Approche semi-automatisée de conception de schémas multidimensionnels valides

Ahlem Soussi*, Jamel Feki**, Faïez Gargouri***

*Centre de Calcul El Khawarizmi,
Campus Universitaire de la Manouba, 2010 ; Manouba, Tunisie
ahlem.soussi@cck.rnu.tn
**Faculté des Sciences Economiques et de Gestion de Sfax
B.P. 1088 ; Sfax, Tunisie
jamel.feki@fsegs.rnu.tn
***Institut Supérieur d'Informatique et du Multimédia de Sfax
Route Mharza km 1,5 ; Sfax, Tunisie
faiez.gargouri@fsegs.rnu.tn

Résumé. Cet article s'inscrit dans le cadre d'une approche semi-automatisée de conception de schémas multidimensionnels. Cette approche accepte des besoins OLAP, exprimés indépendamment de toute source OLTP, sous forme de tableaux n-dimensionnels, puis construit des schémas en étoile *idéaux*. En vue de valider ces étoiles idéales par rapport à une source d'alimentation, nous proposons d'une part, une méthode systématique pour effectuer la correspondance de chaque concept multidimensionnel avec la source et d'autre part, des règles de raffinement des étoiles. Les schémas en étoile valides seront ensuite fusionnés pour générer des schémas en constellation.

1 Introduction

Les entreprises passent à l'ère de l'information : leur défi est de compléter leur système d'information transactionnel OLTP («*On-Line Transaction Processing*»), à vocation de production, par un système d'information décisionnel (SD), à vocation de pilotage. Ceci est facilité avec l'apparition des entrepôts de données (ED) qui permettent désormais à tous les utilisateurs d'accéder à l'information stratégique. Un ED (ou *Data Warehouse*) est le lieu de stockage centralisé d'un extrait des bases de production. Il intègre et « *historise* » les données utiles pour la prise de décision. Son organisation doit faciliter la gestion efficace des données et la conservation des évolutions [Ravat *et al*., 2001]. Dans la majorité des SD, un ED est restructuré en un ensemble de magasins de données (MD) (ou *Data Marts*). Un MD est un extrait de l'entrepôt. Les données extraites sont adaptées à une classe de décideurs ou à un usage particulier ; il est modélisé de manière multidimensionnelle qui s'adapte bien aux analyses OLAP («*On Line Analytical Processing*») [Teste, 2000].

La phase de création d'un schéma d'ED ou de MD passe par les étapes de modélisation conceptuelle, logique et physique [Phipps et Davis, 2002]. Nous nous intéressons dans cet article à la modélisation conceptuelle des schémas de MD et à son automatisation. Dans notre approche, les besoins OLAP sont exprimés sous forme de tableaux n-dimensionnels. Nous partons de ces tableaux pour proposer une méthode systématique de génération et de validation de schémas multidimensionnels en *étoile* ou en *constellation*. Un schéma en étoile est constitué d'un fait central (sujet analysé) et de plusieurs dimensions (axes d'analyses) ; un

schéma en constellation est une généralisation du schéma en étoile [Kimball, 1996] : il regroupe plusieurs faits étudiés selon différentes dimensions éventuellement partagées. Nous commençons par générer des schémas en étoile par fusion des tableaux de besoins. Cette phase génère des schémas reflétant les besoins des utilisateurs finaux indépendamment de toute source. Afin d'aider le concepteur à valider ces schémas, nous étudions une méthode automatique pour effectuer leur correspondance avec les sources d'alimentation, et des règles pour les raffiner. Ensuite, les schémas en étoile valides seront éventuellement fusionnés pour générer des schémas en constellation.

Cet article est organisé comme suit : le paragraphe 2 parcourt et critique l'état de l'art sur la conception des MD, le paragraphe 3 décrit l'architecture fonctionnelle du système de conception de MD, le paragraphe 4 décrit la structure des besoins OLAP, et les paragraphes 5, 6 et 7 détaillent les différents modules de ce système.

2 Etat de l'art sur la conception des magasins de données

Les approches existantes de conception des MD peuvent être classées en trois types : (1) des approches guidées par les besoins des utilisateurs, (2) des approches guidées par la source de données et (3) des approches mixtes.

Les **approches de conception guidées par les besoins** partent des besoins décisionnels d'une activité spécifique pour construire les schémas des MD. Ce type d'approches a été adopté dans [Kimball, 1996] où l'auteur commence par l'identification des faits et des attributs dimensionnels pour proposer un ensemble de schémas en étoile. Certes cette méthode de conception produit une solution qui répond bien aux requêtes exprimées par les utilisateurs finaux, cependant, elle est décrite par des exemples au lieu de procédures explicites de conception, et peut mener à des conceptions incorrectes si le concepteur ne comprend pas correctement les relations existantes entre les données [Moody et Kortink, 2000].

Les **approches de conception guidées par la source** prennent le modèle de données de l'entreprise comme base pour le développement des schémas de MD. Le concepteur peut donc bénéficier des relations existantes entre les données et suivre une approche plus structurée pour concevoir la base de données décisionnelle de l'entreprise. Ce type d'approche a été adopté dans [Golfarelli *et al*, 1998], [Cabibbo et Torlone, 1998], [Moody et Kortink, 2000] et [Hüsemann *et al*, 2000]. Malgré l'importance de l'étape d'identification des faits à partir du modèle de données de l'entreprise, aucune de ces approches n'a présenté une méthode précise pour l'effectuer. Cette étape est soit décrite littérairement, soit traitée uniquement à travers des exemples simples. Ceci constitue une limite à ce type d'approche. Il est également difficile de motiver adéquatement les utilisateurs finaux pour participer à la localisation des faits dans la source décrite par un modèle de données auquel ils ne sont pas familiarisés.

Par conséquent, il s'avère très utile de combiner ces deux approches de conception. Des **approches de conception mixtes** ont été adoptées dans [Phipps et Davis, 2002] et [Bonifati *et al*, 2001]. Dans ces deux approches, la phase de conception guidée par la source génère automatiquement des schémas candidats. Ces schémas sont produits en considérant que chaque entité ou association n-aire, du modèle source, contenant au moins un attribut numérique correspond à un fait potentiel. Cette hypothèse produit un grand nombre de schémas candidats dont la majorité ne décrivent pas des faits à cause des entités contenant uniquement des attributs numériques clés (comme par exemple num_facture,

num_agence…) ou des attributs numériques ne pouvant pas être des indicateurs d'une activité de l'entreprise (comme par exemple num_tél, code_postal…).

Nous proposons une approche mixte de conception de MD travaillant à la fois sur les besoins OLAP et sur les sources OLTP et plaçant le décideur au centre du mécanisme de conception. En effet, nous commençons par la construction des schémas des MD à partir des besoins OLAP exprimés, sous une forme canonique (tableaux n-dimensionnels), indépendamment des sources, puis nous validons et raffinons ces schémas par rapport aux sources OLTP existantes. Cet ordre évite de gérer inutilement tous les schémas candidats. Notre approche se distingue en plus par la génération de schémas multidimensionnels en étoile et en constellation, alors que les autres approches ne produisent que des schémas *individuels* généralement en étoile.

3 Architecture fonctionnelle de notre système de conception de MD

L'architecture fonctionnelle (figure 1) du système de conception semi-automatisé de schémas de MD à partir de besoins OLAP tabulaires montre trois modules :

- Génération des schémas en étoile idéaux par fusion des besoins OLAP ;
- Correspondance entre les schémas en étoiles idéaux et les sources de données, et validation des schémas en étoile associés avec les schémas sources ;
- Génération des schémas en constellation par fusion des schémas en étoile valides.

