Modélisation et manipulation des systèmes OLAP : de l'intégration des documents à l'usager

Olivier Teste

MEMOIRE
pour l'obtention de
l'HABILITATION à DIRIGER des RECHERCHES

Spécialité Informatique

Modélisation et manipulation des systèmes OLAP : de l'intégration des documents à l'usager

Olivier Teste

Soutenue le 7 Décembre 2009 devant la commission d'examen :

C. Cauvet,	Professeur à l'Université Aix-Marseille III	Examinateur
C. Chrisment,	Professeur à l'Université Toulouse III	Examinateur
J. Darmont,	Professeur à l'Université Lyon II	Rapporteur
T. Libourel,	Professeur à l'Université Montpellier II	Rapporteur
M. Miquel,	Maître de Conférences (HDR) à l'INSA Lyon	Rapporteur
G. Zurfluh,	Professeur à l'Université Toulouse I	Directeur de recherche

Institut de Recherche en Informatique de Toulouse (UMR5505)
Université Paul Sabatier (Toulouse III) – 118, Route de Narbonne – F-31062 Toulouse cedex 04

à Léa et Manon

Remerciements

Mes remerciements s'adressent d'abord à Claude Chrisment, Responsable de l'équipe SIG (Systèmes d'Informations Généralisées) et Gilles Zurfluh, Directeur de Recherche, pour m'avoir accueilli au sein de l'équipe m'offrant ainsi d'excellentes conditions de travail. Messieurs je tiens ici à vous exprimer ma profonde gratitude pour votre soutien, vos précieux conseils et tous vos encouragements. Sachez également que j'apprécie la liberté d'action et toute la confiance que vous m'accordez, sans oublier les moments de convivialité que vous avez su créer tout au long de ces années.

Je remercie très sincèrement les rapporteurs de ce travail : Jérôme Darmont, Professeur à l'Université Lyon II, Thérèse Libourel, Professeur à l'Université de Montpellier III et Maryvonne Miquel, Maître de Conférences HDR à l'INSA de Lyon. Leurs remarques, leurs commentaires et nos différents échanges m'ont permis d'améliorer ce mémoire et mon regard sur mes activités. Je les remercie pour l'honneur qu'ils me font en participant au jury.

Je tiens à remercier également Corine Cauvet, Professeur à l'Université d'Aix-Marseille III, pour tout l'intérêt qu'elle a manifesté envers mon travail et pour l'honneur qu'elle m'accorde en participant au jury.

Je souhaite remercier Franck Ravat à plus d'un titre. Franck, je tiens d'abord à t'exprimer toute ma reconnaissance pour ton aide précieuse, tes remarques et les innombrables échanges que nous avons et qui m'ont permis d'améliorer ce mémoire. Je souhaite aussi te remercier pour la qualité de notre collaboration et pour la bonne humeur que nous entretenons dans le bureau. Enfin je souhaite te remercier pour m'avoir fait découvrir La Chapelle-Lariveau et pour toutes les conversations plus ou moins sérieuses qui en découlent…

Je n'oublie pas tous les membres d'équipe SIG. En premier lieu, mes fidèles compères Max Chevalier et Gilles Hubert qui par leurs lectures ont contribué à améliorer ce mémoire, et avec qui je poursuis des collaborations diverses et variées… Je remercie Geneviève Pujolle dont la collaboration, les lectures et les remarques ont permis d'améliorer ce mémoire. Je tiens enfin à remercier tous les « sigs » avec qui j'ai collaboré durant ces années, et qui contribuent à maintenir une atmosphère chaleureuse dans notre équipe.

Mes remerciements s'adressent également à tous les collègues « outre-sig » avec qui je travaille. Je pense en particulier au Professeur Pierre Bazex qui m'a fait l'honneur de me confier des responsabilités dans la formation ISI, et avec qui j'ai toujours eu un grand plaisir de collaborer. Je souhaite également remercier chalereusement mon partenaire Michel Tuffery qui m'a fait partager ses connaissances sur Oracle et plus largement, sur l'enseignement des bases de données. Je remercie tous mes collègues d'autres établissements avec qui je collabore. Je pense en particulier aux collègues d'Aix-Marseille, de Castres, et de Pau avec qui j'ai un réel plaisir de collaborer.

Enfin, je ne peux terminer ces quelques lignes sans remercier ma famille qui dans l'ombre m'apporte depuis si longtemps tout le soutien qui m'est nécessaire, et spécialement, mes pensées vont à Magaly.

Table des matières

Table des figures

Table des tableaux

Chapitre 1 - Introduction

Un peu d'histoire…

Dès les années 60, les données informatisées dans les organisations ont pris une importance qui n'a cessé de croître. Les systèmes informatiques gérant ces données sont utilisés essentiellement pour faciliter l'activité quotidienne des organisations et pour soutenir les prises de décision. La démocratisation de la micro-informatique dans les années 80 a permis un important développement de ces systèmes augmentant considérablement les quantités de données[1] [2] informatisées disponibles. Face aux évolutions nombreuses et rapides qui s'imposent aux organisations, la prise de décision est devenue dès les années 90 une activité primordiale nécessitant la mise en place de systèmes dédiés efficaces [Inmon, 1994].

A partir de ces années, les éditeurs de logiciels ont proposé des outils facilitant l'analyse des données pour soutenir les prises de décision. Les tableurs sont probablement les premiers outils qui ont été utilisés pour analyser les données à des fins décisionnelles. Ils ont été complétés par des outils facilitant l'accès aux données pour les décideurs au travers d'interfaces graphiques dédiées au « requêtage » ; citons le logiciel Business Objects qui reste aujourd'hui encore l'un des plus connus. Le développement de systèmes dédiés à la prise de décision a vu naître des outils E.T.L. (« Extract-Transform-Load ») destinés à faciliter l'extraction et la transformation de données décisionnelles. Dès la fin des années 90, les acteurs importants tels que Microsoft, Oracle, IBM, SAP sont intervenus sur ce nouveau marché en faisant évoluer leurs outils et en acquérant de nombreux logiciels spécialisés ; par exemple, SAP vient d'acquérir Business Objects pour 4,8 milliards d'euros. Ils disposent aujourd'hui d'offres complètes intégrant l'ensemble de la chaîne décisionnelle : E.T.L., stockage (S.G.B.D.), restitution et analyse. Cette dernière décennie connaît encore une évolution marquante avec le développement d'une offre issue du monde du logiciel libre (« open source ») qui atteint aujourd'hui une certaine maturité (Talend[3], JPalo[4], Jasper[5]).

Dominée par les outils du marché, l'informatique décisionnelle est depuis le milieu des années 90 un domaine investi par le monde de la recherche au travers des concepts d'entrepôt de données (« data warehouse ») [Widom, 1995] [Chaudhury, *et al.*, 1997] et d'OLAP (« On-Line Analytical Processing ») [Codd, *et al.*, 1993]. D'abord diffusées dans les grandes manifestations en base de données (VLDB, EDBT, ICDT,…), les recherches dans ce domaine ont permis l'émergence sur le plan international de manifestations (DOLAP créé en 1998, DAWAK créé en 1999) et de journaux (IJDWM créé en 2005) spécialisés. Aujourd'hui

[1] Selon le rapport de mission CNRS de Mai 2009 d'Antoine Petit, le volume des données disponibles en 2010 est estimé à 1000 milliards de gigaoctets. Francine Berman, chercheuse au San Diego Supercomputer Center a évalué à 161 exabytes le volume d'informations créées pour la seule année 2006.

[2] Serge Abiteboul indiquait lors de la séance solennelle de l'Académie des sciences, le 16 juin 2009, que « *du catalogue de vente d'Amazon, aux photos de Flickr ou l'encyclopédie Wikipedia, les bases de données sont au cœur des systèmes du web, avec des problèmes d'échelle étonnants, comme les milliards de pages indexées par Google ou les centaines de millions d'utilisateurs de Facebook* ».

[3] http://www.talend.com/

[4] http://www.jpalo.com/

[5] http://www.jaspersoft.com/

une véritable communauté s'est formée. Au niveau francophone, la conférence EDA offre depuis 2005 un cadre annuel de rencontres entre chercheurs et industriels du domaine.

L'entrepôt de données est reconnu comme le cœur du système décisionnel : il intègre et stocke les données issues des différents domaines fonctionnels d'une organisation pour les rendre facilement accessibles aux processus d'analyses décisionnelles. L'entrepôt de données est défini comme « *une collection de données intégrées, orientées sujets, non volatiles, historisées, résumées et disponibles pour l'interrogation et l'analyse* » [Inmon, 1994]. Les recherches ont initialement porté en majorité sur les techniques d'intégration et de stockage des données au sein de l'entrepôt par le développement du mécanisme de vues matérialisées [Gupta, 1995] [Widom, 1995] ; une vue matérialisée est une vue dont les données calculées sont stockées physiquement pour accélérer l'interrogation. De très nombreux travaux ont été réalisés proposant des algorithmes répondant essentiellement à deux problématiques : la maintenance des vues matérialisées [Gupta, 1995] [Zhuge, *et al*., 1995, 1998] [Huyn, 1997] [Zhou, *et al*., 1996] [Mumick, *et al*., 1997] [Quass, *et al*., 1997] [Yang, *et al*., 1998, 2000] [Labio, *et al*., 1999, 2000] ainsi que la sélection d'un ensemble optimal de vues à matérialiser [Harinarayan, *et al*., 1996] [Baralis, *et al*., 1997] [Gupta, *et al*., 1997] [Yang, *et al*., 1997] [Shukla, *et al*., 1998, 2000] [Kotidis, *et al*., 1999] [Theodoratos, *et al*., 1999]. Ces travaux ont été complétés par la définition de nouveaux opérateurs [Gray, *et al*., 1996] [Li, *et al*., 1996] [Agrawal, *et al*., 1997] [Gyssens, *et al*., 1997] [Cabibbo, *et al*., 1998, 2000] [Abello, *et al*., 2002] spécialisés pour un type d'analyses décisionnelles appelé analyses OLAP. La proposition la plus remarquable est certainement celle de l'opérateur « Cube » [Gray, *et al*., 1996] consistant à opérer efficacement des calculs d'agrégations requis par les processus d'analyses OLAP. Dès la fin des années 90 et durant ces années 2000, les recherches se sont orientées sur les aspects plus méthodologiques avec le développement de solutions visant à faciliter l'élaboration des entrepôts : modèles dédiés de représentation des données [Golfarelli, *et al*., 1998] [Lehner, *et al*., 1998] [Sapia, *et al*., 1998] [Datta, *et al*., 1999] [Tryfona, *et al*., 1999] [Hüsemann, *et al*., 2000] [Abello, *et al*., 2002, 2006] [Franconi, *et al*., 2004] [Malinowski, *et al*., 2006, 2008] et démarches de conception adaptées aux entrepôts de données suivant une approche descendante [Kimball, *et al*., 1998] [Tsois, *et al*., 2001] [Prat, *et al*., 2002], ascendante [Golfarelli, *et al*., 1998] [Cabibbo, *et al*., 1998, 2000] [Hüsemann, *et al*., 2000] [Moody, *et al*., 2000] ou même mixte [Bonifati, *et al*., 2001] [Cavero, *et al*., 2001] [Carneiro, *et al*., 2002] [Luján-Mora, *et al*., 2003, 2006].

Partant de la définition d'un entrepôt de données proposée par [Inmon, 1994] où l'entrepôt doit servir à intégrer les données utiles aux décideurs tout en supportant efficacement les analyses, mes travaux de thèse [Teste, 2000a] se sont appuyés sur une dichotomie des espaces de stockage dans le système décisionnel : l'entrepôt de données servant à intégrer les données (« *une collection de données intégrées (…) non volatiles, historisées, résumées (…)* » [Inmon, 1994]) et les magasins de données servant à répondre aux requêtes analytiques (« *une collection de données (…) orientées sujets (…) et disponibles pour l'interrogation et l'analyse* » [Inmon, 1994]). Ces travaux ont abouti à la définition d'un modèle conceptuel orienté-objet pour les entrepôts supportant l'historisation et l'archivage des données [Ravat, Teste, Zurfluh, 1999, 2000a, 2001b] [Ravat, Teste, 2000b, 2000c, 2001c, 2001d] [Teste, 2000a, 2000b] et d'un outil d'aide à la conception d'entrepôts [Bret, Teste, 1999].

Orientation de mes travaux

Mes travaux de recherche se sont orientés sur l'étude des magasins de données. Mes recherches consistent à proposer des modèles de représentation de données

multidimensionnelles pour les magasins de données qui sont destinés à faciliter l'interrogation et l'analyse de données décisionnelles. Mes travaux portent également sur les mécanismes de manipulation de ces données dans le cadre des analyses OLAP. L'originalité de ma démarche réside dans cette double orientation qui consiste à proposer des mécanismes de description couplés aux mécanismes de manipulation des données.

Ces recherches s'articulent chronologiquement en trois axes majeurs : un premier axe portant sur les systèmes OLAP dont les données sont structurées de manière multidimensionnelle, un deuxième axe, visant à intégrer des données complexes tels que les documents dans les systèmes OLAP, et un troisième axe, visant à développer des mécanismes de personnalisation des systèmes OLAP pour mieux intégrer les besoins inhérents à chaque usage.

Axe 1 : Modélisation et manipulation multidimensionnelle des systèmes OLAP. Une particularité de ce domaine de recherche est qu'il a été initié par le développement de nombreux logiciels dans le monde industriel avant que la communauté scientifique s'empare des problématiques de recherche sous-jacentes. Ceci explique probablement la grande variété qui existe dans les solutions proposées. Au début de la décennie, une convergence sur certains principes existait, mais aucun standard (à l'instar de ce qui existe dans les bases de données relationnelles) n'était unanimement reconnu [Torlone, 2003]. Une des premières tâches que je me suis donc assignées a été de proposer des mécanismes de représentation formalisés selon leur niveau d'abstraction conceptuel ou logique. Parallèlement mes recherches ont consisté à définir une algèbre OLAP afin d'identifier clairement les opérations inhérentes aux analyses indépendamment des variabilités dans les choix d'implantation. En effet, les solutions existantes étaient très diversifiées tant sur le plan des concepts manipulés que sur le plan des définitions des opérations. Sur la base de ce fondement algébrique, j'ai complété ces travaux de recherche par le développement de langages rendant accessibles aux utilisateurs les opérateurs algébriquement formalisés.

Axe 2 : Intégration de documents dans les systèmes OLAP. Les solutions de modélisation et de manipulation développées se montrent robustes devant des données factuelles fortement structurées, mais s'avèrent très vite inadaptées à des données atypiques telles que les documents [Perez, *et al.*, 2008b]. L'intégration des documents dans les systèmes OLAP pose de nombreux problèmes de part la présence plus importante de textes et de part la variabilité des documents d'une collection. Ne pas permettre des analyses sur les documents revient à exclure des données qui pourraient être utiles aux décideurs [Tseng, *et al.*, 2006]. J'ai donc orienté mes recherches dès 2003 sur les problématiques liées à l'intégration de documents, et tout particulièrement, sur les données textuelles. Une première difficulté concerne l'arbitrage entre l'extension des modèles existants et la définition de nouveaux modèles pouvant prendre en compte efficacement les documents. Les travaux menés ont pris en compte des documents structurés sans variabilité (présence ou non des attributs). J'ai également revisité l'algèbre OLAP afin d'étendre les opérateurs existants pour les rendre opérant à la fois sur des données numériques et textuelles. Le principal verrou scientifique qu'il faut résoudre est l'utilisation des mécanismes d'agrégations massivement présents lors des analyses OLAP. Dans ce contexte de nouveaux mécanismes d'agrégation sur des données de type texte doivent être développés.

Axe 3 : Personnalisation des systèmes OLAP. Les recherches menées depuis 15 ans se sont principalement attachées à faciliter l'accès aux données décisionnelles en intégrant ces dernières dans un entrepôt de données et préparant les données au sein de magasins en vue de leur analyse. Ces magasins de données sont généralement conçus pour un groupe d'usagers (« *une collection de données (…) orientées sujets* » [Inmon, 1994]) supposés partager des besoins identiques [Rizzi, *et al*., 2006]. En outre, l'usager décideur fait reposer ses prises de décisions non seulement sur les données décisionnelles mais également sur son « expérience » : connaissances externes, réflexions et discussions, expériences passées… Depuis 2006, j'ai donc élargi mes recherches afin de développer des solutions visant à personnaliser les systèmes décisionnels [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a]. L'objectif suivi est double. En premier lieu, j'ai cherché par l'intégration des préférences utilisateurs [Rizzi, *2007*] à offrir des mécanismes d'assistance au décideur lors de ses analyses, pouvant aller jusqu'à lui recommander certaines manipulations. En deuxième lieu, j'ai cherché à apporter des solutions pour aider l'usager dans ses analyses en capitalisant l'expérience des décideurs. Le patrimoine immatériel, généralement laissé à l'extérieur des systèmes informatiques, correspond à tous les aspects qui ont permis la prise de décision tels que les décisions passées (expériences), les connaissances externes, les commentaires, les discussions et débats entre experts… Cette expérience passée pourrait être alors restituée aux usagers lorsque l'analyse présente est similaire à des analyses antérieures.

Plan du mémoire

Ce mémoire s'articule de la manière suivante.

- Le chapitre 2 introduit le contexte de mes travaux en définissant les différents concepts ayant servi de cadre à mes recherches.
- Le chapitre 3 présente mes principales contributions sur la modélisation multidimensionnelle et la manipulation OLAP de données numériques.
- Le chapitre 4 se consacre à l'intégration des documents dans les bases de données multidimensionnelles. Sont exposés mes propositions en matière de modélisation des documents dans un espace multidimensionnel ainsi que mes solutions pour la manipulation OLAP des documents et pour l'agrégation de données textuelles.
- Le chapitre 5 présente mes travaux sur la personnalisation des systèmes décisionnels pour une meilleure prise en compte de l'usager.
- Le chapitre 6 expose les différents contextes applicatifs et projets ayant permis de valider les propositions des trois axes d'études préalablement décrits : approche classique sur des données numériques, intégration des documents, prise en compte de l'usager.

Chapitre 2 - Les systèmes décisionnels

Le développement de systèmes décisionnels spécialisés pour mesurer et analyser des données afin de mieux soutenir le pilotage des organisations est devenu crucial. L'aide à la décision a connu un tel développement tant au niveau du monde industriel [Inmon, 1993] [Smith, *et al.*, 2004] [Mundy, *et al.*, 2006] qu'à celui de la recherche [Widom, 1995] [Chaudhuri, *et al.*, 1997] [Dinter, *et al.*, 1998] [Vassiliadis, *et al.*, 1999] [Rizzi, 2007] [Pérez, *et al.*, 2008b] qu'il constitue aujourd'hui un axe de recherche important dans le domaine des systèmes d'information et des bases de données.

1 Contexte de l'aide à la décision

1.1 Le système décisionnel dans l'organisation

Toute organisation peut être décrite selon trois systèmes [Mélèse, 1972] :

- le système opérant représentant l'activité productrice de l'organisation qui consiste à transformer les flux primaires pour répondre aux besoins des clients,
- le système de pilotage correspondant à l'ensemble du personnel dirigeant qui régule, pilote et adapte l'organisation par leurs décisions,
- le système d'information permettant de collecter, conserver, traiter et restituer les données produites dans l'organisation.

Le système d'information [Cauvet, *et al.*, 2001] [Libourel, 2003] assure « le lien » entre les systèmes de pilotage et opérant : le système opérant produit des informations stockées dans le système d'information, qui après traitements assure la transmission de ces informations au système de pilotage lui permettant ainsi de connaître l'activité opérationnelle. Les décisions prises sont répercutées vers le système opérant au travers du système d'information.

Face aux importants défis que doivent relever les organisations (concurrence, développement à l'international, émergence de technologies…), le pilotage réclame aujourd'hui des systèmes dédiés efficaces [Miquel, 2005].

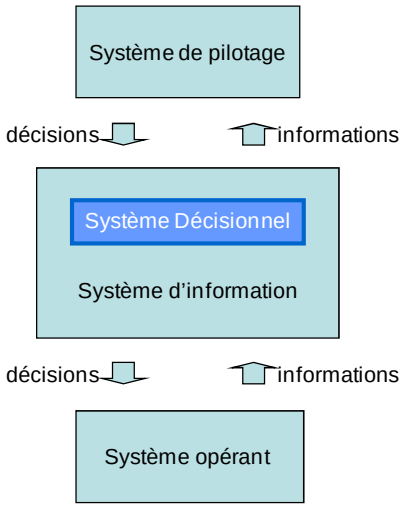


Figure 1 : Positionnement du système décisionnel dans l'organisation.

Définition 1. Nous définissons le *système décisionnel* comme le système dédié au support de la prise de décision (pilotage). Il regroupe l'ensemble des outils informatiques permettant d'extraire et de transformer (E.T.L.), de stocker (S.G.B.D.), d'analyser et de restituer les données décisionnelles d'une organisation.

1.2 Entrepôt et magasins de données

Un système décisionnel est mis en place à partir de sources de données. Ces *sources de données*, extérieures au système décisionnel, correspondent aux bases de données et fichiers présents pour l'essentiel dans le système d'information de l'organisation. Il est possible cependant d'alimenter le système décisionnel à partir de sources externes (filiales et partenaires, institutions gouvernementales, Web…).

L'architecture d'un système décisionnel comporte deux parties essentielles :

- la *préparation des données* où les données extraites des sources sont transformées en données décisionnelles, et
- l'*exploitation des données* où les utilisateurs accèdent aux données décisionnelles au travers d'interfaces graphiques et d'outils d'analyse. Le terme *décideurs* désigne les utilisateurs du système chargés d'analyser les données décisionnelles pour le pilotage de l'organisation. Ils exploitent les données au travers d'interfaces fournies par des outils d'interrogation, d'analyse et de restitution.

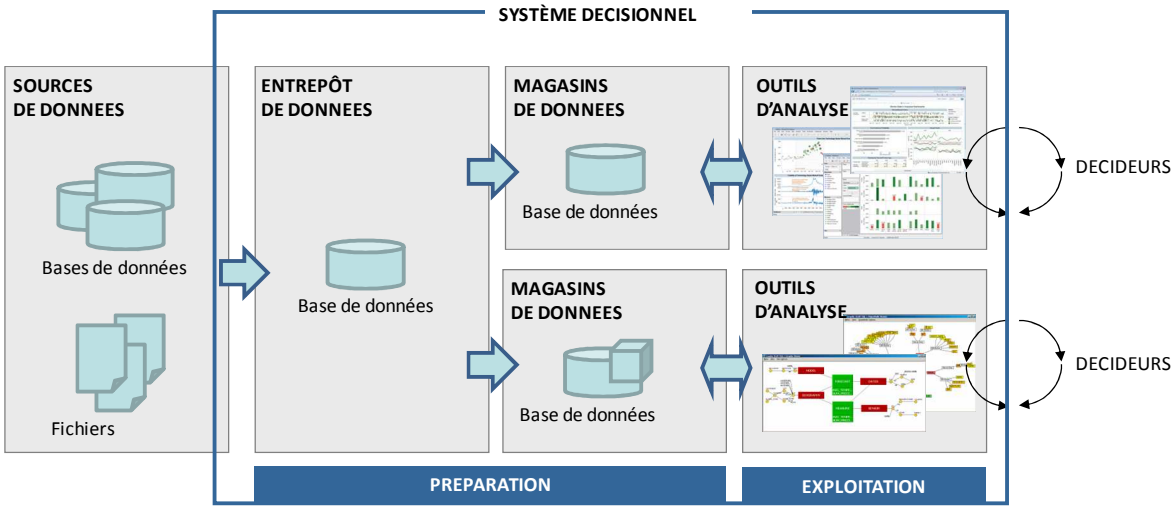


Figure 2 : Architecture d'un système décisionnel.

Un système décisionnel repose donc sur deux catégories d'espaces de stockage [Ravat, Teste, Zurfluh, 1999] : l'entrepôt de données et les magasins de données.

Définition 2. L'*entrepôt de données* (« data warehouse ») constitue l'espace centralisé où les données décisionnelles, identifiées comme pertinentes pour l'aide à la décision, sont stockées de manière homogène, le plus souvent historisées et agrégées, organisées suivant un modèle assurant la gestion efficace des données (cohérence, validité et fraîcheur des données).

Définition 3. Les *magasins de données* (« data marts »), élaborés comme un extrait de l'entrepôt, regroupent les données utiles pour un sujet d'analyse. Les données sont organisées suivant un modèle facilitant l'interrogation et l'analyse des données.

L'organisation des données au sein de l'entrepôt de données et des magasins de données est dirigée par des objectifs antinomiques [Teste, 2000a] : l'entrepôt de données suit une organisation assurant la gestion efficace des données tandis que les magasins de données sont structurés afin que l'interrogation et l'exploitation décisionnelle des données soient améliorées.

1.3 Les systèmes OLAP

La modélisation multidimensionnelle des données [Kimball, 1996] a connu un important développement. En effet, cette approche permet de supporter efficacement le processus décisionnel des organisations [Vassiliadis, *et al*., 1999] reposant sur des analyses OLAP (« On-Line Analytical Processing » [Codd, *et al*., 1993]).

Définition 4. Un *système OLAP* est défini comme un système décisionnel dans lequel les magasins de données suivent une organisation multidimensionnelle des données afin d'assurer un support efficace pour les analyses OLAP.

Notre approche situe la modélisation multidimensionnelle des données au niveau des magasins de données. En effet, nous considérons la modélisation multidimensionnelle comme une solution visant à améliorer l'interrogation et l'exploitation décisionnelle des données tandis que l'entrepôt de données repose sur une organisation des données normalisées afin de faciliter la gestion des données décisionnelles entreposées (notamment l'historisation) et les rafraîchissements périodiques (mises à jour des données par les processus E.T.L. d'alimentation).

Définition 5. Nous désignons par le terme de *Bases de Données Multidimensionnelles* (BDM) un magasin de données suivant une organisation multidimensionnelle.

Pour conclure, il est important de noter que notre architecture des systèmes décisionnels reste modulable. Ainsi, il est possible de concevoir un système décisionnel comportant uniquement un magasin de données, éventuellement organisé de manière multidimensionnelle, construit directement à partir des sources [Annoni, Ravat, Teste, Zurfluh, 2006b].

2 Principes de la modélisation multidimensionnelle

2.1 Métaphore du cube de données

Les analyses OLAP consistent à suivre des indicateurs considérés comme des points observés dans un espace défini par différents axes d'analyse. Cette vision multidimensionnelle des données peut être vue comme un *cube de données* [Gray, *et al.*, 1996]. Le cube de données est formé d'*arête*s représentants les axes d'observations d'indicateurs placés dans les *cellules*. Sur chaque arête, une graduation est choisie afin d'observer les données à un niveau adéquat de granularité.

Exemple. La figure suivante présente un cube de données formé de montants de vente en cellules et de trois arêtes graduées respectivement par des catégories de produits, des villes de magasins et des trimestres. La notion de cube de données ne se limite pas à trois axes mais se généralise en hyper-cube où le nombre d'axes est quelconque pouvant aller jusqu'à plusieurs dizaines.

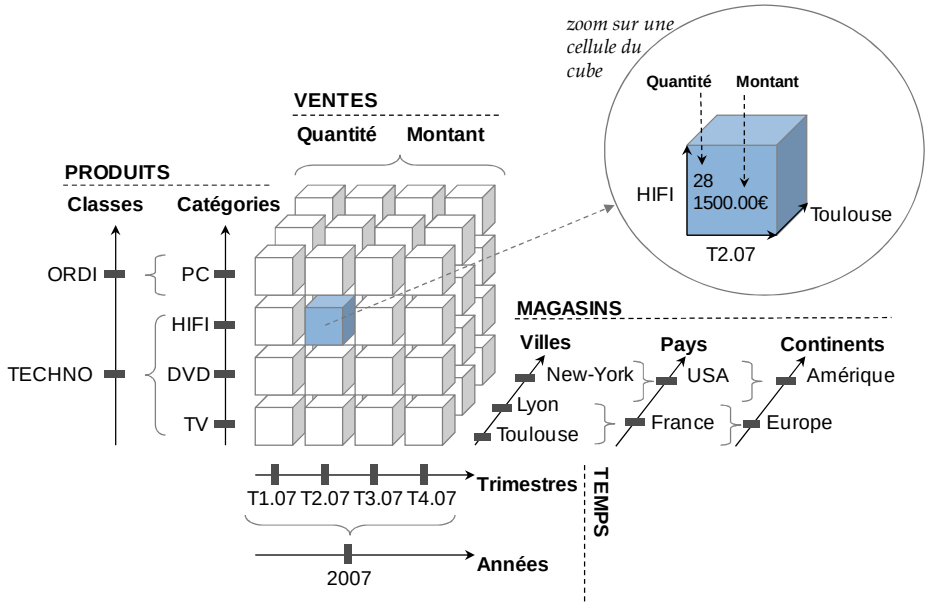


Figure 3 : Cube de données.

2.2 *Modèles de représentation des données*

Nous présentons la modélisation des données au sein d'un système OLAP conformément aux niveaux d'abstraction classiquement définis dans les méthodes de conception des bases de données :

- Les modèles conceptuels consistent à décrire une base de données indépendamment des choix technologiques ;
- Les modèles logiques consistent à définir la base de données en utilisant une technologie informatique (relationnel, objet…).

2.2.1 Modèles conceptuels

Les concepts et les formalismes associés à la modélisation multidimensionnelle existent mais souffrent de l'absence d'un consensus standardisé [Rizzi, *et al.*, 2006]. Cependant, on relève différents concepts reposant sur la métaphore de cube [Gray, *et al.*, 1996]. La modélisation multidimensionnelle consiste à modéliser les sujets d'analyse appelés **faits**, et d'axes d'analyse appelés **dimensions**. Les faits sont des regroupements d'indicateurs d'analyse appelés **mesures**. Elles correspondent aux données des cellules du cube. Les dimensions sont composées de paramètres modélisant les différentes graduations possibles des axes d'analyse. Chaque dimension représente une arête du cube. Enfin, les **paramètres** d'une dimension sont organisés au sein d'une **hiérarchie** conformément aux niveaux de granularité qu'ils représentent. Un fait et ses dimensions associées composent un **schéma en étoile** [Kimball, 1996]. Une généralisation possible consiste à décrire une « constellation d'étoiles » constituée de plusieurs faits et plusieurs dimensions éventuellement partagées formant un **schéma en constellation** [Kimball, 1996].

Exemple. L'exemple de la figure suivante correspond à la modélisation d'une analyse de *ventes* en fonction des *produits* et des *magasins* au cours du *temps*. Le sujet est modélisé par le fait composé des mesures *quantité* et *montant*. Les axes de l'analyse sont représentés par les dimensions *produits*, *magasins* et *temps*. La dimension *magasins* est caractérisée par trois paramètres *villes*, *pays* et *continents* organisés hiérarchiquement : le paramètre *villes* représente une graduation plus fine (permettant d'observer le sujet plus finement) que la graduation *pays*, elle-même étant une graduation plus fine de *continents*.

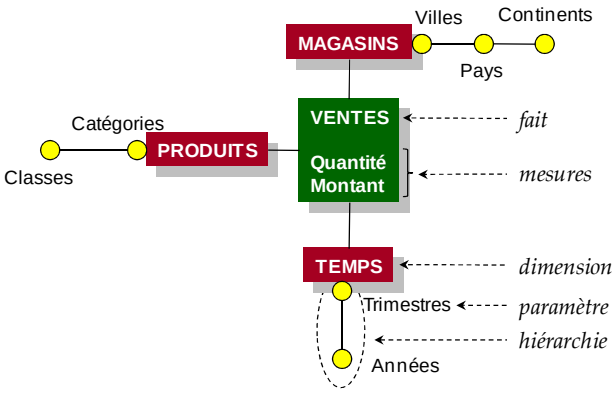


Figure 4 : Exemple de schéma en étoile.

2.2.2 Modèles logiques

R-OLAP

L'approche la plus répandue consiste à stocker les bases de données multidimensionnelles dans un environnement relationnel : on parle de l'approche « Relational OLAP » (R-OLAP) [Kimball, 1996] [Dinter, *et al*., 1998] [Mangisengi, *et al*., 1998]. Dans le contexte relationnel, la BDM est traduite par des relations. Cette approche procure de nombreux avantages : réutilisation des mécanismes de gestion des données éprouvés depuis des décennies et capacité à gérer des volumes de données très importants.

Autres approches

Une autre approche consiste à développer une technologie dédiée à la gestion des structures multidimensionnelles [Dinter, *et al*., 1998]. Cette approche dite « Multidimensional OLAP » (M-OLAP) vise à offrir des niveaux élevés de performance. Les bases M-OLAP stockent les données nativement sous une forme multidimensionnelle : il s'agit d'une application physique du concept de cube. Les bases de données de type M-OLAP restent limitées dans leur capacité à gérer d'importants volumes de données (au-delà du gigaoctet) et se heurtent à la nécessité de développer spécifiquement et entièrement tous les mécanismes des systèmes de gestion de base de données.

Une tendance, principalement présente dans le monde industriel (Oracle Application Server, Microsoft Analysis Services…), préconise une approche hybride, H-OLAP, faisant cohabiter au sein du même système les technologies R-OLAP pour les données détaillées et M-OLAP pour les données agrégées. L'objectif est de bénéficier des avantages des deux approches tout en minimisant leurs faiblesses.

Exemple. La figure 5 présente un exemple de traduction R-OLAP du schéma en étoile conceptuel précédent. Les clés primaires sont soulignées et les clés étrangères sont marquées par le symbole '#'.

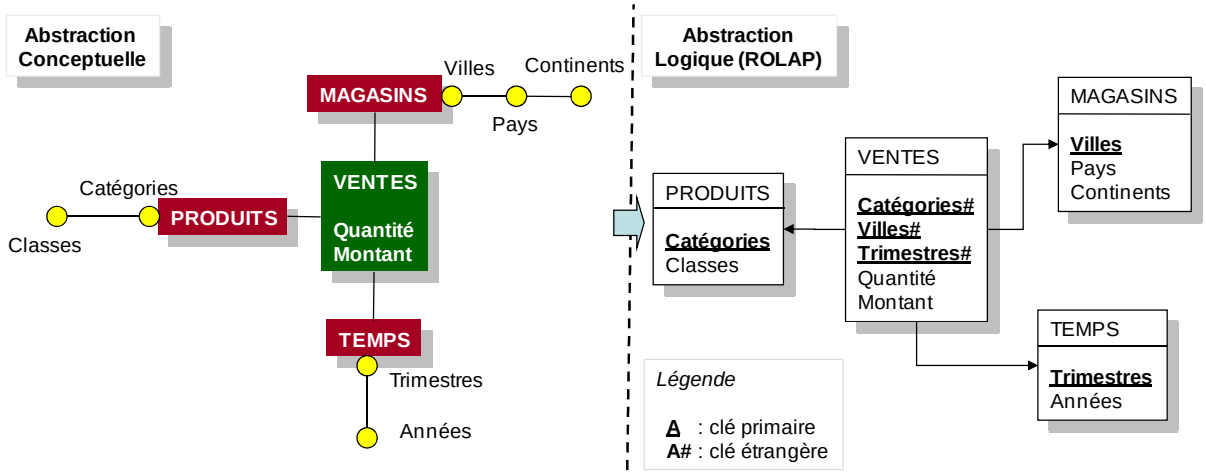


Figure 5 : Traduction R-OLAP d'un schéma conceptuel.

3 Principes de la manipulation OLAP

3.1 Structure de visualisation

La représentation sous forme de tableau est la structure de visualisation qui est le plus souvent retenue [Agrawal, *et al.*, 1997], [Gyssens, *et al.*, 1997] [Lehner 1998]. Il s'agit d'une vision synthétique et précise des données que les décideurs appréhendent facilement. Elle dérive directement de la métaphore du cube de données puisqu'elle peut être considérée comme la tranche du cube de données comme l'illustre la figure suivante.

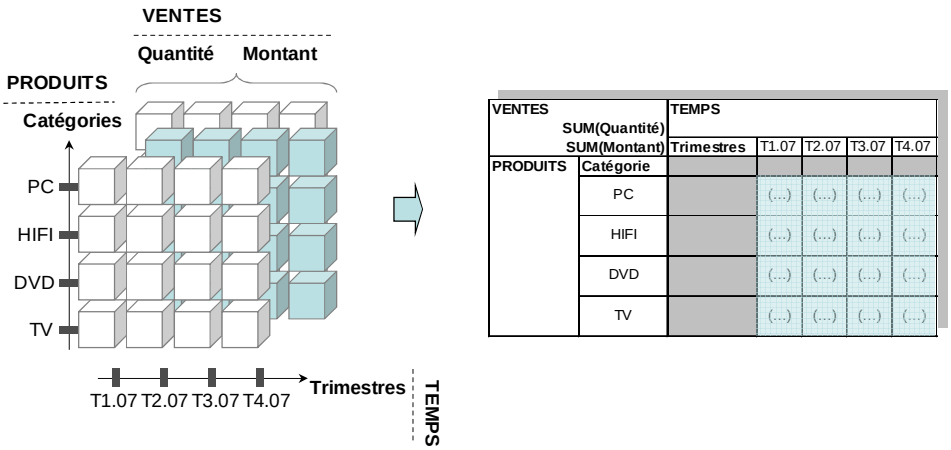


Figure 6 : Notion de tranche du cube de données.

3.2 Opérations de manipulation OLAP

De nombreuses propositions [Rafanelli, 2003] concernent la définition d'opérations de manipulation : cube de données, tranche du cube de données… Il n'existe pas de consensus sur la définition d'un ensemble minimum d'opérateurs assurant l'intégralité des opérations

de manipulation OLAP, mais la plupart des propositions offrent un support partiel des différentes catégories d'opérations.

Parmi ces opérations, les plus emblématiques sont les opérations de forage et les opérations de rotation qui reposent directement sur la métaphore du cube des données. Au-delà de ces opérations, la littérature scientifique et les nombreux logiciels offrent une grande variété d'opérations.

3.2.1 Opérations de forage

Les opérations de *forage* font reposer la navigation sur la structure hiérarchique des axes d'analyses, afin de permettre l'analyse d'un indicateur avec plus ou moins de précision. Le forage vers le haut (« roll-up ») consiste à analyser les données en fonction d'un niveau de granularité moins détaillé tandis que le forage vers le bas (« drill-down »), à l'inverse, permet d'analyser les données avec un niveau plus fin.

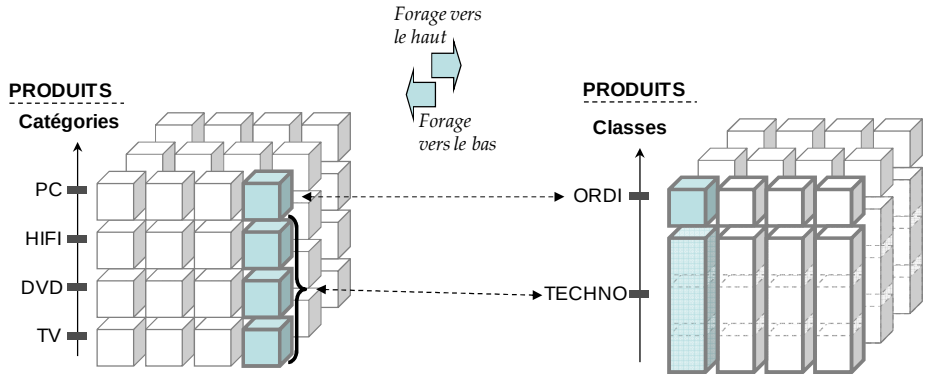


Figure 7 : Principe du forage.

3.2.2 Opérations de rotation

Les opérations de *rotation* réorientent une analyse. L'opération la plus courante consiste à changer l'axe d'analyse en cours d'utilisation (rotation de dimension).

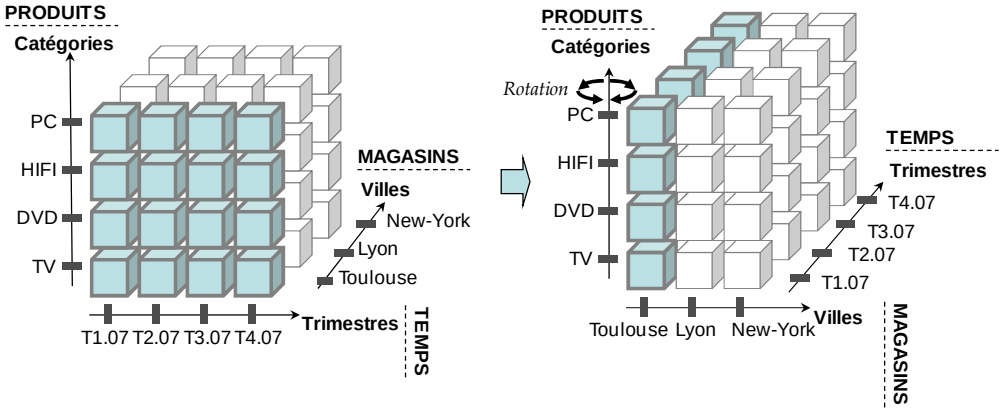


Figure 8 : Principe de rotation.

Un cas particulier de la rotation de dimension consiste à changer de perspective d'analyse (rotation de hiérarchie). Dans ce cas, l'axe en cours d'utilisation est maintenu, mais la manière de le graduer est changée.

Une autre opération apparentée à cette opération de rotation consiste à changer le sujet de l'analyse (rotation de fait ou « drill-across ») dans le contexte d'une constellation. Cette opération nécessite une forte compatibilité entre les dimensions des faits sur lesquels s'applique l'opération.

3.2.3 Autres opérations

Nous proposons de classer les nombreuses autres opérations proposées dans le cadre des manipulations OLAP en différentes catégories.

- Les opérations de *restriction* permettent à un utilisateur de restreindre l'ensemble des données analysées. La spécification d'une tranche de cube (« slice ») consiste à exprimer une restriction sur une des données de l'un des axes d'analyse. La spécification d'un sous-cube (« dice ») consiste à exprimer une restriction sur les données d'un indicateur d'analyse.
- Des opérations de *transformation* permettent l'ajout d'attributs de dimension en tant qu'indicateur d'analyse (« push ») ou de convertir un indicateur d'analyse en paramètre (« pull »).
- Les opérations d'*ordonnancement* permettent de changer l'ordre des valeurs (positions) des paramètres des dimensions (« switch ») ou de réordonner les paramètres d'une hiérarchie (« nest »). Par généralisation, cette dernière permet d'*imbriquer* un attribut dans une autre hiérarchie.
- Certains auteurs proposent aussi l'emploi des opérations binaires ensemblistes (*union*, *différence* et *intersection*) qui nécessitent une très forte compatibilité entre les deux structures multidimensionnelles manipulées. Certains travaux ont aussi proposé la notion de *jointure* inspirée de la jointure relationnelle, mais d'un intérêt limité dans un environnement multidimensionnel.

4 Articulation de mes recherches

La suite de ce mémoire expose les résultats de mes travaux en abordant les trois axes de recherche :

- Le chapitre 3 présente l'axe 1 traitant de la modélisation et la manipulation multidimensionnelles dans les systèmes OLAP. Mes recherches en matière de modélisation visent à définir un cadre précis permettant la description conceptuelle de données multidimensionnelles. Ces travaux sont complétés par l'étude des mécanismes de manipulation. Mes recherches consistent donc également à définir un cadre formalisé permettant de spécifier les opérations de manipulation OLAP des données multidimensionnelles.
- Le chapitre 4 concerne l'axe 2 portant sur l'intégration de documents dans les systèmes OLAP. L'intégration des documents oblige à étendre à la fois les solutions de modélisation et celles de manipulation proposées dans le premier axe. L'objectif de ces recherches est de rendre possible non seulement les analyses sur des valeurs numériques mais également sur le contenu textuel des documents.
- Le chapitre 5 détaille l'axe 3 étudiant la personnalisation dans les systèmes OLAP. Mes recherches consistent à définir des solutions pour personnaliser un système OLAP au regard des besoins de chaque usager. Ces travaux étendent les modèles de représentation des données multidimensionnelles par l'intégration de

préférences utilisateurs, de leur expertise, et proposent des mécanismes de recommandation pour faciliter les manipulations OLAP.

Enfin, le chapitre 6 décrit les contextes applicatifs et les projets dans lesquels se sont inscrites mes recherches.

Chapitre 3 - Modélisation et manipulation multidimensionnelle des systèmes OLAP

Ce chapitre présente les principales contributions de mes recherches pour la modélisation et la manipulation multidimensionnelles dans le contexte classique de données numériques.

1 Problématique

La conception et la mise en œuvre d'un système OLAP consistent, avant toute chose, en l'étude de modèles appropriés de représentation des données et de mécanismes de manipulation adaptés aux besoins d'analyses des décideurs.

Au début de nos travaux de recherche en BDM, des solutions de modélisation existaient [Abelló, *et al.*, 2001] mais aucun modèle standard n'avait été avalisé par la communauté. Les modèles proposés ne reposaient pas sur une formalisation précise, stable et reconnue par l'ensemble de la communauté scientifique. D'autre part, et ce malgré les nombreuses propositions faites pour la manipulation OLAP, une grande confusion existait quant aux opérations de manipulation : formalisations variables, désignations différentes, sémantique non parfaitement identique. Diverses propositions parcellaires ont été faites mais aucune algèbre OLAP n'a été développée avec le succès qu'a connu par exemple l'algèbre relationnelle. Cette absence de formalisation des manipulations a abouti à des outils d'interrogation très disparates, sans structures manipulées standardisées, ni même des opérations communes unanimement définies, à l'image de ce qui peut exister dans les bases de données relationnelles.

L'objet de ce chapitre est de décrire nos propositions en matière de modélisation conceptuelle pour les BDM et de manipulation OLAP. Notre approche vise à définir des solutions conceptuelles afin de se focaliser sur la représentation et l'organisation des données de manière à répondre aux besoins du décideur en faisant abstraction des aspects techniques d'implantation. Dès le début des années 2000, nous avons cherché à répondre à différentes interrogations :

- Quels concepts relèvent de la modélisation multidimensionnelle des données ? A quel niveau d'abstraction sont-ils assignés ?
- Quel est l'ensemble d'opérateurs relevant des manipulations OLAP sur ces données ? Quel est le noyau minimum de ces opérateurs ?

2 Approches existantes

Le tableau suivant dresse de manière chronologique un panorama des modèles conceptuels de description de données multidimensionnelles.

On relève que les premiers modèles proposés reposent directement sur la métaphore de cube [Gray, *et al.*, 1996] [Li, *et al.*, 1996] [Agrawal, *et al.*, 1997] [Datta, *et al.*, 1999] [Gyssens, *et al.*, 1997]. Cette approche supporte une séparation équivoque entre les éléments de structure et les valeurs [Torlone, 2003] : modélisation des axes de l'analyse peu expressive (difficulté à représenter l'organisation hiérarchique des données). Elle se heurte aussi à la représentation des espaces multidimensionnels constitués de plus de trois axes d'analyse. Elle s'avère enfin limitée lorsqu'il s'agit de représenter des constellations de faits et de

dimensions potentiellement partagées. Face à ces limites, d'autres approches sont apparues soit par extension de paradigmes existants, soit par développement de modèles originaux qualifiés de modèles multidimensionnels [Tournier, 2007]. La première tendance repose soit sur le paradigme E/A [Sapia, *et al*., 1998] [Tryfona, *et al*., 1999] [Hahn, *et al*., 2000] [Hüsemann, *et al*., 2000] [Malinowski, *et al*., 2006, 2008], soit sur le paradigme objet utilisant les notations UML [Trujillo, *et al*., 1998] [Pedersen, *et al*., 1999, 2001] [Nguyen, *et al*., 2000] [Abelló, *et al*., 2002, 2006] [Brukner, *et al*., 2001] ou de l'ODMG [Buzydlowski, et al., 1998]. Ces travaux étendent des notations standards par des mécanismes permettant la prise en compte de spécificités des données multidimensionnelles, complexifiant ces approches par un nombre parfois important de concepts. La seconde tendance développe des modèles dits multidimensionnels [Cabibbo, *et al*., 1998, 2000] [Golfarelli, *et al*., 1998] [Ravat, Teste, Zurfluh, 2001a] [Schneider, 2003] qui distinguent les éléments de structuration des valeurs tout en maintenant un nombre limité de concepts : fait, dimension, hiérarchie. Ces modèles restent spécialisés dans la représentation de données multidimensionnelles et ne reposent pas sur des notations standards [Torlone, 2003]. Les concepts et les formalismes associés à la modélisation multidimensionnelle souffrent de l'absence d'un consensus standardisé [Rizzi, *et al*., 2006].

Tableau 1 : Synthèse des travaux sur les modèles multidimensionnels.

		Paradigme existant		**Paradigme spécifique**	
		E/A	**Objet**	**Cube**	**Multidimensionnel**
1	[Gray, *et al* ., 1996]			☑	
2	[Li, *et al* ., 1996]			☑	
3	[Agrawal, *et al* ., 1997]			☑	
4	[Datta, *et al* ., 1997, 1999]			☑	
5	[Gyssens, *et al* ., 1997]			☑	
6	[Buzydlowski, *et al.* , 1998]		☑		
7	[Cabibbo, *et al* ., 1998, 2000][Torlone, 2003]				☑
8	[Golfarelli, *et al* ., 1998]				☑
9	[Lehner, 1998] [Lehner, *et al* ., 1998]			☑	
10	[Pedersen, *et al* ., 1998, 1999, 2001]		☑		
11	[Sapia, *et al* ., 1998]	☑			
12	[Trujillo, *et al* ., 1998, 2003] [Luján-Mora, *et al* ., 2002, 2005, 2006]		☑		
13	[Vassiliadis, *et al* ., 1998]			☑	
14	[Gopalkrishnan, *et al* ., 1999]		☑		
15	[Tryfona, *et al* ., 1999]	☑			
16	[Hurtado, *et al* ., 1999a, 1999b]			☑	
17	[Hahn, *et al* ., 2000]	☑			
18	[Hüsemann, *et al* ., 2000]	☑			
19	[Mangisengi, *et al* ., 2000]		☑		
20	[Mendelzon, *et al* ., 2000, 2003]			☑	
21	[Nguyen, *et al* ., 2000, 2001]		☑		
22	[Abelló, *et al* ., 2001b, 2001c, 2002, 2006]		☑		
23	[Bruckner, *et al* ., 2001]		☑		
24	[Ravat, Teste, Zurfluh, 2001] [Ravat, Teste, Tournier, Zurfluh 2008b]				☑
25	[Tsois, *et al* ., 2001]			☑	
26	[Schneider, 2003, 2007]				☑
27	[Franconi, *et al* ., 2004]			☑	
28	[Malinowski, *et al* ., 2005, 2006, 2008]	☑			
29	[Annoni, 2007]		☑		

Adossés aux différentes propositions de modèles de données multidimensionnelles, de nombreux opérateurs et langages d'interrogation de ces données ont été développés. Ces propositions visent à répondre aux besoins d'analyse OLAP des décideurs en définissant des opérateurs interactifs facilitant la navigation au sein des données multidimensionnelles [Abelló, *et al*., 2003]. Différentes études comparatives ont été réalisées dans [Rafanelli, 2003] [Torlone, 2003], [Abelló, *et al*., 2006] et [Ravat, Teste, Tournier, Zurfluh, 2008b]. Le tableau suivant dresse de manière chronologique un panorama des propositions sur la manipulation OLAP.

