Enrichissement de schéma multidimensionnel en constellation grâce à la Classification Ascendante Hiérarchique

Lucile Sautot*, Sandro Bimonte**,
Ludovic Journaux***, Arnaud Larrère****,
Kevin Saint-Paul****, Bruno Faivre****

*LIRIS, INSA de Lyon, Bât. Blaise Pascal, Campus de La Doua,
69622 Villeurbanne Cedex, lucile.sautot@insa-lyon.fr,
**IRSTEA Centre de Clermont-Ferrand, 9 Avenue Blaise Pascal,
63170 Aubière, sandro.bimonte@irstea.fr
***LE2I, Université de Bourgogne, Allée Alain Savary, 21000 Dijon
ludovic.journaux@agrosupdijon.fr
****Laboratoire Biogéosciences, Université de Bourgogne, 6 Boulevard Gabriel
21000 Dijon, bruno.faivre@u-bourgogne.fr

Résumé. Les hiérarchies sont des structures cruciales dans un entrepôt de données puisqu'elles permettent l'agrégation de mesures dans le but de proposer une vue analytique plus ou moins globale sur les données entreposées, selon le niveau hiérarchique auquel on se place. Cependant, peu de travaux s'intéressent à la construction de hiérarchies, via un algorithme de fouille de données, prenant en compte le contexte multidimensionnel de la dimension concernée. Dans cet article, nous proposons donc un algorithme, implémenté sur une architecture ROLAP, permettant d'enrichir une dimension avec des données factuelles.

1 Introduction

Les entrepôts de données (DW) et les systèmes OLAP sont des technologies permettant l'analyse en ligne de grands volumes de données. Les données entreposées sont organisées selon un modèle multidimensionnel qui définit les concepts de dimensions et de faits. Les dimensions représentent les axes d'analyse, qui sont organisés en hiérarchies, tandis que les faits, qui sont les sujets d'analyse, sont décrits par des indicateurs numériques appelés mesures (Kimball, 1996).

Les hiérarchies sont des structures importantes dans un entrepôt de données car elles permettent d'agréger plus ou moins finement les mesures, selon le niveau hiérarchique auquel on se place. C'est pourquoi plusieurs articles se sont intéressés à la construction de hiérarchies grâce a des algorithmes de fouille de données (Favre et al., 2006; Sautot et al., 2015). Cependant, les méthodes proposées prennent en compte uniquement les membres d'une dimension, et les faits et les autres dimensions du modèle en constellation ne sont pas impactés.

C'est pourquoi, dans cet article, nous présentons un algorithme d'enrichissement d'une dimension par des données dans un modèle multidimensionnel en constellation intégrant un algorithme de fouille de données, permettant la définition de structures hiérarchiques.

2 État de l'art

Plusieurs articles utilisent des algorithmes de fouille de données pour identifier des hiérarchies au sein d'une dimension. (Nguyen et Tjoa, 2000) proposent un système pour construire dynamiquement des hiérarchies à partir de données issues de Twitter. De plus, (Messaoud et al., 2004) présentent un nouvel opérateur OLAP, basé sur une classification ascendante hiérarchique, qui permet d'agréger des faits qui se réfèrent à des objets complexes comme des images. Par ailleurs, (Favre et al., 2006) fournit un système permettant de construire automatiquement des hiérarchies à partir de règles définies par les utilisateurs. Afin de personnaliser un schéma multidimensionnel, (Bentayeb, 2008) propose de créer de nouveaux niveaux dans une hiérarchie avec l'algorithme de K-means. D'autre part, (Leonhardi et al., 2010) proposent d'augmenter les fonctionnalités d'exploration d'un cube OLAP en fournissant à l'utilisateur des algorithmes de fouille de données pour analyser ces dernières. Enfin, Ceci et al. (2011) utilisent une classification hiérarchique pour intégrer des variables continues comme des dimensions dans un schéma OLAP. Cependant, les travaux qui s'intéressent à l'enrichissement de schémas multidimensionnels avec des hiérarchies utilisent soit uniquement des données dimensionnelles, soit uniquement des données d'un fait dépendant de la dimension. Or il est possible que la dimension soit enrichie par une hiérarchie créée en utilisant d'autres dimensions et faits du modèle en constellation.

