Une nouvelle approche mixte d'enrichissement de dimensions dans un schéma multidimensionnel en constellation

Lucile Sautot*, Sandro Bimonte**, Ludovic Journaux***, Bruno Faivre*

*Laboratoire Biogéosciences (UMR CNRS-uB 6282)
6 boulevard Gabriel 21 000 Dijon FRANCE
lucile.sautot@u-bourgogne.fr, bruno.faivre@u-bourgogne.fr
**Irstea- Centre de Clermont-Ferrand
9 avenue Blaise Pascal 63178 Aubière FRANCE
sandro.bimonte@irstea.fr
***Laboratoire Électronique, Informatique et Image (UMR CNRS 6306)
allée Alain Savary 21000 Dijon France
ludovic.journaux@agrosupdijon.fr

Résumé. Les entrepôts de données (DW) et les systèmes OLAP sont des technologies d'analyse en ligne pour de grands volumes de données, basés sur les besoins des utilisateurs. Leur succès dépend essentiellement de la phase de conception où les exigences fonctionnelles sont confrontées aux sources de données (méthodologie de conception mixte). Cependant, les méthodes de conception existantes semblent parfois inefficaces, lorsque les décideurs définissent des exigences fonctionnelles qui ne peuvent être déduites à partir des sources de données (approche centrée sur les données), ou lorsque le décideur n'a pas intégré tous ces besoins durant la phase de conception (approche centrée sur l'utilisateur). Cet article propose une nouvelle méthodologie mixte d'enrichissement de schémas en constellation, où l'approche classique de conception est améliorée grâce à la fouille de données dans le but de créer de nouvelles hiérarchies au sein d'une dimension. Un prototype associé est également présenté.

1 Introduction

Les entrepôts de données (DW) et des 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. Les données entreposées sont ensuite explorées et agrégées en utilisant les opérateurs OLAP (par exemple, Roll-Up, Slice, etc.) (Kimball, 1996).

Le succès des projets de mise en place d'entrepôts de données dépend essentiellement de la phase de conception durant laquelle les besoins fonctionnels sont confrontés aux sources de données (Phipps et Davis, 2002). Trois principaux types de méthodologie de conception ont été développées : Une approche centrée sur les besoins utilisateurs, une centrée sur les sources

de données et une mixte (Romero et Abello, 2009). Les approches centrées sur les besoins utilisateurs mettent les décideurs au centre de la phase de conception en leur proposant des outils pour définir le modèle multidimensionnel uniquement en fonction de leurs besoins analytiques. Habituellement, les approches centrées sur les données proposent de déduire le schéma multidimensionnel à partir de sources de données structurées ou semi-structurées (Mahboubi et al., 2009; Jensen et al., 2004) en exploitant les métadonnées (par exemple les clés étrangères) et quelques valeurs empiriques. Enfin, les approches mixtes fusionnent les deux approches décrites précédemment.

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. Pour ces raisons, plusieurs travaux se penchent sur la définition de hiérarchies grâce a des algorithmes de fouille de données (Favre et al., 2006; Sautot et al., 2014). Cependant, cette phase de conception n'est appliquée qu'une fois que le modèle multidimensionnel a été définit et elle prend en compte uniquement les membres d'une dimension, et les faits et les autres dimensions du modèle en constellation ne sont pas impactées.

De notre point de vue, ces méthodologies présentent une limitation importante car, dans les projets réels d'entrepôt de données, les données qui décrivent les membres d'une dimension, et qui peuvent donc être utilisées pour créer une nouvelle hiérarchie, sont souvent issues de différents faits et dimensions préexistants dans le schéma multidimensionnel. C'est pourquoi, dans cet article, nous présentons une méthodologie mixte d'enrichissement de modèles multidimensionnels en constellation intégrant un algorithme de fouille de données dans une approche classique centrée sur les données. Ceci permet la définition de structures hiérarchiques, selon les spécifications décisionnelles des utilisateurs, qui ne pourraient pas être déduites par une approche centrée sur les données classique. Cette organisation hiérarchique de données dimensionnelles est transposée dans un modèle multidimensionnel et multi-factuel dans le but de représenter aussi bien que possible la sémantique des sources de données.

L'article est organisé selon les sections suivantes : la section 2 introduit l'état de l'art. Un cas applicatif et nos motivation sont présentés dans la section 3. Notre méthode de conception est détaillée dans la section 4 et son implémentation dans la section 5.

2 État de l'art

Trois types d'approches peuvent être utilisés pour la conception d'un entrepôt de données : (i) Les méthodes basées sur les spécifications utilisateurs ou approche centrée sur la demande ; (ii) les méthodes basées sur les données disponibles ou approches centrée sur les données ; (iii) les méthodes mixtes ou approches hybrides. Par exemple, (Jovanovic et al., 2012) est une méthode itérative centrée sur la demande où, à chaque itération, le système recherche les données correspondant le mieux aux informations requises par l'utilisateur en terme de dimensions ou de faits. De plus, plusieurs autres travaux ont proposé des systèmes basés sur une approche hybride, comme (Romero et Abello, 2010), qui propose d'exprimer les besoins fonctionnels grâce à des requêtes SQL.