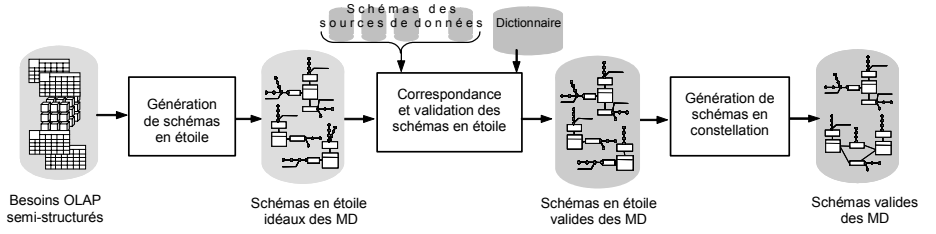

FIG. 1 - *Architecture fonctionnelle du système de conception semi-automatisé de schémas multidimensionnels valides de MD.*

Dans les paragraphes suivants, nous allons décrire l'entrée puis les différents modules de ce système.

4 Les besoins OLAP

Les besoins en analyses décisionnelles de type OLAP peuvent être formulés de différentes manières, généralement sous forme de phrases rédigées en langage naturel décrivant des requêtes types. Dans notre démarche visant une conception automatisée, nous proposons de recueillir ces besoins sous une forme familière aux décideurs, c'est à dire en tant que tableaux structurés [Feki, 2004]. La figure 2 montre la structure générique d'un tableau n-dimensionnel de besoin. Un tableau de besoin définit le fait à analyser, son

domaine, ses mesures, ses dimensions d'analyse, et les paramètres de la hiérarchie choisie pour chaque dimension ; en plus, chaque paramètre (attribut fort) est décrit par un ensemble d'attributs faibles. La figure 4.a montre deux exemples de tableaux analysant le fait «encadrement» du domaine «répartition des enseignements».

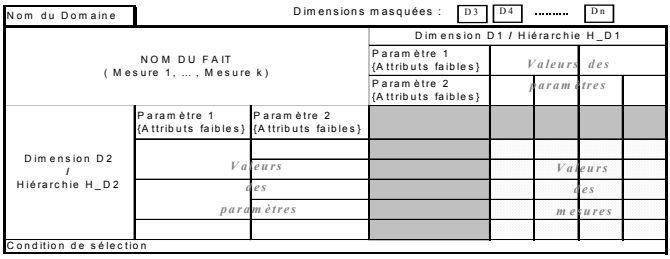

FIG. 2 – *Structure générique du tableau n-dimensionnel de besoin.*

Les besoins décisionnels sont donc spécifiés par un ensemble de tableaux n-dimensionnels définissant chacun un fait et n dimensions d'analyses. Pour analyser un fait selon *n* (n>2) dimensions, nous pouvons utiliser un même tableau et masquer l'une de ses dimensions pour lui ajouter une nouvelle, ou bien utiliser plusieurs tableaux bidimensionnels analysant le même fait. La forme tabulaire est préconisée parce qu'elle est habituellement adoptée par et pour les utilisateurs. En effet, les tableaux constituent pour les décideurs une forme de présentation familière et intuitive.

L'acquisition des besoins OLAP est réalisée, par les décideurs, moyennant une interface graphique et utilise une base de données de spécification des besoins OLAP relatifs à l'application étudiée. Cette base de données est construite suite à une phase d'extraction des concepts multidimensionnels (faits, mesures, dimensions, hiérarchies, paramètres, attributs faibles) à partir de sources de données types pour l'application étudiée. Elle assiste l'utilisateur dans la tâche d'acquisition des besoins et l'aide à lever certaines ambiguïtés telles que de synonymie des noms de faits, dimensions, etc. Le décideur n'a donc qu'à décrire les tableaux de besoins qu'il veut exprimer à partir des différentes listes de choix (de faits, mesures, dimensions, hiérarchies, paramètres, attributs faibles) alimentées par la base de données de spécification des besoins.

Les besoins spécifiés -dits besoins OLAP structurés- sont formulés indépendamment de toute source ; ils sont stockés dans le référentiel décrit par le diagramme de classes UML de la figure 3, représentant le point d'entrée du module de génération des schémas en étoile idéaux.

5 Génération de schémas en étoile

Un tableau de besoin peut être vu comme une description partielle (ou vue multidimensionnelle) d'un schéma en étoile. En conséquence, les schémas en étoile d'un MD sont dérivables à partir d'un ensemble de tableaux analysant le même fait d'un domaine.

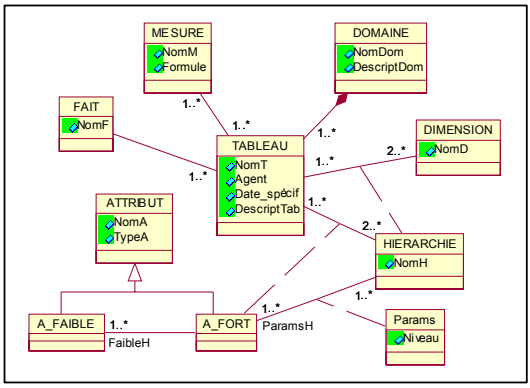

FIG. 3 – *Diagramme de classes UML (simplifié) du référentiel des besoins OLAP structurés.*

Ce module de génération est composé de deux étapes [Soussi *et al*, 2004] [Nabli *et al*, 2005] :

1. Enrichissement des besoins tabulaires spécifiés afin d'élargir l'étendue des requêtes analytiques initiales ;
2. Construction des schémas des MD à partir des besoins tabulaires enrichis.

5.1 Enrichissement des besoins spécifiés

Dans un tableau de besoin OLAP, un décideur peut limiter l'analyse d'un fait à quelques paramètres dans une ou plusieurs hiérarchies. Afin d'offrir une perspective d'analyse plus complète, nous substitutions chaque hiérarchie du tableau par sa hiérarchie maximale correspondante, décrite dans une base de données spécifique à l'application étudiée. Une hiérarchie d'une dimension *D* est dite maximale, si elle ne peut être contenue dans aucune autre hiérarchie possible des paramètres de *D*. Par conséquent, cette étape permet d'élargir les niveaux de granularités des requêtes permettant des forages plus fins (vers le bas) et plus agrégés (vers le haut) et homogénéise le niveau de granularité des mesures.

5.2 Construction des schémas en étoile

Un fait d'un domaine peut être analysé dans plusieurs tableaux de besoins. La construction des schémas en étoile se fait par fusion des tableaux n-dimensionnels $T_i^{F\text{-}dom}$, issus de l'étape précédente, analysant un même fait *F* d'un domaine *dom*. Un schéma en étoile *Sch* est créé à partir des tableaux $T_i^{F\text{-}dom}$. Le schéma *Sch* analysera le fait *F* décrit par l'union des mesures analysées dans les tableaux $T_i^{F\text{-}dom}$ par rapport à toutes les dimensions de ces tableaux. Les hiérarchies de chaque dimension *D* de *Sch* sont obtenues par l'union des hiérarchies maximales de *D* rencontrées dans les $T_i^{F\text{-}dom}$.

La construction des schémas en étoile est effectuée par un opérateur *ETOILE* basé sur l'ajout incrémental de tableaux de besoin appartenant à un même domaine et analysant le même fait.

