Un algorithme distribué pour le clustering de grands graphes

Wissem Inoubli * **, Sabeur Aridhi**
Haithem Mezni*** Mondher Maddouri****, Engelbert Mephu Nguifo ‡

*Faculté des Sciences de Tunis, LIPAH
** Université de Lorraine, LORIA/ Inria NGE, France
*** SMART Lab, Institut supérieur de gestion de Tunis, Tunisie
**** College Of Business, University de Jeddah
‡ Université Clermont Auvergne, LIMOS

Résumé. Le clustering de graphes est l'une des techniques clés qui permet de comprendre les structures présentes dans les données de graphe. La détection des clusters et l'identification des ponts et des bruit sont également des tâches critiques car elles jouent un rôle important dans l'analyse des graphes. Récemment, plusieurs algorithmes de clustering de graphes ont été proposés et utilisés dans de nombreux domaines d'application. La plupart de ces algorithmes sont basés sur les algorithmes de clustering structurel. Néanmoins, ces derniers ont été conçus pour le traitement des petits graphes. D'où, leur performance peut se dégrader dans le cas des graphes larges qui imposent des défis supplémentaires. Dans cet article, nous proposons DSCAN, un algorithme distribué de clustering de graphes qui est basé sur le clustering structurel. Notre algorithme est implémenté sur la base de framework de traitement de grands graphes BLADYG. L'évaluation expérimentale de DSCAN a montré son efficacité et sa compétitivité pour le traitement de grands graphes.

1 Introduction

Plusieurs algorithmes de clustering de graphes sont proposés dans la littérature. À titre d'exemples, les approches basées sur la modularité. Le partitionnement de graphes, le min-cut et le clustering spectral. Ces méthodes fournissent en sortie une liste de clusters (groupes) qui ne sont pas vraiment suffisants pour comprendre le comportement du graphe étudié. Pour relever ce défi, l'algorithme de clustering structurel (SCAN) est proposé dans Xu et al. (2007). Le but de développement de ce dernier est non seulement l'identification des clusters dans le graphe, mais également la génération des informations supplémentaires telles que les ponts (sommets entre un ou plusieurs cluster) et les valeurs bruits (sommets qui n'appartiennent à aucun cluster). Le principe de fonctionnement de SCAN est basé sur la topologie ou la structure du graphe. Son principe de base consiste à regrouper des sommets qui partagent le nombre maximal de voisins. De plus, cet algorithme calcule la similarité entre tous les sommets du graphe afin de réaliser le clustering. Néanmoins, cette étape de calcul de similarité est linéaire et elle dépend du nombre d'arêtes du graphe d'entrée. Ceci dégrade la performance de SCAN en particulier dans le cas de grands graphes. En plus, le clustering structurel de graphes est