Extraction de motifs fréquents dans des arbres attribués

Claude Pasquier[*,**], Jérémy Sanhes[*]
Frédéric Flouvat[*], Nazha Selmaoui-Folcher[*]

[*]Université de Nouvelle Calédonie
PPME, BP R4, F-98851 Nouméa, Nouvelle Caledonie
{jeremy.sanhes,frederic.flouvat, nazha.selmaoui}@univ-nc.nc,
http://ppme.univ-nc.nc
[**]Institut de Biologie Valrose (IBV)
UNS - CNRS UMR7277 - INSERM U1091, F-06108 Nice cedex 2
claude.pasquier@unice.fr
http://ibv.unice.fr

Résumé. L'extraction de motifs fréquents est une tâche importante en fouille de données. Initialement centrés sur la découverte d'ensembles d'items fréquents, les premiers travaux ont été étendus pour extraire des motifs structurels comme des séquences, des arbres ou des graphes. Dans cet article, nous proposons une nouvelle méthode de fouille de données qui consiste à extraire de nouveaux types de motifs à partir d'une collection d'arbres attribués. Les arbres attribués sont des arbres dans lesquels les nœuds sont associés à des ensembles d'attributs. L'extraction de ces motifs (appelés sous-arbres attribués) combine une recherche d'ensembles d'items fréquents à une recherche de sous-arbres et nécessite d'explorer un immense espace de recherche. Nous présentons plusieurs nouveaux algorithmes d'extraction d'arbres attribués et montrons que leurs implémentations peuvent efficacement extraire des motifs fréquents à partir de grands jeux de données.

1 Introduction

L'extraction de motifs fréquents est une tâche importante dans le domaine de la fouille de données. Initialement centrée sur la découverte d'ensemble d'items (itemsets) fréquents (Agrawal et al., 1993), les premiers travaux ont été étendus pour extraire des motifs structurels comme les séquences (Agrawal et Srikant, 1995), les arbres (Chi et al., 2004a) ou les graphes (Washio et Motoda, 2003).

Alors que l'extraction d'itemsets fréquents recherche les combinaisons fréquentes d'items, l'extraction de motifs structurels recherche des sous-structures fréquentes. La plupart des travaux existants se focalisent sur un seul type de problème (fouille d'itemsets ou fouille structurelle). Toutefois, afin de représenter des données plus complexes, il semble naturel de considérer des collections structurées d'itemsets. Dans cet article, nous introduisons le problème de fouille d'arbres attribués. Les arbres attribués sont des arbres dans lesquels les nœuds sont associés à des itemsets.