Exploration de dépendances fonctionnelles et de règles d'association avec OLAP

Pierre Allard[1*], Sébastien Ferré*

*IRISA, Université de Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France
pierre.allard@irisa.fr, ferre@irisa.fr

Dans l'étude des bases de données, il est intéressant de déceler les liens entre attributs. Les Règles d'Association (RA) permettent de savoir quelles valeurs des attributs dépendent d'autres valeurs. Les Dépendances Fonctionnelles (DF) permettent de savoir quels attributs dépendent des autres. Les RA portent sur un sous-ensemble de la relation, tandis que les DF portent sur la relation complète. Les DF conditionnelles (DFC) sont un compromis puisqu'elles permettent de trouver des DF sur un sous-ensemble de valeurs de la relation. Medina et Nourine (2009) ont formalisé une hiérarchie entre les DF et les RA : une DF est l'union de DFC ; une DFC est une union de RA. Les systèmes de recherche de ces règles sont pour la plupart basés sur des algorithmes renvoyant la liste des règles avec leurs mesures.

Les outils d'On-Line Analytical Processing (OLAP, Codd et al. (1993)) permettent de structurer et d'analyser des informations multidimentionnelles, sous la forme d'un *cube* de données. Un cube décrit un ensemble de faits, selon une *mesure* (la valeur à analyser) et un ensemble de *dimensions* (les différentes facettes d'étude). Un ensemble d'opérateurs de navigation permettent à l'utilisateur (i) d'ajouter/supprimer une dimension (*add*/*dice*), (ii) diminuer/augmenter la granularité d'une dimension (*roll-up*/*drill-down*) et (iii) sélectionner une sous-partie du cube (*slice*). Traditionnellement, la mesure est numérique, afin d'aggréger les valeurs des cellules (ex : somme des produits vendus). Dans cet article, nous appliquons les concepts OLAP à une mesure comportant des valeurs quelconques. Notre méthode garde la mesure non aggrégée, sous forme d'un *nuage de tags*, où chaque *tag* est une valeur de la mesure, et où sa taille reflète sa fréquence. Nous montrons que cette utilisation particulière d'OLAP fait apparaître les RA, DFC et DF visuellement et directement dans le cube.

Nous définissons ici la projection d'une relation sur un cube OLAP. Un cube OLAP est défini par (i) une famille de n dimensions $(D_i)_{i\in 1..n}$, chaque dimension correspondant à un attribut de la relation, et (ii) une mesure M correspondant à un attribut de la relation. Le domaine de chaque dimension D et de la mesure est défini par l'ensemble des valeurs possibles de l'attribut correspondant. Chaque cellule $(d_1, \ldots, d_n)$ du cube contient, sous forme d'un nuage de tags, le multi-ensemble des valeurs de M, pour les tuples de la relation tels que $(D_1 = d_1)$, …, $(D_n = d_n)$. Nous utilisons le multi-ensemble afin de garder les fréquences des différentes valeurs, qui servent à évaluer les RA (ex : la confiance) et DF approximatives.

Un cube résultant de la projection fait apparaître simultanément toutes les DF, DFC et RA dont la prémisse est restreinte aux attributs $D_1, \ldots, D_n$ (les dimensions) et dont la conclusion porte sur l'attribut M (la mesure). L'ensemble des résultats suivants est basé sur le fait qu'une

[1]Pierre Allard bénéficie d'une bourse de thèse de la part de la Région Bretagne.

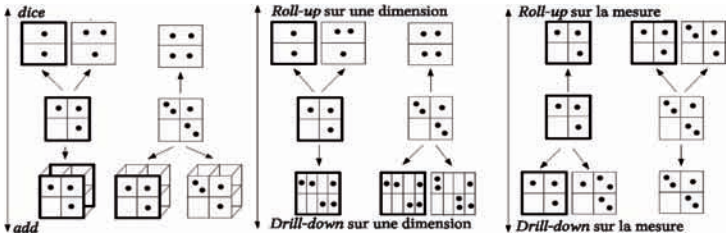

FIG. 1 – *Les différentes opérations OLAP accessibles. Une cellule avec un ou aucun point signifie une RA exacte ; avec deux points, une RA inexacte. Les cubes en gras vérifient une DF.*

cellule contient ou non une ou plusieurs fois une unique valeur. Si la cellule aux coordonnées $(d_1, \ldots, d_n)$ possède cette propriété (avec la valeur m), alors il existe une RA exacte $D_1 = d_1 \wedge \ldots \wedge D_n = d_n \rightarrow M = m$. Si toutes les cellules du cube (resp. une partie du cube) possèdent cette propriété ou sont vides, alors on aura une DF (resp. DFC) $D_1 \ldots D_n \rightarrow M$. Les réciproques de ces 3 conditions sont vérifiées, ce qui démontre un lien fort entre la relation et sa projection. On remarque aussi que la hiérarchie établie par Medina et Nourine (2009) est respectée : en effet, les DF sont une union de DFC, qui sont elles-mêmes une union de RA. Les opérateurs de navigation OLAP font apparaître ou disparaître des DF, DFC ou RA. La figure 1 résume ces comportements de façon synthétique.

Cette méthode d'exploration de DF, DFC et RA comporte quelques limites. Les problèmes de visualisation pour un nombre de dimensions supérieur à 2 sont un sujet récurrent des analystes OLAP, même si plusieurs papiers et logiciels présentent des solutions. Ensuite, c'est à l'utilisateur de choisir les dimensions et la mesure, ce qui implique que l'ensemble des règles affichées n'est pas exhaustif. Cependant, notre méthode s'applique à l'exploration de règle, en proposant un affichage simultané des DF, DFC et RA. Cet affichage structuré rend plus facile la recherche des sous-ensembles du cube où une DF est respectée, afin de trouver les DFC. L'affichage sous forme de nuages de tags permet de voir si quelques valeurs empêchent une DF, DFC ou RA d'être respectée ainsi de trouver les DF, DFC et RA approximatives.

Références

Codd, E., S. Codd, et C. Salley (1993). *Providing OLAP (On-line Analytical Processing) to User-Analysts : An IT Mandate*. San Jose : Codd & Date, Inc.

Medina, R. et L. Nourine (2009). A unified hierarchy for functional dependencies, conditional functional dependencies and association rules. In S. Ferré et S. Rudolph (Eds.), *Formal Concept Analysis*, Volume LNCS 5548. Springer.

Summary

We propose an exploration of functional dependencies and association rules by projecting a database relation into an OLAP cube. Each cube reveals FDs and ARs whose premises are the cube dimensions and whose conclusion is the cube measure.