Exemple
Considérons les deux tableaux T_1 et T_2 de la figure 4.a, analysant le fait « *encadrement* » du domaine « *répartition des enseignements* » par rapport à des hiérarchies maximales. La construction du schéma en étoile correspondant à ces deux tableaux se fait en deux étapes :

- transformer le tableau T_1 en un schéma en étoile *Sch* (figure 4.b),
- ajouter le tableau T_2 au schéma *Sch* (figure 4.c).

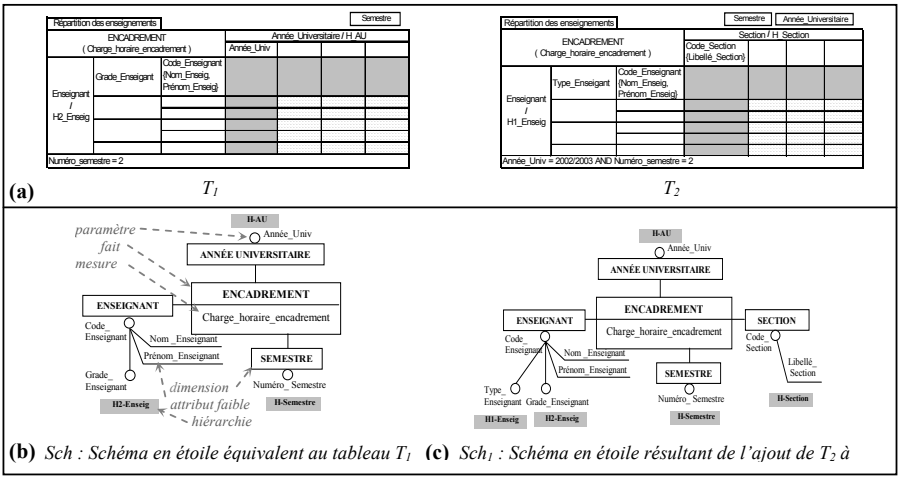

FIG. 4 – *Exemple de construction incrémentale d'un schéma en étoile.*

Remarque : Les schémas multidimensionnels des MD sont représentés en suivant la notation spécifique de [Golfarelli *et al*, 1998].

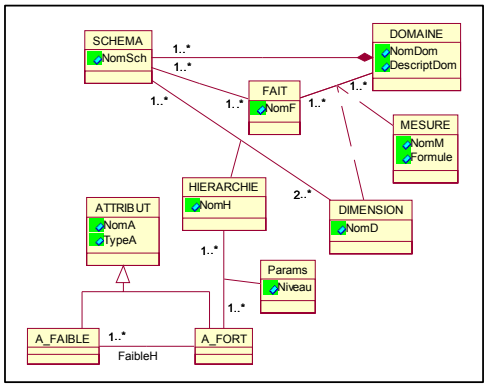

FIG. 5 – *Diagramme de classes UML (simplifié) décrivant les schémas multidimensionnels.*

6 Correspondance et validation des schémas en étoile

Le référentiel des tableaux de besoins (figure 3) est enrichi par la classe SCHEMA pour les schémas multidimensionnels en étoile/constellation, (figure 5). Cette classe est utile pour stocker les schémas en étoile, générés à partir des besoins, et les schémas en constellation, qui seront générés par fusion des étoiles valides.

Les schémas en étoile générés par le premier module reflètent les besoins des utilisateurs finaux indépendamment de toute source, ce sont donc des schémas *idéaux*. Afin d'aider le concepteur à les valider, nous proposons une méthode automatique pour effectuer leur correspondance avec de futures sources d'alimentation. Cette correspondance nécessite un dictionnaire spécifique à l'application étudié et dédié aux EDs. Elle traite les faits et leurs mesures ainsi que les dimensions et leurs hiérarchies. La correspondance d'un schéma en étoile peut être multivaluée, une validation s'avère donc nécessaire. Pour effectuer cette validation, nous proposons un ensemble de règles ; en particulier, nous définissons des métriques de comparaison des solutions de correspondance d'un même schéma en étoile et un algorithme pour le choix d'une solution. En sortie, nous obtenons des schémas en étoile valides associés à des éléments de la source.

Nous détaillons d'abord notre méthode de correspondance entre des schémas en étoile et les schémas des leurs futures sources d'alimentation, et nous l'illustrons par un exemple. Ensuite, nous présentons les étapes de validation des schémas en se basant sur les résultats de leur correspondance.

6.1 Correspondance entre magasin et source de données

La correspondance se fait entre les schémas en étoile des MD et le schéma de la source OLTP choisi pour l'alimentation ; elle identifie pour chaque concept d'un schéma en étoile son (ses) correspondant(s) dans la source.

Notre méthode de correspondance utilise des notions de base simples du schéma source : d'une part, les cardinalités des relations entre les entités ou les entités et les associations; et d'autre part, les types des attributs (numérique, date/heure ou textuel). Ces notions sont présentes dans la quasi-totalité des modèles OLTP, ce qui représente un point fort de la méthode qui reste fonctionnelle pour des sources comme UML, E/R ou relationnelles,… Nous présentons ici notre méthode de correspondance des schémas en étoile idéaux avec des sources E/R. Le choix de sources E/R est justifié par l'adoption de ce modèle par la majorité des entreprises durant les deux dernières décennies [Golfarelli *et al*., 1998]. Cette correspondance traite les faits et leurs mesures ainsi que les dimensions et leurs hiérarchies. Elle se fait, entre un schéma en étoile Sch_k et une source E/R, en quatre phases et selon l'ordre suivant :

1. Correspondance du fait du schéma en étoile Sch_k ;
2. Correspondance des mesures de Sch_k ;
3. Correspondance des dimensions de Sch_k : effectue, pour chaque dimension, la correspondance de son identifiant (paramètre de niveau 1), puis celle de ses attributs faibles ;
4. Correspondance des hiérarchies de Sch_k : effectue, pour chaque hiérarchie d'une dimension de Sch_k, la correspondance de ses paramètres de niveau strictement supérieur à 1 et de leurs attributs faibles.

Chacune des phases 2, 3 et 4 est systématiquement conditionnée par la validation de la phase qui la précède, sachant que :

- La correspondance d'un fait F_k est systématiquement validée, si et seulement si, F_k possède au moins un associé.
- La correspondance des mesures Mes^{Fk} d'un fait F_k est systématiquement validée, si et seulement si, au moins une des mesures de Mes^{Fk} possède un associé.
- La correspondance d'une dimension D_m de Sch_k est systématiquement validée, si et seulement si, le paramètre identifiant de D_m possède un associé.

Dans les quatre phases de correspondance ci-dessus, l'association d'un concept multidimensionnel c_i (i. e. un fait, une mesure, un paramètre identifiant d'une dimension, un paramètre d'une hiérarchie ou un attribut faible) est effectuée en deux étapes :

- ***Extraction des termes potentiels du concept c_i*** : cette étape extrait, à partir d'une source, l'ensemble de *termes potentiels* susceptibles d'être associés au concept multidimensionnel c_i. Cet ensemble est déterminé en se basant sur la cardinalité des liens entre les entités et les associations sources, et sur les types de leurs attributs. Selon le concept multidimensionnel c_i on a un ensemble de *Faits potentiels* ($Fait_{pot}$), de *mesures potentielles* (Mes_{pot}), d'*identifiants potentiels de dimensions* ($IdDim_{pot}$), etc. Les règles de construction de chaque ensemble dépendent du concept c_i, elles sont détaillées dans les sous paragraphes suivants.
- ***Mise en correspondance de c_i*** : établit la correspondance terminologique (à travers un dictionnaire) entre le concept c_i et les éléments de l'ensemble de ses termes potentiels. La correspondance terminologique associe deux dénominations synonymes ou une dénomination et son abréviation.