Les premiers travaux sur les manipulations OLAP ont étendu les opérateurs de l'algèbre relationnelle pour le modèle en cube [Gray, *et al*., 1996] [Li, *et al*., 1996], [Agrawal, *et al*., 1997] [Gyssen, *et a*l., 1997] [Datta, *et al*., 1999]. Pour mieux prendre en compte les structures multidimensionnelles, d'autres travaux ont proposé des opérateurs pour spécifier et manipuler un cube [Cabibbo, *et al*., 1997] [Abelló *et al*., 2003] [Pedersen, *et al*., 2001] [Franconi, *et al*., 2004]. La majorité des travaux repose sur une structure de visualisation simplifiée dans laquelle le concept de hiérarchie n'est pas exploité. Certaines des propositions [Gyssen, *et a*l., 1997] ne supportent qu'un niveau de paramètre en entête des lignes et des colonnes de la structure de visualisation. Ni la structure de visualisation des données, ni l'ensemble des opérateurs OLAP ne font encore aujourd'hui l'objet d'un consensus [Ravat, Teste, Tournier, Zurfluh, 2007b] à l'image des opérateurs de l'algèbre relationnelle. De part l'absence d'un standard de description des opérations OLAP, les outils de manipulation souffrent d'une grande hétérogénéité dans les langages proposés.

Tableau 2 : Synthèse des travaux sur les langages de manipulation OLAP.

		1	2		3	4	5	6	7	8	9	10
		[Gray, *et al*., 1996]	[Li, *et al*., 1996]	[Agrawal, *et al*., 1997]	[Gyssens, *et al*., 1997]	[Cabibbo, *et al*., 1997, 1998]	[Lehner, 1998]	[Datta, *et al*., 1999]	[Pedersen, *et al*., 2001]	[Abelló, *et al*., 2002, 2006]	[Franconi, *et al*., 2004]	[Ravat, Teste, Zurfluh, 2006b] [Ravat, Teste, Tournier, Zurfluh, 2008b]
Forage	**Haut**		☑	☑	☑	☑	☑		☑	☑		☑
	Bas		☑	☑			☑	☑		☑		☑
Rotation	**Hiérarchie**											☑
	Dimension									☑		☑
	Fait							☑	☑	☑		☑
Restriction	**Dimension**			☑	☑	☑		☑			☑	☑
	Fait				☑	☑				☑		☑
Transformation	**Paramètre en Mesure**			☑	☑							☑
	Mesure en Paramètre			☑	☑							☑
Ordonancement	**Position**				☑							☑
	Imbrication Paramètre		☑									☑
Calcul	**Agrégation**	☑										☑
Binaire	**Union**			☑	☑				☑	☑	☑	☑
	Intersection			☑	☑						☑	☑
	Différence			☑	☑						☑	☑
	Jointure		☑						☑			

3 Modélisation multidimensionnelle

L'orientation de mes travaux [Ravat, Teste, Zurfluh, 2001a, 2002] s'inscrit dans la tendance des modèles spécialisés dans la représentation multidimensionnelle limitant le nombre de concepts nécessaires. Mes recherches ont permis de développer une formalisation originale afin de définir l'ensemble des concepts de représentation des données multidimensionnelles. Pour chaque concept, un formalisme graphique associé est proposé constituant les notations du modèle [Ravat, Teste, Tournier, Zurfluh, 2007b, 2008b].

3.1 Concepts

Soient :

- $\mathcal{F}$ l'ensemble des faits,
- $\mathcal{D}$ l'ensemble des dimensions,
- $\mathcal{H}$ l'ensemble des hiérarchies,
- $\mathcal{M}$ l'ensemble des mesures,
- $\mathcal{A}$ l'ensemble des attributs de dimension,
- $\mathcal{P}$ l'ensemble des paramètres,
- $\mathcal{W}$ l'ensemble des attributs faibles.

Nous modélisons une base de données multidimensionnelles au travers d'une constellation [Kimball, 1996] de faits et de dimensions. La constellation constitue une généralisation de la modélisation en étoile. On remarque que si $|\mathcal{F}| = 1$ alors *C* est une étoile.

Définition 1. Une *constellation C* est définie par ($\mathcal{F}$; $\mathcal{D}$; *Star*) où

- $\mathcal{F} = \{F_1, \ldots, F_n\}$ est l'ensemble des faits,
- $\mathcal{D} = \{D_1, \ldots, D_m\}$ est l'ensemble des dimensions,
- *Star* : $\mathcal{F} \rightarrow 2^{\mathcal{D}}$ associe chaque fait à l'ensemble des dimensions en fonction desquelles il est analysable.

Définition 2. $\forall i \in [1..n]$ un *fait* F_i est défini par (NF_i ; $\mathcal{M}_i$) où

- NF_i est le nom identifiant le fait dans la constellation,
- $\mathcal{M}_i = \{m_1, \ldots, m_{xi}\}$ est l'ensemble des mesures.

Définition 3. $\forall i \in [1..m]$ une *dimension* D_i est définie par (ND_i ; $\mathcal{A}_i$; $\mathcal{H}_i$) où

- ND_i est le nom identifiant la dimension dans la constellation,
- $\mathcal{A}_i = \{Id_i, All_i\} \cup \mathcal{P}_i \cup \mathcal{W}_i$ est l'ensemble des *attributs de la dimension*. On distingue les *paramètres* $\mathcal{P}_i \subseteq \mathcal{P}$ représentant les graduations possibles, des *attributs faibles* $\mathcal{W}_i \subseteq \mathcal{W}$ représentant des informations additionnelles associées aux paramètres.
- $\mathcal{H}_i = \{H_1, \ldots, H_{pi}\} \subseteq \mathcal{H}$ est l'ensemble des *hiérarchies*.

Propriétés. Les attributs de la dimension D_i respectent les propriétés suivantes :

- Recouvrement des attributs de la dimension : $\mathcal{A} = \bigcup_{i=1}^{m} \mathcal{A}_i$
- Recouvrement des paramètres : $\mathcal{P} = \bigcup_{i=1}^{m} \mathcal{P}_i$

– Recouvrement des attributs faibles : $W = \bigcup_{i=1}^{m} W_i$

– Disjonction des attributs de dimension : $\forall (j_1, j_2) \in [1..m]^2$, si $j_1 \neq j_2$ alors $A_{j1} \neq A_{j2}$

Définition 4. $\forall H_j \in \mathcal{H}_i$ une *hiérarchie* H_j est définie par (NH_j ; P_{Hj} ; $\prec_{Hj}$; $Weak_{Hj}$) où

– NH_j est le nom identifiant la hiérarchie dans la constellation,

– $P_{Hj} = \{p_1, \ldots, p_y\} \subseteq P$ est l'ensemble des paramètres de la hiérarchie,

– $\prec_{Hj}$ est une relation d'ordre sur P_{Hj} telle que

  - l'ordonnancement des paramètres suit un ordre total
    $\forall p_{k1} \in P_{Hj}, p_{k2} \in P_{Hj}, k_1 \neq k_2, p_{k1} \prec_{Hj} p_{k2} \vee p_{k2} \prec_{Hj} p_{k1}$
  - il existe un paramètre *racine* $\forall p_{k1} \in P_{Hj}, Id_i \prec_{Hj} p_{k1}$
  - il existe un paramètre *extrémité* $\forall p_{k1} \in P_{Hj}, p_{k1} \prec_{H\,j} All_i$

– $Weak_{Hj} : P_{Hj} \rightarrow 2^{W_{Hj}}$ associe les paramètres à un ensemble d'attributs faibles.

Propriétés. Les hiérarchies respectent les propriétés suivantes :

– Recouvrement des hiérarchies : $H = \bigcup_{i=1}^{m} \mathcal{H}_i$

– Disjonction : $\forall i_1 \in [1..m], \forall i_2 \in [1..m]$, si $i_1 \neq i_2$ alors $\mathcal{H}_{i1} \neq \mathcal{H}_{i2}$

3.2 Formalismes graphiques

Associés à l'ensemble des concepts, nous définissons, par extension des notations introduites dans [Golfarelli, *et al.*, 1998], les formalismes graphiques représentant le schéma conceptuel décrivant les structures d'une BDM. Les figures suivantes présentent nos formalismes graphiques.

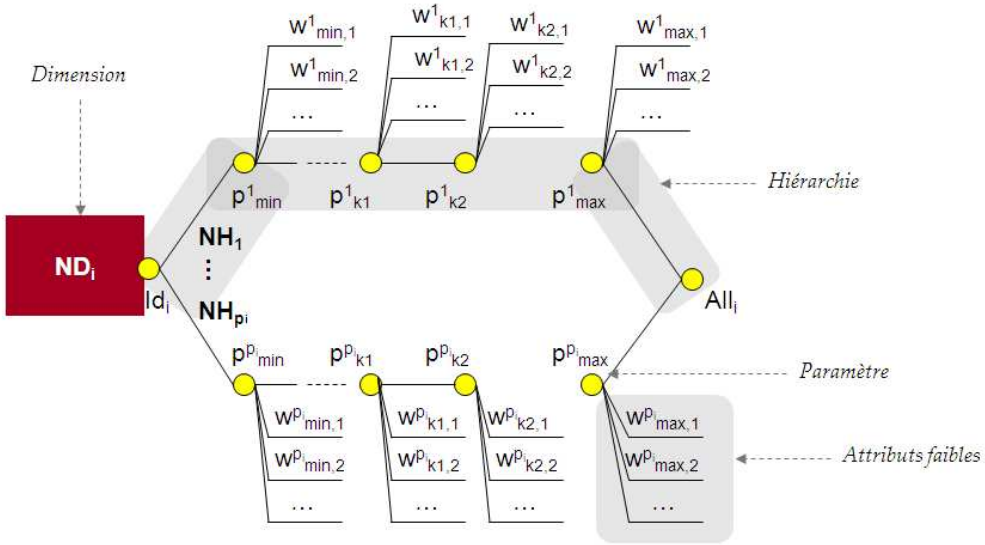


Figure 9 : Formalisme graphique d'une dimension.

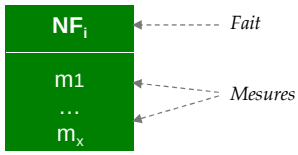


Figure 10 : Formalisme graphique d'un fait.

Mes recherches ont apporté des solutions au problème de visualisation de dimensions à structure complexe où de nombreuses hiérarchies sont définies. Elles ont notamment permis de définir différentes notations permettant de représenter les dimensions sous différentes facettes où les hiérarchies sont présentées de manière plus ou moins synthétique. Ces notations sont présentées en détail dans [Tournier, 2004].

Exemple. L'exemple concerne des analyses relatives à la météorologie, et est extrait de la publication [Ravat, Teste, Tournier, Zurfluh, 2008b] dans *International Journal of Data Warehousing and Mining*. La BDM est modélisée par une constellation de deux faits et quatre dimensions. Conformément à la définition 1, sa description formelle est la suivante (pour simplifier nous désignons le concept de fait F_i, respectivement de dimension D_i et de hiérarchie H_j, par son nom NF_i, respectivement ND_i et NH_j) :

({FORECAST ; MEASURE} ; {MODEL ; GEOGRAPHY ; DATES ; SENSOR} ; {*Star*(FORECAST) = {MODEL ; GEOGRAPHY ; DATES} ; *Star*(MEASURE) = {SENSOR ; GEOGRAPHY ; DATES}})

La représentation graphique de cette constellation est illustrée dans la figure suivante qui présente une copie d'écran issue des interfaces de notre prototype Graphic-OLAP (*cf.* chapitre 6 pour plus de détail sur le logiciel).

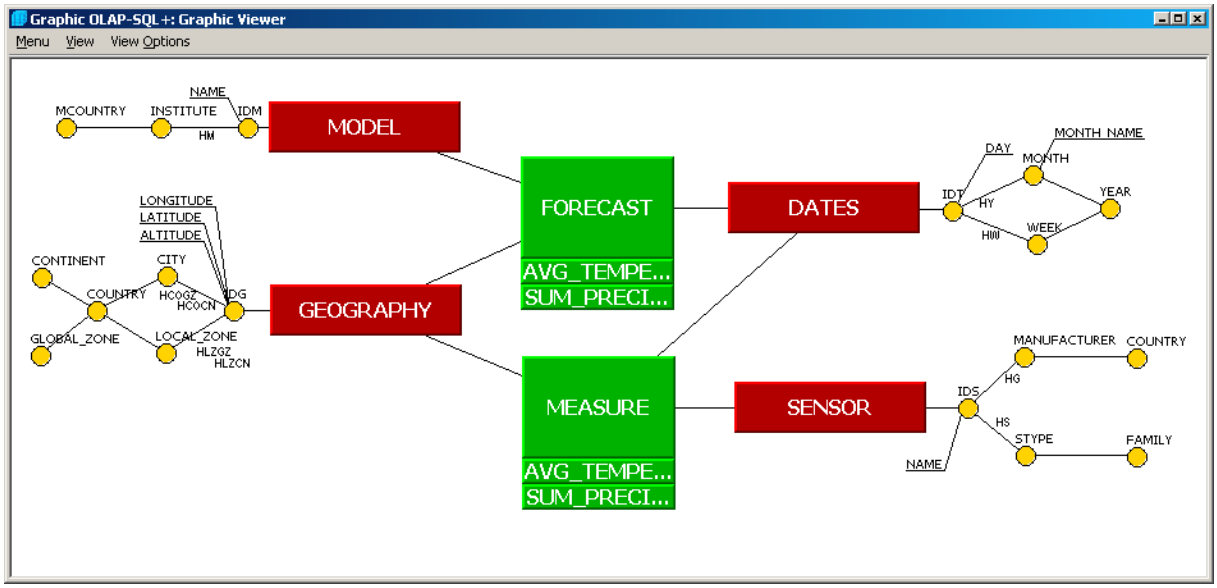


Figure 11 : Exemple d'un schéma en constellation.

Conformément aux définitions de notre modèle, les descriptions formelles du fait FORECAST et de la dimension GEOGRAPHY sont les suivantes :

- (FORECAST; {SUM(PRECIPITATION) ; AVG(TEMPERATURE)}),
- (GEOGRAPHY; {Id$_G$; City; Country; Global_Zone; Continent; Local_Zone; Longitude; Latitude; Altitude; All$_G$}; {H$_{COGZ}$; H$_{COCN}$; H$_{LZGZ}$; H$_{LZCN}$}).

La dimension est constituée des quatre hiérarchies H_{COGZ}, H_{COCN}, H_{LZGZ} et H_{LZCN}. Nous détaillons simplement la définition de H_{COGZ}.

(H_{COGZ}; {Id_G; City; Country; Global_Zone; All_G}; $Id_G \prec_{Hj}$ City $\prec_{Hj}$ Country $\prec_{Hj}$ Global_Zone $\prec_{Hj}$ All_G ; {$Weak_{Hj}(Id_G)$ = {Longitude; Latitude; Altitude}})

3.3 Extensions du modèle

Notre modèle conceptuel offre un socle stabilisant les concepts et les formalismes permettant la représentation d'un espace multidimensionnel pouvant servir de support aux manipulations OLAP. Ce fondement a permis de réaliser différentes extensions à ces recherches. Ces extensions ne sont pas détaillées dans ce mémoire mais elles peuvent se résumer en trois points :

- la prise en compte de contraintes sémantiques pour une meilleure description des données multidimensionnelles dont les travaux ont abouti à la thèse de Faiza Ghozzi [Ghozzi, 2004] ;
- le développement d'éléments de démarche dans le cadre des méthodes mixtes de conception des BDM qui ont fait l'objet de la thèse d'Estella Annoni [Annoni, 2007].
- l'intégration de versions pour gérer les évolutions des données mais également de leurs structures [Ravat, Teste, Zurfluh, 2006a] [Ravat, Teste, 2006d].

4 Manipulations OLAP

4.1 Table multidimensionnelle

Mes travaux [Teste, 2001] [Ravat, Teste, Zurfluh, 2002] [Ravat, Teste, Tournier, Zurfluh 2007b, 2008b] adoptent une structure de visualisation correspondant à une tranche du cube (deux dimensions), mais se distinguent par un respect strict de l'organisation multidimensionnelle des données en intégrant notamment la hiérarchisation des paramètres placés en entête des lignes et des colonnes. Nous définissons la structure de visualisation par le concept de *table multidimensionnelle*.

Définition 5. Une *table multidimensionnelle T* est définie par ($\mathcal{S}$; $\mathcal{L}$; $\mathcal{C}$; $\mathcal{R}$) où

– $\mathcal{S}$ = (F_c, {$f_1(m_1)$,…,$f_t(m_t)$}) représente le sujet analysé $F_c \in \mathcal{F}$ et $\forall i \in [1..t]$, $m_i \in \mathcal{M}_c \wedge f_i \in$ {SUM, AVG, MIN, MAX, COUNT,…},

– $\mathcal{L}$ = (*DL*, *HL*, *PL*) représente l'axe d'analyse $DL \in Star(F_c)$ utilisé en entête de lignes où une hiérarchie courante est désignée *HL* et un ensemble ordonné d'attributs de la dimension est défini *PL*,

– $\mathcal{C}$ = (*DC*, *HC*, *PC*) représente l'axe d'analyse $DC \in Star(F_c)$ utilisé en colonne,

– $\mathcal{R} = \left(\bigwedge_{\forall D_i \in Star(F_c)} D_i.p_j \theta v_k \right) \wedge \left(\bigwedge_{\forall m_i \in M_c} F_c.m_i \theta v_k \right)$ est un prédicat normalisé (conjonction de disjonctions) où $p_j \in \mathcal{A}_i \wedge v_k \in dom(p_j) \wedge \theta \in \{=; <; \leq; >; \geq; \neq\}$.

La figure suivante présente le formalisme que nous adoptons pour représenter une Table Multidimensionnelle (TM).

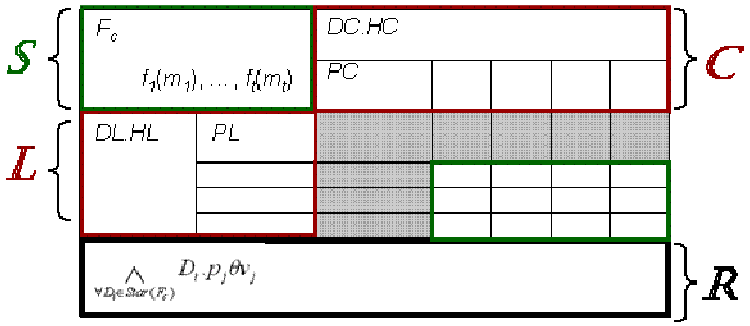


Figure 12 : Formalisme d'une table multidimensionnelle.

4.2 Algèbre OLAP

Les propositions de mes recherches relèvent d'une vision conceptuelle de la BDM (manipulation des faits, des dimensions, des hiérarchies) pour faciliter la compréhension des données [Ravat, Teste, Tournier, Zurfluh, 2007b] en faisant totalement abstraction des structures d'implantation. L'originalité de l'approche réside dans la définition d'une algèbre guidée par la vision du décideur. Les opérations d'interrogation effectuées par les décideurs sont décrites au travers d'une algèbre : description d'opérateurs indépendamment des primitives des langages. Notre algèbre OLAP diffère donc de l'algèbre relationnelle qui relève du niveau logique en décrivant symboliquement les opérations algorithmiques déclenchées par le SGBD en fonction des interrogations de l'utilisateur.

La contribution essentielle de mes recherches est de définir un noyau minimum d'opérateurs fermés pour la manipulation OLAP. Notre objectif est d'offrir un cadre permettant d'exprimer toutes les opérations induites par les navigations dans une BDM. Nous avons également complété notre approche par la définition d'opérateurs de second niveau pouvant simplifier l'expression de certaines opérations de manipulation couramment effectuées [Ravat, Teste, Tournier, Zurfluh, 2008b] et couvrant l'ensemble des propositions existantes dans ce domaine.

Nos travaux portent également sur la définition de langages de manipulation qui s'appuient sur nos propositions algébriques. Dans [Ravat, Teste, Zurfluh, 2002] nous avons proposé un langage textuel tandis que dans [Ravat, Teste, Zurfluh, 2006b] [Ravat, Teste, Tournier, Zurfluh, 2007b, 2008b] nous définissons des langages graphiques mieux adaptés aux décideurs. Le langage graphique que nous développons [Ravat, Teste, Tournier, Zurfluh, 2007b] s'avère complet au regard de notre algèbre OLAP.

Nous détaillons dans la suite nos propositions algébriques.

4.2.1 Constructeur

Le constructeur est essentiel car il permet de définir à partir d'une constellation une première TM initiant le processus exploratoire OLAP. Une TM est obtenue à partir d'un constructeur défini comme suit.

Définition 6. Le *constructeur* est défini par l'opérateur

DISPLAY(F_c, {$f_1(m_1)$,…,$f_t(m_t)$}, *DL*, *HL*, *DC*, *HC*) = T_{RES}

- F_c est le sujet de l'analyse représenté par le fait,
- {$f_1(m_1)$,…,$f_t(m_t)$} est l'ensemble des mesures $\forall i \in [1..t]$, $m_i \in \mathcal{M}_c \wedge f_i \in$ {SUM, AVG, MIN, MAX, COUNT,…},
- $DL \in Star(F_c)$ est l'axe d'analyse en ligne,

- $HL \in H_{DL}$ est la hiérarchie courante utilisée pour graduer les lignes,
- $DC \in Star(F_c)$ est l'axe d'analyse en colonne,
- $HC \in H_{DC}$ est la hiérarchie courante utilisée pour graduer les colonnes,
- $T_{RES} = (S_{RES} ; L_{RES}; C_{RES} ; R_{RES})$ est la TM résultat telle que
  - $S_{RES} = (F_c, \{f_1(m_1),\ldots,f_t(m_t)\})$,
  - $L_{RES} = (DL, HL, <All, p_{Lmax}>)$,
  - $C_{RES} = (DC, HC, <All, p_{Cmax}>$ *et*
  - $R_{RES} = \bigwedge_{\forall D_i \in Star(F_c)} D_i.All = 'all'$.

Exemple. L'expression algébrique suivante permet d'initier une analyse en obtenant une table multidimensionnelle affichant les prévisions météorologiques (quantités de précipitation et températures moyennes en degrés Celsius) par continent et par année.

DISPLAY(FORECAST ; {SUM(*Precipitation*) ; AVG(*Temperature*)} ;
GEOGRAPHY ; H_{COCN} ;
DATES ; H_Y) = T_0

La description formelle de la table multidimensionnelle $T_0 = (S_0 ; L_0; C_0 ; R_0)$ est :

- S_0 = (FORECAST , {SUM(*Precipitation*) ; AVG(*Temperature*)}),
- L_0 = (GEOGRAPHY, H_{COCN}, <*All, Continent*>),
- C_0 = (DATES, H_Y, <*All, Year* > *et*
- R_0 = GEOGRAPHY.ALL_G = 'All' ∧ DAYES.ALL_D = 'All' ∧ MODEL.ALL_M = 'All'.

La figure suivante donne la représentation graphique de la table T_0 résultante.

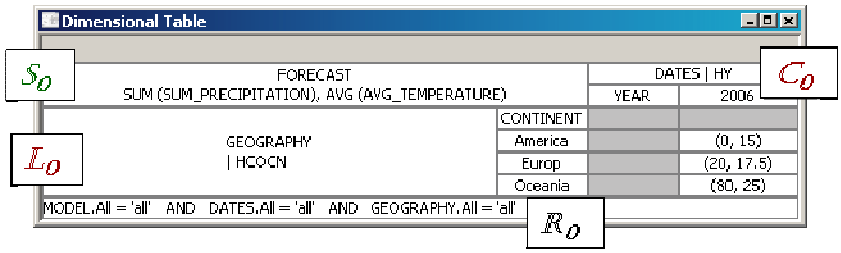


Figure 13 : Exemple d'une table multidimensionnelle.

4.2.2 *Noyau minimum fermé d'opérateurs OLAP*

Le tableau suivant décrit le noyau minimum fermé de l'algèbre de manipulation OLAP sur une TM [Ravat, Teste, Tournier, Zurfluh, 2007b]. Cet ensemble d'opérateurs repose sur une formalisation permettant d'établir les propriétés fondamentales que doit respecter une algèbre (nous détaillons ces propriétés à la sous-section intitulée 'Propriétés du noyau' en page 26). Ainsi afin d'assurer la fermeture du noyau chaque opérateur :

- porte en entrée sur une TM source notée $T_{SRC} = (S_{SRC} ; L_{SRC}; C_{SRC} ; R_{SRC})$, et
- produit en sortie une TM résultat notée $T_{RES} = (S_{RES} ; L_{RES}; C_{RES} ; R_{RES})$.

Les opérateurs binaires [Agrawal, *et al*., 1997] ne sont pas présentés dans ce mémoire. Nous avons fait des propositions sur les opérateurs binaires dans [Ravat, Teste, Zurfluh, 2005] et nous avons montré que les opérations binaires exigent une forte compatibilité entre les TM ce qui limite leur cadre d'utilisation.

Description du noyau

On pose :

- $\mathcal{S}_{SRC} = (F_c, M_c\})$, $M_c = \{f_1(m_1),\ldots,f_t(m_t)\}$
- $L_{SRC} = (DL, HL, PL)$, $PL = <All, p_{Lmax},\ldots, p_{Lmin}>$
- $C_{SRC} = (DC, HC, PC)$, $PC = <All, p_{Cmax},\ldots, p_{Cmin}>$
- $\mathcal{R}_{SRC} = \bigwedge_{\forall D_i \in Star(F_c)} D_i.p_j \theta v_k$

Dans une TM, chaque paramètre a son domaine de définition ordonné ; on notera $dom(p_i) = <v_1, v_2, \ldots>$.

Tableau 3 : Noyau minimum fermé de l'algèbre OLAP.

Opérateurs	
DRILLDOWN(T_{SRC}, D, p_{inf}) = T_{RES}	
Conditions :	$D \in \{DL; DC\}$;
	$D = DL \Rightarrow p_{inf} \in HL \wedge \nexists p_k \in PL \mid level_{HL}(p_k) < level_{HL}(p_{inf})$ (1)
	$D = DC \Rightarrow p_{inf} \in HC \wedge \nexists p_k \in PC \mid level_{HC}(p_k) < level_{HC}(p_{inf})$
Résultat :	$T_{RES} = (\mathcal{S}_{SRC}; L_{RES}; C_{RES}; \mathcal{R}_{SRC})$
	$D = DL \Rightarrow L_{RES} = (DL, HL, <All, p_{Lmax},\ldots, p_{Lmin}, p_{inf}>) \wedge C_{RES} = C_{SRC}$
	$D = DC \Rightarrow L_{RES} = L_{SRC} \wedge C_{RES} = (DC, HC, <All, p_{Cmax},\ldots, p_{Cmin}, p_{inf}>)$
ROLLUP(T_{SRC}, D, p_{sup}) = T_{RES}	
Conditions :	$D \in \{DL; DC\}$;
	$D = DL \Rightarrow p_{sup} \in HL \mid level_{HL}(p_{Lmin}) < level_{HL}(p_{sup})$
	$D = DC \Rightarrow p_{sup} \in HC \mid level_{HC}(p_{Cmin}) < level_{HC}(p_{sup})$
Résultat :	$T_{RES} = (\mathcal{S}_{SRC}; L_{RES}; C_{RES}; \mathcal{R}_{SRC})$
	$D = DL \Rightarrow L_{RES} = (DL, HL, <All, p_{Lmax},\ldots, p_{sup}>) \wedge C_{RES} = C_{SRC}$
	$D = DC \Rightarrow L_{RES} = L_{SRC} \wedge C_{RES} = (DC, HC, <All, p_{Cmax},\ldots, p_{sup}>)$
ROTATE(T_{SRC}, D_{old}, D_{new}, H_{new}) = T_{RES}	
Conditions :	$D_{old} \in \{DL; DC\}$; $D_{new} \in Star(F_c)$; $H_{new} \in \mathcal{H}_{new}$
Résultat :	$T_{RES} = (\mathcal{S}_{SRC}; L_{RES}; C_{RES}; \mathcal{R}_{SRC})$
	$D_{old} = DL \Rightarrow L_{RES} = (D_{new}, H_{new}, <All>) \wedge C_{RES} = C_{SRC}$
	$D_{old} = DC \Rightarrow L_{RES} = L_{SRC} \wedge C_{RES} = (D_{new}, H_{new}, <All>)$
NEST(T_{SRC}, D, p_i, D_{new}, p_{new}) = T_{RES}	
Conditions :	$D \in \{DL; DC\}$; $D_{new} \in Star(F_c)$; $p_{new} \in \mathcal{A}_{new}$
	$D = DL \Rightarrow p_i \in PL$
	$D = DC \Rightarrow p_i \in PC$
Résultat :	$T_{RES} = (\mathcal{S}_{SRC}; L_{RES}; C_{RES}; \mathcal{R}_{SRC})$
	$D = DL \Rightarrow L_{RES} = (DL, HL, <All, p_{Lmax},\ldots, p_i, p_{new},\ldots, p_{Lmin}>) \wedge C_{RES} = C_{SRC}$
	$D = DC \Rightarrow L_{RES} = L_{SRC} \wedge C_{RES} = (DC, HC, <All, p_{Cmax},\ldots, p_i, p_{new},\ldots, p_{Cmin}>)$

AGREGATE(T_{SRC}, D, $f(p_i)$) = T_{RES}

Conditions : $D \in \{DL; DC\}$; $f \in \{\text{SUM, COUNT, MAX, MIN,…}\}$
$D = DL \Rightarrow p_i \in PL$
$D = DC \Rightarrow p_i \in PC$

Résultat : $T_{RES} = (\mathcal{S}_{SRC}; \mathcal{L}_{RES}; \mathcal{C}_{RES}; \mathcal{R}_{SRC})$
$D = DL \Rightarrow \mathcal{L}_{RES} = \mathcal{L}_{SRC}$ où $\text{dom}(p_i) = <v_1, f(v_1), v_2, f(v_2), \ldots> \wedge \mathcal{C}_{RES} = \mathcal{C}_{SRC}$
$D = DC \Rightarrow \mathcal{L}_{RES} = \mathcal{L}_{SRC} \wedge \mathcal{C}_{RES} = \mathcal{C}_{SRC}$ où $\text{dom}(p_i) = <v_1, f(v_1), v_2, f(v_2), \ldots>$

SWITCH(T_{SRC}, D, p_i, v_1, v_2) = T_{RES}

Conditions : $D \in \{DL; DC\}$; $(v_1, v_2) \in \text{dom}(p_i)^2 \mid \text{dom}(p_i) = <\ldots, v_1, \ldots, v_2, \ldots>$

Résultat : $T_{RES} = (\mathcal{S}_{SRC}; \mathcal{L}_{RES}; \mathcal{C}_{RES}; \mathcal{R}_{SRC})$
$D = DL \Rightarrow \mathcal{L}_{RES} = \mathcal{L}_{SRC}$ où $\text{dom}(p_i) = <\ldots, v_2, \ldots, v_1, \ldots> \wedge \mathcal{C}_{RES} = \mathcal{C}_{SRC}$
$D = DC \Rightarrow \mathcal{L}_{RES} = \mathcal{L}_{SRC} \wedge \mathcal{C}_{RES} = \mathcal{C}_{SRC}$ où $\text{dom}(p_i) = <\ldots, v_2, \ldots, v_1, \ldots>$

PULL(T_{SRC}, D, $f(m_i)$) = T_{RES}

Conditions : $D \in \{DL; DC\}$; $f(m_i) \in M_c$

Résultat : $T_{RES} = (\mathcal{S}_{RES}; \mathcal{L}_{RES}; \mathcal{C}_{RES}; \mathcal{R}_{SRC})$
$\mathcal{S}_{RES} = (F_c, M_c \setminus \{f(m_i)\})$ (2)
$D = DL \Rightarrow \mathcal{L}_{RES} = (DL, HL, PL \oplus <f(m_i)>)$ (3) $\wedge \mathcal{C}_{RES} = \mathcal{C}_{SRC}$
$D = DC \Rightarrow \mathcal{L}_{RES} = \mathcal{L}_{SRC} \wedge \mathcal{C}_{RES} = (DC, HC, PC \oplus <f(m_i)>)$

PUSH(T_{SRC}, D, p_i) = T_{RES} (2)

Conditions : $D \in \{DL; DC\}$;
$D = DL \Rightarrow p_i \in PL$
$D = DC \Rightarrow p_i \in PC$

Résultat : $T_{RES} = (\mathcal{S}_{RES}; \mathcal{L}_{RES}; \mathcal{C}_{RES}; \mathcal{R}_{SRC})$
$\mathcal{S}_{RES} = (F_c, M_c \cup \{p_i\})$
$D = DL \Rightarrow \mathcal{L}_{RES} = (DL, HL, PL \setminus \{p_i\}) \wedge \mathcal{C}_{RES} = \mathcal{C}_{SRC}$
$D = DC \Rightarrow \mathcal{L}_{RES} = \mathcal{L}_{SRC} \wedge \mathcal{C}_{RES} = (DC, HC, PC \setminus \{p_i\})$

SELECT(T_{SRC}, *pred*) = T_{RES}

Conditions : $pred = pred_1 \wedge \ldots \wedge pred_t$ est un prédicat en forme conjonctive normale ;
$\forall s \in [1..t]$, $pred_s = E_i.A_j\ \theta\ v_k$ où $E_i \in \{F_c\} \cup Star(F_c)$, $A_j \in \mathcal{M} \cup \mathcal{A}$, $\theta \in \{=; <; \leq; >; \geq; \neq\}$, $v_k \in \text{dom}(A_j)$

Résultat : $T_{RES} = (\mathcal{S}_{SRC}; \mathcal{L}_{SRC}; \mathcal{C}_{SRC}; \mathcal{R}_{RES})$
$\mathcal{R}_{RES} = \text{pred}$

ADDM(T_{SRC}, $f(m_i)$) = T_{RES}

Conditions : $m_i \in \mathcal{M}_c$; $f \in \{\text{SUM, COUNT, MAX, MIN,…}\}$; $f(m_i) \notin M_c$

Résultat : $T_{RES} = (\mathcal{S}_{RES}; \mathcal{L}_{SRC}; \mathcal{C}_{SRC}; \mathcal{R}_{SRC})$
$\mathcal{S}_{RES} = (F_c, M_c \cup \{f(m_i)\})$

DELM(T_{SRC}, $f(m_i)$) = T_{RES}

Conditions : $f(m_i) \in M_c$

Résultat : $T_{RES} = (\mathcal{S}_{RES}; \mathcal{L}_{SRC}; \mathcal{C}_{SRC}; \mathcal{R}_{SRC})$
$\mathcal{S}_{RES} = (F_c, M_c \setminus \{f(m_i)\})$

(1) $\text{level}_H(p)$ est une fonction retournant l'indice de position (entier) du paramètre p dans la hiérarchie H.

(2) $\setminus$ est l'opération de différence d'ensembles : $\{e_1, e_2, e_3\} \setminus \{e_2, e_4\} = \{e_1, e_3\}$

(3) $\oplus$ est l'opération de concaténation de listes : $<e_1, e_2, e_3> \oplus <e_4> = <e_1, e_2, e_3, e_4>$

Exemple. A partir de la table multidimensionnelle T_0 de l'exemple précédent, il est possible d'obtenir une nouvelle table T_1, par combinaison d'opérateurs élémentaires du noyau de l'algèbre OLAP.

DRILLDOWN(T_0 ; GEOGRAPHY ; Country) = T'_0

DRILLDOWN(T'_0 ; GEOGRAPHY ; City) = T''_0

SELECT(T''_0 ; GEOGRAPHY.Continent = 'Europ') = T_1

La figure suivante donne la représentation graphique du résultat T_1.

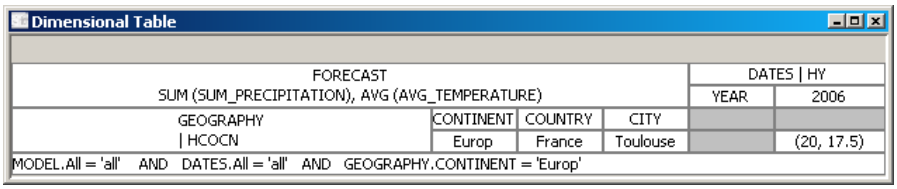


Figure 14 : Exemple de TM résultante de trois opérations algébriques.

Le noyau d'opérateurs algébriques que nous proposons respecte la propriété de fermeture. La série précédente d'opérations peut donc être factorisée dans l'expression suivante :

SELECT(DRILLDOWN(DRILLDOWN(T_0 ; GEOGRAPHY ; Country) ; GEOGRAPHY ; City) ; GEOGRAPHY.Continent = 'Europ') = T_1

Propriétés du noyau

La **fermeture** est l'une des principales propriétés satisfaites par notre algèbre OLAP. Cette propriété, comme l'illustre l'exemple précédent permet d'assurer qu'une manipulation complexe peut être construite par composition d'opérateurs élémentaires. Ceci est vérifié dans le tableau précédent décrivant les opérateurs élémentaires : la structure en entrée et la structure en sortie de chaque opérateur sont des TM.

Ce noyau algébrique fermé permet ainsi de garantir qu'une requête dans un langage complet au regard de notre algèbre peut être construite par composition d'autres requêtes. La notion de complétude correspond à celle introduite par Codd [Codd, 1972] pour comparer le pouvoir d'expression des langages proposés pour le modèle relationnel. Une algèbre satisfait la propriété de complétude relationnelle si pour toute expression algébrique construite à partir de l'algèbre relationnelle, il existe une expression équivalente dans l'algèbre considérée. Nous avons défini un langage graphique (*cf.* chapitre 6 pour plus de détail) complet au regard de notre algèbre OLAP [Ravat, Teste, Tournier, Zurfluh, 2007b, 2008b].

La seconde propriété importante pour une algèbre est la **couverture** du modèle considéré. Cette propriété consiste à garantir que tout élément décrit par le modèle peut être accédé et manipulé au travers des opérateurs de l'algèbre. Dans le tableau précédent décrivant les opérateurs élémentaires, les notations utilisées p_{inf} , p_{sup}, p_i désignent un niveau de graduation : soit un paramètre p, soit un ou plusieurs attributs faibles associés au paramètre $(af_1, af_2,\ldots)$, soit une combinaison des deux $p(af_1, af_2,\ldots)$. Ceci permet donc de manipuler les trois types de propriétés de notre modèle conceptuel : les mesures, les paramètres et les attributs faibles.

Enfin, la propriété de « **minimalité** » de l'algèbre vise à déterminer un ensemble minimum d'opérateurs à partir desquels le pouvoir d'expressivité des langages est garanti.

Nous considérons que l'ensemble d'opérateurs unaires proposé dans le tableau précédent est minimal relativement à l'ensemble des opérations que nous avons identifié dans la littérature scientifique (*cf.* Tableau 2). Toutefois, nous limitons notre exposé aux opérateurs unaires alors que des opérateurs binaires ont été proposés. Ces derniers étendent l'expressivité du langage de manière réduite car ils nécessitent une forte compatibilité des structures en entrée [Ravat, Teste, Zurfluh, 2005]. Nous appuyons cette affirmation suite aux travaux menés dans le cadre du stage de recherche d'Olivier Rouhaud [Rouhaud, 2005] que j'ai encadré.

Dans la section suivante, nous utilisons ce noyau minimum fermé pour définir des opérateurs étendus, définis par une combinaison des opérateurs élémentaires.

4.2.3 Extensions du noyau

Certaines analyses nécessitent de nombreuses combinaisons d'opérateurs élémentaires du noyau. Nous étendons le noyau avec des opérateurs construits par combinaison d'opérateurs élémentaires pour simplifier l'expression de certaines analyses. Il peut être également envisagé d'optimiser l'implantation de ces opérateurs étendus par rapport à la combinaison équivalente d'opérateurs élémentaires.

Le tableau suivant décrit les opérateurs étendus pouvant être exprimés par combinaison d'opérateurs du noyau.

Tableau 4 : Opérateurs OLAP étendus.

Opérateurs	
PLOT(T_{SRC}, D, p_i) = T_{RES}	
Description :	L'opération de projection d'un paramètre consiste à afficher les données suivant un paramètre quelconque d'une dimension
Combinaison du noyau :	DRILLDOWN(ROLLUP(T_{SRC}, D, *All*), D, p_i))=T_{RES}
HROTATE(T_{SRC}, D, H_{new}) = T_{RES}	
Description :	L'opération de rotation de hiérarchies consiste à changer la hiérarchie courante d'une dimension
Combinaison du noyau :	ROTATE(T_{SRC}, D, D, H_{new})
FROTATE(T_{SRC}, F_{new}, {$f_1(m_1)$,…,$f_t(m_t)$}) = T_{RES}	
Description :	L'opération de rotation de faits (Drill-Accross) consiste à utiliser un nouveau fait en conservant les caractéristiques des axes d'analyse en cours
Combinaison du noyau :	History[(1)] (T_{SRC}, F_c, History (T_{SRC}, DL, History (T_{SRC}, DC, DISPLAY(F_c, {$f_1(m_1)$,$f_t(m_t)$}, *DL*, *HL*, *DC*, *HC*))))=T_{RES}
ORDER(T_{SRC}, D, p_i, ω) = T_{RES}	
Description :	L'opération d'ordonnancement croissant (ω='ASC') ou décroissant (ω='DSC') consiste à ordonner les valeurs d'un paramètre
Combinaison du noyau :	SWITCH(…(SWITCH(T_{SRC}, D, p_i, v_1, v_2), …), D, p_i, v_g, v_h)=T_{RES}

(1) History(T_{OLD}, E, T_{SRC})=T_{RES} produit T_{RES} en rejouant dans T_{SRC} l'historique des opérations qui ont été appliquées dans T_{OLD} sur l'objet E (fait ou dimension).

Exemple. Nous considérons la table multidimensionnelle T_1 du précédent exemple visualisant les prévisions météorologiques de l'Europe en fonction des continents, des pays et des villes en ligne, et des années en colonne. Nous souhaitons modifier T_1 afin de visualiser simplement les pays et les villes. En utilisant simplement le noyau minimum de notre algèbre, il est nécessaire de combiner trois opérateurs suivant l'expression :

DRILLDOWN(DRILLDOWN(ROLLUP(T_1 ; GEOGRAPHY ; All_G) ; GEOGRAPHY ; *Country*)) ; GEOGRAPHY ; *City*)= T_2

En utilisant les opérateurs étendus l'expression est simplifiée comme suit :

DRILLDOWN(PLOT(T_1 ; GEOGRAPHY ; *Country*) ; GEOGRAPHY ; *City*)= T_2

4.2.4 *Augmentation du noyau*

Mes recherches sur la manipulation OLAP atteignent aujourd'hui une maturité suffisante pour servir de socle théorique au développement de nouveaux opérateurs. Ainsi nous avons fait une proposition en ce sens par une nouvelle opération autorisant des analyses multigraduelles [Hubert, Teste, 2009] [6]. Les tables multidimensionnelles s'avèrent être un cadre très fortement structuré (intégrant les hiérarchies dans les TM) assurant des manipulations OLAP toujours cohérentes, mais pouvant s'avérer contraignantes.

L'opération baptisée « BLEND » transforme la hiérarchie courante pour autoriser des analyses multi-granulaires ; par exemple il s'agit d'une analyse des températures de la France et de l'Europe où les deux entités géographiques sont considérées de même granularité. L'avantage de cette approche est de rendre possible ce type d'analyse au moment de la manipulation des données alors qu'elle nécessiterait dans le contexte classique la réorganisation complète des données et des processus ETL d'alimentation associés. La phase de construction d'une BDM est une tâche lourde difficilement réitérable en fonction de chaque besoin d'analyse.

Définition. L'opérateur de transformation multigraduelle est défini par l'opérateur

BLEND(T_{SRC}, D, $p_{sup}(s_{sup})$, $p_{inf}(s_{inf})$, *pred*) = T_{RES}

- T_{SRC} est la TM sur laquelle opère la transformation multigraduelle,
- $D \in \{DL, DC\}$ est l'une des dimensions de la table multidimensionnelle T_{SRC},
- p_{sup} et p_{inf} sont des paramètres consécutifs affichés de la dimension D tels que p_{sup} est le paramètre hiérarchiquement supérieur,
- $s_{sup} \in \{+,-\}$ et $s_{inf} \in \{+,-\}$ sont des estampilles indiquant la conservation (+) ou non (-) du paramètre associé dans T_{RES},
- *pred* est un prédicat de sélection permettant de déterminer les valeurs issues des paramètres p_{sup} et p_{inf} pour construire le domaine de définition du nouveau paramètre,
- T_{RES} est la TM transformée.

Le prédicat *pred* sert à calculer les ensembles E_{sup} et E_{inf} servant à la construction du domaine des valeurs du nouveau paramètre : E_{sup} contient les valeurs de p_{sup} sélectionnées par *pred*, et E_{inf} contient les valeurs de p_{inf} sélectionnées par $\neg pred$.

Contraintes.

(1) *pred* est valide si et seulement si $E_{sup} \cap parent(E_{inf}) = \varnothing$ où $parent(E_{inf})$ désigne les valeurs de $dom(p_{sup})$ hiérarchiquement supérieures et liées à E_{inf}. Par abus de langage, nous disons que *pred* doit définir deux ensembles de valeurs « disjoints ».

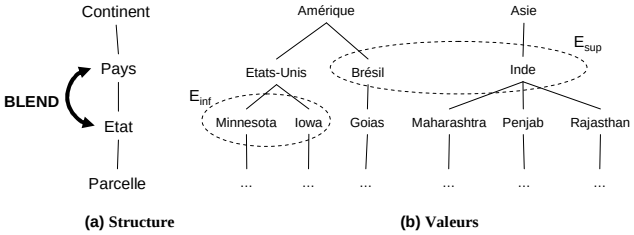
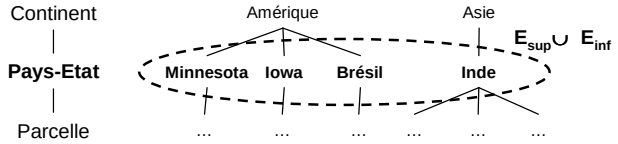
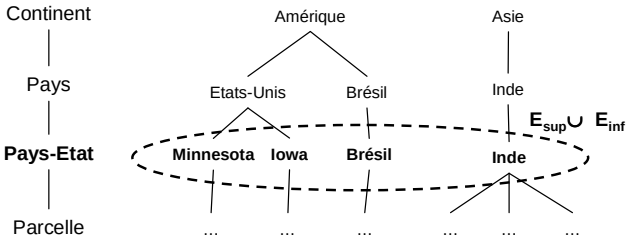
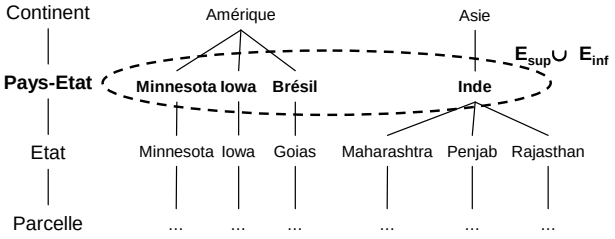
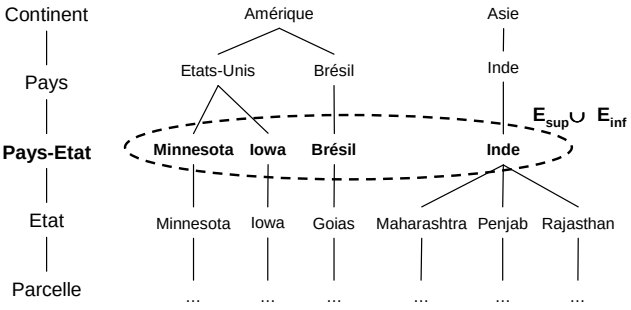
(2) La composition d'opérateurs BLEND n'est pas commutative. L'utilisateur doit construire ses manipulations en tenant compte de l'ordre des paramètres p_{sup} et p_{inf}, mais également de l'ordre des transformations multigraduelles.

Comme illustré dans le tableau suivant, l'opération supporte quatre scenarii :

[6] *Meilleur article académique* EGC'09 (https://lsiit.u-strasbg.fr/egc09/index.php/Prix_meilleurs_articles).

(a) soit le paramètre remplace les deux existants p_{sup} et p_{inf},
(b) soit le paramètre remplace le paramètre p_{inf},
(c) soit le paramètre remplace le paramètre p_{sup},
(d) soit le paramètre s'intercale entre les paramètres p_{sup} et p_{inf}.

Tableau 5 : Transformations multigraduelles par BLEND.

<table>
<tr><td>Entrée :</td><td> 
</td></tr>
<tr><td>Sortie :</td><td>(a) BLEND(T_{SRC}, Geographies, Pays(-), Etat(-),Pays <> 'Etats-Unis')
 
</td></tr>
<tr><td></td><td>(b) BLEND(T_{SRC}, Geographies, Pays(+), Etat(-),Pays <> 'Etats-Unis')
 
</td></tr>
<tr><td></td><td>(c) BLEND(T_{SRC}, Geographies, Pays(-), Etat(+),Pays <> 'Etats-Unis')
 
</td></tr>
<tr><td></td><td>(d) BLEND(T_{SRC}, Geographies, Pays(+), Etat(+),Pays <> 'Etats-Unis')
 
</td></tr>
</table>

L'opérateur BLEND modifie la hiérarchie existante en substituant un nouveau paramètre à l'un des paramètres existants (ou les deux) ou en intégrant un nouveau

paramètre en plus des paramètres existants. Les estampilles ajoutées aux deux paramètres p_{sup} et p_{inf} indiquent le scénario choisi : l'estampille (-) indique que le paramètre ne doit pas apparaître dans le résultat tandis que l'estampille (+) indique le contraire. Ceci permet de transformer deux paramètres en créant un nouveau paramètre multigraduel, tout en choisissant de maintenir tout ou partie des possibilités de navigations initiales (avec les opérations de forage). L'intérêt de l'opération réside dans la transformation de la hiérarchie existante directement dans la TM sans imposer une reconstruction de la BDM.

