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

Pierre Allard, Sébastien Ferré

▶ **To cite this version:**

Pierre Allard, Sébastien Ferré. Recherche de dépendances fonctionnelles et de règles d'association avec OLAP. Extraction et Gestion des Connaissances, Jan 2010, Hammamet, Tunisie. pp.651-652. inria-00549860

HAL Id: inria-00549860
https://inria.hal.science/inria-00549860

Submitted on 13 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Pierre Allard[1]*, Sébastien Ferré*

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

1 Problématique et solutions

Dans l'étude des bases de données, il est intéressant de déceler les liens entre les attributs. Pour ça, plusieurs outils sont disponibles. 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 donnent des informations à un niveau de granularité fin, tandis que les DF portent sur la relation complète. Les DF conditionnelles (DFC, Bohannon et al. (2007)) sont un compromis puisqu'elles permettent de trouver des DF sur un sous-ensemble de la relation. Medina et Nourine (2009) ont formalisé une hiérarchie entre les DF et les RA : une DF est l'union de DFC, qui sont elles-mêmes des unions de RA. Les systèmes de recherche de DF, DFC et RA sont pour la plupart basés sur des algorithmes renvoyant la liste des règles avec leurs mesures.

Les outils d'On-Line Analytical Processing (OLAP) permettent de structurer et d'analyser des informations multidimentionnelles, sous la forme d'un *cube* de données. Le cube décrit un ensemble de faits, selon une *mesure* (la valeur à analyser) et un ensemble de *dimensions* (les différentes facettes d'étude). De plus, un ensemble de liens de navigation permettent à l'utilisateur de (i) sélectionner une sous-partie du cube, (ii) modifier la granularité d'une dimension et (iii) supprimer une dimension d'étude. Traditionnellement, la mesure est numérique, ce qui permet de synthétiser les mesures des celulles, grâce à une fonction d'agrégation (somme, moyenne...). Dans cet article, nous appliquons les concepts d'OLAP à une relation comportant des valeurs quelconques. Pour ça, nous proposons une méthode qui utilise la structure du cube OLAP et ses opérateurs de navigation, mais en gardant la mesure non synthétisée. Nous montrons que cette utilisation particulière d'OLAP implique que les RA, DFC et DF apparaissent visuellement et directement dans le cube.

La première étape est la projection d'une relation sur un cube de données 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 de la relation. Chaque cellule $(d_1, \ldots, d_n)$ de la projection contient le multi-ensemble des valeurs m, pour les tuples de la relation tels que $(D_1 = d_1)$,

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

..., $D_n = d_n$ et $M = m$. Nous utilisons la notion de multi-ensembles afin de garder les cardinalités des différentes valeurs, qui pourront servir à calculer les RA et DF approximatives.

La projection de la relation dans ce cube permet de donner plusieurs informations sur les RA, DFC et DF. L'ensemble des résultats suivants est basé sur le fait que des cellules peuvent contenir une ou plusieurs fois une unique valeur. Premièrement, si dans le cube, aux coordonnées $(d_1, \ldots, d_n)$, la cellule contient une ou plusieurs fois une unique valeur m, alors il existe une RA exacte (i.e. ayant pour confiance 1) $D_1 = d_1 \wedge \ldots \wedge D_n = d_n \rightarrow M = m$. Deuxièmement, si dans une partie du cube, des cellules contiennent une ou plusieurs fois une unique valeur (ou sont vides), alors on aura une DFC $D_1 \ldots D_n \rightarrow M$ vérifiée sur ce sous-ensemble. Enfin, si toutes les cases du cube comportent une ou plusieurs fois une unique valeur (ou sont vides), alors on aura une DF $D_1 \ldots D_n \rightarrow M$. Les réciproques de ces 3 conditions sont vérifiées, ce qui démontre un lien fort entre la relation et sa projection en cube. On remarque aussi que la hiérarchie établie par Medina et Nourine (2009) est respectée avec cette visualisation. En effet, les DF sont une union de DFC, qui sont elles-même une union de RA.

Généralement, la création d'un cube OLAP traditionnel implique de faire un choix définitif sur les dimensions d'étude et la mesure (utilisés pour extraire le premier cube d'une base de données plus large). Pour naviguer dans le cube, nous utilisons les Systèmes d'Information Logiques (LIS), qui permettent de naviguer dans un ensemble de données grâce à des liens de navigation. L'utilisation des LIS permet à l'utilisateur de modifier à tout instant les dimensions de son cube, ainsi que la mesure.

2 Conclusion

Dans le domaine des bases de données, les outils de calcul de règles d'association et de dépendances fonctionnelles affichent traditionnellement les résultats sous forme d'une liste de règles, difficiles à lire. Nous avons proposé dans ce papier de projeter une relation d'une base de données sur un cube de données OLAP, afin d'afficher ces règles de manière plus structurée et plus intuitive. Les points forts d'OLAP (pas de requête, liens de navigation, structure des résultats) sont gardés, et les points faibles (structure rigide et choix définitifs) sont corrigés grâce aux Systèmes d'Information Logiques.

Références

Bohannon, P., W. Fan, F. Geerts, X. Jia, et A. Kementsietsidis (2007). Conditional functional dependencies for data cleaning. *Data Engineering, International Conference on 0*, 746–755.

Medina, R. et L. Nourine (2009). A unified hierarchy for functional dependencies, conditional functional dependencies and association rules. In S. B. . Heidelberg (Ed.), *Formal Concept Analysis*, Volume 5548/2009 of *Lecture Notes in Computer Science*, pp. 98–113.

Summary

In the database study, the tools used to compute association rules and functional dependencies give results as a rules list. We propose to project a relation of a database to an OLAP cube. In this cube, we demonstrate that the association rules and functional dependencies are visible. This projection displays the rules in a more intuitive structure. Finally, we show that OLAP navigation links help users to navigate into the producted rules.