Du relationnel au multidimensionnel : Conception de magasins de données

Jamel Feki, Yasser Hachaichi

Laboratoire MIRACL, Faculté des Sciences Economiques et de Gestion de Sfax
Route de l'Aérodrome km 4, B.P. 1088, 3018 Sfax, Tunisie
Jamel.Feki@fsegs.rnu.tn, Yasserhfr@yahoo.fr

Résumé. En vue d'assister le concepteur décisionnel, nous présentons une méthode ascendante de construction de schémas en étoile à partir d'une source relationnelle. Pour cela, nous étudions la structure des relations et nous proposons une classification en *relation-associations* et *relation-entités* permettant de construire des faits et des dimensions respectivement. Notre méthode a le mérite d'être indépendante de la sémantique du système d'information source. Elle exploite les contraintes de clés primaires et référentielles pour extraire les concepts multidimensionnels et affecte un niveau de pertinence à chaque concept extrait.

1 Introduction

Les systèmes d'information décisionnels (SID) sont dédiés au pilotage de l'entreprise. Ils constituent une synthèse des informations opérationnelles, internes ou externes, choisies pour leur pertinence et leur transversalité fonctionnelles. Ces systèmes sont basés sur des architectures logicielles permettant le stockage et l'interrogation de grands volumes de données complexes (Boussaid, 2006). Ils sont généralement organisés en deux espaces de stockage : l'***entrepôt de données*** (ED) regroupant toute l'information utile à la prise de décision et les ***magasins de données*** (MD). Chaque MD est un extrait de l'entrepôt ; c'est une base de données décisionnelle structurée en fonction d'un métier précis ou d'un usage particulier. L'information y est préparée sous une forme adaptée, dite multidimensionnelle, pour être directement et facilement accessible par les décideurs. Parfois, l'architecture du SID est réduite à des MD et ceci pour des raisons telle que d'économie de coûts, de délais du projet décisionnel, etc. Généralement, les entreprises de petites et moyennes tailles ne se permettent pas de supporter les coûts relativement élevés d'un entrepôt. En conséquence, elles construisent leurs MD directement sur leur base transactionnelle. Le présent travail s'intéresse à cette alternative. Il propose une démarche quasi-automatique d'aide à la construction de schéma de MD en étoile à partir d'une source relationnelle.

Cet article est organisé comme suit : la section 2 étudie l'état de l'art des méthodes de conception de MD et introduit les motivations de cette recherche ; la section 3 présente notre démarche quasi-automatique de construction de schéma de MD ; la section 4 définit la classe conceptuelle d'une relation et introduit un schéma relationnel pour illustrer notre démarche ; la section 5 définit nos heuristiques d'extraction de concepts multidimensionnels ; la section 6 évalue les résultats expérimentaux de notre approche et conclut l'article.

2 Conception de magasins de données : Etat de l'art et motivations

Dans la littérature des SID, il existe trois approches de conception de MD :

L'*approche descendante* (« Top-Down ») : proposée dans (Kimball, 1997) où l'auteur se base sur l'étude des besoins analytiques exprimés par les futurs utilisateurs décisionnels pour construire un ensemble de schémas de MD.

L'*approche ascendante* (« Bottom-Up ») : Construit des schémas de MD en se basant sur le modèle informatique du SI de l'entreprise. Le concepteur peut donc bénéficier des relations existantes entre les entités et suivre une méthode plus structurée pour concevoir la base de données décisionnelle. Ce type d'approche a été adopté dans (Golfarelli et al., 1998a), (Cabbibo et Torlone, 1998), (Moody et Kortink, 2000), (Husemann et al., 2000) et (Pang et al., 2006) .

L'*approche mixte* (« Mixed ») : Combine les deux approches précédentes. Elle a été d'abord adoptée par (Böhnlein et al., 1999) qui construit un Modèle Entité Relation Structuré (SERM) à partir d'un diagramme E/A. Ce modèle identifie les dimensions et les hiérarchies associées à un fait déduit suite à une étude des besoins analytiques et des objectifs. Ensuite, adopté par (Phipps et Davis, 2002) qui appliquent un algorithme de génération partant du modèle de données E/A de l'entreprise pour produire des schémas de MD candidats. Pour déterminer les solutions qui satisfont au mieux les besoins des décideurs, ces schémas candidats sont ensuite évalués par rapport à des requêtes décisionnelles types. Cette approche a également intéressée Bonifati (Bonifati et al., 2001) et (Ghozzi, 2004) qui effectuent une analyse descendante et une analyse ascendante suivie d'une intégration. Aussi, reprise par (Soussi et al, 2005) qui génère des schémas idéaux par fusion de besoins analytiques puis valide ces schémas idéaux obtenus par projection sur une source d'alimentation.

Le bilan de l'état de l'art des ces approches dégage les constatations suivantes :

- les approches descendantes et mixtes produisent des schémas candidats qui répondent aux besoins exprimés par les utilisateurs décisionnels ; néanmoins, elles présentent en pratique deux problèmes : (i) elles exigent du concepteur décisionnel une compétence en SI opérationnel et (ii) sont difficilement automatisables. En effet, ces approches sont présentées à travers des exemples au lieu de procédures de conception ;
- les approches ascendantes actuelles, bien qu'elles soient automatisables, se basent sur des diagrammes E/A que les entreprises n'en disposent pas toujours ou en disposent de versions obsolètes. De plus ces approches construisent des schémas candidats supposés tous équipertinents.
- Aucune de ces approches ne définit un ensemble de transformations formelles pour: (i) *dériver* automatiquement et univoquement les *représentations logiques* possibles d'un modèle conceptuel de MD développé ou (ii) *aider le concepteur* à en sélectionner la plus appropriée (Mazón et al. 2006) ;
- les dimensions se construisent à partir d'entités et parfois sur les attributs temporels (Golfarelli et al. 1998a), (Moody et al. 2000), (Soussi et al. 2005).
- les faits se construisent principalement sur les associations n-aire (Kimball 1997), (Golfarelli et al. 1998a), (Cabibbo et al. 1998), (Soussi et al. 2005) et rarement sur des entités (Moody et al. 2000), (Bonifati et al. 2001) et (Phipps et al. 2002).

