Vers l'OLAP sémantique pour l'analyse en ligne des données complexes

Sabine Loudcher

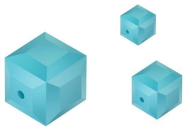
Habilitation à Diriger des Recherches

spécialité : Informatique

présentée par

Sabine LOUDCHER

préparée au sein du laboratoire ERIC – Université Lumière Lyon 2



Vers l'OLAP sémantique pour l'analyse en ligne des données complexes

soutenue publiquement le 29 juin 2011 devant le jury composé de :

M. Daniel LEMIRE	Rapporteur	(Professeur, Université du Québec à Montréal, Canada)
Mme Thérèse LIBOUREL	Rapporteur	(Professeur, Université Montpellier 2)
M. Gilles VENTURINI	Rapporteur	(Professeur, Université de Tours)
M. Pascal PONCELET	Examinateur	(Professeur, Université Montpellier 2)
M. Omar BOUSSAID	Coordonnateur	(Professeur, Université Lyon 2)

ERIC LABO EQUIPE DE RECHERCHE EN INGENIERIE DES CONNAISSANCES

UNIVERSITÉ LUMIÈRE LYON 2 UNIVERSITÉ DE LYON

UNIVERSITÉ DE LYON

Remerciements

Je tiens à remercier madame Thérèse LIBOUREL, monsieur Daniel LEMIRE, monsieur Gilles VENTURINI et monsieur Pascal PONCELET pour avoir accepté d'être membre de mon jury d'habilitation à diriger des recherches et du temps passé à évaluer mon travail.

Les travaux présentés dans ce mémoire ont été réalisés au sein du laboratoire ERIC dont je suis le directeur adjoint depuis 2003. Nicolas et Djamel, tour à tour enseignant, directeur, collègue, je vous remercie sincèrement pour votre confiance, votre soutien et vos encouragements. Merci également à tous les collègues du laboratoire ERIC, Valérie, Julien, Jean-Hugues, Ricco, et à tous ceux que je ne peux pas citer car la liste est trop longue, pour votre gentillesse et disponibilité.

Une habilitation à diriger des recherches est aussi un travail collectif. Merci à tous mes collègues de l'axe ENA-DC, Cécile, Nouria, Fadila, Jérôme et Omar, pour nos projets communs et nos échanges scientifiques parfois très animés mais toujours très riches. Merci à Omar, Stéphane et Jérôme pour vos conseils lors de la rédaction de ce mémoire.

J'ai débuté ma carrière de maître de conférences à l'IUT Lumière en 1998 et je souhaite remercier chaleureusement Paul et Michel pour m'avoir confié, malgré mon manque d'expérience, d'importantes responsabilités pédagogiques et administratives, et pour m'avoir guidée dans cet apprentissage. Merci aussi à Hélène pour sa longue et précieuse collaboration et pour m'avoir permis de me concentrer sur mon HDR ces derniers temps. Que tous les collègues de l'IUT Lumière soient également remerciés.

Je terminerai ces remerciements en pensant aux miens. Je dédie ce mémoire à Denis, Alexandre et Colette qui ont vécu au quotidien la préparation de cette HDR, qui m'ont supportée et qui m'ont donné la force d'aller jusqu'au bout. Merci.

Lyon, juin 2011.

Chapitre 1

Introduction générale

1.1 Contexte des travaux

Dans le contexte de l'informatique décisionnelle dont le but est de collecter, organiser, stocker et analyser l'information pour aider la prise de décision, Inmon introduit les entrepôts de données [Inm96]. Un entrepôt est une collection de données, orientées sujet, intégrées, non volatiles et historisées, organisées comme support d'un processus d'aide à la décision. Les données sont extraites, nettoyées, transformées en un format unique qui les prépare à l'analyse. Au niveau conceptuel, les données sont modélisées de façon multidimensionnelle avec des indicateurs à observer (mesures) et des axes d'observation (dimensions). Au niveau logique, le modèle multidimensionnel est exprimé par un schéma en étoile, en flocon de neige ou en constellation avec les notions de faits, mesures, dimensions et hiérarchies [Kim96]. Pour analyser les données de l'entrepôt de manière multidimensionnelle et interactive, Codd *et al.* introduisent l'analyse OLAP (*On-Line Analytical Processing*) et la définissent selon douze règles [CCS93]. L'OLAP offre la possibilité d'agréger, de visualiser, d'explorer les données à l'aide d'opérateurs. L'ensemble de ce processus est désigné par le terme d'entreposage des données et comprend plusieurs phases pour l'intégration, la structuration (au sens modélisation) et l'analyse en ligne des données.

A l'heure actuelle les entrepôts de données et l'OLAP sont des technologies relativement bien maîtrisées quand il s'agit de données "simples". Cependant, la communauté scientifique s'accorde pour dire que, avec l'avènement du Web et la profusion des données multimédias (son, image, vidéo, texte, ...), les données sont de plus en plus hétérogènes, diverses et qu'elle sont devenues complexes. L'avènement des données complexes a remis en cause le processus d'entreposage et d'analyse des données ; il a induit l'émergence de nouveaux problèmes de recherche comme l'intégration des données complexes dans les entrepôts, le stockage, la représentation

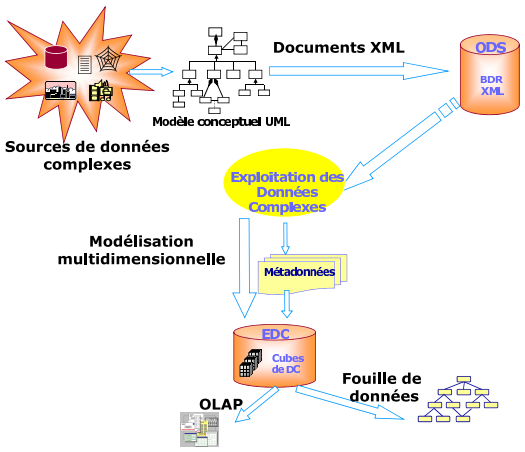


FIGURE 1.1 – Entreposage et analyse des données complexes

ou la modélisation, l'analyse en ligne et la fouille de telles données. Une communauté de recherche a vu le jour et se structure autour d'ateliers associés à des conférences nationales [BBDR05, BT06, YCLM10] ou internationales [ZSD02, HLL+04, ZRT08].

Dans cette communauté, notre équipe de recherche travaille sur les problèmes d'intégration, de modélisation, d'analyse en ligne, de sécurité et de performances des entrepôts de données complexes. Elle propose un processus complet d'entreposage (figure 1.1) des données complexes [BBDR03, DBB+03, BDBLR08, MRL+09, BMM+11].

Une des premières difficultés avec les données complexes réside dans la diversité des formats, diversité qu'il faut prendre en compte dès la phase d'intégration. Une deuxième difficulté est qu'il n'existe pas de modèle unifié pour représenter les données complexes. Notre équipe a choisi d'utiliser le formalisme XML pour structurer, modéliser et stocker les données complexes dans un format unique. XML permet de structurer les données et de stocker leur contenu ; sa flexibilité permet de représenter des données peu structurées, semi-structurées ou très structurées.

Pour la phase d'intégration, des modèles conceptuels, logiques et physiques sont définis. Le modèle conceptuel est traduit sous la forme d'une DTD ou d'un schéma XML. Puis est générée une collection de documents XML. Ces documents sont stockés dans une base de données XML native ou une base de données relationnelle qui sert d'ODS (*Operational Data Storage*) à l'entrepôt de données.

Dans la phase de structuration et pour préparer les données à l'analyse, les données complexes sont modélisées sous forme multidimensionnelle. Un entrepôt de documents XML est ainsi construit. Notre équipe propose également de décrire les cubes OLAP par un schéma XML. Le cube XML est généré automatiquement à partir des besoins de l'utilisateur exprimés par le modèle conceptuel multidimensionnel de l'entrepôt et à partir du corpus de données complexes représentées sous forme de documents XML. Ce cube XML fournit un contexte d'analyse qui peut être analysé par des opérateurs OLAP ou par des méthodes de fouille de données.

Personnellement, dans ce processus, nous travaillons plus particulièrement sur l'analyse en ligne des données complexes. Plusieurs problèmes de recherche non encore ou partiellement résolus se posent. Pour les comprendre, revenons sur les fondements de l'OLAP et sur la spécificité des données complexes.

La vocation de l'OLAP est de réaliser une analyse interactive et multidimensionnelle des données de l'entrepôt. On parle de navigation dans les données et d'analyse exploratoire. Cette analyse en ligne agrège les données pour pouvoir les explorer et les visualiser. Pour cela les données sont représentées dans une structure particulière appelée un cube ou hypercube. L'OLAP dispose d'opérateurs pour résumer les données sous forme d'agrégats (ou au contraire pour détailler les éléments agrégés) et d'opérateurs pour visualiser les informations contenues dans le cube. Ces opérateurs sont dits de navigation et sont classiquement décomposés en opérateurs de structuration (*Rotate*, *Switch*, *Push*, *Pull*), de sélection (*Slice*, *Dice*) et d'agrégation (*Roll-up*, *Drill-down*). Nous qualifions ces opérateurs de classiques.

L'OLAP a fait l'objet de nombreux travaux de recherche et est devenue une technologie bien maîtrisée et largement utilisée dans les systèmes d'informatique décisionnelle. Son apport est de rendre possible la navigation interactive dans les données, la visualisation rapide des informations et l'exploration de la structure multidimensionnelle des données. Cependant, l'exploration est faite par l'utilisateur mais sans outil pour le guider automatiquement dans le cube. C'est à l'utilisateur de naviguer dans les données à la recherche d'informations intéressantes et c'est aussi à l'utilisateur d'évaluer la pertinence des informations découvertes pour savoir si elles constituent de nouvelles connaissances. A l'inverse de la fouille de données (*data mining*), l'OLAP ne permet pas d'extraire automatiquement des connaissances à partir des données [Cha98]. De telles possibilités seraient très utiles à l'utilisateur OLAP. L'idée d'associer les principes de l'OLAP aux méthodes de fouille de données apparait et Han propose le terme de OLAM (*On-Line Analytical Mining*) [Han97, HCC98]. De tous ces constats et idées, émerge un nouveau défi scientifique : "comment enrichir l'OLAP pour ne plus se limiter à l'exploration et à la visualisation des données ?" Une voix peut être trouvée dans la combinaison de l'OLAP et de la fouille de données.

Avec l'avènement des données complexes, l'analyse en ligne doit s'adapter à la nature spécifique des données complexes tout en gardant l'esprit de l'OLAP. Cependant, les opérateurs OLAP sont définis pour des données classiques, c'est à dire souvent numériques et ils deviennent inadaptés quand il s'agit de données complexes par exemple composées de textes, images, son ou vidéos. Plusieurs questions se posent. Comment agréger des données complexes qui comportent des images ou du texte ? Comment visualiser des données complexes dans un cube quand la mesure n'est plus classiquement numérique ? ...

Une autre spécificité des données complexes réside dans la sémantique qu'elles véhiculent. Par exemple, le dossier médical d'un patient est une donnée complexe. Il peut être composé de textes (comptes-rendus médicaux, avis du médecin traitant, ...), d'images (échographie, IRM, scanner, ...), de données numériques (résultats d'analyse, ...) et ce à des moments différents dans le temps. Toutes ces informations décrivent le même patient mais selon des points de vue différents ou complémentaires. Elles forment une entité sémantique qui doit être analysée en tant que telle, à l'instar du médecin qui établit un diagnostic médical. Par conséquent, un nouveau problème émerge : comment prendre en compte la sémantique contenu dans les données complexes lors de l'analyse ?

Pei propose un modèle général (GOLAP) pour pouvoir faire de l'OLAP sur les données complexes, modèle qui généralise le cadre classique de l'OLAP [Pei03]. Il formule deux problèmes : (1) comment partionner les objets de base, c'est-à-dire les faits détaillés, en groupes agrégeables sachant que l'ensemble des groupes doit former une hiérarchie ; (2) comment agréger les groupes ? L'auteur souligne que dans le cas des données complexes, cette fonction d'agrégation n'est pas forcément une moyenne ou une somme mais qu'elle peut être basée sur une régression, une classification, ... Alors qu'il pose les problèmes de modélisation et d'analyse en ligne des données complexes, on peut regretter que Pei n'ait pas poursuivi ces travaux.

