Opérateurs OLAP pour des cubes d'objets complexes: construction, visualisation et analyse

Doulkifli Boukraâ*, Omar Boussaïd**, Fadila Bentayeb**

*Ecole Nationale Supérieure d'Informatique, Oued-Smar, Alger
d_boukraa@esi.dz
** Université Lumière – Lyon 2, 5 avenue Pierre Mendès-France, 69676 Bron Cedex
omar.boussaid@univ-lyon2.fr; Fadila.Bentayeb@univ-lyon2.fr

Résumé. La modélisation multidimensionnelle est aujourd'hui reconnue comme reflétant le mieux la vision des décideurs sur les données à analyser. Cependant, les modèles multidimensionnels classiques ont été pensés pour traiter des données numériques ou symboliques mais échouent dès lors qu'il s'agit de données complexes. Les opérateurs d'analyse en ligne (OLAP) classiques sont alors à redéfinir dans le cadre de données complexes, voire d'autres sont à créer. Dans ce papier, nous proposons deux familles d'opérateurs OLAP pour manipuler un modèle multidimensionnel d'objets complexes que nous avons proposé. La première famille d'opérateurs permet la construction de nouveaux cubes complexes à partir du schéma multidimensionnel de l'entrepôt ou à partir de cubes existants. La deuxième famille d'opérateurs permet de visualiser et d'analyser les données des cubes complexes.

1 Introduction

1.1 Contexte et motivation

La modélisation multidimensionnelle est aujourd'hui reconnue comme reflétant le mieux la vision des décideurs sur les données à analyser. En témoigne la diversité des modèles multidimensionnels proposés dans la littérature (Abelló et al, 2006; Blaschka et al, 1998). Les modèles multidimensionnels sont souvent accompagnés d'opérateurs pour la manipulation des données. Une étude d'une dizaine d'algèbres montre la diversité des opérations ainsi que leur convergence vers un ensemble d'opérateurs pouvant représenter un cadre de référence pour les algèbres OLAP (Romero and Abelló, 2007).

Le besoin d'algèbre OLAP de référence est d'autant plus accentué qu'il s'agit d'entreposer et d'analyses des données non-conventionnelles ou des données complexes. Les opérateurs OLAP classiques qui s'appliquent à des données numériques sont alors à redéfinir dans le cadre de données complexes, voire d'autres sont à créer. Dans ce contexte, nous avons défini un modèle multidimensionnel pour les données complexes (Boussaïd and Boukraâ, 2008; Boukraâ et al, 2009) basé sur le concept d'objet complexe (Boussaïd et al, 2007). Ce modèle répond à des exigences de modélisation que les modèles existants ne satis-

font pas ou satisfont partiellement, comme la complexité des sujets et des axes d'analyse et la prise en compte de deux types de hiérarchies (d'objets et d'attributs). Ensuite, afin de pouvoir manipuler le modèle proposé, nous avons défini un ensemble d'opérateurs permettant de construire des cubes à partir du modèle multidimensionnel (Boukraâ et al, 2010). Dans ce papier, nous étendons cet ensemble d'opérateurs à des opérateurs de manipulation de cubes qui permettent de créer de nouveaux cubes à partir de cubes existants. En plus, nous définissons des opérateurs de visualisation des données de cubes complexes et des opérateurs d'analyse permettant d'effectuer des sélections et des forages.

1.2 Travaux similaires

Ces dernières années, le domaine des entrepôts de données et de l'OLAP a été marqué par la croissance des travaux traitant des données complexes. Une première catégorie de travaux traite de l'intégration des données provenant du Web dans le processus décisionnel (Bhwomick et al., 2003; Xylème, 2001). Une deuxième catégorie de travaux concerne l'entreposage de données de différentes structures dont des données non structurées (Inokuchi and Takeda, 2007; Keith et al, 2006) et des données semi-structurée, représentées notamment en XML (Golfarelli et al, 2001; Vrdoljak et al, 2003; Park et al, 2005). Dans une troisième catégorie de travaux, l'intérêt porte sur la prise en compte de nouveaux formats de données dans les entrepôts de données, notamment les images dans les applications médicales (Wong, 2001) ou dans les applications à caractère spatio-temporel (Bimonte et al, 2006; Gómez et al, 2009). Une dernière catégorie traite d'autres aspects de la complexité des données, comme la temporalité (Teste, 2000; Pedersen and Jensen, 1999; Kondratas and Timko, 2007) et l'incertitude (Pedersen and Jensen, 1999). Sur le plan de la modélisation multidimensionnelle, de nouvelles représentations sont proposées pour les concepts multidimensionnels (dimension, hiérarchie, niveau, fait, mesure, attribut). Dans ce qui suit, nous catégorisons ces travaux par type de concept. Pour ce qui est des dimensions et des hiérarchies, la majorité des travaux modélisent une dimension comme un ensemble d'attributs organisés sous forme d'arbres (Golfarelli et al, 2001; Vrdoljak et al, 2003), de graphes (Inokuchi and Takeda, 2007) ou, comme c'est le cas de beaucoup de travaux, en classes (UML) d'objets (Jensen et al, 2001; Khrouf et Soulé-Dupuy; 2001; Lujàn-Mora, 2002; Trujillo and Palomar, 1998). Dans ces travaux, une dimension est représentée soit par une seule classe d'objets ou par plusieurs classes reliées entre elles. D'autres travaux appliquent des vues sur les faits en les matérialisant (Park et al, 2001) ou pas (Nassis, et al, 2004). L'organisation hiérarchique des dimensions est traduite par des liens entre les sommets dans les modèles d'arbres ou entre les nœuds dans les modèles de graphes. Dans les modèles objet, les hiérarchies sont représentées par des agrégations (Jensen et al, 2001; Khrouf et Dupuy, 2001), des associations ou par des liens d'héritage (Lujàn-Mora, 2002). La représentation des faits, quant à elle, diffère également selon le modèle utilisé. Dans (Golfarelli et al, 2001; Vrdoljak et al, 2003), il s'agit de sommets appartenant à l'arbre du modèle multidimensionnel. Dans les modèles objet, il s'agit soit d'une classe d'objets (Jensen et al, 2001; Khrouf et Dupuy; 2001; Lujàn-Mora, 2002; Trujillo and Palomar, 1998), soit d'un ensemble de classes d'objets définissant un contexte complet d'analyse (Nassis, 2004). Les mesures décrites par les faits sont représentées, comme dans les modèles classiques, par des attributs dont certains peuvent être dérivés. Les valeurs des mesures peuvent être stockées ou calculées à la volée à l'aide de fonctions adéquates lors des opérations d'analyse. Notons qu'au-delà de leur représentation, la nature des mesures dépend du domaine d'application et des spécificités des données. Ain-

