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

domaine pour élaguer sémantiquement certains motifs candidats dans la découverte des motifs fréquents maximaux. L'apport de l'ontologie dans l'approche est d'abord sa terminologie, son expressivité et la puissance de son raisonneur qui permet de bénéficier de plus d'informations structurées afin d'élaguer sémantiquement certains motifs candidats dans le calcul des motifs fréquents maximaux et par la suite les règles d'association non redondantes. Les données massives représentent la source d'alimentation du système MARMO. Ensuite, ces données seront divisées de manière arbitraire pour être ensuite l'entrée des différentes Maps. Après cela, une étape d'élagage sémantique sera introduite ici pour donner naissance aux MFM non-redondants.L'élagage sémantique (Map) élimine du calcul des motifs fréquents maximaux tout candidat dont les éléments sont :Sémantiquement proches d'un concept de l'ontologie, dans la même hiérarchie de concepts de l'ontologie.Ceux-ci plus tard, après avoir été mélangés et triés, ils subissent à nouveau une introduction de l'étape d'élagage sémantique pour donner comme résultat final seulement les RAs utiles et non redondantes. Le recours à l'utilisation d'ontologie est distinguable dans chacune des phases d'élagage sémantique que ce soit dans la fonction ***Map()*** ou ***Reduce()***. L'ontologie est entrée dans le système comme étant un fichier OWL ou RDF. L'API Jena fournit aux utilisateurs la classe "OntModel". Cette dernière hérite de la classe "Model" pour représenter les modèles RDF.

Afin de vérifier l'efficacité et la performance de l'approche proposée, nous comparons MARMO avec l'algorithme MR-Apriori dans les aspects suivants : temps d'exécution, passage à l'échelle et gain de précision. Les résultats expérimentaux montrent bien que notre approche évite de manière significative la production d'un grand nombre de candidats, accélère la vitesse d'extraction des MFM et améliore simultanément le taux d'utilisation des ressources.

Nous travaillons actuellement à définir une démarche adaptée à la visualisation des RAs dans les données massives.

Références

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, Volume 1215, pp. 487–499.

Gahar, R. M., O. Arfaoui, M. S. Hidri, et N. B. Hadj-Alouane (2017). Parallelcharmax : An effective maximal frequent itemset mining algorithm based on mapreduce framework. In *Computer Systems and Applications (AICCSA), 2017 IEEE/ACS 14th International Conference on*, pp. 571–578.

Gahar, R. M., O. Arfaoui, M. S. Hidri, et N. B. Hadj-Alouane (2018). An ontology-driven mapreduce framework for association rules mining in massive data. *Procedia Computer Science 126*, 224–233.