L'objectif de ce mémoire est d'apporter des premières solutions aux problèmes posés par l'analyse en ligne des données complexes, notamment pour :

- enrichir les possibilités de l'analyse OLAP en la dotant de nouvelles possibilités,
- créer une analyse en ligne adaptée aux données complexes,
- faire évoluer l'OLAP vers une analyse sémantique des données.

1.2 Contributions et organisation du mémoire

Notre première idée a été de combiner l'OLAP à la fouille de données pour enrichir les possibilités d'analyse. La fouille de données est un ensemble de techniques destinées à extraire des données celles qui ont le plus de valeur tout en les mettant à la portée des utilisateurs (décideurs) en temps opportun. Les informations extraites prennent la forme de tendances, de modèles de classification ou de régression, d'association, d'écarts par rapport à une norme, ... Une fois validée, l'information extraite devient une connaissance. La fouille de données est un des maillons essentiels du processus d'Extraction des Connaissances à partir des Données (ECD) ou *Knowledge Discovery in Databases* (KDD) en anglais. Le processus d'ECD se compose de plusieurs phases : sélection des données, préparation et transformation des données, fouille des données, validation et interprétation des connaissances [FPSS96]. D'après Zighed *et al.*, on peut répartir les méthodes de fouille de données en trois types selon qu'elles visent à : (1) décrire, réduire, explorer les données ; (2) classifier, structurer les données ; (3) modéliser, expliquer, prévoir [ZR02]. Les méthodes de fouille de données n'ont pas nécessairement besoin d'être associées à un entrepôt de données, mais dans ce cas, il faut envisager séparément les phases d'acquisition, de sélection et de préparation des données, phases essentielles pour garantir la performance des méthodes de fouille de données à la fois en termes de qualité et de temps de réponse.

L'analyse en ligne et la fouille de données sont considérées comme deux champs de recherche séparés et ont connu des évolutions indépendantes. Selon les communautés scientifiques, la fouille de données fait partie du processus d'entreposage et constitue une méthode d'analyse possible comme peut l'être l'OLAP ; ou alors la construction d'un entrepôt de données permet de couvrir les phases de sélection, préparation et transformation des données du processus d'ECD. La fouille de données et l'OLAP étant issues de deux communautés scientifiques différentes, peu de travaux de recherche traitent du couplage entre les deux domaines. Néanmoins dès la fin des années 90, plusieurs auteurs proposent d'associer les principes de l'OLAP aux méthodes de fouille de données pour enrichir l'analyse en ligne [Cha98, Han97, HCC98]. Dans un article sur l'avenir des travaux de recherche dans les entrepôts de données, Rizzi *et al.* pensent qu'il faut prendre en compte les besoins de la fouille de données dès la modélisation et la conception de l'entrepôt [RALT06]. Ils citent un certain nombre de problèmes sur comment l'OLAP et la fouille de données peuvent se compléter mutuellement. Ramakrishnan *et al.* présentent également les problèmes ainsi que de nouvelles approches pour la fouille dans les cubes de données [RC07]. A l'heure actuelle toutes ces questions sont reprises et développées dans [GMH09]. Selon les travaux, le couplage entre l'analyse en ligne et la fouille de données est désigné par les termes de OLAM (*On-Line Analytical Mining*), *OLAP Intelligence*, *Multidimensional Mining*, ...

Nous pensons que la combinaison de l'OLAP et de la fouille de données est une voix intéressante car elle peut déboucher sur une analyse des données plus riche que la fouille d'un côté ou l'OLAP de l'autre. L'avantage de l'OLAP est de fournir une analyse en ligne, une visualisation simple et rapide de l'information, une vision multidimensionnelle des données. La fouille permet d'extraire des connaissances à partir des données et a une grande variété de méthodes avec des objectifs d'analyse différents. On peut regretter d'une part que la fouille ne se fasse généralement pas en ligne et qu'elle ne traite que des données représentées sous forme tabulaire (tableau *individus-variables*) et d'autre part que les opérateurs OLAP classiques permettent seulement d'agréger, de visualiser et d'explorer les données.

Cependant le couplage de l'OLAP et de la fouille de données n'est pas une question triviale ou seulement technologique et de nombreux problèmes scientifiques se posent : comment l'OLAP et les entrepôts peuvent intégrer des algorithmes de fouille ; comment stocker dans un entrepôt les connaissances extraites par une méthode de fouille ; comment exécuter en ligne, sur des cubes parfois de grande taille, des algorithmes de fouille d'une certaine complexité et consommateurs de temps, ... ? Une autre question est liée au fait que la fouille ne travaille que sur des données tabulaires alors que un des intérêts de l'OLAP est d'exploiter l'aspect multidimensionnel des données. Dans des données tabulaires, il n'y a plus de notions de hiérarchie de dimensions, toutes les dimensions ont été "aplaties".

Dans nos réflexions, nous avons identifié trois façons de réaliser le couplage :

- L'adaptation des structures multidimensionnelles. La première façon d'associer l'OLAP et la fouille de données est de transformer les données multidimensionnelles des cubes en données tabulaires. Une fois les données ainsi transformées, il est possible d'appliquer des algorithmes de fouille. Cette première voix est la plus intuitive mais elle peut être réductrice car l'aspect hiérarchique des données est perdu.
- L'extension des opérateurs OLAP. Une autre possibilité consiste à étendre le langage des requêtes OLAP pour simuler des techniques de fouille telles que la détection de règles d'association, la classification, ... Cette possibilité est intéressante mais ne couvre pas toutes les méthodes.
- L'adaptation des algorithmes de fouille de données. La troisième façon de faire le couplage consiste à modifier les algorithmes de fouille de façon à pouvoir les utiliser sur des données multidimensionnelles. Cette voix est la plus prometteuse mais la plus difficile à réaliser.

Pour explorer concrètement ces différentes pistes, nous avons co-dirigé une thèse sur ce sujet [Ben06]. Les résultats des travaux développés dans cette thèse sont très encourageants et nous permettent d'apporter des premières solutions à l'extension de l'OLAP et pour l'analyse des données complexes. Mais pour répondre plus largement

aux questions posées par l'analyse en ligne des données complexes, l'idée d'associer l'OLAP à la fouille de données ne suffit pas car elle ne couvre évidemment pas tous les problèmes. Par exemple, les données complexes peuvent être composées de documents qui contiennent, entre autres, du texte. Se posent alors les questions de "comment modéliser de façon multidimensionnelle de telles données complexes" et "comment faire une analyse en ligne sur ces données"... Nous avons cherché une autre voix. Parmi les techniques maîtrisant le traitement des documents textes, nous nous sommes intéressés à la recherche d'information (*Information Retrieval*). Plusieurs travaux montrent l'intérêt d'associer la recherche d'information (RI) et l'OLAP [LGO02]. Dans nos travaux nous utilisons doublement les principes de la recherche d'information : d'une part pour extraire, des données complexes, des informations qui sont utilisées lors de la modélisation multidimensionnelle des données et de la construction du cube ; d'autre part pour l'analyse du cube. En associant l'OLAP, la recherche d'information et la fouille de données nous proposons une modélisation multidimensionnelle et une analyse en ligne adaptées aux données complexes comportant des documents textes.

Après avoir exploré ces différentes voix que constituent le couplage de l'OLAP à la fouille de données et plus récemment celui de l'OLAP à la recherche d'information, nous pensons qu'il faut poursuivre pour dépasser les limites de l'OLAP. Pour analyser les données complexes, l'OLAP a commencé à s'adapter à leur structure et à leur spécificité, comme par exemple avec XOLAP (XML OLAP) et SOLAP (*spatial* OLAP). Mais il faut aller au delà et nous pensons qu'un des défis est d'extraire et d'analyser (en ligne) la sémantique contenue dans les données complexes. Pour cela, nous voulons faire évoluer l'OLAP vers une nouvelle forme d'analyse : l'OLAP sémantique.

Dans la suite de ce mémoire, nous exposons les propositions et solutions nées de ces réflexions. Le mémoire est organisé en deux grandes parties.

Dans le chapitre 2, nous proposons d'étendre les capacités de l'OLAP et de dépasser la simple exploration avec deux nouvelles possibilités d'analyse : AROX (*Association Rules Operator for eXplication*) qui explique des phénomènes observés à l'aide de règles d'association détectées dans le cube ; OPReT (*Online Prediction by Regression Tree*) qui permet de prévoir la valeur de la mesure des faits.

Dans le chapitre 3, nous montrons comment analyser en ligne les données complexes avec trois opérateurs adaptés : CoDaViC (*Complex Data Visualization by Correspondences*) qui permet la visualisation de faits associés à des données complexes ; OpAC (*Operator for Aggregation by Clustering*) qui agrège de telles données ; ORCA (*Operator for Reorganization by multiple Correspondence Analysis*)

qui détecte les régions intéressantes d'un cube en réorganisant les modalités ou membres d'une dimension.

Pour tester nos propositions, nous avons conçu deux plates-formes logicielles implémentées sous la forme d'applications Web. La première, intitulée MiningCubes, est une plate-forme complète pour l'analyse en ligne des données complexes ; elle comprend l'ensemble des nouveaux opérateurs proposés et est présentée dans le chapitre 2. Pour illustrer plus concrètement les propositions faites pour l'analyse des données complexes, un exemple réel est mis en oeuvre : l'analyse des publications scientifiques d'un laboratoire de recherche. La deuxième plate-forme, intitulée PUMA(*PUblication Modeling and Analysis*), est une application dédiée à l'analyse en ligne des publications et est présentée dans le chapitre 3. Ces deux plates-formes sont actuellement sur un serveur local. L'une est en cours de ré-écriture sous forme de services Web, l'autre est en cours de développement.

Enfin, le chapitre 4 conclut ce mémoire en présentant un bilan général de l'ensemble de nos contributions et en proposant une discussion sur le nouveau cadre d'analyse en ligne que devrait être l'OLAP sémantique.

Les contributions scientifiques et les développements informatiques ont été réalisés en partie dans le cadre de la thèse de Riadh Ben Messaoud [Ben06] que nous avons co-encadrée et de plusieurs stages d'étudiants en master sous notre responsabilité. Depuis septembre 2009, nous co-dirigeons également la thèse de Adbalah Sair à l'Ecole Nationale des Sciences Appliquées d'Agadir au Maroc.

Chapitre 2

Enrichissement de l'analyse OLAP

Résumé

Pour étendre les capacités de l'analyse en ligne, notre idée est d'associer l'OLAP à la fouille de données. Nous proposons dans ce chapitre deux nouvelles possibilités d'analyse dans un cube avec l'explication et la prédiction des données multidimensionnelles. L'opérateur d'explication AROX *(*Association Rules Operator for eXplication*) se fonde sur une recherche guidée de règles d'association. Pour adapter le support et la confiance d'une règle au contexte OLAP, nous proposons une nouvelle définition du support et de la confiance en y intégrant la mesure du cube. Dans le contexte du* What If Analysis*, l'opérateur* OPReT *(*Online Prediction by Regression Tree*) prédit la valeur de la mesure d'une cellule selon une démarche d'apprentissage automatique et utilise les arbres de régression comme méthode de prédiction.*

2.1 Introduction

L'OLAP est une méthode d'analyse multidimensionnelle et interactive des données de l'entrepôt. Selon les fondements définis par Codd *et al.*, elle repose sur une manipulation intuitive des données [CCS93]. L'OLAP est guidée étape par étape par l'utilisateur qui doit se poser des questions précises comme par exemple "les ventes de lecteurs MP3 en 2005 sont-elles plus importantes en France qu'en Espagne ?". La démarche exploratoire de l'OLAP suppose que l'utilisateur ait une expertise importante du domaine analysé s'il veut pouvoir en extraire des connaissances intéressantes. C'est à l'utilisateur de trouver manuellement les connaissances potentiellement contenues dans les données d'un cube. La technologie OLAP se limite donc à des tâches exploratoires et ne fournit pas d'outils automatiques pour aider l'utilisateur à expliquer des valeurs de cellules, des associations existant dans les données multidimensionnelles, à prédire des valeurs dans le cube, ...