si, on trouve des mesures textuelles (Ravat et al, 2007), des mesures géographiques (Bedard, 2005)… Enfin, aux mesures sont associés des opérateurs d'agrégation. Dans certains travaux, l'analyse est précédée d'une phase de construction de la structure multidimensionnelle sur laquelle elle porte (Ravat et al, 2006), ou de définition des mesures à analyser (Park et al, 2005). Dans certains domaines d'application, des opérations spécifiques à la nature des données sont proposées, comme le forage spatial pour les données spatio-temporelles (Bédard, 2005). Ce survol des travaux permet de constater la diversité des modèles multidimensionnels proposés pour tenir compte des aspects de complexité des données. Dans ce contexte, la modélisation objet est pertinente bien qu'il n'existe pas de consensus sur la représentation des concepts multidimensionnels. En outre, dans chacun des travaux, la complexité des données est abordée partiellement à travers quelques aspects (format, structure…). Or, une meilleure prise de décision nécessite d'aborder les mêmes données selon différents aspects simultanément. Par exemple, l'établissement d'un diagnostic médical peut nécessiter d'analyser simultanément des données numériques, des radiologies, des rapports, etc. Il en ressort la nécessité d'un modèle multidimensionnel permettant d'intégrer différents aspects de la complexité ainsi que d'opérateurs OLAP permettant de manipuler le modèle multidimensionnel.

1.3 Objectifs et contributions

L'objectif de notre travail est de proposer un cadre formel pour l'entreposage, la visualisation et l'analyse des données complexes. Dans un travail récent, nous avons proposé une solution d'intégration de données complexes en terme de modèle multidimensionnel d'objets complexes (Boussaïd and Boukraâ, 2008). Dans ce papier, notre contribution majeure porte sur les éléments suivants: (1) Définition d'opérateurs de construction de cubes d'objets complexes. (2) Définition d'opérateurs de visualisation et d'analyse.

Le reste de ce papier est organisé comme suit. Dans la section 2, nous rappelons les concepts et les définitions de notre modèle. Ensuite, nous exposons dans la section 3 les opérateurs OLAP pour la construction de cubes complexes, la visualisation et l'analyse. Dans la section 4, nous décrivons quelques éléments d'implémentation en cours du modèle et des opérateurs. Enfin, nous concluons et nous présentons des perspectives dans la section 5.

2 Rappel du modèle multidimensionnel d'objets complexes

Dans cette section, nous rappelons les principaux concepts du modèle multidimensionnel d'objets complexes. Plus de détails se trouvent dans (Boussaïd and Boukraâ, 2008).

2.1 Définitions des concepts

2.1.1 L'objet complexe

Le concept d'objet complexe (OC) a été proposé par Boussaïd et al (2007) pour l'intégration de données complexes. Selon les auteurs, un OC est une entité abstraite ou physique composée d'un ou de plusieurs sous-documents. Chaque sous-document représente du texte simple ou balisé, une vue relationnelle, une image ou des données temporelles (son, vidéo). Un OC décrit à la base des caractéristiques de bas niveau comme les couleurs d'une

image ou la durée d'une séquence audio. Néanmoins, le modèle d'un OC peut être étendu pour décrire la sémantique portée par les données, comme par exemple le contenu d'une image. Dans le modèle multidimensionnel que nous avons proposé, nous avons utilisé le concept d'OC pour représenter les sujets et les axes d'analyse. En outre, nous avons abstrait l'OC complexe sous la forme d'un ensemble d'attributs, dont un identifiant, et de relations entre attributs. Cependant, au stade actuel de la modélisation, nous nous sommes limités aux seules relations qui organisent les attributs en hiérarchies d'attributs comme cela sera décrit plus loin.

Définition 1. Un OC est un couple $Obj = (ID^{Obj}, SA^{Obj})$ où ID^{Obj} représente l'identifiant de l'objet et $SA^{Obj} = \{A_i^{Obj} / i \in N\}$ représente l'ensemble de ses attributs.

2.1.2 La relation complexe

Une relation complexe (RC) est un lien explicite entre deux OC. Les relations entre objets peuvent représenter des associations, des compositions, des spécialisation/généralisation, etc. Une RC est caractérisée par son nom et par les deux OC qu'elle relie.

Définition 2. Une RC est un couple $R=(Obj_s^R, Obj_t^R)$ où Obj_s^R représente l'objet source de R et Obj_t^R représente son objet cible.

2.1.3 La hiérarchie d'attributs

Une hiérarchie d'attributs (HA) est une relation définie entre un sous-ensemble d'attributs d'un OC en les ordonnant selon leur degré de granularité. Elle est caractérisée par un nom ainsi que par l'ensemble ordonné de ses attributs, chacun étant associé à un niveau au sein de la hiérarchie. Nous supposons que l'attribut le plus fin possède le niveau 0.

Définition 3. Une HA est notée $AH^{Obj} = \{A_i^{AH}\}$ avec $A_i^{AH} \in SA^{Obj} \cup \{ID^{Obj}\} \cup \{All^A\}$ où All^A désigne un attribut factice nécessaire à la formalisation des opérations d'analyse et possédant la plus faible granularité. En outre, nous définissons une fonction $Level_A(A_i^{AH}, AH^{Obj})$ qui retourne le niveau de chaque attribut au sein de la HA. Enfin, nous désignons par $AttHObj(AH^{Obj}) = Obj$ la fonction qui associe la HA AH^{Obj} à l'objet complexe Obj.

2.1.4 La hiérarchie d'objets

Une hiérarchie d'objets (HO) est similaire à une HA mais elle est définie entre plusieurs OC plutôt qu'entre les attributs d'un seul objet. Une HO ordonne les OC selon leur degré de granularité. Une HO est caractérisée par son nom et par l'ensemble ordonné de ses OC, chacun étant associé à un niveau au sein de la hiérarchie. Aussi, nous supposons que l'objet le plus fin possède le niveau 0.

Définition 4. Une HO est notée $OH = \{Obj_i / i \in N\} \cup \{All^{obj}\}$ où All^{Obj} représente un objet factice possédant la plus faible granularité et jouant un rôle similaire à celui de l'attribut All^A. En outre, nous définissons une fonction $Level_{Obj}(Obj, OH)$ qui retourne le niveau de chaque OC au sein de HO.