Nous capitalisons les résultats de nos recherches au sein du prototype Graphic-OLAP qui sert à expérimenter nos propositions par le développement d'interfaces graphiques permettant la définition et la manipulation de BDM au-dessus d'une infrastructure R-OLAP. L'environnement graphique facilite les expérimentations en simplifiant la création, la visualisation et l'interrogation de BDM. Certaines propriétés algébriques telles que la fermeture peuvent être visualisées. Cet environnement nous a servi de plate-forme de tests pour évaluer ce nouvel opérateur BLEND. Nous présentons certaines expériences au chapitre 6 du mémoire.

5 Bilan

5.1 Résultats de nos travaux

La contribution de nos travaux, retracée tout au long de ce chapitre, s'articule en deux thèmes : la modélisation et la manipulation des BDM.

Nos travaux ont dans un premier temps consisté à développer un modèle conceptuel spécialisé dans la représentation des données multidimensionnelles reposant sur les concepts de fait, dimension et hiérarchie. Associé aux concepts, nous avons proposé un formalisme graphique simplifiant la description du schéma d'une BDM en constellation. Ce fondement, inscrit dans la tendance des modèles multidimensionnels spécifiques, nous a permis de proposer des extensions par des contraintes sémantiques [Ghozzi, 2004] et des versions [Ravat Teste, Zurfluh, 2006a] [Ravat, Teste, 2006d] ainsi que le développement d'une démarche de conception [Annoni, 2007].

Nous avons orienté nos travaux vers la définition d'une algèbre OLAP orientée décideur pour l'interrogation de BDM [Ravat, Teste, Tournier, Zurfluh, 2007b, 2008b]. Le terme « orienté décideur » signifie que notre étude s'est focalisée sur l'identification des opérations nécessaires aux décideurs en faisant abstraction des structures d'implantation. Une autre caractéristique de notre approche est la proposition d'une structure de visualisation à double entrée hiérarchisée, appelée table multidimensionnelle, offrant une vision strictement conforme aux structures multidimensionnelles de l'espace d'exploration des données. Nos principaux résultats sont :

- la définition d'un noyau minimum d'opérateurs élémentaires fermés assurant la couverture du modèle,
- l'extension du noyau par des opérateurs définis par composition d'opérateurs élémentaires simplifiant la navigation au sein des données multidimensionnelles,
- l'augmentation du noyau minimum par la proposition d'un nouvel opérateur permettant des analyses multigraduelles par transformation des structures hiérarchiques [Hubert, Teste, 2009].

Nos travaux de recherche sur la modélisation et la manipulation des BDM ont fait l'objet de développement dans plusieurs prototypes de recherche. J'ai notamment développé le prototype GEDOOH [Teste, 2000a], qui fut étendu par Faiza Ghozzi dans le cadre de ces travaux sur les contraintes sémantiques dans les entrepôts avec GMAG [Ghozzi, 2004].

J'ai également développé la première version du prototype, Graphic-OLAP, qui a été ensuite étendu par le développement de plusieurs modules [Annoni, 2003] [Tournier, 2004]. Cette plateforme nous permet de mener des expérimentations tout en capitalisant les résultats de nos recherches. Elle sert également d'environnement de travail aux stages de recherches effectués par les étudiants de Master 2 dont j'assure le suivi et l'encadrement.

5.2 Encadrements et diffusion scientifique

Ces travaux ont permis le déroulement de deux thèses [Annoni, 2007] [Ghozzi, 2004], dont celle d'Estella Annoni que j'ai co-encadrée. Ils ont également permis à différents étudiants de Master (M2 ou D.E.A.), dont j'ai assuré l'encadrement, d'effectuer un stage de recherche.

(1) L. Benakezou « Etude et Conception de bases de données multidimensionnelles temporelles » en 2006,
(2) M. Gargouri « Assistance à l'élaboration incrémentale d'un magasin de données » en 2006,
(3) F. Boucheikh « Méthodologie de conception de systèmes décisionnels » en 2005,
(4) M. Kaddes « Etude de faisabilité d'une modélisation en constellation sous contraintes sémantiques » en 2005,
(5) E. Negre « Evolution de schémas dans une constellation » en 2005,
(6) O. Rouhaud « Bases de données décisionnelles : Fusion de tables multidimensionnelles » en 2005,
(7) A. Tahi « Interface d'interrogation incrémentale de données multidimensionnelles » en 2005,
(8) B. K. Le Thi « Intégration de contraintes dans OLAP-SQL » en 2004,
(9) M. Sallami « Développement d'une politique d'accès aux bases en constellation » en 2004,
(10) R. Tournier « Vers un langage de manipulation graphique des bases multidimensionnelles » en 2004,
(11) E. Annoni « Conception et développement d'un langage assertionnel pour les bases de données multidimensionnelles » en 2003,
(12) L. Bouzguenda « Conception et implantation d'un prototype d'interrogation et de visualisation d'une base multidimensionnelle » en 2002.

Le tableau suivant dresse un panorama des thèmes étudiés, des étudiants encadrés et des publications réalisées dans ce premier axe de mes recherches.

Tableau 6 : Etudiants encadrés et publications de l'axe 1.

Thèmes	Thèses	Master/D.E.A.	Publications
Modélisation en Constellation	F. Ghozzi E. Annoni	L. Benakezou M. Kaddes E. Negre B.K. Le Thi M. Sallami	▪ RI IRECOS [Ravat, Teste, 2006d] ▪ CI DAWAK'06 [Ravat, Teste, Zurfluh, 2006a] ICEIS'03 [Ghozzi, Ravat, Teste, Zurfluh, 2003a] ADBIS'01 [Teste, 2001] ▪ OI DMIDM [Ravat, Teste, Zurfluh, 2006c] ▪ RN RSTI/ISI [Ghozzi, Ravat, Teste, Zurfluh, 2004 ▪ CN EDA'08 [Annoni, Ravat, Teste, Zurfluh, 2008a] INFORSID'08 [Annoni, Ravat, Teste, 2008b] EDA'06 [Annoni, Ravat, Teste, Zurfluh, 2006d] EGC'03 [Ghozzi, Ravat, Teste, Zurfluh, 2003c] EGC'01 [Ravat, Teste, Zurfluh, 2001a]
Algèbre OLAP et Langages	F. Ghozzi	O. Rouhaud A. Tahi R. Tournier E. Annoni L. Bouzguenda	▪ RI IJDWM [Ravat, Teste, Tournier, Zurfluh, 2008b] ▪ CI ADBIS'07 [Ravat, Teste, Tournier, Zurfluh, 2007b] ▪ RN RSTI/ISI [Ravat, Teste, Zurfluh, 2002] ▪ CN EGC'09 [Hubert, Teste, 2009] INFORSID'06 [Ravat, Teste, Zurfluh, 2006b] EGC'05 [Ravat, Teste, Zurfluh, 2005] BDA'03 [Ghozzi, Ravat, Teste, Zurfluh, 2003b]
Démarche de Conception	E. Annoni	M. Gargouri F. Boucheikh	▪ CI SEKE'07 [Annoni, Ravat, Teste, 2007] DAWAK'06 [Annoni, Ravat, Teste, Zurfluh, 2006a] DEXA'06 [Annoni, Ravat, Teste, Zurfluh, 2006b] ▪ RN RSTI/ISI [Annoni, Ravat, Teste, 2006c] RSTI/ISI [Annoni, Ravat, Teste, Zurfluh, 2005a] ▪ CN INFORSID'06 [Annoni, Ravat, Teste, Zurfluh, 2006e] AIM'05 [Annoni, Ravat, Teste, Zurfluh, 2005b] EDA'05 [Ghozzi, Ravat, Teste, Zurfluh, 2005]

Outre l'effort de publication scientifique, j'ai assuré une diffusion plus large de mes recherches en réalisant différentes publications à visée pédagogique pour les Universités et les écoles d'ingénieurs. Certains résultats de mes travaux sont ainsi publiés dans le traité informatique H3870 des Techniques de l'Ingénieur [Chrisment, Pujolle, Ravat, Teste, Zurfluh, 2005] et dans l'Encyclopédie de l'informatique et des systèmes d'information aux éditions Vuibert [Chrisment, Pujolle, Ravat, Teste, Zurfluh, 2006].

5.3 Perspectives

Ces travaux offrent un cadre pour la modélisation et la manipulation OLAP à un niveau conceptuel. L'algèbre OLAP que nous avons développée est orientée décideur. J'envisage de poursuivre ce travail en identifiant un noyau algébrique pertinent d'un point de vue système intégrant l'organisation multidimensionnelle des données et ses optimisations par treillis de vues matérialisées.

Mes travaux ont permis d'augmenter l'algèbre OLAP par un nouvel opérateur. Une autre perspective concerne l'intégration de techniques issues de la fouille de données (« *data mining* ») [Dousset, 2003] dans les processus d'exploration OLAP. Par exemple, nous pensons permettre des ordonnancements des lignes et des colonnes d'une table multidimensionnelle en fonction des données analysées dans les cellules. Cette perspective de travail consiste à proposer une extension de l'algèbre OLAP et des langages associés.

Chapitre 4 - Intégration des documents dans les systèmes OLAP

Ce chapitre se focalise sur mes travaux de recherche visant à intégrer des documents dans les BDM. Mes contributions portent à la fois sur des aspects de modélisation et sur la manipulation OLAP des documents.

1 Problématique

L'analyse multidimensionnelle basée sur des BDM est une tâche bien maîtrisée sur des données factuelles numériques [Sullivan, 2001]. Alors que l'essor des technologies de l'information a considérablement accru la quantité de données et de documents numériques disponibles dans les organisations, seules 20% des données d'un système d'information sont traitées par un système OLAP [Tseng, *et al.*, 2006]. Les 80% restantes, des documents pour l'essentiel, sont hors de portée de la technologie OLAP par absence d'outils et de méthodes adaptés à la gestion de données textuelles. L'analyse de documents n'est pas supportée par les systèmes OLAP actuels [Perez, *et al.*, 2008b] [Ravat, Teste, Tournier, Zurfluh, 2008a]. Il s'agit d'autoriser non seulement les analyses classiques de données numériques additives sur lesquelles les mécanismes d'agrégation sont opérationnels mais également de permettre les analyses de documents et de textes. La figure suivante illustre cet objectif : les analyses classiques ramènent les analyses sur les documents à des dénombrements (valeurs numériques) tandis que nous souhaitons autoriser des analyses directement sur les valeurs textuelles contenues dans les documents.

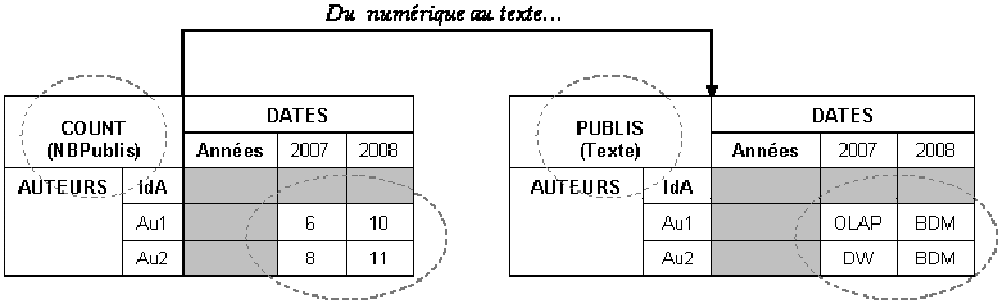


Figure 15 : De l'analyse numérique à l'analyse de textes.

Lors d'une analyse OLAP, le processus d'analyse agrège les données en fonction des niveaux de granularité sélectionnés via une fonction d'agrégation (somme, moyenne, maximum…). Les opérations de forage font un usage intensif de ces fonctions d'agrégation. Ces opérations permettent au décideur de changer le niveau de granularité utilisé pour afficher les données analysées. Ainsi, lors du changement de niveau, les données sont à nouveau agrégées par l'emploi de la fonction d'agrégation selon le nouveau niveau de granularité.

La problématique traitée dans ce chapitre consiste à rendre possible l'analyse de documents en développant un environnement capable non seulement d'analyser quantitativement les données (*nombre de publications*) mais aussi de manière plus qualitative au travers du contenu textuel (*mots-clefs représentatifs de la thématique abordée dans les publications*). Se posent alors plusieurs questions devant être solutionnées :

- Quel type de collection peut-on intégrer ? Quels sont les processus de transformation nécessaires à l'intégration des documents ?

– Doit-on faire évoluer les modèles de représentation existants ou doit-on développer une autre forme de modèles mieux adaptés à la description multidimensionnelle des documents et des indicateurs textuels qui en émanent ?
– Comment appliquer les manipulations OLAP sur ces structures de données textuelles ? Comment effectuer l'agrégation d'indicateurs textuels alors que les fonctions d'agrégation disponibles sont arithmétiques (somme, moyenne…) ?

2 Approches existantes

Les documents ont toujours été considérés comme des données non structurées peu compatibles avec les systèmes d'aide à la décision. Le format XML[7] fournit de nos jours un environnement permettant non seulement de considérer la structure mais aussi le contenu des documents [Fuhr, *et al.*, 2001]. Ainsi, les données textuelles contenues dans les documents XML doivent pouvoir être utilisées dans les systèmes OLAP [Tournier, 2007]. Nous distinguons deux types de documents XML [Kamps, *et al.*, 2004] [Pérez, *et al.*, 2008b] :

– Les *documents XML orientés données* (« data-centric XML documents ») sont constitués de données proches du contenu d'une base de données où les balises XML sont simplement utilisées pour marquer les lignes et les colonnes. Dans ce type de document l'ordre avec lequel sont stockées les données n'importe pas (*cf.* Figure 16) à l'image des relations dans les bases de données.

```
<doc>
<p>
  <nom> Teste </nom>
  <prenom> Olivier </prenom>
</p>
<p>
  <nom> Zurfluh </nom>
  <prenom> Gilles </prenom>
</p>
</doc>
```

=

```
<doc>
<p>
  <prenom> Gilles </prenom>
  <nom> Zurfluh </nom>
</p>
<p>
  <prenom> Olivier </prenom>
  <nom> Teste </nom>
</p>
</doc>
```

Figure 16 : Exemple de documents XML orientés données.

– Les *documents XML orientés documents* (« document-centric XML documents » ou « text-rich XML documents »). Dans ce type de document l'ordre avec lequel les données sont stockées est important (*cf.* Figure 17).

```
<doc>
<titre> Chanson d'automne </titre>
<p>
    Les sanglots longs Des violons De l'automne
    Blessent mon cœur  D'une langueur Monotone.
</p>
<p>
    Tout suffocant Et blême, quand Sonne l'heure,
    Je me souviens Des jours anciens Et je pleure,
</p>
<p>
    Et je m'en vais Au vent mauvais Qui m'emporte
    Deçà, delà Pareil à la Feuille morte.
</p>
<auteur> Paul VERLAINE </auteur>
<ref> Poèmes saturniens (1866) </ref>
</doc>
```

≠

```
<doc>
<auteur> Paul VERLAINE </auteur>
<p>
    Et je m'en vais Au vent mauvais Qui m'emporte
    Deçà, delà Pareil à la Feuille morte.
</p>
<titre> Chanson d'automne </titre>
<p>
    Les sanglots longs Des violons De l'automne
    Blessent mon cœur  D'une langueur Monotone.
</p>
<ref> Poèmes saturniens (1866) </ref>
<p>
    Tout suffocant Et blême, quand Sonne l'heure,
    Je me souviens Des jours anciens Et je pleure,
</p>
</doc>
```

Figure 17 : Exemple de documents XML orientés documents.

[7] Extensible Markup Language (http://www.w3.org/XML/).

En s'appuyant sur cette dichotomie, nous distinguons dans la littérature scientifique deux approches répondant aux problématiques de la modélisation multidimensionnelle intégrant des documents.

La première catégorie concerne la modélisation pour l'analyse de *documents XML orientés données*. L'analyse de tels documents a été spécifiée dans des propositions telles que celles de [Golfarelli, *et al*., 2001], [Pokorný, 2001] et [Boussaïd, *et al*., 2003]. Ces documents font partie des 20% de données sources actuellement exploitées par les systèmes OLAP.

La deuxième catégorie concerne les *documents XML orientés documents*.

- Dans un premier temps, les travaux développés [McCabe, *et al*., 2000] [Jensen, *et al*., 2001] [Mothe, *et al*., 2003] [Keith, *et al*., 2005] [Tseng, *et al*., 2006] ont apporté des solutions pour l'analyse de documents en s'appuyant sur un schéma en étoile classique se cantonnant à des indicateurs numériques et ne permettant donc pas l'analyse du contenu des documents.
- Une approche alternative s'est développée, proposant d'intégrer les documents XML dans des entrepôts de documents spécifiquement dédiés à leur stockage [Khrouf, *et al*., 2004]. Ces travaux sont complétés par des approches visant à adapter des fonctions d'agrégation pour l'analyse de documents à structures hétérogènes : une fonction pour agréger des structures XML, appelée « XAggregation », a été proposée dans [Wang *et al*., 2003, 2005]. Celle-ci a récemment été suivie de l'adaptation de l'opérateur « Cube » [Gray *et al*., 1996] avec l'opérateur nommé « X3 » [Wiwatwattana, *et al*., 2007].

Tous ces travaux sur la modélisation multidimensionnelle de documents XML orientés documents restent cantonés à des analyses numériques se limitant à des comptages d'instances. Par exemple, la Figure 18 illustre l'analyse du *nombre de publications* par *auteurs* et par *dates*. Afin d'avoir une vision plus synthétique, le décideur change le niveau de détail de l'analyse en effectuant un forage des *semestres* vers les *années* ; les valeurs semestrielles sont agrégées en valeurs annuelles pour chaque couple (*auteur*, *année*).

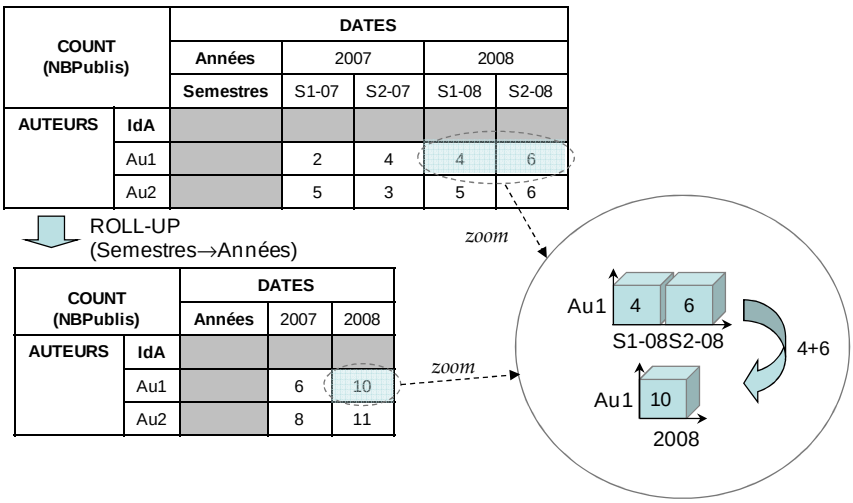


Figure 18 : Forage lors d'une analyse du nombre de publications.

Pour répondre à la problématique de l'analyse du contenu de documents textuels, les auteurs de [Park, *et al*., 2005] proposent de coupler la structure « xfact » [Nassis, *et al*., 2004] avec des fonctions d'agrégation basées sur des techniques de fouilles de données. Cette

approche reste partielle car les auteurs ne fournissent ni une définition formelle, ni une description détaillée de l'implantation de leur solution. Dernièrement dans [Pérez, *et al*., 2008a], une autre approche propose de combiner la modélisation multidimensionnelle et les techniques de recherche d'information pour fournir les documents pertinents relevant de l'analyse en cours. Ces travaux proposent de lier des informations contenues dans des documents à des données multidimensionnelles afin d'expliquer les faits. Notre objectif va plus loin puisque nous souhaitons pouvoir analyser directement les informations contenues dans les documents. Nous avons ainsi orienté nos recherches sur cet axe d'étude.

Aucune solution ne permet l'application de techniques d'exploration et d'analyse OLAP directement sur les documents et les données textuelles qui en émanent. Si l'on considère l'analyse de *publications* par *auteurs* et par *dates*, nous souhaitons permettre d'analyser non seulement le nombre de publications réalisées par les auteurs (Figure 18), mais aussi à rendre possible l'analyse du contenu, notamment textuel, de ces publications ; par exemple les principaux *mots-clefs* qui les caractérisent (Figure 19). Il devient alors nécessaire d'offrir un environnement d'analyse autorisant l'agrégation de textes ; lorsque le décideur change le niveau de détail de l'analyse en effectuant un forage des *semestres* vers les *années*, les valeurs semestrielles des mots-clefs doivent être agrégées en valeurs annuelles.

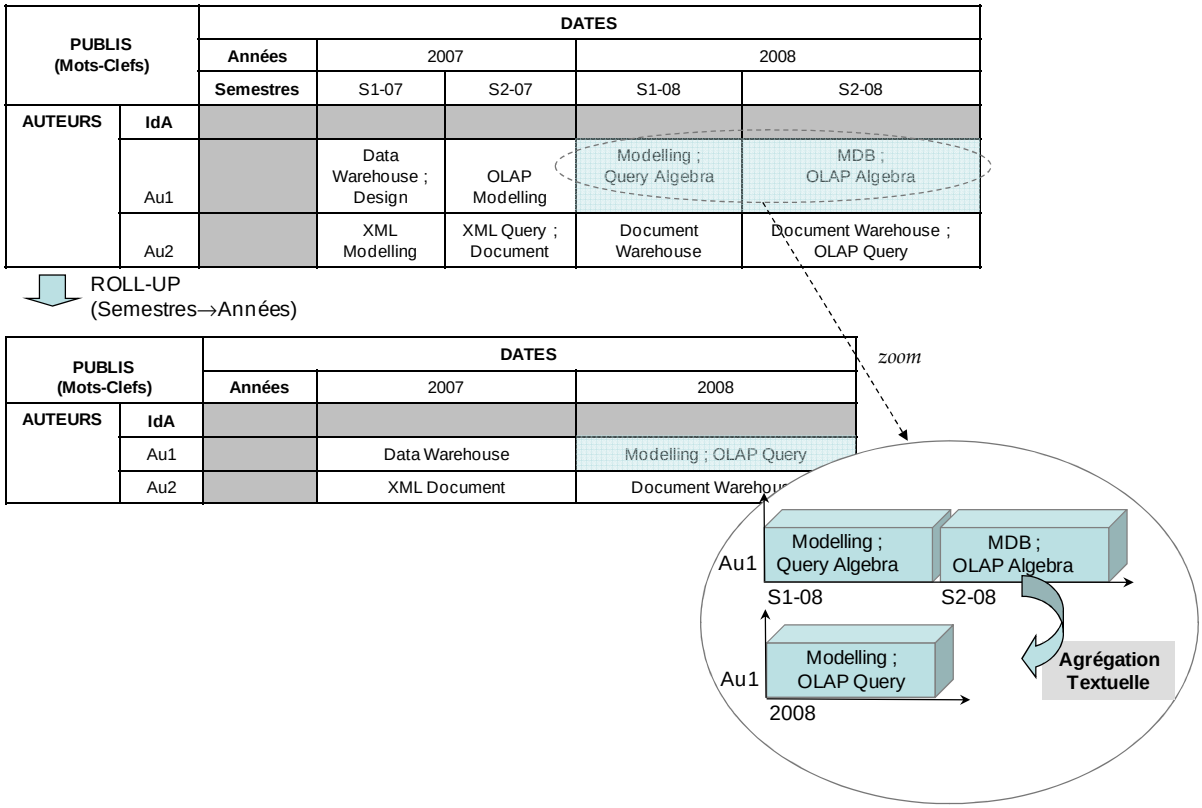


Figure 19 : Forage lors d'une analyse des mots-clefs dans des publications.

Dès 2003, notre objectif a été de fournir un environnement pour l'analyse OLAP du contenu de documents XML orientés données et documents. Afin d'effectuer l'analyse multidimensionnelle de documents, nos travaux ont suivi deux approches :

- une approche consistant à étendre les modèles de représentation multidimensionnelle de données numériques aux documents [Pujolle, Ravat,

Teste, Tournier, 2008] [Ravat, Teste, Tournier, 2008e, 2007e] [Ravat, Teste, Tournier, Zurfluh, 2007c], et,

- une approche consistant à revisiter les principes de la modélisation multidimensionnelle en proposant une nouvelle approche pour répondre aux exigences induites par des données issues de documents [Ravat, Teste, Tournier, Zurfluh, 2008a, 2008f, 2007a, 2007f]. Ces travaux se sont inscrits dans le cadre de la thèse de Ronan Tournier [Tournier, 2007].

Les prétraitements nécessaires à l'intégration des documents (processus E.T.L) ne sont pas détaillés dans ce mémoire ; une présentation est faite dans [Ravat, Teste, Tournier, Zurfluh, 2008f].

3 Extension de la constellation

Une constellation étendue aux documents est un schéma multidimensionnel représentant aussi bien le contenu que la structure des documents XML. Comme dans un schéma en constellation classique, un fait modélise un sujet d'analyse et une dimension modélise un axe d'analyse.

Définition. Une constellation textuelle CT est définie par ($\mathbb{F}^{CT}$; $\mathbb{D}^{CT}$; $Star^{CT}$) où

- $\mathbb{F} = \{F_1, \ldots, F_n\}$ est l'ensemble des faits,
- $\mathbb{D} = \{D_1, \ldots, D_m\}$ est l'ensemble des dimensions,
- $Star : \mathbb{F} \rightarrow 2^{D}$ associe chaque fait à un sous-ensemble des dimensions en fonction desquelles il est analysable.

Cette définition décrit de manière classique la constellation au travers des concepts de fait, de dimension et de hiérarchie (*cf.* chapitre 3). La prise en compte des spécificités des documents s'effectue par une double extension du modèle de représentation :

- au niveau des mesures composant les faits afin de déterminer les fonctions d'agrégation supportées et
- au niveau des dimensions en introduisant des dimensions spécifiquement dédiées à la représentation des documents.

3.1 Typologie de mesures

De manière à décrire les spécificités des données contenues dans les documents sur lesquelles les analyses OLAP doivent être appliquées, une première extension consiste à modéliser différents types de mesures.

Définition. $\forall j \in [1..x_i]$ une *mesure* m_j est définie par (n_j ; t_j ; f_j ;) où

- n_j est le nom de la mesure,
- t_j est le type de la mesure,
- f_j est la liste des fonctions d'agrégations applicables sur cette mesure.

Soit F^{AGG} l'ensemble des fonctions d'agrégation disponibles sur le système OLAP. Le type de la mesure conditionne les fonctions d'agrégation compatibles. Ainsi, pour un type de mesure t_j, l'ensemble des fonctions compatibles est $F^{AGG}{}_j$. Parmi les fonctions compatibles, le concepteur désigne les fonctions qui seront disponibles pour analyser la mesure m_j, c'est-à-dire $f_j \subseteq F^{AGG}{}_j \subseteq F^{AGG}$.

Nous distinguons plusieurs types de mesures [Ravat, Teste, Tournier, 2007e].

- Les *mesures numériques* sont exclusivement composées de données numériques additives ou semi-additives [Kimball, 1996] [Horner, *et al.*, 2004]. Avec des mesures additives toutes les fonctions d'agrégation peuvent être employées. Les mesures semi-additives représentent des valeurs instantanées telles que des températures, des âges de personnes, etc. Effectuer une somme sur des mesures semi-additives est inapproprié ; par exemple, faire la somme des relevés mensuels de température pour obtenir une valeur annuelle n'a pas de sens.
- Les *mesures textuelles* sont composées de données textuelles (chaîne de caractères) pouvant être des termes (ou mots), un ensemble de termes (paragraphes, etc) voire des documents complets. Selon ces différentes possibilités, le processus d'agrégation peut différer.
  - Une *mesure textuelle brute* est une mesure dont le contenu correspond au contenu textuel d'un document ou d'un fragment de document (par exemple, le contenu d'un article scientifique au format XML privé de l'ensemble des balises XML).
  - Une *mesure textuelle élaborée* est obtenue par un prétraitement sur une mesure textuelle brute. Par exemple, une mesure textuelle élaborée de type mots-clefs est composée uniquement de termes spécifiques extraits des documents. Ce type de mesure est obtenu en appliquant des traitements sur une mesure textuelle brute tels que le retrait de mots vides tout en conservant les mots les plus significatifs en prenant en compte le domaine du document.

Exemple. Considérons une BDM supportant l'analyse des publications scientifiques réalisées par les chercheurs. Nous représentons une telle BDM par une constellation réduite à un fait (étoile) et deux dimensions.

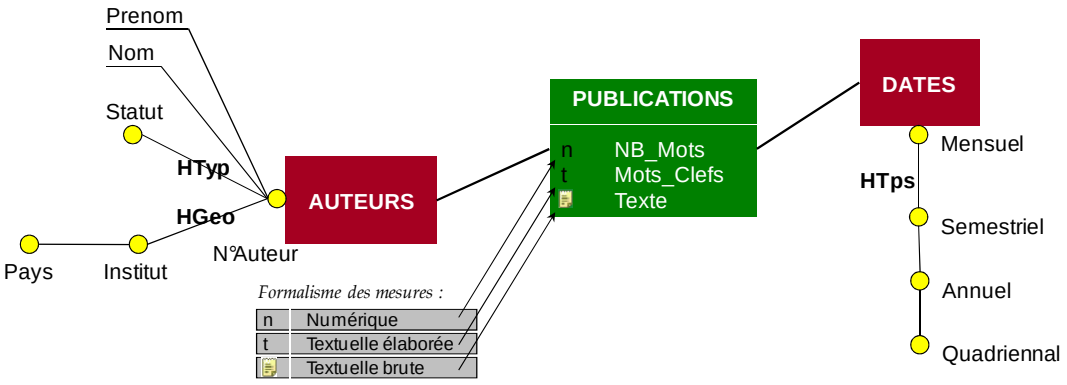


Figure 20 : Exemple d'une constellation textuelle.

En fonction du type de la mesure, toutes les fonctions d'agrégation ne sont pas applicables comme le montre le Tableau 7. La fonction LIST est la fonction identité qui liste toutes les valeurs à agréger. Les fonctions dédiées aux mesures textuelles TOP_KW [Ravat, Teste, Tournier, Zurfluh, 2008c] et AVG_KW [Ravat, Teste, Tournier, 2007d] constituent un des résultats majeurs de nos travaux ; nous détaillons ces fonctions à la section 6 de ce chapitre.

Tableau 7 : Agrégations opérables en fonction du type de mesure.

Typologie des mesures		Fonctions opérables
Numérique	Additive	SUM, AVG, MAX, MIN, COUNT, LIST
	Semi-additive	COUNT, LIST
Textuelle	Brute	COUNT, LIST, TOP_KW
	Elaborée	COUNT, LIST, AVG_KW

3.2 Description des documents par des dimensions spécifiques

La seconde extension induite par la prise en compte des documents concerne la mise en place de dimensions spécifiques : elles sont dédiées à la description d'informations inhérentes aux documents. Plusieurs propositions ont été faites dans la littérature [Mothe, *et al.*, 2003] [Tseng, *et al.*, 2006] [Ravat, Teste, Tournier, 2007e] en fonction de l'origine des données extraites. La modélisation de ces dimensions reste classiquement constituée avec des hiérarchies, des paramètres et des attributs faibles. La contribution de nos travaux concerne la définition de dimensions modélisant la structure des documents. Une dimension « structure » est constituée à partir des structures extraites des documents via la structure arborescente des documents XML (DTD ou XSchema). Chaque paramètre de ces dimensions modélise les différents niveaux de granularité d'un même document. A savoir, la zone de données textuelles qui servira aux fonctions d'agrégation textuelles spécifiques. Ces dimensions modélisent à la fois la structure générique (section, sous-section, paragraphe…) mais aussi la structure spécifique décrivant des types de section comme l'introduction ou la conclusion, ou bien, décrivant des types de paragraphes tels que les définitions, etc. La structure est extraite depuis le balisage qui délimite les éléments dans les documents XML.

Exemple. Reconsidérons la BDM relative à l'analyse de publications scientifiques. On intègre le contenu d'une collection XML d'articles scientifiques. Ces articles sont considérés valides au regard d'une structure commune (DTD) et contiennent un certain nombre de méta-données telles que le nom des auteurs, la date de publication. Les articles sont organisés suivant une structure arborescente. Nous détaillons les prétraitements nécessaires à l'intégration d'une collection de documents dans [Ravat, Teste, Tournier, Zurfluh, 2008f].

La Figure 21 étend la constellation textuelle décrite dans l'exemple précédent ; elle est constituée de quatre dimensions. Le nouveau schéma offre la possibilité d'analyser les méta-données et les mots-clefs des documents comme dans les propositions de [McCabe, *et al.*, 2000], [Mothe, *et al.*, 2003], [Keith, *et al.*, 2005], [Tseng, *et al.*, 2006]. En plus, il permet aussi d'analyser le contenu de documents avec l'emploi de la structure. La dimension STRUCTURES modélise la structure des documents de la collection : chaque paramètre désigne une granularité (paragraphe, sous-section, section) selon laquelle peut être observée la mesure textuelle nommée TEXTE. Cette dimension STRUCTURES est construite à partir de la DTD des fichiers XML représentant les documents [Ravat, Teste, Tournier, Zurfluh, 2008f] tandis que la mesure représente le texte contenu dans les documents (*cf.* Figure 22).

L'intégration de dimensions dédiées à la représentation de la structure des documents induit une problématique d'incomplétude des hiérarchies [Malinowski, *et al.*, 2006]. Il s'agit d'un problème dû à l'hétérogénéité des documents qui provoque l'absence de certaines valeurs ; par exemple dans la Figure 22, le paragraphe P2 de la section S1 n'appartient à aucune sous-section tandis que la hiérarchie prévoit qu'un paragraphe appartient à une sous-section qui elle-même est placée dans une section. Cette variabilité est

habituellement gérée par le niveau logique en utilisant des valeurs virtuelles ; par exemple, le paragraphe P2 de la section S1 est placé dans la sous-section virtuelle S1.0.

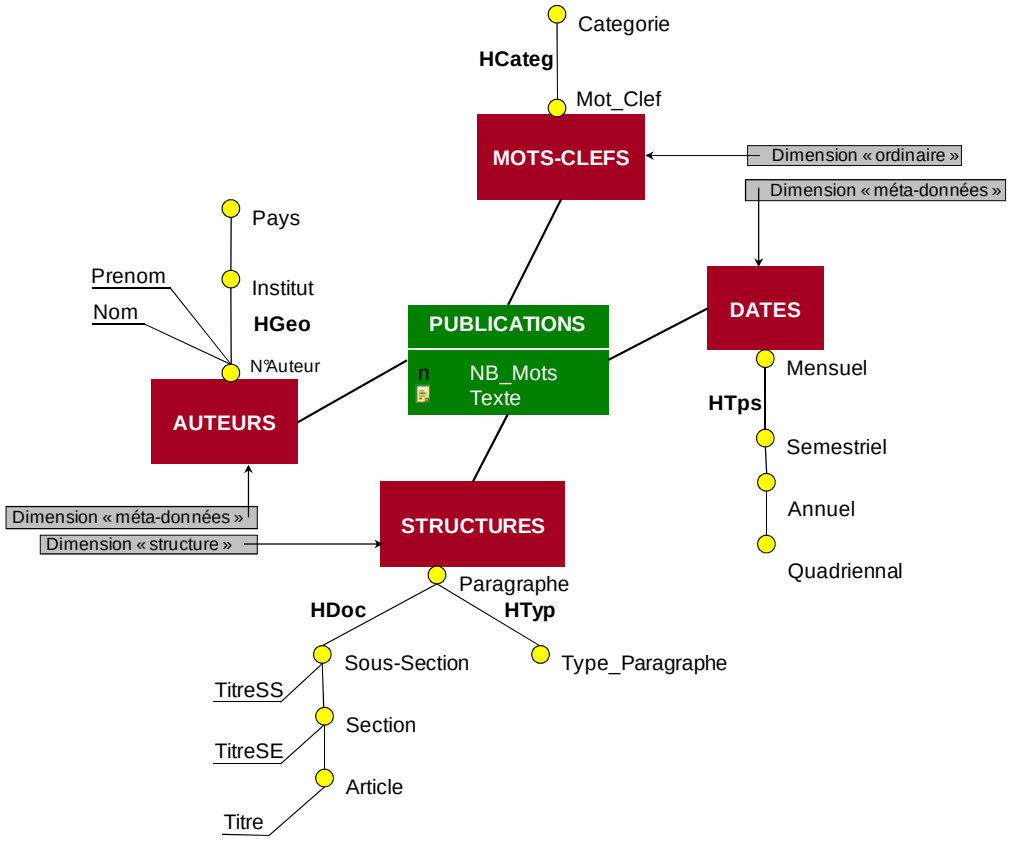
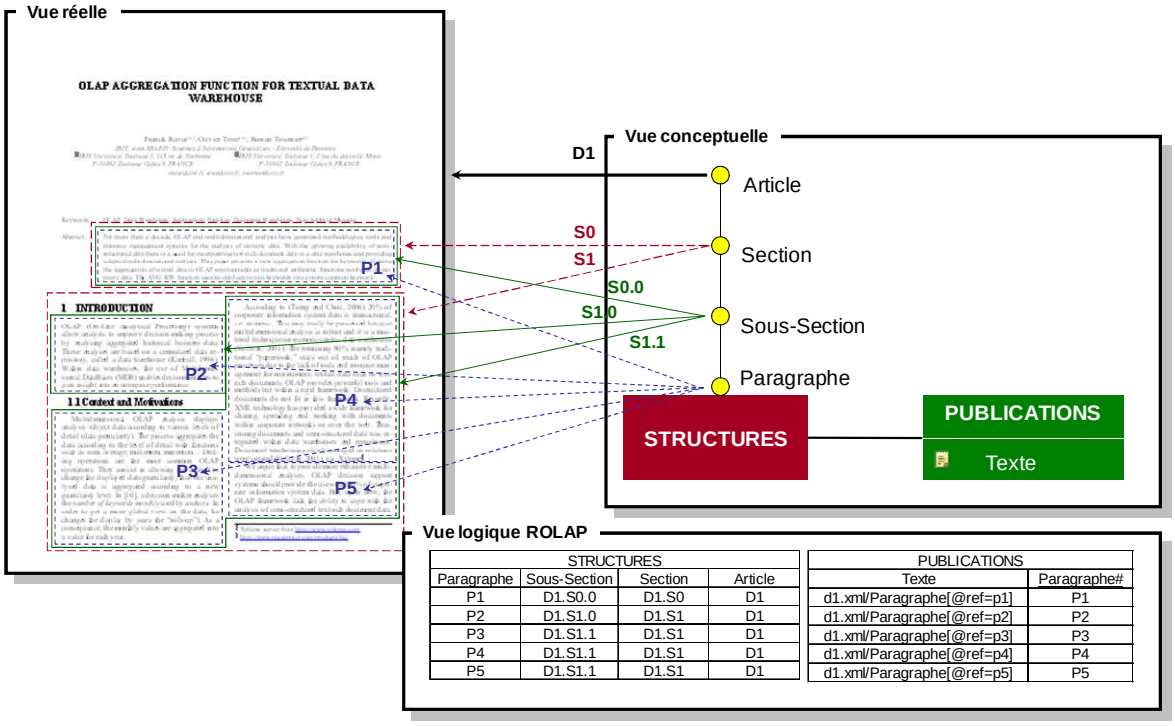


Figure 21 : Exemple d'une constellation textuelle.



STRUCTURES			
Paragraphe	Sous-Section	Section	Article
P1	D1.S0.0	D1.S0	D1
P2	D1.S1.0	D1.S1	D1
P3	D1.S1.1	D1.S1	D1
P4	D1.S1.1	D1.S1	D1
P5	D1.S1.1	D1.S1	D1

PUBLICATIONS	
Texte	Paragraphe#
d1.xml/Paragraphe[@ref=p1]	P1
d1.xml/Paragraphe[@ref=p2]	P2
d1.xml/Paragraphe[@ref=p3]	P3
d1.xml/Paragraphe[@ref=p4]	P4
d1.xml/Paragraphe[@ref=p5]	P5

Figure 22 : Exemple d'une dimension « structures ».

L'approche par extensions de modèles conserve autant que possible les mécanismes de description existants, évitant ainsi de redéfinir les opérations de manipulation OLAP. Cette approche se heurte néanmoins à des solutions pouvant s'avérer difficilement utilisables. Une difficulté importante concerne le choix de définir une donnée en tant que sujet ou axe de la BDM : dans les exemples précédents, les mots-clefs ont été modélisés soit par une mesure textuelle élaborée, soit par une dimension ordinaire, alors que les mots-clefs peuvent jouer le rôle de paramètre ou de mesure en fonction de l'analyse.

4 Nouvelle modélisation

Pour pallier les difficultés des modèles par extensions, nous avons développé un nouveau modèle appelé modèle en « Galaxie » [Tournier, 2007] [Ravat, Teste, Tournier, Zurfluh, 2007a]. Pour solutionner le problème de dualité que jouent certaines données au sein des documents, **notre modélisation en galaxie repose sur l'idée originale d'utiliser un concept unique pour représenter les données qui peuvent être employé de manière symétrique en tant que sujet ou axe d'analyse.** L'objectif de cette simplification est de rendre la définition de la BDM plus souple pour le concepteur.

4.1 Concept de galaxie

Définition. Une *galaxie G* est définie par (D^G; $Star^G$; Lk^G) où

– $D^G = \{D_1, \ldots, D_n\}$ est un ensemble de dimensions,

– $Star^G : D^G \rightarrow 2^{D^G}$ est une fonction qui associe chaque dimension $D_i \in D^G$ à ces dimensions liées $D_{j_x \neq i} \in D^G$.

$$Star^G : D^G \rightarrow 2^{D^G}$$
$$D_i \rightarrow Star^G(D_i) = \{ D_{j_1}, \ldots, D_{j_n} \}$$

– $Lk^G = \{l_1, \ldots, l_u\}$ est un ensemble de liens intra ou inter-documents.

On note :

– $D^{c_k} = \{D_i \mid \exists\ D_{j \neq i} \in D^G,\ D_i \in Star^G(D_j) \wedge D_j \in Star^G(D_i)\}$ un ensemble de dimensions toutes reliées entre elles deux à deux. D^{c_k} représente un sous-graphe complet appelé clique du graphe (ensemble de sommets adjacents) défini par $Star^G$.

– D^C l'ensemble des cliques d'une galaxie obtenues à partir de $Star^G$.

Ainsi nous distinguons les dimensions partagées des dimensions non partagées :

– Une dimension non partagée *Di* est une dimension n'apparaissant que dans une seule clique D^{c_k} c'est-à-dire, si $Di \in D^{c_k}$ alors $\nexists\ D^{c_l} \in D^C\ (l \neq k) \mid Di \in D^{c_l}$.

– Une dimension partagée *Di* est une dimension apparaissant dans au moins deux cliques D^{c_k} et D^{c_l} c'est-à-dire si $Di \in D^{c_k}$ alors $\exists\ D^{c_l} \in D^C\ (l \neq k) \mid Di \in D^{c_l}$.

Exemple. Considérons la BDM relative à l'analyse des publications scientifiques réalisées par les chercheurs. Nous utilisons le concept de galaxie pour décrire les *chercheurs* auteurs de *publications* scientifiques parues dans des *conférences* à différentes *dates* ainsi que les *projets* auxquels participent ces *chercheurs* appartenant aux *instituts* de recherche. La figure suivante décrit la galaxie formée de six dimensions organisées selon deux cliques.

D^G = {CONFERENCES, ARTICLES, AUTEURS, DATES, INSTITUTS, PROJETS} ;

$Star^G$ = {(CONFERENCES, {ARTICLES, AUTEURS, DATES}),
(ARTICLES, {CONFERENCES, AUTEURS, DATES}),

(AUTEURS, {CONFERENCES, ARTICLES, DATES, INSTITUTS, PROJETS}),
(DATES, {CONFERENCES, ARTICLES, AUTEURS, INSTITUTS, PROJETS}),
(INSTITUTS, {AUTEURS, DATES, PROJETS}),
(PROJETS, {AUTEURS, DATES, INSTITUTS})} ;

$D^C = \{D^{c1}, D^{c2}\}$ où

D^{c1} = {CONFERENCES, ARTICLES, AUTEURS, DATES} ;

D^{21} = {AUTEURS, DATES, INSTITUTS, PROJETS}}.

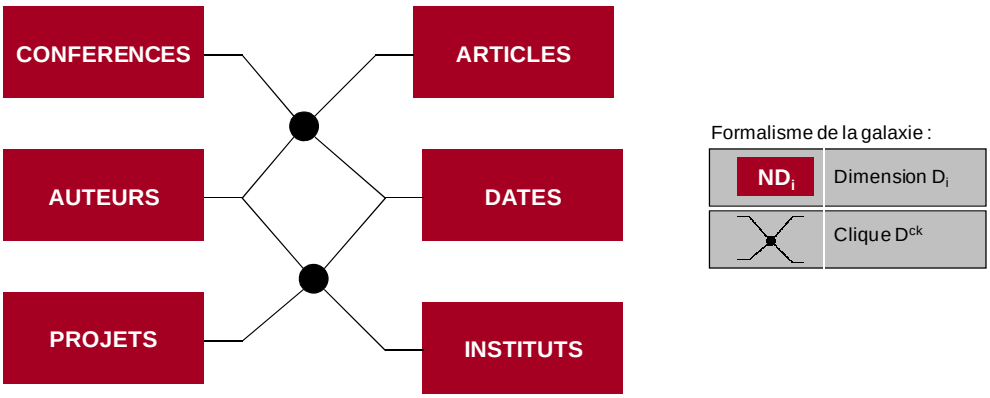


Figure 23 : Exemple d'une galaxie.

4.2 *Concept unique de dimension*

L'originalité de la description des données par une galaxie repose sur le concept unique de « dimension » représentant à la fois un axe d'analyse et un sujet d'analyse. Comme dans les approches classiques, le concept de hiérarchie organise les attributs d'une dimension selon les niveaux de granularité qu'ils représentent (*cf.* chapitre 3).

Définition. Une *dimension* D_i est définie par $(ND_i ; A^{D_i} ; H^{D_i} ; I^{D_i} ; IStar^{D_i})$ où

- ND_i est le nom identifiant la dimension dans la galaxie,
- $A^{D_i} = \{a^{D_i}_1, \ldots, a^{D_i}_r\} \cup \{Id_i, All_i\}$ est un ensemble d'attributs,
- $H^{D_i} = \{H^{D_i}_1, \ldots, H^{D_i}_s\}$ est un ensemble de hiérarchies,
- $I^{D_i} = \{i^{D_i}_1, \ldots, i^{D_i}_t\}$ est un ensemble d'instances de dimension. On note $\forall k \in [1..t]$, $i^{D_i}_k = [Id_i : id_{x_k} ; a^{D_i}_1 : v_{k1}; \ldots ; a^{D_i}_r : v_{kr}; All_i : \text{all}]$.
- $IStar^{D_i} = \bigcup_{\forall D^{ck} \in D^C} IStar^{D_i}_{D^{ck}}$ est l'ensemble des fonctions associant les instances de D_i aux instances des autres dimensions liées, pour chacune des cliques D^{c_k} auxquelles elle appartient. On note $\forall D^{c_k} \in D^C$, $IStar^{D_i}_{D^{ck}} : I^{D_i} \rightarrow 2^{\Im}$ tel que $\Im = \prod_{\forall D_j \in D^{ck} - \{D_i\}} I^{D_j}$ l'ensemble des fonctions associant les instances de D_i aux instances des dimensions de la clique D^{c_k}.

Exemple. La Figure 24 illustre la fonction $IStar^{CONFERENCES}_{D^{c_1}}$ en utilisant les références [Ravat, Teste, Tournier, 2007d] et [Ravat, Teste, Tournier, Zurfluh, 2008c] de ce mémoire.

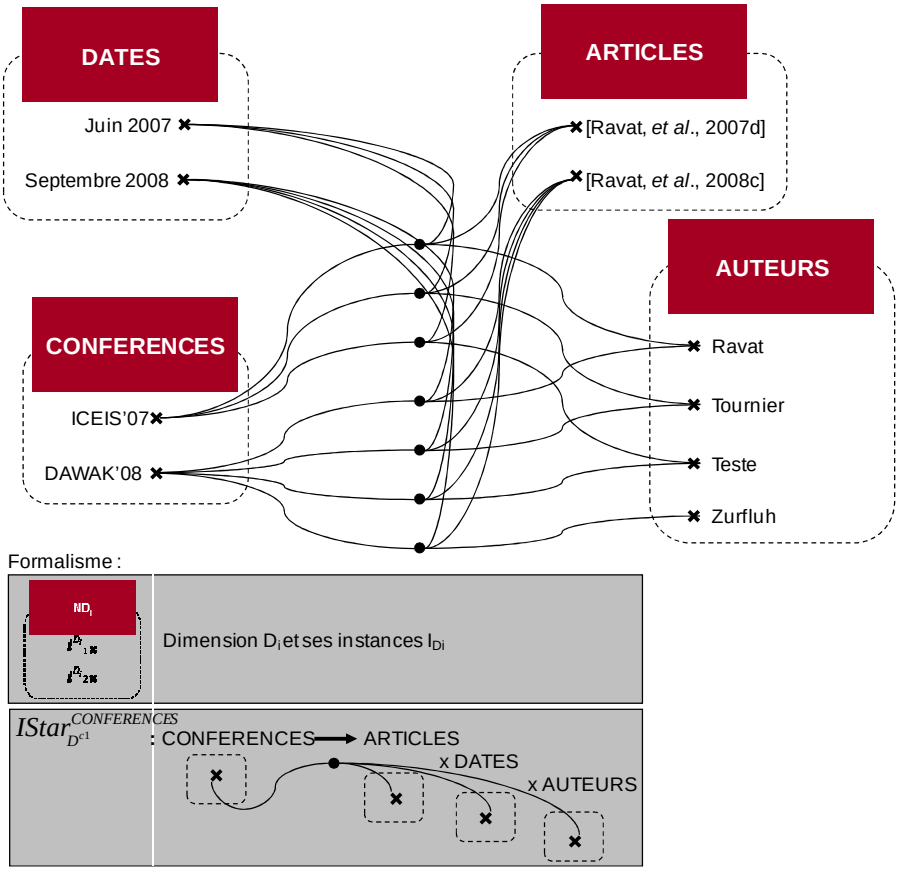


Figure 24 : Exemple d'association des instances des dimensions.