Les approches centrées sur les données relationnelles déduisent des structures multidimensionnelles (faits ou dimensions) à partir de modèles conceptuels (Phipps et Davis, 2002) et/ou de modèles logiques (Carme et al., 2010; Jensen et al., 2004). Plusieurs travaux explorent en

		Sources de données utilisée pour construire la hiérarchie		
		Schéma en étoile		Schéma en constellation
		Attributs d'une dimension	Faits	Faits et attributs de dimensions
Algorithme utilisé	K-means	(Bentayeb, 2008)	(Bentayeb, 2008)	
	Classification Hiérarchique	(Ceci et al., 2011)	(Messaoud et al., 2004)	Notre proposition
	Autre	(Favre et al., 2006; Nguyen et Tjoa, 2000)	(Leonhardi et al., 2010)	

TAB. 1 – *Récapulatif de l'état de l'art sur la construction de hiérarchies*

particulier la découverte automatique de faits en utilisant des heuristiques (Carme et al., 2010). Concernant les dimensions, d'autres travaux proposent d'utiliser les métadonnées logiques d'une base de données comme par exemple les clés étrangères (Jensen et al., 2004).

D'autres articles utilisent des algorithmes plus complexes 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 nommé OPAC qui permet d'agréger des faits qui se réfère à des objets complexes comme des images. Cet opérateur est basé sur une classification ascendante hiérarchique. 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, toutes les méthodologies de création de schéma multidimensionnel en constellation s'arrêtent une fois tous les faits et les dimensions définies. 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, comme le montre la table 1. Or il est possible que la dimension soit enrichie par une hiérarchie créée en utilisant d'autres dimensions et faits du modèle en constellation. Cela signifie donc que la définition d'une nouvelle hiérarchie peut également impliquer la redéfinition de faits et de dimensions de tout le modèle en constellation.

Nous allons expliciter cette problèmatique dans la section suivante en utilisant, un vrai cas d'étude environnemental.

3 Cas applicatif

Afin de décrire nos motivations à proposer une nouvelle méthodologie d'enrichissement d'entrepôt de données, nous présenterons dans cette section un véritable cas d'étude concer-

nant les biodiversité des oiseaux (Sautot et al., 2014). Ce jeu de données a été collecté pour analyser les changements spatiotemporels au sein des communautés d'oiseaux le long de la Loire (France) et pour identifier les facteurs environnementaux locaux et globaux qui peuvent expliquer ces changements. Les données sont stockées dans une base de données relationnelle (PostGIS). En appliquant l'approche centrée sur les données proposée par (Romero et Abello, 2010), nous obtenons le schéma en constellation présentée sur la Figure 1, qui contient deux faits, décrits ci-après. Le premier fait est l'abondance, qui peut être analysée selon trois dimensions (une instance est présentée dans la Table 3) : (i) la dimension "Espèces", qui stocke les noms des espèces et leurs caractéristiques, (ii) la dimension "Années", qui correspond aux années de recensement des oiseaux, et (iii) la dimension spatiale "Points d'écoute", qui décrits les points d'écoute le long de la rivière. En utilisant un tel modèle, les décideurs peuvent répondre à des requêtes comme "*Quel est le nombre total d'oiseaux par année et par point d'écoute ?*" ou "*Quel est le nombre total d'oiseaux par année et par altitude ?*". Pour compléter le recensement des oiseaux, le paysage et la rivière ont été décrits autour de chaque point d'écoute. Ces descriptions environnementales représentent un autre fait, qui est associé à la dimension temporelle et à la dimension spatiale. Avec ce modèle, il est possible de décrire les points d'écoute. Par exemple, une requête OLAP possible est : "*Quel est le pourcentage de forêt autour de chaque point d'écoute en 2002 ?*".

Il faut noter que les attributs descriptifs des points d'écoute qui ne sont pas dépendants du temps, tels que l'altitude et la géologie, sont utilisés dans la dimension spatiale comme des niveaux de hiérarchies, tandis que les attributs dépendants du temps sont représentés comme des mesures d'un fait (par exemple le pourcentage de forêt). Malheureusement, les abondances d'une espèce d'oiseau n'ont pas de sens si elles ne sont pas corrélées avec les données environnementales décrivant le point d'écoute correspondant. Dans cette situation, une opération Drill-Accross n'est pas adéquate car elle conduit à masquer la dimension "Espèces". En effet, avec les opérateurs Drill-Accross, les faits sont joints uniquement sur les dimensions communes. De plus, le modèle multidimensionnel de la Figure 1 ne fournit pas aux décideurs la possibilité de construire une requête OLAP qui agrège les abondances selon les classes d'une variable environnementale, telle que par exemple "*Quel est le nombre total d'oiseaux par an et par groupe de points d'écoute avec 30% de forêt ?*" ou "*Quel est le nombre total d'oiseaux par an et par groupe de points d'écoute avec 50% d'eau ?*", car les paramètres environnementaux n'apparaissent comme des niveaux, mais comme des mesures, interdisant ainsi les requêtes GROUP-BY.