D'autre part, à travers notre examen des travaux de la littérature nous confirmons l'hypothèse de construction des faits à partir d'association et nous constatons que les entités ayant générées des faits sont réellement des associations ; en effet, le formalisme E/A permet de représenter une association par le formalisme d'entité quand elle possède un identifiant (différent de la concaténation des clés de ses entités liées). Le tableau 1 donne l'origine conceptuelle des faits pour quelques exemples. Notons que dans (Golfarelli et al. 1998a) l'entité ADMISSION (tableau 1, ligne 4) est en réalité une association entre les entités DIAGNOSIS, WARD, D.R.G. et PHYSICIAN ; mais du fait qu'elle possède un identifiant (numéro séquentiel) elle a été modélisée comme entité.

Source E/A	Fait extrait et validé	Représentation conceptuelle
Activité commerciale (Golfarelli et al. 1998b)	VENTE	Assoc. ternaire
Répartition des charges des Enseignants (Soussi et al. 2005)	ENCADREMENT	Assoc. quaternaire
	ENSEIGNEMENT	Assoc. quaternaire
« Flight reservation system » (Böhnlein et al. 1999)	"BOOKING"	Assoc. binaire
« Hospital » (Golfarelli et al. 1998a)	"ADMISSION"	Entité[1]

TAB. 1 – *Origines des faits extraits de sources E/A.*

La suite de cet article détaille notre approche ascendante de construction assistée de schéma de MD partant d'une base de données relationnelle. Ce choix est motivé par le fait que le relationnel constitue encore le noyau fondamental pour la majorité des systèmes OLTP. De plus, le schéma de la source peut être facilement extrait à partir du SGBD ce qui nous permet de contourner les problèmes d'absence de la documentation classique (i.e. diagramme E/A) et essentiellement de son obsolescence.

3 Notre démarche

Pour construire quasi-automatiquement des schémas de MD candidats à partir d'une source relationnelle nous comptons exploiter au mieux la '*sémantique structurelle des relations*' disséminée essentiellement dans les définitions des clés primaires et contraintes référentielles. Nous proposons alors une démarche qui adhère aux points suivants :

- être indépendante de la sémantique du SI et de celle de son schéma dans la mesure où aucune signification des relations ni de leurs attributs ou des liens inter-relations n'est nécessaire,
- se base sur une distinction des relations du SI en : *relation-entité* et *relation-association* décrivant respectivement les entités et les associations ;
- exploite les liens structurels inter-relations exprimés par les clés primaires et les contraintes référentielles,
- affecte à chaque concept multidimensionnel un niveau de pertinence, et assiste le concepteur décisionnel.

Notre démarche se compose de trois étapes :

[1] Association conceptuellement représentée sous forme d'une entité.

Pré-construction. Elle extrait le schéma de la source (noms des tables, noms des colonnes et leur type, contraintes de clé primaire et référentielles) à partir du dictionnaire du SGBD. Aussi, elle attribue à chaque relation sa classe conceptuelle (entité ou association) ; ceci optimise ultérieurement l'identification des faits et des dimensions.

Construction des schémas de MD. Elle extrait les concepts multidimensionnels : les faits et leurs mesures ainsi que les dimensions avec leurs attributs organisés en hiérarchies puis, construit des schémas de MD candidats. Pour cela, nous avons défini des règles d'extraction qui associent à chaque concept extrait sa relation source et qui tiennent compte de la finesse (i.e. dépendance des mesures par rapport aux dimensions) des attributs identifiés comme mesures candidates.

L'association *concept-source* prépare le passage vers le niveau logique en garantissant la faisabilité des opérations ultérieures de mapping et de chargement. Egalement, nous proposons d'agréger des mesures lorsque ces dernières sont enregistrées de façon plus détaillée dans le système opérationnel que dans le fait (cf. § 5).

Validation. Elle permet au concepteur de valider les schémas en étoile candidats construits, c'est-à-dire, de les adapter aux besoins analytiques du système de pilotage (e.g. supprimer ou renommer des schémas ou des éléments…).

4 Pré-construction des schémas de MD

Etant donné un schéma relationnel, la pré-construction consiste essentiellement à déterminer la classe conceptuelle de chacune de ses relations.

Classe conceptuelle d'une relation. Rappelons que dans la littérature de la conception de MD, les dimensions se construisent à partir d'entités (ou d'attributs temporels) alors que les faits se construisent principalement sur les associations *n-aire* et rarement sur des entités matérialisant réellement des associations (cf. § 2).

Puisque notre démarche est guidée par une source relationnelle d'une part, et que en relationnel 'tout est relation' d'autre part, il se pose alors le problème de déterminer la *classe conceptuelle* d'une relation, c'est-à-dire, de savoir si une relation décrit une entité ou une association. La détermination de la classe conceptuelle nous a amené à effectuer un examen des structures des relations et notamment de leurs clés pour répartir l'ensemble *S* des relations d'un SI en deux sous ensembles :

- *Sa*: les relations de *S* décrivant des associations ; nous les appelons *relation-association* (R-a). En général, une *relation-association est reconnue par sa clé primaire composée d'au moins une clé étrangère*.
- *Se*: les relations de *S* décrivant des entités ; nous les appelons *relation-entité* (R-e). , En général, une *relation-entité est reconnue par sa clé primaire ne contenant aucune clé étrangère.*

Naturellement, la qualité du résultat de l'étape de construction dépendra de la qualité de cette classification, c'est-à-dire, de la bonne formation des deux sous ensembles *Sa* et *Se* qui doivent vérifier les trois propriétés suivantes :

- ***disjonction*** : $Sa \cap Se = \varnothing$;
- ***complétude*** : $Sa \cup Se = S$;

- ***exactitude*** : $\forall R \in Sa$, R n'est pas une entité et, $\forall R \in Se$, R n'est pas une association ; toute relation doit être correctement classée.

