Vers le traitement à grande échelle de données symboliques

Omar Merroun*, Edwin Diday*, Philippe Rigaux*

*Univ. Paris Dauphine
omar.merroun@gmail.com, diday@ceremade.dauphine.fr, rigaux@lamsade.dauphine.fr

1 Introduction

L'Analyse de Données dites Symboliques (ADS) [DN07] a pour but d'analyser des unités statistiques de haut niveau appelées « concepts ». Ces concepts sont décrits par des données dites « symboliques » : intervalles, histogrammes, diagrammes, etc. Les méthodes implantées dans SODAS2[1] pour manipuler des Données Symboliques sont peu adaptées au traitement de grandes masses de données. De plus, elles sont complexes et non décomposables en opérateurs atomiques et clos. Cela empêche d'établir des stratégies d'optimisation globales pour évaluer ces méthodes. Nous proposons un modèle de données et une algèbre pour pallier ces problèmes. Nous visons à combiner un niveau logique où l'utilisateur exprime des méthodes d'ADS sous forme d'expression d'opérateurs algébriques clos, et un niveau physique d'évaluation, indépendant du premier, supportant des techniques efficaces d'évaluation.

2 Algèbre Symbolique

On s'intéresse à des *individus* qui sont des objets identifiables du monde réel. Ces individus forment une population Ω et sont décrits par des *variables* associées à des *types symboliques*. Ce modèle a été aussi adopté par d'autres types de bases de données : Statistiques et OLAP [Sho97]. Les variables forment un espace E de description des sous ensembles de Ω tel que chaque sous ensemble non vide est associé à un vecteur de description dans le domaine de E.

On s'inspire de l'algèbre des relations emboîtées [GG88] pour proposer notre algèbre. On définit les opérateurs atomiques de notre structure algébrique en se basant sur la notion de *résumé symbolique* qui est un ensemble de vecteurs de description d'une partition de Ω.

Ces opérateurs sont des opérateurs ensemblistes : ils s'appliquent sur des résumés pour produire un autre résumé dans E. La propriété de fermeture des opérateurs apporte de l'expressivité et permet de composer ces opérateurs sous forme d'une expression, dite *expression symbolique*.

[1] http ://www.ceremade.dauphine.fr/˜touati/sodas-pagegarde.htm