Représentation graphique des hiérarchies contextuelles : modèles avec satellites

Cécile Favre, Anne Laurent, Yoann Pitarch, Pascal Poncelet

Représentation Graphique des Hiérarchies Contextuelles : Modèle avec Satellites

Cécile Favre*, Anne Laurent**
Yoann Pitarch**, Pascal Poncelet**

*ERIC, EA 3083, Université Lyon 2, Lyon, France
cecile.favre@univ-lyon2.fr
http://eric.univ-lyon2.fr
** LIRMM, UMR 5506, Université Montpellier 2, Montpellier, France
{pitarch,laurent,poncelet}@lirmm.fr,
http://www.lirmm.fr

Résumé. Les modèles d'entrepôts dits classiques (étoile, etc.) ont émergé et connu un vif succès au sein des entreprises de part leur présentation graphique facile à lire. Ceci est nécessaire dans un contexte où la modélisation multidimensionnelle doit être confrontée à l'avis des décideurs. Dans des travaux antérieurs, nous avons mis en avant un certain manque d'expressivité de ces modèles. Par exemple, dans le cas d'un entrepôt de données médicales, il n'était pas possible de modéliser le fait que la tension artérielle d'un patient soit « faible », « normale » ou « élevée » (hiérarchisation de la mesure) dépend de son âge et du fait qu'il fume ou non. Ayant développé une formalisation de hiérarchies dites « contextuelles » pour pallier à ce problème, nous proposons dans ce papier un modèle graphique, pour faciliter la lisibilité du modèle auprès des décideurs, que nous baptisons le « modèle avec satellites ».

1 Introduction

Les entrepôts de données (Inmon, 1996) permettent de consolider, stocker et organiser des données à des fins d'analyse. Des faits peuvent alors être analysés à travers des indicateurs (les mesures) selon différents axes d'analyse (les dimensions). En s'appuyant sur des mécanismes d'agrégation, les outils OLAP (On Line Analytical Process)(Agrawal et al., 1997; Chen et al., 1996; Han, 1997) permettent de naviguer aisément le long des hiérarchies des dimensions. La puissance de ces outils place les entrepôts au centre des systèmes d'information décisionnels (Mallach, 2000). Ces considérations justifient l'émergence d'entrepôts dans des domaines aussi variés que l'analyse de ventes, la surveillance de matériel, le suivi de données médicales (Einbinder et al., 2001)... Dans cet article, nous considérons cette dernière application des données médicales [1] pour illustrer la problématique traitée et la solution apportée.

Considérons un entrepôt de données médicales enregistrant les paramètres vitaux (*e.g.*, le pouls, la tension artérielle...) des patients d'un service de réanimation. Afin de réaliser un suivi

1. Ce travail a été réalisé dans le cadre du projet ANR MIDAS (ANR-07-MDCO-008).

efficace des patients, un médecin souhaiterait par exemple connaître ceux qui ont eu une tension artérielle basse au cours de la nuit. Pouvoir formuler ce type de requête suppose l'existence d'une hiérarchie sur la tension artérielle dont le premier lien d'agrégation serait une catégorisation de la tension artérielle (*e.g.*, basse, normale, élevée). Toutefois, cette catégorisation est délicate car elle dépend à la fois de la tension artérielle mesurée mais aussi de certaines caractéristiques physiologiques (âge du patient, fumeur ou non, ...). Dès lors, une même tension peut être généralisée différemment selon le contexte d'analyse considéré. Par exemple, une valeur de 13 pour la tension artérielle est *élevée* chez un *nourrisson* alors qu'il s'agit d'une tension *normale* chez un *adulte*.

Dans des travaux précédents, nous avons tout d'abord introduit ces hiérarchies dites contextuelles (Pitarch et al., 2009), puis nous avons proposé une solution de mise en œuvre (Pitarch et al., 2010a,b) en évoquant l'aspect implémentation et en nous focalisant sur comment exprimer les connaissances experts et les exploiter pour permettre des analyses flexibles.

Dans ce travail, nous nous plaçons davantage dans la phase de modélisation qui nécessite une prise en compte des utilisateurs, en partant de l'hypothèse d'un besoin d'échange avec les experts sur le modèle. Cette phase de modélisation est cruciale. Nous distinguons trois types d'approches de modélisation : les approches orientées données (ascendantes), les approches orientées besoins (descendantes), et les approches hybrides combinant les deux types précédents. Les approches hybrides permettent de combiner une réalité des données avec la satisfaction des besoins utilisateurs, satisfaction permettant d'assurer l'utilisation du système. Pour cette prise en compte des besoins utilisateurs, un échange avec les futurs utilisateurs, ou du moins des personnes connaissant bien le domaine métier, apparaît nécessaire. Pour ce faire, le recours à un modèle graphique facilement lisible est alors crucial.

Rappelons qu'historiquement, les premiers modèles d'entrepôt de données (étoile, constellation proposés par Kimball (1996)) ont émergé et connu un vif succès au sein des entreprises. Il est assez naturel de faire l'hypothèse que la simplicité de lecture/d'interprétation de ces modèles graphiques a sans doute participé à leur succès. D'un point de vue support d'échange entre concepteur et utilisateur final (décideurs/experts du domaine), il paraît assez naturel de faire un parallèle entre le modèle entité/association dans le contexte des bases de données et les modèles en étoile/flocon/constellation dans le contexte des entrepôts de données.

Le modèle entité-association est qualifié consensuellement de modèle conceptuel ; cependant, dans le domaine des entrepôts, aucun consensus sur la modélisation n'a encore émergé. Nous ne cherchons pas ici à discuter du débat conceptuel/logique. Nous partons simplement du principe que les notions de tables de faits/tables de dimension dans la modélisation multidimensionnelle constitue la base de discussion entre concepteur et décideur. Il nous paraît alors nécessaire de pouvoir étendre les modèles « classiques » pour prendre en compte les hiérarchies contextuelles, tout en conservant le fait qu'ils soient compréhensibles et discutables par les décideurs.