4.3 *Concept de liens de navigation*

Les sources de données comme les documents comportent des liens hypertextes permettant de naviguer entre documents. Nous souhaitons conserver ce cadre « navigationnel » dans la BDM pour exploiter ces liens lors des analyses OLAP et ainsi accroître le potentiel d'analyses en facilitant les corrélations entre les données analysées. Par exemple, pour obtenir les domaines que les auteurs font intervenir dans leurs recherches, il est possible de centrer l'analyse des articles d'un auteur sur la section des références aux autres articles. Ainsi les principaux mots-clefs des articles référencés peuvent être intégrés dans l'analyse en utilisant simplement le lien (un exemple exploitant les liens de navigation est présenté à la section 5.4 du chapitre).

Nous considérons un lien comme étant une relation « correspond à » entre des valeurs de ces deux attributs. Ce lien peut être matérialisé dans les documents sources, comme un renvoi hypertexte vers un autre document.

Définition. Un *lien* l_i est une fonction reliant les valeurs des attributs $a^{D_i}{}_u$ et $a^{D_i}{}_v$:

$$l_i : dom(a^{D_i}{}_u) \rightarrow dom(a^{D_i}{}_v)$$

S'il existe au moins une valeur de $a^{D_i}{}_u$ qui n'est pas liée à au moins une valeur de $a^{D_i}{}_v$, l_i est défini sur un sous-ensemble de $dom(a^{D_i}{}_u)$. La restriction de la fonction l_i, notée $l_{i|P}$, est définie comme suit

$$l_{i|P} : P \rightarrow dom(a^{D_i}{}_v) \text{ où } P \subseteq dom(a^{D_i}{}_u)$$

Exemple. Nous limitons la galaxie aux dimensions ARTICLES, AUTEURS et INSTITUTS. La Figure 25 comporte deux liens $Lk^G = \{l_{References}, l_{Appartenances}\}$, nommés REFERENCES et APPARTENANCES.

- Le lien $l_{References}$ est un lien intra-dimension qui relie deux attributs au sein d'une même dimension. Il est défini par :
  $l_{References}$: *dom(ARTICLES.Paragraphe)* → *dom(ARTICLES.Document).*
- Le lien $l_{Appartenances}$ est un lien inter-dimension qui relie deux attributs de deux dimensions différentes. Il est défini par :
  $l_{Appartenances}$: *dom(AUTEURS.Institut)* → *dom(Instituts.Nom).*

Dans un article, certains paragraphes représentent une référence à un autre article ; il s'agit des paragraphes d'une section de références par exemple. Le lien $l_{References}$ ne concerne donc qu'un sous-ensemble des paragraphes : les paragraphes de type référence (TypeP = 'Ref'). Par conséquent, le lien REFERENCES est une restriction de la fonction $l_{References}$, notée $l_{References|R}$, définie comme suit :

- $l_{References|R}$: $R \rightarrow dom(ARTICLES.Document)$ où
- $R = \{v_i \mid \forall i^{ARTICLES}{}_j \in I^{ARTICLES}, i^{ARTICLES}{}_j = [\ldots; Paragraphe : v_i; \ldots; TypeP : 'Ref'; \ldots]\}$

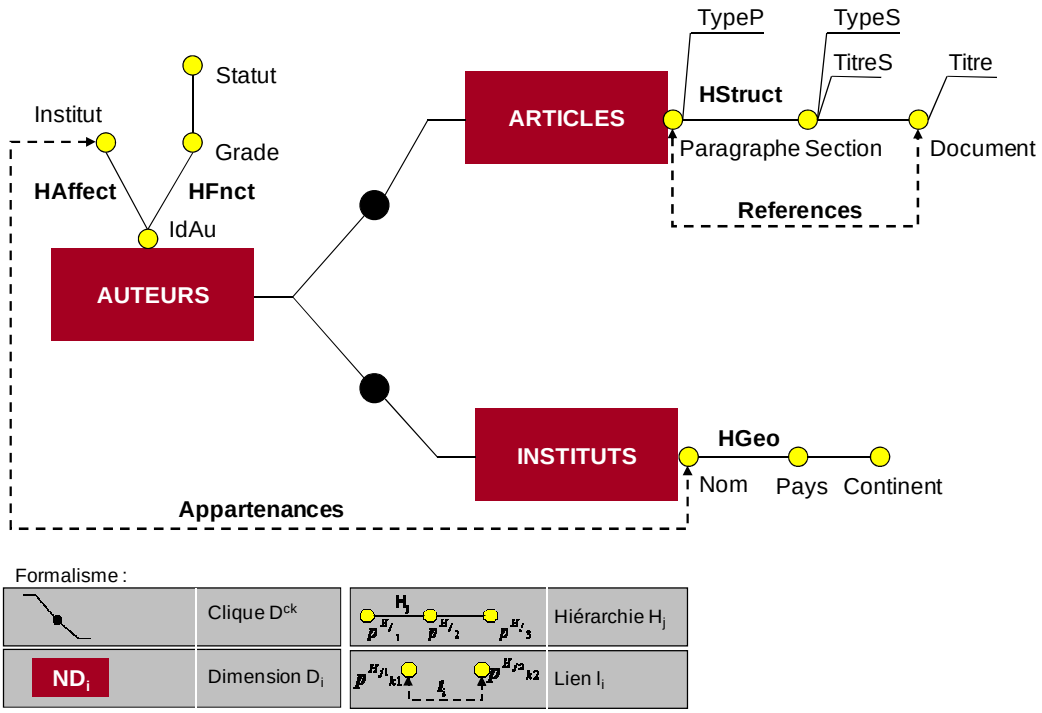


Figure 25 : Exemple de liens de navigation.

5 Manipulation multidimensionnelle des documents

Un aspect important de nos travaux concerne l'étude des manipulations OLAP dans le contexte des BDM intégrant des documents. L'approche par extension a l'avantage de supporter directement les opérations classiques de manipulation et ne nécessite pas une structure spécifique pour visualiser les données tandis que l'approche en galaxie impose de revisiter les opérations de manipulation et d'étendre ces dernières au sein d'une table multidimensionnelle étendue [Tournier, 2007].

Le Tableau 8 compare l'ensemble des opérateurs utilisables sur une constellation textuelle et sur une galaxie. Nous limitons l'étude aux opérateurs les plus emblématiques de la manipulation multidimensionnelle couvrant les besoins essentiels des décideurs :

- l'opération de construction d'une table multidimensionnelle (DISPLAY),
- l'opération de restriction (SELECT) spécialisant l'analyse,
- les deux opérations de forage (DRILLDOWN, ROLLUP) permettant de modifier le niveau de détail des données analysées,
- l'opération de rotation (ROTATE) qui réoriente l'analyse en changeant un axe d'analyse, voire le sujet de l'analyse,
- l'opération d'imbrication (NEST) qui élargit les analyses en imbriquant un attribut en provenance de n'importe quelle dimension liée au sujet analysé.

Tableau 8 : Opérations de manipulation multidimensionnelle des documents.

Constellation textuelle	Galaxie
DISPLAY(F_C, {$f_1(m_1)$,…,$f_t(m_t)$}, *DL*, *HL*, *DC*, *HC*) = T_{RES}	**FOCUS(*DS.HS*, f_i, *DL.HL*, *DC.HC*) = S^G**
DRILLDOWN(T_{SRC}, *D*, p_{inf}) = T_{RES} ROLLUP(T_{SRC}, *D*, p_{sup}) = T_{RES}	DRILLDOWN(S^G, *D*, p_{inf}) = $S^G{}_{RES}$ ROLLUP(S^G, *D*, p_{sup}) = $S^G{}_{RES}$
ROTATE(T_{SRC}, D_{old}, D_{new}, H_{new}) = T_{RES}	ROTATE(S^G, D_{old}, D_{new}, H_{new}) = $S^G{}_{RES}$ **ROTATE(S^G, *DS*, D_{new}, H_{new}, f_i) = $S^G{}_{RES}$**
SELECT(T_{SRC}, *pred*) = T_{RES}	SELECT(S^G, *pred*) = $S^G{}_{RES}$
NEST(T_{SRC}, *D*, p_i, D_{new}, p_{new}) = T_{RES}	NEST(S^G, *D*, p_i, D_{new}, p_{new}) = $S^G{}_{RES}$

Le choix de la galaxie nécessite de redéfinir quelques opérations (FOCUS et ROTATE) mais offre de nouvelles perspectives à l'utilisateur :

- la modélisation d'un sujet par une dimension munie de hiérarchies facilite le *traitement symétrique des paramètres et des indicateurs* d'analyse [Agrawal, *et al.*, 1997] [Cabibbo, *et al.*, 1997] [Gyssens, *et al.*, 1997],
- la modélisation des liens facilite des analyses en *navigant* au sein des dimensions et entre différentes dimensions.

Nous détaillons ces aspects originaux dans les sections suivantes.

5.1 Généralisation de la table multidimensionnelle

Notre étude repose sur une généralisation du concept de table multidimensionnelle. La structure de visualisation des données organisées en galaxie comporte trois dimensions constituant un sujet d'analyse (axe^S) et deux axes d'analyses (axe^x représentant les colonnes et axe^y les lignes) ainsi qu'un ensemble possible de restrictions (*res*) des domaines de valeurs de l'analyse. Cette structure 2D se veut être un compromis entre simplicité et puissance pour visualiser les données d'une BDM [Gyssens, *et al.*, 1997]. Il est possible néanmoins d'élargir la vision 2D par l'opérateur NEST qui permet à un décideur expérimenté d'emboîter en ligne et en colonne des paramètres d'autres dimensions pouvant ainsi construire une analyse multidimensionnelle avec plus de deux dimensions (se reporter à l'exemple présenté au Tableau 9).

Définition. Une table multidimensionnelle généralisée S^G est définie par (axe^S, axe^x, axe^y, R) où :

– $axe^S = (DS, HS, PS)$ est le sujet analysé $DS \in D^G$ où une hiérarchie courante est désignée $HS \in H^{DS}$ et un ensemble ordonné d'attributs est défini PS,

– $axe^x = (DL, HL, PL)$ est l'axe d'analyse $DL \in Star^G(DS)$ utilisé en entête de lignes où une hiérarchie courante est désignée $HL \in H^{DL}$ et un ensemble ordonné d'attributs de la dimension est défini PL,

– $axe^y = (DC, HC, PC)$ est l'axe d'analyse $DC \in Star^G(DS)$ utilisé en entête de colonnes,

– R est un prédicat normalisé (conjonction de disjonctions).

La particularité de cette structure est la hiérarchisation du sujet d'analyse axe^S qui se traduit par la division des cellules dans la table bidimensionnelle en « sous-cellules ». L'intérêt de représenter le sujet d'analyse par une dimension hiérarchisée réside dans la capacité octroyée ainsi à la structure de supporter les opérations de forage non seulement sur les axes d'analyse mais aussi sur le sujet d'analyse.

Exemple. La figure suivante présente une table multidimensionnelle étendue. Elle permet de visualiser le principal mot-clef des articles publiés par les auteurs de l'institut de recherche IRIT par période quadriennale.

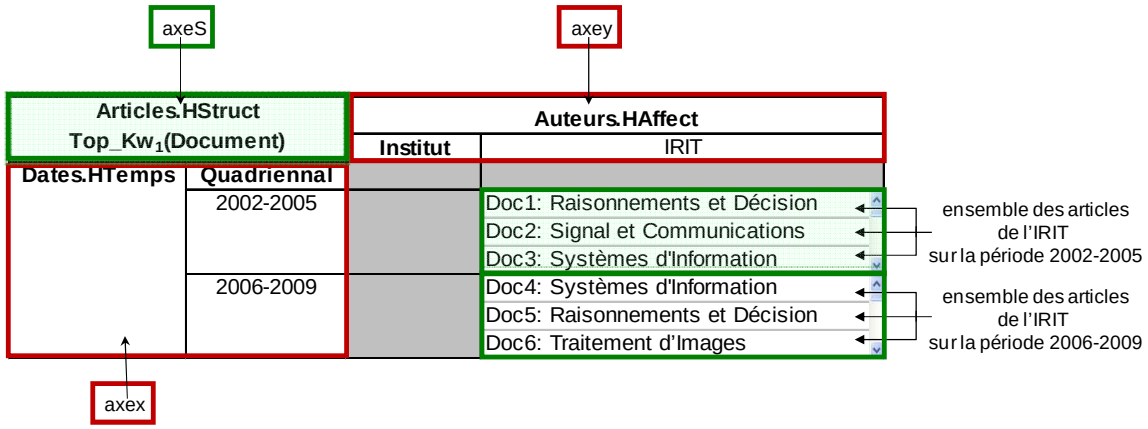


Figure 26 : Exemple d'une table multidimensionnelle généralisée.

Les sous-sections suivantes présentent le constructeur de table généralisée (section 5.2), les extensions des opérateurs OLAP à la galaxie (section 5.3) et l'exploitation des liens entre les documents (section 5.4). En effet, au-delà de l'unicité du concept de représentation des données en galaxie, notre modèle vise à conserver au sein de la BDM les spécificités des documents au travers des dimensions de type STRUCTURE et des liens de navigation.

5.2 *Opération de construction*

A partir de la représentation en galaxie, le décideur peut spécifier des analyses multidimensionnelles par l'intermédiaire d'une opération de focalisation. L'utilisateur focalise l'analyse sur un sujet et projette les données du sujet sur plusieurs axes d'analyse. Les données projetées sont agrégées par une fonction d'agrégation pour en donner une vision synthétique.

Définition 4. La *focalisation* est définie par l'opérateur

FOCUS(*DS.HS*, f_i, *DL.HL*, *DC.HC*) = S^G

- $DS \in D^G$ est le sujet de l'analyse avec la hiérarchie courante $HS \in H^{DS}$,
- f_i est une fonction d'agrégation calculant les valeurs des cellules,
- $DL \in Star^G(DS)$ est l'axe en ligne muni d'une hiérarchie courante $HL \in H^{DL}$,
- $DC \in Star^G(DS)$ est l'axe en colonne muni d'une hiérarchie courante $HC \in H^{DC}$,
- $S^G = (axe^S, axe^x, axe^y, R)$ est la table résultat telle que
  - $axe^S = (DS, HS, <All, f_i(p_{Smax})>)$ où p_{Smax} désigne le paramètre extrémité HS,
  - $axe^x = (DL, HL, <All, p_{Lmax}>)$,
  - $axe^y = (DC, HC, <All, p_{Cmax}>$ *et*
  - $R = DS.All='all' \wedge \left(\bigwedge_{\forall D_i \in Star^G(DS)} D_i.All = 'all' \right)$.

Exemple. Au sein de la galaxie, décrite à la section précédente, le décideur peut sélectionner n'importe quelle dimension en tant que sujet d'analyse. Supposons qu'il focalise son analyse sur l'activité de publication (dimension *Articles*) des chercheurs (dimension *Auteurs*) au cours du temps (dimension *Temps*). Le résultat de l'opération de focalisation ci-dessous est présenté Figure 26.

FOCUS(Articles.HStruct, TOP_KW1, Dates.HTemps, Auteurs.HAffect) = S^{G_1}

La fonction *Top_KW1* utilisée dans l'opération de focalisation permet d'agréger les articles en visualisant les mots-clefs caractérisant ces derniers (k=1 limite l'affichage à un mot-clef) ; une présentation détaillée de la fonction se trouve section 6.1 du chapitre.

5.3 Traitement symétrique des données

Le modèle en galaxie nécessite de revisiter certains opérateurs OLAP mais procure le grand avantage de faciliter le traitement symétrique des données.

- Les opérations de rotation ne sont applicables dans les modèles en constellation qu'aux axes d'analyse (axe^x et axe^y), nécessitant le développement d'une opération spécifique pour appliquer la rotation sur le sujet d'analyse [Abelló, *et al*., 2003] ou pour transformer les axes d'analyse en sujet d'analyse et inversement [Agrawal, *et al*., 1997] [Gyssens, *et al*., 1997]. Dans le cadre d'une galaxie, ces opérations sont généralisées par l'opération ROTATE qui s'applique symétriquement sur des dimensions en tant que sujet ou axe d'analyse.
- Les opérations de forage peuvent être appliquées non seulement sur les axes en ligne (axe^x) et en colonne (axe^y) mais également sur l'axe « focalisé » (axe^S) puisque ce dernier est muni d'une hiérarchie courante. Cette fonctionnalité donne une nouvelle vision du processus d'agrégation en effectuant des agrégations à un nouveau niveau de détails.

Exemple. Le Tableau 9 présente une séquence d'opérations multidimensionnelles réalisée par un décideur.

Tableau 9 : Séquence de manipulation multidimensionnelle des documents.

<table>
<tr><th>Opération</th><th>Résultat</th></tr>
<tr><td>DRILLDOWN
(S^{G_1}, Auteurs, IdAu)
=
S^{G_2}</td><td>
<table>
<tr><td colspan="2" rowspan="2">Articles.HStruct
Top_Kw$_1$(Document)</td><td colspan="4">Auteurs.HAffect</td></tr>
<tr><td>Institut</td><td colspan="3">IRIT</td></tr>
<tr><td colspan="2">IdAu</td><td>Teste</td><td>Tournier</td><td>…</td></tr>
<tr><td rowspan="6">Dates.HTemps</td><td rowspan="3">Quadriennal</td><td rowspan="6"></td><td colspan="3"></td></tr>
<tr><td>2002-2005</td><td>Doc3: Système d'Information</td><td></td><td></td></tr>
<tr><td></td><td>Doc7: Entrepôt</td><td></td><td></td></tr>
<tr><td></td><td>Doc9: Manip. OLAP</td><td></td><td></td></tr>
<tr><td rowspan="3">2006-2009</td><td>Doc8: Entrepôt XML</td><td>Doc8: Entrepôt XML</td><td></td></tr>
<tr><td>Doc12: Entrepôt</td><td>Doc10: Document OLAP</td><td></td></tr>
<tr><td>Doc10: Document OLAP</td><td>Doc11: Manip. OLAP</td><td></td></tr>
</table>
</td></tr>
<tr><td>SELECT
(S^{G_2}, Auteurs.Nom = 'Teste')
=
S^{G_3}</td><td>
<table>
<tr><td colspan="2" rowspan="2">Articles.HStruct
Top_Kw$_1$(Document)</td><td colspan="2">Auteurs.HAffect</td></tr>
<tr><td>Institut</td><td>IRIT</td></tr>
<tr><td colspan="2">IdAu</td><td>Teste</td></tr>
<tr><td rowspan="6">Dates.HTemps</td><td rowspan="3">Quadriennal
2002-2005</td><td rowspan="6"></td><td>Doc3: Système d'Information</td></tr>
<tr><td>Doc7: Entrepôt</td></tr>
<tr><td>Doc9: Manip. OLAP</td></tr>
<tr><td rowspan="3">2006-2009</td><td>Doc8: Entrepôt XML</td></tr>
<tr><td>Doc12: Entrepôt</td></tr>
<tr><td>Doc10: Document OLAP</td></tr>
<tr><td colspan="4">Auteurs.Nom = 'Teste'</td></tr>
</table>
</td></tr>
<tr><td>DRILLDOWN
(S^{G_3}, Articles, Section)
=
S^{G_4}</td><td>
<table>
<tr><td colspan="2" rowspan="2">Articles.HStruct
Top_Kw$_1$(Document:Section)</td><td colspan="2">Auteurs.HAffect</td></tr>
<tr><td>Institut</td><td>IRIT</td></tr>
<tr><td colspan="2">IdAu</td><td>Teste</td></tr>
<tr><td>Dates.HTemps</td><td>Quadriennal</td><td></td><td></td></tr>
<tr><td rowspan="6">Dates.HTemps</td><td rowspan="3">2002-2005</td><td rowspan="6"></td><td>Doc3:S1: Décisionnel</td></tr>
<tr><td>Doc3:S2: BDM</td></tr>
<tr><td>…</td></tr>
<tr><td rowspan="3">2006-2009</td><td>Doc7:S1: BD Décisionnelle</td></tr>
<tr><td>Doc8:S1: Entrepôt XML</td></tr>
<tr><td>Doc8:S2: BDM</td></tr>
<tr><td colspan="4">Auteurs.Nom = 'Teste'</td></tr>
</table>
</td></tr>
<tr><td>ROTATE
(S^{G_4}, Articles, Conférences, HConf, AVG)
=
S^{G_5}</td><td>
<table>
<tr><td colspan="2" rowspan="2">Conférences.HConf
AVG(Tx_Accept)</td><td colspan="2">Auteurs.HAffect</td></tr>
<tr><td>Institut</td><td>IRIT</td></tr>
<tr><td colspan="2">IdAu</td><td>Teste</td></tr>
<tr><td rowspan="3">Dates.HTemps</td><td>Quadriennal</td><td></td><td></td></tr>
<tr><td>2002-2005</td><td></td><td>15</td></tr>
<tr><td>2006-2009</td><td></td><td>23</td></tr>
<tr><td colspan="4">Auteurs.Nom = 'Teste'</td></tr>
</table>
</td></tr>
<tr><td>NEST(S^{G_5}, Dates, Quadriennal, Conférences, IdC)
= S^{G_6}</td><td>
<table>
<tr><td colspan="3" rowspan="2">Conférences.HConf
AVG(Tx_Accept)</td><td colspan="2">Auteurs.HAffect</td></tr>
<tr><td>Institut</td><td>IRIT</td></tr>
<tr><td colspan="3">IdAu</td><td>Teste</td></tr>
<tr><td>Dates.HTemps</td><td>Quadriennal</td><td>Conférences.IdC</td><td></td><td></td></tr>
<tr><td rowspan="4">Dates.HTemps</td><td>2002-2005</td><td>ICEIS'03</td><td></td><td>15</td></tr>
<tr><td rowspan="3">2006-2009</td><td>ER'07</td><td></td><td>22</td></tr>
<tr><td>DAWAK'08</td><td></td><td>33</td></tr>
<tr><td>ICEIS'09</td><td></td><td>13</td></tr>
<tr><td colspan="5">Auteurs.Nom = 'Teste'</td></tr>
</table>
</td></tr>
</table>

– Le décideur détaille dans un premier temps l'analyse en visualisant les données pour chaque auteur (DRILLDOWN).

- Il centre son analyse à un auteur particulier en le sélectionnant (SELECT).
- Ensuite, il détaille à nouveau l'analyse en opérant un forage sur le sujet analysé pour visualiser le détail des mots clefs caractérisant les sections des articles de l'auteur (DRILLDOWN). Cette nouvelle fonctionnalité est disponible puisque la galaxie modélise le sujet de l'analyse comme une dimension hiérarchisée.
- Il change le sujet analysé en observant le taux d'acceptation moyen des publications de l'auteur (ROTATE). Cette opération, traditionnellement disponible pour les seuls axes d'analyse, s'applique de manière symétrique au sein d'une galaxie entre le sujet d'analyse et les axes. L'opération de rotation appliquée sur le sujet de l'analyse impose la spécification de la nouvelle fonction d'agrégation ; cette extension de l'opération classique permet une plus grande souplesse, sinon l'opération ne pourrait s'appliquer qu'en cas de compatibilité entre l'ancienne et la nouvelle mesure.
- Enfin, pour comprendre la hausse du taux d'acceptation, le décideur imbrique l'identifiant des conférences en entête de ligne (NEST) obtenant ainsi une analyse tridimensionnelle. Il observe alors un accroissement des publications, avec des taux d'acceptation plus ou moins faibles.

L'opération de forage s'applique sur le sujet d'analyse sans extension, mais se heurte à une difficulté importante dans le cadre des données textuelles : les fonctions d'agrégations textuelles n'opèrent pas de la même manière que les fonctions d'agrégations numériques. Ainsi, l'extraction des principaux mots-clefs d'une section ne correspond pas nécessairement à l'extraction des principaux mots-clefs des articles. Ceci est un problème connu des fonctions holistiques [Gray, *et al*., 1996] qui ne peuvent être calculées à partir de résultats intermédiaires (par exemple la fonction qui calcule la médiane).

5.4 Navigation au sein des données

Le modèle en galaxie intègre dans la structure multidimensionnelle les liens existant dans les documents sources pour être exploités durant les analyses. Les liens au sein de la galaxie sont principalement employés pour accéder à un ensemble particulier de données. Ceci permet une plus grande flexibilité lors de la désignation de sous éléments de documents.

Exemple. Afin d'avoir une meilleure vision des domaines que les auteurs font intervenir dans leurs recherches, il est possible de construire une analyse des mots-clefs des articles référencés par les publications des chercheurs. Dans la séquence suivante, les publications d'un auteur sont analysées. L'analyse porte sur un auteur particulier : les mots-clefs des articles référencés dans les publications (articles) de l'auteur sont analysés.

- Focalisation de l'analyse sur les articles par auteurs et périodes quadriennales. FOCUS(Articles.HStruct, TOP_KW1, Dates.HTemps, Auteurs.HAffect) = S^{G_7}
- Restriction de l'analyse à un auteur et forage sur l'auteur en colonne. DRILLDOWN(SELECT(S^{G_7}, Auteurs.Nom = 'Teste'), Auteurs, IdAu) = S^{G_8}
- Restriction de l'analyse aux références bibliographiques des articles. SELECT(S^{G_8}, Articles.TypeS = 'Reference') = S^{G_9}
- Forage sur le sujet analysé pour observer les mots-clefs des articles référencés par les articles de l'auteur (vision thématique des travaux connexes aux recherches de l'auteur).
  DRILLDOWN(S^{G_9},Articles,Paragraphe.References.HStruct.Document)=$S^{G_{10}}$

Pour accéder aux articles référencés, le lien REFERENCES est exploité dans le forage au travers de l'expression suivante : Paragraphe.References.HStruct.Document.

La Figure 27 présente le résultat obtenu à partir de la séquence d'opérations.

Articles.HStruct Top_Kw $_1$(Document: Paragraphe.References.Document)		Auteurs.HAffect	
		Institut	IRIT
		IdAu	Teste
Dates.HTemps	Quadriennal		
	2002-2005		Doc3:P34:Doc24: Data Warehousing
			Doc3:P35:Doc33: Modelling
			…
			Doc7:P25:Doc24: Data Warehousing
	2006-2009		Doc8:P53:Doc24: Data Warehousing
			Doc8:P55:Doc14: OLAP
			…
			Doc12:P30:Doc40: OLAP Modelling
Auteurs.Nom = 'Teste' $\wedge$ Articles.TypeS = 'Reference'			

Figure 27 : Exemple d'une analyse avec navigation sur un lien.

6 Agrégation de données textuelles

La manipulation multidimensionnelle, notamment le forage, nécessite des fonctions d'agrégation pour synthétiser les informations textuelles. Ces fonctions d'agrégation sont opérantes sur les valeurs numériques mais inopérantes sur du texte (*cf.* Figure 19). Un résultat important de nos recherches [Pujolle, Ravat, Teste, Tournier, 2008] a été de fournir une nouvelle approche pour l'agrégation de données textuelles dans un environnement pour l'analyse OLAP du contenu de documents XML. Nos travaux ont abouti à la définition de deux fonctions d'agrégation, l'une adaptée aux mesures brutes, l'autre dédiée aux mesures élaborées (*cf.* Tableau 7) :

- *TOP_KW$_k$* [Ravat, Teste, Tournier, Zurfluh, 2008c, 2008d] permet l'agrégation d'un ensemble de documents en ses k termes les plus représentatifs,
- *AVG_KW* [Ravat, Teste, Tournier, 2007d] permet de résumer un ensemble de mots-clefs issus d'un vocabulaire contrôlé par un ensemble limité de termes plus généraux. Cette fonction repose sur une ontologie légère de domaine.

Nous détaillons dans la suite les principes développés par ces deux fonctions d'agrégation dédiées aux données textuelles.

6.1 *Fonctions Top_Kw$_k$*

La fonction d'agrégation *TOP_KW$_k$* extrait les k termes les plus représentatifs d'une mesure textuelle brute constituée de n termes (ou mots).

Définition. La fonction *TOP_KW$_k$* est définie par

$$Top\_Kw_k : \quad W^n \longrightarrow T^k$$
$$\{w_1, \ldots, w_n\} \mapsto \langle t_1, \ldots, t_k \rangle$$

- W^n est un ensemble de n termes,
- $T^k \subseteq W^n$ est un sous-ensemble ordonné des k termes les plus représentatifs.

Afin de déterminer les k termes les plus représentatifs, nous avons adapté au contexte OLAP des techniques bien maîtrisées en Recherche d'Information (RI) [Baeza-Yates, *et al.*, 1999] [Boughanem, 2000] [Mothe, 2000] [Soulé-Dupuy, 2001] qui ordonnancent les termes selon leur représentativité en fonction de poids. Pour ce faire, il est nécessaire de connaître la

représentativité d'un terme vis-à-vis de la collection (intégralité des autres documents). Dans le contexte OLAP, il n'est pas nécessaire de connaître cette représentativité vis-à-vis de la collection complète, mais plutôt selon les documents qui seront agrégés par la fonction. Le problème est alors d'opérer sur une liste variable de documents qui change à chaque manipulation multidimensionnelle.

La restitution d'une analyse OLAP est effectuée au moyen d'une table multidimensionnelle (*cf.* chapitre 3). Chaque cellule c_{ij} de la TM stocke la valeur issue de l'agrégation des valeurs de la mesure analysée en fonction de la valeur des attributs en i^{ème} ligne et en j^{ème} colonne. A chaque cellule c_{ij} correspond un ensemble de documents (ou de fragments de documents) D_{ij} composé de d_{ij} documents et un nombre total $n_{ij}(d)$ de termes dans chacun des documents $d \in D_{ij}$. Au sein de chaque cellule, des poids sont assignés à chacun des termes des d_{ij} documents de D_{ij}. Afin « d'ordonner » ces termes, nous employons la fonction de pondération *tf.idf* [Robertson, 2004] qui correspond au produit entre la représentativité d'un terme vis-à-vis d'un document (*tf* : « term frequency ») avec l'inverse de sa représentativité vis-à-vis de l'ensemble des autres documents de la collection (*idf* : « inverse document frequency »). La fonction a été adaptée à notre contexte, c'est-à-dire que l'*idf* est calculé seulement vis-à-vis de l'ensemble des documents de la cellule c_{ij} concernée. Ainsi, pour chaque cellule c_{ij} et chaque terme t correspond un nombre d'occurrences $n_{ij}(d,t)$ du terme t dans le document d de c_{ij} et un nombre de documents $d_{ij}(t)$ qui contiennent le terme t, parmi les documents de c_{ij} : $d_{ij}(t) \leq d_{ij}$. Ainsi notre fonction repose sur les calculs suivants :

- $tf_{ij}(d,t) = \frac{n_{ij}(d,t)}{n_{ij}(d)}$ représente le nombre de fois où le terme t est présent dans un fragment de texte normalisé par le nombre total de termes du fragment.

- $idf_{ij}(t) = \log \frac{d_{ij}+1}{d_{ij}(t)}$ est l'inverse du rapport entre le nombre de documents qui contiennent le terme t et le nombre total de documents contenus dans la cellule c_{ij}. Le 1 assure que $d_{ij} > d_{ij}(t)$, *i.e.* $idf = 0$, cas rare mais possible dans les TM.

- $w_{ij}(d,t) = tf_{ij}(d,t) \times idf_{ij}(t)$ est le poids du terme t pour un document d de c_{ij}.

- $w_{ij}(t) = \sum_{d} w_{ij}(d,t)$ est le poids du terme t pour la cellule c_{ij}.

Les termes pondérés sont alors ordonnés dans une liste L_{ij}, selon leur poids : $L_{ij} = \langle t_1,\ldots, t_n \rangle \mid w_{ij}(t_1) > w_{ij}(t_2) > \ldots > w_{ij}(t_n)$. La fonction d'agrégation TOP_KW$_k$ extrait ensuite les k premiers termes de la liste L_{ij} et les renvoie à la cellule c_{ij} correspondante de la TM. Ainsi le résultat de l'agrégation dans la cellule c_{ij} est : $R_{ij} = \langle t_1,\ldots, t_k \rangle$ avec $R_{ij} \subseteq L_{ij}$.

Exemple. L'exemple est extrait de notre publication [Pujolle, Ravat, Teste, Tournier, 2008] dans la *Revue des Sciences et Technologies de l'Information*.

Un décideur analyse les principaux termes d'articles scientifiques selon l'auteur et l'année de publication. La fonction TOP_KW$_2$ utilisée retourne les deux termes les plus représentatifs (k = 2). Elle est appliquée sur quatre groupes de documents : un pour chaque cellule de la table multidimensionnelle, chacune correspondant à l'un des couples : (Au1, 2005), (Au1, 2006), (Au2, 2005) et (Au2, 2006).

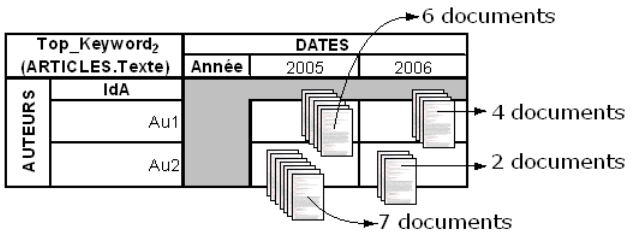


Figure 28 : Analyse de contenu de documents à agréger.

Ainsi dans la figure suivante, la cellule correspondant à (Au1, 2006) affiche le résultat de l'application de la fonction d'agrégation sur les documents (d1, d2, d3 et d4). Le *tf* est calculé pour chaque terme pour chaque document et l'*idf* l'est pour chaque document. Le poids associé aux termes est la somme des *tf.idf* du terme des documents de la cellule. Les termes sont ordonnés en fonction du poids global, puis les 2 termes avec les poids les plus élevés sont retenus. Le procédé de calcul est détaillé dans la figure suivante. Seuls les premiers termes retenus sont représentés : *OLAP*, *Query*, *SQL* et *Definition*.

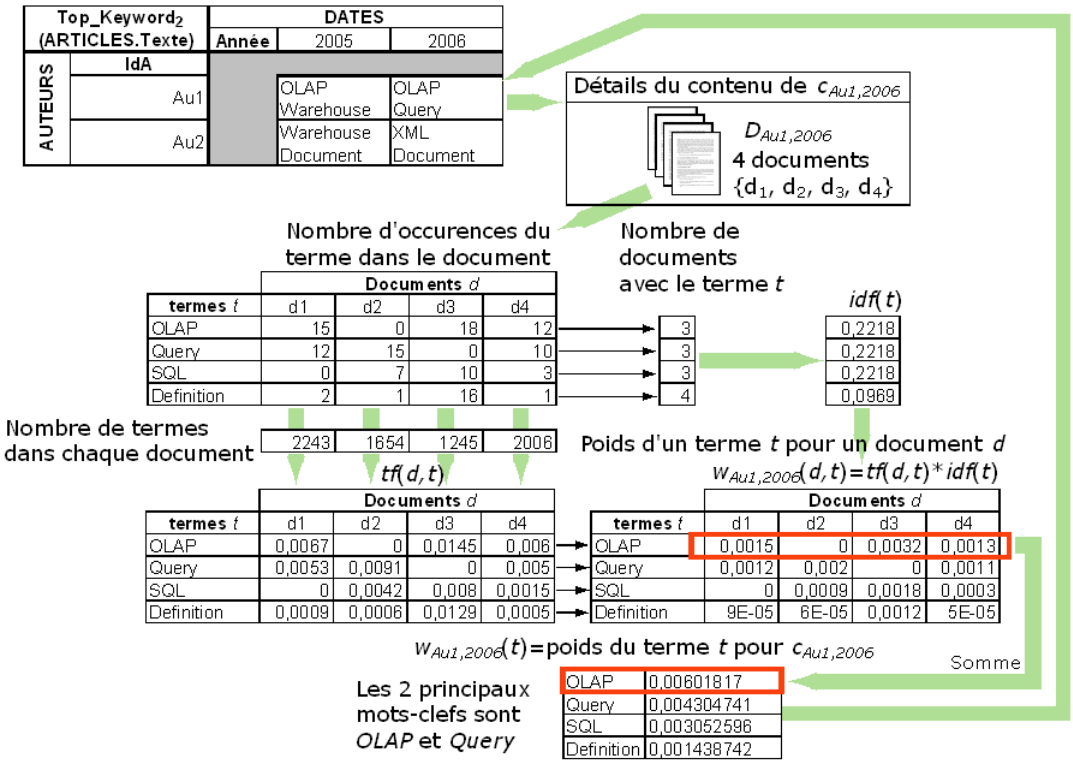


Figure 29 : Processus d'agrégation de la fonction TOP_KW2.

6.2 *Fonctions Avg_Kw*

La fonction d'agrégation *AVG_KW* synthétise par un calcul de pseudo-moyenne les termes d'une mesure élaborée en un ensemble restreint de termes plus généraux.

Il est nécessaire de disposer d'une « loi » autorisant l'agrégation tout comme la moyenne permet de le faire dans un environnement numérique. Pour établir cette règle, nous utilisons une ontologie légère dotée de liaisons « est-un » informelles [Lassila, *et al.*, 2001]. Ce type d'ontologie correspond à une hiérarchie de concepts d'un domaine où chaque nœud représente un concept (un terme) et chaque arc entre les nœuds modélise une relation « est-un ». Étant donnée une ontologie *O*, le *domaine* de *O*, noté *dom*(*O*), est l'ensemble des termes

non redondants (sans homographes) de *O*. La *profondeur* (« depth ») d'une ontologie est le nombre maximum d'arcs entre le nœud racine et une feuille. Deux opérations sont définies pour permettre la manipulation des nœuds de l'ontologie lors du processus d'agrégation :

- Le plus petit ancêtre commun *lca* (« Least Common Ancestor ») est une fonction qui retourne le nœud représentant le plus petit ancêtre commun (n_{LCA}) au sein de *O* entre n_1 et n_2.

$$\begin{array}{rccc} lca: & dom(O)^2 & \longrightarrow & dom(O) \\ & (n_1, n_2) & \mapsto & n_{LCA} \end{array}$$

- La distance *d* entre deux nœuds n_1 et n_2 est une fonction qui retourne le nombre d'arcs entre le nœud représentant le plus petit ancêtre commun (n_{LCA}) et le nœud le plus bas dans la hiérarchie entre n_1 et n_2. Il s'agit d'une mesure de similarité, à savoir une distance structurelle calculant un nombre d'arcs.

$$\begin{array}{rccc} d: & dom(O)^2 & \longrightarrow & N \\ & (n_1, n_2) & \mapsto & \max(d(n_1, lca(n_1, n_2)), d(n_2, lca(n_1, n_2))) \end{array}$$

La fonction *AVG_KW* se base sur un processus d'agrégation utilisant une ontologie de domaine. Pour chaque paire de termes, la fonction calcule le plus petit ancêtre commun. Lors de l'agrégation de termes très éloignés dans l'ontologie, il y a une forte probabilité de retourner le nœud racine de l'ontologie ; en outre, plus les termes sont éloignés, plus l'agrégation se traduit par une perte de sens. Une limite est donc imposée dans le processus d'agrégation : la fonction emploie une distance maximale autorisée lors de l'agrégation de termes : D_{MAX}. Des heuristiques suggèrent une distance comprise entre 3 et 5 ; une ontologie généraliste telle que WordNet[8], D_{MAX} est généralement fixée à 3.

Dans la TM qui visualise les données analysées, à chaque intersection entre les lignes et les colonnes, la cellule contient un ensemble de termes pour la mesure textuelle élaborée. La fonction *AVG_KW* est appliquée sur le contenu de chaque cellule pour produire un nouvel ensemble composé de termes agrégés et éventuellement de termes de la cellule initiale si le processus d'agrégation échoue suite à une distance excessive.

Définition. La fonction *AVG_KW* est définie par

$$\begin{array}{rccc} AVG\_KW: & X^n & \longrightarrow & X^m \\ & \langle x_1, ..., x_n \rangle & \mapsto & \langle x'_1, ..., x'_m \rangle \end{array} \quad \text{avec } m \leq n$$

– X^n est un ensemble ordonné de termes associés à une distance tel que $X = dom(kw) \times \mathbb{N}$. Les termes correspondant aux nœuds les plus éloignés de la racine sont en premier $\forall x_i \in X,\ x_j \in X \mid i<j,\ d(x_i, x_{ROOT}) \leq d(x_j, x_{ROOT})$ (avec $x_i = (kw_i, d_i)$ où $kw_i \in dom(O)$ et $d_i \leq D_{MAX}$).

– X^m est un ensemble ordonné de termes agrégés associés à une distance qui représente la distance avec le terme agrégé le plus lointain. Ces termes sont agrégés par une fonction conditionnelle basée sur le *lca* :

$$(x_i, x_j) \mapsto \begin{cases} \text{si } l(x_i, x_j) \leq D_{MAX},\ x_{LCA} = (kw_{LCA}, l(x_i, x_j)) \\ \text{sinon } (x_i, x_j) \end{cases}$$

où

- $l(x_i, x_j) = d(kw_i, kw_j) + d_i + d_j$ et
- $kw_{LCA} = LCA(kw_i, kw_j)$.

[8] WordNet : Ontologie lexicale anglaise disponible sur http://wordnet.princeton.edu/

Si x_i et x_j sont agrégés en x_{LCA} alors x_i et x_j sont retirés de l'ensemble d'entrée X et x_{LCA} est ajouté à X. Le processus d'agrégation est itéré sur X jusqu'à ce qu'aucune nouvelle agrégation n'ait pu avoir lieu : $\forall (x_i, x_j) \in X^2, \nexists x_{LCA} \mid l(x_i,x_j) \leq D_{MAX}$. Pour un x_k donné de X^m, si $d_k=0$, alors le terme correspondant kw_k n'a pas été agrégé durant le processus ($\exists x_i \in X^n \mid x_i=x_k$).

Exemple. L'exemple est extrait de notre publication [Pujolle, Ravat, Teste, Tournier, 2008] dans la *Revue des Sciences et Technologies de l'Information*.

La figure suivante présente un jeu de données. Six documents ont été sélectionnés quatre ont été écrits par l'auteur Au1 et deux par Au2, à deux périodes différentes (semestre et année). Pour simplifier, seulement deux mots-clefs ont été extraits du contenu de chaque document.

Documents	Keywords	Date	Author
Doc_1	Conceptual Model Architecture	S1 2006	Au1
Doc_2	Data Warehouse Conceptual Model	S1 2006	Au1
Doc_3	Logical Fact Table	S1 2006	Au1
Doc_4	Document Warehouse Algebra	S2 2006	Au1
Doc_5	Architecture Conceptual Model	S1 2006	Au2
Doc_6	OLAP Document Warehouse	S2 2006	Au2

Figure 30 : Exemple de données pour l'agrégation AVG_KW.

Un décideur analyse les publications des auteurs Au1 et Au2 durant l'année 2006. La vision des publications par semestre dans une table multidimensionnelle génère quatre cellules : premier auteur (Au1), semestre 1 (S1) ; premier auteur (Au1), semestre 2 (S2) ; deuxième auteur (Au2) semestre 1 (S1) et deuxième auteur (Au2), semestre 2 (S2). La Figure 31 présente dans la table supérieure la répartition des six documents du jeu de données par auteur et par semestre de publication dans les cellules de la table multidimensionnelle :

- $c_{Au1,S1}$={Doc_1, Doc_2, Doc_3} et $c_{Au2,S1}$={Doc_5} pour le semestre S1 ;
- $c_{Au1,S2}$={Doc_4}et $c_{Au2,S2}$={Doc_6} pour le semestre S2.

Ces deux ensembles de documents correspondent à des mots-clefs « agrégeables ». La table inférieure de la figure suivante montre l'ensemble des mots à agréger par cellule. Les mots-clefs à agréger sont les suivants (ces mots-clefs sont associés à une distance de 0 par défaut) :

- x_1=(kw_1='Document Warehouse', d_1=0) ;
- x_2=(kw_2='Algebra', d_2=0) ;
- x_3=(kw_3='Data Warehouse', d_3=0) ;
- x_4=(kw_4='Conceptual Model', d_4=0) ;
- x_5=(kw_5='Logical', d_5=0) ;
- x_6=(kw_6='Fact Table', d_6=0) ;
- x_7=(kw_7='Architecture', d_7=0) ;
- x_8=(kw_4='Conceptual Model', d_8=0).

AVG_KW (Mots_Clefs)		DATES	
	Annee	2006	
	Semestre	S1	S2
AUTEURS	Auteur		
	Au_1	Doc_1, Doc_2, Doc_3	Doc_4
	Au_2	Doc_5	Doc_6

AVG_KW (Mots_Clefs)		DATES	
	Annee	2006	
	Semestre	S1	S2
AUTEURS	Auteur		
	Au_1	Conceptual model, Architecture, Data warehouse, Logical, Fact table	Document warehouse, Algebra
	Au_2	Architecture, Conceptual model	OLAP, Document warehouse

Figure 31 : Répartition dans les cellules des documents et des mots-clefs.

Seules les données du premier auteur seront détaillées (cellules $c_{Au1,S1}$ et $c_{Au1,S2}$).

La partie gauche de la figure suivante montre les positions des différents mots-clefs des publications de l'auteur Au1 au sein de l'ontologie qui nous sert d'exemple. Ces mots-clefs sont entourés d'un rectangle. Dans la partie droite de la figure, les agrégations possibles des mots-clefs extraits des documents sont indiquées. Les flèches représentent les agrégations avec les distances (le nombre d'arcs entre les nœuds agrégés). Pour cet exemple, la distance maximale d'agrégation a été fixée à 3 arcs : $D_{MAX} = 3$.

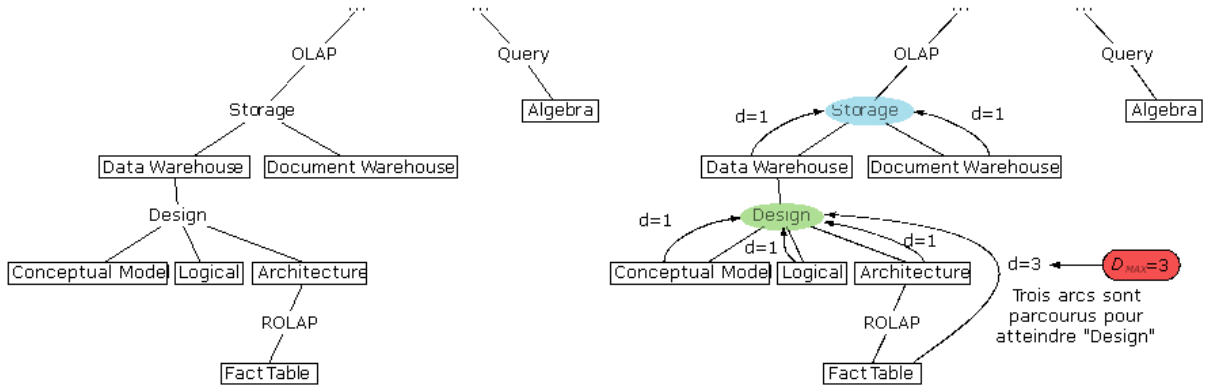


Figure 32 : Positionnement des mots-clefs dans l'ontologie de domaine.

Le décideur affiche les résultats de l'analyse par semestre. Le processus d'agrégation s'opère pour chaque cellule de la table de manière indépendante en commençant par agréger les mots-clefs les plus éloignés de la racine :

- AVG_KW(x_3, x_4, x_5, x_6, x_7, x_8) = (x_3, x_9) avec x_9 = (kw_9='Design', d_9=3)
- AVG_KW(x_1, x_2) = (x_1, x_2) car $d(x_1, x_2) > D_{MAX}$

Les mots-clefs des trois publications de Au1, du semestre S1 sont agrégés : quatre d'entre eux (Fact table, Conceptual model, Logical et Architecture) sont agrégés en un seul (Design) comme l'indique la figure suivante. Le résultat kw_9 est un mot-clef agrégé ayant atteint D_{MAX}, il ne peut plus être agrégé avec un mot-clef de niveau supérieur dans l'ontologie. Le cinquième mot-clef, par contre, est trop éloigné des autres et ne peut être agrégé (Data Warehouse) : la distance d qui serait obtenue par le processus d'agrégation serait supérieure au maximum D_{MAX} autorisé. Concernant l'autre Au1, les deux mots-clefs qui en ont été extraits sont trop éloignés l'un de l'autre et ne peuvent être agrégés.

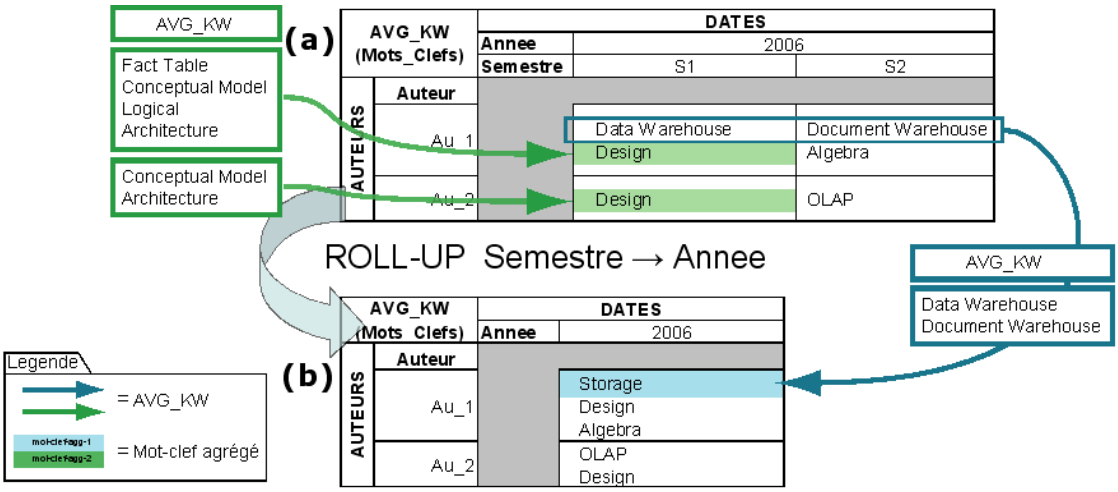


Figure 33 : Analyse des mots-clefs par semestres, puis par années.

Par la suite, le décideur effectue un forage vers le haut afin d'obtenir une vision d'ensemble. Le forage se fait selon la dimension DATES et le niveau de détail des données passe de Semestre à Année. Ainsi les publications analysées sont regroupées en deux groupes (au lieu de quatre) correspondant aux publications de l'auteur Au1 durant l'année 2006 et celles de l'auteur Au2 durant la même année. La fonction AVG_KW tente alors d'agréger l'ensemble des mots-clefs. Les mots-clefs x_4 et x_8 seront fusionnés car ils sont identiques et contenus dans la même cellule. Lors du forage, le processus est réitéré :

$AVG\_KW(x_1, x_2, x_3, x_9) = (x_2, x_9, x_{10})$ avec $x_{10} = (kw_{10}$='Storage', $d_{10}=1)$

Ainsi après agrégation l'ensemble des six mots-clefs du départ se résume à trois mots-clefs plus généraux : 'Algebra', 'Design' et 'Storage'.