3 Proposition théorique et implémentation

Dans cette section, nous présentons notre proposition pour l'enrichissement d'un schéma multidimensionnel grâce à la classification ascendante hiérarchique. L'idée principale est de fournir un algorithme qui transforme le schéma multidimensionnel en constellation en éliminant un noeud factuel et en intégrant les données factuelles dans une dimension associée, où elles seront utilisées pour créer de nouveaux niveaux.

Pour appuyer notre proposition, nous utiliserons un exemple de schéma multidimensionnel en constellation présenté sur la Figure 1 (haut). Bien que nous ayons développé notre proposition théorique sur un exemple réel, à propos de la biodiversité des oiseaux (Sautot et al., 2015), nous proposons notre démonstration sur un exemple artificiel de schéma multidimensionnel en constellation, qui présente une complexité plus intéressante.

3.1 Proposition théorique

Nous représentons un modèle multidimensionnel grâce à un graphe multidimensionnel. Un graphe multidimensionnel est un graphe dirigé M_G avec des noeuds dimensionnels (qui représentent les dimensions), des noeuds factuels (qui représentent les faits) et des arcs [1], dirigés uniquement d'un noeud factuel vers un noeud dimensionnel. De plus, il ne peut y avoir aucun

1. Dans le reste de cet article, la notation (f_i, d_j) désignera un arc sortant de f_i vers d_j.

noeud isolé, sans connexion avec un autre noeud, au sein de M_G. Cependant, M_G peut être constitué de plusieurs sous-graphes déconnectés entre eux, si chaque sous-graphe contient au moins un noeud factuel. Un exemple de graphe multidimensionnel est présenté sur la Figure 1 (haut).

Dans notre approche, nous souhaitons enrichir une dimension avec de nouvelles hiérarchies, calculées à partir de données factuelles. Cette dimension, la dimension Cible, notée d_t, d'un graphe multidimensionnel est une dimension telle que d_t est liée à au moins deux faits, dont l'un va être supprimé et utilisé pour créer de nouveaux niveaux au sein de d_t. Un exemple possible de dimension cible est la dimension d_t (Figure 1). Les données utilisées pour enrichir la dimension cible sont issues d'un noeud factuel, appelé "fait Source". La seule contrainte concernant le choix d'un noeud factuel comme fait Source est que ce noeud soit lié à la dimension cible. Un exemple de fait source possible est le noeud factuel f_s (Figure 1, haut). Ainsi, les données du fait source seront intégrées à la dimension cible, puis le fait source sera supprimé du graphe multidimensionnel, ce qui implique de redéfinir ce graphe et de manipuler les dimensions associées à ce noeud factuel.

Pour chaque noeud factuel $f_i \mid i \neq s$ tel que $\exists(f_i, d_t)$, il est possible de définir trois types parmi les dimensions liées à ce noeud : (i) la dimension cible d_t ; (ii) les dimensions Contextuelles D_i^c ; (iii) les dimensions Non-Contextuelles D_i^{nc}. Les dimensions contextuelles D_i^c sont les dimensions de M_G qui sont partagées par f_i et par f_s, le fait source. Dans le futur graphe modifié, les utilisateurs analyseront les mesures de f_i selon d_t, la dimension cible. Mais les données utilisées pour calculer les nouvelles hiérarchies de d_t proviennent de f_s et sont donc dépendantes des dimensions de D_i^c. C'est pourquoi nous devons nous assurer que les données utilisées pour créer la hiérarchie proposée à l'utilisateur sont cohérentes avec les données factuelles qu'il consulte pendant son analyse OLAP. Dans cet esprit, nous proposons un algorithme qui calcule des hiérarchies selon un contexte, ce contexte étant défini grâce à D_c^i. Par exemple, l'ensemble des dimensions contextuelles du noeud factuel f_3 est $\{d_3, d_4\}$. Les dimensions non-contextuelles D_i^{nc} sont les dimensions de f_i qui ne sont pas partagées avec f_s. Il n'y a donc pas de risque d'incohérence concernant ces dimensions. Pour supprimer une de ces dimensions, il est possible d'utiliser l'opérateur "Dice", classique en OLAP, qui est une agrégation des données factuelles au plus haut niveau d'une dimension. Par exemple, l'ensemble des dimensions non-contextuelles du noeud factuel f_1 est $\{d_6\}$.