C'est pourquoi, dans notre cas d'étude, les décideurs ont besoin d'une nouvelle méthode de conception qui groupe les points d'écoute (données dimensionnelles) selon les paramètres environnementaux (données factuelles) et les années (données dimensionnelles).

Le modèle multidimensionnel permettant les analyses OLAP correctes est présenté sur la Figure 2 (Miquel et al., 2002). Ce schéma multidimensionnel présente un seul fait et la dimension spatiale est enrichie avec plusieurs niveaux représentant des groupes basés sur des paramètres environnementaux pour chaque année. En effet, les paramètres environnementaux décrivant les points d'écoutes en 2001 peuvent être différents de ceux de 2002, ce qui implique que le même point d'écoute n'est pas groupé dans le même niveau pour les deux différentes années, comme nous pouvons le constater dans la Table 3.

Par exemple, les données décrivant les activités agricoles autour des points d'écoute ne sont disponibles que pour l'année de recensement 2002. C'est pourquoi, il est important de prendre

Années	Points d'écoute	Agences	Pourcentage de forêt	Pourcentage de prairies
2002	1	LE2I	0.176	0.250
2002	1	ONEMA	0.356	0.261
2002	2	LE2I	0.311	0.420
2002	2	ONEMA	0.255	0.574
2011	1	LE2I	0.189	0.278
2011	1	ONEMA	0.241	0.385
2011	2	LE2I	0.322	0.568
2011	2	ONEMA	0.257	0.575

TAB. 2 – *Données factuelles du noeud "Environnements"*

en compte ces différentes classifications lors d'une session d'analyse OLAP impliquant une navigation dans la dimension temporelle. Par exemple, la requête "*quel est le nombre total d'oiseaux en 2002 et dans les points d'écoutes ayants les même paramètres environnementaux ?*" doit utiliser le niveau " *environnement type 2002*". Ainsi une requête OLAP utilisant le niveau environnement type 2002 et le membre temporel 2011 n'est pas cohérente car elle associe le nombre d'oiseaux en 2011 avec la configuration géographique et environnementale de 2002, pouvant ainsi induire des interprétations erronées.

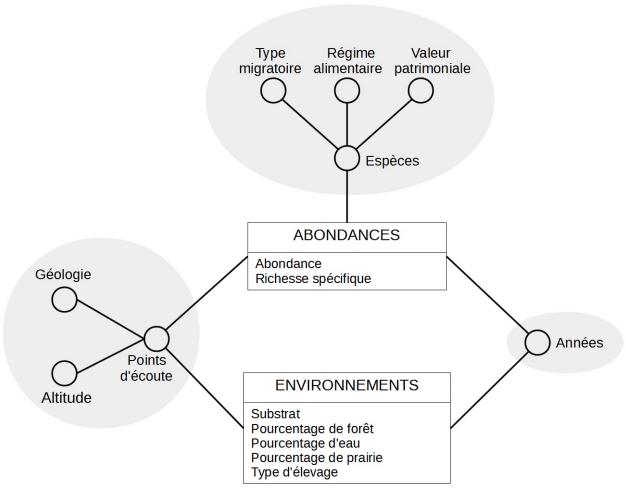

FIG. 1 – *Schéma en constellation de notre entrepôt de données sur la biodiversité des oiseaux*

4 Notre proposition

Dans cette section, nous introduirons notre méthodologie pour l'enrichissement d'un schéma multidimensionnel grâce à une approche mixte. L'idée principale est d'utiliser une méthodologie existante centrée sur les données comme première étape, afin d'obtenir un schéma

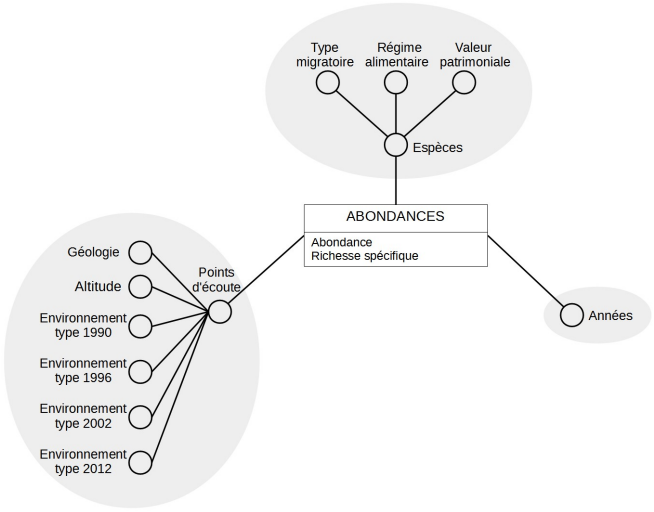

FIG. 2 – *Le schéma multi-versions couvrant les besoins des décideurs*

Années	Points d'écoute	Espèces	Abondance
2002	1	Bruant jaune	1.5
2002	1	Mésange noire	0.5
2002	2	Bruant jaune	1.5
2002	2	Mésange noire	0
2011	1	Bruant jaune	1
2011	1	Mésange noire	3
2011	2	Bruant jaune	1
2011	2	Mésange noire	2