Remarque. Au stade actuel des travaux, les hiérarchies supportées par notre modèle sont des hiérarchies parallèles et multiples. Les HA et HO peuvent être comparées respectivement aux hiérarchies des schémas multidimensionnels en étoile et en flocons de neige. Toutefois, dans ces schémas, le choix du type de la hiérarchie répond à un souci d'optimisation (de normalisation). Par contre, dans notre modèle, nous modélisons ces deux types d'hiérarchies d'un point de vue conceptuel. En d'autres termes, une HO ne peut pas être transformée en HA par fusion de ses membres en un seul objet. Une telle fusion engendrerait la perte de l'existence propre de chaque OC et la réduction du nombre de cubes qu'il est possible de construire. De même, une HA ne peut pas être transformée en HO en éclatant chaque OC en sous-objets, ce qui produirait beaucoup de sous-objets et un schéma multidimensionnel difficile à gérer.

2.1.5 Le schéma multidimensionnel

Le schéma multidimensionnel d'OC est composé des éléments suivants: (1) l'ensemble des OC, (2) l'ensemble des RC, (3) l'ensemble des HA et (4) l'ensemble des HO.

Définition 5. Le schéma multidimensionnel d'OC est noté $SCM = (SO, SR, SAH, SOH)$ où $SO = \{Obj_i / i \in N\}$, $SR = \{R_j / j \in N\}$, $SAH = \{AH_k / k \in N\}$ et $SOH = \{OH_m / m \in N\}$.

2.2 Exemple

Un laboratoire de recherche veut entreposer les données concernant sa production scientifique pour répondre à différents besoins d'analyse, comme par exemple (1) l'évaluation de la qualité des publications selon différents critères comme les notes qui leur sont attribuées, (2) l'analyse de la qualité de la production d'un auteur à travers la fréquence de ses publications. Les données du domaine des publications scientifiques sont des données complexes. Elles sont publiées dans de multiples sources (ex: DBLP[1], PubZone[2]), elles peuvent avoir des formats différents (ex: les images contenues dans les sites web des conférences), elles peuvent renfermer diverses structures (ex: les fichiers de présentation des publications sont semi-structurés), etc. Afin de répondre aux différents besoins d'analyse, ces données sont alors modélisées selon le schéma multidimensionnel suivant.

1. Les objets: *Publication*, *Author*, *Proceedings*, *Conference*, *Journal_number*, *Journal_volume*, *Journal*, *Date*. Quant aux attributs, nous nous contentons de citer des exemples d'attributs de l'objet *Publication*, comme *title*, *pages*, *keyword*, *type* ainsi que l'identifiant *publication_id*.
2. Les relations: *Authored_by* entre *Publication* et *Author*; *Date_pub* entre *Publication* et *Date*; *Publi_conf* entre *Publication* et *Proceedings*; *Publi_journal* entre *Publication* et *Journal_number*.
3. Les hiérarchies d'attributs: *H_pub* associée à *Publication* et composée des attributs *publication_id* et *type* et *H_time* associée à *Date* et composée des attributs *date_id*, *month* et *year*.
4. Les hiérarchies d'objets: *H_conf* composée des objets *Proceedings* et *Conference* et la hiérarchie *H_journal* composée des objets *Journal_number*, *Journal_volume* and *Journal*.

[1] http://dblp.uni-trier.de
[2] http://www.pubzone.org

La figure 1 illustre la structure du schéma dimensionnel décrit ci-haut.

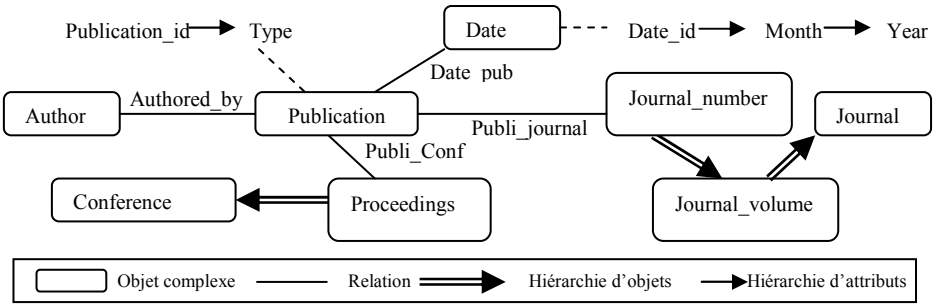

Fig. 1 – *Exemple de schéma multidimensionnel des données complexes*

3 Les opérateurs OLAP

Dans notre modèle, le schéma multidimensionnel est, par définition, indépendant de tout contexte d'analyse. Autrement dit, il n'existe pas d'axe et de sujet d'analyse prédéfinis. Ainsi, nous traitons les dimensions et mesures de manière symétrique (Pedersen and Jensen, 1999). Les axes et le sujet d'analyse sont alors définis au moment de l'analyse en appliquant des opérations de projection sur un ensemble des composants du schéma multidimensionnel. Ces opérations de projection produisent une nouvelle structure appelée cube d'objets complexes. Les données du cubes peuvent alors être matérialisées et visualisées afin d'effectuer des analyses. En outre, la matérialisation des données d'un cube permet de construire de nouveaux cubes à partir de cubes existants. Dans ce qui suit, nous présentons les opérations de construction de cubes et de visualisation.

3.1 Opérateurs de construction de cubes

3.1.1 Projection cubique

L'objectif de cette opération est de construire un cube complexe à partir du schéma multidimensionnel. Dans (Boukraâ et al, 2010), nous avons défini cette opération comme étant composée de plusieurs opérateurs consécutifs. Dans ce papier, nous fusionnons ces opérateurs en un seul afin de rendre homogènes les résultats de chaque opération, i.e. sous forme de cubes. La projection cubique consiste à projeter le schéma multidimensionnel sur un objet complexe pour lui faire jouer le rôle de sujet d'analyse (fait). Cette projection produit une structure de cube et entraine la projection des éléments suivants:

1. un objet fait représentant le sujet d'analyse;
2. un ensemble de mesures de l'objet-fait. Chaque mesure est caractérisée par un nom et est associée à
   (a) un attribut du fait contenant les valeurs de base de la mesure;
   (b) une fonction d'agrégation des valeurs de l'attribut associé à la mesure;
   (c) un ensemble de relations par rapport auxquelles les valeurs de l'attribut associé à la mesure peuvent être agrégées simultanément;

3. l'ensemble des relations reliant le fait aux autres objets. Autrement dit, sont exclues les relations reliant deux objets si aucun d'eux ne joue le rôle de fait.
4. l'ensemble des objets reliés directement au fait;
5. l'ensemble des hiérarchies d'objets contenant les objets projetés en 4;
6. l'ensemble des hiérarchies d'attributs associés aux objets projetés en 4 et 5.