Ainsi, nous calculerons plusieurs versions d'une hiérarchie : une version par combinaison de membres des dimensions de D_i^c. Nous utilisons la classification ascendante hiérarchique pour construire les nouvelles hiérarchie car le résultat de cet algorihtme a une structure proche d'un hiérarchie définie dans un schéma multidimensionnel (Messaoud et al., 2004). Les hiérarchies générées automatiquement sont strictes, onto et couvrantes (Malinowski et Zimányi, 2006). Une fois les versions calculées, elles peuvent être gérées comme des versions temporelles classiques d'une hiérarchie (Saroha et Gosain, 2014).

3.2 Implémentation

A présent que nous avons défini les dimensions contextuelles et non-contextuelles, nous proposons de décrire notre algorithme (Voir Algorithme 1). Le paramètre d'entrée de cet algorithme est le graphe multidimensionnel M_G présenté sur la Figure 1 (haut).

La sortie de cet algorithme est le graphe multidimensionnel présenté sur la Figure 1 (bas). On peut noter que f_s a été supprimé et qu'il y a de nouvelles dimensions basées sur le modèle de d_t et complétées par des hiérarchies contextuelles.

L'outil d'enrichissement a été développé sous Matlab®. Il se connecte à un serveur OLAP (Mondrian), qui interroge un entrepôt de données implémenté sur PostgreSQL. Nous avons utilisé le client OLAP Saiku pour construire des requêtes multidimensionnelles sur notre entrepôt de données.

Nous proposons une démonstration du fonctionnemnent de notre algorithme sous Matlab®. Nous montrerons les différentes étapes du processus d'enrichissement d'une dimension par des données factuelles et les principales fonctionnalités de l'outil :

1. la connexion du prototype à l'entrepôt de données ;
2. la sélection de la dimension cible et du fait source par un utilisateur ;
3. l'identification automatique des dimensions contextuelles ;
4. la génération des requêtes permettant de récupérer les instances de chaque contexte ;
5. l'exécution de ces requêtes ;
6. le calcul automatique des hiérarchies ;
7. la mise à jour de l'entrepôt de données et des cubes OLAP associés.

L'outil proposé fonctionne de maniére complètement automatique une fois que l'utilisateur a selectionné le fait source et la dimension cible.

Input : E un entrepôt de données, M_G un graphe multidimensionnel, d_t une dimension cible et f_s un fait source

for *chaque fait* f_i *dans* M_G **do**
  if *(*f_i *est lié à* d_t*) et (*f_i *n'est pas* f_s*)* **then**
    Ajouter une nouvelle table T dans E;
    Ajouter une nouvelle dimension d_{ti} dans M_G;
    Trouver l'ensemble D_i^C des dimensions contextuelles de f_i pour d_t;
    for *chaque combinaison* I *de membres de* D_i^C **do**
      Construire une requête R telle que :
      R = "SELECT Ensemble des mesures de f_s ON COLUMNS, Membres du plus bas niveau de d_t ON ROWS FROM Cube associé à f_s WHERE I";
      $Donnees$ = Exécuter(R);
      $Hierarchie$ = ClassificationAscendanteHiérarchique($Donnees$);
      Mettre à jour T avec $Hierarchie$;
      Mettre 'a jour d_{ti} avec $Hierarchie$;
    end
  end
end
return E, M_G

Algorithme 1 : L'algorithme principal

4 Conclusion et perspectives

La conception d'un entrepôt de données est une tâche complexe et cruciale, qui dépend des sources de données disponibles et des besoins en termes d'analyses décisionnelles. Une des

