délestage naïf de triplets RDF [1] (Sujet Prédicat Objet) dans (Jain et al., 2013), conduit à la destruction des liens sémantiques qui relient les données du flux, impliquant une réduction de leur niveau sémantique. Pour pallier cet inconvénient, nous proposons dans cet article un délestage orienté graphe.

2 L'approche orientée graphe pour l'échantillonnage de flux de données sémantiques

La Figure 1 illustre l'effet de la suppression de deux triplets RDF (en pointillé) en détruisant les liens reliant les nœuds (2) à (7) et (4) à (10). Ceci a pour conséquence de rendre les données du sous-graphe [7, 10, 13, 14, 15, 16, 17] inaccessibles (zone hachurée), malgré leur présence en mémoire, ce qui représente plus de 33% des données. Cet exemple illustre ainsi les effets négatifs de l'application naïve de l'approche orientée triplet RDF pour échantillonner un flux sémantique.

Notre approche consiste à élever la granularité de la donnée à considérer par les algorithmes d'échantillonnage, en traitant des sous-graphes RDF (ensemble contigu de triplets RDF liés) au lieu de triplets RDF, tels que [7, 13, 14 et 15] ou [10, 15, 16 et 17] de la Figure 1. De cette

1. http ://www.w3.org/standards/techs/rdf#w3c_all