De plus, une des règles fondatrices, précise que l'OLAP doit extraire et gérer les valeurs manquantes dans la représentation multidimensionnelle du cube. Aucun opérateur OLAP standard ne permet de faire cette gestion.

Ces différents constats nous permettent de dire qu'il faut faire évoluer l'OLAP vers d'autres possibilités d'analyse. Parmi toutes les extensions possibles, nous choisissons d'aborder le problème de la recherche d'explication et celui de la prédiction dans un cube. Notre idée est de combiner l'OLAP à la fouille de données et d'étendre ainsi les possibilités de l'OLAP. En effet, dans le cadre de la fouille de données, l'utilisateur peut se poser des questions moins précises ("quelle est la structure des ventes de lecteur MP3 ?" ou "combien de ventes de lecteurs MP3 peut-on espérer l'année prochaine en France ?") et l'utilisateur a moins besoin de connaissances sur le domaine. Les méthodes de fouille sont nombreuses et visent des objectifs différents (description, structuration, explication, prédiction, ...). Elles recherchent automatiquement les informations potentiellement utiles pour répondre aux questions de l'utilisateur. Après validation, elles transforment ces informations en connaissances.

Pour la recherche d'explication des valeurs des cellules du cube, nous avons exploré l'idée d'extraire des règles d'association dans le cube. L'extraction de règles d'association est une technique de fouille dont l'objectif est proche de celui que nous nous sommes fixés où l'utilisateur cherche à comprendre les relations (ou associations) entre les valeurs de la mesure avec celles des dimensions.

Pour la prédiction de cellules dont la valeur de la mesure est manquante, nous nous sommes tournés vers les arbres de régression. Cette méthode de prédiction ne suppose pas d'hypothèse sur les données et elle est adaptée au contexte de la prédiction de la valeur de la mesure (classiquement quantitative) par celles des dimensions (souvent qualitatives).

Pour explorer ces pistes, nous avons encadré différents travaux de recherche en thèse ou en master. Les travaux ont débouché sur deux nouvelles possibilités d'analyse dans un cube OLAP : l'explication des données multidimensionnelles qui se fonde sur les règles d'association avec l'approche AROX (*Association Rules Operator for eXplication*) et la prédiction de la valeur de la mesure d'une cellule avec OPReT (*Online Prediction by Regression Tree*) en utilisant les arbres de régression. Ces travaux sont validés par plusieurs publications dans des conférences ou ouvrages [BBLR06c, BLRBM06, BLRMB07, BNBLRB08, SEE+11] et sont concrétisés par deux nouveaux opérateurs OLAP.

Dans ce chapitre, nous présentons le résultat de ces recherches. Dans la section 2.2, nous introduisons la notion de cube OLAP ainsi que les notations qui seront utilisées dans tout le mémoire. Nous présentons ensuite la recherche d'explication (section 2.3) et la prédiction dans un cube (section 2.4). Nous terminons le chapitre (section 2.5) en

décrivant la plate-forme logicielle MiningCubes dans laquelle les nouveaux opérateurs OLAP sont développés.

2.2 Définitions et notations

2.2.1 Généralités

La notion de fait désigne l'objet que l'on veut analyser. Sur le fait, on observe une ou des mesures (indicateurs). A chaque fait est associé un ensemble de valeurs, valeurs prises par le fait pour chaque mesure. Les mesures sont le plus souvent numériques. Par exemple, si on veut étudier les ventes d'une grande entreprise, les faits seront les ventes et les mesures peuvent être le chiffre d'affaires et le montant des bénéfices. A chaque mesure est associée une fonction d'agrégation (somme, moyenne, minimum, maximum, comptage, ...) qui permet de calculer, à partir des données détaillées, la valeur des mesures aux différents niveaux hiérarchiques des dimensions. La valeur de la mesure ainsi calculée s'appelle un agrégat. Dans notre exemple des ventes, la fonction d'agrégation peut être la somme.

Une dimension définit un axe d'analyse et offre à l'utilisateur des points de vue différenciés pour analyser ou observer les faits. Une dimension est constituée d'un ou de plusieurs niveaux hiérarchiques, appelés aussi attributs. Ces niveaux correspondent aux différents niveaux de détail qu'il est possible d'observer sur les faits, c'est à dire sur les données gérées par l'entrepôt. Un niveau est composé d'éléments appelés membres ou modalités. Les liens entre les niveaux peuvent avoir des cardinalités différentes, c'est ce qui détermine les liens entre les modalités des différents niveaux. Par exemple, les ventes peuvent être observées selon plusieurs dimensions : Lieu, Produit, Temps, Profession du client, Sexe du client et Promotion. La dimension Lieu peut comporter deux niveaux Pays et Continent. Le niveau Continent est défini par les modalités Asie, Europe et Amérique ; et ainsi de suite pour les autres niveaux et dimensions.

L'instance d'un modèle multidimensionnel définit un hypercube. Un hypercube ou un cube OLAP représente les valeurs (détaillées ou agrégées) de la mesure dans un espace multidimensionnel défini par l'utilisateur. L'utilisateur choisit les dimensions selon lesquelles il veut analyser les faits et choisit pour chaque dimension le niveau hiérarchique sur lequel il veut travailler. La combinaison des modalités des dimensions sélectionnées forment un ensemble de coordonnées qui désigne les cellules du cube. Chaque cellule contient la valeur de la ou des mesures correspondant à la combinaison des modalités. Selon les niveaux hiérarchiques choisis, la valeur de la mesure d'une cellule est une valeur détaillée ou une valeur agrégée. Par exemple, dans le cube de la figure 2.1, l'utilisateur a choisi d'observer les ventes selon trois dimensions Lieu, Produit et Temps et avec les niveaux respectifs Pays, Article et Année.

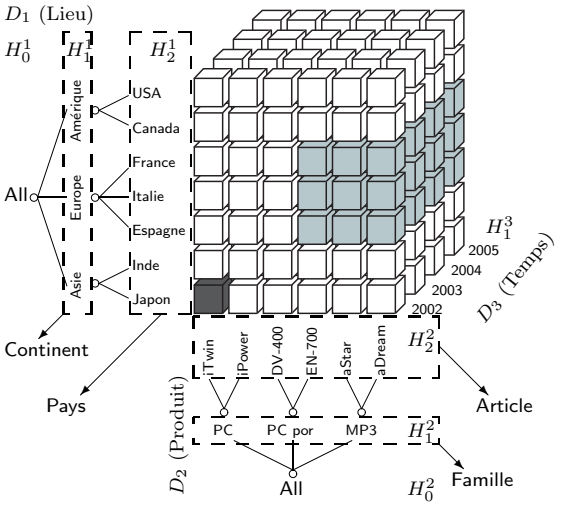


FIGURE 2.1 – Exemple de cube et sous-cube de données

L'algèbre OLAP introduite par Codd *et al.* définit des opérateurs qui permettent la navigation entre les cellules, les modalités ou les niveaux hiérarchiques du cube [CCS93]. Les opérateurs OLAP standards sont répartis en différentes catégories :

1. Les opérateurs de manipulation qui sont liés à la structure du cube et qui permettent de réorienter la vue multidimensionnelle ou d'en changer l'agencement (opérateurs *Rotate*, *Switch*, *Split*, *Nest*, *Push* et *Pull*).
2. Les opérateurs de sélection qui permettent de ne travailler que sur une partie du cube (opérateurs *Slice* et *Dice*).
3. les opérateurs de forage qui calculent des informations plus agrégées ou plus détaillées à des niveaux de granularité différents avec le *Roll-up* et le *Drill-down*.

2.2.2 Notations

Dans la suite du document, nous adoptons les notations suivantes :

Soit $\mathcal{C}$ un cube de données avec un ensemble non vide de d dimensions $\mathcal{D} = \{D^1, ..., D^i, ..., D^d\}$ et m mesures $\mathcal{M} = \{M_1, ..., M_q, ..., M_m\}$. Le cas $m = 0$ correspond à la situation où il n'y a pas de mesure définie. $\mathcal{H}^i$ est l'ensemble des niveaux hiérarchiques de la dimension D^i. H_j^i est le $j^{ème}$ niveau hiérarchique de la dimension D^i. Dans l'exemple de la figure 2.1, la dimension Lieu D^1 contient deux niveaux : le niveau Continent noté H_1^1 et le niveau Pays noté H_2^1.

$\mathcal{A}^{ij} = \{a_1^{ij}, ..., a_t^{ij}, ..., a_l^{ij}\}$ représente l'ensemble des l modalités a_t^{ij} du niveau hiérarchique H_j^i de la dimension D^i. Le niveau Continent (H_1^1) contient trois

modalités : Asie notée a_1^{11}, Europe notée a_2^{11} et Amérique notée a_3^{11}.

Nous considérons également qu'une cellule dans un cube $\mathcal{C}$ est pleine (respectivement, vide) si elle contient au moins un fait (respectivement, ne contient pas de faits).

D'une manière générale, un cube permet de représenter un ensemble de faits, en présentant les valeurs prises par une mesure M_q (avec $M_q \in \mathcal{M}$) selon l'ensemble de modalités $\mathcal{A}^{ij}$ des dimensions $\{D^1, ..., D^i, ..., D^d\}$ qui caractérisent les faits pour un niveau d'agrégation donné H_j^i.

L'utilisateur définit le cube $\mathcal{C}'$ sur lequel il veut travailler, ce cube est un sous-cube du cube $\mathcal{C}$.

Soit $\mathcal{D}'$ un sous-ensemble non vide de $\mathcal{D}$ avec p dimensions $\{D^1, ..., D^p\}$ ($\mathcal{D}' \subseteq \mathcal{D}$ et $p \leq d$). Le p-uplet $(\Theta^1, ..., \Theta^p)$ définit un sous-cube selon $\mathcal{D}'$ si $\forall i \in \{1, ..., p\}$, $\Theta^i \neq \emptyset$ et s'il existe un indice $j \geq 1$ tel que $\Theta^i \subseteq \mathcal{A}^{ij}$.

Un sous-cube $\mathcal{C}'$ correspond à une portion du cube $\mathcal{C}$. Parmi les d dimensions existantes, seules p sont choisies. Pour chaque dimension retenue $D^i \in \mathcal{D}'$, un niveau hiérarchique H_j^i est fixé et un sous-ensemble Θ^i non vide de modalités est sélectionné parmi l'ensemble des modalités $\mathcal{A}^{ij}$ du niveau. Soit M_q (avec $M_q \in \mathcal{M}$) la mesure choisie par l'utilisateur. Par souci de simplification des notations, la mesure M_q choisie par l'utilisateur est notée M.

Par exemple, dans la figure 2.1, parmi les 6 dimensions existantes ($d = 6$), l'utilisateur a choisi d'observer les ventes selon trois dimensions ($p = 3$) : Lieu notée D^1, Produit notée D^2 et Temps notée D^3 avec les niveaux respectifs Continent noté H_1^1, Article noté H_2^2 et Année noté H_1^3. $\mathcal{D}'$ est le sous-ensemble de dimensions avec $\mathcal{D}' = \{D^1, D^2, D^3\}$. Soient le sous-ensemble de modalités Θ^1={Europe} du niveau H_1^1 (Continent) de la dimension D^1 (Lieu), le sous-ensemble de modalités Θ^2={EN-700, aStar, aDream} du niveau H_2^2 (Article) de la dimension D^2 (Produit) et le sous-ensemble de modalités Θ^3={2002, 2003, 2004, 2005} du niveau H_1^3 (Année) de la dimension D^3 (Temps).

