Découverte de motifs à la demande dans une base de données distribuée

Lamine Diop[*,**], Cheikh Talibouya Diop[**], Arnaud Giacometti[*]
Dominique Li[*], Arnaud Soulet[*]

[*]Université de Tours, France
{arnaud.giacometti, dominique.li, arnaud.soulet}@univ-tours.fr
[**]Université Gaston Berger de Saint-Louis, Sénégal
{diop.lamine3,cheikh-talibouya.diop}@ugb.edu.sn

Résumé. De nombreuses applications s'appuient sur des bases de données distribuées. Pourtant, peu de méthodes de découverte de motifs ont été proposées pour les extraire sans centraliser les données. Il faut dire que cette centralisation est souvent moins coûteuse que la communication des motifs extraits. Pour contourner cette difficulté, cet article adopte une approche parcimonieuse en coûts de communication en fournissant à l'utilisateur des motifs à la demande. Plus précisément, nous proposons l'algorithme DDSAMPLING qui tire un motif dans une base de données distribuée proportionnellement à son intérêt. Nous démontrons son exactitude et analysons sa complexité en temps et en communication soulignant son efficacité. Enfin, une étude expérimentale montre sur plusieurs jeux de données la robustesse de DDSAMPLING face aux défaillances d'un site ou du réseau.

1 Introduction

De nombreuses applications requièrent un stockage et une manipulation de bases de données distribuées (Özsu et Valduriez, 2011). Le plus souvent, la centralisation des données est impossible à cause de contraintes légales ou techniques. Ainsi, Zhang et Zaki (2006) soulignent l'importance d'étendre la découverte de connaissances aux bases de données distribuées. Par exemple, les données du web sémantique sont réparties sur plusieurs triplestores accessibles uniquement via des requêtes SPARQL. Dans ce contexte, les propriétés décrivant une même entité (e.g., Paris) sont réparties sur plusieurs sites (e.g., Wikidata ou GeoNames). Cet article vise à extraire directement des motifs au sein de telles bases de données distribuées.

Peu de travaux de la littérature se sont intéressés à la découverte de motifs dans des bases de données distribuées (Cheung et al., 1996; Otey et al., 2003; Jin et Agrawal, 2006; Kum et al., 2006). Ces propositions se sont focalisées sur une extraction exhaustive des motifs en fusionnant les extractions réalisées localement sur chacun des sites. Malheureusement, le volume de données à transmettre entre les différents sites exige un coût de communication bien supérieur à la centralisation des données car les motifs sont nombreux par nature et les multiples extractions génèrent de multiples doublons. De plus, le coût de calcul de ces extractions parallèles est prohibitif même si des techniques d'élagage les diminuent sensiblement en contrepartie de