Une dimension est composée de l'ensemble des membres des hiérarchies d'objets auquel appartient l'objet directement relié au fait. Notons que lors de la projection cubique, les hiérarchies d'objets (respectivement les hiérarchies d'attributs) subiront une réduction de leurs membres de sorte à ce que le membre de plus fine granularité soit l'objet directement relié au fait (respectivement l'identifiant de l'objet associé à la hiérarchie d'attributs).

Définition 6. Soit $SCM = (SO, SR, SAH, SOH)$ un schéma multidimensionnel et $SAF = \{af_i / i \in N\}$ un ensemble de fonctions d'agrégation. La projection cubique de *SCM* sur un objet *Obj* est notée $\Pi_C\ Obj\ (SCM) = C = (F, SM, SR^C, SD, SAH^C, SOH^C)$, où:

- F représente l'objet fait avec $F \in SO$;
- SM représente l'ensemble des mesures avec $SM = \{M_i / i \in N\}$ où M_i représente une mesure. En outre, nous définissons les trois fonctions suivantes: (1) la fonction *AttM qui* associe la mesure M_i à un attribut $A^{Mi} \in \{ID^F\} \cup SA^F$, (2) la fonction *AggFun* qui associe la mesure M_i à une fonction d'agrégation $af^{Mi} \in SAF$ et (3) la fonction *AggRel* qui associe la mesure M_i à un ensemble de relations $SR^{Mi} \subseteq SR^C$;
- $SR^C = \{R_i^C / i \in N\}$ représente l'ensemble des relations du cube où $SR^C \subseteq SR$;
- $SD = \{D_j / j \in N\}$ représente l'ensemble des objets membres des dimensions où $SD \subseteq SO$;
- $SOH^C = \{AH_k^C / k \in N\}$ représente l'ensemble des hiérarchies d'objets réduites;
- $SAH^C = \{OH_m^C / m \in N\}$ représente l'ensemble des hiérarchies d'attributs réduites.

Exemple. Dans le schéma multidimensionnel précédent (Fig. 1), noté *SCM_pub*, l'utilisateur veut analyser le maximum des notes des publications par auteur et par année et leurs principaux mots-clés par auteur et par conférence. Le cube correspondant à ce besoin est obtenu par projection cubique de *SCM_pub* sur l'objet *Publication*. Les autres éléments du modèle sont projetés conséquemment.

On écrit $C\_pub = \Pi_C\ publication\ (SCM\_pub) = C\_pub = (F, SM, SR^{C\_pub}, SD, SAH^{C\_pub}, SOH^{C\_pub})$ où $F = Publication$, $SM = \{max\_rating, top\_keyword\}$ où $AttM(max\_rating) = Rating$, $AggRel(max\_rating) = \{Authored\_by, Date\_Pub\}$, $AggFun(max\_rating) = max$, $AttM(top\_keyword) = keyword$, $AggRel(top\_keyword) = (Authored\_by, Publi\_conf)$, $AggFun(top\_keyword) = top\_keyword$, $SR^{C\_pub} = \{Authored\_by, Date\_pub, Publi\_journal, Publi\_conf\}$, $SD = \{Time, Author, Proceeding, Conference, Journal\_number, Journal\_volume, Journal\}$, $SAH^{C\_pub} = \{H\_time\}$ et $SOH^{C\_pub} = \{H\_conf, H\_journal\}$.

3.1.2 Construction de cube à partir de cubes existants

Les opérations de construction de nouveaux cubes à partir de cubes existants se déclinent en deux ensembles: (1) des opérations liées à la structure (c.f. TAB. 1) qui permettent de modifier la structure des cubes existants en ajoutant ou en supprimant des composants et (2) des opérations liées aux données (c.f. TAB. 2) qui permettent d'obtenir des cubes de mêmes

structures mais contenant des données différentes. Remarquons que dans le premier tableau, les opérations de projection sont dérivées des opérations de base, i.e. d'ajout et de suppression. Les opérations de projection permettent alors de simplifier les expressions combinant plusieurs ajouts et/ou suppressions.

Opération	Définition
Ajout / suppression d'une relation	$ADD_R(C, R_+) \mid REM_R(C, R_-^C)$
Ajout / suppression d'une hiérarchie d'attributs	$ADD_{AH}(C, AH_+) \mid REM_{AH}(C, AH_-^C)$
Ajout / suppression d'une hiérarchie d'objets	$ADD_{OH}(C, OH_+) \mid REM_{OH}(C, OH_-^C)$
Ajout / suppression d'une mesure	$ADD_M(C, M_+) \mid REM_M(C, M_-)$
Projection dimensionnelle	$\Pi_D R_n, R_{n+1}, \ldots, R_P(C)$
Projection hiérarchique d'attributs	$\Pi_{AH} AH_n, AH_{n+1}, \ldots, AH_P(C)$
Projection hiérarchique d'objets	$\Pi_{OH} OH_n, OH_{n+1}, \ldots, OH_P(C)$
Projection de mesures	$\Pi_M M_n, M_{n+1}, \ldots, M_P(C)$

TAB. 1 – *Opérateurs liés à la structure pour la construction de cube complexes*

Opération	Définition
Sélection des données d'un cube selon un prédicat portant sur un objet	$\delta_C\ C(P(Obj_\delta))$
Union des données de deux cubes	$C_1 \cup_C C_2$
Différence entre deux cubes sur la base d'un objet	$C_1 -_C C_2(Obj_-)$
Intersection des données de deux cubes sur la base d'un objet	$C_1 \cap_C C_2(Obj_\cap)$

TAB. 2 – *Opérateurs liés aux données pour la construction de cubes complexes*

Exemple. Soit le cube *C_pub* de l'exemple précédent. Supposons que l'utilisateur veut créer un cube pour analyser la seule mesure *max_rating* par auteur et par date. Cette opération s'exprime comme suit $C\_pub_1 = REM_{OH}(REM_{OH}(REM_R(REM_R(REM_M(C\_pub, top\_keyword), Publi\_Journal), Publi\_conf), H\_conf), H\_journal)$. En se basant sur C_pub_1, l'utilisateur peut remplacer *max_rating* par *top_keyword* et analyser cette dernière par auteur et par journal. On peut alors écrire $C\_pub_2 = ADD_{OH}(ADD_R(REM_R(REM_M(ADD_M(C\_pub_1, top\_keyword), max\_rating), Date\_pub), Publi\_journal), H\_journal)$.