Pratiquement, l'exactitude n'est pas garantie lorsque la clé primaire d'une association :
(a) n'est pas la concaténation de ses clés étrangères ; cette clé primaire peut être un attribut artificiel tel qu'un numéro séquentiel (e.g. l'asso. ADMISSION du tableau 1), ou
(b) est la concaténation d'attributs venant d'entités vides : ces attributs ne sont pas des clés étrangères puisqu'une entité vide ne se transforme pas en une relation (e.g. dans la figure 1, la relation CHARGE_EXIGEE est en réalité une association reliant deux entités vides de clés respectifs NAT_ENSMT et GRAD_ENS).

Dans notre approche nous proposons d'identifier automatiquement la classe conceptuelle d'une relation. Dans les deux situations (a) et (b), une relation-association sera incorrectement identifiée comme relation-entité. Pour pallier à cette inexactitude, notre approche (i) construit aussi des faits sur des entités et les présente avec un faible niveau de pertinence, et (ii) offre au concepteur la possibilité de modifier la classe conceptuelle.

Pour la situation (a) la règle suivante peut assister le concepteur : '*Si pour une relation r* $\in$ *Se une clé candidate est définissable sur ses clés étrangères alors r est une association*'.

Afin d'illustrer notre démarche, nous considérons la source relationnelle modélisant la répartition des charges des enseignants à la FSEG[2]. La figure 1 montre le modèle logique de cette source où les clés primaires sont soulignées et, pour chaque clé étrangère la relation référencée est reportée ; la colonne gauche indique la classe conceptuelle (R-a ou R-e) affectée automatiquement à chaque relation de ce schéma.

R-e	**SECTION** (COD_SEC, INT_SEC)
R-e	**AUDITOIRE** (COD_AUD, INT_AUD, NUM_CYC, COD_SEC : SECTION)
R-e	**ETUDIANT** (NUM_ETUD, NOM_ETUD, PRENOM_ETUD)
R-e	**MATIERE** (COD_MAT, COD_AUD : **AUDITOIRE**, INT_MAT, VOL_HOR_MAT)
R-e	**ENSEIGNANT** (NUM_ENS, NOM_ENS, PRE_ENS, NUM_TEL_F, NUM_TEL_M, E_MAIL, TYP_ENS, GRAD_ENS)
R-a	**INSCRIT_DANS** (COD_AUD : **AUDITOIRE**, NUM_ETUD : **ETUDIANT**, AN_UNIV)
R-a	**PEUT_ENSEIGNER** (NUM_ENS : **ENSEIGNANT**, NAT_ENSMT, COD_MAT : **MATIERE**)
R-a	**ENCADRER** (NUM_ENS : **ENSEIGNANT**, NUMGRP, AN_UNIV)
R-a	**SOUS_GROUPE** (NUM_ETUD : **ETUDIANT**, AN_UNIV, NUM_GRP, NUM_SOUS_GRP)
R-a	**COMPORTE_GRP** (AN_UNIV, COD_MAT : **MATIERE**, NAT_ENSMT, NBR_GRP)
R-a	**CONTIENT_HEUR** (COD_ MAT : **MATIERE**, NAT_ENSMT, VOL_HOR_ENS_NAT)
R-a	**REGROUPE_ETUD** (AN_UNIV, COD_MAT : **MATIERE**, NBR_ETUD)
R-a	**ENSEIGNEMENT_ASSURE** (COD_MAT : **MATIERE**, NUM_ENS : **ENSEIGNANT**, NUM_SEM, NAT_ENSMT, AN_UNIV, NBR_GRP_ENS)
R-e	**CHARGE_EXIGEE** (NAT_ENSMT, GRAD_ENS, CHARG_HOR_EXI)
R-e	**CH_ENCADREM** (COD_SEC : **SECTION**, NUM_ENS : **ENSEIGNANT**, NUM_SEM, AN_UNIV, CH_ENCD_R)
R-e	**PLANIFIE_DANS** (DATE_PLANIF, NUM_SEM, COD_MAT : **MATIERE**)

FIG. 1 - *S1 Schéma relationnel de la répartition des charges des Enseignants à la FSEG.*

[2] Faculté des Sciences Economiques et de Gestion de l'Université de Sfax

5 Construction des schémas de MD

Pour construire les schémas des MD notre approche identifie les concepts multidimensionnels moyennant un ensemble d'heuristiques d'extraction qui attribue à chaque concept extrait un niveau de pertinence ainsi que son attribut et sa relation source.

5.1 Extraction des faits

Le fait représente un centre d'intérêt pour la prise de décision. En effet, il modélise un sujet d'analyse représentant un événement qui se produit au sein d'une organisation.

Pour extraire les faits, le critère de (Kimball 1997), (Bonifati et al. 2001), amélioré par (Soussi et al. 2005) construit un ensemble composé des représentations conceptuelles (entités ou associations) possédant au moins un attribut numérique non clé (primaire ou étrangère) et considère que les éléments de cet ensemble sont tous équipertinents. Pour une source relationnelle, nous définissons deux heuristiques (Hf1 et Hf2) qui construisent deux ensembles de faits de deux niveaux de pertinence. La première identifie les faits issus des *relation-associations* et la deuxième identifie les faits issus des *relation-entités*.

Etant construits sur des associations, nous considérons que les faits obtenus par Hf1 sont plus pertinents que ceux obtenus par Hf2 (cf. § 2). Cette distinction, absente dans les autres approches, offre une meilleure assistance au concepteur lors de la sélection des faits à retenir.

Hf1 : Toute relation-association *R* contenant au moins un attribut numérique non clé (primaire ou étrangère) est un fait candidat pertinent nommé *R*.

Hf2 : Toute relation-entité *R* contenant au moins un attribut numérique non clé (primaire ou étrangère) est un fait candidat de faible pertinence nommé *R*.

Tout au long de cet article, nous adoptons les notations suivantes :

- S : une source relationnelle en troisième forme normale,
- R : une relation appartenant à S,
- Ω_R : l'ensemble des attributs de R,
- $\Omega_{R/NUM}$: le sous-ensemble des attributs numériques de Ω_R,
- $\Omega_{R/BOL}$: le sous-ensemble des attributs booléens de Ω_R,
- $\Omega_{R/TEM}$: le sous-ensemble des attributs temporels (de type date ou temps) de Ω_R,
- Pk_R : l'ensemble des attributs formant la clé primaire de R ($Pk_R \subseteq \Omega_R$) et,
- Fk_R : l'ensemble des attributs clés étrangères de R ($Fk_R \subseteq \Omega_R$).

