Parallélisation de l'échantillonnage de motifs séquentiels

Lamine Diop*, Cheikh Ba**

*Université de Tours, France, lamine.diop@univ-tours.fr
**Université Gaston Berger de Saint-Louis, Sénégal, cheikh2.ba@ugb.edu.sn

Résumé. Durant ces 10 dernières années, le domaine de la fouille de données a connu d'importants travaux sur la découverte de motifs par échantillonnage en sortie. Très récemment, ces méthodes d'échantillonnage ont été appliquées sur des données séquentielles qui sont d'une nature complexe. La complexité de ces données réside sur leur structure qui a un impact notoire sur la rapidité du calcul et notamment sur le pré-traitement. A cela s'ajoute la taille des bases de données qui, de nos jours, deviennent très volumineuses. Dans ce papier, nous avons montré comment bénéficier du modèle de programmation BSP (Bulk Synchronous Parallel) pour améliorer l'efficacité des méthodes d'échantillonnage en sortie sur les données séquentielles. En effet, nous proposons un algorithme distribué et parallèle qui s'opère sur des bases de données séquentielles sciemment distribuées afin d'accélérer le temps de calcul. Les analyses que nous avons faites montrent l'impact positif du framework sur le temps d'exécution de la méthode.

1 Introduction

La découverte de motifs séquentiels (Agrawal et Srikant, 1995) est une tâche très étudiée dans de nombreux domaines de la vie réelle telles que la médecine, la biologie, l'exploration Web, etc. Au début, l'objectif principal était de trouver de manière exhaustive tous les motifs séquentiels qui ont une fréquence supérieure à un seuil donné. Les méthodes d'extraction exhaustive de motifs fréquents rencontrent deux problèmes majeurs. 1) Il n'est pas du tout facile de régler le seuil minimal de fréquence. S'il est trop élevé, la sortie risque d'être vide et s'il est très petit, l'ensemble des motifs renvoyés devient trop grand et alors difficile à analyser. 2) Le temps de calcul devient très élevé si la base de données est très volumineuse. Par conséquent, des méthodes de fouille de motifs dites optimales (Fournier Viger et al., 2017) ont été proposées pour supprimer le seuil tout en étant très rapides. Cependant, ces dernières rencontrent un problème sur la diversité des motifs produits. Par exemple, l'ensemble vide qui est le plus fréquent dans une base de données n'est pas souvent intéressant pour l'utilisateur.

Dans ce contexte, les méthodes d'échantillonnage en sortie (Al Hasan et Zaki, 2009; Boley et al., 2011; Diop et al., 2018) ont été proposées pour trouver des motifs très rapidement. Le but de ces méthodes est d'obtenir un échantillon de motifs où chaque motif est tiré avec une probabilité proportionnelle à une mesure d'intérêt choisie par l'utilisateur. Dans le cas des données séquentielles, l'algorithme CSSAMPLING (Diop et al., 2018) qui tire des motifs en fonction de la fréquence et sous des contraintes de normes minimales et maximales pour éviter la malédiction longue traîne a été étendu à un algorithme plus générique appelé NUSSAMPLING