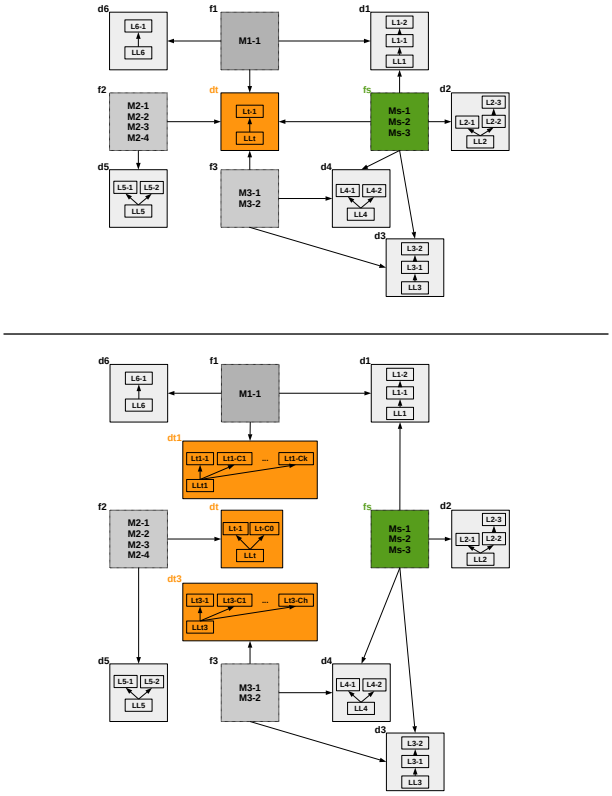

FIG. 1 – *Le graphe multidimensionnel M_G intial (haut) et à la sortie de notre algorithme (bas)*

étapes de cette démarche de conception est la définition de hiérarchies. Les travaux existants exploitent peu l'environnement factuel de la dimension considérée pour créer automatiquement des hiérarchies complexes. Ainsi, dans cet article, nous avons présenté un algorithme d'enrichissement d'un schéma multidimensionnel, qui transforme un schéma en constellation, en définissant de nouvelles hiérarchies grâce à la Classification Ascendante Hiérarchique. De plus, nous avons présenté une implémentation de cet algorithme sur une architecture ROLAP. Nos travaux en cours consistent en une extension de la méthodologie proposée dans cet article, afin de simplifier et de réduire le nombre de niveaux créés pendant le processus d'enrichissement, afin de proposer aux utilisateurs une exploration aisée des données lors d'une analyse OLAP et une mise en place simplifiée au sein d'une architecture ROLAP.

Références

Bentayeb, F. (2008). K-means based approach for olap dimension updates. In *Proceedings of the 10th International Conference on Enterprise Information Systems (ICEIS)*, pp. 531–534.

Ceci, M., A. Cuzzocrea, et D. Malerba (2011). Olap over continuous domains via density-based hierarchical clustering. In *Proceedings of the 15th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES)*, Volume 2, pp. 559–570.

Favre, C., F. Bentayeb, et O. Boussaid (2006). A knowledge-driven data warehouse model for analysis evolution. *Frontiers in Artificial Intelligence and Applications 143*, 2–71.

Kimball, R. (1996). *The Data Warehouse Toolkit : Practical Techniques for Building Dimensional Data Warehouses*. Wiley.

Leonhardi, B., B. Mitschang, R. Pulido, C. Sieb, et M. Wurst (2010). Augmenting olap exploration with dynamic advanced analytics. In *Proceedings of the 13th International Conference on Extending Database Technology (EDBT)*, pp. 687–692.

Malinowski, E. et E. Zimányi (2006). Hierarchies in a multidimensional model : From conceptual modeling to logical representation. *Data and Knowledge Engineering 59*(2), 348–377.

Messaoud, R. B., O. Boussaid, et S. Rabaséda (2004). A new olap aggregation based on the ahc technique. In *Proceddings of the 17th ACM International Workshop on Data Warehousing and OLAP (DOLAP)*, pp. 2004.

Nguyen, T. B. et A. M. Tjoa (2000). An object oriented multidimensional data model for olap. In *Proceedings of 1st International Conference on Web-Age Information Management (WAIM), number 1846 in LNCS*, pp. 69–82. Springer.

Saroha, K. et A. Gosain (2014). Multi-version data warehouse : A survey. In *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference* -, pp. 40–45.

Sautot, L., B. Faivre, L. Journaux, et P. Molin (2015). The hierarchical agglomerative clustering with gower index : a methodology for automatic design of olap cube in ecological data processing context. *Ecological Informatics 26*, 217–230.

Summary

Hierarchies are important structures in a data warehouse, because they offer several levels of precision on the analytical view of warehoused data. Most of actual methodologies for hierarchy building with data mining algorithms don't take account the multidimensional context of the modified dimension. Therefore, in this paper, we present an algorithm enriching a dimension with factuel data. This algorithm has been implemented on a ROLAP architecture.