En se basant sur C_pub_1, l'utilisateur peut créer deux cubes C_1 et C_2 contenant respectivement les publications dont le titre contient le mot 'database' et celles écrites par des auteurs dont le nom de famille commence par 'A'. On écrit respectivement $C_1 = \delta_C\ C\_pub_1$ (*Contains* (*Publication.Title*, 'database')) et $C_2 = \delta_C\ C\_pub_1$ (*FirstLetter* (*Author.FamilyName* = 'A')). En se basant sur C_1 and C_2, l'utilisateur peut créer les cubes contenant les données suivantes. (1) Les publications dont le titre contient le mot 'database' et rédigées, entre autres, par des auteurs dont le nom de famille commence par 'A': $C_1 \cap_C C_2$ (*Publication*). (2) Les auteurs dont le nom de famille commence par 'A' et ayant écrit, entre autres, des publications dont le titre contient le mot 'database': $C_1 \cap_C C_2$ (*Author*). (3) Les publications dont le titre contient le mot 'database' ou rédigées, entre autres, par des auteurs dont le nom de famille commence par 'A': $C_1 \cup_C C_2$. (4) Les publications dont le titre contient le mot 'database' mais n'ayant pas été écrites par des auteurs dont le nom de famille commence par 'A': $C_1 -_C C_2$ (*Publication*).

3.2 Opérateurs de visualisation de cubes

Afin d'analyser les données d'un cube d'objets complexes, nous définissons un espace multidimensionnel de visualisation composé d'un ensemble d'éléments d'observation appelés dimensions de la vue, plus d'un ensemble de mesures décrites par le fait. Chaque dimension de la vue correspond à une relation appartenant au cube complexe. Le principe de la vue est de rester abstrait par rapport à toute solution concrète. Ainsi, il convient à l'utilisateur de choisir la solution de visualisation adéquate selon la nature des données à analyser.

3.2.1 Projection de vue sur un cube

Une vue sur un cube d'objets complexes est obtenue par une opération de projection du cube sur les éléments suivants:

- un fait de la vue (FV) qui correspond au fait du cube complexe. Un FV possède un nom et est décrit par un ensemble de caractéristiques. Une caractéristique est un attribut du fait que l'utilisateur souhaite afficher sur la vue;
- un ensemble de mesures à afficher, choisies parmi les mesures du cube;
- une dimension de la vue (DV) par relation du cube et correspondant à une dimension du cube. Une DV possède un nom et est décrites par un ensemble de caractéristiques. Une caractéristique d'une DV est un attribut appartenant à l'un des objets composant la dimension du cube correspondante.

En outre, afin de définir les niveaux d'agrégation des valeurs des mesures, nous associons la DV à deux éléments.

- Un OC appartenant à une HO de la dimension du cube. A défaut de HO, la DV est associée à l'objet qui est directement lié au fait. On note cet objet OA ;
- Un attribut appartenant à une HA liée à OA. A défaut de HA, la DV est associée à l'identifiant de OA. On note cet attribut AA.

Les agrégations des valeurs des mesures sont calculées par rapport à OA puis par rapport à AA. Les caractéristiques à afficher pour la DV dépendent de OA et de AA.

La notion de vue sur un cube d'objets complexes est illustrée en figure 2 où FV est représenté par deux caractéristiques A_1^{FV} et A_2^{FV} qui décrivent les mesures M_1 et M_2. Ces mesures sont observées par rapport aux dimensions de la vue DV_1, DV_2 et DV_3 lesquelles correspondent respectivement aux relations R_1, R_2 et R_3 et sont décrites par les caractéristiques $A_1^{DV_1}$, $A_2^{DV_1}$, $A_1^{DV_2}$, $A_2^{DV_2}$, $A_3^{DV_2}$ et $A_1^{DV_3}$.

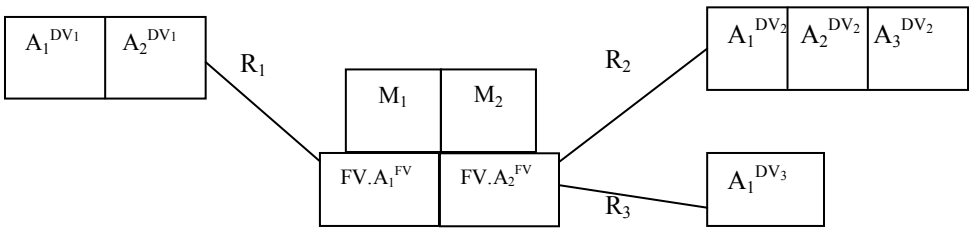

Fig. 2 – *Illustration de la notion de vue sur un cube complexe*

Définition 5. Soit $C = (F, SM, SR^C, SD, SAH^C, SOH^C)$ le schéma d'un cube complexe. L'opération de projection de vue sur C est notée $\Pi_V(C) = V^C = (F^V, SM^V, SD^V)$ où
- $SM^V = \{M_i^V / i \in N\} \subseteq SM$ représente l'ensemble des mesures affichées;
- $SD^V = \{D^{VRj} / j \in N\}$ où D^{VRj} désigne une dimension de la vue correspondant à la relation R_j de C. En outre, nous définissons la fonction $ViewObj(D^{VRj}) = VO^{VRj}$ qui associe D^{VRj} à un objet appartenant à la dimension du cube qui correspond à la relation R_j. Enfin, nous définissons la fonction $ViewAtt(D^{VRj}) = VA^{VRj}$ qui associe D^{VRj} à un attribut appartenant à l'une des hiérarchies d'attributs associées éventuellement à VO^{VRj}.
- F^V représente le fait de la vue avec $F^V = \{A_p^{FV} / p \in N\}$ *où* $F^V \subseteq \{SA^F\} \cup \{ID^F\}$ si la vue affiche les données de bases du cube (aucune agrégation n'est appliquée) et F^V est positionné à une constante *Undefined* s'il existe au moins une mesure de la vue pour laquelle les valeurs sont agrégées. En effet, les caractéristiques des faits décrivent les valeurs détaillées de toutes les mesures. Ainsi, dès que les valeurs détaillées d'une mesure sont agrégées, les valeurs des caractéristiques des faits deviennent indéfinies.

Exemple. L'utilisateur veut analyser *max_rating* par auteur et par année et *top_keyword* par auteur et par *proceedings*. Il veut afficher les titres des publications, les noms et prénoms des auteurs et les titres des proceedings. La vue correspondant à cette analyse est illustrée en figure 3. L'ensemble des caractéristiques du fait est indéfini car les valeurs de la mesure *max_rating* sont agrégées au niveau de l'attribut *année* de la hiérarchie *H_time*.