Dans ce cas, $(\Theta^1, \Theta^2, \Theta^3)$=(Europe, {EN-700, aStar, aDream}, {2002, 2003, 2004, 2005}) correspond au sous-cube grisé de la figure 2.1.

2.3 Explication dans un cube OLAP

2.3.1 Motivation et positionnement

L'OLAP ne fournit pas d'outil automatique pour expliquer les relations et les associations potentiellement existantes dans les données. Par exemple, un utilisateur peut observer, à partir d'un cube de données de ventes, que le niveau de ventes des sacs de couchage est particulièrement élevé dans une ville donnée. En revanche, cette

exploration ne permet pas d'expliquer automatiquement les raisons de ce niveau de ventes élevé alors qu'il est, par exemple, expliqué par une association entre une saison estivale et une clientèle relativement jeune. Ce constat nous a motivé pour proposer une nouvelle possibilité d'analyse en ligne : l'explication. Nous nous sommes orientés vers le principe des règles d'association.

Plusieurs travaux soulignent l'intérêt d'explorer automatiquement les cubes avec des algorithmes de recherche des règles d'association [KHC97, IKA02, GC98]. Les différents auteurs (Kamber, Imieliński, Goil, Choudhary, ...) pensent que la structure multidimensionnelle du cube, avec ses agrégats pré-calculés, est un contexte favorable pour la recherche de règles. Les agrégats calculés avec la fonction d'agrégation COUNT permettent d'avoir directement les fréquences pour le calcul du support et de la confiance des règles. Ils évitent ainsi de parcourir plusieurs fois toutes les données.

Nous avons conduit une étude bibliographique sur la problématique de la recherche des règles d'association dans les cubes OLAP. Une synthèse et une comparaison des travaux ont été faites et elles sont présentées dans [BLRMB07]. Dans ce mémoire, nous ne reprenons que quelques points.

A notre connaissance, Kamber *et al.* sont les premiers à chercher des règles d'association dans des données multidimensionnelles et ils introduisent la fouille guidée de règles (*metarule-guided mining*) [KHC97]. Un modèle général, appelé méta-règle, définit le contenu des règles recherchées avec une conjonction de prédicats pour l'antécédent et avec un prédicat instancié à une modalité d'une dimension précise pour le conséquent. Ainsi la méta-règle va permettre de piloter le processus d'extraction vers des règles inter-dimensionnelles répondant aux besoins d'analyse de l'utilisateur.

Zhu distingue trois types de règles d'association qui peuvent être extraites à partir d'un cube de données [Zhu98] :

- les règles inter-dimensionnelles, constituées de prédicats non répétitifs où chaque instance de prédicat provient d'une dimension distincte du cube ;
- les règles intra-dimensionnelles, constituées de prédicats répétitifs, c'est-à-dire de modalités provenant d'une même dimension du cube ;
- les règles hybrides.

A la différence de l'approche de Kamber *et al.* (où les règles sont extraites directement de la structure multidimensionnelle des données), Zhu aplatit le cube et le transforme selon un format tabulaire approprié, puis recherche les motifs fréquents en utilisant l'algorithme Apriori et génère enfin les règles d'association.

Imieliński *et al.* proposent une approche de généralisation des règles d'association appelée cubegrades [IKA02]. Un cubegrade est un formalisme qui calcule le différentiel de mesure des agrégats entre un cube initial et un cube obtenu suite à des opérations de spécialisation (*drill-down*), de généralisation (*roll-up*) ou de permutation (changement de modalité dans une dimension) sur le cube initial. Les auteurs reprochent

aux règles d'association classiques de n'exploiter que les comptages (fonction COUNT) dans la recherche des règles. Ils proposent ainsi d'exploiter dans les cubegrades d'autres agrégations de mesure.

Tjioe et Taniar proposent une approche d'extraction de règles d'association inter-dimensionnelles qui consiste en un ensemble de procédures de pré-traitement des données afin de les préparer pour la phase de fouille [TT05]. Ces procédures de pré-traitement partent des dimensions choisies par l'utilisateur et avec des requêtes SQL élaguent dans la table des faits les données jugées inutiles pour le processus de fouille. Les données filtrées sont aplaties selon un format tabulaire (*initialized table*) à partir duquel des règles inter et intra-dimensionnelles sont extraites avec des algorithmes de type Apriori.

Les règles d'association inter-dimensionnelles sont les plus exploitées dans l'ensemble des approches. Dans de telles règles, les prédicats sont généralement non répétitifs sauf dans les travaux d'Imieliński *et al.* et de Tjioe et Taniar. Dans les cubegrades d'Imieliński *et al.*, il peut y avoir des prédicats de plusieurs dimensions dont les instances peuvent être redondantes dans l'antécédent et le conséquent de l'implication. Quant à Tjioe et Taniar, ils sont les seuls à proposer des règles d'association inter-dimensionnelles qui peuvent avoir, soit des prédicats répétitifs, soit des prédicats non répétitifs.

Dans notre approche, nous souhaitons découvrir des règles qui expriment des associations entre les différentes dimensions du cube car nous pensons que c'est le type de règles le plus intéressant pour l'utilisateur. Nous nous plaçons donc dans le contexte des règles inter-dimensionnelles avec prédicats non répétitifs. De plus nous utilisons le principe de meta-règle de Kamber *et al.* car il permet de guider la recherche des règles selon les besoins de l'utilisateur. La recherche des règles est faite directement dans le cube de données sans transformation préalable de ce dernier.

Hormis dans les cubegrades, toutes les propositions se fondent sur le comptage de la fréquence des données, avec la fonction COUNT, pour calculer le support et la confiance des règles. Nous pensons qu'il faut modifier la définition du support et de la confiance pour l'adapter au contexte de l'OLAP. Nous proposons de les calculer en fonction d'une mesure quantitative (positive) quelconque du cube. Ainsi, contrairement à toutes les autres approches, une règle d'association n'est plus évaluée selon le nombre d'occurrences des faits qu'elle supporte mais est évaluée selon la mesure des faits qu'elle supporte. Cette nouvelle définition du support et de la confiance enrichit le sens des règles d'association extraites et les adapte au contexte de l'analyse OLAP. Ces différents choix et objectifs constituent les fondements de notre opérateur d'explication AROX (*Association Rules Operator for eXplication*). Afin de valoriser les connaissances induites par les règles extraites, nous mettons également en place un codage graphique des règles extraites pour les visualiser dans le cube.

2.3.2 Définitions

Agrégation SUM. L'utilisateur définit le sous-cube $\mathcal{C}'$ sur lequel il veut travailler. Pour notre approche, nous supposons que la mesure M choisie par l'utilisateur est quantitative, additive et qu'elle ne prend que des valeurs positives. Nous supposons également que la fonction d'agrégation est la somme (SUM) de la mesure M. L'agrégat noté $\mathsf{SUM}_M(\Theta^1, \ldots, \Theta^p)$ est la somme de toutes les valeurs de la mesure M des faits présents dans le sous-cube.

Par exemple, le *chiffre d'affaires* des ventes du sous-cube grisé de la figure 2.1 est l'agrégat $\mathsf{SUM}_{\textit{Chiffre Affaires}}(\mathsf{Europe}, \{\mathsf{EN\text{-}700}, \mathsf{aStar}, \mathsf{aDream}\}, \{2002, 2003, 2004, 2005\})$ qui est la somme des chiffres d'affaires de toutes les cellules du sous-cube.

Prédicat inter-dimensionnel. Un prédicat dimensionnel dans une dimension D^i, noté x_i, est un prédicat de la forme $\langle a \in \mathcal{A}^{ij} \rangle$. Un prédicat dimensionnel est la valeur d'une modalité d'une dimension. Par exemple dans la dimension D^1, $x_1 = \langle a \in \{\mathsf{Amérique}, \mathsf{Europe}, \mathsf{Asie}\} \rangle$ est un prédicat dimensionnel.

$(x_1 \wedge \cdots \wedge x_p)$ est un prédicat inter-dimensionnel si $\forall i \in \{1, \ldots, p\}$, x_i est un prédicat dimensionnel de D^i, avec $(2 \leq p \leq d)$. Un prédicat inter-dimensionnel est une conjonction de prédicats dimensionnels non répétitifs. Par exemple, $(x_1 \wedge x_p)$ avec $(\langle a_1 \in \mathcal{A}^{11} \rangle \wedge \langle a_2 \in \mathcal{A}^{21} \rangle)$ est un prédicat inter-dimensionnel.

Méta-règle inter-dimensionnelle. Plantevit *et al.* proposent une partition des dimensions du cube en trois sous-ensembles $\mathcal{D}_\mathcal{C}, \mathcal{D}_\mathcal{A}$ et $\mathcal{D}_\mathcal{I}$ [PCL+05] :

- $\mathcal{D}_\mathcal{C}$ est un sous-ensemble de p dimensions dites de *contexte*. Les règles d'association sont extraites à partir des dimensions du contexte $\mathcal{D}_\mathcal{C}$;
- $\mathcal{D}_\mathcal{A}$ est un sous-ensemble de $(s+r)$ dimensions d'*analyse* dans lequel les prédicats d'une méta-règle inter-dimensionnelle sont choisis ;
- $\mathcal{D}_\mathcal{I}$ est le sous-ensemble des dimensions restantes et inutilisées. Ces dimensions sont totalement agrégées et n'interviennent ni dans la définition du contexte ni dans la définition de la méta-règle.

Une méta-règle inter-dimensionnelle est un modèle de règles défini par l'utilisateur. Elle est de la forme :

$$\mathcal{R} \;\Big|\; \begin{array}{l} \text{Dans le contexte } (\Theta^1, \ldots, \Theta^p) \\ (x_1 \wedge \cdots \wedge x_s) \Rightarrow (y_1 \wedge \cdots \wedge y_r) \end{array}$$

où $(\Theta^1, \ldots, \Theta^p)$ est un sous-cube de $\mathcal{C}$ défini selon le sous-ensemble des dimensions $\mathcal{D}_\mathcal{C}$. Ce sous-cube désigne la portion du cube dans laquelle les règles d'association sont cherchées. A la différence de la méta-règle proposée par Kamber *et al.* [KHC97], notre méta-règle permet de cibler un contexte d'analyse précis dans le cube. De plus, de part

sa construction, notre méta-règle induit des règles d'association inter-dimensionnelles avec des prédicats non répétitifs. x_k (respectivement y_k) est un prédicat dimensionnel dans une dimension de $\mathcal{D}_\mathcal{A}$. La conjonction des prédicats $(x_1 \wedge \cdots \wedge x_s) \wedge (y_1 \wedge \cdots \wedge y_r)$ est un prédicat inter-dimensionnel dans $\mathcal{D}_\mathcal{A}$. Le nombre de prédicats $(s + r)$ dans la méta-règle est égal au nombre de dimensions dans $\mathcal{D}_\mathcal{A}$.

Par exemple, en plus des trois dimensions représentées dans la figure 2.1, il y a trois autres dimensions : Profession du consommateur (D^4), Sexe du consommateur (D^5) et Promotion (D^6). Supposons que l'utilisateur ait défini la partition suivante :
- $\mathcal{D}_\mathcal{C} = \{D^4, D^5\}$ ={Profession, Sexe} ;
- $\mathcal{D}_\mathcal{A} = \{D^1, D^2, D^3\}$ ={Lieu, Produit, Temps} ;
- $\mathcal{D}_\mathcal{I} = \{D^6\}$ ={Promotion} ;

ainsi que la méta-règle inter-dimensionnelle :

> Dans le contexte (Etudiant, Femme)
> $\langle a_1 \in$ Continent$\rangle \wedge \langle a_3 \in$ Année$\rangle \Rightarrow \langle a_2 \in$ Produit$\rangle$

Selon cette méta-règle, les règles d'association sont extraites à partir du sous-cube concernant seulement la population des femmes étudiantes. La dimension Promotion n'intervient pas dans le processus de découverte. L'algorithme va chercher des associations entre les modalités du niveau Continent de D^1 et celles du niveau Année de D^3 pour impliquer les modalités du niveau Produit de D^2.