TAB. 3 – *Données factuelles du noeud "Abondances"*

multidimensionnel en constellation. Après cela, nous enrichirons le modèle multdimensionnel obtenu : nous collecterons les besoins des utilisateurs à propos des hiérarchies qui ne peuvent pas être déduites par l'analyse de dépendances fonctionnelles puis, ces besoins utilisateur seront exprimés sous la forme de faits existant dans le modèle multidimensionnel à intégrer à une dimension. En fait, notre 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 elle seront utilisées pour créer de nouveaux niveaux.

Pour réaliser cela, nous formaliserons le modèle multidimensionnel sous forme de graphe multidimensionnel.

Dans la section suivante, nous exposerons des définitions à propos du graphe multidimensionnel (4.1), nous détailleront l'algorithme principal dans la section 4.2 et le calcul de nouvelles hiérarchies versionnées sera expliqué dans la section 4.3.

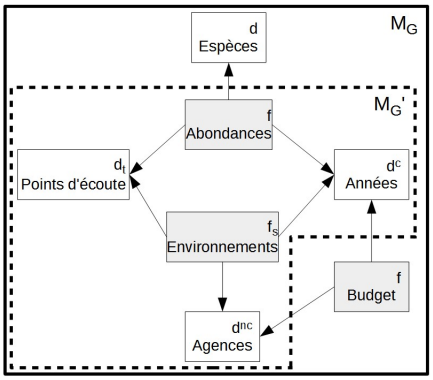

FIG. 3 – *Le graphe multidimensionnel M_G associé à notre jeu de données*

4.1 Préliminaires

Dans cette sous-section, nous présenterons quelques définitions préliminaires.
Nous représentons un modèle multidimensionnel grâce à un graphe multidimensionnel.

Définition 1. **Graphe multidimensionnel.** Un graphe multidimensionnel est un graphe dirigé $M_G =< D, F, A >$avec :
$D = \{d_1, ..., d_m\}$, les noeuds dimensionnels qui représentent les dimensions.
$F = \{f_1, ..., f_n\}$, les noeuds factuels qui représentent les faits.
$A = \{a_1, ..., a_p\} \mid \forall i \in [1, p], a_i = (f_j, d_k)$ [1] avec $j \in [1, n]$ et $k \in [1, m]$, les arcs, ce qui signifie que les arcs sont dirigés uniquement d'un noeud factuel vers un noeud dimensionnel. De plus, M_G ne contient aucun noeud isolé, sans liaison vers un autre noeud, mais peut contenir plusieurs sous-graphe déconnectés, si chaque sous-graphe contient au moins un noeud factuel.
Exemple. Un exemple de graphe multidimensionnel est présenté sur la Figure 3, où nous avons ajouté à notre cas d'étude un nouveau fait et une nouvelle dimension dans le but de mieux décrire notre algorithme.

Dans notre approche, nous faisons l'hypothèse qu'un décideur souhaite enrichir une dimension avec de nouvelles hiérarchies utilisant des données factuelles. Cette dimension est appelée dimension Cible.

Définition 2. **Dimension Cible.** La dimension Cible, notée d_t, d'un graphe multidimensionnel est une dimension telle que : $d_t \in D \mid \exists (f_1, d_t), ..., (f_u, d_t)\ avec\ u \in [2, n]$. Cela signifie 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.
Exemple. Un exemple possible de dimension cible est la dimension "Points d'écoute".

A présent, formalisons la définition du noeud factuel utilisé pour créer la nouvelle hiérarchie.

Définition 3. **Noeud Source.** Le noeud Source d'un graphe multidimensionnel M_G avec une dimension cible d_t est un noeud factuel f_s tel que $\exists a \in A \mid a = (f_s, d_t)$.

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

Année	Point d'écoute	Pourcentage de forêt	Pourcentage de prairies
2002	1	0.266	0.256
2002	2	0.283	0.497
2011	1	0.215	0.332
2011	2	0.290	0.572

TAB. 4 – *Données factuelles du noeud "Environnements' agrégées selon la dimension "Agences"*

Exemple. Avec la dimension "Points d'écoute" comme noeud cible, un exemple de noeud source possible est le noeud factuel "Environnements".

