DynaClose : Une approche de data mining pour la sélection des index de jointure binaires dans les entrepôts de données

Hamid Necir*, Ladjel Bellatreche**, Rokia Missaoui***

* Université de Bab Ezzouar BP 32 El Alia Bab Ezzouar ALGERIE
ncrhmd@yahoo.fr
** Université de Poitiers - LISI/ENSMA – FRANCE
bellatreche@ensma.fr
*** Université du Québec en Outaouais (UQO) – CANADA
rokia.missaoui@uqo.ca[1]

Résumé. L'indexation est l'une des techniques d'optimisation redondantes qui accélère les requêtes OLAP. Deux types d'index sont disponibles : les mono-index (B-tree, index binaire, projection, etc.) et les multi-index (index de jointure). Pour un entrepôt représenté par un schéma en étoile, les index de jointure binaires sont souvent utilisés pour accélérer les requêtes de jointure en étoile connues pour leur nombre important d'opérations de jointure. La sélection des index de jointure binaires est un problème difficile vu le nombre important des attributs candidats participant à la construction des index. Pour surmonter cette difficulté, nous proposons la démarche suivante : (1) nous adaptons d'abord un algorithme de fouille de données, appelé, *Close* qui permet de générer un ensemble d'itemsets fermés fréquents qui représentent les attributs candidats pour le processus de sélection des index. (2) Une fois les attributs candidats générés, nous proposons un algorithme itératif qui sélectionne un ensemble d'index de jointure binaires en prenant en compte l'ensemble des attributs candidats. Ces index doivent minimiser le coût d'exécution d'un ensemble de requêtes fréquentes et respecter une contrainte de stockage. Finalement, notre approche est validée par une étude expérimentale en la comparant avec les solutions existantes.

1 Introduction

Les entrepôts de données et les bases de données de grande taille sont souvent accédés par des requêtes complexes et coûteuses en terme de temps de calcul par le fait qu'elles nécessitent des opérations de jointure. Les entrepôts de données sont souvent représentés par un schéma en étoile constitué par une table des faits et de tables de dimension. Les requêtes typiques définies sur ce schéma sont appelées les *requêtes de jointure en étoile* (star join queries) qui ont les caractéristiques suivantes : (1) elles possèdent des jointures multiples entre la table des faits ayant une taille importante et les tables de dimension, (2) elles n'ont aucune jointure entre les tables de dimension (toute opération de jointure passe par la table

[1] Ce travail a été réalisé lors d'un congé sabbatique à l'Université Blaise Pascal à Clermont-Ferrand en France.