Dans ce travail, au-delà d'étendre le principe de généralisation contextuelle aux attributs de dimension, il s'agit donc surtout de donner toute son importance à la représentation graphique en apportant comme contribution, une visualisation qui permettra de discuter des données et connaissances à mettre en œuvre dans l'entrepôt de données grâce à ce modèle graphique volontairement simple, à l'image des premiers modèles d'entrepôt, tout en permettant une puissance d'expressivité sur un aspect plutôt complexe dans les hiérarchies, à savoir les hiérarchies contextuelles. Nous baptisons ce modèle : « le modèle avec satellites ».

La suite de l'article est organisée de la façon suivante. Dans la section 2 nous présentons une étude de cas qui permet de motiver notre travail. Nous proposons ensuite dans la section 3 un état de l'art permettant de positionner les travaux connexes. Nous reprenons ensuite dans la section 4 le modèle formel qui permet de représenter les hiérarchies contextuelles de dimension et de mesure. Puis, nous présentons notre modèle avec satellites dans la section 5 et le discutons. Enfin, nous concluons et dressons les perspectives de ce travail dans la section 6.

2 Etude de cas

Pour illustrer la problématique et la solution apportée, nous considérons tout au long de ce papier le cas d'un hôpital qui souhaiterait mettre en œuvre un entrepôt de données enregistrant pour chaque patient de son service de réanimation sa tension ainsi que les médicaments prescrits au fil du temps. Ces valeurs seraient mesurées par des capteurs et alimenteraient directement l'entrepôt.

Ainsi, l'entrepôt permettrait d'observer deux faits : la tension et la posologie. Les dimensions considérées sont les suivantes. La dimension temps (partagée par les deux faits), la dimension médicament rattachée au fait posologie et la dimension patient (partagée également par les deux faits). Le schéma correspondant, en adoptant le formalisme graphique introduit par Golfarelli et al. (1998b) puis étendu par Ravat et al. (2007b), est présenté dans la figure 1. Nous nous attardons plus spécifiquement sur la dernière dimension. Chaque patient possède un identifiant unique et est décrit par son nom, son âge, son sexe, la ville où il habite et par un attribut fumeur qui indique si le patient fume ou non. L'âge du patient peut être considéré selon trois niveaux de détail différents : Age, SubCatAge et CatAge.

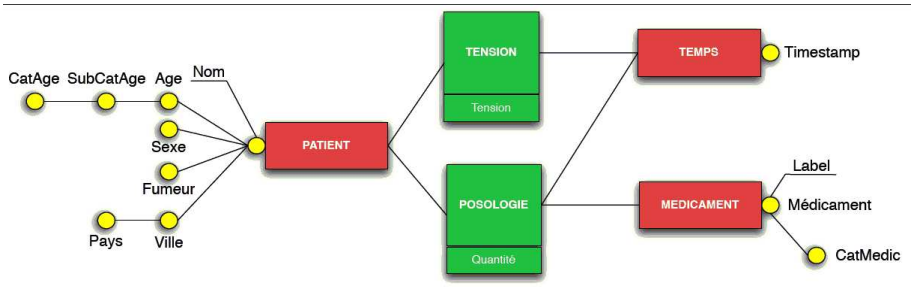


FIG. 1 – *Schéma classique de l'entrepôt de données pour l'analyse de la tension et de la posologie.*

Afin d'assurer un suivi efficace des patients du service, il est souhaitable de pouvoir formuler des requêtes telles que « Quels sont les patients dont la tension artérielle a été élevée pendant la nuit ? » ou « Quels sont les patients qui se sont vu prescrire une quantité trop importante de médicament X ? » ou encore « Quelle est la tension moyenne des enfants ? ».

Malheureusement, les modèles traditionnels d'entrepôts (celui de la figure 1 par exemple) ne permettent pas de répondre à de telles requêtes pour deux raisons. Premièrement, la notion de tension (resp. posologie) élevée peut être considérée comme une généralisation de la tension

mesurée (resp. de la quantité prescrite). Dans la mesure où les modèles classiques ne permettent pas d'établir une hiérarchie sur les mesures, ces requêtes ne peuvent être formulées.

De plus, même si l'on suppose que de telles requêtes sont formulables, la généralisation correcte de valeurs numériques est bien souvent contextuelle. Dans ce cas d'étude, nous considérons que les notions de tension élevée, de posologie élevée ou de patient jeune sont directement liées à certaines caractéristiques des patients et/ou des médicaments prescrits. Par exemple, un bébé ne doit pas recevoir la même quantité d'un médicament qu'un adulte. Ainsi, une même posologie pourra être considérée comme faible, normale ou élevée selon l'âge du patient considéré. Une connaissance experte est alors nécessaire pour (1) définir quels sont les attributs qui impactent sur la généralisation d'un attribut (attributs contextualisant) et (2) décrire cette généralisation en fonction des valeurs prises par ces attributs contextualisant.

Dans la suite de ce papier, nous nous focalisons sur la catégorisation d'une tension et celle de l'âge pour illustrer l'approche proposée. Le tableau 1 présente quelques exemples de connaissances expertes sur la catégorisation d'une tension en fonction des attributs SubCatAge et Fumeur d'un patient. Par exemple, une tension à 13 est normale chez un adulte fumeur mais est élevée chez un nourrisson.

SubCatAge	Fumeur	Tension	CatTension
Nourrisson	Oui ou Non	$>$12	Elevée
Adulte	Oui	$>$14	Elevée
$3^{ème}$ âge	Oui ou Non	$>$ 16	Elevée
Nourrisson	Oui ou Non	Entre 10 (inclus) et 12 (inclus)	Normale
Adulte	Oui	Entre 12 (inclus) et 14 (inclus)	Normale
...	...	...	...