Formellement, soit C le cube complexe défini par ce contexte d'analyse et soit *V_RK* la vue de cet exemple. Donc $V\_RK = \Pi_V(C) = (F^{V\_RK}, SM^{V\_RK}, SD^{V\_RK})$ avec F^{V_RK} = Undefined, SM^{V_RK} = {*max_rating*, *top_keyword*} et SD^{V_RK} = {*Authors*, *Time*, *Conferences*} où *Authors* = {*Author.fname*, *Author.lname*} avec *ViewObj*(*Authors*) = *Author* et *ViewAtt*(*Authors*) = *Author.Author_id*, *Time* = {*Date.Year*} avec *ViewObj*(*Time*) = *Date*, ViewAtt(*Time*) = *Date.Year*, *Conferences* = {*Proceedings.Name*} avec *ViewObj*(*Conferences*) = *Proceedings* et *ViewAtt*(*Conferences*) = *Proceedings.Proceedings_ID*.

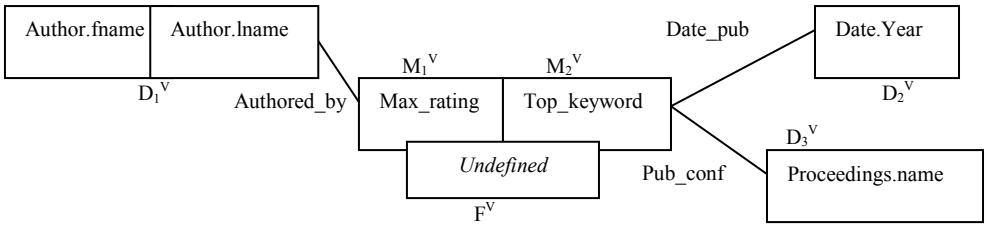

Fig. 3 – *Exemple de vue sur un cube complexe.*

3.2.2 Structuration et restriction des données des vues

Restructurer une vue consiste à ajouter ou supprimer des caractéristiques ou des mesures de l'affichage. Cela permet à l'utilisateur d'avoir plus ou moins d'informations au niveau de la vue. La restriction des données d'une vue est similaire aux opérations de *slice_and_dice* dans les outils OLAP. Elle permet de restreindre les données affichées par la vue selon un prédicat de sélection de données. La restriction peut porter aussi bien sur les valeurs des caractéristiques des faits ou des dimensions que sur les valeurs agrégées ou détaillées des

mesures. Inversement, la levée de restriction permet d'afficher toutes les valeurs de caractéristiques ayant été restreintes. Ces opérations sont résumées dans le tableau 3.

Opération		Définition
Ajout / suppression de mesure		$ADD_{VM}(V^C, M_+) \mid REM_{VM}(V^C, M_-)$
Ajout / suppression de caractéristique	de fait	$ADD_{FF}(V^C, FF_+) \mid REM_{FF}(V^C, FF_-)$
	de dimension	$ADD_{DF}(V^C, DF_+) \mid REM_{DF}(V^C, DF_-)$
Restriction / levée de restriction des données	de fait	$\delta_{DF}V^C(P(DF)) \mid \mu_{DF}V^C(DF)$
	de dimension	$\delta_{FF}V^C(P(FF)) \mid \mu_{FF}V^C(FF)$
	de mesure	$\delta_M V^C(P(M)) \mid \mu_M V^C(M)$

TAB. 3 – *Opérateurs de structuration et de restriction des données des vues*

Exemple. A partir de la vue *V_RK* précédente, l'utilisateur peut par exemple (1) compléter les titres des proceedings par le titre de leurs conférences: $ADD_{DF}(V\_RK, Confrence.Conference.Name)$, (2) supprimer les prénoms des auteurs: $REM_{DF}(V\_RK, Authors.Author.fname)$, (3) afficher seulement les publications de l'année 2000: $\delta_{DF}V^C(Time.Date.year = '2000')$, (4) afficher les groupes de publications dont les maximum des notes sont inférieurs à 3: $\delta_M V^C(max\_rating < 3)$.

3.3 Opérateurs d'analyse

Pour ce qui est des opérations d'analyse, nous nous contentons de présenter les opérations liées à la granularité (forage vers le haut et forage vers le bas). Ces deux opérations s'effectuent sur une vue du cube d'objets complexes. Une opération de forage consiste à changer la vue en cours d'affichage par une nouvelle vue. En particulier, il s'agit, pour une dimension donnée de la vue, d'en modifier l'objet complexe associé ou l'attribut associé par un nouvel objet ou attribut. Les opérations de forage se déclinent en trois variantes tel que résumé dans le tableau 4.

1. Le forage à base d'hiérarchie d'objets: cette opération est applicable si l'OA fait partie d'au moins une HO. Le forage consiste alors à changer l'OA courant par l'objet de niveau directement supérieur (forage vers le haut) ou directement inférieur (forage vers le bas) selon une HO choisie. Les valeurs des mesures sont agrégées ou détaillés conséquemment. L'AA de la nouvelle vue est alors l'identifiant du nouvel OA.
2. Le forage à base d'hiérarchie d'attributs: cette opération est applicable si l'AA fait partie d'au moins une HA. Le forage consiste alors à changer l'AA courant par l'attribut ayant le niveau directement supérieur (forage vers le haut) ou directement inférieur (forage vers le bas) par rapport à une HA choisie. Les valeurs des mesures sont agrégées ou détaillées conséquemment. L'OA de la nouvelle vue reste le même que celui la vue précédente.
3. Le forage sans hiérarchie: cette opération est applicable lorsque l'OA (resp. l'AA) ne fait partie d'aucune HO (resp. d'aucune HA). Dans le cas du forage vers le haut, l'OA doit obligatoirement être celui qui est lié au fait et l'AA doit obligatoirement être l'identifiant de OA. Le forage vers le haut consiste alors à retirer la dimension de la vue. A l'inverse, le forage vers le bas consiste à réintroduire dans la vue une dimension déjà retirée. Formellement, une dimension de vue retirée est associée à l'objet factice All^{Obj} et à l'attribut factice All^{A}.