Comme nous l'avons précisé précédemment, notre algorithme supprime le noeud source du graphe. C'est pourquoi une partie de la structure du graphe est transformée. Il faut noter que seuls les noeuds liés au noeud source sont affectés par cette transformation. Nous avons donc défini un sous-graphe impliqué dans la transformation :

Définition 4. **Sous-graphe multidimensionnel Source-Cible.** Soit M_G un graphe multidimensionnel contenant une dimension cible d_t et un noeud source f_s, alors le sous-graphe multidimensionnel Source-Cible M'_G est un graphe multidimensionnel tel que : $M'_G =< D', F', A' >$ avec

$F' = \{f_i \in F \mid \exists\,(f_i, d_t)\}$

$D' = \{d_i \in D \mid \exists\,(f_s, d_i)\}$

$A' = \{(f_i, d_j) \mid f_i \in F', d_j \in D'\}$

Ainsi, M'_G contient seulement des noeuds factuels liés à d_t et des noeuds dimensionnels liés à f_s. Tous les noeuds factuels de M'_G sont donc liés au moins à une dimension et tous les noeuds dimensionnels de ce sous-graphe sont liés au mois à un fait. De plus, il n'y a pas de noeud isolé au sein de M'_G. M'_G est donc un graphe multidimensionnel bien formé.

Exemple. Un exemple de sous-graphe multidimensionnel Source-Cible utilisant l'exemple précédent est présenté sur la Figure 3.

Dans le but de formaliser les paramètres d'entrée de la classification ascendante hiérarchique utilisés lors de la création de nouveaux niveaux dans la dimension cible, nous avons défini le concept d'instance d'un noeud factuel, qui représente des données factuelles agrégées selon un ensemble de dimension.

Définition 5. **Instance d'un noeud factuel.** Soit M_G un graphe multidimensionnel. Soit m_i un membre de la dimension d_i. L'instance d'un noeud factuel f, notée $I\,(f, d_1.m_1, ..., d_n.m_n)$, est alors un ensemble de tuples représentant les faits de f agrégés selon les membres des n dimensions associées $d_1.m_1$, ..., $d_n.m_n$.

Exemple. Soit la Table 2, représentant l'instance du noeud factuel "Environnements", alors la Table 4 représente les faits agrégés selon le membre ALL de la dimension "Agences" : $I\,("Environnements", "Agences.ALL", "Années.1990", "Pointsd'écoute".*)$ [2].

2. '*' signifie "tous les membres de la dimension".

4.2 Algorithme

Dans cette section, nous décrirons en détail notre approche et nous la formaliserons.

Supprimer un noeud noeud factuel au sein d'un graphe multidimensionnel implique une redéfinition de ce graphe. Ainsi, le principe de notre approche est de travailler, dans un premier temps, uniquement sur le sous-graphe multidimensionnel Source-Cible, de transformer ce sous-graphe en ajoutant de nouveaux niveaux au sein de la dimension cible, de supprimer le noeud source, et, dans un second temps, de réintégrer le sous-graphe transformé au sein de l'ensemble de du graphe multidimensionnel original.

Supprimer le noeud source implique de manipuler les dimensions associées à ce noeud factuel. Il est possible de distinguer trois types parmi les dimensions liées au noeud source :

- La dimension cible d_t.
- Les dimensions Non-Contextuelles D_{nc}.
- Les dimensions Contextuelles D_c.

Les dimensions non-contextuelles D_{nc} sont les dimensions liées uniquement au noeud source dans M'_G. 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.

Exemple. Un exemple de dimension non-contextuelle est le noeud "Agences". La Table 4 est un exemple d'utilisation de l'opérateur "Dice" sur la dimension "Agences", qui est une dimension non-contextuelle.

Formellement,

Définition 6. **Dimension Non-Contextuelle.** Soit un sous-graphe multidimensionnel Source-Cible $M'_G =< D', F', A' >$, l'ensemble des dimensions non-contextuelles D_{nc} est défini par :

$$D_{nc} = \{d_1^{nc}, ..., d_v^{nc}\} \subset D' \mid \forall i \in [\![1, v]\!] \exists! (d_i^{nc}, f_j) \mid f_j \in F'$$

Notons que, dans la formule précédente, tous les noeuds dimensionnels dans D_{nc} ne sont liés qu'au noeud source f_s. En effet, tous les noeuds dimensionnels de M'_G sont liés à f_s et tous les noeuds dimensionnels de D_{nc} sont liés à un unique noeud factuel.

Les dimensions contextuelles D_c sont les dimensions de M'_G qui sont associées à f_s, le noeud source, et à une autre noeud factuel f. Dans le future graphe perfectionné, les utilisateurs analyseront les faits de f 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_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 système qui calcule des hiérarchies selon un contexte, ce contexte étant défini grâce à D_c.

Formellement,

Définition 7. **Dimensions Contextuelles.** Soit un sous-graphe multidimensionnel Source-Cible M'_G, alors l'ensemble des dimensions contextuelles D_c est défini par

$$D_c \subset D' \mid D_c = D' - (D_{nc} \cup \{d_t\})$$

avec d_t, la dimension cible.

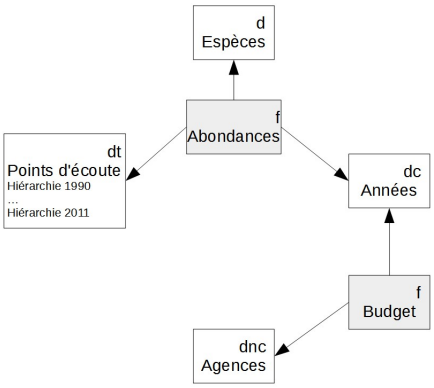

FIG. 4 – *Le graphe multidimensionnel M_G à la sortie de notre algorithme*

Exemple. Un exemple de dimension contextuelle est le noeud “Années”. Dans la Table 3, nous présentons des données provenant du noeud “Abondances” : ces données sont dépendantes du noeud dimensionnel “Années”.