2.3.3 Support et confiance basés sur la mesure

Classiquement, le support (Supp) d'une règle d'association $X \Rightarrow Y$, dans une base de transactions, est la fréquence d'avoir des transactions contenant à la fois les items X et Y. La confiance (Conf) de cette règle est la fréquence conditionnelle d'avoir dans une transaction l'item Y sachant que la transaction contient déjà l'item X. Une règle d'association est dite *fréquente* si son support est supérieur ou égal au support minimal (*minsupp*) fixé par l'utilisateur. De même, une règle est considérée intéressante lorsque sa confiance est supérieure ou égale à la confiance minimale (*minconf*) fixée par l'utilisateur.

Dans le contexte de l'analyse en ligne, le fait OLAP joue le rôle d'une transaction. Le cube contient des agrégats pré-calculés qui correspondent aux valeurs de la mesure pour tous les croisements possibles des modalités des dimensions. Ces agrégats sont d'une grande utilité dans le calcul du support et de la confiance car ils évitent de parcourir plusieurs fois toutes les données. En particulier, la fonction d'agrégation COUNT permet d'avoir directement les fréquences pour le calcul du support et de la confiance classiques.

Avec la méta-règle $\mathcal{R}$, R_1 est un exemple de règle qui pourrait être découverte :

$$R_1 \mid \begin{array}{l}\text{Dans le contexte (Etudiant, Femme)}\\ \text{Continent=Amérique} \wedge \text{Année=2004} \Rightarrow \text{Article=aStar}\end{array}$$

avec comme support

$$\text{SUPP}(R_1) = \frac{\text{COUNT(Amérique,aStar,2004,All,Etudiant,Femme,All)}}{\text{COUNT(All,All,All,All,Etudiant,Femme,All)}}$$

et comme confiance

$$\text{CONF}(R_1) = \frac{\text{COUNT(Amérique,aStar,2004,All,Etudiant,Femme,All)}}{\text{COUNT(Amérique,All,2004,All,Etudiant,Femme,All)}}$$

Dans ces expressions, le support et la confiance dépendent des fréquences des faits et sont calculés avec la fonction d'agrégation COUNT. D'un point de vue statistique, cela revient à étudier la population des faits en fonction de leurs occurrences. Cependant, du point de vue de l'analyse en ligne, les faits OLAP sont, le plus souvent, observés selon des mesures plus intéressantes que leurs simples fréquences. Par exemple, un directeur commercial est plus intéressé d'observer les ventes d'un produit en fonction du chiffre d'affaires plutôt qu'en fonction du nombre de produits vendus. Il nous paraît donc important de prendre en compte la mesure des faits dans la génération des règles d'association.

	2004		2005	
	Amérique	Europe	Amérique	Europe
PC	1200	800	950	500
PC por	2500	2700	2800	3200
MP3	10600	5900	11400	9100

(a)

	2004		2005	
	Amérique	Europe	Amérique	Europe
PC	$ 60000	$ 33000	$ 28000	$ 10000
PC por	$ 500000	$ 567000	$ 420000	$ 544000
MP3	$ 116000	$ 118000	$ 57000	$ 91000

(b)

FIGURE 2.2 – Fragment du cube des ventes selon (a) les fréquences et selon (b) le chiffre d'affaires

Les ventes sont représentées selon leur fréquence dans la figure 2.2 (a)et selon leur chiffre d'affaires dans la figure 2.2 (b). Pour un support minimum donné, on se rend compte que certains motifs sont fréquents selon la fréquence des faits alors qu'ils ne le sont pas selon le chiffre d'affaires et *vice versa*. Par exemple, pour un support minimum $minsupp = 0,2$, les motifs ($\langle$ Amérique $\rangle$, $\langle$ MP3 $\rangle$, $\langle$ 2004 $\rangle$) et ($\langle$ Amérique $\rangle$, $\langle$ MP3 $\rangle$, $\langle$ 2005 $\rangle$) sont fréquents selon la fréquence des faits (cellules grisées dans le

tableau de la figure 2.2 (a)). En revanche, ces mêmes motifs ne sont pas fréquents selon le chiffre d'affaires. Les motifs fréquents selon cette mesure sont ($\langle$ Europe $\rangle$, $\langle$ PC por $\rangle$, $\langle$ 2004 $\rangle$) et ($\langle$ Europe $\rangle$, $\langle$ PC por $\rangle$, $\langle$ 2005 $\rangle$) (cellules grisées dans le tableau de la figure 2.2 (b)).Ces derniers motifs sont plus intéressants et plus cohérents dans le contexte de l'analyse en ligne.

Pour ces raisons, nous modifions dans notre approche les définitions du support et de la confiance pour qu'elles se fondent sur la mesure des faits. Notre proposition consiste à étendre la notion de fréquence qui est un rapport du nombre d'unités statistiques, à la notion de concentration qui est un rapport de mesure avec des unités de masse [Gin21]. Par analogie à cette notion de concentration, nous supposons que la mesure M est quantitative, additive et qu'elle ne prend que des valeurs positives. Nous définissons le support et la confiance d'une règle d'association inter-dimensionnelle comme :

$$\textsc{Supp}(R) = \frac{\mathsf{SUM}_M(x_1,...,x_s,y_1,...,y_r,\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}{\mathsf{SUM}_M(\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}$$

$$\textsc{Conf}(R) = \frac{\mathsf{SUM}_M(x_1,...,x_s,y_1,...,y_r,\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}{\mathsf{SUM}_M(x_1,...,x_s,\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}$$

Dans ces nouvelles expressions, nous ramenons la population étudiée à la population des unités de masse mesurées. Notre définition du support et de la confiance est plus générale que celle du cas classique avec la population des unités de faits. Le cas classique est en réalité un cas particulier où la mesure M est égale à l'unité et où l'agrégation SUM est équivalente à la fonction COUNT.

2.3.4 Démarche

Recherche des motifs fréquents et des règles d'association. Classiquement, les algorithmes d'extraction des règles d'association comportent deux grandes étapes : la recherche des motifs fréquents puis, à partir des motifs fréquents, la génération des règles d'association. La génération des règles s'effectue de la même manière dans tous les algorithmes ; la complexité de l'algorithme dépend principalement de comment est effectuée la recherche des motifs fréquents. Dans la littérature, il existe de nombreux algorithmes pour générer les motifs fréquents. Ils sont des variantes de l'algorithme Apriori et reposent sur la double propriété d'anti-monotonie du support : (i) tout sous-ensemble d'un motif fréquent est fréquent ; (ii) tout sur-ensemble d'un motif non fréquent est non fréquent. En revanche, ils diffèrent, entre autres, selon la stratégie de recherche adoptée : (1) une recherche descendante (*Top-Down*) qui consiste à générer les motifs fréquents en partant des grands motifs vers les plus petits en utilisant la propriété que tout sous-ensemble d'un motif fréquent est fréquent ; (2) une recherche ascendante (*Bottom-Up*) qui génère les motifs fréquents en partant des petits motifs

vers les motifs les plus grands en utilisant la propriété que tout sur-ensemble d'un motif non fréquent est non fréquent.

Dans le cadre d'un contexte OLAP, un cube de données est souvent "creux" avec des données éparses. Dans ce cas, il y aura peu de motifs fréquents et les motifs non fréquents ont plus de chance d'être plus nombreux que les motifs fréquents. La deuxième propriété d'anti-monotonie est alors plus efficace et réduit considérablement l'espace de recherche. Dans ce cas, il est préférable d'adopter une stratégie ascendante lors de la recherche des motifs fréquents, stratégie qui élimine dès le départ les petits motifs non fréquents.

Dans notre approche, nous adoptons une stratégie ascendante, nous utilisons l'algorithme Apriori en l'adaptant aux données multidimensionnelles et nous y incluons notre définition du support et de la confiance. La nouvelle définition du support conserve la propriété d'anti-monotonie.

Soit un cube de données $\mathcal{C}$ avec ses dimensions réparties en sous-ensembles $(\mathcal{D}_\mathcal{C}, \mathcal{D}_\mathcal{A}, \mathcal{D}_\mathcal{I})$, une méta-règle inter-dimensionnelle $\mathcal{R}$, une mesure M, un support minimum *minsupp* et une confiance minimale *minconf*.

Soit k, l'indice de l'itération en cours de l'algorithme. Un k-itemset candidat, noté $C(k)$, est un motif de longueur k ($k \geq 1$) c'est-à-dire une conjonction de k prédicats (items). Un k-itemset fréquent, noté $F(k)$, est un motif de longueur k dont le support est supérieur ou égal au support minimum (*minsupp*).

Après une étape d'initialisation, l'extraction des motifs fréquents est itérative avec trois étapes par itération. La figure 2.3 illustre le processus de recherche de règle d'association sur notre exemple.

Etape d'initialisation. Les 1-itemsets candidats $C(1)$ sont initialisés à partir des dimensions d'analyse $\mathcal{D}_\mathcal{A}$. Chaque 1-itemset représente une instance d'un prédicat dimensionnel x_k ou y_k de la méta-règle. Les éléments de $C(1)$ sont les modalités des ensembles $\mathcal{A}^{ij}$.

Pour chaque itération k, si $C(k) \neq \emptyset$ et $k \leq (s + r)$, la **première étape** cherche les k-itemsets fréquents $F(k)$ à partir des k-itemsets candidats $C(k)$.

La **deuxième étape** consiste à extraire les règles d'association à partir des k-itemsets fréquents $F(k)$. Pour générer une règle à partir d'un k-itemset fréquent $F(k)$, l'algorithme cherche dans $F(k)$, un sous-ensemble d'items A et un sous-ensemble d'items B non vides, tels la règle $A \Rightarrow B$ est conforme à la méta-règle inter-dimensionnelle $\mathcal{R}$ et a une confiance supérieure ou égale à la confiance minimale *minconf*.

La **troisième étape** génère à partir des k-itemsets fréquents $F(k)$ un nouvel ensemble de $(k+1)$-itemsets candidats $C(k+1)$. Un $(k+1)$-itemset candidat est généré selon l'union de deux k-itemsets fréquents A et B de $F(k)$, où A et B vérifient trois conditions : (1) A et B doivent partager $(i-1)$ éléments communs ; (2) tout sous-itemset non vide de $A \cup B$ doit être une instance d'un prédicat inter-dimensionnel dans $\mathcal{D}_{\mathcal{A}}$; et (3) tout sous-itemset non vide de $A \cup B$ doit être fréquent.

Enfin, l'algorithme s'**arrête** quand l'une des deux conditions suivantes est vérifiée : (1) il y a eu $(s+r)$ itérations, où $(s+r)$ correspond au nombre de dimensions d'analyse $\mathcal{D}_{\mathcal{A}}$; (2) il n'y a plus de motif candidat.

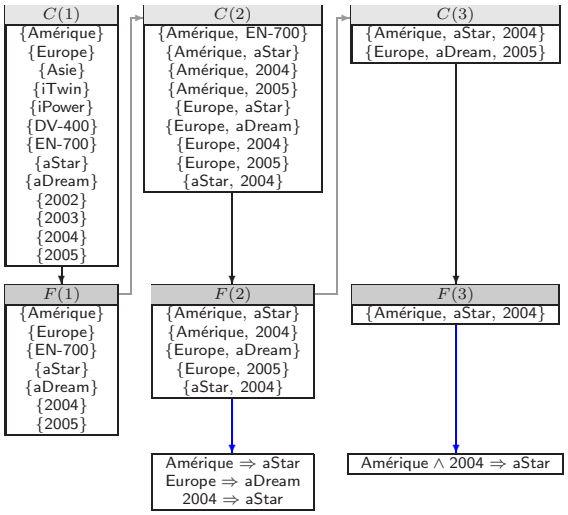


FIGURE 2.3 – Exemple de recherche ascendante de règles d'association inter-dimensionnelles

Qualité des règles. Lors de la génération, le support et la confiance sont les indicateurs d'évaluation des règles mais ils conduisent souvent à un trop grand nombre de règles dont la plupart sont inintéressantes. D'autres indicateurs de la qualité des règles sont donc nécessaires pour limiter le nombre de règles. Dans la littérature scientifique, sont distinguées deux grandes catégories d'indicateurs d'évaluation des règles d'association : *les indices descriptifs* et *les indices statistiques* ; le choix des indicateurs à utiliser dépend de l'utilisateur et de la structure des données [Bla05, LVL05, LVL06].