Nous développons au sein du laboratoire le prototype XML-GOLAP qui sert à valider nos propositions et de plate-forme de tests pour évaluer le coût des agrégations textuelles. Des expérimentations sur la fonction AVG_KW sont décrites au chapitre 6 du mémoire.

7 Bilan

Dans ce chapitre nous avons détaillé nos contributions dans l'intégration des documents au sein d'une BDM. Mes travaux se sont particulièrement intéressés aux documents XML orientés-documents qui restaient inexploités dans les analyses OLAP. L'objectif a été d'étendre les systèmes OLAP dans leur capacité à supporter non seulement des analyses quantitatives sur le contenu numérique des documents mais également des analyses plus qualitatives sur le contenu textuel des documents.

7.1 Résultats de nos travaux

Pour permettre l'application d'analyses OLAP sur le contenu de documents, mes travaux ont suivi dans un premier temps une approche par extension des modèles multidimensionnels existants (« constellation étendue ») [Ravat, Teste, Tournier, 2008e, 2007e] [Ravat, Teste, Tournier, Zurfluh, 2007c] afin de conserver l'ensemble des manipulations multidimensionnelles (algèbre OLAP) classiques applicables. Une constellation étendue repose essentiellement sur l'intégration *de mesures textuelles*.

Dans un second temps, nous avons généralisé nos travaux dans le développement d'une approche originale qui revisite les principes de la modélisation multidimensionnelle : le modèle en Galaxie [Tournier, 2007]. Il s'agit d'une contribution importante de nos travaux

dans cette thématique [Ravat, Teste, Tournier, Zurfluh, 2008a, 2008f, 2007a, 2007f]. En effet, le modèle en galaxie repose sur deux idées novatrices pour les systèmes OLAP :

- *Unicité du mécanisme de description des données* analysées (la dimension),
- *Navigabilité dans les données* au travers de liens.

Nous avons également étudié la manipulation des documents décrits de manière multidimensionnelle soit dans une constellation étendue, soit dans une galaxie. Dans le cadre d'une galaxie, l'algèbre multidimensionnelle a nécessité le développement d'opérateurs supplémentaires, principalement la définition d'un constructeur (FOCUS) permettant de projeter les données à analyser dans une structure de visualisation généralisée. Ces travaux ont montré que la galaxie et l'ensemble étendu des opérateurs de manipulation OLAP sont une généralisation de la constellation textuelle. En outre, la galaxie offre de nouvelles perspectives d'analyses au travers de l'extension des opérateurs de forage vers le bas et de rotation. En outre nous avons montré comment l'exploitation des liens au sein de la galaxie peut servir à mener des analyses qualitatives complexes sur les documents.

La contribution de ces travaux qui me semble la plus notable concerne le développement d'une nouvelle approche pour l'*agrégation de données textuelles* [Pujolle, Ravat, Teste, Tournier, 2008]. Nous avons développé deux fonctions d'agrégation adaptées aux mesures textuelles brutes ou élaborées :

- *TOP_KW*$_k$ [Ravat, Teste, Tournier, Zurfluh, 2008c, 2008d] exploite la fonction de pondération *tf.idf* issues de la recherche d'information,
- *AVG_KW* [Ravat, Teste, Tournier, 2007d] repose sur une ontologie légère de domaine.

Ces travaux ont fait l'objet de validations au sein du prototype XML-GOLAP [Abdelhedi, 2009] que nous développons au sein du laboratoire. Nous présentons les principaux aspects du prototype XML-GOLAP dans le chapitre 6 de ce mémoire.

7.2 Encadrements et diffusion scientifique

Ces travaux ont été menés dans le cadre de la thèse de Ronan Tournier [Tournier, 2007] que j'ai co-encadrée et de stages de recherche (M2R) dont j'ai assuré l'encadrement :

(1) F. Abdelhedi, « Etude et développement d'un mécanisme d'agrégation pour l'analyse OLAP de documents XML » en 2009,
(2) C. Koussa, « Bases de données multidimensionnelles : construction d'un magasin de données à partir des sources XML » en 2007.

Le tableau suivant dresse un panorama des thèmes étudiés, des étudiants encadrés et des publications réalisées dans ce second axe de mes recherches.

Tableau 10 : Etudiants encadrés et publications de l'axe 2.

Thèmes	Thèses	Master/D.E.A.	Publications
Modélisation par Extensions et Galaxie	R. Tounier	C. Koussa	▪ RI AoIS [Ravat, Teste, Tournier, Zurfluh, 2008a] ▪ CI ER'07 [Ravat, Teste, Tournier, Zurfluh, 2007a] SEKE'07 [Ravat, Teste, Tournier, Zurfluh, 2007c] ▪ OI ADWM [Ravat, Teste, Tournier, 2008e] ▪ RN Doc. Num. [Ravat, Teste, Tournier, 2007e] ▪ CN MADSI'08 [Ravat, Teste, Tournier, Zurfluh, 2008f] EDA'07 [Ravat, Teste, Tournier, Zurfluh, 2007f]
Manipulation OLAP et Agrégations Textuelles	R. Tounier	F. Abdelhedi	▪ CI DAWAK'08 [Ravat, Teste, Tournier, Zurfluh, 2008c] ICEIS'07 [Ravat, Teste, Tournier, 2007d] ▪ RN RSTI/ISI [Pujolle, Ravat, Teste, Tournier, 2008] ▪ CN EDA'08 [Ravat, Teste, Tournier, Zurfluh, 2008d]

7.3 Perspectives

Les travaux que nous avons réalisés sont cantonnés à des documents fortement structurés. Les perspectives directes de cette approche sont triples : l'intégration de documents semi-structurés [Sèdes, 1998], de documents multi-facettes [Djemal, *et al.*, 2008] et enfin, la prise en compte de leur évolution au cours du temps. Un second point concerne l'agrégation de textes. Nous avons proposé deux fonctions d'agrégation adaptées aux données textuelles. D'autres fonctions peuvent être imaginées. J'ai comme ambition de proposer un environnement générique permettant l'intégration de nouvelles fonctions : un concepteur pourrait ajouter des fonctions, adaptées à des besoins spécifiques d'analyse.

A plus long terme, en profitant de l'expérience de l'équipe SIG dans le domaine de la recherche d'information [Boughanem, 2000] [Mothe, 2000] [Soulé-Dupuy, 2001], j'envisage d'appliquer un environnement d'analyse OLAP à des données indexées issues de pages Web. Cette perspective vise à répondre à l'une des difficultés des moteurs de recherche sur le Web, qui demandent à l'utilisateur de savoir ce qu'il cherche, rendant l'accès à l'information difficile lorsque l'on ignore ce que l'on cherche. L'idée est donc d'inverser ce principe en autorisant un utilisateur à retourner la question au moteur de recherche. Ainsi l'utilisateur pourra, via une vision synthétique, parcourir le Web, et employer les opérations OLAP (comme le forage) pour analyser les thèmes et les concepts plus en détails.

Enfin, mes travaux sur l'intégration des documents dans les systèmes OLAP nous ont permis d'acquérir une expérience dans les problématiques liées à l'ingénierie guidée par les modèles à l'image de la modélisation en galaxie. Notre participation au groupe de travail MADSI est l'une des illustrations de cette préoccupation. Je compte appliquer mes recherches dans le domaine médical, s'inscrivant ainsi dans les axes définis par le

laboratoire, et plus globalement, dans un des pôles de compétitivité de la région. J'envisage par exemple d'utiliser nos solutions pour explorer le dossier patient. Plus complexe encore, je souhaite adapter les systèmes OLAP pour l'exploration et la navigation dans des bases de données biomédicales [Darmont, *et al*. 2008] comme les bases de gènes.

Chapitre 5 - Personnalisation des systèmes OLAP

Ce chapitre présente mes travaux de recherche sur la personnalisation d'une BDM afin que le système OLAP réponde de façon plus adaptée aux besoins de l'usager. L'approche développée propose d'assister le décideur par des recommandations lors de la manipulation multidimensionnelle.

1 Problématique

Les systèmes OLAP facilitent l'analyse en offrant un espace de représentation multidimensionnelle des données que les décideurs explorent interactivement par une succession d'opérations OLAP [Choong, *et al*., 2003] [Dittrich, *et al*., 2005]. Cette approche a connu un développement important grâce à sa capacité à permettre un accès direct et dynamique aux données analysées. Cependant ces systèmes OLAP sont élaborés pour un groupe de décideurs ou un sujet d'analyse (« subject-oriented » [Inmon, 1994]) pour lesquels sont présumés des besoins parfaitement identiques. Cette simplification rend les systèmes OLAP parfois mal adaptés à un usage particulier. Le décideur se trouve alors confronté à un espace multidimensionnel avec lequel il doit opérer un nombre important de manipulations afin d'obtenir un résultat le plus proche possible de son besoin. Au-delà du résultat parfois imparfait, le décideur se trouve confronté à des bases de données multidimensionnelles très vastes car les concepteurs augmentent souvent l'espace multidimensionnel par des ajouts au fur et à mesure des demandes des usagers, complexifiant alors la base de données multidimensionnelles.

La problématique traitée dans ce chapitre consiste à personnaliser les systèmes OLAP en fonction de besoins analytiques individuels. Les mécanismes de personnalisation dans les systèmes OLAP, visant à mieux prendre en compte l'usager, ne sont que très peu étudiés [Rizzi, 2007] [Golfarelli, *et al*., 2009]. Nous proposons d'utiliser la personnalisation afin de rendre possible au décideur des recommandations de manipulations OLAP comme l'illustre la figure suivante [Jerbi, Ravat, Teste, Zurfluh, 2009c]. Il s'agit pour le système OLAP d'être capable :

(1) d'assister l'usager en lui proposant d'enrichir la requête en cours d'élaboration,
(2) de simplifier la navigation en proposant à l'usager des requêtes anticipées,
(3) de suggérer à l'usager des requêtes alternatives à sa requête pour découvrir des données de l'espace multidimensionnel jugées pertinentes.

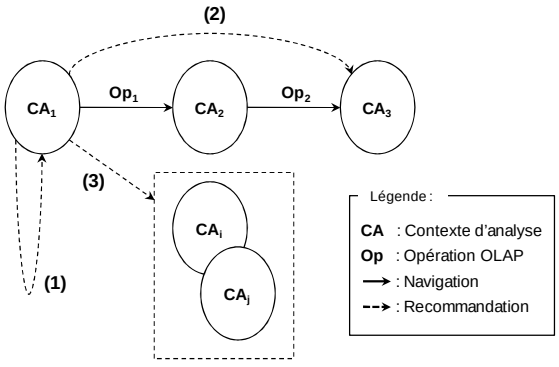


Figure 34 : Scenarii de recommandation.

Les prises de décisions reposent non seulement sur les données brutes manipulées mais également sur les réflexions, les commentaires des décideurs voire la confrontation de différentes interprétations. Les systèmes OLAP sont réduits à la mise à disposition efficace des données décisionnelles, mais les décideurs doivent analyser ces données sur la base immatérielle de leur expertise.

Se posent alors plusieurs questions concernant à la fois la modélisation et les manipulations multidimensionnelles :

- Comment modéliser les caractéristiques individuelles des décideurs dans le système OLAP ? Comment intégrer ces caractéristiques (ou préférences) dans une base de données en constellation ?
- Comment conserver le patrimoine immatériel que représente son expertise ?
- Comment exploiter les préférences ? Quels mécanismes doivent être développés lors des manipulations OLAP pour assister l'usager individuellement ?

2 Approches existantes

La **personnalisation** est définie comme un mécanisme « *...providing an overall customized, individualized user experience by taking into account the needs, preferences and characteristics of a user or group of users* » [Ioannidis , *et al*., 2005]. Généralement, la personnalisation d'un système consiste à définir, puis à exploiter un **profil utilisateur** [Korfhage, 1997] pouvant s'apparenter à une modélisation de l'usager. Un profil regroupe un ensemble de caractéristiques servant à configurer ou à adapter le système à l'utilisateur, afin de lui fournir des réponses plus adaptées. Aucun consensus n'existe sur la définition de profil ; on peut relever cependant la proposition d'un profil générique multidimensionnel visant à couvrir une majorité de contextes [Bouzeghoub, *et al*., 2005].

Nous proposons de caractériser un profil selon l'implication de l'utilisateur et les fonctions systèmes [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a]. Dans le cadre d'une implication explicite, l'utilisateur doit effectuer des interactions avec le système tandis que lors d'une implication implicite, le système s'adapte automatiquement à l'utilisateur. Les fonctions systèmes liées au profil consistent à définir le profil, puis à exploiter ce dernier pour une meilleure prise en compte de l'utilisateur. A partir de ces caractéristiques, la figure suivante décrit les principes mis en jeu lors de la personnalisation.

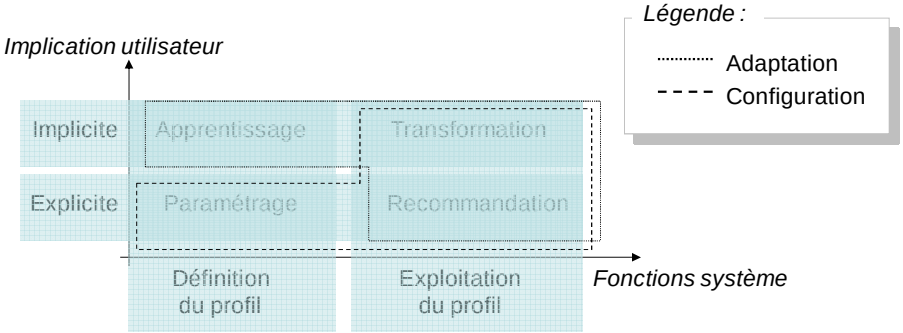


Figure 35 : Principes de la personnalisation.

La définition d'un profil réalisée de façon explicite correspond à la configuration (paramétrage) d'un système tandis que la définition implicite s'apparente à l'apprentissage. La **configuration** consiste pour l'utilisateur à paramétrer explicitement son profil tandis que l'**adaptation** consiste pour le système à définir implicitement le profil de l'utilisateur.

L'exploitation du profil peut soit nécessiter l'intervention explicite de l'utilisateur qui transforme le système par des choix de recommandations du système, soit induire une transformation automatique du système.

Alors que la personnalisation a fait l'objet de très nombreux travaux en recherche d'information et bases de données [Ioannidis, *et al*., 2005], très peu de propositions visent à personnaliser les systèmes OLAP [Rizzi, 2007] [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a]. Le tableau suivant dresse chronologiquement un panorama des quelques travaux existants.

Tableau 11 : Synthèse des travaux sur la personnalisation dans les systèmes OLAP.

		Définition profil	**Exploitation profil**
1	[Sapia, 2000]	Explicite	Transformation
2	[Espil, *et al*., 2001]	Explicite	Transformation
3	[Thalhammer, *et al*., 2001]	Explicite	Transformation
4	[Bellatreche, *et al*., 2005]	Explicite	Transformation
5	[Favre, *et al*., 2007]	Explicite	Transformation
6	[Giacometti, *et al*., 2008, 2009]	Implicite	Recommandation

On relève que la quasi-totalité des travaux produits jusqu'à maintenant visent à transformer le système OLAP en fonction de préférences utilisateur explicitement collectées. Seuls les travaux dans [Giacometti, *et al*., 2008, 2009] consistent à recommander des requêtes sur la base de calculs opérés sur l'historique des navigations réalisées par un groupe d'utilisateurs. Ces travaux traitent uniquement de requêtes anticipées et recommandent des requêtes complètes déjà jouées par le groupe social d'usagers, introduisant des approximations pouvant être importantes.

Dès 2006, j'ai orienté mes recherches sur l'étude des mécanismes de personnalisation dans les systèmes OLAP pour assister le décideur lors des analyses. Mes travaux ont suivi deux approches [Agrawal, *et al*., 2000] :

- une approche quantitative consistant à exprimer les préférences en fonction de scores numériques sur les données [Ravat, Teste, Zurfluh, 2007g] [Ravat, Teste, 2008g],
- une approche qualitative définissant les préférences directement à l'aide de relations binaires entre les données [Jerbi, Ravat, Teste, Zurfluh, 2008, 2009a, 2009b, 2009c].

3 Approche quantitative

Nos propositions pour la personnalisation quantitative d'une base de données multidimensionnelles reposent sur un mécanisme dynamique d'affectation de poids sur les propriétés (mesures, paramètres et attributs faibles) d'une constellation [Ravat, Teste, Zurfluh, 2007g] [Ravat, Teste, 2008g]. Ce mécanisme est complété par la possibilité offerte aux décideurs d'annoter les composants d'une constellation [Cabanac, Chevalier, Ravat, Teste, 2006a, 2006b, 2007] : une annotation permet de conserver dans le système OLAP les commentaires, remarques, questions et/ou réponses formulés par les décideurs sur les données décisionnelles. Ce cadre permet aussi bien un usage personnel (lecture active des données décisionnelles) qu'un usage collectif facilitant les prises de décisions (fil de discussion liant des annotations).

3.1 Constellation personnalisée

Nous avons étendu le concept de constellation afin de permettre sa personnalisation et d'associer des annotations aux composants de celle-ci.

Définition. Une constellation personnalisée *CP* est définie par ($\mathbb{F}^{CP}$; $\mathbb{D}^{CP}$; $Star^{CP}$; $Rule^{CP}$; $Annotate^{CP}$) où

- $\mathbb{F}^{CP} = \{F_1,\ldots, F_n\}$ est l'ensemble des faits,
- $\mathbb{D}^{CP} = \{D_1,\ldots, D_m\}$ est l'ensemble des dimensions,
- $Star^{CP} : \mathbb{F} \rightarrow 2^{\mathbb{D}}$ associe chaque fait à un sous-ensemble des dimensions en fonction desquelles il est analysable,
- $Rule^{CP} = \{R^{CP}_1, R^{CP}_2,\ldots\}$ est l'ensemble de règles actives de personnalisation,
- $Annotate^{CP} = \{AD^{CP}_1, AD^{CP}_2,\ldots\}$ est l'ensemble des annotations décisionnelles.

Cette définition décrit de manière classique la constellation au travers des concepts de fait, de dimension et de hiérarchie (*cf.* chapitre 3). L'extension consiste en une personnalisation par règles $Rule^{CP}$ et par annotations $Annotate^{CP}$. Les règles permettent d'exprimer les préférences de l'usager tandis que les annotations servent à conserver l'expertise de l'usager. Nous détaillons dans la suite nos propositions.

3.2 Règles de personnalisation

Nous définissons la personnalisation de manière quantitative par des règles actives fixant les propriétés préférées d'un usager qui seront utilisées de manière prioritaire par le système lors de la manipulation de la constellation. Notre approche consiste à associer un poids à chaque propriété p_i de la constellation. Ce poids, noté w_i, modélise l'importance que l'utilisateur souhaite associer à p_i. Afin de faciliter son exploitation, chaque poids est normalisé ($0 \leq w_i \leq 1$).

Nous proposons de personnaliser la constellation en intégrant le contexte d'utilisation des propriétés au travers d'un mécanisme dynamique de type ECA (Evènement – Condition – Action) [Widom, 1992] [Tchounikine, 1994].

Définition. Une règle R^{CP}_x est définie par (N^{Rx}; S^{Rx}; E^{Rx}; C^{Rx}; A^{Rx}) où

- N^{Rx} est le nom identifiant la règle,
- S^{Rx} désigne la portée de la règle. Les règles peuvent être associées soit à un fait (NF_i), soit à une dimension (ND_i), soit à une hiérarchie (NH_j).
- E^{Rx} est le contexte de la manipulation déclenchant la règle. Nous définissons ce contexte par rapport aux opérations de manipulation qui sont appliquées sur les composants de la constellation : '*display*', '*rotate*', '*drilldown*', '*rollup*',…
- C^{Rx} est une condition définissant si la règle est déclenchée au travers d'une fonction *current*(*E*) : *boolean* qui détermine si un élément *E* de la constellation est en cours de manipulation ; $E \in \{ND_i, ND_i.NH_j, ND_i[.NH_j].p_k, NF_i, NF_i.f_j, NF_i[.f_j].m_k\}$.
- A^{Rx} est la séquence d'actions déclenchée. Les actions s'appliquent sur les composants (faits, mesures, dimensions,…) de la constellation. Nous définissons la procédure *setWeight*(*E*, w_i) permettant d'associer contextuellement un poids à chaque attribut (mesure, paramètre, attribut faible) : si $E = ND_i$, le poids w_i est affecté à toutes les propriétés de la dimension, si $E = ND_i.NH_j$, w_i est affecté à tous les attributs de la hiérarchie…

La combinaison de E^{Rx} et C^{Rx} permet de distinguer deux types de personnalisation contextuelle : soit le **contexte de manipulation**, soit le **contexte d'utilisation**. Le contexte de manipulation permet de spécifier l'opération suivant laquelle les priorités sont fixées tandis que le contexte d'utilisation spécifie l'état courant de la constellation suivant lequel les priorités doivent être fixées. Ces deux contextes peuvent être utilisés simultanément au sein d'une même règle.

Pour faciliter l'expression des règles de personnalisation, nous avons défini un langage comportant un ordre textuel de définition des règles.

Définition. La commande de définition d'une règle $R^{CP}{}_x$ est la suivante :

CREATE RULE N^{Rx}
ON S^{Rx}
WHEN E^{Rx} [**IF** C^{Rx}] **THEN BEGIN** A^{Rx} **END**;

Exemple. L'exemple est extrait de notre publication [Ravat, Teste, 2008g] dans *Annals of Information Systems*.

Considérons qu'un décideur utilise très fréquemment les paramètres CITY et COUNTRY d'une dimension CUSTOMERS, moins fréquemment les paramètres FIRSTNAME et LASTNAME, plus rarement encore les paramètres ZONE et IDC (identifiant du client), et jamais le paramètre CONTINENT. Ses préférences peuvent être modélisées par la règle suivante :

$R_1 = (N^{R1}; S^{R1}; E^{R1}; C^{R1}; A^{R1})$ où

- N^{R1} = CustomerRule
- S^{R1} = CUSTOMERS
- E^{R1} = Displayed $\vee$ Rotated
- C^{R1} = Current(ACCOUNTS)
- A^{R1} = {IdC←0.5; Firstname←0.8; Lastname←0.8; City←1; Country←1; Continent←0; Zone←0.5}.

Cette règle est associée à la dimension CUSTOMERS et se déclenche lorsqu'une opération de construction de table multidimensionnelle ou une rotation est effectuée. Il s'agit d'une règle contextuelle puisqu'elle est déclenchée si et seulement si le fait ACCOUNTS est utilisé.

La définition textuelle de cette règle correspond à l'ordre suivant :

```
CREATE RULE CustomerRule
ON Customers
WHEN Display OR Rotate
IF current('Accounts')
THEN
    BEGIN
      setWeight('IdC',0.5);
      setWeight('Firstname',0.8);
      setWeight('Lastname',0.8);
      setWeight('City',1);
      setWeight('Country',1);
      setWeight('Continent',0);
      setWeight('Zone',0.5);
    END;
```

3.3 Annotations

Les prises de décisions reposent non seulement sur les données brutes mais également sur les réflexions, les commentaires des décideurs voire la confrontation de différentes interprétations. Notre proposition consiste à modéliser au travers d'annotations le capital immatériel mentalement associé aux données par les décideurs. Les annotations visent à conserver les commentaires et discussions formulés lors des analyses et du processus de prise de décisions. Ce cadre informatique permet d'exploiter et de partager les données multidimensionnelles tout en supportant des fonctionnalités d'annotation permettant d'enrichir interactivement les composants d'une constellation. Les décideurs sont alors des usagers actifs créant leur propre système de repérage au travers de signes graphiques (surlignage, cerclage…), de commentaires et de réponses (affirmation, infirmation) pouvant impliquer des fils de discussions (communication asynchrone).

L'expertise que véhiculent ces annotations est utilisée à des fins personnelles ou collectives et elles peuvent contribuer à améliorer les analyses futures. Les annotations contiennent des informations de plusieurs natures :

- les informations subjectives regroupant le contenu et le type de l'annotation,
- les informations objectives telles que l'identifiant, la date de création, le créateur, des références à une annotation mère dans le cas d'un fil de discussion, un point d'ancrage.

Définition. Une annotation $AD^{CP}{}_x$ est définie par (IS^{ADx}; IO^{ADx}) où

- IS^{ADx} est un ensemble d'informations subjectives regroupant :
  - le contenu textuel saisi par le décideur qui annote,
  - le type de l'annotation caractérisant son contenu (commentaire, question, réponse à une annotation, conclusion).
- IO^{ADx} est un ensemble d'informations objectives comportant :
  - son identifiant,
  - sa date de création permettant de caractériser sa position dans le fil de discussions ordonnées chronologiquement,
  - l'identifiant de son créateur (décideur),
  - référence à une annotation père,
  - son point d'ancrage spécifiant la localisation précise de l'annotation.

Le point d'ancrage peut être de deux natures :

- un point d'ancrage global localisé sur un concept dans une constellation (l'annotation sera présente dans toutes les tables multidimensionnelles intégrant le concept annoté globalement), ou bien,
- un point d'ancrage local localisé sur un élément dans une table multidimensionnelle (l'annotation n'est présente que dans le contexte local de la table multidimensionnelle).

Définition. Un point d'ancrage α est défini par (S; D_1; D_2) où

- $S = \{C \mid TM\}[.NF_i[.f(m)[=val]?]?]?$ désigne un ancrage relatif au fait F_i,
- $D_1 = \lambda \mid ND_{i1}\ [.NH_{j1}[/p_{k1}[=pos_1]?]?]^*]?$ désigne un ancrage relatif à une dimension,
- $D_2 = \lambda \mid ND_{i2}\ [.NH_{j2}[/p_{k2}\ [=pos_2]?]?]^*]?$ désigne un ancrage relatif à une dimension.

Notons que *C* désigne une constellation, *TM* désigne une table multidimensionnelle, *f*(*m*) est une mesure associée à une fonction d'agrégation, *val* représente une valeur prise par la mesure, p_{k1} (respectivement p_{k2}) désigne un paramètre, pos_1 (respectivement pos_{k2}) représente une valeur prise par le paramètre p_{k1} (respectivement p_{k2}).

La sauvegarde et la manipulation de l'expertise des décideurs au travers d'annotations permettent d'intervenir à deux niveaux :

- **Au niveau schéma**. Les annotations facilitent le processus de prise de décision par une plus grande compréhension de la sémantique des composants et des instances d'une base de données multidimensionnelle.
- **Au niveau analyses décisionnelles**. La spécification d'analyses décisionnelles au travers d'une table multidimensionnelle accompagnée d'une réflexion critique s'apparente au concept de lecture active [Adler, *et al.*, 1972]. La tâche des décideurs est alors facilitée par la conservation de la réflexion critique du décideur induite par l'analyse, mais également par le partage de ces réflexions entre les différents décideurs et experts.

L'usage des annotations suit donc soit une modalité personnelle, soit une modalité collective :

- **Usage personnel**. Les annotations matérialisent la réflexion et l'analyse de l'usager décideur rendant leur réutilisation possible.
- **Usage collectif**. Lorsque l'analyse est complexe, l'avis d'un autre expert est souvent sollicité, ce qui peut donner lieu à des débats argumentés visant à atteindre un consensus pour une prise de décision collégiale. Le support par des annotations de cet échange permet de sauvegarder et réutiliser les expertises.

Exemple. L'exemple est extrait de notre publication [Cabanac, Chevalier, Ravat, Teste, 2007][9] dans *International Conference on Data Warehousing and Knowledge Discovery*.

La figure suivante présente un exemple de constellation comportant des annotations globales (A1 à A6). Un premier usager décideur U1 annote localement la table multidimensionnelle (A7 à A10), puis un second usager U2 annote cette même table multidimensionnelle (A11) en réponse à l'annotation A9.

Les ancres de ces annotations sont définies par les expressions suivantes (où C1 désigne la constellation et MT1 la table multidimensionnelle) :

- A1: (λ, CUSTOMERS, λ) ou (λ,λ, CUSTOMERS).
- A2: (λ,CUSTOMERS.HGEO, λ) ou (λ, λ, CUSTOMERS.HGEO).
- A3: (λ, DATES.HYEAR/YEAR, λ) ou (λ, λ, DATES.HYEAR/YEAR).
- A4: (λ, CUSTOMERS/IDC/LASTNAME, λ).
- A5: (C1.ACCOUNTS, λ, λ).
- A6: (C1.ACCOUNTS/SUM_BALANCES, λ, λ).
- A7: (MT1, DATES.HYEAR/YEAR='2007', λ).
- A8: (MT1, CUSTOMERS.HGEO/COUNTRY='France'/CITY='Toulouse', λ).
- A9: (MT1.ACCOUNTS/SUM(SUM_BALANCES), DATES.HYEAR/YEAR = '2007', CUSTOMERS.HGEO/COUNTRY = 'France'/CITY = 'Toulouse').
- A10: (MT1, DATES.HYEAR/YEAR='2007', λ).

[9] Article sélectionné 'top 10 DaWaK'07' pour parution en version étendue dans la série *Advances in Data Warehousing and Mining* (ADWM), 2009.

– A11: L'ancre est identique à celui de A9, seul le contenu des annotations diffère.

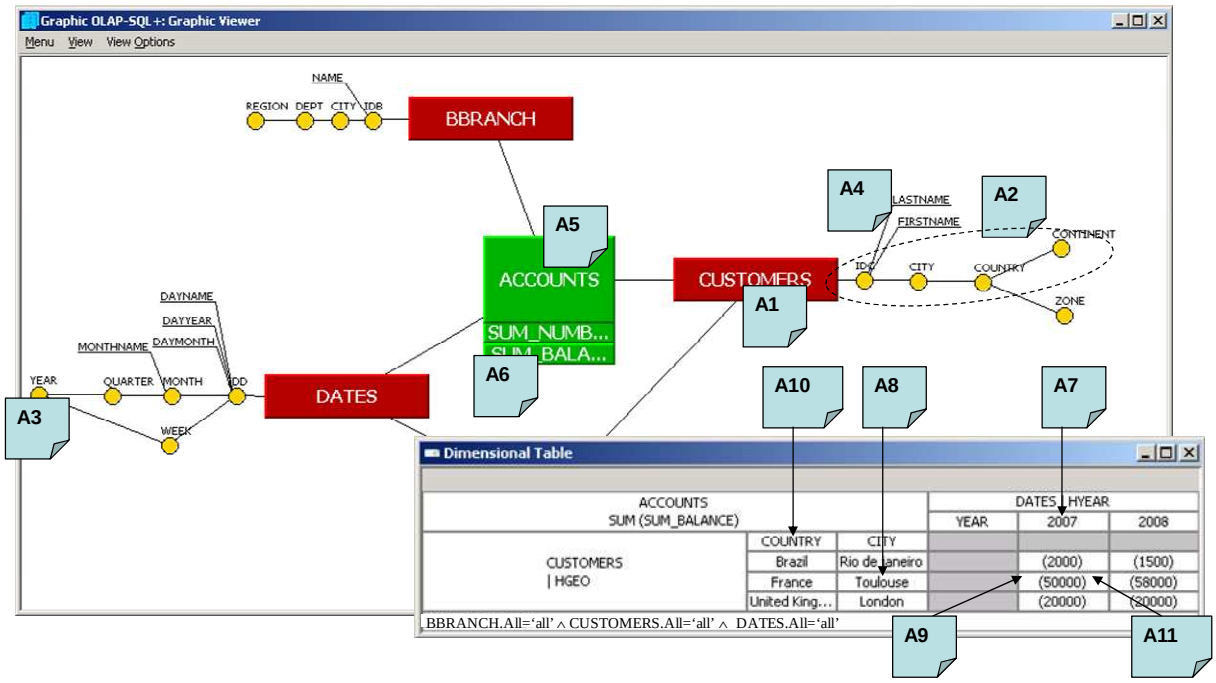


Figure 36 : Exemple d'annotations globales et locales.

3.4 Manipulations OLAP personnalisées

La constellation personnalisée (*Rule*^{CP}) annotée (*Annotate*^{CP}) sert de support aux manipulations OLAP de l'usager qui sont effectuées de manière classique (*cf.* chapitre 3). Le décideur spécifie un seuil qui fixe le niveau pour qualifier les propriétés préférées. Les annotations qui représentent l'expertise des décideurs, enrichissent les données affichées dans les tables multidimensionnelles résultats.

Exemple. Nous poursuivons les exemples précédents. L'usager décideur affiche les comptes en fonction des dates et des clients. La table multidimensionnelle est obtenue par l'expression algébrique suivante :

DISPLAY(Accounts, {SUM(Balance)}, Customers, HGeo, Dates, HYear) = TRES

Dans le contexte classique sans personnalisation, la table multidimensionnelle TRES obtenue est définie par :

– SRES = (Accounts, {SUM(Balance)}),
– LRES = (Customers, HGeo, <All, Continent>),
– CRES = (Dates, HYear, <All, Year>),
– RRES = Customers.All = 'all' ∧ Dates.All = 'all' ∧ BBranch.All = 'all'.

Dans le contexte personnalisé, l'expression algébrique utilise un seuil de 0.9 et la table multidimensionnelle TRES obtenue est identique à l'exception des lignes :

DISPLAY(Accounts, {SUM(Balance)}, Customers, HGeo, Dates, HYear, **0.9**) = TRES avec LRES = (Customers, HGeo, <All, Country, City>).

Le système OLAP prend en compte les propriétés dont le poids est supérieur 0.9 (seuil fixé dans la règle `CustomerRule`) : `setWeight('Country',1)` et `setWeight('City',1)`. La figure suivante décrit la table multidimensionnelle obtenue avec les annotations associées.

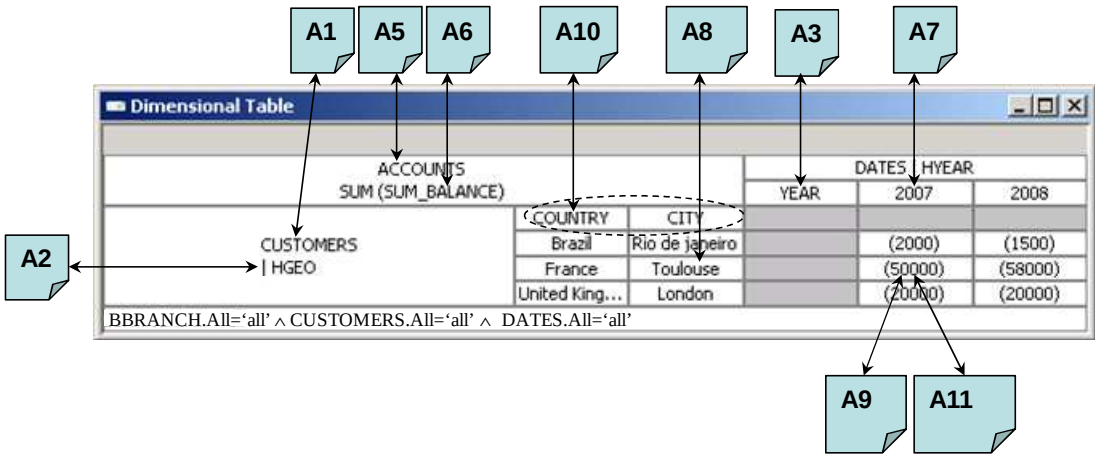


Figure 37 : Exemple de table multidimensionnelle annotée.

Un avantage de la personnalisation concerne la simplification de la navigation pour l'utilisateur. En effet, dans un contexte non personnalisé, l'usager doit réaliser la séquence d'opérations suivante pour obtenir la même table multidimensionnelle tandis que dans le contexte personnalisé, le système OLAP est capable d'anticiper les besoins de l'usager en fonction des préférences contextuelles qu'il a exprimées.

DISPLAY(Accounts, {(SUM(Balance)}, Customers, HGeo, Dates, HYear) = T_1

ROLLUP(T_1, Customers, All) = T_2

DRILLDOWN(T_2, Customers, Country) = T_3

DRILLDOWN(T_3, Customers, City) = T_{RES}

En modifiant le seuil, il est possible d'obtenir des résultats différents prenant en compte d'autres propriétés préférées. Ainsi si le seuil est fixé à 0.75, la table multidimensionnelle comporte alors des entêtes de lignes définies par L_{RES} = (Customers, HGeo, <All, Country, City, (Firstname, Lastname)>). Non seulement les propriétés FIRSTNAME et LASTNAME sont ajoutées dans la table résultante, mais toutes les annotations (A4 dans notre exemple) ancrées à ces nouvelles propriétés sont rendues disponibles et visibles.

Nous présentons la mise en œuvre R-OLAP des constellations personnalisées annotées dans le chapitre 6 de ce mémoire avec le prototype Personal-GOLAP.

4 Approche qualitative

L'approche quantitative facilite les traitements par l'expression des préférences utilisateur au travers de valeurs numériques. L'utilisation de poids et de seuils permet de développer des mécanismes de personnalisation peu coûteux pour le système OLAP. Cependant il est parfois difficile de configurer de manière adéquate ces valeurs face aux besoins de l'usager, d'autant plus qu'il peut voir ses besoins varier suivant le contexte d'utilisation et l'analyse. L'approche qualitative définit les préférences de l'usager sur les données sans utiliser un score pouvant s'avérer difficilement ajustable [Agrawal, *et al.*, 2000].

4.1 Contexte d'analyse

Comme évoqué à la section 1, les décideurs explorent interactivement l'espace multidimensionnel par une succession d'opérations de manipulation OLAP [Choong, *et al.*, 2003] [Dittrich, *et al.*, 2005]. Compte tenu de la propriété de fermeture de notre algèbre OLAP (*cf.* chapitre 3), chaque opération s'applique sur une table multidimensionnelle (à

l'exception du constructeur DISPLAY) et en produit une nouvelle résultant de la transformation opérée. Ces tables comportent à la fois les valeurs analysées et les éléments de structures (fait, dimension, hiérarchie). Afin de rendre notre approche indépendante de la structure de visualisation, nous définissons le concept de **contexte d'analyse** par une description arborescente des tables multidimensionnelles [Jerbi, Ravat, Teste, Zurfluh, 2008].

Définition. Un contexte d'analyse *CA* est défini par (C^F; $\mathcal{C}^D$; $\mathcal{C}^R$) où

– $C^F = NF_i[/f(m) \in \{[val]+\}]+$ représente le sujet (fait F_i) en cours d'analyse,

– $\mathcal{C}^D = \{C^{D1}, \ldots, C^{Du}\}$ représente les axes de l'analyse en cours avec $\forall i \in [1..u]$, $C^{Di} = ND_i[.NH_j]?[/(p_{k1},p_{k2}) \in \{[(val_1,val_2)]+\}]+$ où $p_{k1} \in A^{Di}$, $p_{k2} \in A^{Di}$ et $(val_1,val_2) \in dom(p_{k1}) \times dom(p_{k2})$,

– $\mathcal{C}^R = \{pred^F, pred^{D1}, \ldots, pred^{Du}\}$ est l'ensemble des prédicats définissant les restrictions sur les valeurs analysées.

Exemple. Considérons une analyse où l'usager décideur affiche (DISPLAY) les comptes en fonction des dates et des clients, puis il spécialise (SELECT) son analyse sur les années 2007 et 2008. La table multidimensionnelle obtenue détermine le contexte d'analyse suivant :

– C^F = ACCOUNTS/SUM(SUM_BALANCE),

– $\mathcal{C}^D$ = {CUSTOMERS.HGEO
/(ALL,COUNTRY)∈{(*All, Brazil*), (*All, France*), (*All, United Kingdom*)}
/(COUNTRY,CITY)∈{(*Brazil, Rio de janeiro*), (*France, Toulouse*), (*United Kingdom, London*)},
DATES.HYEAR
/(ALL,YEAR)∈{(*All, 2007*), (*All, 2008*)}},

– $\mathcal{C}^R$ = {ACCOUNTS.ALL = *all*,
CUSTOMERS.ALL = *all*,
DATES.YEAR = *2007* ∨ DATES.YEAR = *2008*,
BBRANCH.ALL = *all*}.

Associé au concept de contexte d'analyse, nous définissons un formalisme permettant de représenter un contexte sous une forme arborescente comme l'illustre la figure suivante. L'arbre proposé comporte des nœuds représentant les composants structurels (fait, mesure, dimension, hiérarchie, paramètre et attribut faible) ainsi que des nœuds ordonnés représentant les valeurs. Sur les nœuds représentant les propriétés, il est possible de placer des prédicats de restriction du domaine des valeurs analysées dans le contexte d'analyse modélisé.

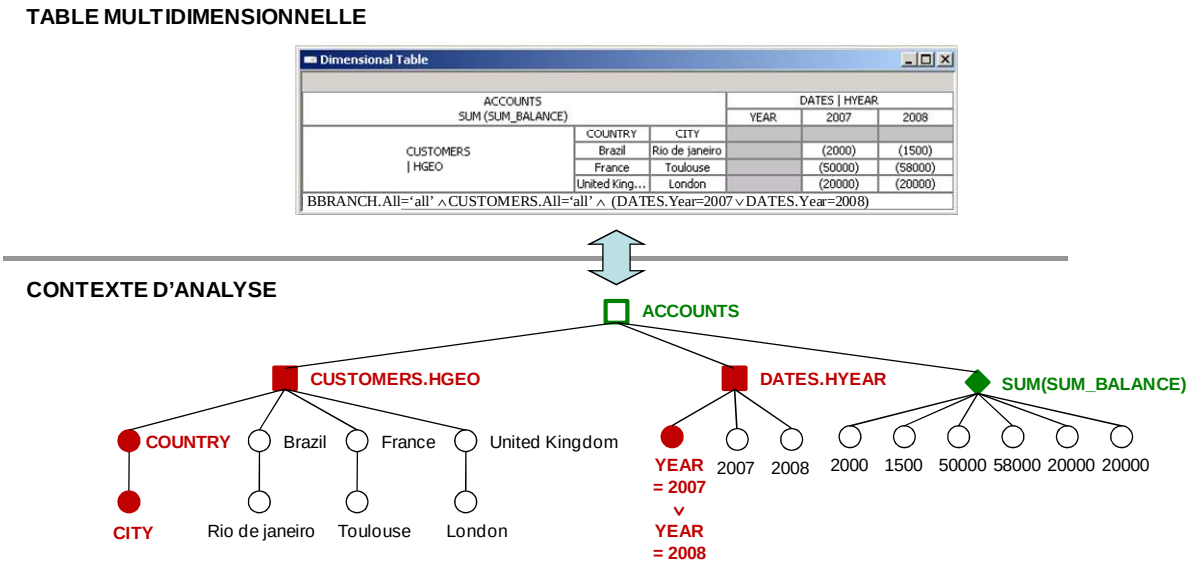


Figure 38 : Exemple de contexte d'analyse.

Ainsi un contexte d'analyse représente l'état courant de l'analyse menée par un usager décideur. L'analyse OLAP est donc vue comme une succession de contextes d'analyse sur lesquels le décideur applique des opérations de manipulation OLAP formant un **graphe de contextes** d'analyse où les nœuds sont les contextes d'analyse et les arcs sont les opérations de manipulation OLAP appliquées. La figure suivante illustre ce principe [Jerbi, Ravat, Teste, Zurfluh, 2009b].

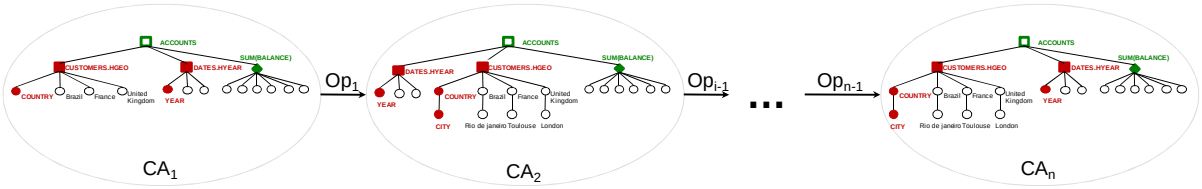


Figure 39 : Principe de navigation dans un graphe d'analyse.

4.2 *Constellation à base de préférences contextuelles*

Nous étendons le concept classique de constellation en intégrant les préférences usagers.

Définition. Une constellation personnalisée *CP* est définie par ($\mathbb{F}^{CP}$; $\mathbb{D}^{CP}$; $Star^{CP}$; $Preference^{CP}$) où

- $\mathbb{F}^{CP} = \{F_1,\ldots, F_n\}$ est l'ensemble des faits,
- $\mathbb{D}^{CP} = \{D_1,\ldots, D_m\}$ est l'ensemble des dimensions,
- $Star^{CP} : \mathbb{F} \rightarrow 2^{\mathbb{D}}$ associe chaque fait à un sous-ensemble des dimensions en fonction desquelles il est analysable,
- $Preference^{CP} = \{P^{CP}_1, P^{CP}_2,\ldots\}$ est l'ensemble de règles actives de personnalisation.

L'usager peut exprimer des **préférences**, notées simplement P_i, sur les éléments de **structure** d'une constellation et/ou sur les **valeurs**. Ces préférences peuvent être absolues ou contextuelles : une préférence absolue est toujours prise en compte par le système, tandis

qu'une préférence contextuelle est prise en compte par le système lorsque le contexte d'analyse courant couvre le contexte de la préférence.

Définition. Une préférence P_i est définie par ($\succ_{Pi}$; C^{Pi}) où

- $\succ_{Pi}$ est une relation d'ordre sur un ensemble E d'éléments.
  - Si les éléments de E sont des éléments de structure de la constellation, $E \in \{F; D; H; M; A\}$, alors P_i est une **préférence de structures**,
  - Si les éléments de E sont des prédicats sur les propriétés de la constellation, alors P_i est une **préférence de valeurs**.
- $C^{Pi} = (C^F; \mathcal{G}^D; \mathcal{G}^R)$ est le contexte de la préférence. Le contexte de préférence est défini comme un contexte d'analyse, mais pouvant comporter des parties vides.

Exemple. On considère les préférences d'un décideur :

(P1) « Je m'intéresse aux soldes des comptes bancaires »

(P2) « Je préfère analyser les soldes des comptes par pays, puis par villes des clients »

(P3) « Je m'intéresse aux clients européens lorsque mon analyse porte sur les comptes bancaires de l'année courante (2009) »

Les préférences P1 et P2 portent sur la structure de la constellation tandis que la préférence P3 concerne les valeurs. En outre, la préférence P1 est une préférence absolue valide pour tous les contextes (le contexte vide est représenté par λ) tandis que P2 et P3 sont contextuelles. Conformément aux définitions préalables, les préférences sont définies par les expressions suivantes :

- P1 : (Sum_Balance ;
  λ)
- P2 : (Country $\succ_{P2}$City ;
  (ACCOUNTS/SUM(SUM_BALANCE) ; {CUSTOMERS} ;_))
- P3 : (CUSTOMERS.CONTINENT='Europe' ;
  (ACCOUNTS; {CUSTOMERS}; DATES.YEAR = *2009*))

4.3 Recommandations

Notre objectif de conception d'un système OLAP de recommandation est d'aider l'usager dans sa navigation au sein de l'espace multidimensionnel. Soit un contexte d'analyse courant CA_i. Le décideur transforme CA_i en CA_{i+1} par l'application d'une opération OLAP Op_i. Comme l'illustre la Figure 34, le système de recommandation OLAP cherche à déterminer un ensemble de recommandations $\mathcal{R}_i$:

- en enrichissant le contexte d'analyse CA_{i+1} obtenu par l'opération Op_i de l'utilisateur, $\mathcal{R}_i$,={CA^+_{i+1}},
- en déterminant par anticipation un futur contexte d'analyse CA_{i+j} que l'usager est susceptible de construire, $\mathcal{R}_i$,={CA_{i+j}},
- en suggérant les contextes d'analyse CA_k alternatifs à la navigation de l'usager, $\mathcal{R}_i$={CA_{k1},…, CA_{km}}.

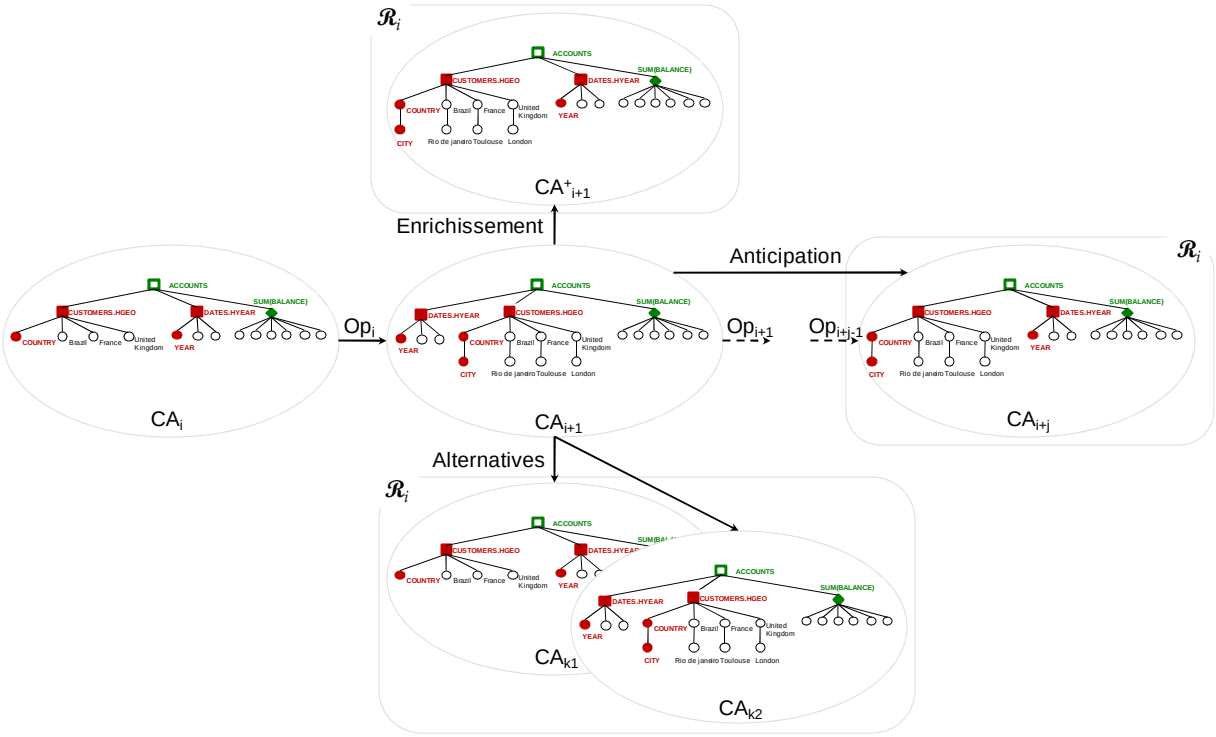


Figure 40 : Principe des recommandations de contextes d'analyse.