TAB. 1 – *Exemple de règles expertes décrivant la catégorie d'une tension (CatTension) en fonction de la tension mesurée, de la classe d'âge d'un patient (SubCatAge) et de l'attribut Fumeur.*

La généralisation d'un âge (qui est un attribut de dimension et non une mesure) au niveau SubCatAge peut, elle aussi, être contextuelle. Nous supposons dans ce cas d'étude qu'elle dépend à la fois de l'âge mais aussi de la valeur associée à l'attribut Pays. Ceci permet de signifier que l'espérance de vie n'est pas la même dans les différents pays et que de ce fait la catégorie d'âge d'une personne peut varier en fonction de cette caractéristique. Le tableau 2 présente d'ailleurs un extrait des règles permettant une généralisation correcte. Dès lors, généraliser correctement les tensions mesurées implique au préalable d'avoir correctement généralisé l'âge du patient au niveau SubCatAge.

Avant de pouvoir exprimer les connaissances elles-mêmes, il s'agit de pouvoir définir le modèle de l'entrepôt et donc de disposer d'un modèle graphique à discuter aisément avec les experts (donc facilement interprétable), modèle qui mette en exergue les connaissances à représenter. Après avoir illustré notre problématique de modélisation à travers une étude de cas, nous revenons sur les travaux connexes à notre proposition.

Age	Pays	SubCatAge
Entre 20 et 70 (inclus)	France	Adulte
Entre 20 et 30 (inclus)	Swaziland	Adulte
...	...	...

TAB. 2 – *Exemple de règles expertes décrivant la généralisation au niveau SubCatAge en fonction de l'âge du patient et de son pays.*

3 Etat de l'art

3.1 Conception du schéma de l'entrepôt

Récemment, une étude présentant les méthodologies de modélisation multidimensionnelle a été proposée par Romero et Abelló (2009). Au-delà d'une description des travaux, une étude comparative est menée selon différents critères. Parmi ces critères figure ce que les auteurs ont appelé « paradigme». Du point de vue de la conception du schéma de l'entrepôt, nous distinguons dans la littérature trois grandes approches par rapport à ce paradigme : celle guidée par les données, qualifiée également d'ascendante ; celle guidée par les besoins d'analyse, dénommée également descendante et l'approche mixte qui combine les deux précédentes (Soussi et al., 2005).

L'approche orientée données ignore les besoins d'analyse a priori. Elle concerne en particulier les travaux sur l'automatisation de la conception de schéma. En effet, cette approche consiste à construire le schéma de l'entrepôt à partir de ceux des sources de données et suppose que le schéma qui sera construit pourra répondre à tous les besoins d'analyse. Par exemple, Golfarelli et al. (1998a) proposent une méthodologie semi-automatique pour construire un schéma d'entrepôt de données à partir des schémas entité-relation qui représentent les bases de données sources. Peralta et al. (2003) représentent les connaissances sur la construction de l'entrepôt de données sous forme de règles. Un algorithme gère alors l'ordre d'exécution des règles qui permettent une succession de transformations sur le schéma source pour obtenir le schéma logique final de l'entrepôt.

Les approches orientées besoins d'analyse, quant à elles, proposent de définir le schéma de l'entrepôt en fonction des besoins d'analyse et supposent que les données disponibles permettront la mise en œuvre d'un tel schéma, ou tout du moins que la confrontation avec les données réelles se fera dans un second temps (Prat et al., 2006). Parmi les approches orientées besoins d'analyse, List et al. (2002) proposent une distinction entre les approches guidées par les buts et les approches guidées par les utilisateurs.

L'approche orientée buts suppose que le schéma de l'entrepôt est défini selon les objectifs d'analyse de l'entreprise (Boehnlein et vom Ende, 2000). Ainsi, on suppose que tous les employés de l'entreprise ont des besoins d'analyses similaires vis-à-vis de l'exploitation de l'entrepôt de données. Autrement dit, tous les employés ont la même vision analytique de l'entrepôt de données.

Dans l'approche orientée utilisateurs, ces derniers sont interrogés afin de collecter l'ensemble de leurs besoins d'analyse avant de construire l'entrepôt de données, ce qui permet de garantir l'acceptation du système par les utilisateurs. Cependant la durée de vie de ce schéma peut être courte, étant donné que le schéma dépend beaucoup des besoins des personnes impli-

quées dans le processus de développement de l'entrepôt. Pour y remédier, Poe (1996), propose une méthodologie pour conduire les entretiens avec les utilisateurs pour la collecte des besoins. Il est alors recommandé d'interroger différents groupes d'utilisateurs pour avoir la vision la plus complète possible des besoins des utilisateurs. Mais reconnaissons qu'il est non seulement difficile de déterminer de façon exhaustive les besoins d'analyse pour l'ensemble des utilisateurs à un instant donné, mais qu'il est encore moins facile de déterminer leurs besoins à venir.

Dans ces deux approches, nous considérons tout simplement qu'il s'agit de besoins d'analyse qui sont exprimés, certains sont plus globaux, au sens ou l'ensemble des utilisateurs partage ce besoin, d'autres plus spécifiques. Enfin, l'approche mixte considère à la fois les besoins d'analyse et les données pour la construction du schéma. Cette approche est celle qui fait l'objet de plus d'investigations aujourd'hui. L'idée générale est de construire des schémas candidats à partir des données (démarche ascendante) et de les confronter aux schémas définis selon les besoins (démarche descendante) (Bonifati et al., 2001; Phipps et Davis, 2002; Soussi et al., 2005). Quant à Romero et Abelló (2010), ils proposent une méthode de dérivation des schémas multidimensionnels en fonction des besoins d'analyse, méthode incrémentale guidée par les examples. Ainsi, le schéma construit constitue une réponse aux besoins réels d'analyse et il est également possible de le mettre en œuvre avec les sources de données. Il apparaît donc important dans le cadre de cette démarche de pouvoir discuter des schémas compréhensibles avec les utilisateurs (experts du domaine).