Formulation. Selon cette notation, les faits candidats extraits à partir de *S* sont qualifiés par :

Hf1 détermine $\{R \in Sa : \Omega_{R/NUM} - (Pk_R \cup Fk_R) \neq \phi\}$

Hf2 détermine $\{R \in Se : \Omega_{R/NUM} - (Pk_R \cup Fk_R) \neq \phi\}$

La figure 2 montre les faits candidats extraits de la source *S1*.

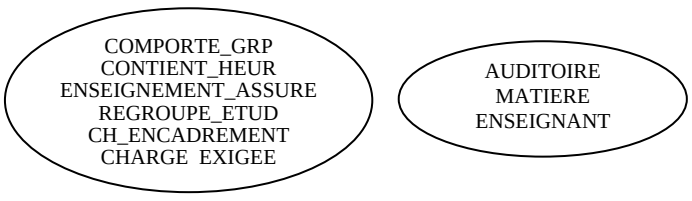


FIG. 2 – *Faits extraits de la source S1*:

5.2 **Extraction des mesures**

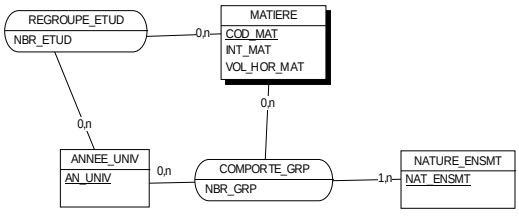
Un fait comporte un nombre fini de mesures qui sont des attributs numériques issus soit de *relation-fait*[3] soit de *relations parallèles*[4].

Extraction de mesures à partir d'une relation-fait. Ces mesures sont extraites par l'heuristique suivante :

Hm1 : Les attributs numériques *non clés* appartenant à une relation-fait *F* et *n'appartenant pas à d'autres relations* sont des mesures candidates pour *F*.

Nous écartons les clés de l'ensemble des mesures candidates car elles sont artificielles, redondantes et ne tracent pas l'activité de l'entreprise. Egalement, nous écartons les attributs non clés de *F* appartenant à d'autres relations car ils sont réellement des clés d'entités vides. Par exemple, pour le fait issu de la relation ENSEIGNANT nous écartons l'attribut GRAD_ENS de l'ensemble des mesures candidates car cet attribut appartient aussi à la relation CHARGE_EXIGEE. En réalité GRAD_ENS est la clé d'une entité vide (GRADE). Le tableau 2 montre les mesures extraites par cette heuristique pour la source *S1*.

Extraction de mesures à partir d'une relation-parallèle. Dans un schéma E/A, les mesures peuvent provenir des associations parallèles (Soussi et al., 2005). Une association *A1 m-aire* est dite parallèle à une association *A2 n-aire* ($m \leq n$) si et seulement si toutes les entités reliées par *A1* sont aussi reliées par *A2*. Cas de la figure 3-a



(*a*)

REGROUPE_ETUD (AN_UNIV, COD_MAT : **MATIERE**, NBR_ETUD)
COMPORTE_GRP (AN_UNIV, COD_MAT : **MATIERE**, NAT_ENSMT, NBR_GRP)
MATIERE (COD_MAT, COD_AUD : **AUDITOIRE**, INT_MAT, VOL_HOR_MAT)

(*b*)

FIG. 3 – *Exemple d'associations parallèles et leur transformation en relationnel.*

Nous identifions une relation-association *R1* parallèle à une autre relation-association *R2* par le fait que l'ensemble des attributs de la clé primaire de *R1* est inclus ou égal à celui de *R2* ($Pk_{R1} \subseteq Pk_{R2}$). Par exemple, dans la figure 3-b la relation-association REGROUPE_ETUD est parallèle à la relation-association COMPORTE_GRP.

Soient *R1* et *R2* deux relation-faits. A travers Hm1 nous identifions des mesures dans *R1* et dans *R2*. En outre, si *R1* est parallèle à *R2* alors le fait bâti sur *R1* peut recevoir d'autres mesures provenant de *R2* à condition de préserver leur dépendance par rapport aux dimensions de *R1*. Nous distinguons deux cas :

[3] Relation-fait est une relation identifiée comme fait

[4] Relation-parallèle est une relation transformée d'une association parallèle

Cas 1 : Dimension (R1) ⊂ Dimension (R2)

Pour ce cas, les mesures de *R2* sont enregistrées dans la source de façon plus détaillée que celles de *R1* ; ainsi pour les ajouter à *R1*, elles doivent être agrégées sur l'ensemble de dimensions Dimension (*R2*) - Dimension (*R1*), d'où l'heuristique suivante :

Hm2.1 : Si une relation-fait *R1* de classe *R-a* est parallèle à une autre relation-fait *R2* de classe *R-a* et si l'ensemble des dimensions de *R1* est strictement inclus dans celui de *R2* alors les mesures de *R2* déterminées par Hm1 s'ajoutent comme *mesures agrégées* à *R1*.

Dans notre méthode, les mesures à agréger ainsi que les dimensions de leur agrégation sont automatiquement identifiées, mais la fonction d'agrégation est à la charge du concepteur puisque son choix dépend de la sémantique d'analyse.

Exemple : Dans la relation-fait COMPORTE_GRP (cf. figure 3-a), la mesure NBR_GRP est enregistrée par année universitaire (AN_UNIV), nature d'enseignement (NAT_ENSMT) et matière (COD_MAT). Pour inclure cette mesure dans la relation-fait REGROUPE_ETUD, il faut l'agréger sur la dimension nature d'enseignement (DCA_NAT_ENSMT).

Cas 2 : Dimension (R1) = Dimension (R2)

A égalité de dimensions, *R1* est alors parallèle à *R2* et réciproquement. Ainsi, les mesures de *R1* et de *R2* ont le même niveau de détail. L'heuristique suivante illustre ce cas :

Hm2.2 : Si deux relation-faits *R1* et *R2* (de même classe *R-a*) sont parallèles et possèdent les mêmes dimensions alors les mesures de *R1* (respectivement de *R2*) extraites par Hm1 s'ajoutent à *R2* (respectivement à *R1*).