Notons que la correspondance d'une mesure, d'un paramètre ou d'un attribut faible ne peut s'établir qu'avec un attribut source au plus puisque nous supposons que la source est minimale. Tandis que pour un fait, la correspondance peut s'établir avec n ($n \geq 0$) éléments (entité ou association) de la source. Si au fait F_k d'un schéma en étoile Sch_k correspondent plusieurs faits potentiels sources $Fp_1 \ldots, Fp, \ldots, Fp_n$ (figure 6), nous continuons par établir la correspondance des autres concepts de Sch_i (mesures, dimensions, hiérarchies) avec les concepts potentiels sources accessibles à partir de chaque Fp_i. Le choix du fait le plus approprié nécessite des métriques de comparaison des différentes solutions (se référer au § 6.2.1).

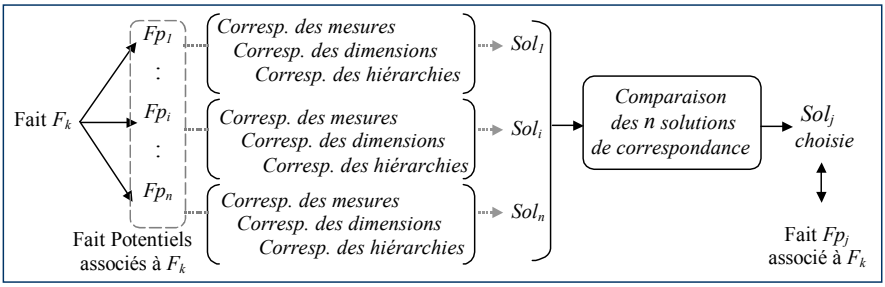

FIG. 6 – *Principe de la correspondance du fait F_k d'un schéma en étoile.*

6.1.1 Correspondance des faits

La correspondance d'un fait nécessite la détermination de l'*ensemble des faits potentiels*.

Définition 1. L'ensemble des ***faits potentiels*** $Fait_{pot}(S)$ d'une source de données de schéma *S*, est formé par tous les faits qu'on peut extraire à partir de *S*.

La recherche du(es) correspondant(s) d'un fait F_k d'un schéma en étoile dans *S*, se limite à une recherche dans l'ensemble $Fait_{pot}(S)$ déterminé une seule fois, d'où une optimisation de cette étape.

Extraction des faits potentiels : l'ensemble $Fait_{pot}(S)$ des faits potentiels qu'on peut extraire à partir d'une source *S* est formé par les entités et les associations de *S* vérifiant l'une des deux règles suivantes :

Rf1. Les associations n-aires contenant au moins un attribut numérique non-clé[1] [Kimball, 1997].

Rf2. Les entités contenant au moins un attribut numérique non-clé.

Notons que cette règle recouvre chacun des critères « *entités fréquemment mises à jour* » de [Golfarelli *et al.*, 1998]**,** « *entités de transaction* » de [Moody et Kortink, 2000] et améliore la règle d'extraction des faits potentiels dans [Bonifati *et al.,* 2001] en écartant les entités dont les seuls attributs numériques sont des clés.

Mise en correspondance des faits : cette correspondance compare, terminologiquement, le fait F_k d'un schéma en étoile avec les éléments de l'ensemble $Fait_{pot}(S)$ et détermine les éléments Fp_i de $Fait_{pot}(S)$ qu'on peut associer à F_k.

A fin d'illustrer les étapes de notre démarche, considérons la source *S1* modélisant la répartition des enseignements décrite par son schéma E/R de la figure 7, et le schéma en étoile idéal Sch_1 (figure 4.c) à alimenter à partir de *S1*.

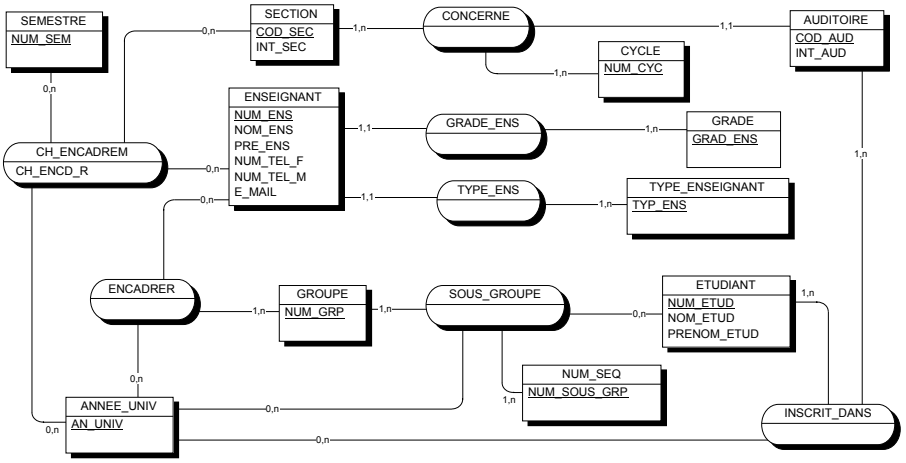

FIG. 7– *S1 : Schéma E/R (simplifié) source décrivant la répartition des enseignements.*

[1]. Primaire ou étrangère, partout dans le reste.

Exemple : Considérons la source *S1* pour alimenter le schéma en étoile Sch_1. La correspondance établie entre le fait « *encadrement* » de Sch_1 et *S1* est mono-valuée, elle est illustrée par la figure 8.

NOTE. — Dans la figure 8 (et dans 9, 10, 11, et 13), chaque élément extrait de la source est précédé par le nom de la règle qui l'a identifié.

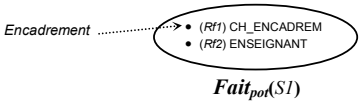

FIG. 8– *Faits potentiels extraits de la source S1 et correspondance du fait Encadrement.*

Dans la suite, les termes *concept*, *lien (1,n)* et *fermeture transitive* signifient :

- un *concept* désigne une entité ou une association porteuse de données ;
- un concept *c2* peut être lié à un concept *c1* par *un lien (1,n)* avec ($n \geq 1$), où 1 et *n* désignent respectivement les cardinalités maximales du coté de *c1* et de *c2*.
- La *fermeture transitive* d'un concept *c* est formé par tous les concepts qui lui sont directement ou transitivement liés par un lien (1,1) ou (1,n) avec n>1.

6.1.2 Correspondance des mesures

Rappelons que la correspondance des mesures est opérée pour chaque fait associé au fait du schéma en étoile. Elle nécessite la détermination de l'*ensemble des mesures potentielles*.

Définition 2. L'ensemble des ***mesures potentielles*** $Mes_{pot}(F)$ d'un fait *F* est formé par toutes les mesures, qu'on peut extraire à partir d'une source *S* et, pouvant être associées aux mesures de *F*.

Notation :

- F_k : un fait d'un schéma en étoile ;
- $F_{k\text{-}ass}$: un parmi les faits potentiels associés à F_k dans l'étape précédente (§6.1.1);
- $Mes_{pot}(F_{k\text{-}ass})$: l'ensemble des mesures potentielles de $F_{k\text{-}ass}$, à déterminer.

Extraction des mesures potentielles : l'ensemble $Mes_{pot}(F_{k\text{-}ass})$ est déterminé, selon que $F_{k\text{-}ass}$ est une entité ou une association n-aire comme suit :

– Si $F_{k\text{-}ass}$ est une entité alors $Mes_{pot}(F_{k\text{-}ass})$ est formé par l'union des attributs numériques non-clés appartenant à :

Rm1. L'entité $F_{k\text{-}ass}$;

Rm2. Les concepts liées à $F_{k\text{-}ass}$ par un lien (1,1) ;

Rm3. Les entités directement liées à $F_{k\text{-}ass}$ par un lien (1,n).