Dans le travail que nous réalisons ici, nous nous sommes focalisés sur l'expression des besoins utilisateurs à travers un modèle compréhensible. Ainsi, il s'agit d'avantage d'une approche descendante. Nous ne nous focalisons pas ici sur la problématique de l'alimentation de l'entrepôt de données, mais seulement sur la phase de modélisation.

3.2 Modélisation des hiérarchies

Différents travaux se sont intéressés à la modélisation des hiérarchies de dimension. Malinowski et Zimányi (2004) ont proposé une classification des différents types de hiérarchies, en se basant sur des situations réelles. Ils proposent des notations graphiques basées sur les modèles Entités/Associations, en exploitant entre autres les cardinalités. Différents types de hiérarchies sont représentés : symétriques / asymétriques, strictes / non strictes, multiples, parallèles, etc.

Un autre travail a été développé par Ghozzi et al. (2003) autour des bases de données multidimensionnelles contraintes en exprimant différents types de contraintes sur les hiérarchies de dimension (contrainte inter-dimensions et intra-dimension) avec une proposition de représentation graphique au niveau des opérations sur les hiérarchies, permettant d'assurer la consistance des données de l'entrepôt.

Ces travaux mettent en avant l'intérêt d'une modélisation graphique pour une meilleure compréhension du modèle. Toutefois, la modélisation des hiérarchies contextuelles telles que nous en avons besoin pour représenter la réalité de nos données médicales n'est pas prise en charge par ces formalismes.

3.3 Discussion

Au travers des travaux présentés, nous avons tout d'abord mis en avant l'importance d'impliquer les utilisateurs dans la phase de conception du modèle. D'un point de vue méthode, il s'agit de pouvoir échanger sur un modèle facilement compréhensible qui représentent bien la réalité des données et la manière de les analyser. Nous constatons alors l'émergence de deux caractéristiques importantes qui peuvent apparaître contradictoires d'un premier abord, à savoir : l'expressivité d'un modèle (à quel point l'on peut exprimer au travers du modèle des situations complexes posées par la réalité des données) et la simplicité du modèle (pour permettre la discussion entre le concepteur qui maîtrise les méthodes de conceptions et les utilisateurs qui connaissent le domaine métier). Une présentation graphique du modèle permet sans aucun doute d'accéder à cette simplicité (Moody et Shanks, 1994).

Le travail de Malinowski est un travail très intéressant du point de vue de ces deux aspects : permettre une représentation graphique en augmentant l'expressivité du point de vue des hiérarchies de dimension. Mais le problème est à présent de pouvoir représenter nos hiérarchies contextuelles, qui ne se limitent pas d'ailleurs aux hiérarchies de dimension, mais concernent également les hiérarchies de mesure. Revenons tout d'abord sur la formalisation de ces hiérarchies dans la section suivante, avant de les représenter graphiquement par la suite.

4 Modèle formel avec hiérarchies contextuelles

Dans cette section, nous proposons la formalisation qui permet de représenter des hiérarchies contextuelles aussi bien au niveau des dimensions que des mesures, ce qui constitue une extension des travaux précédents qui se focalisaient sur la généralisation de mesure. Rappelons en préambule que la notion de contexte ne correspond ni au concept traditionnel dans les entrepôts de données de contexte d'analyse, ni au contexte au sens où l'on peut l'entendre lorsqu'un processus de personnalisation est mis en œuvre. Il s'agit bien ici d'exprimer le fait que pour une hiérarchie, au moins un des liens de généralisation entre deux niveaux ne peut être simplement déterminé grâce à un lien un à plusieurs prédéfini dans la mesure où d'autres informations (relevant d'autres dimensions par exemple) sont nécessaires.

Pour supporter le processus qui vise à la prise en compte de contextes par rapport à la détermination de la valeur de certains attributs généralisant les mesures ou des attributs de dimension, il est alors crucial de disposer d'un modèle d'entrepôt qui retrace cette contextualisation, par conséquent un modèle plus flexible.

Dans cette formalisation, nous nous basons sur un modèle en constellation que nous étendons par le concept de contexte.

Définition 1 (Dimensions) *Soit* $\{D_i, i = 1..s\}$ *l'ensemble des* s *dimensions.*

Notons Id_i *l'identifiant la dimension* D_i*.*

Soit $\{AF_{iu}, i = 1..s, u = 1..u_i\}$ *l'ensemble des* u_i *attributs dits « faibles » de la dimension* D_i*.*

Soit $\{A_{ihl}, i = 1..s, h = 1..h_i, l = 1..l_i\}$ *l'ensemble des attributs dits « paramètres » de la dimension* D_i*,* h *dénotant la hiérarchie de* D_i *et* l *le niveau dans la hiérarchie en question.*

Exemple 1 *Dans notre étude,*

$s = 3$ *:* $D_1 \equiv PATIENT$*,* $D_2 \equiv TEMPS$*,* $D_3 \equiv MEDICAMENT$*.*

$Id_1 \equiv IdPatient$, $Id_2 \equiv Timestamp$, $Id_3 \equiv IdMedicament$
$AF_{11} \equiv Nom$, $AF_{31} \equiv Label$, *etc.*
$A_{111} \equiv Age$, $A_{112} \equiv SubCatAge$, *etc.*

Définition 2 (Faits) *Soit $\{F_j, j = 1..t\}$ l'ensemble des t faits.*

Chaque fait est déterminé structurellement par un ensemble de dimensions et de mesures. De façon parallèle, on peut retrouver la notion de mesure faible au même titre que les attributs de dimension faibles.

Soit $\{MF_{jv}, j = 1..t, v = 1..v_j\}$ l'ensemble des v_j mesures dites « faibles » du fait F_j.