Dans notre contexte OLAP, les cubes peuvent contenir un grand nombre de faits, nous choisissons donc d'utiliser deux indices descriptifs : le *Lift* [BMS97] et l'indice de

Loevinger [Loe47]. Un indice descriptif est insensible au nombre de données (faits) alors que les indices statistiques dépendent fortement de la taille des données étudiées. Quand le nombre de données est très important, un indice statistique perd son pouvoir discriminant et tend artificiellement vers 1. De plus, un indice statistique fait référence à un modèle probabiliste, la loi de probabilité des données devant être connue. Ceci est rarement le cas dans un contexte d'analyse en ligne. Dans notre approche, cela nous amène à écarter les indices statistiques et à préférer des indices descriptifs. D'où notre choix. De plus, pour garder le même raisonnement que dans la nouvelle définition du support et de la confiance, nous introduisons dans le calcul du *Lift* et l'indice de *Loevinger* la mesure à la place des fréquences.

Soit une règle R de la forme $X \Rightarrow Y$, avec $X = (x_1 \wedge \cdots \wedge x_s)$ et $Y = (y_1 \wedge \cdots \wedge y_r)$ des conjonctions de prédicats dimensionnels.

P_X (respectivement P_Y, P_{XY}) est la proportion de l'agrégation SUM de la mesure M des faits vérifiant X (respectivement Y, X et Y) dans le sous-cube de données.

$P_{Y/X}$ est la proportion de l'agrégation SUM de la mesure M des faits (dans le sous-cube) vérifiant Y sachant qu'ils vérifient déjà X.

$P_{\overline{X}} = 1 - P_X$ (respectivement $P_{\overline{Y}} = 1 - P_Y$) est la proportion de l'agrégation SUM de la mesure M des faits ne vérifiant pas X (respectivement Y) dans le sous-cube.

On peut écrire :

$$P_X = \frac{M(x_1,...,x_s,\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}{M(\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}$$

$$P_Y = \frac{M(\mathsf{All},...,\mathsf{All},y_1,...,y_r,\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}{M(\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}$$

$$\textsc{Supp}(R) = P_{XY} = \frac{M(x_1,...,x_s,y_1,...,y_r,\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}{M(\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}$$

$$\textsc{Conf}(R) = P_{Y/X} = \frac{M(x_1,...,x_s,y_1,...,y_r,\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}{M(x_1,...,x_s,\mathsf{All},...,\mathsf{All},\Theta_1,...,\Theta_p,\mathsf{All},...,\mathsf{All})}$$

Le *Lift* d'une règle d'association représente l'écart à l'indépendance mesurée par le produit $P_X P_Y$:

$$\textsc{Lift}(R) = \frac{P_{XY}}{P_X P_Y} = \frac{\textsc{Supp}(R)}{P_X P_Y}$$

Classiquement, le *Lift* s'interprète comme le quotient du nombre d'exemples observés par celui attendu sous l'hypothèse d'indépendance entre X et Y. Par exemple, une règle ayant un *Lift* égal à deux signifie que le nombre d'exemples de la règle est deux fois plus grand que celui attendu sous l'hypothèse d'indépendance, ce qui implique que si l'antécédent X se réalise alors il y a deux fois plus de chance que Y se réalise. Dans le cadre de notre approche, dans le contexte (Etudiant, Femme), si la règle inter-dimensionnelle Continent=Amérique $\wedge$ Année=2004 $\Rightarrow$ Article=aStar a un *Lift* égal à deux, ceci signifie que les produits achetés par les étudiantes en 2004 sur le continent américain ont deux fois plus de chance d'être des lecteurs MP3 aStar que

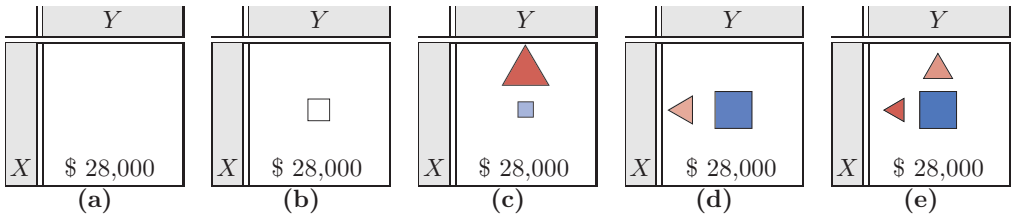


FIGURE 2.4 – Exemples de visualisation de règles d'un cube

dans le cas général. Dans cet exemple précis, le *Lift* peut aussi s'interpréter comme le coefficient multiplicateur de la part du marché des lecteurs MP3 aStar due à une clientèle d'étudiantes sur le continent américain en 2004.

L'indice de *Loevinger* permet d'améliorer l'interprétation de la confiance d'une règle par normalisation de cette dernière. Il normalise la confiance centrée de la règle par rapport aux exemples ne vérifiant pas le conséquent de la règle :

$$\text{LOEV}(R) = \frac{P_{Y/X} - P_Y}{P_{\overline{Y}}} = \frac{\text{CONF}(R) - P_Y}{P_{\overline{Y}}}$$

Visualisation des règles inter-dimensionnelles. Une fois extraites, nous proposons à l'utilisateur OLAP de visualiser les règles dans le cube. Cette visualisation est un moyen de rendre encore plus exploitables les connaissances induites par la méthode de fouille et répond bien au principe de visualisation de l'OLAP. Nous utilisons les principes de la sémiologie graphique de Bertin [Ber99]. Ces derniers consistent à représenter une ou plusieurs informations en faisant varier des éléments visuels comme la *position*, la *taille*, la *luminosité*, la *texture*, la *couleur*, l'*orientation* et la *forme*.

Une règle d'association $X \Rightarrow Y$ est représentée selon un codage graphique qui tient compte de son support, de se confiance, du *Lift* et l'indice de *Loevinger*. Le motif $\{X, Y\}$ est codé par un carré bleu alors que l'implication $X \Rightarrow Y$ est codée par un triangle rouge. Le support du motif $\{X, Y\}$ est codé par la surface du carré et la confiance de la règle $X \Rightarrow Y$ est codée par la surface du triangle. Plus le support (respectivement, la confiance) est grand, plus la surface du carré (respectivement, du triangle) est grande. Le *Lift* et l'indice de *Loevinger* sont représentés par l'intensité de la couleur de la forme : plus la couleur est foncée, plus la valeur de l'indice est grande. Le *Lift* est codé par l'intensité du carré bleu, l'indice de *Loevinger* est codé par l'intensité du triangle rouge.

Selon les propriétés du motif $\{X, Y\}$, il y a un code graphique différent dans la

cellule (figure 2.4) :

- si $\{X, Y\}$ n'est pas fréquent, seule la valeur de la mesure M, si elle existe, est affichée ;
- si $\{X, Y\}$ est fréquent et ne génère pas de règle d'association, un carré blanc est affiché. La surface du carré est proportionnelle au support du motif ;
- si $\{X, Y\}$ est fréquent et génère la règle d'association $X \Rightarrow Y$, un carré bleu et un triangle isocèle rouge sont affichés. Le triangle pointe vers Y selon le sens d'implication de la règle (et réciproquement pour la règle $Y \Rightarrow X$) ;
- si $\{X, Y\}$ est fréquent et génère les deux règles d'association $X \Rightarrow Y$ et $Y \Rightarrow X$, un carré bleu et deux triangles isocèles rouges sont affichés, chacun symbolisant une implication.

2.3.5 Expérimentation

Afin d'évaluer les performances de notre algorithme d'extraction de règles d'association inter-dimensionnelles, nous avons mené un ensemble d'expériences.

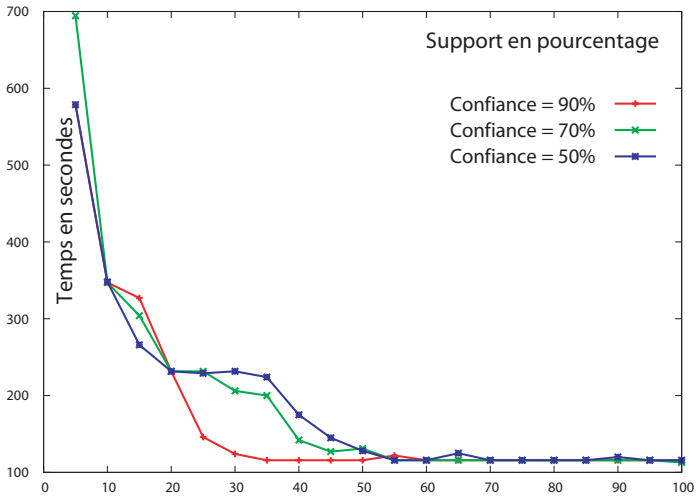


FIGURE 2.5 – Temps d'exécution de l'algorithme en fonction du support minimum pour différentes confiances minimales

La figure 2.5 montre le temps d'exécution de notre algorithme en fonction du support minimum selon différents seuils de confiance minimale. On remarque que le temps d'exécution de l'algorithme décroît, en général, en fonction du support minimum. Grâce à la propriété d'anti-monotonie, quand le support minimum augmente l'algorithme devient plus rapide (élagage des motifs non fréquents dès les premières itérations de l'algorithme). On remarque aussi que plus la confiance

minimale est élevée plus le temps d'exécution de l'algorithme baisse quand le support minimum est faible.

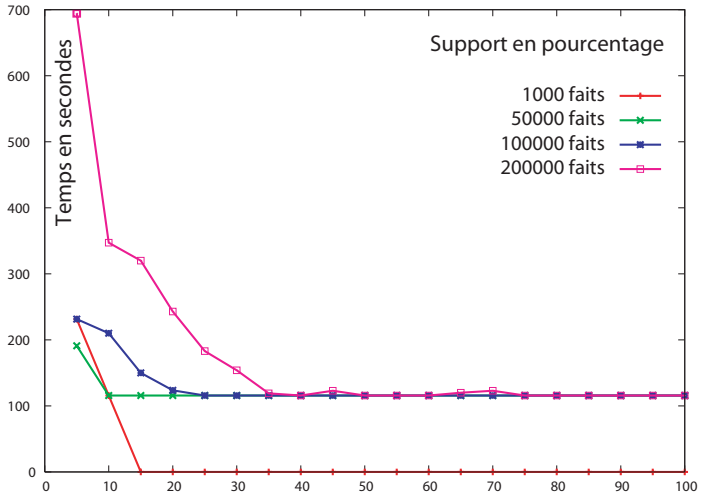


FIGURE 2.6 – Temps d'exécution de l'algorithme en fonction du support minimum pour différentes tailles de cube

La figure 2.6 résume des tests de performance de l'algorithme pour des cubes de données de taille différente en fonction du support minimum. On remarque que pour les petites valeurs du support minimum (moins de 40%), le nombre de faits dans le cube est un élément déterminant pour la rapidité de l'algorithme. En revanche, pour les grandes valeurs du support minimum, le nombre de faits n'a pratiquement aucune influence sur le temps d'exécution de l'algorithme.

2.3.6 Perspectives

Les premiers résultats sont encourageants et montrent l'intérêt d'utiliser une recherche guidée de règles d'association pour expliquer des phénomènes observés dans un cube. Mais des améliorations et de nouvelles pistes de recherche méritent d'être étudiées.

L'algorithme Apriori n'est pas forcément le plus performant. Il faut envisager l'utilisation d'un autre algorithme comme par exemple avec la génération des motifs fermés fréquents.

Les règles extraites ne portent que sur les faits détaillés non pas sur les agrégats.

On pourrait imaginer de générer les règles à partir des agrégats.