Nous détaillons dans la suite les principes permettant de calculer $\mathcal{R}_i$ selon les trois types recommandations.

4.3.1 *Anticipation*

L'anticipation vise à déterminer un contexte d'analyse futur que l'utilisateur devrait construire durant son processus d'analyse. Pour ce faire, le contexte d'analyse CA_{i+1} qui est calculé sert à déterminer un contexte d'analyse à recommander sur la base de préférences candidates $\mathcal{P}_i$. Parmi l'ensemble des préférences candidates $\mathcal{P}_i$, un sous-ensemble des préférences les plus couvrantes $\mathcal{P}^{Max}_i$ est déterminé. Chaque préférence $P_j \in \mathcal{P}^{Max}_i$ est intégrée dans CA_{i+1} formant ainsi le contexte d'analyse anticipé ; le symbole $\oplus$ représente l'intégration d'une préférence dans un contexte d'analyse. Nous présentons ci-dessous l'algorithme permettant de calculer les recommandations basées sur l'anticipation.

Algorithme de recommandation par anticipation.

```
Entrées :
   CAi
   Opi
Sortie :
   ℛi

Début
   CAi+1←ConstruireContexte(CAi; Opi);
   (𝒫i; χ^Max)←PreferencesCandidates(CAi+1; Preference^CP);
   𝒫^Max_i←PreferencesCouvrantes(𝒫i; χ^Max);
   CAi+j←CAi+1;
   Pour Chaque Pj∈𝒫^Max_i Faire
```

```
        CA_{i+j} ← CA_{i+j} ⊕ P_j;
    FinPour;
    ℛ_i ← {CA_{i+j}};
Fin.
```

- **PreferencesCandidates(CA_{i+1}; $Preference^{CP}$)**.

Le système construit l'ensemble des préférences candidates $\mathcal{P}_i$ en sélectionnant dans l'ensemble des préférences $Preference^{CP}$, les préférences dont le contexte est couvert par le contexte d'analyse CA_{i+1}. La couverture χ^{Pj} d'une préférence P_j correspond au nombre d'arcs en commun entre le contexte d'analyse et celui de la préférence :

$\chi^{Pj} = N^{CPj} / N^{CAi+1}$

avec N^{CPj} : nombre d'arcs de P_j présents dans le contexte d'analyse CA_{i+1}

N^{CAi+1} : nombre d'arcs du contexte d'analyse CA_{i+1}

Actuellement, nous ne considérons que les cas de **couverture totale** : un contexte de préférence cp_j est couvert par le contexte d'analyse CA_{i+1} si et seulement si tous les arcs (v_{k1}, v_{k2}) appartiennent à CA_{i+1}. Lorsqu'il s'agit de nœuds avec prédicats, des conflits peuvent survenir : deux prédicats $pred_{k1}$ et $pred_{k2}$ (prédicats normalisés) sont incompatibles si et seulement s'ils portent sur la même propriété (mesure ou attribut de dimension) et $pred_{k1} \wedge pred_{k2} = \varnothing$. L'**incompatibilité de prédicats** induit une non prise en compte de la préférence.

- **PreferencesCouvrantes($\mathcal{P}_i$; χ^{Max}).**

L'ensemble des préférences les plus couvrantes $\mathcal{P}^{Max}_i$ est obtenu en utilisant la couverture maximale χ^{Max}.

Afin d'affiner le processus de recommandation, il est possible d'introduire certaines heuristiques dans le calcul des préférences candidates telles que :

- définir une politique de priorité sur les structures de la constellation afin de pondérer les arcs (le fait analysé est prioritaire sur les dimensions, elles-mêmes prioritaires sur les hiérarchies…),
- classer les préférences *ex aequo* en fonction de leur date de création (la plus récente, la plus ancienne…).

Exemple. Considérons une table multidimensionnelle représentant les soldes de comptes bancaires en fonction des clients et des dates. On suppose que l'utilisateur modifie le contexte d'analyse CA_i par l'opération Op_i suivante :

DRILLDOWN(T_i ; CUSTOMERS ; Country)

Parmi les préférences de la constellation $Preference^{CP}$, le système OLAP détermine les préférences candidates $\mathcal{P}_i$ = {P1 ; P2}, puis l'ensemble des préférences les plus couvrantes $\mathcal{P}^{Max}_i$ = {P2}. La préférence P2 est choisie car elle est « la plus couvrante » : $\chi^{P1} = 0$ et $\chi^{P2} = 2$. Comme l'illustre la figure suivante, le contexte d'analyse CA_{i+1} est complété par la préférence P2 formant une recommandation $\mathcal{R}_i$ = {CA_{i+j}} permettant de déterminer la table multidimensionnelle résultat.

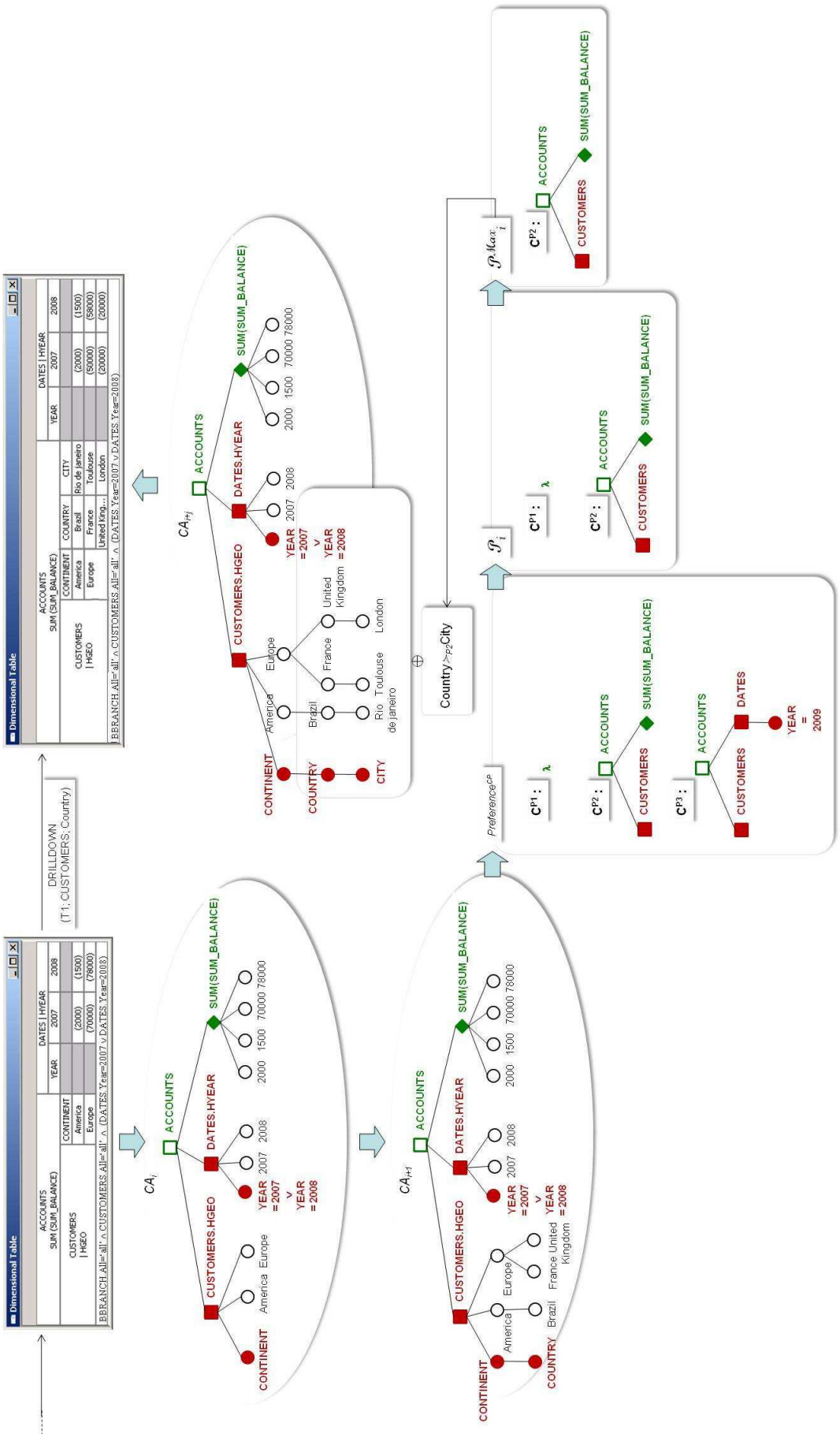


Figure 41 **:** Exemple de recommandation par anticipation.

4.3.2 *Alternatives*

La recommandation d'alternatives consiste à proposer des contextes d'analyse considérés comme nouveaux et alternatifs dans l'analyse de l'usager. Pour ce faire, le contexte d'analyse CA_{i+1} qui est calculé sert à déterminer un ensemble $\mathcal{R}_i$ de contextes d'analyse à recommander sur la base de préférences candidates $\mathcal{P}_i$. A partir des préférences candidates $P_j \in \mathcal{P}_i$, une recommandation CA_k est déterminée en intégrant la préférence dans CA_{i+1} formant au fur et à mesure un ensemble de recommandations $\mathcal{R}_i$.

Algorithme de recommandation d'alternatives.

```
Entrées :
   CA_i
   Op_i
Sortie :
   R_i

Début
   CA_{i+1} ← ConstruireContexte(CA_i ; Op_i);
   (P_i ; χ^Max) ← PreferencesCandidates(CA_{i+1} ; Preference^CP);
   R_i ← ∅;
   Pour Chaque P_j ∈ P_i Faire
      CA_k ← CA_{i+1} ⊕ P_j;
      R_i ← R_i ∪ {CA_k};
   FinPour;
Fin.
```

Exemple. Considérons l'exemple précédent en appliquant cette fois le processus calculant les recommandations alternatives. On suppose toujours que l'utilisateur modifie le contexte d'analyse CA_i par l'opération de forage Op_i suivante :

DRILLDOWN(T_i ; CUSTOMERS ; Country)

Parmi les préférences de la constellation $Preference^{CP}$, le système OLAP détermine l'ensemble des préférences candidates $\mathcal{P}_i$ = {P1 ; P2}. Comme l'illustre la figure suivante, ces préférences candidates sont utilisées pour calculer les recommandations $\mathcal{R}_i$ = {CA_{k1} ; CA_{k2}} en les intégrant à tour de rôle dans le contexte d'analyse courant CA_{i+1}.

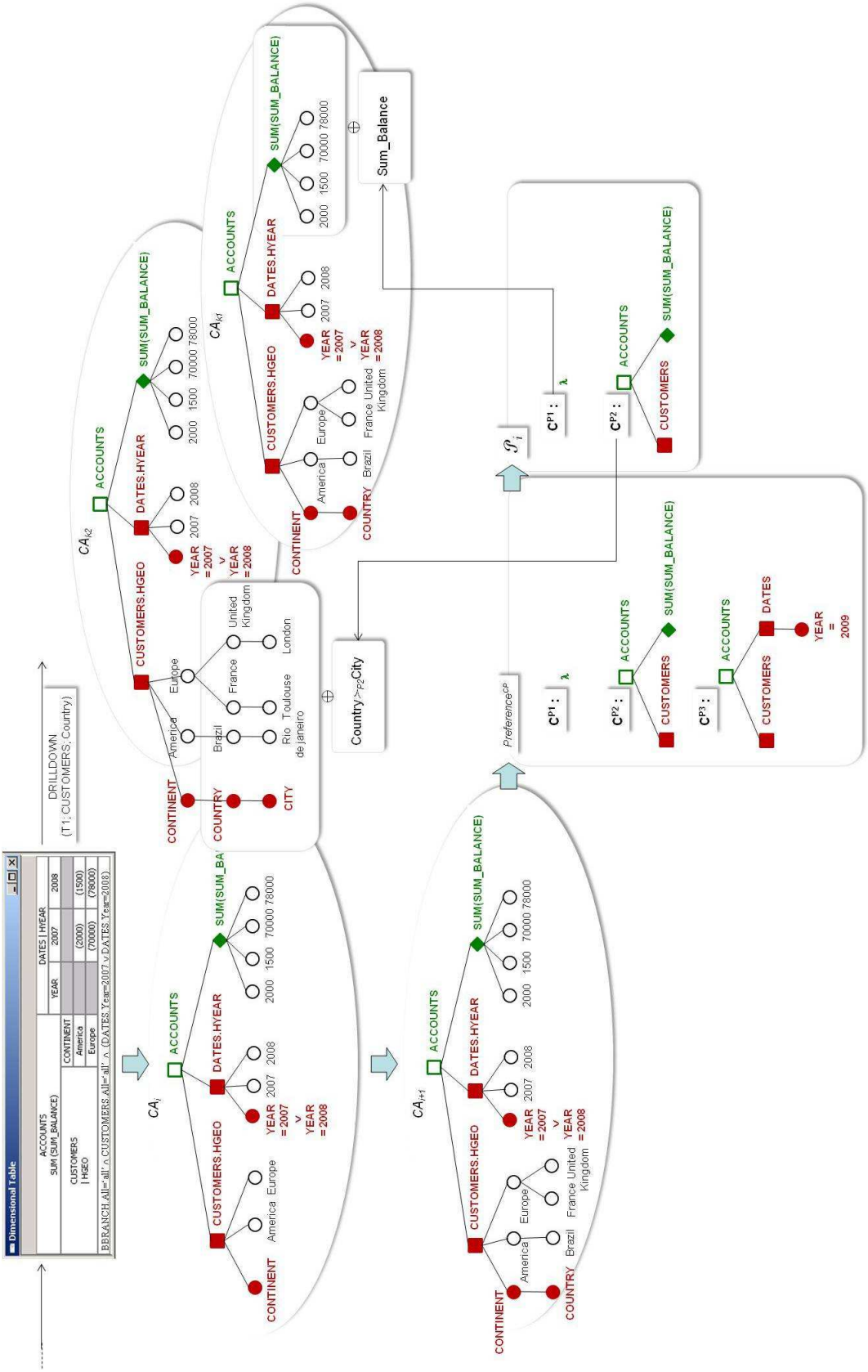


Figure 42 : Exemple de recommandations alternatives.

4.3.3 *Enrichissement*

L'enrichissement vise à déterminer un contexte d'analyse complet CA^{+}_{i+1} alors que l'utilisateur n'a fourni qu'une opération OLAP incomplète formant un contexte d'analyse CA_{i+1} potentiellement partiel (un contexte d'analyse partiel est non affichable sous forme de table multidimensionnelle). Pour ce faire, est calculé l'ensemble des préférences candidates $\mathcal{P}_i$, et le sous-ensemble des préférences les plus couvrantes $\mathcal{P}^{Max}_i$. Chaque préférence $P_j \in \mathcal{P}^{Max}_i$ enrichit le contexte d'analyse partiel CA_{i+1}. A l'issue du processus, le contexte d'analyse peut rester partiel, dans ce cas, de nouvelles itérations du processus sont réalisées jusqu'à obtenir un contexte complet ou avoir utilisé toutes les préférences candidates disponibles : les ensembles $\mathcal{P}_i$ et $\mathcal{P}^{Max}_i$ sont recalculés en fonction du contexte enrichi mais toujours partiel, puis les préférences les plus couvrantes sont intégrées. Les préférences intégrées dans le contexte d'analyse ne sont plus prises en compte dans les itérations suivantes.

Algorithme de recommandation par enrichissement.

```
Entrées :
   CA_i
   Op_i
Sortie :
   R_i

Début
   CA_{i+1}←ConstruireContexte(CA_i; Op_i);
   (P_i; χ^Max)←PreferencesCandidates(CA_{i+1}; Preference^CP);
   CA+_{i+1}←CA_{i+1};
   continu←vrai;
   Tant Que continu=vrai Et P_i≠∅ Faire
      continu←faux;
      P^Max_i←PreferencesCouvrantes(P_i; χ^Max);
      Pour Chaque P_j∈P^Max_i Faire
         CA+_{i+1}←CA+_{i+1}⊕P_j;
         Si P_j intégrée Alors
            Preference^CP = Preference^CP \ {P_j};
         FinSi;
      FinPour;
      Si CA+_{i+1} partiel Alors
         (P_i; χ^Max)←PreferencesCandidates(CA_{i+1}; Preference^CP);
         continu←vrai;
      FinSi;
   FinTantQue;
   R_i←{CA+_{i+1}};
Fin.
```

Exemple. Nous reprenons l'exemple précédent en le complétant par deux nouvelles préférences :

(P4) « l'usager s'intéresse d'abord aux clients européens puis américains lorsqu'une analyse porte sur les soldes des comptes bancaires de l'année 2008 et que les clients sont visualisés en fonction des villes ». Cette nouvelle préférence est définie par l'expression : (CUSTOMERS.CONTINENT='Europe' $\succ_{P4}$ CUSTOMERS.CONTINENT = 'America' ; (ACCOUNTS/

SUM(SUM_BALANCE) ; {CUSTOMERS.HGEO / (ALL, CONTINENT) / (CONTINENT, COUNTRY) / (COUNTRY, CITY)} ; DATES.YEAR = *2008*)).

(P5) « l'usager préfère la hiérarchie géographique des clients durant les analyses des soldes des comptes bancaires ». Cette nouvelle préférence est définie par l'expression : (HGEO ; (ACCOUNTS/SUM(SUM_BALANCE) ; {CUSTOMERS} ;_))

L'utilisateur exprime une opération de rotation de hiérarchies incomplète Op_i pour modifier le contexte d'analyse CA_i : ROTATE(T_i ; CUSTOMERS ; CUSTOMERS ; _). Le contexte d'analyse CA_{i+1} obtenu est un contexte partiel, non affichable, puisque la hiérarchie n'est pas spécifiée. Le système ne connaît donc pas les paramètres sur l'axe en ligne des clients. L'algorithme de recommandation réalise deux itérations afin d'obtenir un contexte d'analyse enrichi complet CA^{+}_{i+1}.

Durant la première itération, deux préférences sont qualifiées : P5 peut être intégrée, mais P2 non car la dimension est sans hiérarchie courante.

Tableau 12 : Première itération de l'enrichissement.

Preference^{CP}	$\mathcal{P}_i$	$\mathcal{P}Max_i$
C^{P1} : λ C^{P2} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE) C^{P3} : ACCOUNTS — CUSTOMERS, DATES — YEAR = 2009 C^{P4} : ACCOUNTS — CUSTOMERS — CONTINENT — COUNTRY — CITY ; DATES — YEAR = 2009 C^{P5} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE)	C^{P1} : λ C^{P2} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE) C^{P5} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE)	C^{P2} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE) C^{P5} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE)
Contexte d'analyse enrichi	CA_{i+1} : ACCOUNTS — CUSTOMERS.HGEO, DATES.HYEAR, SUM(SUM_BALANCE) DATES.HYEAR : YEAR = 2007 ∨ YEAR = 2008, 2007, 2008 SUM(SUM_BALANCE) : 2000, 1500, 70000, 78000	

A l'issue de la seconde itération, le contexte d'analyse enrichi CA^{+}_{i+1} est complet. Le processus s'arête et le contexte d'analyse obtenu est affiché.

Tableau 13 : Deuxième itération de l'enrichissement.

Preference^{CP}	$\mathcal{P}_i$	$\mathcal{P}Max_i$
C^{P1} : λ C^{P2} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE) C^{P3} : ACCOUNTS — CUSTOMERS, DATES — YEAR = 2009 C^{P4} : ACCOUNTS — CUSTOMERS — CONTINENT — COUNTRY — CITY; DATES — YEAR = 2009	C^{P1} : λ C^{P2} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE)	C^{P2} : ACCOUNTS — CUSTOMERS, SUM(SUM_BALANCE)
Contexte d'analyse enrichi	CA^{+}_{i+1} : ACCOUNTS — CUSTOMERS.HGEO — COUNTRY — Brazil (Rio de janeiro), France (Toulouse), United Kingdom (London) ; DATES.HYEAR — YEAR = 2007 ∨ YEAR = 2008 (2007, 2008) ; SUM(SUM_BALANCE) — 2000, 1500, 70000, 78000	

5 Bilan

Ce chapitre a détaillé mes contributions à la personnalisation des systèmes OLAP. Mes travaux ont abouti à la définition de mécanismes permettant la constitution d'une mémoire de l'expertise et des recommandations lors de la navigation multidimensionnelle.

5.1 Résultats de nos travaux

Mes travaux sur la personnalisation des BDM [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a] proposent de modéliser les préférences des décideurs selon deux approches complémentaires : quantitative [Ravat, Teste, 2008g] [Ravat, Teste, Zurfluh, 2007g] et qualitative [Jerbi, Ravat, Teste, Zurfluh, 2009a, 2009b, 2009c, 2008].

La modélisation des préférences de l'usager par une approche quantitative consiste simplement à représenter l'intérêt d'une donnée par une pondération associée. Une originalité de mon approche est la mise en place de mécanismes dynamiques de type ECA pour définir des préférences contextuelles [Ravat, *et al.*, 2008g]. Une autre proposition originale de ces recherches concerne la définition du concept de mémoire d'expertise

[Cabanac, Chevalier, Ravat, Teste, 2009] à base d'annotations décisionnelles. Elles permettent de conserver au sein du système OLAP le patrimoine immatériel (commentaires, discussions et débats, prise de décision…) non pris en compte dans les systèmes classiques, et d'ancrer ces éléments matérialisés aux données analysées.

Pour simplifier l'acquisition des préférences, j'ai orienté mes recherches vers le développement d'une approche qualitative qui consiste à définir les préférences de manière relative (relation d'ordre) [Jerbi, Ravat, Teste, Zurfluh, 2009a, 2009b, 2009c, 2008]. Le résultat important de cette approche est l'élaboration de recommandations contextuelles [Jerbi, Ravat, Teste, Zurfluh, 2009b] lors des analyses. Nous identifions trois types de recommandation : par enrichissement, anticipation et alternatives.

Pour valider ces propositions, nous avons développé le prototype PERSONAL-GOLAP [Jerbi, 2007] [Ghalamallah, 2008] présenté dans le chapitre 6 de ce mémoire.

5.2 Encadrements et diffusion scientifique

Ces travaux ont été menés dans le cadre de la thèse de Houssem Jerbi que je co-encadre. Ils ont permis également d'offrir un cadre d'étude à des stages de Master Recherche (M2R) dont j'ai assuré l'encadrement :

(1) F. Atigui, « Personnalisation de requêtes dans une base de données multidimensionnelle » en 2009,
(2) A. Ghalamallah, « Personnalisation d'une base de données multidimensionnelle » en 2008,
(3) H. Jerbi, « Mémoire d'expertises décisionnelles à base d'annotations » en 2007.

Le tableau suivant dresse un panorama des thèmes étudiés, des étudiants encadrés et des publications effectuées dans le cadre de ce troisième axe de nos travaux portant sur l'intégration de l'usager dans les bases de données multidimensionnelles.

Tableau 14 : Etudiants encadrés et publications de l'axe 3.

Thèmes	Thèses	Master/D.E.A.	Publications
Modélisation Quantitative de l'Usager et Annotations		A. Ghalamallah H. Jerbi	▪ RI AoIS [Ravat, Teste, 2008g] ▪ CI DAWAK'07 [Cabanac, Chevalier, Ravat, Teste, 2007] ▪ OI ADWM [Cabanac, Chevalier, Ravat, Teste, 2009] ▪ CN INFORSID'07 [Ravat, Teste, Zurfluh, 2007g] EDA'06 [Cabanac, Chevalier, Ravat, Teste, 2006b] EGC'06 [Cabanac, Chevalier, Ravat, Teste, 2006a]
Modélisation Qualitative des Usagers et Recommandations	H. Jerbi	F. Atigui	▪ CI DAWAK'09 [Jerbi, Ravat, Teste, Zurfluh, 2009c] ICEIS'09 [Jerbi, Ravat, Teste, Zurfluh, 2009b] ICDIM'08 [Jerbi, Ravat, Teste, Zurfluh, 2008] ▪ CN EDA'09 [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a] EGC'09 [Jerbi, Ravat, Teste, Zurfluh, 2009a]

5.3 Perspectives

Les travaux que nous avons réalisés proposent des mécanismes de recommandation basés à la fois sur une structure de visualisation des données unique et sur le seul contexte d'analyse courant, sans tenir compte du chemin de navigation dans le graphe d'analyse. Une première perspective envisageable est donc de proposer différentes structures permettant de visualiser les données sous différentes facettes en fonction de préférences contextuelles. Une seconde perspective concerne les extensions à apporter aux mécanismes de recommandation. Actuellement, notre mécanisme de recommandation est identique pour chaque type de recommandation : enrichissement, anticipation et alternatives. Je souhaite développer des mécanismes différents en fonction du type de recommandation. Par exemple, il serait intéressant d'étendre les mécanismes de recommandation par anticipation en exploitant l'historique des contextes d'analyse ainsi que les opérations permettant d'obtenir ces contextes d'analyse, tandis qu'il serait pertinent de développer des recommandations alternatives en utilisant des contextes d'analyses d'autres utilisateurs, formant ainsi un système OLAP collaboratif permettant à chaque usager de profiter de l'expertise d'un groupe.

Au-delà de ces perspectives directes, il est nécessaire de développer des mécanismes permettant l'apprentissage des préférences des décideurs rendant la collecte des caractéristiques utilisateurs (profil) automatique et transparente à ce dernier [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a]. Cette perspective se heurte à la définition d'une (des) mesure(s) de similarité dans les entrepôts de données. Il s'agit d'un véritable verrou scientifique relatif au problème de la constitution d'une mesure de similarité permettant de comparer et d'aligner les profils des utilisateurs ainsi que les contextes d'analyse lors des navigations dans l'espace multidimensionnel. Une fois ce verrou levé il sera alors envisageable de proposer un système OLAP adaptatif.

Chapitre 6 - Validations et cadres applicatifs

L'objectif de ce chapitre est de présenter les différents contextes applicatifs qui ont permis de valider mes recherches depuis ma nomination en qualité de maître de conférences à l'Université Paul Sabatier de Toulouse.

1 Equipe de recherche et thèses encadrées

Mes recherches s'effectuent dans l'équipe Systèmes d'Informations Généralisées (SIG) dirigée par le Professeur Claude Chrisment au sein de l'Institut de Recherche en Informatique de Toulouse (IRIT) – UMR 5505 CNRS. Mes travaux se déroulent plus spécifiquement dans la composante SIG-ED (Conception d'Entrepôts de Données) sous la direction du Professeur Gilles Zurfluh en collaboration avec Geneviève Pujolle et Franck Ravat.

Durant ces années, j'ai participé à l'encadrement d'étudiants sur des sujets couvrant les différents axes thématiques de mes recherches :

- 2001-2004, Faiza Ghozzi, « Conception et manipulation de bases de données dimensionnelles à contraintes »,
- 2004-2005, Estella Annoni, « Eléments méthodologiques pour le développement des systèmes décisionnels dans un contexte de réutilisation »,
- 2005-2007, Ronan Tournier, « Analyse en ligne (OLAP) de documents »,
- 2007-…, Houssem Jerbi, « Personnalisation dans les bases de données multidimensionnelles »,
- 2009-…, Faten Atigui, « Entrepôts de données médicales ».

2 Projets et partenaires

Mes travaux ont consisté à proposer des solutions en matière de modélisation et de manipulation au sein des systèmes OLAP. Durant ces années, je suis intervenu dans différents projets impliquant des partenaires industriels et institutionnels.

2.1 Partenaires industriels

CS : Société CS spécialisée en Conception et Développement de Systèmes Critiques (http://www.c-s.fr/).

IBM : Société Informatique (http://www.ibm.com/fr/fr/).

ICR: Institut Claudius Regaud, centre régional de lutte contre le cancer (http://www.claudiusregaud.fr/).

ID-6: Société de Service en Informatique Décisionnelle (http://www.i-d6.com/accueil/novedia.php)

OUTCOME-REA : Association de professionnels de santé (http://outcomerea.org/).

THALES : Société Aérospatiale, Espace, Défense et Sécurité (http://www.thalesgroup.com/).

2.2 Partenaires institutionnels

Cédric : Centre d'Étude et de Recherche en Informatique du Cnam (EA 1395) (http://cedric.cnam.fr/).

CRI : Centre de Recherche en Informatique de Paris 1 (http://crinfo.univ-paris1.fr/).

ERIC : Equipe de Recherche en Ingénierie des Connaissances de Lyon 2 (http://recherche.univ-lyon2.fr/eric/).

LAMIH : Laboratoire d'Automatique, de Mécanique et d'Informatique industrielles et Humaines – UMR 8530 CNRS (http://www.univ-valenciennes.fr/LAMIH2/).

LIRMM : Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier – UMR CNRS (http://www.lirmm.fr/).

LIG : Laboratoire d'Informatique de Grenoble – UMR 5217 CNRS (http://liglab.imag.fr/).

LIPN : Laboratoire d'Informatique de Paris-Nord (http://lipn.fr/).

LRI : Laboratoire de Recherche en Informatique de Paris-Sud – UMR 8623 CNRS (http://www.lri.fr/).

LISI : Laboratoire d'Informatique Scientifique et Industrielle (http://www.lisi.ensma.fr/).

PRiSM : Laboratoire de recherche en informatique de Versailles Saint-Quentin-en-Yvelines – UMR 8144 CNRS (http://www.prism.uvsq.fr).

2.3 Description et implication personnelle dans les projets

Une synthèse dressant un panorama des contextes applicatifs dans lesquels s'inscrivent mes recherches est présentée dans le Tableau 15.

2000-2001 : REANIMATIC

- *Cadre* :
  Financement MENRT (2MF).
- *Partenaires* :
  Laboratoires Cédric, LAMIH, LIPN, LIRMM, LISI, PRiSM.
  Association de médecins OUTCOM-REA.
- *Objectif* :
  Modélisation des entrepôts de données médicales évolutives (collectées à partir des bases opérationnelles des services de réanimation) afin d'améliorer la qualité des soins et le devenir des patients dans les services de réanimation des hôpitaux.
- *Implication* :
  Participant

  Mon rôle a été de définir les tâches à réaliser par notre équipe dans le projet. A ce titre, j'ai participé et dirigé différentes réunions avec les membres de l'équipe. J'ai également représenté l'IRIT lors de réunions plénières à Paris.
  Ce projet a servi de cadre applicatif à mes propositions sur l'architecture des systèmes décisionnels et les modèles associés. J'ai ainsi validé certaines de mes

propositions sur les modèles de représentation des données dédiés aux entrepôts et aux magasins dans un contexte relationnel de données médicales.

2001-2002 : GafoDonnées

- *Cadre* :
  Action Spécifique n°20 STIC/CNRS
- *Partenaires* :
  Laboratoires ERIC, LRI
- *Objectif* :
  Rassembler les communautés françaises en base de données et apprentissage automatique, dégager et approfondir des thèmes prioritaires communs.
- *Implication* :
  Responsable IRIT

  Mon rôle a consisté à participer activement aux différentes réunions (à Paris) au cours desquelles j'ai présentés oralement les travaux de l'équipe SIG. Mon implication m'a amené à co-rédiger le rapport final du groupe de travail GafOLAP [Laurent, Marcel, Ravat, Teste, Zurfluh, 2002].
  Ce projet a permis d'étudier l'état de l'art en matière de modélisation multidimensionnelle et de manipulation OLAP. Nous avons pu établir des critères de comparaison, me permettant de positionner et de définir les contributions de mes recherches (*cf.* Tableau 1 et Tableau 2).

2004-2007 : ID-6

- *Cadre* :
  ANRT (contrat 766/2003).
- *Partenaires* :
  Société ID-6.
- *Objectif* :
  Développer une méthode dédiée à la conception de systèmes d'aide à la décision par réutilisation.
- *Implication* :
  Co-encadrant de la thèse d'Estella Annoni [Annoni, 2007].

  Cette collaboration avec la société ID-6 a servi de cadre aux travaux de recherche que j'ai co-encadrés d'Estella Annoni. La société ID-6, spécialisée dans la conception de systèmes décisionnels, nous a offert un vaste champ d'expérimentations pour nos propositions de démarche de conception par capitalisations des systèmes décisionnels. Nous avons pu ainsi appliquer la démarche sur différents domaines fonctionnels et la confronter aux évaluations de différentes personnes : la plate-forme BIPAD [Annoni, 2007] a été conçue, mise en place et utilisée pour permettre la capitalisation au travers d'un catalogue de patrons de conception par des collaborateurs expérimentés et débutants de la société.

2007-2009 : MADSI

- *Cadre* :
  Groupe de travail GdR I3.
- *Partenaires* :
  Laboratoires CRI, LIG...
- *Objectif* :
  Fédération des chercheurs francophones de la communauté des « systèmes d'information » autour des méthodes avancées d'ingénierie.
- *Implication* :
  Co-Responsable IRIT

  Je dirige (co-responsabilité) la participation de l'IRIT au groupe de travail MADSI. A ce titre, je suis amené à participer aux réunions et aux journées (MADSI'07 et MADSI'08) organisées dans le cadre du congrès INFORSID. J'ai ainsi présenté oralement à MADSI'08 les résultats de nos recherches [Ravat, Teste, Tournier, Zurfluh, 2008f] concernant la conception en galaxie des bases de données multidimensionnelles. Cette étude a permis d'établir les critères permettant de qualifier notre approche d'ingénierie guidée par les modèles.

2007-2009 : IAPA

- *Cadre* :
  Projet Laboratoire IRIT / Labellisation Pôle Compétitivité Bio-Santé
  Financement BQR/UPS.
- *Partenaires* :
  ICR – Sociétés CS et IBM.
- *Objectif* :
  Conception d'une infrastructure intégrée dédiée aux professionnels de la santé facilitant l'accès, le partage et surtout des analyses croisées de données biomédicales dans le cadre de la lutte contre le cancer.
- *Implication* :
  Co-Responsable du *Work Package 6*

  J'ai dirigé (co-responsabilité) le lot 6 intitulé « Knowledge discovery: data mining and multidimensional analysis ». De part cette responsabilité j'ai dû rédiger l'ensemble des documents relatifs au lot 6 et participer aux différentes réunions durant lesquelles j'ai exposé oralement les activités du lot 6.
  Ce projet a permis notamment d'étudier les données médicales et d'identifier les problématiques de recherche induites. L'étude des données « anonimisées » (comptes rendus de médecins, extractions de bases de données hospitalières) fournies par le Docteur Courbon de l'Institut Claudius Regaud a permis de dégager des problématiques de recherches : (1) le développement de systèmes décisionnels pour l'analyse multidimensionnelle et la fouille de données appliquées à des documents semi-structurés voire non-structurés, ainsi que (2) l'élaboration de mécanismes pour l'assistance contextuelle des usagers d'un système décisionnel personnalisé de données médicales.
  Une thèse liée à la problématique (1) débute cette année (F. Atigui) tandis que la problématique (2) fait l'objet d'une thèse débutée dès 2007 (H. Jerbi). J'assure le

co-encadrement de ces deux thèses sous la direction scientifique du Professeur Gilles Zurfluh.

2009-2010 : EDM

- *Cadre* :
  Projet Laboratoire IRIT.
- *Partenaires* :
  Société THALES.
- *Objectif* :
  Elaboration d'un entrepôt de données pour la métrologie des projets dans l'aérospatiale.
- *Implication* :
  Responsable IRIT

  Cette année, j'ai initié une collaboration (EDM : Entrepôt de Données pour la Métrologie) avec la société Thalès, qui vise à définir un entrepôt de données et les processus ETL d'alimentation de cet entrepôt. L'objectif est d'homogénéiser les données afin de mettre en place à terme des métriques communes applicables sur l'ensemble des projets informatiques développés dans l'entreprise. Ce projet offre un cadre applicatif permettant de valider mes recherches sur la modélisation en constellation [Ravat, Teste, Tournier, Zurfluh, 2008b].
  J'envisage par la suite l'élaboration des métriques par l'adaptation de techniques de fouilles et d'analyse des données dans le cadre fortement structuré de l'entrepôt de données multidimensionnelles. Cette approche d'OLAP Mining nécessite de redéfinir l'algèbre OLAP pour assurer la cohérence des techniques exploratoires et de la fouille ainsi intégrées. Ce projet servira de cadre pour des coopérations, avec notamment l'équipe SIG-EVI, qui possède une solide expertise en Text Mining (Tetralogie [Dousset, 2003]).

Tableau 15 : Synthèse de mes recherches depuis 2000.

Axes	Etudiants	Publications RI | CI RN | CN	Projets		
			Noms	**Partenaires**	**Rôle personnel**
Depuis 2000 - Axe 1 : Bases de données multidimensionnelles					
Modélisation en Constellation	4 stages Master Thèse F. Ghozzi Thèse E. Annoni	1 | 3 1 | 5	▪ 2000/2001 : REANIMATIC National (MENRT) ▪ 2001/2002 : GafoDonnées National (AS n°20/CNRS) ▪ 2009/2010 : EDM Laboratoire	▪ Cédric, LAMIH, LIPN, LIRMM, LISI, PRiSM OUTCOME-REA ▪ ERIC, LRI ▪ THALES	Participant Responsable IRIT Co-Responsable
Algèbre OLAP et Langages	6 stages Master Thèse F. Ghozzi	1 | 1 1 | 4			
Démarche	3 stages Master Thèse E. Annoni	| 3 2 | 3	▪ 2004/2007 : CIFRE ID-6 National (ANRT)	▪ ID-6	Co-Responsable
Depuis 2003 - Axe 2 : Intégration des documents					
Modélisation par Extensions et Galaxie	1 stage Master Thèse R. Tournier	1 | 2 1 | 2	▪ 2007/2009 : MADSI National (GT GdR I3)	▪ CRI, LIG…	Co-Responsable IRIT
Manipulation OLAP et Agrégations Textuelles	1 stage Master Thèse R. Tournier	| 2 1 | 1			
Depuis 2006 - Axe 3 : Prise en compte de l'usager					
Modélisation Quantitative de l'Usager et Annotations	2 stages Master	1 | 1 | 3	▪ 2007/2009 : IAPA Laboratoire (BQR/UPS)	▪ ICR, CS, IBM	Responsable WP6
Modélisation Qualitative des Usagers et Recommandations	1 stage Master Thèse H. Jerbi	| 3 | 2			

3 Prototypes

Au-delà des projets et partenariats, nous réalisons des développements au sein du laboratoire afin de valider nos recherches. Ce contexte applicatif nous permet également d'offrir un cadre d'étude aux étudiants en Master et en thèse.

Nous déclinons dans la suite trois prototypes associés à chaque axe de mes recherches :

- GRAPHIC-OLAP pour l'axe 1,
- XML-GOLAP pour l'axe 2
- PERSONAL-GOLAP pour l'axe 3.

3.1 GRAPHIC-OLAP

Depuis 2001, nous développons l'outil GRAPHIC-OLAP permettant de définir, de manipuler et d'interroger des bases de données multidimensionnelles. Il repose sur un langage assertionnel OLAP-SQL et sur un langage graphique. Les détails d'implantation sont présentés dans [Annoni, 2003] et [Tournier, 2004].

3.1.1 Architecture

L'architecture de GRAPHIC-OLAP décrite à la figure suivante comporte trois niveaux.

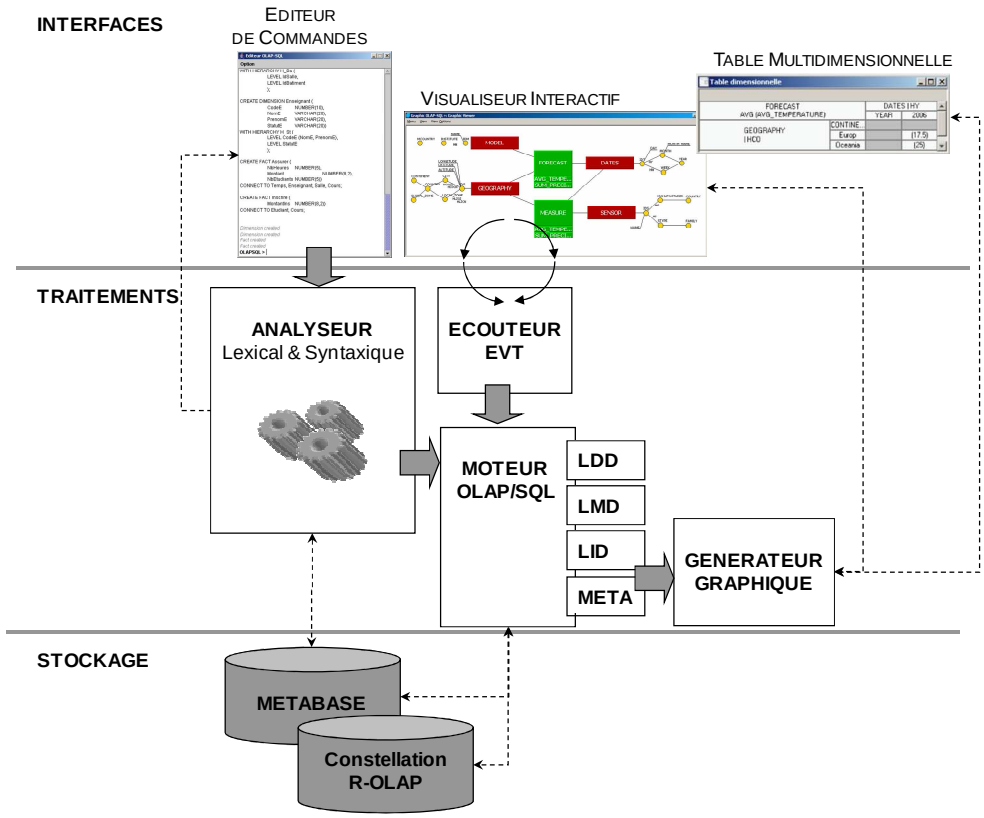


Figure 43 : Architecture de GRAPHIC-OLAP.

- Le niveau « INTERFACES » offre un ensemble d'interfaces permettant à l'utilisateur d'interagir avec la base de données multidimensionnelle. Le « visualiseur interactif » affiche le schéma en constellation de la base de données

multidimensionnelle. L'éditeur de commandes OLAP-SQL permet de saisir textuellement des ordres assertionnels et d'obtenir des réponses (messages d'erreurs,…). Le résultat d'une commande d'interrogation permet d'obtenir une table multidimensionnelle contenant les données de l'analyse souhaitée.

- Le niveau « TRAITEMENTS » regroupe quatre composants. Chaque commande OLAP-SQL de définition (LDD), de manipulation (LMD) et d'interrogation (LID) est analysée lexicalement et syntaxiquement afin d'être validée. La commande valide (décomposée dans une représentation interne pseudo-algébrique) est prise en charge par le moteur OLAP/SQL chargé des transcriptions vers le SGBD relationnel de stockage. Les interactions graphiques de l'utilisateur sont également captées par l'écouteur d'événements qui peut communiquer au moteur OLAP/SQL la traduction interne des manipulations graphiques effectuées. Les ordres d'interrogations sont calculés et leurs résultats sont mis en forme par le générateur graphique dans une table multidimensionnelle.
- Le niveau « STOCKAGE » regroupe deux bases de données. La métabase décrit les structures multidimensionnelles de la constellation tandis que les données des faits et des dimensions sont stockées dans la base R-OLAP.

3.1.2 *Langage assertionnel*

L'outil GRAPHIC-OLAP offre à l'utilisateur un langage de commandes assertionnelles qu'il saisit textuellement au travers de l'éditeur de commande. L'analyseur lexical et syntaxique valide la commande, puis retourne un message à l'utilisateur. Dans le cas d'une interrogation, le logiciel affiche en plus une table multidimensionnelle graphique. Le fonctionnement de l'outil est détaillé dans [Annoni, 2003] tandis que les primitives du langage assertionnel OLAP-SQL sont définies dans [Ravat, Teste, Zurfluh, 2002].

La Figure 44 illustre le fonctionnement de l'éditeur de commande :

- une commande invalide provoquant un message d'erreur indiquant que la dimension est inconnue,
- une commande valide aboutissant à l'affichage d'une table multidimensionnelle.

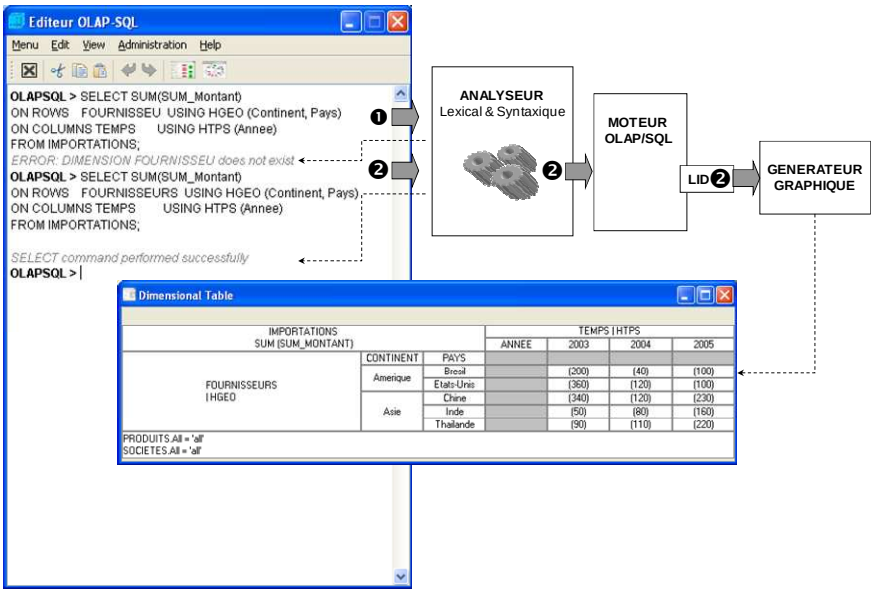


Figure 44 : Fonctionnement de GRAPHIC-OLAP.

3.1.3 *Langage graphique*

L'outil GRAPHIC-OLAP offre une vision conceptuelle de la BDM au travers de diverses interfaces :

- arborescente (représentation classique),
- graphique (représentation conforme aux formalismes de notre modèle en constellation), et
- hyperbolique (représentation dédiée aux grandes constellations).

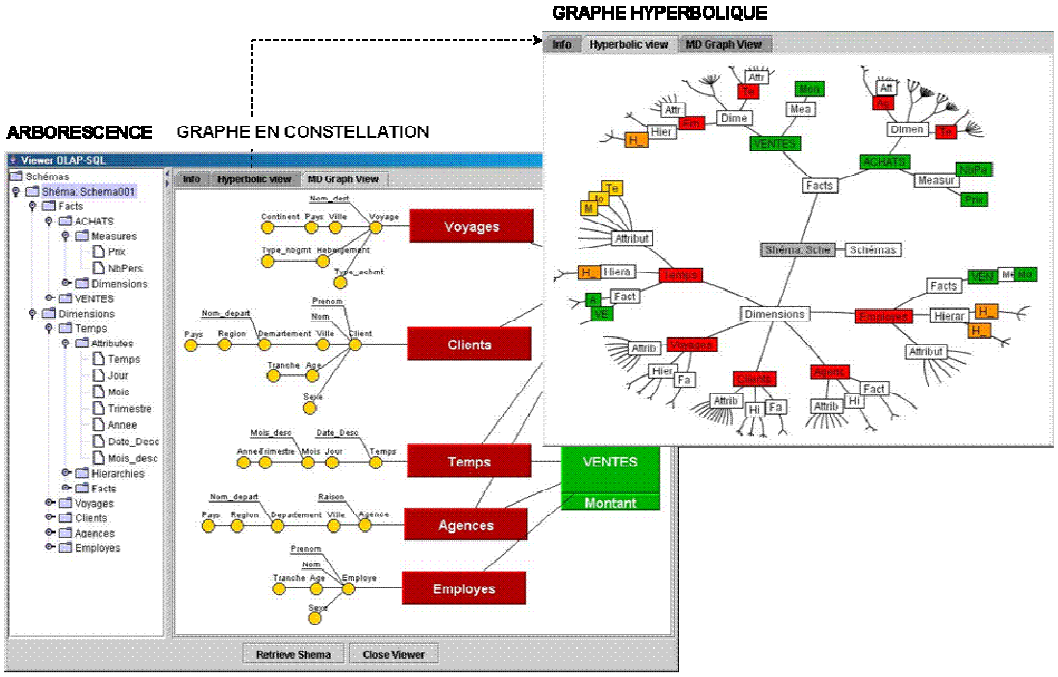


Figure 45 : Interfaces de visualisation dans GRAPHIC-OLAP.

Le fonctionnement de la partie graphique de l'outil GRAPHIC-OLAP est détaillé dans [Tournier, 2004]. L'interface de visualisation en graphe de la constellation sert de support à un langage d'interrogation graphique permettant à l'utilisateur d'élaborer par manipulation directe du graphe. Les primitives du langage graphique sont définies dans [Ravat, Teste, Tournier, Zurfluh, 2007b] et [Ravat, Teste, Tournier, Zurfluh, 2008b].

La figure suivante montre les manipulations graphiques permettant de construire une table multidimensionnelle. Ces manipulations correspondent à l'opération DISPLAY définie dans notre algèbre.

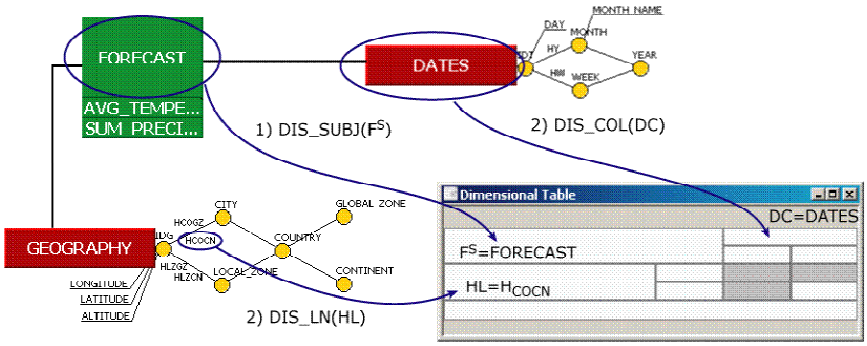


Figure 46 : Construction graphique d'une table multidimensionnelle.

La figure qui suit présente l'enchaînement de trois manipulations graphiques équivalentes aux opérations algébriques SELECT, DRILLDOWN et DRILLDOWN.