Soit $\{M_{jpn}, j = 1..t, p = 1..p_j, n = 1..n_j\}$ l'ensemble des mesures (qui vont êtres hiérarchisées) du fait F_j, p dénotant la hiérarchie de F_j et n le niveau dans la hiérarchie en question.

Ainsi, $F_j = (\mathcal{D}_|, \mathcal{M}_|)$ avec $\mathcal{D} = \{Id_i, i = 1..s_j, s_j <= s\}$ les identifiants des s_j dimensions décrivant le fait F_j et $\mathcal{M} = \{M_{jpn}, j = 1..t, p = 1..p_j, n = 1..n_j\} \cup \{MF_{jv}, j = 1..t, v = 1..v_j\}$.

Exemple 2 *Pour l'étude de la tension, on a :*
$Id_1 \equiv IdPatient$, $Id_2 \equiv Timestamp$
$M_{111} \equiv Tension$, $M_{112} \equiv CatTension$, $M_{113} \equiv NormaliteTension$
$F_1 = (\{IdPatient, Timestamp\}, \{Tension, CatTension, NormaliteTension\})$

Définition 3 (Attributs contextualisés et contextualisant) *Un attribut est dit **contextualisé** si sa valeur dépend des valeurs prises par un ensemble d'autres attributs de l'entrepôt (qualifiés alors de **contextualisant**). Cette définition est valable aussi bien pour les attributs de dimension que pour les mesures.*

Exemple 3 *Dans l'étude de la tension, l'attribut mesure $CatTension$ est un attribut contextualisé puisque sa valeur dépend des attributs contextualisant $Fumeur$, $SubCatAge$ et $Tension$. L'attribut de dimension $SubCatAge$ est un attribut contextualisé puisque sa valeur dépend des attributs contextualisant Age et $Pays$.*

Notons qu'un attribut donné peut être tour à tour contextualisant et contextualisé. C'est le cas par exemple de l'attribut $SubCatAge$ qui est contextualisé et contextualisant selon les cas, ou plus précisément selon les contextes, contextes dont nous donnons à présent la définition.

Définition 4 (Contexte : structure) *Soit $\mathcal{C}$ l'ensemble des contextes de l'entrepôt de données, à la fois pour les mesures et les dimensions.*

$\mathcal{C}^D_{ihl} \in \mathcal{C}$, où $\mathcal{C}^D_{ihl}$ dénote la structure du contexte relative à la définition de l'attribut de dimension A_{ihl}, qui se trouve donc être un attribut contextualisé, $A_{ihl}/i = 1..s, h = 1..h_i, l = 1..l_i$ correspondant à l'attribut de la dimension D_i, h dénotant la hiérarchie de D_i et l le niveau dans la hiérarchie en question.

$\mathcal{C}^F_{jpn} \in \mathcal{C}$, où $\mathcal{C}^F_{jpn}$ dénote la structure du contexte relative à la définition de l'attribut de mesure M_{jpn}, qui se trouve donc être un attribut contextualisé, $M_{jpn}/j = 1..t, p = 1..p_j, n = 1..n_j$ correspondant à la mesure du fait F_j, p dénotant la hiérarchie de F_j et n le niveau dans la hiérarchie en question.

La structure du contexte $\mathcal{C}^D_{ihl}$ est notée de la façon suivante :

$(A_{ihl}, \{C^D_{ihl\alpha}, \alpha > 1\})$ *où* A_{ihl} *est l'attribut contextualisé et* $\{C^D_{ihl\alpha}, \alpha > 1\}$ *est l'ensemble des attributs contextualisant pour* A_{ihl}.

De façon similaire, la structure du contexte $\mathcal{C}^F_{jpn}$ *est notée de la façon suivante :*

$(M_{jpn}, \{C^F_{jpn\beta}, \beta > 1\})$ *où* M_{jpn} *est l'attribut contextualisé et* $\{C^F_{jpn\beta}, \beta > 1\}$ *est l'ensemble des attributs contextualisant pour* M_{jpn}.

Note : α *et* β *sont strictement supérieurs à 1, car si ce n'était pas le cas, on serait dans le cas classique des hiérarchies.*

Exemple 4 $\mathcal{C}^D_{112} = (SubCatAge, \{Age, Pays\})$
$\mathcal{C}^F_{112} = (CatTension, \{Fumeur, SubCatAge, Tension\})$

Définition 5 (Contexte : instances) *Chaque structure de contexte est ensuite instanciée.*

Notons $c^D_{112}\omega$ *(resp.* $c^F_{jpn}\psi$*) l'instance* ω *de la structure du contexte* $\mathcal{C}^D_{ihl}$ *(resp.* $\mathcal{C}^F_{jpn}$*). Elle est définie par l'instanciation de chacun des attributs.*

Etant donné que dans ce papier nous nous focalisons sur l'aspect modélisation graphique, nous ne développons pas davantage l'aspect instanciation puisque c'est la structure qui est importante dans la modélisation graphique. Nous renvoyons le lecteur à nos travaux précédents sur l'expression des connaissances elles-mêmes (Pitarch et al., 2010b) qui détaillent également comment mettre en œuvre ce modèle dans un contexte relationnel grâce à l'implémentation de tables pour stocker les contextes et leurs instances. Nous nous focalisons ici sur la représentation graphique que nous présentons à présent.

5 Modèle graphique

Nous rappelons ici que notre objectif est de fournir une représentation graphique d'un modèle qui permettrait de représenter les hiérarchies contextuelles, modèle utile lors de la phase de conception pour permettre des échanges fructueux entre concepteur(s) et décideurs/utilisateurs. En effet, la formalisation, que nous venons de développer, permet certes de représenter ces hiérarchies contextuelles mais demeure assez lourde en terme de notation et n'est donc pas forcément facilement interprétable.

