TP DATA MINING

Résumé

Etant donnée le nombre croissant de grandes bases de données et de la complexité des algorithmes mis en œuvre pour leurs exploitation, la question d'optimisation préoccupe de plus en plus les chercheurs du domaine de la fouille de données.

L'une des techniques les plus utilisées pour extraire ces connaissances est la méthode des règles d'association.

Toutefois, la plupart des algorithmes d'extraction des itemsets fermés fréquents voient leurs performances se dégrader lorsque la taille des données augmente. Pour maintenir les performances de ces algorithmes, l'utilisation de méthodes et outils distribués apparait comme une solution naturelle.

Le projet demandé, vise à améliorer les performances de l'un des meilleurs algorithmes d'extraction des itemsets fermés fréquents a savoir l'algorithme Close.

Il s'agit de mettre en place une version de l'algorithme Close, dans le cas d'un contexte où les données seront partitionnées. Ceci permettra d'extraire des resultats locaux relatifs a chaque partition. Par la suite une seconde etape permettra de retrouver les résultats finaux a partir de ceux déjà extrait.

Dans notre cas, on considère ainsi n sources de données S1, S2,…, Sn. Soit D la taille de l'ensemble de nos données et di la taille de chaque source. Pour un itemset X donné, soit Sup(X) et Supi(X) le support de X dans Si respectivement. Dans ce cas, Sup(X) est appelé le support « global » et Supi(X) le support « local » de l'itemset X dans Si. Pour un support minimum donné minsup, X est globalement fréquent si: Sup(X) >=minsup * D.

De même, X est localement fréquent dans Si: si Supi(X) >= minsup * di.

Ainsi, le problème d'extraction des règles d'association se réduit à trouver tous les itemsets globalement fréquents et de générer ensuite les règles d'associations correspondantes.

Objectifs:

Il est demandé d'implémenter l'algorithme Close dans sa nouvelle version et de rédiger un rapport expliquant votre démarche et les différents choix fait.