A présent que nous avons défini les dimensions contextuelles et non-contextuelles, nous proposons de décrire notre algorithme (Voir Algorithme 1). Nous ferons l’hypothèse qu’il n’y a qu’une dimension contextuelle.

Le paramètre d’entrée de cet algorithme est le graphe multidimensionnel M_G présenté sur la Figure 3.

Input : M_G un graphe multidimensionnel, d_t une dimension cible et f_s un fait source

$M'_G \leftarrow GetSubGraph(M_G, d_t, f_s)$;
$d_t \leftarrow BuildHierarchies(M'_G, d_t, f_s)$;
$M_G \leftarrow Delete(f_s)$;
$M_G \leftarrow Clean(M_G)$;
return M_G

Algorithme 1 : L’algorithme principal

La sortie de cet algorithme est le graphe multidimensionnel présenté sur la Figure 4. On peut noter que f_s a été supprimé et qu’il y a de nouvelles hiérarchies dans le noeud “Points d’écoute”.

Notre approche permet donc d’enrichir une dimension avec une nouvelle hiérarchie, qui intègre différentes versions, et cela implique une transformation du modèle en constellation. Dans notre exemple, le graphe initial de la Figure 3 a perdu un fait ("Environnements") et les associations de ce fait avec ses dimensions ("Points d’écoute", "Années" et "Agences") (Figure 4). En revanche, nous avons enrichi la dimension spatiale avec une hiérarchie. En d’autre termes, l’enrichissement d’une hiérarchie correspond à une nouvelle phase de conception, qui impacte tout le modèle en constellation.

4.3 Création automatique de hiérarchies

Dans cette section, nous décrirons comment on peut créer de nouveaux niveaux dans la dimension cible.

La méthodologie complète permettant de créer de nouvelles hiérarchies dans un modèle multidimensionnel avec la Classification Ascendante Hiérarchique est présentée dans (Sautot et al., 2014). Les principales étapes de cet algorithme sont : (1) Le calcul des distances entres les individus. (2) Le choix des deux individus les plus proches. (3) L'agrégation des deux individus les plus proches au sein d'un cluster. Le cluster est ensuite considéré comme un individu. (4) Un retour de boucle vers l'étape 1, qui tourne tant qu'il y a plus d'un individu.

Dans notre approche, le clustering (CAH) a pour paramètres d'entrée l'instance du noeud source f_s évalué pour chaque membre de la dimension contextuelle et agrégé pour chaque dimension non-contextuelle.

Formellement, l'étape 2 de notre algorithme est réalisée par l'Algorithme 2.

Input : M'_G un sous graphe multidimensionnel Cible-Source, d_t une dimension cible, f_s un fait source

$d_c \leftarrow GetContext(M'_G, d_t, f_s)$;
$d_{nc} \leftarrow GetNonContext(M'_G, d_t, f_s)$;
for *each member* m *of* d_c **do**
  $I \leftarrow GetInstance(f_s, d_{nc}.ALL, d_c.m, d_t.*)$;
  $H \leftarrow CAH(I)$;
  $d_t \leftarrow SetNewHierarchy(d_t, H)$;
end
return d_t

Algorithme 2 : L'algorithme de construction de hiérarchies

En résumé, cet algorithme fonctionne de la façon suivante : dans le noeud source f_s, elle sélectionne les faits pour un n-uplet de membres des dimensions contextuelles (ensemble de dimensions D_c), les agrègent selon les dimensions non-contextuelles (ensemble de dimensions D_{nc}) et les affichent pour chaque membre de la dimension cible d_t. La méthode effectue ensuite un clustering hiérarchique des membres de d_t grâce aux données sélectionnées dans f_s comme expliqué ci-avant. Le résultat de ce clustering hiérarchique devient une nouvelle hiérarchie de d_t. La méthode recommence ensuite pour un autre n-uplet de membres de dimensions contextuelles.

Un exemple est présenté dans la Figure 5. Il faut noter que deux hiérarchies ont été créées pour la dimension spatiale, correspondant aux années 2002 et 2011. Notons de plus que, comme le schéma transformé est un graphe multidimensionnel bien formé, il est implémentable dans une architecture ROLAP.

5 Expérimentation et validation

La Classification Ascendante Hiérarchique, que nous également avons implémentée sous Matlab®, présente des performances suffisantes pour une utilisation lors d'une phase de design hors ligne (Sautot et al., 2014).

