# 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 $\Omega$ 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 $\Omega$ 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 $\Omega$.

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

## 3 Evaluation des expressions symboliques

Outre les opérations ensemblistes des opérateurs algébriques, on peut appliquer des fonctions symboliques sur des variables de $E$. On définit deux formes de représentation des résumés symboliques : intentionnelle et extensionnelle. Ces représentations permettent de dissocier l'évaluation des fonctions symboliques appliquées sur des variables, de l'évaluation des opérateurs algébriques sur un résumé. La représentation intentionnelle retarde l'évaluation de la fonction symbolique, et ne garde que l'expression syntaxique des fonctions appliquées sur des variables de description du résumé symbolique. En revanche, la représentation extentionnelle évalue immédiatement les fonctions symboliques. L'évaluation retardée d'une expression est possible si elle n'intervient pas dans l'évaluation d'un opérateur algébrique. À défaut, on utilise la représentation extensionnelle de la variable.

## 4 Conclusion et perspectives

Notre approche vise, grâce à cette structure algébrique, à orienter l'ADS vers le traitement à grande échelle. Ces opérateurs algébriques clos apportent de l'expressivité à l'ADS et permettent de définir de nouvelles méthodes à un niveau logique. L'implantation de cette algèbre est en cours sur une base de données relationnelle. Par ailleurs, les représentations intentionnelles et extensionnelles permettent d'établir des stratégies d'évaluation adaptées au volume grandissant des données à traiter.

## Références

[DN07] E. Diday and M. Noirhomme. *Symbolic Data Analysis and the SODAS software*. Wiley, 2007.

[GG88] Marc Gyssens and Dirk Van Gucht. The powerset algebra as a result of adding programming constructs to the nested relational algebra. In *Proc. ACM Intl. Conf. on Data Management (SIGMOD)*, pages 225–232, 1988.

[Sho97] Arie Shoshani. Olap and statistical databases : similarities and differences. In *PODS '97 : Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 185–196, New York, NY, USA, 1997. ACM.

## Summary

This paper presents a first step towards large-scale manipulation of Symbolic Data. We introduce a data model and algebraic operators to support Symbolic Data Analysis. The model allows to evaluate symbolic operators in closed form, and brings the flexibility to support a lazy or opportunistic evaluation of symbolic functions and scalable operators.