5.1 Modèle avec satellites

Pour introduire la contextualisation dans les généralisations d'attributs, nous introduisons le concept de satellite.

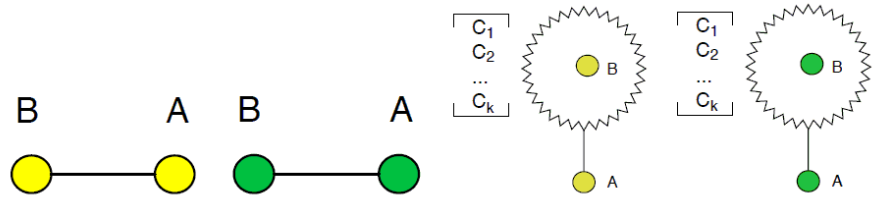
Définition 6 (Représentation des chemins de généralisation) *Soit* $G = A \xrightarrow{\mathcal{C}} B$ *un chemin de généralisation entre deux attributs.*

La représentation graphique associée à G sera différente en fonction de sa nature :

- *Si G est un chemin classique entre attributs de dimension, le formalisme graphique associé est celui présenté dans la figure 2(a). Les attributs sont représentés par des petits cercles jaunes reliés entre eux par un trait plein) ;*
- *Si G est un chemin classique entre mesures, le formalisme graphique associé est celui présenté dans la figure 2(b). Les mesures sont représentées par des petits cercles verts reliés entre eux par un trait plein) ;*

– *Si G est un chemin contextuel entre attributs de dimension, le formalisme graphique associé est celui présenté dans la figure 2(c). L'attribut de dimension A est représenté par un cercle jaune et l'attribut contextualisé B par un satellite entourant un cercle jaune. La liste des éléments contextualisant (i.e., les « membres » de $\mathcal{C}$) est adossée au satellite et le lien entre A et B est représenté par un trait plein ;*
– *Si G est un chemin contextuel entre mesures, le formalisme graphique associé est celui présenté dans la figure 2(d). La mesure A est représentée par un cercle vert et la mesure contextualisée B par un satellite entourant un cercle vert. La liste des éléments contextualisant (i.e., les « membres » de $\mathcal{C}$) est adossée au satellite et le lien entre A et B est représenté par un trait plein.*

Note : Rappelons ici la correspondance avec les notations employées dans la formalisation précédente (les « membres » de $\mathcal{C}$ n'étant autres que les attributs contextualisant des contextes) :
– A_{ihl}, lorsqu'il est contextualisé, est représenté par un cercle jaune et un satellite l'entourant, avec entre crochets les différents éléments $\{C^D_{ihl\alpha}, \alpha > 1\}$;
– M_{jpn}, lorsqu'elle est contextualisée, est représentée par un cercle vert et un satellite l'entourant, avec entre crochets les différents éléments $\{C^F_{jpn\beta}, \beta > 1\}$.



(a) Chemin classique entre attributs (b) Chemin classique entre mesures (c) Chemin contextuel entre attributs (d) Chemin contextuel entre mesures

FIG. 2 – *Représentation graphique d'un chemin de généralisation.*

Ainsi, en appliquant cette représentation et en adoptant la formalisation définie précédemment, on obtient dans la figure 3 la modélisation graphique d'une dimension.

De même, nous obtenons la modélisation graphique d'un fait dans la figure 4

Si nous appliquons cette représentation graphique à notre cas d'étude, nous obtenons le modèle de la figure 5.

Nous pouvons relever la présence de trois hiérarchies contextuelles. Une d'entre elles, (IdPatient ; Age ; SubCatAge ; CatAge ; ALLPatient), est une hiérarchie contextuelle dimension alors que les deux autres, i.e., (Tension ; CatTension ; NormaliteTension) et (Quantite ; CatQuantite ; NormaliteQuantite), sont des hiérarchies contextuelles de mesures. Un tel formalisme graphique rend immédiate la compréhension de l'entrepôt modélisé.

5.2 Discussion

Le modèle graphique obtenu est facilement discutable avec des décideurs. En effet, le formalisme reste très proche des modèles classiques, en présentant une extension aisément interprétable grâce à la représentation du concept de satellite, qui demeure dans la métaphore