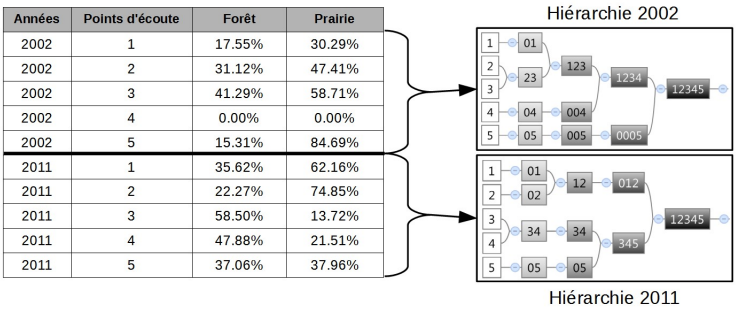

FIG. 5 – *Hiérarchies contextuelles pour la dimension "Points d'écoute"*

Dans cette section, nous traiterons de l'implémentation et de l'expérimentation des algorithmes proposés. Nous avons notamment développé un outil d'enrichissement de dimension et nous avons implémenté la Classification Ascendante Hiérarchique. Une évaluation qualitative et une évaluation des perfromances sont détaillées respectivement dans les sections 5.1 et 5.2.

L'outil d'enrichissement a été développé sous Matlab®. Cet outil permet de définir un graphe via une interface visuelle simple. Le graphe multidimensionnel traité est présenté sur la partie supérieure de l'écran, tandis que la partie inférieure permet à l'utilisateur de rentrer les paramètres d'entrée demandé par le système, via une fenêtre de commande.

5.1 Évaluation de la qualité du graphe produit

Dans cette section, nous d'écrirons les plus-values de notre méthodologie en termes de design (c'est-à-dire, nous évaluerons si notre méthodologie d'enrichissement produit un graphe qui répond aux besoin des décideurs). Pour cela, nous avons traité deux questions : 1) Les faits et les dimensions obtenues grâce à notre méthodologie correspondent-ils aux besoins analytiques des décideurs ? ; 2) Les hiérarchies produites grâce à notre méthodologie augmentent-elles les possibilités d'analyses ?

C'est pourquoi nous avons décidé de comparer le résultat obtenu avec notre méthodologie avec celui obtenu grâce aux travaux proposés dans (Miquel et al., 2002). En effet, (Miquel et al., 2002) proposent une méthode manuelle pour obtenir un schéma multidimensionnel multi-version. Or, quand la dimension temporelle est choisie comme dimension contextuelle, le résultat de notre approche est un schéma multidimensionnelle multi-version. Le résultat de cette validation montre que les schémas multidimensionnels obtenus manuellement et grâce à notre méthodologie sont identiques.

De plus, dans le but de valider la cohérence sémantique l'utilisation de la CAH dans la définition de hiérarchies, nous avons demandé aux écologues impliqués dans le projet de choisir entre une dimension spatiale avec un unique niveau, et une dimension spatiale avec une hiérarchie obtenue grâce à la CAH. Tant que le nombre de niveaux créés est inférieur à 5, les décideurs préfèrent avoir des hiérarchies qui peuvent contenir des motifs spatiaux intéressants, comme, par exemple, les profils agricoles des points d'écoute. Par exemple, les données dans la table de faits "Environnements" décrivent (entre autres) les pratiques agricoles autour de chaque point d'écoute pour chaque année. Une classification non-supervisée sur ces données

peut regrouper les points d'écoute et permet aux décideurs d'analyser l'impact des pratiques agricoles sur la biodiversité des oiseaux. Par exemple, les décideurs peuevent analyser la biodiversité selon l'occupation des sols autour des points d'écoute, en utilisant la simple requête OLAP : "Quelle est la valeur de la biodiversité pour chaque groupe de points d'écoute (le premier niveau de la hiérarchie obtenue via la clustering) en 2002 et en 2012 ?". Cette requête peut révéler que, pour la même année, la biodiversité est très affectée par les paramétres agricoles.

5.2 Évaluation des performances

Dans cette section, nous proposons de tester les performances en termes de temps de calcul, afin de valider la faisabilité de notre méthodologie, dans le cadre d'un projet de déploiement d'un entrepôt de données.

En particulier, nous avons étudié les performances de : 1) l'algorithme d'enrichissement, et 2) la création de hiérarchies grâce à la CAH.

Dans le but de tester le premier point, nous avons créé un ensemble de 200 schéma en constellation simulé, contenant entre 2 et 100 dimensions, sachant que les chémas multidimensionnels réels utilisables ont au maximm entre 3 et 10 dimensions (Kimball, 1996). Finalement, le pire temps d'exécution obtenu sur ces schémas est de 15,23 secondes. Nous avons mesuré les temps d'exécution de notre algorithme sur 200 graphes multidimensionnels et avons obtenu un temps moyen d'exécution de 11,7 secondes.