Les règles présentées génèrent des mesures avec un degré de pertinence décroissant à chaque fois qu'on s'éloigne de la relation-fait (application successive de Hm1 et (Hm2.1 ; Hm2.2)) Nous considérons alors que Hm1 est plus pertinente que Hm2.1 et Hm2.2.

Exemple : Pour le fait REGROUPE_ETUD la mesure NBRE_ETUD est plus pertinente que la mesure NBR_GRP.

Note : Les relation-associations parallèles identifiées comme faits candidats peuvent former un bon schéma en constellation.

Formulation. Si *R1* est une relation-fait de *S* alors l'ensemble de ses mesures candidates est l'union des trois ensembles obtenus par :

Hm1 : $\Omega_{R1/NUM} - \left(Pk_{R1} \cup Fk_{R1} \cup \bigcup_{\substack{Rj \in S, \\ Rj \neq R1}} \Omega_{Rj} \right)$

Hm2.1 : $\bigcup_{\substack{R2 \in Sa \\ R1 // R2 \\ D(R1) \subset D(R2)}} \left(\bigcup_{m \in M_{Hm1}^{R2}} \Omega_{R2} \{ Agr(m, D(R1) - D(R2)) \} \right)$ avec :

M_{Hm1}^{R} : Mesures de *R* extraites par Hm1.

$D(R1)$: Dimensions de la relation-fait *R1*.

$Agr(m, \{d1, d2, ..., dn\})$: Agrégation de la mesure *m* sur les dimensions *di*.

Hm2.2 : $\bigcup_{\substack{R2 \in Sa \\ R1 // R2 \\ D(R1) = D(R2)}} \left(\bigcup_{m \in M_{Hm1}^{R2}} \Omega_{R2} \{ m \} \right)$

Illustration. Le tableau 2 énumère les mesures candidates pour chaque fait de la figure 2.

Fait	Mesure extraite *m*	Heuristique	Formule d'extraction
COMPORTE_GRP	NBR_GRP	Hm1	-
	NBR_GRP_ENS	Hm2.1	Agr(*m*,ENSEIGNANT, DCA_NUM_SEM)
CONTIENT_HEUR	VOL_HOR_ENS_NAT	Hm1	-
	NBR_GRP	Hm2.1	Agr(*m*, DCA_NUM_SEM
	NBR_GRP_ENS	Hm2.1	Agr(*m*, ENSEIGNANT, DCA_NUM_SEM, DCA_AN_UNIV)
REGROUPE_ETUD	NBR_ETUD	Hm1	-
	NBR_GRP	Hm2.1	Agr(*m*, DCA_NAT_ENSMT
	NBR_GRP_ENS	Hm2.1	Agr(*m*, ENSEIGNANT, DCA_NUM_SEM, DCA_NAT_ENSMT)
ENSEIGNEMENT_ASSURE	NBRE_GRP_ENS	Hm1	-
CH_ENCADREMENT	CH_ENC_R	Hm1	-
AUDITOIRE	Num_CYC	Hm1	-
MATIERE	VOL_HOR_MAT	Hm1	-
ENSEIGNANT	NUM_TEL_F	Hm1	-
	NUM_TEL_M	Hm1	-
CHARGE_EXIGEE	CHARG_HOR_EXI	Hm1	-

TAB. 2 – *Mesures Extraites pour chaque fait de la figure 2.*

Nous complétons la construction des schémas en étoile par l'extraction des dimensions.

5.3 Extraction des dimensions

Naturellement, un fait est lié à un nombre *n (n>1)* fini de dimensions représentant les axes d'analyses. Une dimension est caractérisée par un nom et possède une liste d'attributs dont un identifiant. L'ensemble des dimensions candidates d'un fait *F* est construit soit à partir de relations (celles référencées par *F*) soit à partir d'attributs (Hachaichi et Feki, 2006).

5.3.1 Dimension construite à partir d'une relation

Hd1 : Toute relation-entité *R* directement référencée par une relation-fait *F* est une dimension candidate pour *F*. Le nom de cette dimension est celui de *R*, son identifiant est la clé primaire de *R*.

Exemple : La relation MATIERE est une dimension pour le fait COMPORTE-GRP, son identifiant est COD_MAT.

Nous désignons par *relation-dimension* toute relation identifiée comme dimension.

5.3.2 Dimension construite à partir d'un attribut

Cas d'un attribut booléen. Un attribut booléen appartenant à une relation-fait répartit ses tuples en deux sous ensembles et peut donc constituer un axe d'analyse. D'où Hd2 :

Hd2 : Tout attribut booléen appartenant à une relation-fait *F* donne naissance à une dimension candidate pour *F* dont il est l'identifiant.

Par exemple, un MD comportant une dimension *fidélité client* construite sur un attribut booléen renforce les combinaisons d'analyse.

Cas d'un attribut temporel. Dans le modèle dimensionnel, la dimension temps figure systématiquement dans tout entrepôt (Kimball 1997) considéré comme une série temporelle.

Hd3 : Tout attribut temporel (date ou temps) appartenant à une relation-fait *F* estampille le fait *F* et construit alors une dimension temporelle dont il est l'identifiant.

Exemple : Pour le cas traité par (Böhnlein et al., 1999), Hd3 construit une dimension sur l'attribut BOOKING_DATE pour le fait BOOKING comme prévue par l'auteur.