– Si $F_{k\text{-}ass}$ est une association n-aire, alors $Mes_{pot}(F_{k\text{-}ass})$ est formé par l'union des attributs numériques non-clés appartenant à :

Rm4. L'association n-aire $F_{k\text{-}ass}$;

Rm5. Les associations m-aires *parallèles* à $F_{k\text{-}ass}$, sachant qu'une association m-aire R1 est dite *parallèle* à une association n-aire R2 ($n \geq m$) si et seulement si *toutes* les entités reliées par R1 sont aussi reliées par R2 ;

Rm6. Les entités reliées par $F_{k\text{-}ass}$.

Mise en correspondance des mesures : la correspondance d'une mesure M_s d'un fait F_k détermine dans $Mes_{pot}(F_{k\text{-}ass})$ l'attribut qui correspond terminologiquement à M_s. Si aucune des mesures de F_k ne possède de correspond dans la source, alors le schéma en étoile analysant F_k ne peut pas être alimenté ; en conséquence les opérations de correspondance des dimensions et des hiérarchies seront abandonnées pour le fait $F_{k\text{-}ass}$.

Exemple (suite) : La correspondance de la mesure du fait « encadrement » avec la source *S1* est illustrée dans la figure 9.

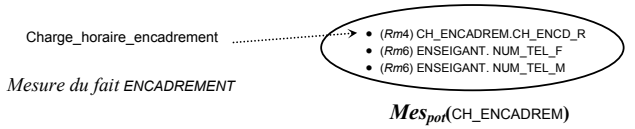

FIG. 9– *Mesures potentielles de ENCADREMENT et correspondance des mesures.*

6.1.3 Correspondance des dimensions

Cette étape concerne les dimensions des faits ayant au moins une mesure associée. La recherche du correspondant d'une dimension revient à la recherche du correspondant de son identifiant. Si cette correspondance ne peut être établie, alors aucun des autres paramètres de cette dimension ne peut exister dans la source : ceci s'explique par le fait qu'un système OLTP contient des données décrivant les détails les plus fins de l'activité de l'entreprise, et donc si le niveau le plus fin d'une dimension n'est pas présent alors aucun de ses niveaux agrégés ne peut l'être.

Dans le cas où le paramètre identifiant la dimension possède un correspondant, nous passons à la correspondance de ses attributs faibles.

(a) Correspondance de l'identifiant d'une dimension

La correspondance d'un identifiant de dimension nécessite la détermination de l'*ensemble des identifiants potentiels des dimensions*.

Définition 3. L'ensemble des ***identifiants potentiels des dimensions*** d'un fait *F,* noté $IdDim_{pot}(F)$, est formé de tous les attributs, qu'on peut extraire d'une source *S*, pouvant être des identifiants pour les dimensions de *F*.

Notation :

- F_k : un fait d'un schéma en étoile ;
- $F_{k\text{-}ass}$: un des faits potentiels déjà associés à F_k ;
- $Mes_{ass}(F_k)$: l'ensemble des mesures associées aux mesures de F_k dans l'étape précédente ;
- D_m : une dimension de F_k ;
- $IdDim_{pot}(F_{k\text{-}ass})$: ensemble des identifiants potentiels des dimensions de $F_{k\text{-}ass}$, à déterminer.

Extraction des identifiants potentiels des dimensions : l'ensemble $IdDim_{pot}\ (F_{k\text{-}ass})$ est formé par l'union des attributs qui ne sont pas des mesures, c'est-à-dire n'appartenant pas à $Mes_{ass}(F_k)$, et appartenant à/aux :

Rid1. $F_{k\text{-}ass}$,

Rid2. Concepts directement liés à $F_{k\text{-}ass}$ par un lien (1,1) ou un lien (1,n),
Rid3. Concepts appartenant à la fermeture transitive des concepts trouvés dans *Rid2*.

Mise en correspondance des identifiants des dimensions : la correspondance de l'identifiant Id_m d'une dimension D_m d'un fait F_k détermine dans $IdDim_{pot}(F_{k\text{-}ass})$ un attribut de même type (ou de type compatible) que Id_m et qui correspond terminologiquement à Id_m.

Si Id_m existe dans un schéma en étoile Sch_j déjà traité et a été associé à un correspondant $Id_{m\text{-}ass}$, alors, deux cas peuvent se présenter :

- $Id_{m\text{-}ass} \in IdDim_{pot}(F_{k\text{-}ass})$: dans ce cas, nous prenons le résultat de correspondance obtenu pour D_m et ses hiérarchies à partir de la correspondance de Sch_j.
- $Id_{m\text{-}ass} \notin IdDim_{pot}(F_{k\text{-}ass})$: dans ce cas, la dimension D_m ne possède pas un correspondant dans le schéma en étoile analysant le fait F_k. Cela signifie que D_m est une mauvaise dimension d'analyse de F_k.

Exemple (suite) : Continuons l'exemple par la correspondance des identifiants de toutes les dimensions de Sch_1 avec la source *S1*. La figure 10 montre l'ensemble des identifiants potentiels des dimensions et les correspondances effectuées.

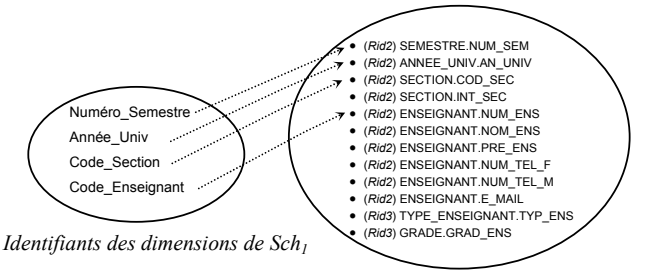

FIG. 10– *Exemple de correspondance des dimensions d'un schéma en étoile.*

(b) Correspondance des attributs faibles de l'identifiant d'une dimension

Cette étape concerne chaque identifiant de dimension possédant un correspondant et ayant des attributs faibles.

Notation :

- F_k : un fait d'un schéma en étoile ;
- D_m : une dimension liée à F_k ;
- Id_m : l'identifiant de D_m ;
- $Id_{m\text{-}ass}$: l'attribut associé à Id_m dans l'étape précédente ;
- $Faible_{pot}(Id_{m\text{-}ass})$: ensemble des attributs faibles potentiels de $Id_{m\text{-}ass}$, à déterminer ;
- $Mes_{ass}(F_k)$: l'ensemble des mesures déjà associées aux mesures de F_k.

Extraction des attributs faibles potentiels de l'identifiant d'une dimension : l'ensemble $Faible_{pot}(Id_{m\text{-}ass})$ est formé par l'union des attributs non-clés, n'appartenant pas à $(Mes_{ass}(F_k) \cup \{Id_{m\text{-}ass}\})$ et appartenant au(x) :

Ra1. Concept *c* contenant $Id_{m\text{-}ass}$,
Ra2. Concepts directement liés à *c* par un lien (1,1).

Mise en correspondance des attributs faibles de l'identifiant d'une dimension : La correspondance d'un attribut faible Af_b de Id_m détermine dans $Faible_{pot}(Id_{m\text{-}ass})$ l'attribut de même type (ou de type compatible) que Af_b et qui correspond terminologiquement à Af_b.