Nous pensons qu'il pourrait également être intéressant d'intégrer la mesure dans la règle inter-dimensionnelle en tant que prédicat dimensionnel. La mesure pourrait aussi faire l'objet d'un codage graphique intégré dans celui de la règle et ainsi offrir à l'utilisateur une visualisation complète des faits et des liens entre ces faits.

Le codage graphique des règles dans le cube est une façon de visualiser la connaissance extraite par la méthode de fouille. Mais cela n'est pas facile de proposer un codage graphique simple et non surchargé. De plus, le codage que nous proposons suppose que l'utilisateur connaisse le principe des règles d'association. La visualisation dans le cube des connaissances extraites est un problème ouvert qui mérite d'explorer d'autres possibilités.

Fort de cette tentative d'extension des capacités de l'OLAP, nous poursuivons notre réflexion avec la volonté d'étendre l'OLAP à la prédiction en ligne.

2.4 Prédiction en ligne

2.4.1 Positionnement et principe

Pour motiver et introduire la 2ème extension de l'OLAP que nous proposons, nous prenons un nouvel exemple.

Pour étudier les résultats d'élèves à un examen, un utilisateur peut construire un cube de données à trois dimensions, Filières, Sexe et Matières, et prendre pour mesure la moyenne des notes des élèves. Selon les filières suivies, toutes les matières ne sont pas enseignées. Il y a donc des cellules vides. L'utilisateur peut se demander quelle serait la note moyenne de certains élèves s'ils suivaient telle ou telle matière. Il connaît les modalités des dimensions des cellules vides et les valeurs de la mesure des cellules voisines. Il peut donc se demander, comment, à partir des cellules pleines voisines, affecter une valeur à la cellule vide. Les opérateurs OLAP classiques ne permettent pas d'assister l'utilisateur dans cette tâche. Pour répondre à sa question, l'utilisateur a besoin d'une nouvelle possibilité d'analyse : la prédiction.

Pour apporter une solution, nous nous plaçons dans le cadre du *What If Analysis*, défini par Golfarelli *et al.* [GRP06] sur la base des travaux de Philippakis [Phi88]. Ils soulignent que les entrepôts de données permettent d'analyser les données du passé mais qu'ils ne sont pas adaptés pour anticiper les événements du futur. Pour répondre à ce besoin, ils introduisent le *What If Analysis*, une démarche centrée sur

l'utilisateur, qui mesure comment des changements dans un ensemble de variables indépendantes ont un impact sur un ensemble de variables dépendantes. De plus, les auteurs soulignent que peu d'outils commerciaux permettent de faire du *What If Analysis*. Sans proposer une réelle méthode pour faire de la prédiction dans les cubes OLAP, les auteurs proposent une méthodologie pour faire des analyses de type *What If Analysis*. Dans [GR08], ils introduisent YAM^2 UML un formalisme basé sur une extension de UML pour modéliser l'entrepôt et les cubes de données mais aussi pour exprimer le modèle de simulation c'est-à-dire le modèle de prédiction.

Dans ce contexte, nous pensons que le couplage entre l'OLAP et la fouille de données peut permettre de prédire la valeur de la mesure pour des faits inexistants ou des faits avec une valeur manquante.

Quelques équipes dans le monde travaillent ou ont travaillé sur ce problème. Sans vouloir être exhaustive, nous pouvons citer l'équipe de Han avec le logiciel DBMiner et les travaux de S. Cheng [Che98] ; l'équipe de Sarawagi [SAM98] et les travaux connexes de B.C. Chen [CCLR05, CRST06] ; l'équipe de Y. Chen et J. Pei [CDH$^+$06, HPDW01, HWD$^+$02] et l'équipe de Palpanas [PK01, PKM05]. Dans la plupart des travaux, l'objectif est prédire la mesure existante mais certains créent un nouveau cube alors que d'autres complètent le cube existant. Deux équipes utilisent le principe de la prédiction mais pour atteindre d'autres objectifs : Sarawagi *et al.* cherchent à guider l'utilisateur dans l'exploration du cube en détectant des valeurs exceptionnelles de la mesure [SAM98] alors que Palpanas *et al.* utilisent la prédiction pour prédire les faits détaillés d'un cube à partir des faits agrégés [PK01].

Dans tous les travaux, un des défis est la quantité des données due à la taille du cube et aux hiérarchies. De plus, le modèle doit pouvoir être obtenu pour les différents niveaux de granularité selon les besoins de l'utilisateur. Une réduction des données et une optimisation des algorithmes sont souvent envisagées dans les différents travaux. Une démarche exhaustive serait de calculer la prédiction de la valeur de la cellule pour tous les cubes possibles définis par toutes les combinaisons possibles des différents niveaux hiérarchiques. Compte tenu des temps de calculs prohibitifs, Sarawagi *et al.* proposent de calculer le modèle de prédiction à un niveau en combinant les modèles des niveaux inférieurs. S. Cheng a également réfléchi à comment adapter le modèle de prédiction quand l'utilisateur fait une opération de navigation comme un *drill-down* ou un *roll-up* et il propose une solution alternative pour re-calculer le modèle de prédiction [Che98]. B.C. Chen *et al.* introduisent la notion de *Bellwether Analysis* [CRST06]. Ils segmentent le cube en sous-ensembles c'est-à-dire en régions. Pour chaque région est construit un modèle de prédiction de la mesure. Quand on veut prédire la mesure pour un nouveau fait, l'algorithme cherche, selon un critère de coût, la meilleure région pour faire la prédiction. Cette région est appelée une région *bellwether*. Pour éviter une explosion combinatoire du nombre de régions candidates,

les auteurs ont un algorithme efficace de recherche et d'élagage des régions et ils utilisent un arbre des régions (*bellwether tree*).

Pour pouvoir parler d'intégration de la prédiction dans l'OLAP, il faut pouvoir proposer un modèle de prédiction, l'interpréter et l'associer à la sémantique OLAP. L'utilisateur doit disposer d'indicateurs sur la qualité de la prédiction et doit pouvoir interpréter le modèle de prédiction avec des résultats lisibles et exploitables et ce sans avoir nécessairement de connaissances théoriques sur le modèle. Y. Chen et J. Pei introduisent une nouvelle fonction d'agrégation basée sur la régression et l'utilisent pour compresser le cube de données [CDH+06]. En partant du cube de données le plus détaillé possible, les faits de chaque cellule sont condensés par une équation linéaire. Chaque cellule est résumée par un modèle de régression linéaire. Cette équipe de recherche a largement contribué à la problématique de compression des cubes, mais dans un contexte de prédiction et d'analyse en ligne, on peut regretter qu'il soit difficile à l'utilisateur d'interpréter le contenu des cellules.

La plupart des travaux, utilisent comme méthode de prédiction un modèle de régression (linéaire, log-linéaire, logistique, ...) qui suppose des hypothèses sur les données, comme l'homoscédasticité, la normalité, ..., hypothèses rarement vérifiées ou non vérifiables dans un contexte OLAP. Dans le cas d'une vraie intégration de la fouille dans l'OLAP, le processus classique d'apprentissage doit être respecté avec les phases de pré-traitement des données, de sélection des variables explicatives, de construction du modèle et de validation. Peu de travaux réalisent une phase de validation du modèle.

La particularité de nos travaux est de proposer une démarche qui respecte à la fois les besoins de l'utilisateur, qui reprend la méthodologie des méthodes de fouille de données et qui respecte la philosophie de l'analyse en ligne.

Les objectifs de notre démarche sont de :
- permettre à l'utilisateur de prédire la valeur d'une mesure pour un nouveau fait et de compléter ainsi le cube,
- placer l'utilisateur OLAP au coeur du processus de prédiction en lui donnant des indicateurs sur la qualité de la prédiction et un modèle de prédiction interprétable dans un environnement OLAP,
- fournir des résultats qui ne dépendent pas d'hypothèse et qui ne nécessitent pas de connaissances spécifiques pour être interprétés,
- intégrer un processus d'apprentissage automatique complet dans l'analyse OLAP.

Notre approche est basée sur une démarche d'apprentissage automatique (*machine learning*) et utilise les arbres de régression. Cette approche se concrétise par l'opérateur OPReT (*Online Prediction by Regression Tree*).

Nous nous démarquons des travaux de [Che98], [SAM98], [CRST06], [CDH+06]

où l'utilisateur doit connaître les principes parfois complexes des modèles de régression. Dans notre approche, l'utilisateur n'a pas besoin d'avoir des connaissances approfondies sur les arbres de régression.

A l'inverse de [Che98], [CRST06], [CDH+06], nous ne souhaitons pas utiliser une méthode qui nécessite des hypothèses sur les données. Les arbres de régression ne supposent aucune hypothèse.

De plus, les arbres de régression fournissent des résultats validés et des indicateurs pour que l'utilisateur mesure la qualité des valeurs prédites. Ainsi notre approche intègre un processus complet d'apprentissage avec une phase de sélection des variables explicatives, une phase d'apprentissage et une phase de validation, phase que l'on trouve rarement dans les différents travaux. Nous nous plaçons dans le cas d'un cube de données volumineux, où l'utilisateur définit un contexte d'analyse dans le cube. Le contexte d'analyse correspond à un sous-cube avec les niveaux hiérarchiques et les dimensions sur lesquels l'utilisateur souhaite prédire de nouveaux faits. Dans un premier temps, pour simplifier l'approche, notre proposition considère que le cube comporte une seule mesure.

2.4.2 Démarche

Dans un cube de données, les dimensions sont souvent des variables qualitatives et les faits sont généralement mesurés par des variables quantitatives continues. Parmi les méthodes de prédiction, les arbres de régression répondent à ces caractéristiques et ils n'ont pas besoin d'hypothèses statistiques sur les données. Les arbres de régression proposent, par une discrimination des variables explicatives, une prédiction pour des cellules vides. De plus, ils fournissent des résultats compréhensibles et se fondent sur des techniques ne s'apparentant pas à une boite noire pour l'utilisateur.

Pour illustrer notre démarche, nous utilisons l'exemple très simple d'un cube avec trois dimensions, *Sexe* (f, m), *Filières* (S, ES, L), *Matières* (A, B, C, D, E) et avec comme mesure la moyenne des notes. On considère que l'utilisateur veuille prédire la mesure de quatre cellules vides (celles qui sont en grisé dans la figure 2.7(b)). Nous reviendrons sur les figures 2.7(b) et (c) lors du déroulement de la démarche.

Sous-cube de données. Le point de départ de notre démarche est un contexte d'analyse défini par l'utilisateur. A partir du cube original $\mathcal{C}$, l'utilisateur construit $\mathcal{C}'$, le sous-cube $(\Theta^1, ..., \Theta^p)$ dans lequel il veut travailler. Soit n le nombre de faits détaillés ayant une valeur pour la mesure quantitative M dans le sous-cube $\mathcal{C}'$. Dans notre exemple de la figure 2.7(a), le sous-cube $\mathcal{C}'$ est défini par $(\Theta^1, \Theta^2, \Theta^3)$=({*S, ES, L*},{a, b, c, d, e},{*m, f*}).

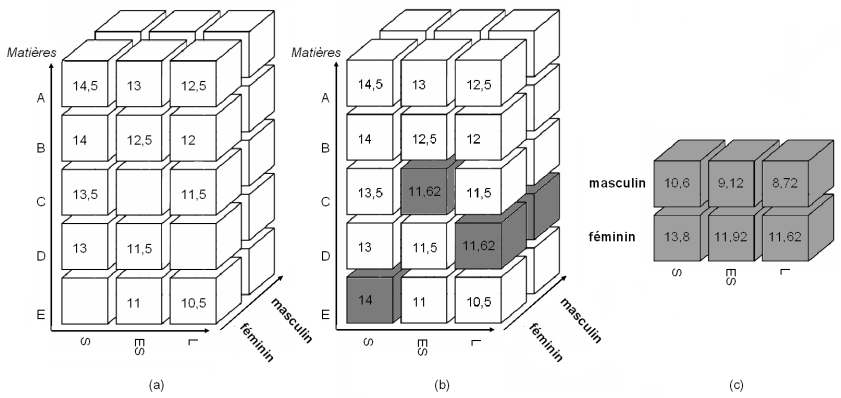


FIGURE 2.7 – Prédiction dans un cube OLAP.