Cas d'un attribut représentant une entité vide. Généralement la transformation d'une entité vide (i.e. réduite à sa clé) en relationnel n'engendre pas une relation mais se traduit par migration de sa clé vers une autre relation. Cette clé peut jouer le rôle d'une dimension. Afin de localiser cette dimension dans une source relationnelle, nous avons distingué deux cas. Pour chacun nous présentons un exemple de motivation suivi de son heuristique :

***Cas 1** : Entité vide liée par une association multivaluée porteuse de données*

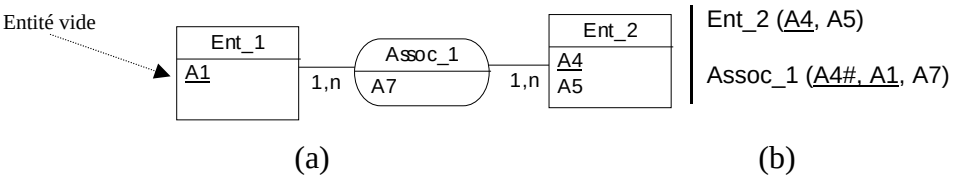


FIG. 4 – *Entité vide liée par une association multivaluée (a) et sa représentation relationnelle (b).*

Dans un contexte E/A (figure 4-a), si l'attribut *A7* est numérique alors Assoc_1 sera un fait candidat. Ce fait doit avoir parmi ses dimensions l'entité Ent_1. Mais Ent_1 ne se transforme pas en une relation mais sa clé compose la clé de la relation Assoc_1 (figure 4-b).

Ce cas type nous a permis de définir l'heuristique suivante :

Hd4 : Si un attribut de la clé primaire d'une relation-fait *F* de classe *R-a* n'est pas une clé étrangère, alors cet attribut construit une dimension candidate dont il est l'identifiant.

Exemple : Dans la figure 5-a, l'attribut NAT_ENSMT est identifié comme dimension associée au fait CONTIENT_HEUR. En réalité, cet attribut représente l'entité vide NATURE_ENSMT de la figure 5-b.

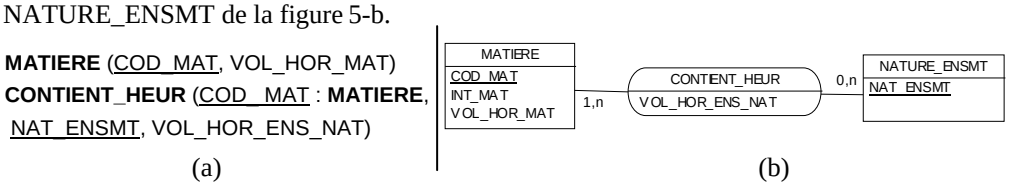


FIG. 5 – *Entité NATURE_ENSMT vide.*

***Cas 2** : Entité vide liée par une association CIF (Contrainte d'intégrité fonctionnelle)*

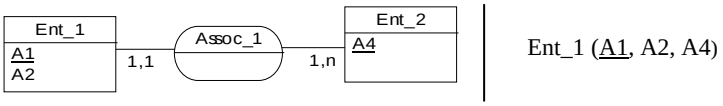


FIG. 6 – *Entité vide liée par une association CIF.*

Pour ce cas (figure 6), si l'attribut *A2* est numérique alors l'entité Ent_1 sera un fait candidat ayant comme dimension candidate l'entité Ent_2. Mais ce diagramme se transforme en une seule relation. Ceci nous fait perdre la méta-donnée « *A4 décrit l'entité ENT_2* ». En conséquence il serait malheureusement impossible de définir une heuristique permettant de déduire automatiquement que *A4* est l'identifiant d'une dimension candidate.

Exemple : Dans la figure 7, l'attribut TYP_ENS de l'entité vide TYPE_ENSEIGNANT a migré vers la relation ENSEIGNANT comme attribut simple

ENSEIGNANT (NUM_ENS, NOM_ENS, PRE_ENS, NUM_TEL_F, NUM_TEL_M, E_MAIL, *TYP_ENS*, GRAD_ENS)

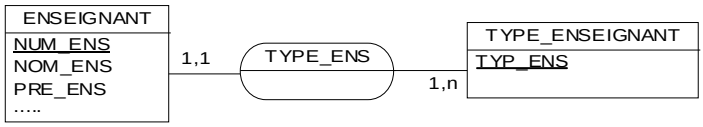


FIG. 7 – *Entité TYPE_ENSEIGNANT vide.*

Par ailleurs, nous pouvons définir l'attribut A4 comme une dimension si son entité Ent_2 est liée à au moins une autre entité (cf. figure 8). En effet, dans ce cas, A4 sera présent dans une autre relation en plus de sa présence dans Ent_1 c'est-à-dire :

- soit dans la relation Ent_3 (A7,…, A4) (si y = 1),
- soit dans la relation Assoc_2 (A4, A7) (si y>1).

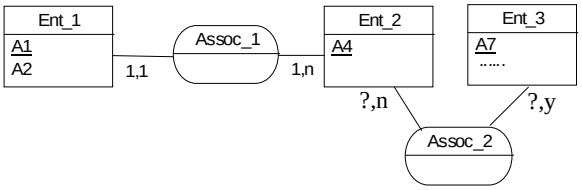


FIG. 8 – *Entité vide liée par une association CIF et à d'autres associations.*

Ce cas type nous a permis de définir l'heuristique suivante :

Hd5 : Tout attribut *a* non clé (primaire ou étrangère) appartenant à une relation-fait *F* et à d'autre(s) relation(s) est un identifiant candidat d'une dimension de *F* construite sur *a*.

Exemple : l'attribut GRAD_ENS appartenant aux relations ENSEIGNANT et CH_ENCADREM, est un identifiant de dimension définie sur le fait ENSEIGNANT.

Remarque : Les dimensions obtenues par Hd1 à Hd3 sont plus pertinentes que celles déduites par Hd4 et Hd5. En effet, ces dernières considèrent comme dimension la transformée d'une entité réduite à sa clé primaire. Or, ce type d'entité représente parfois une entité particulière qui ne mérite pas d'être un axe d'analyse comme, par exemple, une entité COMPTEUR générant des numéros séquentiels.

Nous convenons de nommer une dimension construite sur un attribut *a* par DCA_*a*.

Formulation. Si *R1* est une relation-fait d'une source *S* alors l'ensemble de ses dimensions candidates est l'union des cinq ensembles obtenus comme suit :

Hd1 : $\{R \in Se : Pk_R \cap Fk_{R1} \neq \phi\}$; est l'identifiant de la dimension bâtie sur *R est* Pk_R,

Hd2 : $\bigcup_{a \in \Omega_{R1/BOL}} \{DCA\_a\}$; l'identifiant de la dimension DCA_a est l'attribut a,

Hd3 : $\bigcup_{a \in \Omega_{R1/TEM}} \{DCA\_a\}$; l'identifiant de la dimension DCA_a est l'attribut a,

Hd4 : $\bigcup_{a \in (Pk_{R1} - Fk_{R1})} \{DCA\_a\}$; l'identifiant de la dimension DCA_a est l'attribut a.