Exemple (suite) : La correspondance des attributs faibles de l'identifiant « code_enseignant » de la dimension « enseignant » est illustrée dans la figure 11.

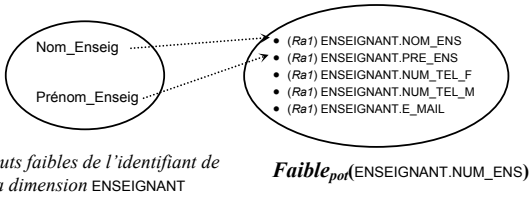

FIG. 11– *Exemple de correspondance des attributs faibles d'une dimension.*

6.1.4 Correspondance des hiérarchies

Cette étape concerne les paramètres de rang strictement supérieur à 1 de chacune des hiérarchies d'une dimension D_m validée, et les attributs faibles de ces paramètres.

Notation :
- h_q^{Dm} : une hiérarchie d'une dimension D_m,
- $Param_{valid}(h_q^{Dm})$: l'ensemble des paramètres valides (qui ont un correspondant dans la source) de h_q^{Dm}.

La correspondance d'une hiérarchie h_q^{Dm} se déduit de celle de ses paramètres comme suit :
- Si $Param_{valid}(h_q^{Dm}) = \emptyset$, alors la hiérarchie est absente dans la source.
- Si $Param_{valid}(h_q^{Dm}) \neq \emptyset$, alors la hiérarchie h_q^{Dm} possède un correspondant, si et seulement si, il n'existe aucune hiérarchie $h_p^{Dm} \neq h_q^{Dm}$ de D_m tel que $Param_{valid}(h_q^{Dm}) \subseteq Param_{valid}(h_p^{Dm})$, autrement dit, si et seulement si, h_q^{Dm} possède au moins un paramètre valide n'appartenant pas à une autre hiérarchie valide de D_m.

Le cas d'inclusion d'une hiérarchie h_q^{Dm} dans une hiérarchie h_p^{Dm} est illustré par la figure suivante :

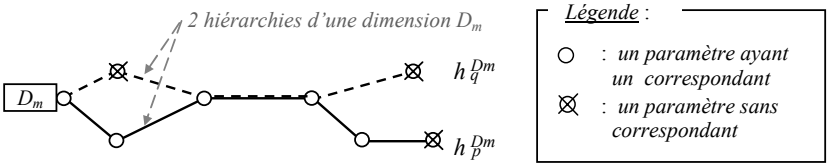

FIG. 12– *Correspondance des hiérarchies : cas d'inclusion de h_q^{Dm} dans h_p^{Dm}.*

Notation :
- F_k : un fait d'un schéma en étoile,
- D_m : une dimension liée à F_k ,
- $Mes_{ass}(F_k)$: l'ensemble des attributs sources associées aux mesures de F_k,
- h_q^{Dm} : une hiérarchie de D_m,
- P_i : le paramètre de rang i dans h_q^{Dm} ,
- P_d : le dernier paramètre des précédents de P_i dans h_q^{Dm} ayant un correspondant,
- $P_{d\text{-}ass}$: l'attribut source associé à P_d,
- $Faible_{ass}(P_{d\text{-}ass})$: l'ensemble des attributs sources associés aux attributs faibles de $P_{d\text{-}ass}$.

(a) Correspondance d'un paramètre d'une hiérarchie

Nous commençons par déterminer l'ensemble $Param_{pot}(h_q^{Dm}, P_i)$ des paramètres potentiels du paramètre P_i de h_q^{Dm} .

Extraction de l'ensemble $Param_{pot}(h_q^{Dm}, P_i)$: cet ensemble est formé par l'union des attributs, de même type (ou de type compatible) que P_i, n'appartenant pas à $(Mes_{ass}(F_k) \cup \{P_{d\text{-}ass}\} \cup Faible_{ass}(P_{d\text{-}ass}))$ et appartenant au(x) :

Rp1. concept c contenant $P_{d\text{-}ass}$,

Rp2. concepts directement liés à c par un lien (1,1) ou (1,n),

Rp3. concepts appartenant à la fermeture transitive des concepts trouvés dans *Rp2*.

Mise en correspondance d'un paramètre : la correspondance d'un paramètre P_i de h_q^{Dm} retrouve dans $Param_{pot}(h_q^{Dm}, P_i)$ l'attribut $P_{i\text{-}ass}$ qui correspond terminologiquement à P_i.

Si P_i existe dans une hiérarchie (h_f^{Dm} de D_m) déjà visitée, et y possède un correspondant $P_{i\text{-}ass}$, alors deux cas peuvent se présenter :
- $P_{i\text{-}ass} \in Param_{pot}(h_q^{Dm}, P_i)$: nous retenons la correspondance de P_i (et celle de ses attributs faibles).
- $P_{i\text{-}ass} \notin Param_{pot}(h_q^{Dm}, P_i)$: P_i de h_q^{Dm} ne possède pas de correspondant.

Exemple (suite) : la correspondance du paramètre « Type_enseignant » de rang 2 dans la hiérarchie «H1-Enseig» (figure 4.c) est illustrée dans la figure 13.

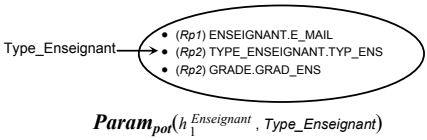

FIG. 13– *Exemple de correspondance d'un paramètre d'une hiérarchie.*

(b) Correspondance des attributs faibles de chaque paramètre

Cette étape concerne les attributs faibles de chaque paramètre P_i ayant un correspondant.
Notation :
- $P_{i\text{-}ass}$: l'attribut source associé à P_i déterminé dans l'étape précédente,
- $Faible_{pot}(P_{i\text{-}ass})$: l'ensemble des attributs faibles potentiels de P_i, à déterminer.

Extraction des attributs faibles potentiels d'un paramètre : l'ensemble $Faible_{pot}(P_{i\text{-}ass})$ est formé par l'union des attributs non-clés, n'appartenant pas à ($Mes_{ass}(F_k) \cup \{P_{d\text{-}ass}\} \cup Faible_{ass}(P_{d\text{-}ass}) \cup \{P_{i\text{-}ass}\}$) et appartenant au(x) :

Rap1. concept *c* contenant $P_{i\text{-}ass}$,

Rap2. concepts liés à *c* par un lien (1,1).

Mise en correspondance des attributs faibles d'un paramètre : la correspondance d'un attribut faible Af_b de P_i retrouve dans $Faible_{pot}(P_{i\text{-}ass})$ l'attribut de même type (ou de type compatible) que Af_b et qui correspond terminologiquement à Af_b.

Exemple (suite) : La correspondance du schéma en étoile Sch_1 avec la source *S1* est stockée dans la table de correspondance de la figure 14.