Opération		Définition
Forage vers le haut	à base d'hiérarchie d'objets	$RollUp_{OH}(V^C, D^{V\,RRU}, OH^C)$
	à base d'hiérarchie d'attributs	$RollUp_{AH}(V^C, D^{V\,RRU}, AH^C)$
	sans hiérarchie	$RollUp_{HL}(V^C, D^{V\,RRU})$
Forage vers le bas	à base d'hiérarchie d'objets	$DrillDown_{OH}(V^C, D^{V\,RRU}, OH^C)$
	à base d'hiérarchies d'attributs	$DrillDown_{AH}(V^C, D^{V\,RRU}, AH^C)$
	sans hiérarchie	$DrillDown_{HL}(V^C, D^{V\,RRU})$

TAB. 4 – *Opérateurs d'analyse liés à la granularité*

Exemple. Reprenons l'exemple de la vue *V_RK*. A partir de cette vue, l'utilisateur peut effectuer les opérations suivantes. (1) Forage vers le haut à base de la hiérarchie d'attributs *H_time* et selon la relation *Date_pub* pour afficher les *max_rating* des auteurs pour toutes les périodes: $RollUp_{AH}$(*V_RK*, *Time*, *H_time*). (2) Forage vers le haut à base de la hiérarchie d'objets *H_conf* pour afficher les *top_keyword* par auteur et par conférence: $RollUp_{OH}$(*V_RK*, *Conferences*, *H_conf*). (3) Forage vers le haut sans hiérarchie pour afficher les *max_rating* par année et les *top_keyword* des proceedings tout auteur confondu: $RollUp_{HL}$(*V_RK*, *Authors*). Les opérations de forage vers le bas consistent à changer les vues résultant des trois opérations précédentes par la vue *V_RK*.

4 Implémentation

Afin de valider notre proposition de schéma multidimensionnel et d'opérateurs OLAP associés, une plateforme d'entreposage et d'analyse est en cours de développement (fig. 4) dont certains modules ont été réalisés. Ainsi, le modèle conceptuel a été traduit en un modèle logique puis physique en utilisant XML. Notre choix d'utiliser XML est justifié par sa capacité de décrire des données de diverses natures et de différents formats, ce qui convient à la description des données complexes. Pour valider fonctionnellement nos propositions, nous utilisons la base de données DBLP disponible en libre téléchargement en format XML[3]. La base DBLP constitue notre source de données de l'entrepôt. Pour la modélisation logique en XML, nous avons développé une algèbre XML décrivant au niveau métadonnées le schéma d'un entrepôt de données et le schéma d'un cube. Ce schéma XML a été instancié en termes de document XML (dblp_xwh.xml) décrivant la structure (métadonnées) de l'entrepôt des références bibliographiques de DBLP. Au niveau des données, nous avons développé des algèbres XML pour chaque classe de composant du modèle multidimensionnel (objet complexe, relation, hiérarchie d'objets, hiérarchie d'attributs). D'autres modules sont en cours de développement, à savoir (1) un module ETL permettant d'obtenir des fichiers XML décrivant les données des différents composants de l'entrepôt, (2) un module de stockage qui permet de stocker les fichiers de l'entrepôt dans une base XML native (eXist) et (3) un module pour la spécification et la génération des cubes XML qui serviront par la suite de base aux modules de construction de cubes, de visualisation et d'analyse dont le développement est inscrit comme travail futur.

[3] http://dblp.uni-trier.de/xml

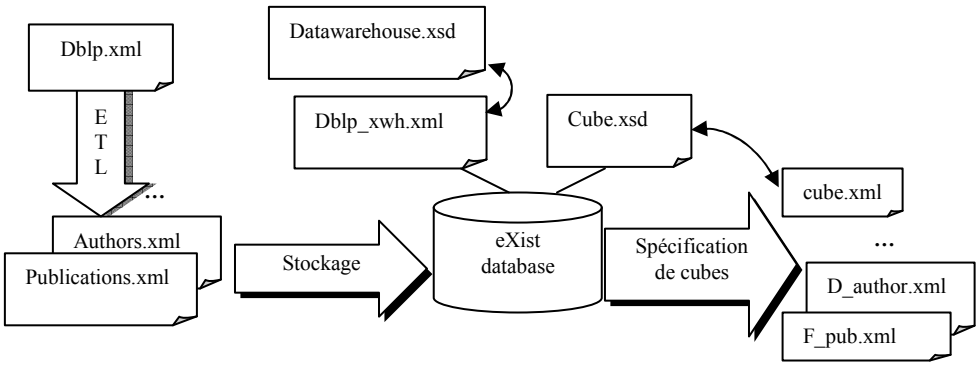

Fig. 4 – *Eléments d'implémentation de la plateforme*

5 Conclusion

Dans ce papier, nous avons rappelé les principaux concepts du modèle multidimensionnel à base d'objets complexes et avons défini un ensemble d'opérateurs OLAP associés. Un premier opérateur de construction permet de faire une projection du schéma multidimensionnel sur un objet complexe et produit un cube. Les opérateurs de construction permettent de construire des cubes d'objets complexes à partir de cubes existants. Les opérateurs de construction orientés structure produisent de nouveaux cubes ayant des structures différentes. Par contre, les opérateurs orientés données produisent des cubes ayant la même structure mais avec un contenu différent. Nous avons également défini des opérateurs de visualisation des données d'un cube. Le premier opérateur permet d'afficher des attributs de faits ou de dimensions et des mesures. Les opérateurs définis sur la vue permettent de restructurer la vue en termes de caractéristiques à afficher ou de restreindre les données affichées. Enfin, les opérateurs d'analyse permettent d'avoir des vues plus ou moins détaillées sur les données.

Les travaux présentés dans cet article ouvrent plusieurs perspectives. En premier lieu, il sera question de considérer des mesures plus complexes, i.e. ne se limitant pas à de simples attributs. En second lieu, il s'agira de prendre en compte les notions d'hiérarchies d'objets et d'hiérarchies d'attributs liées au fait lesquelles sont actuellement exclues de notre modèle. Cela permettra d'ajouter et d'étendre l'opération de forage au sein d'un fait et donc d'éclater et de fusionner la structure des mesures selon qu'il s'agisse de forage vers le bas ou vers le haut. En troisième lieu, il s'agira d'exploiter la structure interne d'un objet complexe au cours de la phase d'analyse OLAP. En effet, notre proposition actuelle consiste à aligner les attributs d'un objet sur un même niveau dans une vue. Nous pensons que l'affichage des relations entre les composants d'un objet permettra de mieux interpréter les valeurs affichées.

Références

Abellò A., J. Samos, and F. Saltor (2006). Yam^2: A Multidimensional Conceptual Model Extending UML. *Inf. Syst.*, 31(6):541-567.

Bédard Y., M.J. Proulx, et S. Rivest (2005). Enrichissement du OLAP pour l'Analyse Géographique: Exemples de Réalisation et Différentes Possibilités Technologiques. In *1ère*

Journée Francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'05), *Lyon*, *France*.

Bhowmick S.S., S.K. Madria, and W.K. Ng (2003). *Web Data Management: A Warehouse Approach*. New York, USA: Springer Verlag.

Bimonte S., A. Tchounikine, and M. Miquel (2006). GeoCube, A Multidimensional Model and Navigation Operators Handling Complex Measures: Application in Spatial OLAP. In *Proceedings the 4th International Conference on Advances in Information Systems*, *Izmir*, *Turkey*, Volume 4243 of *LNCS*, pp. 100-109. Springer.