Hd5 : $\bigcup_{a \in \Omega_{R1} - (Pk_{R1} \cup Fk_{R1})} \{DCA\_a : \exists R \in S - \{R1\} \wedge \Omega_{R1} - (Pk_{R1} \cup Fk_{R1}) \subseteq R\}$;

Illustration. L'application de ces heuristiques sur les faits candidats de la figure 1 produit l'ensemble des dimensions candidates du tableau 3.

Fait	Dimension extraite	Identifiant de dimension	Heuristique utilisée
AUDITOIRE	SECTION	COD_SEC	Hd1
MATIERE	AUDITOIRE	COD_AUD	Hd1
ENSEIGNANT	DCA_GRAD_ENS	GRAD_ENS	Hd5
COMPORTE_GRP	MATIERE	COD_MAT	Hd1
	DCA _AN_UNIV	AN_UNIV	Hd4
	DCA_NAT_ENSMT	NAT_ENSMT	Hd4
CONTIENT_HEUR	MATIERE	COD_MAT	Hd1
	DCA_NAT_ENSMT	NAT_ENSMT	Hd4
REGROUPE_ETUD	MATIERE	COD_MAT	Hd1
	DCA_AN_UNIV	AN_UNIV	Hd4
ENSEIGNEMENT_ASSURE	MATIERE	COD_MAT	Hd1
	ENSEIGNANT	COD_ENS	Hd1
	DCA_NUM_SEM	NUM_SEM	Hd4
	DCA_NAT_ENSMT	NAT_ENSMT	Hd4
	DCA_AN_UNIV	AN_UNIV	Hd4
CH_ENCADREM	SECTION	COD_SEC	Hd1
	ENSEIGNANT	COD_ENS	Hd1
	DCA_NUM_SEM	NUM_SEM	Hd4
	DCA _AN_UNIV	AN_UNIV	Hd4

TAB. 3 – *Dimensions extraites pour les faits du tableau* 2.

Notons que nous considérons "douteux" tout fait sans dimensions et nous l'éliminons. Par exemple, le fait CHARGE_EXIGEE de la figure 2 est supprimé.

5.4 Extraction des hiérarchies

Une hiérarchie organise les paramètres d'une dimension du plus fin au plus général conformément à leur niveau de détail. Par ailleurs, toute hiérarchie d'une dimension d part de l'identifiant de d qui est le paramètre le plus fin (de rang 1) déjà extrait avec la dimension. Nous continuons alors à extraire les paramètres de rang supérieur à 1.

Notons que les identifiants trouvés pour une dimension candidate proviennent soit d'une relation (par Hd1) soit d'un attribut (par Hd2 à Hd5). En réalité, une dimension construite sur un attribut ne peut s'étendre à d'autres niveaux de paramètres. Par contre, une dimension d construite sur une relation pourrait référencer d'autres relations ; ces dernières fourniront des paramètres pour les hiérarchies de d.

Nous présentons quatre heuristiques d'extraction des paramètres de rang 2, ensuite nous traitons l'extraction des paramètres de rang supérieur à 2.

Hh1 : Si la clé primaire *pk* d'une relation-entité est directement référencée par une relation-dimension *d* alors *pk* est un paramètre candidat de rang 2 d'une hiérarchie de *d*.

Par analogie à l'extraction des dimensions, un attribut booléen ou temporel, présent dans une relation-dimension peut constituer une hiérarchie.

Hh2 : Tout attribut booléen ou temporel appartenant à une relation-dimension est un paramètre candidat terminal de rang 2 d'une hiérarchie définie sur cette dimension.

Pour prendre en considération la transformation des entités vides liées par une association CIF (cf. figure 7) lors de l'extraction des paramètres, nous définissons l'heuristique Hh3.

Hh3 : Tout attribut non clé (primaire ou étrangère) appartenant simultanément à une relation-dimension *d* et à d'autre(s) relation(s) est un paramètre candidat de rang 2 d'une hiérarchie définie sur *d*.

Plus généralement, l'application récursive des heuristiques Hh1 à Hh3 sur les relations obtenues par Hh1 produit des paramètres de rang supérieur à 2.

Formulation. Si $R1 \in S$ est une relation-dimension (resp. une relation dont la clé est un paramètre de rang i >1) alors les paramètres de rang 2 (resp. de rang i+1) sont définis par :

Hh1 : $\{R \in Se : Pk_R \cap Fk_{R1} \neq \phi\}$

Hh2 : $\bigcup_{a \in (\Omega_{R1/BOL} \cup \Omega_{R1/TEM})} \{a\}$

Hh3 : $\bigcup_{a \in \Omega_{R1} - (Pk_{R1} \cup Fk_{R1})} \{a : \exists R \in S - \{R1\} \wedge \Omega_{R1} - (Pk_{R1} \cup Fk_{R1}) \subseteq R\}$

La figure 9 présente les hiérarchies de chaque dimension déduite par Hd1.

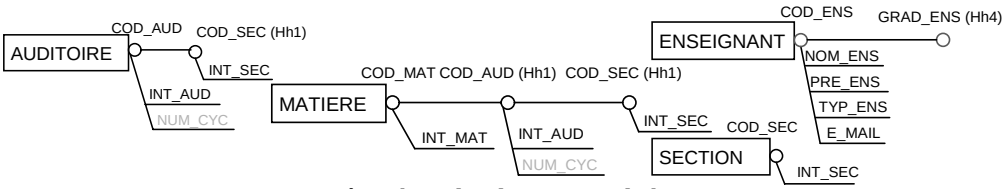


FIG. 9 – *Hiérarchies des dimensions de la source S1.*

5.5 Extraction des attributs faibles

Un attribut faible est en dépendance fonctionnelle élémentaire et directe du paramètre qu'il décrit. En conséquence, il se trouve dans la même relation que son paramètre. Vu que les attributs faibles sont descriptifs (Golfarelli, et al., 1998a), nous privilégions les attributs textuels en les considérant plus significatifs que ceux numériques. Haf extrait ces attributs.