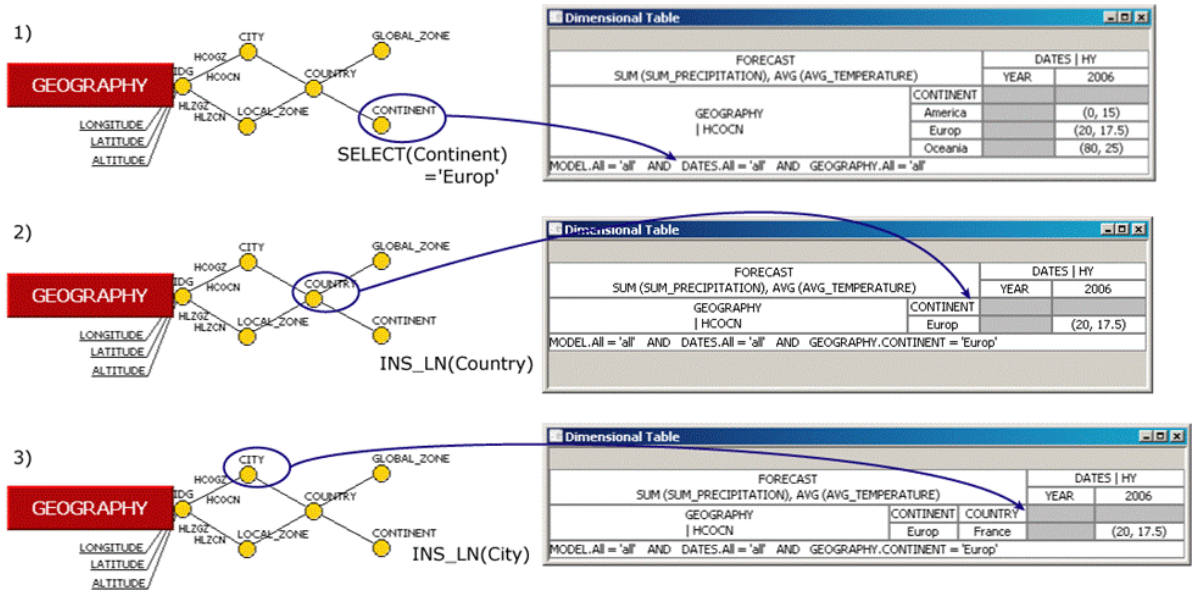


Figure 47 : Exemple d'opérations graphiques dans GRAPHIC-OLAP.

3.1.4 *Etudes expérimentales de l'opérateur BLEND*

Au-delà de l'étude conceptuelle de l'opérateur algébrique BLEND, nous avons testé notre proposition en menant une série d'expériences [Hubert, Teste, 2009]. Notre objectif était non seulement de montrer la faisabilité de l'opération en environnement R-OLAP, mais également d'analyser le coût que représente cette opération.

Collection : Nous avons utilisé une constellation R-OLAP comportant trois tables de dimensions et une table de fait. Pour simplifier l'expérience, nous n'avons utilisé aucune vue matérialisée (Gupta, *et al.*, 1995) pour optimiser la BDM.

DATES(id_dates, moisn, moisl, trimestre, annee, quadriennal)

ORGANISMES(id_organismes, variete, categorie, typeorganisme)

GEOGRAPHIES(id_geographies, parcelle, etat, region, pays, densite, continent)

REPARTITION(id_repartition, id_dates#,id_organismes#,id_geographies#, superficie)

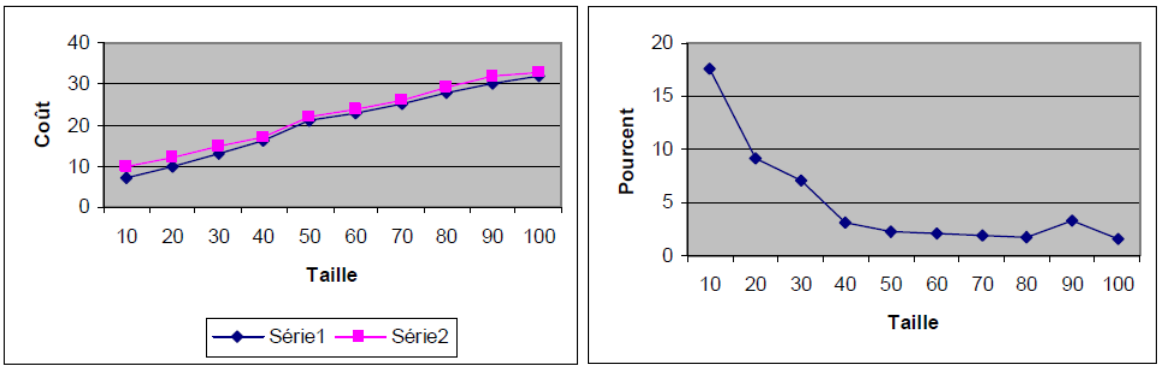
Nous avons généré des enregistrements dans les relations de notre BDM, complétées par des index multiples sur les clés étrangères : |ORGANISMES| = 250, 10 ≤ |GEOGRAPHIES| ≤ 100, |REPARTITION| = |ORGANISMES| x |GEOGRAPHIES|. Les valeurs des enregistrements ont été générées aléatoirement en veillant à ce que les deux sous-ensembles E_{sup} et E_{inf} (*cf.* chapitre 3) soient de taille homogène.

Protocole : La comparaison a consisté à observer le coût théorique du calcul de deux requêtes :

- une requête utilisant un attribut stockant le résultat de la transformation multigraduelle simulant l'utilisation d'une BDM modélisée en fonction du besoin équivalent à la transformation multigraduelle (série 1),
- une requête effectuant dynamiquement la transformation multigraduelle simulant l'utilisation de notre opérateur BLEND (série 2).

Ces deux requêtes ont été appliquées sur la base R-OLAP en faisant varier le nombre de n-uplets dans GEOGRAPHIES (10 à 100) et REPARTITION (250x10 à 250x100). La taille de la relation ORGANISMES est maintenue constante à 250 n-uplets.

Résultats : Les valeurs du coût correspondent au coût théorique calculé par le SGBD (coût fourni par le « *explain plan* » d'Oracle Server Application 11g) puisque notre objectif est d'étudier le surcoût provoqué par un calcul du BLEND. La Figure 48(a) compare les deux requêtes : le calcul dynamique (série 2) est évidemment plus coûteux que l'utilisation de la transformation préalablement stockée (série 1), cependant, ce coût apparaît faible (entre 18% et 2%). Ce résultat est d'autant plus encourageant que nous montrons à la Figure 48(b) que ce surcoût tend à diminuer avec l'accroissement du nombre d'enregistrements (taille) dans la relation transformée.



(a) (b)

Figure 48 : Expérimentations sur le coût du BLEND.

3.2 XML-GOLAP

Afin de valider nos propositions relevant de l'axe 2, nous avons développé le prototype XML-GOLAP [Abdelhedi, 2009] permettant le stockage et la manipulation de documents XML au sein d'une base de données multidimensionnelles.

3.2.1 Architecture

Le prototype repose sur une architecture présentée dans la figure suivante.

- Le niveau « INTERFACES » offre un éditeur de commandes OLAP-SQL permettant de saisir des requêtes d'interrogation via un ordre OLAP-SQL intégrant potentiellement les fonctions d'agrégation textuelle. Le résultat d'une commande d'interrogation permet d'obtenir une table multidimensionnelle étendue aux mesures textuelles contenant les données de l'analyse souhaitée.
- Le niveau « TRAITEMENTS » regroupe quatre composants. Chaque interrogation OLAP-SQL (LID) est analysée par l'analyseur lexical et syntaxique, puis est prise en charge par le moteur OLAP/SQL effectuant les transcriptions vers le SGBD relationnel de stockage. Les agrégations textuelles sont gérées par un module additionnel appelé « agrégateur textuel ». Ce dernier effectue des interactions entre la galaxie R-OLAP, les documents entreposés et l'ontologie de domaine stockée sous forme de fichiers XML autonomes. Une fois l'ordre d'interrogation valide calculé, le résultat est mis en forme par le générateur graphique dans une table multidimensionnelle étendue aux mesures textuelles.

– Le niveau « STOCKAGE » regroupe deux bases de données. La métabase regroupe la description des structures multidimensionnelles de la galaxie tandis que les données sont stockées dans la base R-OLAP contenant les documents XML. En plus, l'ontologie utilisée pour l'agrégation de texte est stockée sous forme XML dans le système de gestion de fichiers.

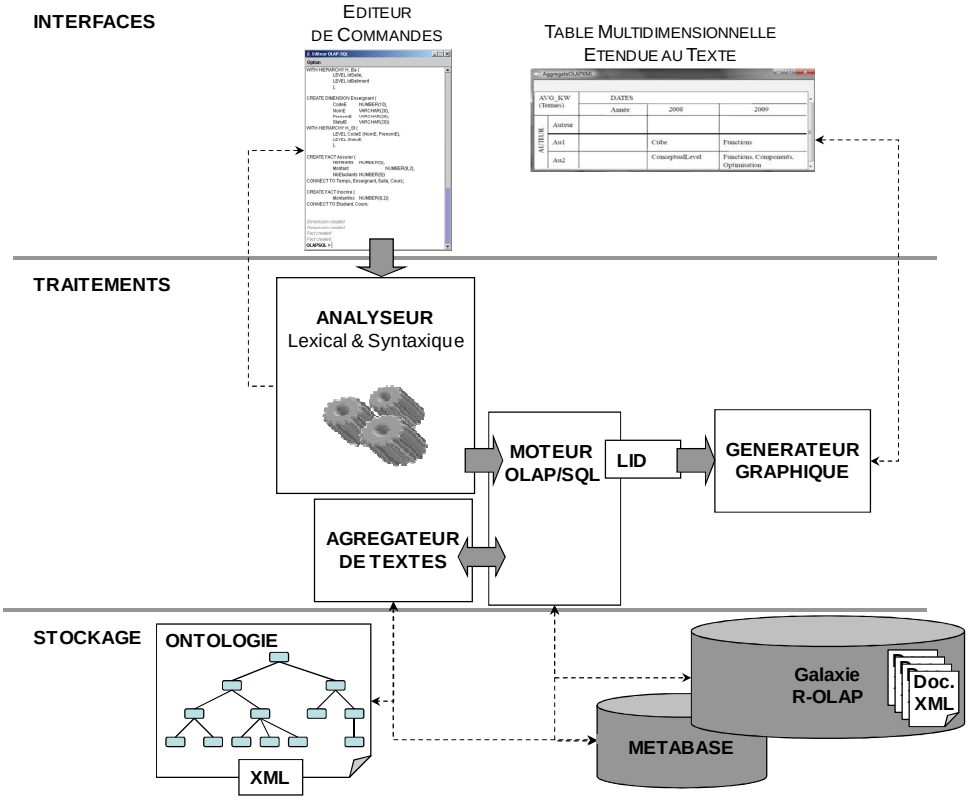


Figure 49 : Architecture de XML-GOLAP.

3.2.2 *Expérimentations sur l'agrégation AVG_KW*

Le prototype XML-GOLAP nous sert à expérimenter la faisabilité de nos propositions sur l'intégration des documents dans les BDM, ainsi qu'à évaluer le coût des agrégations textuelles que nous développons.

Nous avons effectué une première expérience visant à montrer la faisabilité de l'agrégation textuelle AVG_KW.

Collection : Nous avons utilisé une très petite collection de 4 documents (extraits des références du mémoire de thèse de Ronan Tournier [Tournier, 2007]) du même domaine, mais de structures très hétérogènes : il s'agit de documents *pdf* issus des éditeurs Springer et Elsevier que nous avons convertis en fichiers XML.

Protocole : Nous avons appliqué notre fonction d'agrégation sur cette collection réduite à 4 documents.

Résultats : Les documents comportaient initialement environ 13000 termes. A l'issue du processus d'agrégation le nombre de termes produits est situé entre 5 et 9 termes par document. Ces résultats montrent que notre fonction permet une réduction significative du nombre de termes, rendant envisageables des analyses et des interprétations humaines.

La seconde série d'expériences que nous avons réalisée s'est attachée à étudier le coût du calcul de l'agrégation.

Collection : Nous avons utilisé une collection de documents XML issus des bases des campagnes de tests INEX ; la collection utilisée regroupe des articles scientifiques de la revue TKDE (*Transactions on Knowledge and Data Engineering*) de l'IEEE entre 1995 et 2004, représentant environ 850 documents formant une collection de 75MB.

Protocole : Nous avons appliqué notre fonction d'agrégation en faisant varier la taille de la collection à 1, 10, 68, 100, 175, 451 et 843 documents comme le montre la **Figure** 50.

Résultats : Les documents comportaient initialement environ 200000 termes. A l'issue des différents lancements, les documents ont été agrégés au plus en 9 termes, et généralement en 5 ou 6 termes confirmant les premiers résultats que nous avions obtenus. La Figure 50 donne les coûts de calcul de l'agrégation qui sont compris entre 800ms et 400s. Nous distinguons deux types de coût :

- Le coût « off-line » qui correspond au coût des processus d'intégration des documents et de préparation (pré-calculs indépendants de la requête) de la couche « STOCKAGE ».
- Le coût « on-line » qui correspond réellement au calcul réalisé dynamiquement par le module « agrégateur textuel ».

On observe que le coût « off-line » suit une progression importante en fonction du nombre de documents tandis que le coût « on-line » reste maitrisé entre 20ms et 1,75s. Ce résultat est là encore encourageant puisqu'il garantit un temps de calcul réel de l'agrégation textuelle ne dépassant pas 2 secondes. Il faut en plus noter que nous n'avons pas utilisé de techniques de pré-calculs (treillis de vues matérialisées) ; ceci est l'une des perspectives qui nous semble prometteuse pour nos futurs travaux de recherche.

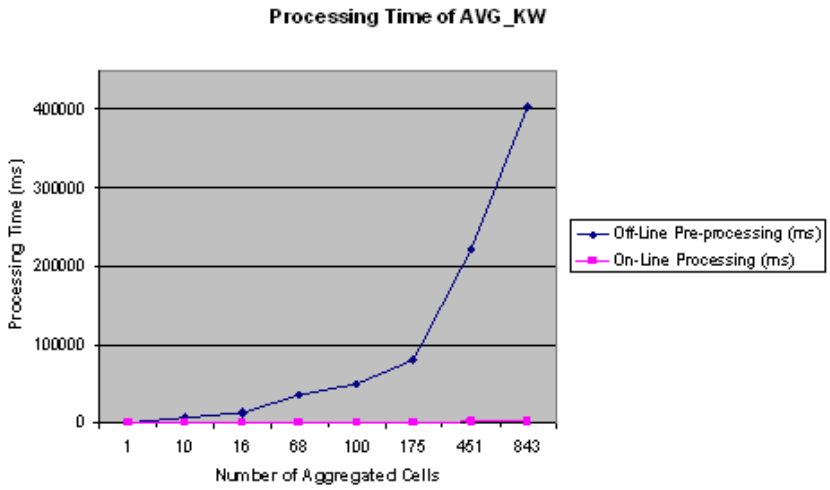


Figure 50 : Coûts de l'agrégation AVG_KW.

Il faut relever également que ces résultats ont été obtenus en utilisant une ontologie de « petite » taille. L'ontologie que nous avons développée est tirée d'une classification des articles de l'ACM. Notre ontologie a une profondeur maximale de 7 nœuds, elle comporte plus de 150 termes relatifs au domaine des systèmes d'information et les termes composés ont une largeur maximale de 3 sous-termes (par exemple « geographic information system »). Elle repose également sur une centaine de synonymes additionnels ainsi que

différentes formes pour chaque terme. Dans les ontologies généralistes telles que WorldNet8 [Baziz, *et al*., 2005], la distance maximale d'agrégation (D_{MAX}) est fixée généralement entre 3 et 5 (3 pour WorldNet8). Pour nos expériences, nous avons fixé arbitrairement la distance maximale d'agrégation à 2 ($D_{MAX} = 2$) compte tenu de notre domaine d'étude spécialisé et de la relative faible profondeur de notre ontologie.

3.3 PERSONAL-GOLAP

Afin de valider nos propositions relevant de l'axe 3, nous avons développé le prototype Personal-GOLAP [Jerbi, 2007] [Ghalamallah, 2008] [Atigui, 2009] permettant de personnaliser une base de données multidimensionnelles en intégrant à la fois l'expertise décisionnelle et les préférences de l'usager.

3.3.1 Architecture

A l'image des différents développements que nous réalisons, ce prototype repose sur une architecture modulaire présentée dans la figure suivante.

- Le niveau « INTERFACES » offre l'éditeur de commandes pour saisir sous la forme de règles les préférences d'un utilisateur. Ces règles peuvent être définies soit de manière quantitative par des poids affectés aux propriétés de la constellation, soit de manière qualitative en ordonnant les propriétés en fonction de la priorité accordée par l'utilisateur à ces dernières. L'éditeur de commande permet également de saisir des requêtes d'interrogation via un ordre OLAP-SQL classique (conforme à la grammaire du langage assertionnel OLAP-SQL). Parallèlement, l'éditeur d'annotation offre la possibilité de créer, éditer, supprimer et modifier des annotations ancrées dans la constellation. Enfin, les tables multidimensionnelles produites en réponses des requêtes d'interrogation comportent les annotations ancrées aux propriétés de la table. Le système complète la table multidimensionnelle en recommandant des tables personnalisées.
- Le niveau « TRAITEMENTS » regroupe cinq composants. Chaque commande de définition de règles (LDR) ou d'interrogation (LID) est analysée par l'analyseur lexical et syntaxique. La commande valide est prise en charge par le moteur OLAP/SQL chargé des transcriptions vers le SGBD relationnel de stockage : les règles validées de personnalisation sont stockées dans la base de données relationnelles. Le module de mémoire d'expertise décisionnelle, appelé « Moteur MED » gère les différentes annotations exprimées par l'utilisateur. Les ordres d'interrogations sont calculés et le résultat est mis en forme par le générateur graphique dans une table multidimensionnelle intégrant les annotations associées au contexte. Le générateur graphique reçoit des recommandations sous forme de tables multidimensionnelles personnalisées venant s'ajouter à la table répondant à l'interrogation initiale.
- Le niveau « STOCKAGE » regroupe deux bases de données. La métabase regroupe la description des structures multidimensionnelles de la constellation ainsi que les préférences contextuelles exprimées par l'utilisateur. La base R-OLAP contient les données décisionnelles de la constellation enrichies d'annotations de l'utilisateur.

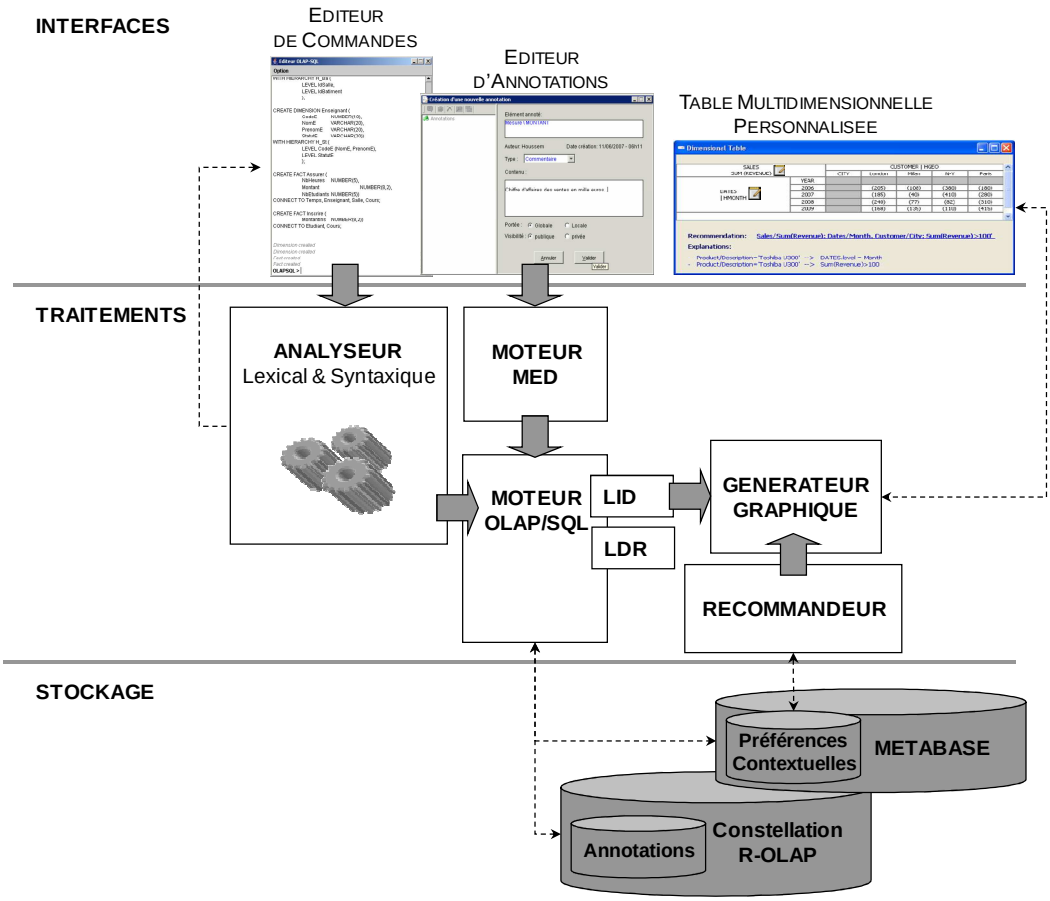


Figure 51 : Architecture de Personal-GOLAP.

3.3.2 Langage de définition de préférences

Les préférences de l'utilisateur sont acquises par le système au travers d'un langage de règles. Actuellement, Personal-GOLAP prend en compte les préférences individuellement pour chaque utilisateur sans gérer les aspects collaboratifs. Les règles permettent donc à un utilisateur de personnaliser la constellation en exprimant de manière quantitative [Ghalamallah, 2008] ou qualitative [Atigui, 2009] ses préférences.

Personnalisation quantitative

L'ordre définissant quantitativement les préférences est présenté dans [Ravat, Teste, Zurfluh, 2007g] [Ravat, Teste, 2008g] et les détails de sa mise en œuvre se trouvent dans [Ghalamallah, 2008]. La figure suivante présente un exemple de règle définissant les préférences d'affichage des paramètres MOIS et ANNEE sur une dimension DATES. Une fois cette personnalisation mise en place, le système OLAP affiche à l'utilisateur des tables multidimensionnelles personnalisées en fonction du contexte d'interrogation (DISPLAY) et d'un seuil (fixé dans l'exemple à 0.5) déterminant les règles utilisées. La table multidimensionnelle ainsi obtenue est personnalisée en affichant les paramètres préférés ANNEE et MOIS (au lieu du simple paramètre extrémité ANNEE dans un fonctionnement classique).

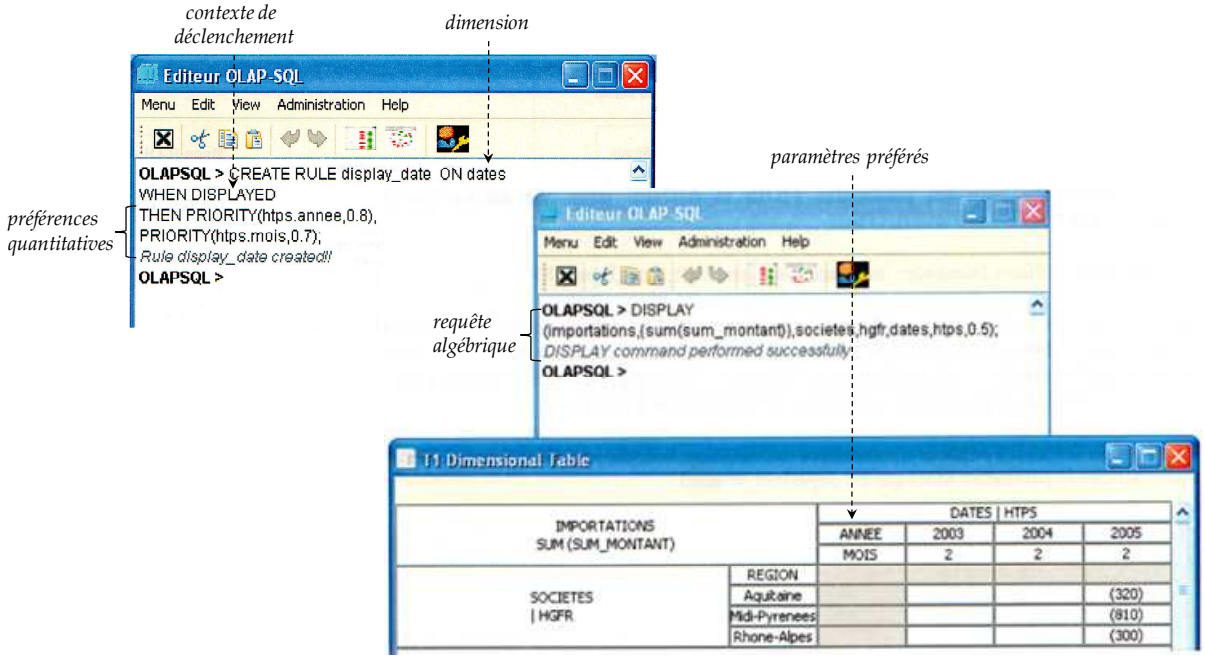


Figure 52 : Exemple de personnalisation quantitative.

Personnalisation qualitative

Pour définir une préférence qualitativement, nous proposons une règle de définition dont les détails de la mise en œuvre sont exposés dans [Atigui, 2009]. La figure suivante montre la définition des deux règles qualitatives : une définissant un ordre d'affichage sur le domaine des valeurs du paramètre ANNEE, et l'autre définissant la valeur préférée TOULOUSE sur un paramètre VILLE. L'affichage des montants des ventes en fonction des années et des villes est alors personnalisé en fonction des préférences du décideur.

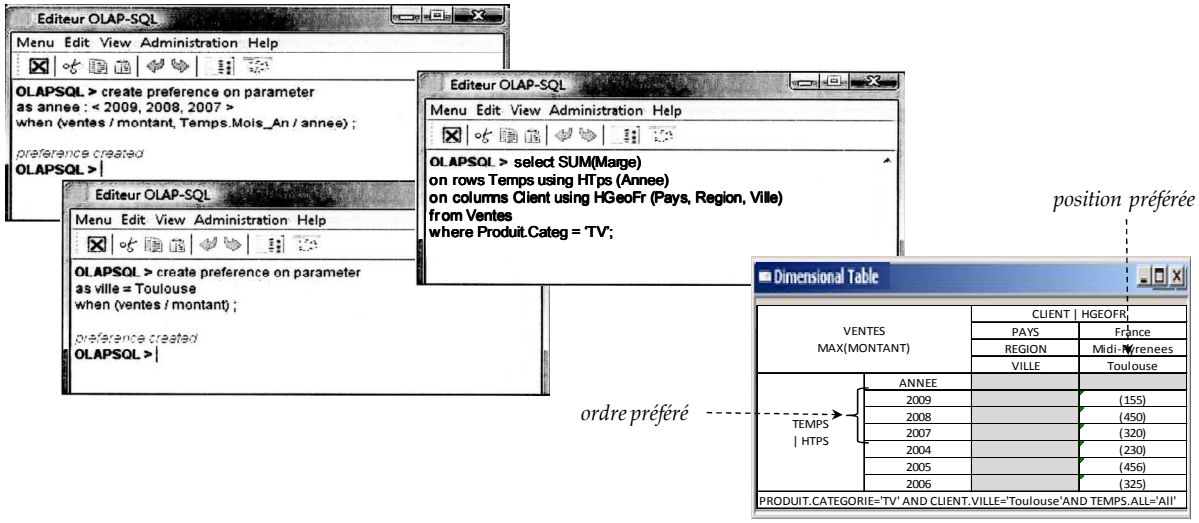


Figure 53 : Exemple de personnalisation qualitative.

3.3.3 *Interface de gestion des annotations*

Les détails d'implantation de l'éditeur d'annotations sont décrits dans [Jerbi, 2007] permettant au système OLAP de conserver la mémoire d'expertise décisionnelle (MED) des utilisateurs [Cabanac, Chevalier, Ravat, Teste, 2009].

La figure suivante présente une copie écran de l'éditeur d'annotation. L'arbre d'édition permet de visualiser les annotations ancrées dans la constellation : chaque type d'annotation (Commentaire, Réponse,…) est distingué par une icône spécifique et les annotations sont organisées afin de faire apparaître les fils de discussion. Lorsqu'une annotation est éditée, les informations objectives (ancre, date de création…) et subjectives (contenu, type…) sont visualisées (*cf.* chapitre 5).

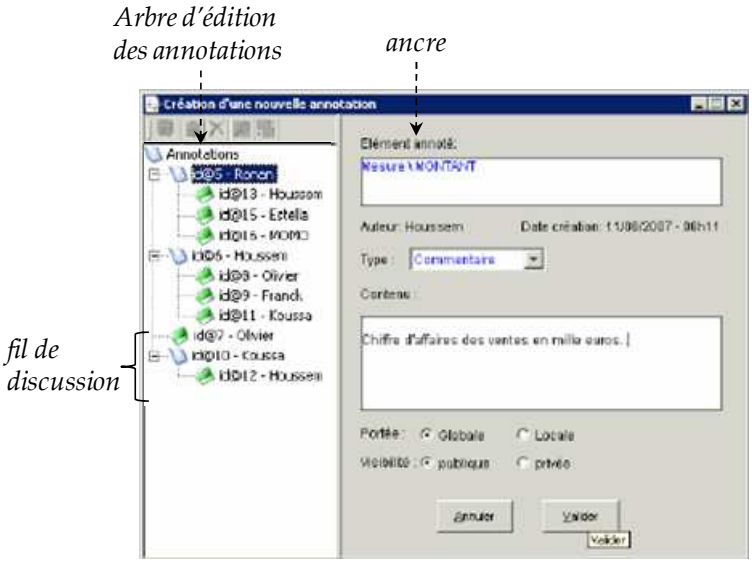


Figure 54 : Exemple d'annotations.

Chapitre 7 - Conclusion et perspectives

1 Synthèse

Mes travaux de recherches ont consisté à proposer des solutions en matière de modélisation et de manipulation des données au sein des systèmes OLAP. Ils s'articulent selon trois axes principaux :

- l'axe 1 traite de la modélisation des bases de données multidimensionnelles et de la manipulation OLAP,
- l'axe 2 s'attache à intégrer des documents dans les systèmes OLAP,
- l'axe 3 vise à personnaliser le système OLAP par la prise en compte de l'expertise décisionnelle et des préférences de l'usager.

Axe 1 : Modélisation et manipulation multidimensionnelle des systèmes OLAP.

Les travaux que nous avons réalisés visent à définir des modèles de représentation des données décisionnelles ainsi que des langages de manipulation supportant efficacement les analyses.

Ces travaux ont abouti à la définition d'un modèle conceptuel spécialisé dans la représentation des données multidimensionnelles [Ravat, Teste, Zurfluh, 2001a] [Teste, 2001]. Associé aux concepts, nous avons proposé un formalisme graphique [Ravat, Teste, Tournier, Zurfluh, 2008b] permettant une description simple du schéma en constellation d'une BDM. L'objectif a été de constituer dès 2001 une représentation uniforme et complète des différents concepts partiellement décrits dans les propositions qui existaient en distinguant clairement les niveaux d'abstraction. Nous avons étendu ce modèle par diverses propositions : enrichissement par des contraintes sémantiques [Ghozzi, 2004], démarche de conception [Annoni, 2007] par capitalisation, intégration de versions [Ravat, Teste, 2006d] pour gérer les évolutions temporelles.

Nos travaux sur la modélisation ont servi de socle à l'élaboration d'une algèbre OLAP [Teste, 2001]. Ces travaux ont abouti à la définition d'un noyau algébrique minimum fermé d'opérateurs élémentaires assurant la couverture du modèle multidimensionnel [Ravat, Teste, Tournier, Zurfluh, 2008b]. Ce fondement théorique a servi de support pour différentes contributions :

- définition d'opérateurs étendus par composition des opérateurs élémentaires du noyau algébrique simplifiant l'interrogation des données multidimensionnelles [Ravat, Teste, Zurfluh, 2005],
- proposition d'un nouvel opérateur [Hubert, Teste, 2009] d'analyses multigraduelles permettant de relaxer la modélisation multidimensionnelle (réorganisation des paramètres dans les hiérarchies),
- définition du langage assertionnel OLAPSQL [Ravat, Teste, Zurfluh, 2002] supportant toutes les fonctionnalités nécessaires aux décideurs : définition, manipulation et interrogation une BDM,
- définition d'un langage graphique d'interrogation complet au regard de notre algèbre OLAP [Ravat, Teste, Tournier, Zurfluh, 2007b] reposant sur des interactions directes avec la constellation.

Le prototype GRAPHIC-OLAP que nous développons au sein de l'équipe nous sert de plateforme pour expérimenter nos propositions : modèle en constellation reposant sur un formalisme graphique, langage assertionnel OLAPSQL, langage graphique reposant sur l'algèbre OLAP, opérateur « BLEND » d'analyse multigraduelle.

Axe 2 : Intégration de documents dans les systèmes OLAP.

Les travaux que nous avons réalisés pour intégrer les documents au sein des bases de données multidimensionnelles avaient pour objectif de rendre possible l'analyse OLAP sur des documents. Le résultat de nos travaux a été de rendre possible non seulement des analyses quantitatives sur le contenu numérique des documents mais également des analyses plus qualitatives sur le contenu textuel des documents.

Ces travaux ont abouti à la redéfinition de modèles de représentation des documents dans un espace multidimensionnel [Tournier, 2007]. La principale contribution de nos travaux sur la modélisation est la définition du modèle en galaxie [Ravat, Teste, Tournier, Zurfluh, 2008a, 2008f, 2007a, 2007f]. Ce modèle en galaxie repose sur plusieurs idées originales :

- L'unicité du mécanisme de description des données analysées décrivant de manière symétrique les sujets et les axes de l'analyse. Cette flexibilité simplifie la définition de la BDM pour le concepteur ;
- Le support de mesures textuelles permettant de faire porter les analyses non seulement sur les données numériques mais également sur les données textuelles ;
- L'intégration de liens navigationnels sur les données pouvant servir à analyser les relations entre les documents.

Nos travaux ont montré la nécessité de généraliser l'algèbre OLAP à la galaxie. La contribution la plus remarquable de ces travaux concerne le développement d'une nouvelle approche pour l'agrégation de données textuelles [Pujolle, Ravat, Teste, Tournier, 2008]. Ces travaux ont permis de développer deux fonctions d'agrégation :

- *TOP_KW*$_k$ [Ravat, Teste, Tournier, Zurfluh, 2008c, 2008d] exploite la fonction de pondération *tf.idf* issue de la recherche d'information et permet d'agréger les valeurs des mesures textuelles brutes lors de l'analyse OLAP ;
- *AVG_KW* [Ravat, Teste, Tournier, 2007d] repose sur une ontologie légère de domaine et s'attache à rendre possible l'agrégation de mesures textuelles élaborées (données extraites des documents telles que les mots-clefs).

Nous avons expérimenté nos propositions au sein du prototype XML-GOLAP. Nous avons notamment constitué une BDM intégrant des documents conformément à nos principes de modélisation. Cette base a servi de support pour expérimenter la fonction *AVG_KW* prouvant la faisabilité des agrégations textuelles lors des manipulations OLAP.

Axe 3 : Personnalisation des systèmes OLAP.

L'objectif de nos travaux sur la personnalisation dans les systèmes OLAP [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a] est double. Il s'agit dans un premier temps de rendre disponible et accessible toute information ayant permis d'aboutir à une décision. Dans un deuxième temps, nous souhaitons mieux prendre en compte les préférences de l'usager en termes de données.

Nos travaux ont abouti à un premier résultat reposant sur le concept de *mémoire d'expertise* [Cabanac, Chevalier, Ravat, Teste, 2009] afin de conserver le patrimoine immatériel des décideurs au sein du système OLAP. En effet, l'information utile lors du processus d'analyse décisionnelle ne se trouve pas uniquement dans les bases de données multidimensionnelles, mais une partie importante est habituellement immatérielle : il s'agit de « l'expertise » du décideur. Nous avons proposé de modéliser sous la forme d'annotations [Cabanac, Chevalier, Ravat, Teste, 2007, 2006a, 2006b] ancrées dans la base de données multidimensionnelles toutes ces informations immatérielles relevant de l'expertise de l'usager décideur (commentaires, discussions, prises de décision…).

Nous avons complété notre approche par la définition de modèles de préférence pour mieux représenter les besoins de l'usager en matière de données analysées. Nos travaux reposent sur deux approches complémentaires : l'approche quantitative et l'approche qualitative. L'approche quantitative [Ravat, Teste, 2008g] [Ravat, Teste, Zurfluh, 2007g] consiste à représenter l'intérêt pour l'usager d'une propriété de la constellation par une pondération. Nous avons proposé de définir les préférences de manière contextuelle par un mécanisme de type ECA. Cette approche de préférence contextuelle exprimée quantitativement est facilement utilisable par le système OLAP. L'approche qualitative [Jerbi, Ravat, Teste, Zurfluh, 2009a, 2008] représentant les préférences de l'usager par une relation d'ordre exprimée sur les données. Ces préférences sont alors simplement définies les unes par rapport aux autres. Nous avons représenté son contexte d'analyse pour déterminer durant l'analyse les préférences relevant de l'analyse en cours. Cette « contextualisation » des préférences permet lors des manipulations OLAP des recommandations contextuelles [Jerbi, Ravat, Teste, Zurfluh, 2009b, 2009c] qui assistent l'usager dans son exploration de l'espace multidimensionnel. L'assistance que nous proposons consiste à recommander à l'usager :

- des enrichissements de sa requête pour compléter le résultat qu'il cherche,
- des requêtes anticipées pour obtenir plus directement le résultat attendu, et
- des requêtes alternatives auxquelles il ne pense pas.

Nos recherches montrent que l'approche quantitative facilite les traitements mais rend l'acquisition des préférences difficile tandis que l'approche qualitative simplifie l'acquisition des préférences au détriment de traitements plus couteux. Nous expérimentons ces techniques de personnalisation des systèmes OLAP en développant le prototype PERSONAL-GOLAP. Ces développements ont permis de tester certaines propositions : annotations décisionnelles, personnalisation quantitative et qualitative.

Un panorama des contributions scientifiques qui nous semblent les plus significatives, est proposé dans la Figure 55 en fonction des trois axes suivis dans nos recherches.

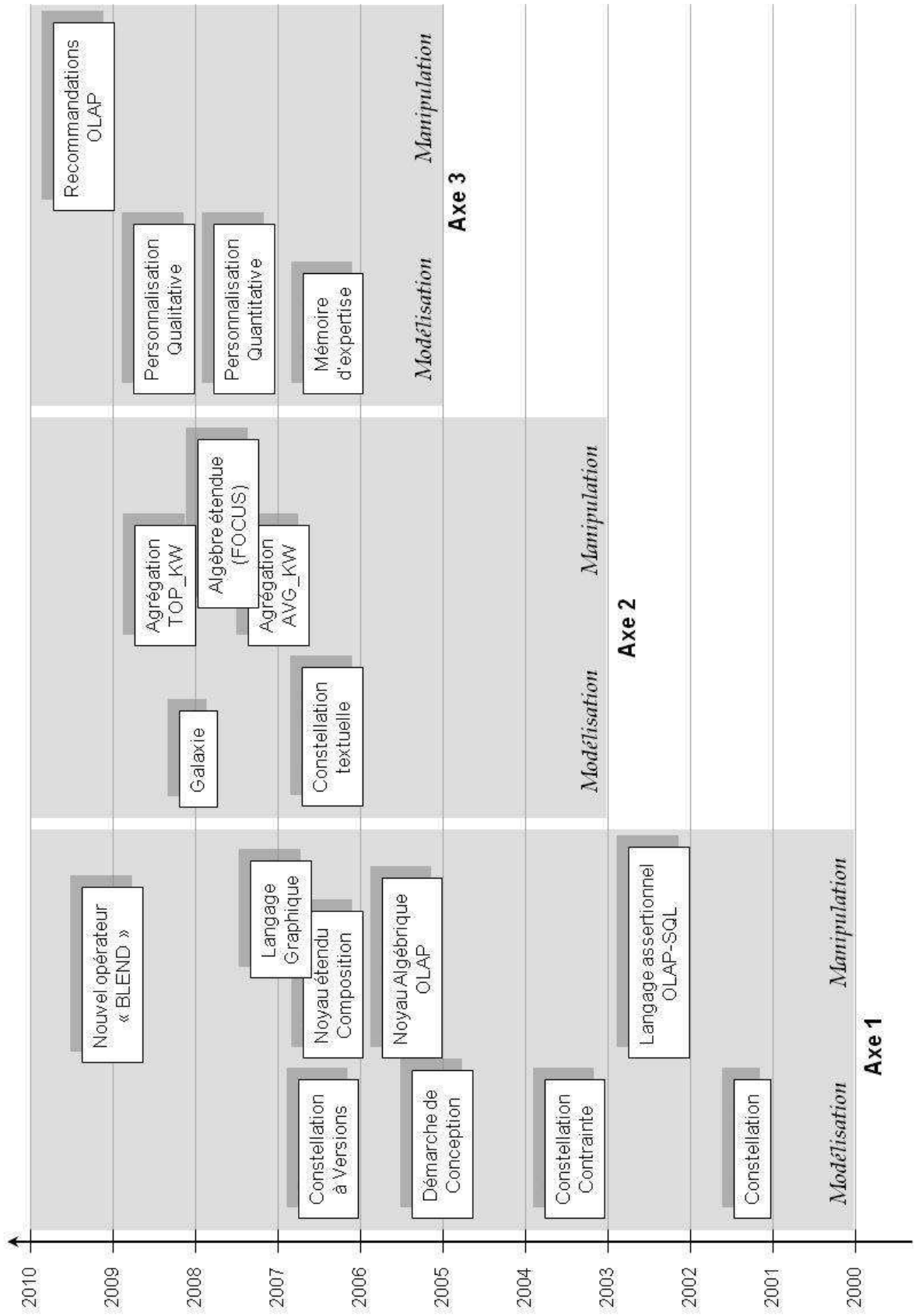


Figure 55 : Panorama chronologique des principaux résultats.

2 Perspectives

Les suites que j'envisage de donner à ces recherches sont nombreuses et s'orientent dans plusieurs directions.

Systèmes OLAP Mining. Les analyses multidimensionnelles supportées par les systèmes OLAP et la fouille de données (« data mining ») visent toutes les deux à analyser des données pour en améliorer la connaissance et alimenter généralement un processus décisionnel. Ces deux approches sont la plupart du temps menées de manière séparée. Le couplage des systèmes OLAP avec les techniques de fouilles et d'analyses de données [Dousset, 2003] reste une piste toujours ouverte afin de développer de véritables systèmes OLAP Mining. Cet objectif ouvre un champ de recherche à la fois au niveau de la modélisation et celui de la manipulation des données.

Une première perspective que j'envisage est la définition d'une algèbre OLAP couplant les opérateurs OLAP à de nouveaux opérateurs nécessaires aux manipulations de fouille de données. Cette étude devra permettre de définir un ensemble d'opérateurs complexes d'OLAP Mining mais également la mise au point de langages graphiques de manipulation destinés à des décideurs. A titre d'exemple, les travaux de [Sureau, *et al*., 2009] à EDA'09 ont montré comment des algorithmes de fouille de données pouvaient être utilisés pour réorganiser les données au sein de tables multidimensionnelles et améliorer la visualisation des données. Une autre voie de recherche de l'OLAP Mining concerne la modélisation multidimensionnelle. Il s'agit d'utiliser les techniques de fouilles pour proposer des structures multidimensionnelles cachées dans les données ; par exemple, les travaux de [Bentayeb, *et al*., 2009b] à DAWAK'09 proposent d'exploiter des techniques de fouilles de données pour faire émerger de nouvelles hiérarchies au sein de l'espace multidimensionnel. Depuis 2008, j'ai initié une collaboration avec l'équipe du laboratoire ERIC. Cette collaboration a permis de publier un article de prospective [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a]. Cette collaboration se poursuit actuellement.

Systèmes Web OLAP et Multimédia. La dernière révolution informatique est probablement le Web qui offre une masse importante de documents encore largement inexploitée par les systèmes OLAP. Les travaux sur l'intégration des documents dans les systèmes OLAP en sont encore à leurs débuts [Pérez, *et al*., 2008b], restant cantonnés à des analyses numériques sur des documents structurés ni variables, ni évolutifs.

Une perspective directe à nos travaux est l'intégration de documents semi-structurés [Sèdes, 1998], de documents multi-facettes [Djemal, *et al*., 2008] et enfin, la prise en compte de leur évolution au cours du temps. Elle nécessite de revoir la modélisation multidimensionnelle, par exemple par de nouvelles structures hiérarchiques [Malinowski, *et al*., 2006], ainsi que la manipulation OLAP afin de permettre les analyses OLAP sur ces documents complexes à structures hétérogènes [Mbarki, *et al*., 2007]. Les documents XML ouvrent en effet des perspectives de recherche sur l'étude des opérateurs OLAP [Wiwatwattana, *et al*., 2007] et la définition d'algèbres XML OLAP [Hachicha, et al., 2008]. Cette année nous avons débuté une collaboration avec l'équipe SIG-EVI de l'IRIT afin de développer de nouvelles opérations OLAP [Hubert, Teste, 2009]. Nous envisageons à terme d'appliquer les techniques de navigation multidimensionnelle sur des données XML hétérogènes, semi-structurées, multi-facettes voire évolutives telles que les résultats de recherche d'information. A plus long terme, ces travaux doivent permettre une nouvelle approche dans l'usage du Web que nous qualifions d'approche par inversion : le cadre multidimensionnel peut permettre cette inversion où l'utilisateur ne demanderait plus ce

qu'il souhaite, mais le Web donnerait ce qu'il contient par de nouveaux moyens OLAP d'explorer cette masse de données.

Le contexte Web oblige également à prendre en compte les aspects multimédias des documents. Les données de type image, vidéo et son restent mal exploitées par les systèmes OLAP. Ces données sont massivement disponibles sur le Web, mais aussi, dans des domaines comme par exemple le domaine médical [Bentayeb, *et al*., 2009c] avec le dossier patient par exemple. Un objectif ambitieux est d'élaborer un système OLAP multimédia capable d'analyser des données classiques mais également des données atypiques multimédias. Cet objectif pose de nombreux problèmes tant au niveau de la modélisation que celui de la manipulation des données. Ces travaux s'inscrivent dans le long terme, puisqu'il est nécessaire solutionner les problématiques liées à l'application des techniques OLAP sur chaque média : image, son… Un tel système devrait permettre par exemple d'effectuer des rotations ou des forages sur des images. Ensuite seulement, il pourra être possible d'envisager le développement d'opérations OLAP véritablement multimédia faisant intervenir simultanément plusieurs médias à l'image des propositions de systèmes Spatiaux-Temporels OLAP [Bimonte, *et al*., 2008] [Vaisman, *et al*., 2009] qui intègrent les approches Spatial-OLAP [Rivest, *et al*., 2001] et Temporal-OLAP [Mendelzon, *et al*., 2000].

Usagers et systèmes décisionnels. La personnalisation des systèmes OLAP reste un champ de recherche émergeant dans le domaine OLAP [Golfarelli, *et al*., 2009]. Les travaux que nous menons s'inscrivent dans cette perspective.

Une extension directe de nos travaux est la prise en compte de l'aspect collaboratif qui est largement exploité en RI tandis que dans les systèmes OLAP cette approche reste peu développée [Morfonios, et *al*., 2008]. L'apport du web social peut présenter un réel intérêt dans une perspective de personnalisation sociale des bases de données multidimensionnelles. Ainsi, la définition d'un système OLAP proposant des recommandations pourrait servir à aider un décideur en tirant bénéfice de l'expertise des autres membres du groupe d'utilisateurs.

A l'heure actuelle, il n'existe pas non plus de réelle solution permettant la mise en place d'un système OLAP adaptatif [Rizzi, 2007] [Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a]. Une difficulté concerne l'acquisition automatique des préférences des décideurs. Le principal verrou scientifique qu'il est nécessaire de solutionner, concerne la mise en oeuvre d'une mesure de similarité pour comparer et aligner les profils des utilisateurs ainsi que les contextes d'analyse lors des navigations dans l'espace multidimensionnel. Notre collaboration avec l'équipe SIG-D2S2 a abouti à un premier résultat définissant la notion de mémoire d'expertise à base d'annotations [Cabanac, Chevalier, Ravat, Teste, 2009]. La définition d'une mesure de similarité pourra être élargie de manière à aligner également les annotations sur les contextes d'analyse des décideurs.

Systèmes OLAP Streaming. Les données provenant d'environnements dynamiques tels que les systèmes de surveillance, de télécommunication ou des systèmes embarqués génèrent des masses importantes de données qui évoluent dynamiquement et rapidement dans le temps. Face à ce type de données, les systèmes OLAP s'avèrent inadaptés [Cuzzocrea, 2009] car les structures multidimensionnelles sur lesquelles ils reposent nécessitent le calcul de pré-agrégats et des processus E.T.L. complexes.

Ce type de données en « flux continus » induit des mises à jour nombreuses et fréquentes afin de rafraîchir les espaces de stockage du système décisionnel. Cela remet en cause le principe même de l'approche OLAP qui repose sur une préparation importante des données analysées.

Références[10]

A

[Abdelhedi, 2009] F. Abdelhedi, « Etude et développement d'un mécanisme d'agrégation pour l'analyse OLAP de documents XML », Mémoire Master 2 RIBD, Université Paul Sabatier, Toulouse III, Juin 2009.

[Abelló, 2001] A. Abelló, J. Samos, F. Saltor, « A Framework for the Classification and Description of Multidimensional Data Models », 12th International Conference on Database and Expert Systems Applications (DEXA'01), pp.668-677, Munich (Germany), Septembre 2001.

[Abelló, *et al.*, 2002] A. Abelló, J. Samos, F. Saltor, « YAM2 (Yet Another Multidimensional Model): An Extension of UML », International Database Engineering and Applications Symposium (IDEAS'02), pp. 172-181, Edmonton (Canada), Juillet 2002.

[Abelló, *et al.*, 2003] A. Abelló, J. Samos, F. Saltor, « Implementing operations to navigate semantic star schemas », DOLAP'03, pp. 56-62, 2003.

[Abelló, *et al.*, 2006] A. Abelló, J. Samos, F. Saltor, « YAM2: a multidimensional conceptual model extending UML », Information Systems, 31(6), pp541-567, 2006.

[Adler, *et al.*, 1972] M.J. Adler, C. Van Doren, « How to Read a Book », Simon & Shuster, 1972.

[Agrawal, *et al.*, 1997] R. Agrawal, A. Gupta, S. Sarawagi, « Modeling Multidimensional Databases », 13th International Conference on Data Engineering (ICDE'97), pp.232-243, Birmingham (U.K), Avril 1997.

