Fouille de données et Visualisation : des ensembles fréquents maximaux aux dépendances fonctionnelles

Responsable : Romain Bourqui et Nicolas Hanusse

lieu : LaBRI

téléphone : 05.40.00.26.04

e-mail : `hanusse,romain.bourqui@labri.fr`

équipe : MaBioVis/CombAlgo, thème : Visualisation, Graphes projet : GRAVITE/CEPAGE

Mots-clés : fouille de données, visualisation, algorithmes de graphes

Description du sujet

Un des algorithmes de références en fouille de données porte sur le calcul des ensembles fréquents maximaux. Une *transaction* est constituée d'une liste d'items et une *table de transactions* comporte des milliers de transactions. l'objectif est de découvrir quelles sont les combinaisons d'items les plus fréquentes qui apparaissent dans une transaction.

Par exemple, une transaction peut représenter un ticket de caisse d'un magasin et la table de transactions peut constituer l'ensemble des tickets de caisse de l'année. Ainsi, il a été montré en utilisant un calcul de fréquents maximaux que de nombreux paires de famille achetaient simultanément bière et couche lors des courses hebdomadaires. De manière plus générale, le calcul de fréquents maximaux peut aider à batir des hypothèses sur les caractéristiques de maladies, la découvertes des chemins les plus employées dans les réseaux, ...

Un des obstacles de l'usage de tels algorithmes porte sur l'interaction avec les données : les résultats de calculs de fréquents maximaux sont souvent représentés par de grandes listes et seules les premières entrées sont regardées. D'autre part, le moindre changement de paramêtres implique un recalcul des fréquents maximaux et il devient impossible d'effectuer des comparaisons avec les résultats précédents.

l'objectif de ce travail est de proposer des mécanismes d'intéraction et visualisation des fréquents maximaux et de leur généralisation : les dépendances fonctionnelles. Par exemple, la représentation en graphe ou coordonnées parallèles pourront être envisagées. Des expérimentations sous Tulip permettront de faire des comparaisons sur des jeux de données réels.