Haf : Les attributs textuels (ou numériques) non clés appartenant à une relation dont la clé est un paramètre *p* (de rang quelconque) et n'appartenant pas à d'autres relations sont des attributs faibles pour *p*.

Formulation. Si $R1 \in S$ est une relation ayant sa clé un paramètre alors l'ensemble des attributs faibles de ce paramètre est déterminé par l'ensemble obtenu par :

Haf : $\bigcup_{a \in \Omega_{R1/TEX} - (Pk_{R1} \cup Fk_{R1})} \{a\} \cup \bigcup_{a \in \Omega_{R1/NUM} - (Pk_{R1} \cup Fk_{R1})} \{a\}$

Illustration. La figure 9 montre l'ensemble des attributs faibles de la source *S1*

6 Evaluation et conclusion

Nous avons présenté une méthode quasi-automatique d'aide à la construction de schémas de MD en étoile à partir d'une source relationnelle. Pour cela, nous avons examiné les structures des relations pour les classer en relation-associations et relation-entités pour faciliter la localisation des faits et des dimensions dans la source. Ensuite, pour l'extraction de chaque concept multidimensionnel, nous avons défini un ensemble d'heuristiques indépendantes de la sémantique des relations et exploitant les liens interrelations (clé primaire/étrangère). Par ailleurs, nos heuristiques classent les concepts multidimensionnels extraits d'un même type (e.g. faits, mesures…) par niveau de pertinence; ce qui assiste le concepteur décisionnel dans le choix des schémas les plus intéressants.

Nous avons illustré notre méthode sur un cas type et nous avons obtenus de bons résultats. Par ailleurs, et pour montrer son efficacité, nous l'avons appliqué sur des sources choisies dans la littérature des systèmes décisionnels pour lesquelles les schémas en étoiles sont construits manuellement par leur auteur. Le tableau 4 compare les résultats obtenus par notre méthode avec les résultats des auteurs des cas référencés.

Critère	**Taux de couverture des MD**	**Pertinence des MD obtenus et non envisagées**		**Taux de couverture des mesures et des dimensions des MD couverts**	
		E*	**F***	**Mes**	**Dim**
Cas 1 : Activité commerciale (Golfarelli et al. 1998b)	1/1	0	2	2/2	3/3
Cas 2 : «Flight reservation system» (Böhnlein et al. 1999)	1/1	2	0	2/2	4/4
Cas 3 : « Hospital » (Golfarelli et al. 1998a)	1/1	0	0	2/4	7/10

TAB. 4 – *Evaluation des résultats de notre approche (* E : Elevé ; F : Faible).*

Cette étude nous a mené au constat suivant :

- Notre méthode est capable d'extraire tous les faits qu'une analyse ascendante peut extraire (taux de couverture 100 %).
- A l'exception des mesures calculables, toutes les mesures sont couvertes.
- Les niveaux hiérarchiques que génère notre méthode sont très détaillés et les hiérarchies obtenues incluent les attributs de type booléen.
- Il génère également d'autres schémas pertinents qui ne correspondaient pas à des besoins projetés par les auteurs (cas des analyses descendantes).
- A l'exception des dimensions construites sur les clés des entités vides non liées à d'autres entités (figures 5 et 6), les dimensions que nous extrayons couvrent toutes les dimensions possibles pour un fait donné.

Sur le plan pratique un outil supportant notre approche est en cours de finalisation. En effet, nous comptons aboutir à un produit qui aide les petites et moyennes entreprises à se doter de leur propre système décisionnel avec des coûts de conception raisonnables.

Références

Böhnlein, M., Ulbrich-vom Ende.A (1999). *Deriving Initial Data Warehouse Structures from the Conceptual Data Models of the Underlying Operational Information Systems.*

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

Boussaid, O. (2006). *Evolution de l'entreposage des données complexes*, Memoire de HDR, université lumière Lyon2.

Cabibbo, L., and Torlone, R. (1998). *A Logical Approach to Multidimensional Databases*. Conference on Extended Database Technology, Valencia, Spain pp. 187-197.

Golfarelli, M., Maio, D., and Rizzi, S. (1998a). *Conceptual Design of Data Warehouses from E/R Schemas.* Conference on System Sciences,Vol. VII, Kona, Hawaii.

Golfarelli, M., Maio, D., and Rizzi, S. (1998b). *The dimensional fact model : a conceptual model for data warehouses.* International Journal of Cooperative Information Systems 7, 2-3 (1998), 215-247.

Ghozzi, F. (2004), *Conception et manipulation des bases de données dimensionnelles à contraintes*, Thèse de Doctorat, Université Paul Sabatier, France.

Hachaichi, Y., Feki, J., (2006). *Heuristiques de construction de MD à partir d'une source OLTP relationnelle*, Atelier des Systèmes Décisionnels (ASD 06), Agadir, Maroc.

Kimball, R. (1997). *The Data Warehouse Toolkit*, John Wiley and Sons, Inc.

Mazón, J.-N., Trujillo, J. (2007). *An MDA approach for the development of data warehouses*, Decisional Support System.

Moody, L.D., and Kortink, M.A.R. (2000). *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.

Pang, C., Taylor, K., Zhang, X. et Cameron, M. (2004). *Generating Multidimensional Schemata from Relational Aggregation Queries.* LNCS 3306, pp. 584–589,

Phipps, C., Davis, K. (2002). Automating data warehouse conceptual schema design and evaluation. DMDW'02, Canada.

Soussi, A., Feki, J., Gargouri. F (2005). *Approche semi-automatisée de conception de schémas multidimensionnels valides*, EDA 05, Revue RNTI vol B-1, P71-90.

Summary

To assist the decisional designer, we present a bottom-up method for the construction of data mart star schema starting from a relational data source. For this, we study the structure of relations and we propose a classification of relations into *R-relationships* and *R-entities* those allow construction of facts and dimensions respectively. Our method has the merit of being independent of the semantics of the source database. Indeed, it exploits the constraints of primary and referential keys to extract the multidimensional concepts and assigns a level of pertinence to each extracted concept.