CONCEPTS DU SCHEMA Sch_1							CONCEPTS ASSOCIES Sol_1 DE MAPPING
FAIT						ENCADREMENT	CH_ENCADREM
MESURE(S)						*Charge_horaire_encadrement*	CH_ENCADREM.CH_ENCD_R
DIMENSIONS	ENSEIGNANT	NIVEAU 1 (Identifiant)			paramètre	*Code Enseignant*	ENSEIGNANT.NUM_ENS
					attributs faibles	*Nom Enseignant*	ENSEIGNANT.NOM_ENS
						Prénom Enseignant	ENSEIGNANT.PRE_ENS
		HIERARCHIES	H1_Enseig	NIVEAU 2	paramètre	*Type Enseignant*	TYPE_ENSEIGNANT.TYP_ENS
			H2_Enseig	NIVEAU 2	paramètre	*Grade Enseignant*	GRADE.GRAD_ENS
	SECTION	NIVEAU 1 (Identifiant)			paramètre	*Code Section*	SECTION.COD_SEC
					attributs faibles	*Libellé Section*	SECTION.INT_SEC
	ANNEE UNIVERSITAIRE	NIVEAU 1 (Identifiant)			paramètre	*Année Universitaire*	ANNEE_UNIV.AN_UNIV
	SEMESTRE	NIVEAU 1 (Identifiant)			paramètre	*Numéro Semestre*	SEMESTRE.NUM_SEM

FIG. 14– *Table de correspondance du schéma en étoile Sch_1 avec le schéma source S1.*

6.2 Validation et raffinement des schémas en étoile

Le résultat de la correspondance de chaque schéma en étoile est représenté dans une table de correspondance ayant une colonne par solution de correspondance.

Après la correspondance des schémas en étoile avec la source, leur validation s'effectue en trois étapes :

1. *Ajustement des correspondances effectuées* : c'est une opération qui nécessite l'intervention manuelle du concepteur des MD. Elle consiste à :
   (a) valider la correspondance automatiquement effectuée et corriger/supprimer les correspondances incorrectes.
   (b) correspondre manuellement (si possible) les mesures et les attributs n'ayant pas de correspondants avec leurs attributs potentiels sources.
2. *Choix de la correspondance la plus appropriée* : c'est une opération semi-automatique. Elle consiste à déterminer pour chaque schéma en étoile la solution de correspondance la plus appropriée. Nous détaillons cette étape dans le paragraphe 6.2.1.
3. *Amélioration de la correspondance retenue* : il s'agit de parfaire les correspondances retenues de l'étape précédente. Elle consiste à :
   (a) éliminer (automatiquement) les éléments (mesures, dimensions, hiérarchies, paramètres et attributs faibles) n'ayant pas de correspondant dans la solution retenue de chaque schéma en étoile.
   (b) enrichir (manuellement) les mesures des faits en ajoutant des mesures calculables à partir des mesures valides (par exemple les mesures calculant

des totaux), ou à partir des attributs sources (par exemple la mesure de la durée d'une activité est calculée par différence de deux dates).

(c) ajouter (manuellement) aux schémas en étoile obtenus des dimensions et des attributs supplémentaires, disponibles dans la source, afin d'offrir d'autres axes d'analyses aux utilisateurs finaux des MD. Les dimensions ajoutées à un schéma en étoile doivent respecter la contrainte sur la granularité du fait [Kimball, 1996]. Les paramètres ajoutés doivent respecter l'ordre dans les hiérarchies correspondantes.

6.2.1 Choix d'une correspondance

Si un fait F_k d'un schéma en étoile Sch_k est associé à plusieurs faits potentiels (Fp_1, ..., Fp_i, ...,Fp_n), alors une solution de correspondance de Sch_k est possible en effectuant les étapes de correspondance des mesures, des dimensions et des hiérarchies à partir de chaque Fp_i associé à F_k (figure 6). Pour évaluer les solutions de correspondance candidates de Sch_k, nous observons les nombres de leurs composants (mesures, dimensions, hiérarchies et paramètres) ayant un correspondant. Pour cela, nous définissons les métriques suivantes :

- La métrique *M1* des mesures : c'est le nombre de mesures de Sch_k qui possèdent un correspondant dans la source.
- La métrique *M2* des *dimensions* : c'est le nombre de dimensions de Sch_k qui possèdent un correspondant dans la source.
- La métrique *M3* des *hiérarchies* : c'est le nombre de hiérarchies des dimensions de Sch_k qui possèdent un correspondant dans la source.
- La métrique *M4* des *paramètres* : c'est le nombre de paramètres des dimensions de Sch_k qui possèdent un correspondant dans la source.

L'algorithme de principe pour le choix d'une solution de correspondance à partir des *m* solutions trouvées pour un même schéma est inspiré de l'exemple étudié dans [Bonifati *et al.,* 2001]. Il utilise une matrice des métriques Met[m,4] tel que Met[i,j] est la valeur de la métrique *Mj* dans la solution à la ligne *i*.

```
Entrée
Une table de correspondance d'un schéma en étoile.
Début
- Calculer la matrice Met
- Supprimer de Met les lignes i correspondant à des solutions sans
  dimensions c-à-d Met[i,2]=0
- Chercher les lignes de Met qui dominent toutes les autres pour les
  quatre métriques
 Si aucune ligne trouvée Alors
   - Supprimer de Met les lignes dominées pour les quatre métriques
   - Chercher les lignes de Met qui dominent toutes les autres pour
     les trois premières métriques (M1, M2, M3)
   Si aucune ligne trouvée Alors
      - Supprimer de Met les lignes dominées pour les trois premières
      métriques
      - Chercher une solution de Met qui domine toutes les autres
        pour les deux premières métriques(M1, M2)
```

```
      Si aucune ligne trouvée Alors
        - Supprimer de Met les lignes dominées pour les deux
        premières métriques
        - Donner la main au concepteur pour choisir une des solutions
          restantes dans Met
      Fin Si
    Fin Si
  Fin Si
- Valider le résultat par le concepteur des MD
Fin
```

NOTE. — La solution de la ligne *k* est dite dominante sur les *n* premières métriques ($n \leq 4$), si chacune des *n* premières valeurs de *k* est la plus grande de sa colonne.

7 Construction des schémas en constellation

Un schéma de MD peut être en étoile ou en constellation. Une constellation est une fusion de plusieurs schémas en étoile [Teste, 2000]. Il regroupe plusieurs faits étudiés selon différentes dimensions éventuellement partagées. L'avantage de la modélisation en constellation est de permettre des opérations de forage transversal ("*DrillAcross*"). Ces opérations comparent des faits d'un schéma en constellation par rapport aux dimensions qu'ils se partagent. Par exemple, un schéma en constellation analysant les faits *Vente* et *Achat* se partageant les dimensions *Produit* et *Temps*, permet de comparer les *montants* des ventes et les montants des achats par rapport aux axes d'analyses.

Le module de génération de schémas en constellation prend en entrée l'ensemble des schémas en étoile valides, issus de l'étape précédente. Les schémas en étoile d'un même domaine et ayant des dimensions communes peuvent être intégrés pour construire des schémas en constellation. De plus, les schémas résultants peuvent aussi être intégrés pour générer d'autres schémas en constellation. Ce processus itératif de constellation de schémas traite à chaque itération les schémas les plus pertinents à fusionner. Pour mesurer la pertinence de cette fusion, nous définissons un coefficient de similitude entre schémas multidimensionnels inspiré de l'intégration de schémas de bases de données.