Étudions à présent les performances de la CAH. Dans ce paragraphe, les "objets classés" sont les points d'écoute (qui sont les membres dela dimensions "Points d'écoute", la dimension cible) et les "attributs" sont les faits agrégés issus du noeud factuel "Environnements" (qui est le fait source). La CAH , que nous également avons implémentée sous Matlab®, présente des performances suffisantes pour une utilisation lors d'une phase de design hors ligne. En effet, en utilisant notre cas applicatif, nous avons réalisé 2090 tests, avec un nombre d'objets classifiés entre 10 et 190, et un nombre d'attributs entre 10 et 100. Le temps de calcul moyen est égal à 0.072 secondes avec un écart-type de 0.002 secondes. Pour compléter notre évaluation, nous avons simulé un jeu de données avec 10 000 objets classifiés et 150 attributs. Dans ce cas, la CAH calcule une hiérarchie en 147.36 secondes, avec un écart-type de 4.03 secondes, et un temps de calcul maximal de 214 secondes. Ce temps de calcul (approximativement 4 minutes) est cohérent avec une phase de design hors ligne.

6 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 étapes de cette démarche de conception est la définition de hiérarchies. Les travaux existant exploitent peu l'environnement factuel de la dimension considérée pour créer automatiquement des hiérarchies complexes. Ainsi, dans cet article, nous avons présenté une méthodologie mixte 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.

Nous avons testé la méthodologie que nous proposons sur un cas applicatif réel, issu de l'étude de la biodiversité au sein des peuplements d'oiseaux. En fait, les méthodologies au-

tomatiques actuelles de design multidimensionnel ne peuvent pas produire un schéma multidimensionnel qui couvre les besoins des décideurs, en raison de la complexité des données. Notre méthodologie propose d'enrichir une dimension avec des données factuelles, et par ce moyen, transforme le schéma multidimensionnel afin de rendre possible de nouvelles analyses des données.

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.

Remerciements L'acquisition des données a bénéficié un support financier de la part du FEDER Loire, de l'établissement Public Loire, de la DREAL Bassin-Centre, de la Région Bourgogne (PARI, Projet Agrale 5) et du Ministère francais chargé de l'Agriculture. Nous les en remercions.

Références

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

Carme, A., J.-N. Mazon, et S. Rizzi (2010). A model-driven heuristic approach for detecting multidimensional facts in relational data sources. In T. Pedersen, M. Mohania, et A. M. Tjoa (Eds.), *Proceedings of 12th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, Volume LNCS 6263, pp. 13–24.

Ceci, M., A. Cuzzocrea, et D. Malerba (2011). Olap over continuous domains via density-based hierarchical clustering. In *15th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2011)*, 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*, 271.

Jensen, M. R., T. Holmgren, et Torben (2004). Discovering multidimensional structure in relational data. In *Data Warehousing and Knowledge Discovery : 6th International Conference (DaWaK)*.

Jovanovic, P., O. Romero, A. Simitsis, et A. Abelló (2012). Ore : An iterative approach to the design and evolution of multi-dimensional schemas. In *Proceedings of the Fifteenth International Workshop on Data Warehousing and OLAP*, DOLAP '12, New York, NY, USA, pp. 1–8. ACM.

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 *13th International Conference on Extending Database Technology (EDBT 2010)*.

Mahboubi, H., J.-C. Ralaivao, S. Loudcher, O. Boussaïd, F. Bentayeb, J. Darmont, et al. (2009). X-wacoda : an xml-based approach for warehousing and analyzing complex data.

Data Warehousing Design and Advanced Engineering Applications : Methods for Complex Construction, 38–54.

Messaoud, R. B., O. Boussaid, et S. Rabaséda (2004). A new olap aggregation based on the ahc technique. In *DOLAP 2004, ACM Seventh International Workshop on Data Warehousing and OLAP*, pp. 65–72.

Miquel, M., Y. Bédard, et A. Brisebois (2002). Conception d'entrepôts de données géospatiales à partir de sources hétérogènes. exemple d'application en foresterie. *Ingénieries des Systèmes d'information 7*(3), 89–111.

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

Phipps, C. et K. C. Davis (2002). Automating data warehouse conceptual schema design and evaluation. In *Proceedings of the 4th International Workshop on Design and Management of Data Warehouses (DMDW)*, Volume 2.

Romero, O. et A. Abello (2009). A survey of multidimensional modeling methodologies. *International Journal of Data Warehousing and Mining 5*(2), 1–23.

Romero, O. et A. Abello (2010). Automatic validation of requirements to support multidimensional design. *Data and Knowledge Engineering 69*, 917–942.

Sautot, L., B. Faivre, L. Journaux, et P. Molin (2014). The hierarchical agglomerative clustering with gower index : a methodology for automatic design of olap cube in ecological data processing context. *Ecological Informatics*. In Press.

Summary

Data warehouses (DW) and OLAP systems are business intelligence technologies allowing the on-line analysis of huge volume of data according to users' needs. Their success depends mainly on the design phase in which functional requirements meet data sources (mixed design methodology). However, existing design methods sometimes seem inefficient when decision makers define functional requirements that can not be deduced from the data sources (data driven approach), or when the decision maker has not integrated all these requirements during the design phase (user driven approach). This paper deals with a new mixed refinement methodology of constellation schemes, where classical design approach is enhanced with data mining in order to create new hierarchies in a dimension. An associate prototype is also presented.