[Agrawal, *et al.*, 2000] R. Agrawal, E. L. Wimmers, « A Framework for Expressing and Combining Preferences », International Conference on Management of Data, Dallas, USA, 2000, pp. 297-306.

[Annoni, 2003] E. Annoni, « Conception et développement d'un langage assertionnel pour les bases de données multidimensionnelles », Mémoire D.E.A. 2IL, Université Paul Sabatier, Toulouse III, Juin 2003.

[Annoni, 2007] E. Annoni, « Eléments méthodologiques pour le développement des systèmes décisionnels dans un contexte de réutilisation », Thèse de doctorat, Université Paul Sabatier, Toulouse III, Juillet 2007.

❖[Annoni, Ravat, Teste, Zurfluh, 2005a] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « Une approche d'analyse et de conception de SID à base de patrons », Revue des Sciences et Technologies de l'Information, ISI (Ingénierie des Systèmes d'Information), Hermès, Vol.10, N°6, p.81-106, 2005.

❖[Annoni, Ravat, Teste, Zurfluh, 2005b] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « BIPAD : Une méthode d'analyse et de conception des systèmes d'information décisionnels par réutilisation de patron », Association Information and Management (AIM'05), Toulouse, septembre 2005.

❖[Annoni, Ravat, Teste, Zurfluh, 2006a] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « Towards Multidimensional Requirement Design », 8th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'06), Springer-Verlag, LNCS 4081, A. Min Tjoa, J. Trujillo, p.75-84, Krakow (Poland), septembre 2006.

[10] Toutes les références marquées par le symbole ❖ sont celles dont l'auteur du mémoire est un des auteurs classés par ordre alphabétique.

❖[Annoni, Ravat, Teste, Zurfluh, 2006b] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « Automating the Choice of Decision Support System Architecture », 17th International Conference on Database and Expert Systems Applications (DEXA'06), Springer-Verlag, LNCS 4080, S. Bressan, J. Küng, R. Wagner, p.244-253, Krakow (Poland), septembre 2006.

❖[Annoni, Ravat, Teste, 2006c] E. Annoni, F. Ravat, O. Teste, « Traitements à l'origine des systèmes d'information décisionnels », Revue des Sciences et Technologies de l'Information, ISI (Ingénierie des Systèmes d'Information), Hermès, Vol.11, N°6, p.115-143, 2006.

❖[Annoni, Ravat, Teste, Zurfluh, 2006d] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « Modélisation adaptée aux besoins utilisateurs dans le développement des systèmes d'information décisionnels », 2ème journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'06), Revue des Nouvelles Technologies de l'Information, RNTI-B-2, Cépaduès, p.23-38, juin 2006.

❖[Annoni, Ravat, Teste, Zurfluh, 2006e] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « Méthode de Développement des Systèmes d'Information Décisionnels : Roue de Deming », XXIVème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'06), p.657-673, Hammamet (Tunisie), mai 2006.

❖[Annoni, Ravat, Teste, 2007] E. Annoni, F. Ravat, O. Teste, « Data and Process analyses of Data Warehouse Requirements », 19th International Conference on Software Engineering and Knowledge Engineering (SEKE'07), Knowledge Systems Institute, p.191-196, Boston (Massachusetts, USA), juillet 2007.

❖[Annoni, Ravat, Teste, Zurfluh, 2008a] E. Annoni, F. Ravat, O. Teste, G. Zurfluh, « Modélisation intégrée de la dynamique des systèmes d'information décisionnels », 4ème journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'08), Revue des Nouvelles Technologies de l'Information, RNTI-B-4, Cépaduès Editions, p.35-43, juin 2008.

❖[Annoni, Ravat, Teste, 2008b] E. Annoni, F. Ravat, O. Teste, « Modélisation de la Structure Complexe des Faits et des Mesures », XXVIème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'08), Fontainebleau, p.231-248, mai 2008.

[Atigui, 2009] F. Atigui, « Personnalisation de requêtes dans une base de données multidimensionnelle », Mémoire Master 2 RIBD, Université Paul Sabatier, Toulouse III, Juin 2009.

B

[Baeza-Yates, *et al*., 1999] R. Baeza-Yates, B., Ribeiro-Neto, « Modern Information Retrieval », Addison-Wesley, ACM Press, 1999

[Baralis, *et al*., 1997] Baralis E., Paraboschi S., Teniente E., « Materialized view selection in a multidimensional database », Proc. VLDB '97.

[Baziz, *et al*., 2005] M. Baziz, M. Boughanem, N. Aussenac-Gilles, « Evaluating a Conceptual Indexing Method by Utilizing WordNet », 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Revised Selected Papers, Lecture Notes in Computer Science, Vol. 4022, Springer-Verlag, Septembre 2005.

[Bellatreche, *et al*., 2005] L. Bellatreche, A. Giacometti, P. Marcel, H. Mouloudi, D. Laurent, « A Personalization Framework for OLAP Queries », DOLAP'05, pp. 9–18, 2005.

❖[Bentayeb, Boussaid, Favre, Ravat, Teste, 2009a] F. Bentayeb, O. Boussaid, C. Favre, F. Ravat, O. Teste, « Personnalisation dans les entrepôts de données : bilan et perspectives », 5ème

journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'09), Revue des Nouvelles Technologies de l'Information, RNTI-B-5, Cépaduès Editions, juin 2009.

[Bentayeb, *et al*., 2009b] F. Bentayeb, C. Favre, « RoK: Roll-Up with the K-Means Clustering Method for Recommending OLAP Queries », Database and Expert Systems Applications (DEXA'09), pp.501-515, Linz, Austria, 2009.

[Bentayeb, *et al*., 2009c] F. Bentayeb, O. Boussaid, J. Darmont, N. Harbi, S. Loudcher (Eds.), « Warehousing and Mining Complex Data: Applications to Biology, Medicine, Behavior, Health and Environment », International Journal of Biomedical Engineering and Technology, Vol. 3 (1-2), Inderscience, Geneva, Switzerland, 2009

[Bimonte, *et al*., 2008] S Bimonte, A. Tchounikine, M. Miquel, R. Laurini, « Introduction de l'analyse spatiale et de l'information géographique dans l'analyse multidimensionnelle », Revue des Nouvelles Technologies de l'Information (RNTI), Editions Cépaduès, 2008.

[Bonifati, *et al*., 2001] Bonifati, A., Cattaneo, F., Ceri, S., Fuggetta, A., Paraboschi, S., « Designing data marts for data warehouses », ACM Trans. Softw. Eng. Methodol., 10(4), pp.452–483, 2001.

[Boughanem, 2000] M. Boughanem, « Formalisation et Spécification de Systèmes de Recherche et de Filtrage d'Information », Habilitation à diriger des recherches, Université Paul Sabatier, novembre 2000.

[Boussaïd, *et al*, 2003] O. Boussaid, R.B. Messaoud, R. Choquet S. Anthoard, « X-Warehousing: An XML-Based Approach for Warehousing Complex Data », 10th East European Conf. on Advances in Databases and Information Systems (ADBIS), LNCS 4152, Springer, pp. 39–54, 2006.

[Bouzeghoub, *et al*., 2005] M. Bouzeghoub, D. Kostadinov, « Personnalisation de l'information : aperçu de l'état de l'art et définition d'un modèle flexible de profils », CORIA'05, pp. 201–218, 2005.

❖[Bret, Teste, 1999] F. Bret, O. Teste, « Construction Graphique d'Entrepôts de données », XVIIème Congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'99), p.165-184, La Garde (France), juin 1999.

[Bruckner, *et al*., 2001] R.M. Bruckner, T.W. Ling, O. Mangisengi, A.M. Tjoa, « A Framework for a Multidimensional OLAP Model using Topic Maps », 2nd International Conference on Web Information Systems Engineering (WISE'01), pp.109-118, Kyoto (Japan), Décembre 2001.

[Buzydlowski, *et al*., 1998] J.W. Buzydlowski, I.Y. Song, L. Hassell, « A Framework for Object-Oriented On-line Analytical Processing », 1st International Workshop on Data Warehousing and OLAP (DOLAP'98), pp.10-15, Bethesda (Maryland, USA), Novembre 1998.

C

❖[Cabanac, Chevalier, Ravat, Teste, 2006a] G. Cabanac, M. Chevalier, F. Ravat, O. Teste, « Modèle conceptuel pour bases de données multidimensionnelles annotées », 6ème journées Extraction et Gestion des Connaissances (EGC'06), Revue des Nouvelles Technologies de l'Information, RNTI-E-6, Cépaduès Editions, p.119-124, janvier 2006.

❖[Cabanac, Chevalier, Ravat, Teste, 2006b] G. Cabanac, M. Chevalier, F. Ravat, O. Teste, « Méta-modélisation des bases de données multidimensionnelles annotées », 2ème journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'06), Revue des Nouvelles Technologies de l'Information, RNTI-B-2, Cépaduès Editions, p. 39-54, juin 2006.

❖[Cabanac, Chevalier, Ravat, Teste, 2007] G. Cabanac, M. Chevalier, F. Ravat, O. Teste, « An Annotation Management System for Multidimensional Databases », 9th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'07), Springer-Verlag, LNCS 4654, I-Y. Song, J. Eder, T.M. Nguyen, p. 89-98, Regensburg (Germany), septembre 2007.

❖[Cabanac, Chevalier, Ravat, Teste, 2009] G. Cabanac, M. Chevalier, F. Ravat, O. Teste, « Decisional Annotations: Integrating and Preserving Decision-Makers' Expertise in Multidimensional Systems », Chapitre IV de l'ouvrage "Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications" de la série "Advances in Data Warehousing and Mining", IGI Publishing, Nguyen Manh Tho, ISBN 9781605667485, Juillet 2009.

[Cabibbo, *et al*., 1997] L. Cabibbo, R. Torlone, « Querying Multidimensional Databases », 6th International Workshop Database Programming Languages, DBPL'97, pp.319-335, 1997.

[Cabibbo, *et al*., 1998] L. Cabibbo, R. Torlone, « A Logical Approach to Multidimensional Databases », 6th International Conference on Extending Database Technology (EDBT'1998), pp.183-197, Valencia (Spain), Mars 1998.

[Cabibbo, *et al*., 2000] L. Cabibbo, R. Torlone, « The Design and Development of a Logical System for OLAP », 2nd International Conference on Data Warehousing and Knowledge Discovery (DaWaK'00), pp.1-10, London (UK), Septembre 2000.

[Carneiro, *et al*., 2002] Carneiro, L., Brayner, A.,« X-META : A methodology for data warehouse design with metadata management », Design and Management of Data Warehouses (DMDW)., pages 13–22, 2002.

[Cauvet, *et al*., 2001] C. Cauvet, C. Rosenthal-Sabroux, « Ingénierie des systèmes d'information », Hermes Science Publication, ISBN 2-7462-0219-0, 2001.

[Cavero, *et al*., 2001] Cavero, J. M., Piattini, M., Marcos, E., « Midea : A multidimensional data warehouse methodology », ICEIS, pages 138–144, 2001.

[Chaudhury, *et al*., 1997] S. Chaudhuri, U. Dayal, « An Overview of Data Warehousing and OLAP Technology », SIGMOD Record 26(1), pp.65-74, 1997.

[Choong, *et al*., 2003] Choong, Y.W., Laurent, D., Marcel, P. , « Computing Appropriate Representations for Multidimensional », Data & knowledge Engineering Journal 45(2), 181–203, 2003.

❖[Chrisment, Pujolle, Ravat, Teste, Zurfluh, 2005] C. Chrisment, G. Pujolle, F. Ravat, O. Teste, G. Zurfluh, « Les entrepôts de données », Traité Informatique des Techniques de l'Ingénieur (H3870), G. Zurfluh, 2005.

❖[Chrisment, Pujolle, Ravat, Teste, Zurfluh, 2006] C. Chrisment, G. Pujolle, F. Ravat, O. Teste, G. Zurfluh, « Bases de données décisionnelles », Encyclopédie de l'informatique et des systèmes d'information. J. Akoka, I. Comyn-Wattiau, Vuibert, p.533-546, 2006.

[Codd, 1972] E. F. Codd, « Relational Completeness of Data Base Sublanguages », R. Rustin (ed.): Database Systems: 65-98, Prentice Hall and IBM Research Report RJ 987, San Jose, California, 1972.

[Codd, *et al*., 1993] E.F. Codd, S.B. Codd, C.T. Salley, « Providing OLAP to User-Analysts », IT Mandate, 1993.

[Cuzzocrea, 2009] A. Cuzzocrea, « CAMS: OLAPing Multidimensional Data Streams Efficiently », DAWAK, pp.48-62, 2009.

[Darmont, *et al*. 2008] J. Darmont, E. Olivier, « Biomedical Data Warehouses », Encyclopaedia of Healthcare Information Systems, IGI Publishing, Hershey, PA, USA, May 2008, 149-156.

[Datta, *et al*., 1999] A. Datta, H. Thomas, « The cube data model: A conceptual model and algebra for on-line analytical processing in data warehouses », Decision Support Systems, 27(3), pp.289-301, 1999.

[Dinter, *et al*., 1998] B. Dinter, C. Sapia, G. Höfling, M. Blaschka, « The OLAP Market: State of the Art and Research Issues ». 1st International Workshop on Data Warehousing and OLAP (DOLAP'98), pp.22-27, Bethesda (Maryland, USA), Novembre 1998.

[Dittrich, *et al*., 2005] M. Dittrich, D. Kossmann, A. Kreutz, « Bridging the gap between OLAP and SQL », International Conference on Very Large Data Bases, pp. 1031–1042, 2005.

[Djemal, *et al*., 2008] K. Djemal, C. Soulé-Dupuy, N. Vallès-Parlangeau, « Formal modeling of multistructured documents », IEEE International Conference on Research Challenges in Information Science, RCIS'08, pp.227-236, Marrakech, Maroc, 2008.

[Dousset, 2003] B. Dousset, « Intégration de méthodes interactives de découverte de connaissances pour la veille stratégique », Habilitation à diriger des recherches, Université Paul Sabatier, novembre 2003.

[Espil, *et al*., 2001] M. Espil, A. Vaisman, « Efficient Intensional Redefinition of Aggregation Hierarchies in Multidimensional Databases », DOLAP'01, pp. 1–8, 2001.

[Favre, *et al*., 2007] C. Favre, F. Bentayeb, O. Boussaid, « Evolution et personnalisation des analyses dans les entrepôts de données : une approche orientée utilisateur », INFORSID'07, pp. 308–323, 2007.

[Franconi, *et al*., 2004] E. Franconi, A. Kamble, « A Data Warehouse Conceptual Data Model », 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), pp.435-436, Santorini Island (Greece), Juin 2004.

[Fuhr, *et al*., 2001] N. Fuhr, K. Grojohann, « XIRQL: a query language for information retrieval in XML documents », SIGIR, W.B. Croft, D.J. Harper, D.H. Kraft, J. Zobel (Eds.), ACM Press, New York, 2001, pp. 172–180.

[Ghalamallah, 2008] A. Ghalamallah, « Personnalisation d'une base de données multidimensionnelle », Mémoire Master 2 RIBD, Université Paul Sabatier, Toulouse III, Juin 2008.

[Gargouri, 2006] M. Gargouri, « Assistance à l'élaboration incrémentale d'un magasin de données », Mémoire Master 2, Université Paul Sabatier, Toulouse III, Juin 2006.

[Ghozzi, 2004] F. Ghozzi, « Conception et manipulation de bases de données dimensionnelles à contraintes », Thèse de doctorat, Université Paul Sabatier, Toulouse III, Novembre 2004.

❖[Ghozzi, Ravat, Teste, Zurfluh, 2003a] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, « Constraints and multidimensional databases », 5th International Conference on Enterprise Information Systems (ICEIS'03), p.104-111, Angers (France), avril 2003.

❖[Ghozzi, Ravat, Teste, Zurfluh, 2003b] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, « Contraintes pour modèle et langage multidimensionnels », 19ème Journées Bases de Données Avancées (BDA'03), C. Chrisment, p.383-402, Lyon (France), octobre 2003.

❖[Ghozzi, Ravat, Teste, Zurfluh, 2003c] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, « Modèle Multidimensionnel à Contraintes », 3ème journées Extraction et Gestion des

Connaissances (EGC'03), M-S. Hacid, Y. Kodratoff, D. Boulanger, Revue des Sciences et Technologies de l'Information, Série RIA-ECA, Hermès, Vol.17, N°1-3, janvier 2003.

❖[Ghozzi, Ravat, Teste, Zurfluh, 2004] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, « Contraintes pour modèle et langage multidimensionnels », Revue des Sciences et Technologies de l'Information, ISI (Ingénierie des Systèmes d'Information), Hermès, Vol.9, N°1, juin 2004.

❖[Ghozzi, Ravat, Teste, Zurfluh, 2005] F. Ghozzi, F. Ravat, O. Teste, G. Zurfluh, « Méthode de conception d'une base multidimensionnelle contrainte », 1ère journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'05), Revue des Nouvelles Technologies de l'Information - Entrepôts de Données et l'Analyse en ligne, RNTI-B-1, Cépaduès, p.51-70, juin 2005.

[Giacometti, *et al.*, 2008] A. Giacometti, P. Marcel, E. Negre, « A Framework for Recommending OLAP Queries », DOLAP'08, pp. 73–80, 2008.

[Giacometti, *et al.*, 2009] A. Giacometti, P. Marcel, E. Negre, « Recommending Multidimensional Queries », DAWAK, pp.453-466, 2009

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

[Golfarelli, *et al.*, 2001] M. Golfarelli, S. Rizzi, B. Vrdoljak, « Data Warehouse Design from XML Sources », 4th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2001), November 9, 2001, Atlanta, Georgia, USA, 2001.

[Golfarelli, *et al.*, 2009] M. Golfarelli, S. Rizzi, « Expressing OLAP Preferences », SSDBM'09, Springer Verlag LNCS 5566, pp.83-91, 2009.

[Gray, *et al.*, 1996] J. Gray, A. Bosworth, A. Layman, H. Pirahesh, « Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total », 12th International Conference on Data Engineering (ICDE'96), IEEE Computer Society, pp.152-159, New Orleans (Louisiana, USA), Mars 1996.

[Gupta, 1995] Gupta A., Mumick I.S., « Maintenance of Materialized Views: Problems, Techniques, and Applications », IEEE Data Engineering Bulletin, 1995.

[Gupta, 1997] Gupta H., « Selection of Views to Materialize in a Data Warehouse », International Conference on Database Theory, Athens, Greece, January 1997.

[Gyssen, *et al.*, 1997] M. Gyssens, L.V.S. Lakshmanan, « A Foundation for Multi-dimensional Databases », 23rd International Conference on Very Large Data Bases (VLDB'97), pp.106-115, Août 1997, Athens (Greece).

H

[Hachicha, *et al.*, 2008] M. Hachicha, H. Mahboubi, J. Darmont, « Expressing OLAP operators with the TAX XML algebra »,, 3rd International Workshop on Database Technologies for Handling XML Information on the Web (DataX-EDBT 08), Nantes, France, March 2008

[Hahn, *et al.*, 2000] K. Hahn, C. Sapia, M. Blaschka, « Automatically Generating OLAP Schemata from Conceptual Graphical Models », 3rd international workshop on Data warehousing and OLAP (DOLAP'00), pp.9-16, Washington (DC, USA), Novembre 2000.

[Harinarayan, *et al.*, 1996] V. Harinarayan, A. Rajaraman, J.D. Ullman, « Implementing Data Cubes Efficiently », International Conference on Management of Data, SIGMOD Record 25(2), pp. 205-216, Montreal, Quebec, Canada, June 1996.

[Horner, *et al.*, 2004] J. Horner, I-Y. Song,P.P. Chen, « An analysis of additivity in OLAP systems », 7th ACM Int. Workshop on Data Warehousing and OLAP (DOLAP 2004), ACM Press, pp. 83–91, 2004.

❖[Hubert, Teste, 2009] G. Hubert, O. Teste, « Analyse multigraduelle OLAP », Extraction et Gestion des Connaissances (EGC'09), Revue des Nouvelles Technologies de l'Information, RNTI-E-5, Cépaduès Editions, p.241-252, janvier 2009.

[Hüsemann, *et al*., 2000] B. Hüsemann, J. Lechtenbörger, G. Vossen, « Conceptual data warehouse modeling », Second International Workshop on Design and Management of Data Warehouses (DMDW'00), Stockholm (Sweden), Juin 2000.

[Huyn, 1997] Huyn N., « Multiple-View Self-Maintenance in Data Warehousing Environments », 23rd International Conference on Very Large Data Bases - VLDB'97, Athens (Greece), August 25-29 1997.

I-J

[Inmon, 1994] W.H. Inmon, « Building the Data Warehouse », John Wiley & Sons, ISBN 0471-14161-5, 1994.

[Ioannidis , *et al*., 2005] Ioannidis, Y., G. Koutrika, « Personalized Systems: Models and Methods from an IR and DB Perspective », VLDB'05, pp. 1365–1365, 2005.

[Jensen, *et al*., 2001] M.R. Jensen, T.H. Møller, T.B. Pedersen, « Specifying OLAP Cubes On XML Data », 13th Int. Conf. on Scientific and Statistical Database Management (SSDBM), IEEE Computer Society, pp.101–112, 2001.

[Jerbi, 2007] H. Jerbi, « Mémoire d'expertises décisionnelles à base d'annotations », Mémoire Master 2 RIBD, Université Paul Sabatier, Toulouse III, Juin 2007.

❖[Jerbi, Ravat, Teste, Zurfluh, 2008] H. Jerbi, F. Ravat, O. Teste, G. Zurfluh, « Management of context-aware preferences in multidimensional databases », 3rd International Conference on Digital Information Management (ICDIM'08), IEEE, p.669-675, Londres (UK), novembre 2008.

❖[Jerbi, Ravat, Teste, Zurfluh, 2009a] H. Jerbi, F. Ravat, O. Teste, G. Zurfluh, « Modèle de préférences contextuelles pour les analyses OLAP », 4ème journées sur les Entrepôts de Extraction et Gestion des Connaissances (EGC'09), Revue des Nouvelles Technologies de l'Information, RNTI-E-5, Cépaduès Editions, p.253-258, janvier 2009.

❖[Jerbi, Ravat, Teste, Zurfluh, 2009b] H. Jerbi, F. Ravat, O. Teste, G. Zurfluh, « Applying Recommendation Technology in OLAP Systems », 11th Intl. Conference on Enterprise Information Systems (ICEIS'09), Springer LNBIP 24, J. Filipe, J. Cordeiro, p.220-233, Milan (Italie), 2009.

❖[Jerbi, Ravat, Teste, Zurfluh, 2009c] H. Jerbi, F. Ravat, O. Teste, G. Zurfluh, « Preference-Based Recommendations for OLAP Analysis », 10th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'09), Springer-Verlag, septembre 2009.

K

[Kamps, *et al*., 2004] J. Kamps, M. Marx, M. de Rijke, B. Sigurbjörnsson, « Best-Match Querying from Document-Centric XML », Seventh Int'l Workshop the Web and Databases (WebDB '04), pp. 55-60, 2004.

[Keith, *et al*., 2005] S. Keith, O. Kaser, D. Lemire, « Analyzing Large Collections of Electronic Text Using OLAP », APICS 29th Conf. in Mathematics, Statistics and Computer Science, Acadia University, pp. 17–26, 2005.

[Khrouf, *et al*., 2004] K. Khrouf, C. Soulé-Dupuy, « A Textual Warehouse Approach: A Web Data Repository », M. Mohammadian (Ed.), Intelligent Agents for Data Mining and Information Retrieval, Idea Publishing Group, pp. 101–124, 2004.

[Kimball, 1996] R. Kimball, « The data warehouse toolkit: practical techniques for building dimensional data warehouses », John Wiley & Sons, ISBN 0-471-15337-0, 1996.

[Kimball, *et al*., 1998] Kimball, R., Reeves, L., Thornthwaite, W., Ross, M., Thornwaite, W. « The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing », John Wiley & Sons, Inc., New York, NY, USA. ISBN 0-471-25547-5, 1998.

[Korfhage, 1997] Korfhage, R. R., « Information Storage and Retrieval », JohnWiley & Sons, 1997.

[Kotidis, *et al*., 1999] Kotidis Y., Roussopoulos N., « DynaMat: A Dynamic View Management System for Data Warehouses », ACM SIGMOD '99, Philadephia (Pennsylvania, USA), December 1999.

L

[Labio, *et al*., 1999] Labio W. J., Yerneni R., Garcia-Molina H., « Shrinking the Warehouse Update Window », ACM SIGMOD Conference, Phildelphia (USA), May 1999.

[Labio, *et al*., 2000] Labio W. J., Yang J., Cui Y., Garcia-Molina H., Widom J., « Performance Issues in Incremental Warehouse Maintenance », VLDB 2000: 461-472.

[Lassila, *et al*., 2001] Lassila O., McGuinness D.L., « The Role of Frame-Based Representation on the Semantic Web », Knowledge Systems Laboratory Report KSL-01-02, Stanford University, 2001.

❖[Laurent, Marcel, Ravat, Teste, Zurfluh, 2002] A. Laurent, P. Marcel, F. Ravat, O. Teste, G. Zurfluh, « Entrepôts de données et OLAP : un aperçu orienté recherche », Rapport Final AS 20 GafoDonnées - GT GafOLAP, CNRS-STIC, juin 2002.

[Lehner, 1998] W. Lehner, « Modelling Large Scale OLAP Scenarios », 6th International Conference on Extending Database Technology (EDBT'98), pp.153-167, Valencia (Spain), Mars 1998.

[Lehner, *et al*., 1998] W. Lehner, J. Albrecht, H. Wedekind, « Normal Forms for Multidimensional Databases », 10th International Conference on Scientific and Statistical Database Management (SSDBM'98), pp.63-72, Capri (Italy), Juillet 1998.

[Li, *et al*., 1996] C. Li, X.S. Wang, « A Data Model for Supporting On-Line Analytical Processing », Fifth International Conference on Information and Knowledge Management (CIKM'96), pp.81-88, Rockville (Maryland, USA), Novembre 1996.

[Libourel, 2003] T. Libourel, « Autour de la conception des systems complexes : Modélisation, Evolution, Infrastructures », Habilitation à diriger des recherches de Montpellier II, 2003.

[Luján-Mora, *et al*., 2003] Luján-Mora, S., Trujillo, J., « A comprehensive method for data warehouse design », Design and Management of Data Warehouses, DMDW'03, Berlin, Germany, September 8, 2003.

[Luján-Mora, *et al*., 2006] S. Luján-Mora, Juan Trujillo, I.Y. Song, « A UML profile for multidimensional modeling in data warehouses », Data Knowl. Eng. 59(3): 725-769, 2006.

M-N

[Malinowski, *et al*., 2006] E. Malinowski, E. Zimányi, « Hierarchies in a multidimensional model: From conceptual modeling to logical representation », Data & Knowledge Engineering, 59(2), pp.348-377, 2006.

[Malinowski, *et al*., 2008] E. Malinowski, E. Zimányi, « A conceptual model for temporal data warehouses and its transformation to the ER and the object-relational models », Data & Knowledge Engineering, 64(1), pp.101-133, 2008.

[Mangisengi, *et al*., 1998] O. Mangisengi, A.M. Tjoa, « A multidimensional modeling approach for OLAP within the framework of the relational model based on quotient relations », 1st international workshop on Data warehousing and OLAP (DOLAP'98), pp.40-46, Bethesda (Maryland, USA), Novembre 1998.

[Mbarki, *et al*., 2007] M. Mbarki, C. Soulé-Dupuy, N. Vallès-Parlangeau, « A document repository architecture for heterogeneous business information management », ICEIS (1) pp.192-198, 2007.

[McCabe, *et al*., 2000] C. McCabe, J. Lee, A. Chowdhury, D.A. Grossman, O. Frieder, « On the design and evaluation of a multi-dimensional approach to information retrieval », 23rd Int. ACM Conf. on research and development in Information Retrieval (SIGIR), ACM Press, pp. 363– 365, 2000.

[Meleze, 1972] J. Mélèse, « L'analyse modulaire des systèmes de gestion, AMS », Editions Hommes et Techniques, Paris, 1972.

[Mendelzon, *et al*., 2000] Mendelzon, A., Vaisman, A., « Temporal queries in OLAP », VLDB, pp.242–253, 2000.

[Miquel, 2005] M. Miquel, « Contribution à la modélisation des données et des traitements dans le contexte décisionnel », Habilitation à diriger des recherches, INSA de Lyon, 2005.

[Moody, *et al*., 2000] Moody, D. L., Kortink, M. A. R., « From enterprise models to dimensional models: a methodology for data warehouse and data mart design », Second Intl. Workshop on Design and Management of Data Warehouses, DMDW 2000, Stockholm, Sweden, June 5-6, 2000.

[Morfonios, et *al*., 2008] K. Morfonios, G. Koutrika G., « OLAP Cubes for Social Searches: Standing on the Shoulders of Giants? », 11th International Workshop on Web and Databases (WebDB'08), Vancouver, Canada, June 13, 2008.

[Mothe, 2000] J. Mothe, « Recherche et exploration d'informations -Découverte de connaissances pour l'accès; à l'information », Habilitation à diriger des recherches, Université Paul Sabatier, décembre 2000.

[Mothe, *et al*., 2003] J. Mothe, C. Chrisment, B. Dousset, J. Alau, « DocCube: Multi-dimensional visualisation and exploration of large document sets », Journal of the American Society for Information Science and Technology (JASIST), vol.54(7), Wiley Periodicals, pp. 650–659, 2003.

[Mumick, *et al*., 1997] Mumick I., Quass D., Mumick B., « Maintenance of Data Cubes and Summary Tables in a Warehouse », ACM SIGMOD Conference, Tuscon (Arizona, USA), May, 1997.

[Mundy, *et al*., 2006] J. Mundy, W. Thornthwaite, « The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset », John Wiley & Sons, ISBN 978-0-471-26715-7, 2006.

[Nassis, *et al*., 2004] V. Nassis, R. Rajugan, T.S. Dillon, J. Wenny Rahayu , « Conceptual Design of XML Document Warehouses », 6th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 3181, Springer, pp.1–14, 2004.

[Nguyen, *et al*., 2000] T.B. Nguyen, A. Min Tjoa, R. Wagner, « An Object Oriented Multidimensional Data Model for OLAP », 1st Intl. Conference on Web-Age Information Management, WAIM'00, Shangaï, China, LNCS 1846, pp. 69-82, June 2000.

[Park, *et al*., 2005] B.K. Park, H. Han, I.Y. Song, « XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses », 7th Int. Conf. on Data Warehousing and Knowledge Discovery (DaWaK), LNCS 3589, Springer, pp.32–42, 2005.

[Pedersen, *et al*., 1999] Pedersen T.B., Jensen C.S, « Multidimensional Data Modeling for Complex Data », International Conference on Data Engineering - ICDE'99, March 1999.

[Pedersen, *et al*., 2001] Pedersen T.B., Jensen C.S, Dyreson C. E., « A foundation for capturing and querying complex multidimensional data », Information Systems (IS), vol.26(5), Elsevier, p. 383–423, juillet 2001.

[Pérez, *et al*., 2008a] J.M. Pérez, R.B. Llavori, M.J. Aramburu, T.B. Pedersen, « Contextualizing data warehouses with documents », Decision Support Systems, Elsevier, n°45, pp.77–94, 2008.

[Pérez, *et al*., 2008b] J.M. Pérez, R.B. Llavori, M.J. Aramburu, T.B. Pedersen, « Integrating Data Warehouses with Web Data: A Survey », Transactions on Knowledge and Data Engineering, IEEE, 20(7), pp.940-955, 2008.

[Pokorný, 2001] J. Pokorný, « Modelling Stars Using XML », Proc. 4th ACM Int. Workshop on Data Warehousing and OLAP (DOLAP 2001), pp.24-31, 2001.

[Prat, *et al*., 2002] N. Prat, J. Akoka, « From uml to rolap multidimensional databases using a pivot model », 18e Journées Bases de Données Avancées, BDA'02, pp. 21-25, Evry, 2002.

❖[Pujolle, Ravat, Teste, Tournier, 2008] G. Pujolle, F. Ravat, O. Teste, R. Tournier, « Fonctions d'agrégation pour l'analyse en ligne (OLAP) de données textuelles », Revue des Sciences et Technologies de l'Information, ISI (Ingénierie des Systèmes d'Information), Hermès, Vol.13, N°6, p.61-84, 2008.

[Quass, *et al*., 1997] Quass D., Widom J., « On-Line Warehouse View Maintenance for Batch Updates », ACM SIGMOD Conference, Tuscon (Arizona, USA), May 1997.

[Rafanelli, 2003] M. Rafanelli, « Operators for Multidimensional Aggregate Data », Chapitre V, Multidimensional Databases: Problems and Solutions, IGI Publishing Group, ISBN 1-59140-053-8, p. 116–165, 2003.

❖[Ravat, Teste, Zurfluh, 1999] F. Ravat, O. Teste, G. Zurfluh, « Towards Data Warehouse Design », 8th International Conference on Information and Knowledge Management (CIKM'99), ACM Press Susan Gauch, p.359-366, Kansas City (Missouri, USA), novembre 1999.

❖[Ravat, Teste, Zurfluh, 2000a] F. Ravat, O. Teste, « Object-Oriented Decision Support System », 2nd International Conference en Enterprise information Systems (ICEIS'00), B. Sharp, J. Cordeiro, J. Filipe, p.79-84, Stafford (UK), juillet 2000.

❖[Ravat, Teste, 2000b] F. Ravat, O. Teste, « An Object Data Warehousing Approach: a Web Site Repository », Enlarged 4th East-European Conference on Advances in Databases and Information Systems (ADBIS/DASFAA'00), Symposium on Advances in Databases and Information Systems, Matfyz Press, p.128-137, Prague (Czech Republic), septembre 2000.

❖[Ravat, Teste, 2000c] F. Ravat, O. Teste, « A Temporal Object-Oriented Data Warehouse Model », 11th International Conference on Database and Expert Systems (DEXA'00), Springer-Verlag, LNCS 1873, M. Ibrahim, J. Jüng, N. Revell, p.583-592, London (UK), 2000.

❖[Ravat, Teste, Zurfluh, 2001a] F. Ravat, O. Teste, G. Zurfluh, « Modélisation multidimensionnelle des systèmes décisionnels », 1ère journées Extraction et Gestion des

Connaissances (EGC'01), H. Briand, F. Guillet, Revue des Sciences et Technologies de l'Information, RIA-ECA (Extraction des Connaissances et Apprentissage), Hermès, Vol.1, N°1-2, p.201-212, 2001.

❖[Ravat, Teste, Zurfluh, 2001b] F. Ravat, O. Teste, G. Zurfluh, « Modélisation et extraction de données pour un entrepôt objet », 16ème journées Bases de Données Avancées (BDA'00), A. Doucet, p.119-138, Blois (France), octobre 2000.

❖[Ravat, Teste, 2001c] F. Ravat, O. Teste, « Object-Oriented Decision Support System », dans "Enterprise Information System II", B. Sharp, J. Filipe, J. Cordeiro, Kluwer Academic Publishers, p.42-48, 2001.

❖[Ravat, Teste, 2001d] F. Ravat, O. Teste, « Modélisation et manipulation de données historisées et archivées dans un entrepôt orienté objet », 17ième Journées Bases de Données Avancées (BDA'01), Cépaduès Editions, N. Mouaddib, p.243-256, Agadir (Maroc), octobre 2001.

❖[Ravat, Teste, Zurfluh, 2002] F. Ravat, O. Teste, G. Zurfluh, « Langages pour Bases Multidimensionnelles : OLAP-SQL », Revue des Sciences et Tech. de l'Information, ISI (Ingénierie des Systèmes d'Information), Hermès, Vol.7, N°3, p.11-38, 2002.

❖[Ravat, Teste, Zurfluh, 2005] F. Ravat, O. Teste, G. Zurfluh, « Manipulation et fusion de données multidimensionnelles », 5ème journées Extraction et Gestion des Connaissances (EGC'05), S. Pinson, N. Vincent, Revue des Nouvelles Technologies de l'Information, RNTI-E-3, Cépaduès Editions, p.349-354, janvier 2005.

❖[Ravat, Teste, Zurfluh, 2006a] F. Ravat, O. Teste, G. Zurfluh, « A multiversion-based multidimensional model », 8th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'06), Springer-Verlag, LNCS 4081, A. Min Tjoa, J. Trujillo, p.65-74, Krakow (Poland), septembre 2006.

❖[Ravat, Teste, Zurfluh, 2006b] F. Ravat, O. Teste, G. Zurfluh, « Algèbre OLAP et langage graphique », XXIVème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'06), p.1039-1054, Hammamet (Tunisie), mai 2006.

❖[Ravat, Teste, Zurfluh, 2006c] F. Ravat, O. Teste, G. Zurfluh, « Constraint-Based Multi-Dimensional Databases », dans "Database Modeling for Industrial Data Management: Emerging Technologies and Applications", Zongmin Ma, IGI Publishing, p.323-368, 2006.

❖[Ravat, Teste, 2006d] F. Ravat, O. Teste, « Supporting Data Changes in Multidimensional Data Warehouses », International Review on Computers and Software, Praize Worthy Prize, Wantag - USA, Vol.1 N°3, p.251-259, 2006.

❖[Ravat, Teste, Tournier, Zurfluh, 2007a] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « A Conceptual Model for Multidimensional Analysis of Documents », 26th International Conference on Conceptual Modeling (ER'07), Springer-Verlag, LNCS 4801, C. Parent, K.-D. Schewe, V. C. Storey, B. Thalheim, p.550-565, Auckland (New Zealand), novembre 2007.

❖[Ravat, Teste, Tournier, Zurfluh, 2007b] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Querying Multidimensional Databases », 11th East-European Conference on Advances in Databases and Information Systems (ADBIS'07), Springer-Verlag, LNCS 4690, Y.E. Ioannidis, B. Novikov, B. Rachev, p.298-313, Varna (Bulgarie), septembre 2007.

❖[Ravat, Teste, Tournier, Zurfluh, 2007c] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Integrating Complex Data into a Data Warehouse », 19th International Conference on Software Engineering and Knowledge Engineering (SEKE'07), Knowledge Systems Institute, p.483-486, Boston (Massachusetts, USA), juillet 2007.

❖[Ravat, Teste, Tournier, 2007d] F. Ravat, O. Teste, R. Tournier, « OLAP Aggregation Function for Textual Data Warehouse », 9th International Conference on Enterprise Information Systems (ICEIS'07), INSTICC Press, Vol. DISI, J. Cardoso, J. Cordeiro, J. Filipe, Funchal (Madeira, Portugal), p.151-156, juin 2007.

❖[Ravat, Teste, Tournier, 2007e] F. Ravat, O. Teste, R. Tournier, « Analyse multidimensionnelle de documents via des dimensions OLAP », Document numérique, Hermès, Numéro spécial "Entreposage de documents et données semi-structurées", Vol.9, 2007.

❖[Ravat, Teste, Tournier, Zurfluh, 2007f] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Modèle conceptuel pour l'analyse multidimensionnelle de documents », 3ème journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'07), Revue des Nouvelles Technologies de l'Information, RNTI-B-3, Cépaduès Editions, p.161-175, juin 2007.

❖[Ravat, Teste, Zurfluh, 2007g] F. Ravat, O. Teste, G. Zurfluh, « Personnalisation de bases de données multidimensionnelles », XXVème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'07), p.121-136, Perros-Guirec, mai 2007.

❖[Ravat, Teste, Tournier, Zurfluh, 2008a] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Designing and Implementing OLAP Systems from XML Documents », Annals of Information Systems, Vol. 3, issue "New Trends in Data Warehousing and Data Analysis", S. Kozielski, R. Wrembel, 2008.

❖[Ravat, Teste, Tournier, Zurfluh, 2008b] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Algebraic and graphic languages for OLAP manipulations », International Journal of Data Warehousing and Mining, IGI Publishing, Vol.4, N°1, p.17-46, 2008.

❖[Ravat, Teste, Tournier, Zurfluh, 2008c] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « A Top-keyword extraction method for OLAP document », 9th International Conference on Data Warehousing and Knowledge Discovery (DAWAK'08), Springer-Verlag, LNCS 5182, Il Yeol Song, Johan Eder, Tho Manh Nguyen, p.55-64, Torino (Italy), septembre 2008.

❖[Ravat, Teste, Tournier, Zurfluh, 2008d] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Top_keywords : agrégation de mots-clefs dans un environnement d'analyse en ligne (OLAP) », 4ème journées sur les Entrepôts de Données et l'Analyse en ligne (EDA'08), Revue des Nouvelles Technologies de l'Information, RNTI-B-4, Cépaduès Editions, p.85-98, juin 2008.

❖[Ravat, Teste, Tournier, 2008e] F. Ravat, O. Teste, R. Tournier, « Multidimensional Analysis of XML Document Contents With OLAP Dimensions », Chapitre VI de l'ouvrage "Progressive Methods in Data Warehousing and Business Intelligence: Concepts and Competitive Analytics" de la série "Advances in Data Warehousing and Mining", IGI Publishing, David Taniar, 2008.

❖[Ravat, Teste, Tournier, Zurfluh, 2008f] F. Ravat, O. Teste, R. Tournier, G. Zurfluh, « Systèmes OLAP pour l'analyse de documents textuels XML : Conception et implantation d'une base de données multidimensionnelle », 2ème atelier Méthodes Avancées de Développement des Systèmes d'Information (MADSI'08), organisé dans le cadre du congrès INFORSID'08, Fontainebleau, 27 mai 2008.

❖[Ravat, Teste, 2008g] F. Ravat, O. Teste, « Personalization and OLAP Databases », Annals of Information Systems, Vol. 3, issue "New Trends in Data Warehousing and Data Analysis", S. Kozielski, R. Wrembel, 2008.

[Rivest, et al., 2001] Rivest, S., B´edard, Y., Marchand, P., « Toward better suppport for spatial decision making: Defining the characteristics of spatial on-line analytical processing (SOLAP) », Geomatica 55(4), pp.539–555, 2001.

[Rizzi, 2007] S. Rizzi, « OLAP Preferences: a Research Agenda », 10th International Workshop on Data Warehousing and OLAP (DOLAP'07), ACM, pp.99-100, Lisbon (Portugal), Novembre 2007.

[Rizzi, *et al.*, 2006] S. Rizzi, A. Abelló, J. Lechtenbörger, J. Trujillo, « Research in data warehouse modeling and design: dead or alive? 9th International Workshop on Data Warehousing and OLAP, pp.3-10, Arlington, Virginia, USA, November 10, 2006.

[Robertson, 2004] S. Robertson, « Understainding Inverse Document Frequency: On theoretical arguments for IDF », Journal of Documentation, 60(5), Emerald Publishing Group, p. 503–520, 2004.

[Rouhaud, 2005] O. Rouhaud, « Bases de données décisionnelles : Fusion de tables multidimensionnelles », Mémoire Master 2, Université Paul Sabatier, Toulouse III, Juin 2005.

S

[Sapia, 2000] Sapia, C., « PROMISE: Predicting Query Behavior to Enable Predictive Caching Strategies for OLAP Systems », DaWaK'00, LNCS 1874, pp. 224–233, Heidelberg (2000).

[Sapia, *et al.*, 1998] C. Sapia, M. Blaschka, G. Höfling, B. Dinter, V, « Extending the E/R Model for the Multidimensional Paradigm »,,ER Workshops, pp. 105-116, 1998.

[Schneider, 2003] M. Schneider, « Well-formed data warehouse structures », 5th Intl. Workshop on Design and Management of Data Warehouses, DMDW'03, Germany, 2003.

[Sèdes, 1998] F. Sèdes, « Bases documentaires-Hyperbases : Proposition d'un modèle générique et contribution à la spécification d'un langage pour l'intégration et la manipulation d'informations semi-structurées », Habilitation à diriger des recherches, Université Paul Sabatier, décembre 1998.

[Shukla, *et al.*, 1998] Shukla A., Deshpande P.M., Naughton J.F., « Materialized View Selection for Multidimensional Datasets », VLDB 1998: 488-499

[Shukla, *et al.*, 2000] Shukla A., Deshpande P.M., Naughton J.F., « Materialized View Selection for Multi-Cube Data Models », EDBT 2000: 269-284

[Smith, *et al.*, 2004] P. Smith, L. Hobbs, S. Hillson, S. Lawande, « Oracle 10g Data Warehousing », Elseiver, ISBN 1-55558-322-9, 2004.

[Soulé-Dupuy, 2001] C. Soulé-Dupuy, « Bases d'informations textuelles : des modèles aux applications », Habilitation à diriger des recherches, Université Paul Sabatier, décembre 2001.

[Sullivan, 2001] D. Sullivan, « Document Warehousing and Text Mining », Wiley John & Sons, ISBN: 0471399590, 2001.

[Sureau, *et al.*, 2009] F. Sureau, F. Bouali, G. Venturini, « Optimisation heuristique et génétique de visualisations 2D et 3D dans OLAP : premiers résultats », Journées Entrepôts et Analyses en ligne (EDA'09), Montpellier, juin2009.

T

[Tchounikine, 1994] A. Tchounikine, « Activité dans les Bases de Données Objets : le concept de Schéma Actif », Thèse de l'Université Paul Sabatier, Juillet 1993.

❖[Teste, 2001] O. Teste, « Towards Conceptual Multidimensional Design in Decision Support Systems », 5th East-European Conference on Advances in Databases and Information Systems (ADBIS'01), Research Communications Vol.1, A. Caplinskas, J. Eder, p.77-88, Vilnius (Lithuania), septembre 2001.

❖[Teste, 2000a] O. Teste, « Modélisation et manipulation d'entrepôts de données complexes et historisées », Thèse de l'Université Paul Sabatier, décembre 2000.

❖[Teste, 2000b] O. Teste, « Elaboration d'entrepôts de données complexes », XVIIIème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'00), p.229-245, Lyon (France), mai 2000.

[Thalhammer, *et al.*, 2001] T. Thalhammer, M. Schrefl, M. Mohania, « Active DataWarehouses: Complement-ing OLAP with Analysis Rules », Data and Knowledge Engineering 39(3), 241–269, 2001.

[Theodoratos, *et al.*, 1999] Theodoratos D., Sellis T., « Designing Data Warehouses », DKE 31, (3): 279-301, 1999.

[Torlone, 2003] R. Torlone, « Conceptual Multidimensional Models », Chapitre 3 de l'ouvrage Multidimensional Databases: Problems and Solutions, pp. 69–90, IGI Publishing Group, 2003.

[Tournier, 2004] R Tournier, « Vers un langage de manipulation graphique des bases multidimensionnelles », Mémoire D.E.A. 2IL, Université Paul Sabatier, Toulouse III, Juin 2004.

[Tournier, 2007] R Tournier, « Analyse en ligne (OLAP) de documents », Thèse de doctorat, Université Paul Sabatier, Toulouse III, Décembre 2007.

[Trujillo, *et al.*, 1998] J. Trujillo, M. Palomar, « An Object-Oriented Approach to Multidimensional Database Conceptual Modeling », DOLAP 1998: 16-21.

[Tryfona, *et al.*, 1999] N. Tryfona, F. Busborg, J. G. Borch Christiansen, « starER: A Conceptual Model for Data Warehouse Design », 2nd ACM Intl. Workshop on Data Warehousing and OLAP (DOLAP), ACM Press, p. 3–8, 1999.

[Tseng, *et al.*, 2006] Tseng, F.S.C., Chou, A.Y.H., « The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence », J. of Decision Support Systems (DSS), vol.42(2), Elsevier, pp. 727–744, 2006.

[Tsois, *et al.*, 2001] Tsois, A., Karayannidis, N., and Sellis, T. K., « Mac : Conceptual data modeling for olap », 3rd Intl. Workshop on Design and Management of Data Warehouses, DMDW'01, Interlaken, Switzerland, June 4, 2001.

U-V-W

[Vaisman, *et al.*, 2009] A. A. Vaisman, E. Zimányi, « What Is Spatio-Temporal Data Warehousing? », DAWAK, pp 9-23, 2009.

[Vassiliadis, *et al.*, 1999] P. Vassiliadis, T.K. Sellis, « A Survey of Logical Models for OLAP Databases », SIGMOD Record, ACM, 28(4), pp.64-69, 1999.

[Wang, *et al.*, 2003] Wang, H., Li, J., He, Z., Gao, H., « Xaggregation: Flexible Aggregation of XML Data », 4th Intl. Conf. on Advances in Web-Age Information Management (WAIM), LNCS 2762, Springer, pp. 104–115, 2003.

[Wang, *et al.*, 2005] Wang, H., Li, J., He, Z., Gao, H., « OLAP for XML Data », 5th Intl. Conf. on Computer and Information Technology (CIT), IEEE Computer Society, pp. 233–237, 2005.

[Widom, 1992] J. Widom, « The Starburst Rule System: Language Design, Implementation, and Applications », IEEE Data Eng. Bull., vol.15(1-4), pp15-18, 1992.

[Widom, 1995] J. Widom, « Research problems in data warehousing », 4th International Conference on Information and Knowledge Management (CIKM'95), ACM, Baltimore (Maryland, USA), Novembre 1995.

[Wiwatwattana, *et al*., 2007] Wiwatwattana, N., Jagadish, H.V., Lakshmanan, L.V.S., Srivastava, D., « X3: A Cube Operator for XML OLAP », 23rd Intl. Conf. on Data Engineering (ICDE), IEEE Computer Society, pp. 916–925, 2007.

[Yang, *et al*., 1997] Yang J., Karlapalem K., Li Q., « Algorithms for materialized view design in data warehousing environment », Proc. VLDB '97.

[Yang, *et al*., 1998] Yang J., Widom J., « Maintaining Temporal Views Over Non-Temporal Information Sources For Data Warehousing », 6th International Conference on Extending Database Technology, Valencia, Spain, March 1998.

[Yang, *et al*., 2000] Yang J., Widom J., « Temporal View Self-Maintenance in a Warehousing Environment », 7th International Conference on Extending Database Technology - EDBT 2000, Konstanz (Germany), March 2000.

[Zhou, *et al*., 1996] Zhou G., Hull R., King R., « Generating Data Integration Mediators that Use Materialization », Journal of Intelligent Information Systems, Volume 6(2), pages 199-221, May 1996.

[Zhuge, *et al*., 1995] Zhuge Y., Garcia-Molina H., Hammer J., Widom J., « View Maintenance in a Warehousing Environment », ACM SIGMOD Conference, San Jose (California, USA), May 1995.

[Zhuge, *et al*., 1998] Zhuge Y., Garcia-Molina H., Wiener J. L., « Consistency Algorithms for Multi-Source Warehouse View Maintenance », Journal of Distributed and Parallel Databases, volume 6, number 1, January 1998.