Vers une approche heuristique distribuée à base d'ontologie pour la fouille des règles d'association dans les données massives

Rania Mkhinini Gahar[*,***] Olfa Arfaoui[*,****] Minyar Sassi Hidri[*,**,****]
Nejib Ben Hadj-Alouane[*,***]

[*]Université de Tunis El Manar,
École Nationale d'Ingénieurs de Tunis, BP. 37, Le Belvèdère 1002, Tunis, Tunisie
{rania.mkhininigahar,olfa.arfaoui,minyar.sassi}@enit.rnu.tn, nejib_bha@yahoo.com
[**]Imam Abdulrahman Bin Faisal University, Arabie Saoudite
[***]UR-OASIS, ENIT
[****]LR-RISC, ENIT

Étant considérée comme la plus cruciale dans l'analyse Big Data, la prédiction devient un générateur d'innovation dans l'extraction de valeur. La fouille des Motifs Fréquents Maximaux (MFM) est une sorte de prédiction importante. Pour aller plus loin, nous pouvons aller au-delà de la phase de recherche des MFM pour générer des règles d'association (Agrawal et al., 1994; Gahar et al., 2017) qui est une tâche importante pour l'exploration du Big Data. En effet, l'extraction des RA est un processus itératif et interactif constitué de plusieurs phases allant de la sélection et la préparation des données à l'interprétation des résultats, en passant par la phase de recherche des connaissances : data mining. La plupart des approches proposées pour l'extraction des itemsets fréquents sont basées sur quatre étapes : (1) Préparation des données, (2) Extraction des motifs fréquents, (3) Génération de RA, et (4) Interprétation des résultats. Face aux quatre étapes précédentes, plusieurs problèmes peuvent être posés. Le temps de réponse de l'extraction des RA dépend principalement du temps d'extraction des itemsets fréquents car plusieurs balayages de contexte doivent être effectués en comptant pour chaque items et potentiel fréquent le nombre d'objets du contexte dans lequel il est contenu. Le nombre des motifs à considérer et la taille des jeux de données (contexte d'extraction) sont importants aussi. De plus, le flot de données massives pose le problème de générer un nombre prohibitif des RA extraites qui sont pour la plupart redondantes et inintéressantes. De plus, les règles redondantes représentent pour la plupart des types de données la majorité des règles extraites, c'est pourquoi leur suppression peut réduire considérablement le nombre de règles à gérer lors de la visualisation.Afin de répondre aux différents enjeux posés, nous proposons une nouvelle approche heuristique distribuée pour extraire les principales RA utiles à travers des MFM basés sur le framework MapReduce et enrichis par une ontologie. Une étape d'élagage sémantique est introduite dans les deux tâches et se réduit en tant que phases de pré-traitement et de post-traitement pour donner de la crédibilité et de l'efficacité à notre approche.

Notre approche distribuée Gahar et al. (2018), baptisée MARMO (MapReduce-based Association Rules approach through Maximal Fequent Itemsets for big Ontological data processing) qui repose essentiellement sur la recherche des MFM pour générer les RAs utiles. L'originalité de l'approche consiste à intégrer de manière explicite les éléments d'une ontologie de