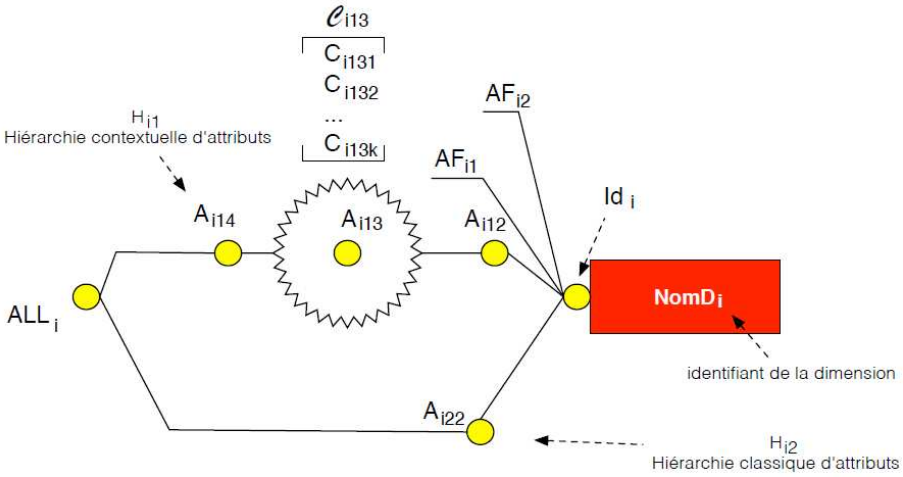


FIG. 3 – *Représentation graphique d'une dimension Di.*

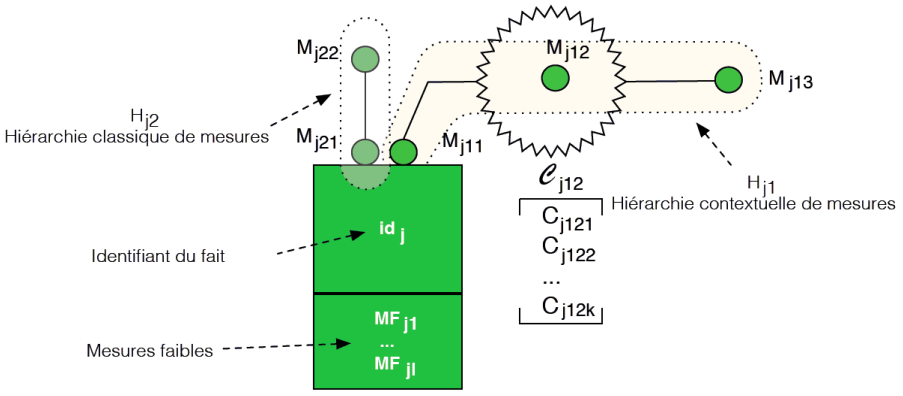


FIG. 4 – *Représentation graphique d'un fait Fj.*

Modèle avec Satellites

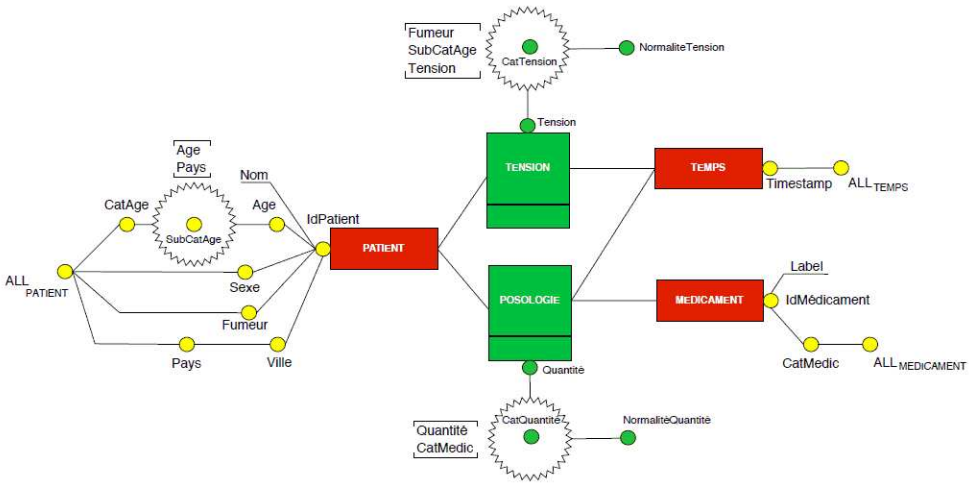


FIG. 5 – *Représentation graphique de l'entrepôt de données médicales.*

classique céleste (étoile, flocon, constellation et galaxie de Ravat et al. (2007a) pour le plus récent). Ainsi, l'objectif de favoriser les interactions entre les concepteurs et les experts peut être atteint (Moody et Shanks, 1994).

Bien évidemment, ce modèle graphique se situe au niveau de la représentation structurelle. Cela permet de structurer les connaissances à exprimer et de les confronter avec tous les acteurs de la modélisation. Pour l'explicitation des connaissances, il s'agit de s'intéresser ensuite aux instances, aspects que nous avons évoqués par ailleurs dans des travaux précédents (Pitarch et al., 2010a). En effet, la modélisation et le stockage des connaissances assurés, il s'agit par la suite de se focaliser sur le processus d'exploitation des données, et entre autres le processus d'agrégation (opération roll-up par exemple), prenant en compte les hiérarchies contextuelles.

6 Conclusion

Dans cet article, nous avons dans un premier temps étendu nos précédents travaux de hiérarchies contextuelles pour la généralisation d'attributs de dimension. Dans un second temps, nous avons proposé un modèle graphique d'entrepôt de données à l'instar des modèles en étoile/flocon de neige/constellation, en enrichissant ces modèles dans leur possibilité d'expression grâce au concept de satellite. Ce nouveau concept a pour but de mettre en œuvre visuellement la proposition de formalisation de hiérarchies contextuelles proposés dans nos précédents travaux. Cette notion de hiérarchie contextuelle répond à un besoin issu de la réalité des données et il nous paraissait important de pouvoir concrétiser cela sous la forme d'une représentation graphique pour une meilleure discussion avec les experts du domaine (en l'occurrence, le domaine médical pour ce travail).

Ce travail ouvre de nombreuses perspectives parmi lesquelles nous pouvons citer un travail important sur les opérateurs pour la navigation. En effet, il s'agit à présent de pouvoir étudier

l'impact de cette modélisation sur les processus de navigation au sein des données en proposant des solutions adéquates en terme à la fois d'utilisabilité et de performances. Une deuxième perspective réside dans l'intégration de cette modélisation dans un outil d'aide à la conception pour permettre un retour sur l'utilisabilité, la compréhensibilité, l'expressivité, etc. du modèle. Ceci implique également de s'intéresser à la validation automatique du modèle à travers la vérification des liens de contextualisation (détection de conflit entre autres).

Références

Agrawal, R., A. Gupta, et S. Sarawagi (1997). Modeling multidimensional databases. In *13th International Conference on Data Engineering (ICDE'97)*, pp. 232–243.

Boehnlein, M. et A. U. vom Ende (2000). Business Process Oriented Development of Data Warehouse Structures. In *Data Warehousing 2000 - Methoden Anwendungen, Friedrichshafen, Germany*. Physica Verlag.

Bonifati, A., F. Cattaneo, S. Ceri, A. Fuggetta, et S. Paraboschi (2001). Designing Data Marts for Data Warehouses. *ACM Transactions on Software Engineering and Methodology 10*(4), 452–483.