Blaschka M., C. Sapia, G. Höfling, and B. Dinter (1998). Finding Your Way through Multidimensional Data Models. DEXA Workshop, pp 198-203.

Boukraâ D., O. Boussaïd, F. Bentayeb, and S. Loudcher (2010). OLAP Operators For A Complex Object-Based Multidimensional Model. *International Journal of Business Intelligence and Data Mining* (to appear).

Boukraâ D., R. Ben Messaoud, and O. Boussaïd (2009). *Modeling XML Warehouses For Complex Data: The New Issues*, pp 287-307. Open and Novel Issues in XML Database Applications: Future Directions and Advanced Technologies. Hershey, PA, USA: IGI publishing.

Boussaïd, O., and D. Boukraâ (2008). *Multidimensional Modeling of Complex Data*, pp 1358-1364. Encyclopedia of Data Warehousing and Mining, 2nd Edition. Hershey, PA, USA: IGI Publishing.

Boussaïd O., A. Tanasescu, F. Bentayeb, and J. Darmont (2007), Integration and Dimensional Modelling Approaches for Complex Data Warehousing. *Journal of Global Optimization 35*(4), 571-591.

Golfarelli M., S. Rizzi, and B. Vrdoljak (2001). Data warehouse Design from XML Sources. In *Proceedings of the 3rd ACM International Workshop on Data Warehousing and OLAP (DOLAP'01)*, *Atlanta*, *USA*.

Gómez L. I., B. Kuijpers, B. Moelans, and A. A. Vaisman (2009). A Survey of Spatio-Temporal OLAP. *International Journal of Data Warehousing and Mining*, *5*(3).

Inokuchi A., and K. Takeda (2007). A Method for Online Analytical Processing of Text Data. In M. J. Silva, A. H. F. Laender, R. A. Baeza Yates, D. L. McGuiness, B. Olstad, ∅. H. Olsen, and A. O. Falcão (Eds.), *Proceedings of the 16th Conference on Information and Knowledge Management (CIKM'07), Lisbon, Portugal*, pp. 455-464. ACM.

Jensen M. R., T. H. Møller, and T. B. Pedersen (2001). Specifying OLAP Cubes on XML Data. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management, Virginia, USA*, pp. 101-112. IEEE Computer Society.

Keith S., O. Kaser, and D. Lemire (2006). Analyzing Large Collections of Electronic Text Using OLAP, *CoRR abs/cs/0605127*.

Khrouf K., and C. Soulé-Dupuy (2001). Conception d'Entrepôts de Documents Décisionnels. In *Actes du XIXème Congrès INFORSID*, *Martigny*, *Suisse*, pp. 387-401.

Kondratas E., and I. Timko (2007). CT-OLAP: Temporal Multidimensional Data Model and Algebra for Moving Objects. In I.-Y Song and T. B. Pedersen (Eds.), *10th ACM International Workshop on Data warehousing and OLAP (DOLAP'07)*, *Lisbon*, *Portugal*, pp 81-88. ACM.

Luján-Mora S. (2002). Multidimensional Modeling using UML and XML. In *16th European Conference on Object-Oriented Programming (ECOOP'02)*, *Málaga, Spain*. Volume 2548 of *LNCS*, pp. 48-49. Springer.

Nassis V., R. Rajugan, T. S. Dillon, and W. Rahayu (2004). Conceptual Design of XML Document Warehouses. In Y. Kambayashi, M. K. Mohania, and W. Wöß (Eds.), *Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'04)*, *Zaragoza*, *Spain*. Volume 3181 of *LNCS*, pp. 1-14. Springer.

Park B.K., H. Han, and I-Y. Song (2005). XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. In A. M. Tjoa and J. Trujillo (Eds), *Proceedings of the 7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'05)*, *Copenhagen*, *Denmark*. Volume 3589 of *LNCS*, pp. 32-42. Springer.

Pedersen T. B., and C. S. Jensen (1999). Multidimensional Data Modeling for Complex Data. In *Proceedings of the 15th International Conference on Data Engineering (ICDE'99)*, *Sydney*, *Australia*. pp. 336-345.

Ravat F., O. Teste, and R. Tournier (2007). Analyse Multidimensionnelle de Documents via des Dimensions OLAP. *Document numérique, 10*(2), pp. 85-104.

Ravat F., O. Teste, and G. Zurfluh (2006), Algèbre OLAP et Langage Graphique. In *Actes du XXIVème Congrès INFORSID, Hammamet, Tunisie,* pp. 1039-1054.

Romero O., and A. Abelló (2007). On the Need of a Reference Algebra for OLAP. In I-Y. Song, J. Eder, and T. M. Nguyen (Eds.), *Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'07)*, *Regensburg*, *Germany*. LNCS, pp. 99-110. Springer.

Teste O. (2000), Elaboration d'Entrepôts de Données Complexes. In *Actes du XVIIIème Congrès INFORSID, Lyon, France,* pp. 229-245.

Trujillo J., and M. Palomar (1998). An Object-Oriented Approach to Multidimensional Database Conceptual Modeling. In *Proceedings of the 1st ACM International Workshop on Data warehousing and OLAP (DOLAP'98), Bethesda, Maryland, USA*, pp 16-21. ACM.

Vrdoljak B., M. Banek, and S. Rizzi (2003). Designing Web warehouses from XML Schemas. In In Y. Kambayashi, M. K. Mohania, and W. Wöß (Eds.), *Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery (DaWaK'03)*, *Prague*, *Czech Republic*, Volume 2737 of *LNCS*, pp. 89-98. Springer.

Wong S.T.C., K.S. Jr Hoo, R.C. Knowlton, K.D. Laxer, X. Cao, R.A. Hawkins, W.P. Dillon, and R.L. Arenson (2001). Design and Applications of a Multimodality Image Data Warehouse Framework. *The journal of the American Medical informatics Association*.

Xylème L. (2001). A Dynamic Warehouse for XML Data of the Web. *IEEE Data Eng. Bull. 24*(2): 40-47.

Summary

Nowadays, multidimensional modeling is recognized to best reflect the decision makers' analytical view of data. However, the classical multidimensional models were meant to handle numerical or symbolic data but they fail regarding complex data. Particularly, the classical OLAP operators are to be redefined or new ones will be created in the context of complex data warehousing. In this paper, we propose two families of OLAP operators in order to manipulate a complex object-based multidimensional model proposed previously. The first family of OLAP operators allows constructing complex cubes from the multidimensional model or from existing cubes. The second family allows visualizing and analyzing the complex data cubes.