Notre méthode de calcul de la similarité entre deux schémas Sch_i et Sch_j classe les coefficients de similitude en huit partitions selon la position du nombre de dimension commune (*p*) par rapport aux nombres de dimensions de chacun des deux schémas (*n* et *m*). Le coefficient de similitude de chaque partition est donné par le tableau suivant :

Condition	$Sim(Sch_i, Sch_j)$	Description
si p=0	0	Aucune dimension commune à Sch_i et Sch_j
si p=n=m	1	Toutes les dimensions sont communes
si p=1	1/5	Une seule dimension commune
si p=n et n<m	3/4	Toutes les dimensions d'un schéma sont incluses dans l'autre
si p=n/2 et n=m	1/2	La moitié des dimensions est commune
si p≥m/2 et n<m	2/3	Le nombre de dimensions communes est au moins égal à la moitié des dimensions du grand schéma (ayant le plus de dimensions)

si $n/2 \leq p < m/2$	1/3	Le nombre de dimensions communes est au moins égal à la moitié des dimensions du petit schéma (ayant le moins de dimensions), et n'atteint pas la moitié des dimensions de l'autre
si $p < n/2$ et $n \leq m$	1/4	Le nombre de dimensions communes n'atteint pas la moitié des dimensions de chaque schéma

TAB 1 – *Coefficient de similitude entre deux schémas multidimensionnels.*

Nous avons proposé dans [Soussi, *et al* 2004] et [Nabli *et al*, 2005] un algorithme itératif et interactif pour la génération des schémas en constellation des MD. Cet algorithme fait intervenir le concepteur pour choisir le nombre maximal de faits par schéma (*NFait*) et la valeur de similitude (*Vsimilt*) au-delà de laquelle la constellation sera réalisée. *NFait* est, en pratique, de l'ordre d'une demi-dizaine [Gouarné, 1998]. La constellation peut être forte ou faible selon la valeur de similitude *Vsimilt* choisie. Cette valeur doit être strictement supérieure à 1/5 (cas p=1) car il est n'est pas intéressant de fusionner deux schémas ayant une seule dimension commune (généralement, la dimension Temps est présente dans tous les schémas de MD).

A chaque itération de l'algorithme, nous relevons les schémas les plus similaires, c'est à dire ayant un coefficient de similitude maximal. Si ce coefficient est supérieur à *Vsimilt* et le nombre de fait des schémas à consteller ne dépasse pas *NFait,* alors la constellation de ces schémas est effectuée. Les schémas ayant participés à une constellation n'interviennent plus. Les schémas en constellation obtenus serviront pour la prochaine itération. L'algorithme s'arrête lorsque la constellation n'est plus possible, ou sur demande du concepteur.

La constellation de deux schémas multidimensionnels Sch_1 et Sch_2 produit un schéma en constellation Sch_3 ayant pour faits l'union des faits analysés dans Sch_1 et Sch_2 et pour dimensions l'union de leurs dimensions. Chaque fait de Sch_3 reste lié aux mêmes dimensions que dans son schéma initial. Les hiérarchies de chaque dimension *D* de Sch_3 sont obtenues par l'union des hiérarchies de *D* dans Sch_1 et Sch_2.

8 Conclusion

Notre travail s'inscrit dans le contexte de développement d'une approche de conception semi-automatique d'entrepôts de données en partant des besoins OLAP spécifiés sous forme de tableaux n-dimensionnels. Nous avons présenté dans ce papier une approche semi-automatique de construction de schémas multidimensionnels valides de MD. Notre approche est composée de trois modules : (a) génération de schémas en étoile idéaux, (b) correspondance/validation des schémas en étoile, et (c) génération des schémas en constellation à partir des étoiles validées. La correspondance effectuée entre les schémas en étoile et les schémas des sources OLTP s'appuie sur des règles d'extraction par concept du schéma en étoile, des règles de correspondance avec les sources de données, et des règles de validation. Nos règles de correspondance sont suffisamment détaillées et ont le mérite d'être automatisables. De plus, la correspondance effectuée pour valider les schémas conceptuels des MD facilite la phase de création des schémas logiques et la génération des procédures de chargement.

Nous travaillons actuellement sur l'implémentation de la correspondance entre les schémas des MD et des sources logiques relationnelles. L'intégration des MD pour construire le schéma de l'entrepôt fait l'objet d'une recherche en parallèle [Feki *et al.,* 2005].

Références

[Bonifati *et al*, 2001] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, et S. Paraboschi. Designing Data Marts for Data Warehouse. *ACM Transaction on Software Engineering and Methodology*, *ACM*, vol. 10, Octobre 2001, p. 452-483.

[Cabibbo et Torlone, 1998] L. Cabibbo et R. Torlone. A logical approach to multidimensional databases. *Proc. 6th EDBT* 1998, LNCS 1377, 183–197.

[Feki, 2004] J. Feki. Vers une conception automatisée des entrepôts de données : Modélisation des besoins OLAP et génération de schémas multidimensionnels. *8th MCSEAI*, 9-12 Mai 2004, Sousse-Tunisie.

[Feki *et al.*, 2005] J. Feki, J. Majdoubi et F. Gargouri. A Two-Phase Approach for Multidimensional Schemes Integration. *The 17th International Conference on Software Engineering and Knowledge Engineering* Juillet 14~16, 2005, Taipei, Taiwan, République de Chine (à paraître).

[Golfarelli *et al*, 1998] M. Golfarelli, D. Maio et S. Rizzi. Conceptual Design of Data Warehouses from E/R Schemas. *Proc. of the 31st Hawaii Int'l Conference on System Sciences*, Vol. VII, Kona, Hawaii, 1998, pp. 334-343.

[Gouarné, 1998] J. M. Gouarné. Le projet décisionnel. Eyrolles 1998.

[Hüsemann *et al.,* 2000] B. Hüsemann, J. Lechtenbörger et G. Vossen. Conceptual Data Warehouse Design. *Proc. of the Int'l Workshop on Design and Management of Data Warehouses*, Stockholm, Sweden, 2000, pp. 6.1-6.11.

[Kimball, 1996] R. Kimball. The Data Warehouse Toolkit. John Wiley and Sons, Inc., New York, 1996.

[Kimball, 1997] R. Kimball. A Dimensional Modelling Manifesto. *DBMS Magazine*, Juillet, 1997.

[Moody et Kortink, 2000] L. D. Moody et M. A. R. Kortink. From Enterprise Models to Dimensional Models: A Methodology for Data Warehouses and Data Mart Design. *Proc. of the Int'l Workshop on Design and Management of Data Warehouses*, Stockholm, Sweden, 2000, pp. 5.1-5.12.

[Nabli *et al*, 2005] A. Nabli, A. Soussi, J. Feki, H. Ben Abdallah et F. Gargouri. Towards an automatic Data Mart Design. *ICEIS*, 2005 (à paraître).

[Phipps et Davis, 2002] C. Phipps et K. Davis. Automating data warehouse conceptual schema design and evaluation. *DMDW'02*, Canada, 2002.

[Ravat *et al*, 2001] F. Ravat, O. Teste et G. Zurfluh. Modélisation multidimensionnelle des systèmes décisionnels. *Revue ECA*, Volume 1 - n°1-2/2001 - ISBN 2-7462-0216-6.

[Soussi *et al*, 2004] A. Soussi, A. Nabli, J. Feki et F. Gargouri. Construction de schémas de Data Marts par fusion de besoins Olap : Génération de modèles multidimensionnels à hiérarchies multiples. *GEI'04*, Tunisie, Monastir, Mars 2004.

[Soussi, *et al* 2005] A. Soussi, J. Feki et F. Gargouri. Génération et validation automatiques de schémas de magasins de données. *GEI'05*, Tunisie, Sousse, Mars 2005.

[Teste, 2000] O. Teste. *Modélisation et manipulation d'entrepôts de données complexes et historisées*. Thèse de l'Université Paul Sabatier (Toulouse III), Décembre 2000.

Summary

This paper is within the scope of a semi-automatic design approach for multidimensional schemes. In this approach, OLAP requirements are specified as n-dimensional sheets. This specification is data source independent. Since these requirements are used to build star schemes, the generated stars are ideal (theoretical) and therefore should be validated according to a real data source. In order to validate these stars, we propose first, a systematic method to match each multidimensional concept with a given data source schema, and secondly, some rules to refine this matching. After that, the validated stars are merged to generate constellation schemes.