Chen, M., J. Han, et P. S. Yu (1996). Data mining : An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering : TKDE 8*(6), 866–883.

Einbinder, J. S., K. W. Scully, R. D. Pates, J. R. Schubart, et R. E. Reynolds (2001). Case study : a data warehouse for an academic medical center. *Journal of Healthcare Information Management : JHIM 15*(2), 165–175.

Ghozzi, F., F. Ravat, O. Teste, et G. Zurfluh (2003). Constraints and Multidimensional Databases. In *Vth International Conference on Enterprise Information Systems (ICEIS'03), Angers, France*, Volume 1, pp. 104–111.

Golfarelli, M., D. Maio, et S. Rizzi (1998a). Conceptual Design of Data Warehouses from E/R Schemes. In *XXXIst Annual Hawaii International Conference on System Sciences (HICSS'98), Big Island, Hawaii, USA*, Volume 7, pp. 334–343.

Golfarelli, M., D. Maio, et S. Rizzi (1998b). The Dimensional Fact Model : A Conceptual Model for Data Warehouses. *International Journal of Cooperative Information Systems 7*(2-3), 215–247.

Han, J. (1997). OLAP mining : An integration of OLAP with data mining. *7th IFIP 2.6 Working Conference On Database Semantics (DS-7)*, 1—9.

Inmon, W. H. (1996). *Building the Data Warehouse, 2nd Edition* (2 ed.). Wiley.

Kimball, R. (1996). *The Data Warehouse Toolkit*. John Wiley & Sons.

List, B., R. Bruckner, K. Machaczek, et J. Schiefer (2002). A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse. In *XIIIth International Conference on Database and Expert Systems Applications (DEXA'02), Aix-en-Provence, France*, Volume 2453 of *LNCS*, pp. 203–215. Springer.

Malinowski, E. et E. Zimányi (2004). OLAP Hierarchies : A Conceptual Perspective. In *XVIth International Conference on Advanced Information Systems Engineering (CAiSE'04), Riga, Latvia*, Volume 3084 of *LNCS*, pp. 477–491. Springer.

Mallach, E. G. (2000). *Decision Support and Data Warehouse Systems*. McGraw-Hill Higher Education.

Moody, D. L. et G. G. Shanks (1994). What Makes a Good Data Model ? Evaluating the Quality of Entity Relationship Models. In *13th International Conference on the Entity-Relationship Approach (ER'94), Manchester, U.K.*, Volume 881 of *LNCS*, pp. 94–111. Springer.

Peralta, V., A. Illarze, et R. Ruggia (2003). On the Applicability of Rules to Automate Data Warehouse Logical Design. In *Ist International Workshop on Decision Systems Engineering (DSE'03), in conjunction with the XVth International Conference on Advanced Information Systems Engineering (CAiSE'03), Klagenfurt/Velden, Austria*, Volume 75 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Phipps, C. et K. C. Davis (2002). Automating Data Warehouse Conceptual Schema Design and Evaluation. In *IVth International Workshop on Design and Management of Data Warehouses (DMDW'02), Toronto, Canada*, Volume 58 of *CEUR Workshop Proceedings*, pp. 23–32. CEUR-WS.org.

Pitarch, Y., C. Favre, A. Laurent, et P. Poncelet (2010a). Analyse flexible dans les entrepôts de données : quand les contextes s'en mêlent. In *6èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA'10), Djerba, Tunisie*, Volume B-6 of *RNTI*, Toulouse, pp. 191–205. Cépaduès.

Pitarch, Y., C. Favre, A. Laurent, et P. Poncelet (2010b). Context-aware generalization for cube measures. In *ACM 13th International Workshop on Data Warehousing and OLAP (DOLAP'10), Toronto, Ontario, Canada*, pp. 99–104.

Pitarch, Y., A. Laurent, et P. Poncelet (2009). A conceptual model for handling personalized hierarchies in multidimensional databases. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, France, pp. 107–111. ACM.

Poe, V. (1996). *Building a Data Warehouse for Decision Support*. Prentice Hall.

Prat, N., J. Akoka, et I. Comyn-Wattiau (2006). A uml-based data warehouse design method. *Decision Support System 42*, 1449–1473.

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2007a). A Conceptual Model for Multidimensional Analysis of Documents. In *International Conference on Conceptual Modeling (ER'07), Auckland, New Zealand*, Number 4801 in LNCS, pp. 550–565. Springer.

Ravat, F., O. Teste, R. Tournier, et G. Zurfluh (2007b). Graphical querying of multidimensional databases. In *11th East European Conference on Advances in Databases and Information Systems (ADBIS'07), Varna, Bulgaria*, Volume 4690 of *LNCS*, pp. 298–313. Springer.

Romero, O. et A. Abelló (2009). A Survey of Multidimensional Modeling Methodologies. *International Journal of Data Warehousing and Mining : IJDWM 5*(2), 1–23.

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

Soussi, A., J. Feki, et F. Gargouri (2005). Approche semi-automatisée de conception de schémas multidimensionnels valides. In *Ière journée sur les Entrepôts de Données et l'Analyse en ligne (EDA'05), Lyon*, Volume B-1 of *Revue des Nouvelles Technologies de l'Information*, pp. 71–90. Cépaduès Editions.

Summary

Classical data warehouse models (star schema, etc.) have emerged and have been very successful in business because of their graphic presentation easy to read. This is necessary in a context where multi-dimensional modeling should be confronted with the decision makers. In previous work, we have highlighted a certain lack of expressiveness of these models. For example, in the case of a medical data warehouse, it was not possible to model the fact that the blood pressure of a patient is "low", "normal" or "high" (measure hierarchy) depends on his/her age and if he/she smokes or not. We developped a formalization of hierarchies called "contextual" to solve this problem. Then, we propose in this paper a graphical model, to facilitate the readability of the model to decision makers, we call this model "schema with satellites".