Construction et validation du modèle de prédiction. En utilisant le vocabulaire des méthodes d'apprentissage supervisé, les dimensions jouent le rôle de variables explicatives et la mesure M correspond à la variable à prédire. Les n faits détaillés du sous-cube $\mathcal{C}'$ sont segmentés aléatoirement en échantillon d'apprentissage et en échantillon test : 70% des faits servent à la phase d'apprentissage pendant laquelle le modèle est construit, 30% des faits sont réservés à l'évaluation du modèle lors de la phase de test.

Pendant la phase d'apprentissage, la construction d'un arbre de régression se fait selon un principe de partitionnement récursif. A chaque étape, l'algorithme cherche la meilleure variable explicative pour prédire la variable à expliquer. Pour cela, les valeurs ou modalités des variables explicatives sont regroupées en deux ou k sous-ensembles. La prédiction associée à un sous-ensemble (ou sommet de l'arbre) est la moyenne des observations appartenant au regroupement et l'homogénéité d'un regroupement est mesurée par la variance de la variable à prédire dans le sommet. Au moment de segmenter un sommet en deux ou k sous-ensembles, on cherche à minimiser la variance intra-groupe ou à maximiser la variance inter-groupe. L'opération est réitérée séparément dans chaque sous-ensemble ou sommet ainsi constitué.

Les critères d'évaluation d'un arbre de régression sont le taux d'erreur moyen et la réduction de l'erreur. Le taux d'erreur indique en moyenne l'écart entre la valeur observée et la vraie valeur de la variable à prédire. Si le taux d'erreur est proche de 0 alors cela signifie que le modèle de prédiction (l'arbre) ne fait pas d'erreur systématique. La réduction de l'erreur ($1-R^2$, avec R^2 le coefficient de détermination qui mesure la proportion de variance expliquée par le modèle c'est-à-dire la qualité de la régression) indique si l'arbre prédit mieux que le modèle par défaut (l'arbre réduit à sa racine) où on utiliserait uniquement la moyenne de la mesure pour prédire

les valeurs de la mesure. Le modèle donne une prédiction parfaite si la réduction de l'erreur est égale à 0. Si le modèle construit est pire que le modèle par défaut alors le coefficient R^2 devient négatif et la réduction de l'erreur est supérieure à un. Dans ce cas, les variables explicatives (dimensions) ne permettent pas de prédire la mesure.

Une fois l'arbre de régression construit lors de la phase d'apprentissage, le modèle doit être évalué lors d'une phase de test. L'arbre d'apprentissage est utilisé sur l'échantillon test afin de prédire la valeur de la mesure pour chaque observation ou fait. Si l'erreur moyenne et la réduction de l'erreur en phase de test sont faibles et proches de celles obtenues en phase d'apprentissage, alors le modèle est validé.

Parmi les méthodes permettant de construire un arbre de régression, les deux techniques les plus connues sont CART (*Classification and Regression Tree*) [BFOS84] et AID (*Automatic Interaction Detection*) [MS63]. Les principales différences entre AID et CART se situent dans la phase d'apprentissage lors de la construction de l'arbre : (1) AID ne construit pas forcément des arbres binaires. A l'aide du test de Fisher de comparaison des variances de groupes, les valeurs des variables explicatives sont regroupées par deux ou plus ou sont laissées seules. L'arbre ainsi construit est k-aire. La difficulté est alors de conserver des sommets avec suffisamment d'observations. Dans CART, l'arbre est forcément binaire, à chaque partitionnement, les valeurs ou modalités des variables explicatives sont systématiquement regroupées en deux sous-ensembles. (2) Dans CART, la phase d'apprentissage se fait en deux étapes, l'échantillon d'apprentissage est alors lui même divisé en deux : *growing set* et *pruning set*. Une première étape, appelée en expansion, maximise l'homogénéité des groupes sur le sous-ensemble *growing set*. La seconde est une étape d'élagage de l'arbre et consiste à minimiser l'erreur de prédiction sur le sous ensemble appelé *pruning set*. AID ne possède pas d'étape d'élagage pendant la phase d'apprentissage. En revanche, lors de la construction de l'arbre, est utilisé un critère d'arrêt, critère basé sur la notion de probabilité critique.

Dans un premier temps, nous laissons à l'utilisateur le choix d'utiliser CART ou AID. Par la suite, nous comptons réaliser des expériences avec les deux méthodes pour voir si l'une donne systématiquement des meilleurs résultats que l'autre.

Dans notre exemple, nous utilisons CART pour construire l'arbre de regression. La moyenne de l'erreur est de 0,243 ce qui est acceptable.

Interprétation du modèle prédictif. Après la construction et la validation du modèle, l'arbre de régression retourne λ règles de décision ($\lambda \geq 0$). L'ensemble des règles d'un modèle est noté $\mathcal{R} = \{R_1, R_2, ..., R_t, ..., R_\lambda\}$.

Soit $R_t(X \Rightarrow Y; S; \sigma)$ une règle de décision $\in \mathcal{R}$. X est une conjonction et/ou une

disjonction de modalités correspondant aux antécédents de la règle. Y est la valeur moyenne prédite pour la mesure M sachant X. S est le support de X et σ est l'écart type de la mesure M dans l'ensemble d'apprentissage vérifiant X.

En plus des deux critères de fiabilité du modèle que sont la moyenne et la réduction de l'erreur, deux autres indicateurs permettent d'évaluer la qualité d'une règle. Le premier est le support ou l'effectif relatif S des faits qui supportent la règle. Le deuxième est l'écart type σ de M, qui indique l'homogénéité des faits supportant la règle. Plus l'écart type est faible, plus les faits supportant la règle sont homogènes et la règle fiable.

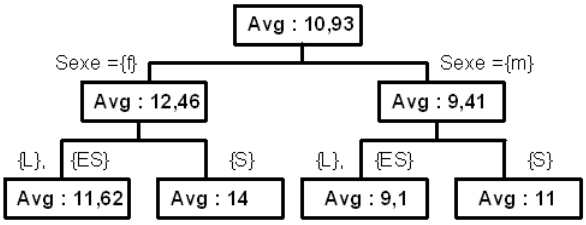


FIGURE 2.8 – Arbre de régression

Dans notre exemple, nous obtenons l'arbre de régression de la figure 2.8 et les règles suivantes :

- $R_1(m \wedge (ES \vee L) \Rightarrow 9,1 \ ; \ 33\% \ ; \ 0,83)$
- $R_2(m \wedge S \Rightarrow 11 \ ; \ 17\% \ ; \ 0,91)$
- $R_3(f \wedge (ES \vee L) \Rightarrow 11,62 \ ; \ 33\% \ ; \ 0,84)$
- $R_4(f \wedge S \Rightarrow 14 \ ; \ 17\% \ ; \ 0,64)$

Chaque règle correspond à une feuille terminale de l'arbre. La règle R_1 indique que si l'élève est de sexe masculin et qu'il est en filière ES ou L alors sa moyenne sera de 9,1. Cette règle repose sur 33% des faits présents dans la base d'apprentissage. L'écart type est de 0,83. Le sexe et la filière sont les variables les plus discriminantes. Elles sont explicatives des résultats des élèves, contrairement aux matières qui ne sont pas déterminantes.

Exploitation OLAP du modèle prédictif. Dans le sous-cube $\mathcal{C}'$, l'utilisateur désigne la ou les cellules vides qu'il souhaite prédire. Les règles adaptées aux cellules sont sélectionnées et appliquées pour prédire la valeur de la mesure dans ces cellules.

Soit c la cellule désignée par l'utilisateur $c = (\theta^1, ..., \theta^i, ..., \theta^p)$ où chaque θ^i est un singleton contenant une seule modalité pour la dimension à laquelle il est rattaché. On note $M(c)$ la valeur de la mesure M que prend la cellule c.

Pour chaque cellule c désignée par l'utilisateur, telle que $M(c) = Null$ (c'est-à-dire que la cellule est vide), on recherche la règle $R_t \subset \mathcal{R}$ telle que son antécédent X a l'ensemble de ses modalités inclus dans l'ensemble des modalités décrivant la cellule c. Il s'agit donc de comparer l'ensemble des modalités décrivant la cellule aux antécédents X des règles de l'arbre de régression. Si $X \subset (\theta^1, ..., \theta^p)$ alors, la valeur moyenne Y de la règle peut être affectée comme valeur de la mesure de la cellule c. On note $M(c) \longleftarrow Y$.

L'opération est réitérée pour chaque cellule désignée par l'utilisateur pour la prédiction.

Sur la figure 2.7(b), nous avons prédit la valeur de la mesure pour quatre cellules vides. Par exemple, lorsque nous avons ciblé la cellule décrite par les modalités (f, S, E) pour les dimensions, respectivement, *Sexe*, *Filière* et *Matières*, la règle $R_4(f \wedge S \Rightarrow 14\ ;\ 17\%\ ;\ 0,64)$ a été sélectionnée, $(f \wedge S) \subset (f, S, E)$. Nous constatons que les élèves de sexe féminin en filière S auront en moyenne 14/20 dans la matière E. Pour un autre exemple, en terme de requête de type *What If Analysis*, l'arbre de régression permet de savoir quelle serait la moyenne des élèves de filière littéraire selon leur sexe, si l'on ouvrait une nouvelle matière, D. Ainsi nous voyons que les filles en filière littéraire auraient 11,62 de moyenne pour cette nouvelle matière D.

Cette intégration de la prédiction dans l'OLAP permet aussi à l'utilisateur d'appréhender les valeurs prévues des agrégats pour un niveau hiérarchique supérieur. Les agrégats sont recalculés en considérant les valeurs prédites. Par exemple en faisant un *Roll-up* total sur les matières, la moyenne des notes est exprimée selon le sexe et la filière. Dans la figure 2.7(c) on obtient comme résultat par exemple que la moyenne prévue pour l'ensemble des matières du tronc commun, pour les garçons de la filière L, est de 8,72/20. Cette moyenne tient compte de la note prédite pour les garçons si l'on ouvre la matière D à l'enseignement en filière L.

Une extension envisagée est d'utiliser des indicateurs visuels pour l'utilisateur. Sur les figures 2.7(b) et 2.7(c), nous utilisons une nuance de gris pour une valeur prédite ou un agrégat recalculé à partir des valeurs prédites. Nous pensons que selon les critères de qualité d'une règle (support et écart type), nous pouvons nuancer ce code de couleur. Ainsi l'utilisateur peut directement interpréter les prédictions au sein du cube de données.

2.4.3 Expérimentation

Pour évaluer notre travail, nous avons mené une série d'expérimentations sur des données réelles. Nous utilisons un jeu de données médicales relatifs au dépistage du cancer du sein. Les données sont extraites de la base DDSM (*Digital Database*

for Screening Mammography[1]). La base DDSM est une ressource libre mise à disposition des communautés de chercheurs dans le domaine médical et dans le domaine de la fouille de données. La base DDSM contient 2 604 dossiers où chaque dossier correspond à une patiente. Un dossier est constitué d'un ensemble d'images (mammographies) et de fichiers textes (informations d'ordre général sur la patiente et annotations des mammographies). Dans le cas d'une application médicale, l'objectif peut être de prévenir le cancer du sein et nous pouvons imaginer qu'un cancérologue cherche à étudier les régions suspectes dans une mammographie. Ces régions peuvent correspondre à des tumeurs cancéreuses ou non. Une région suspecte est décrite par un ensemble de propriétés médicales et spatiales (position dans la mammographie, frontière, ...). Dans [BBLR06a], est proposée une modélisation multidimensionnelle des régions suspectes selon un schéma en étoile. Dans l'entrepôt issu de ce modèle, il y a 87 399 faits où chaque fait correspond à une région avec les propriétés de la région et les données de la patiente à laquelle elle appartient (figure 2.9).

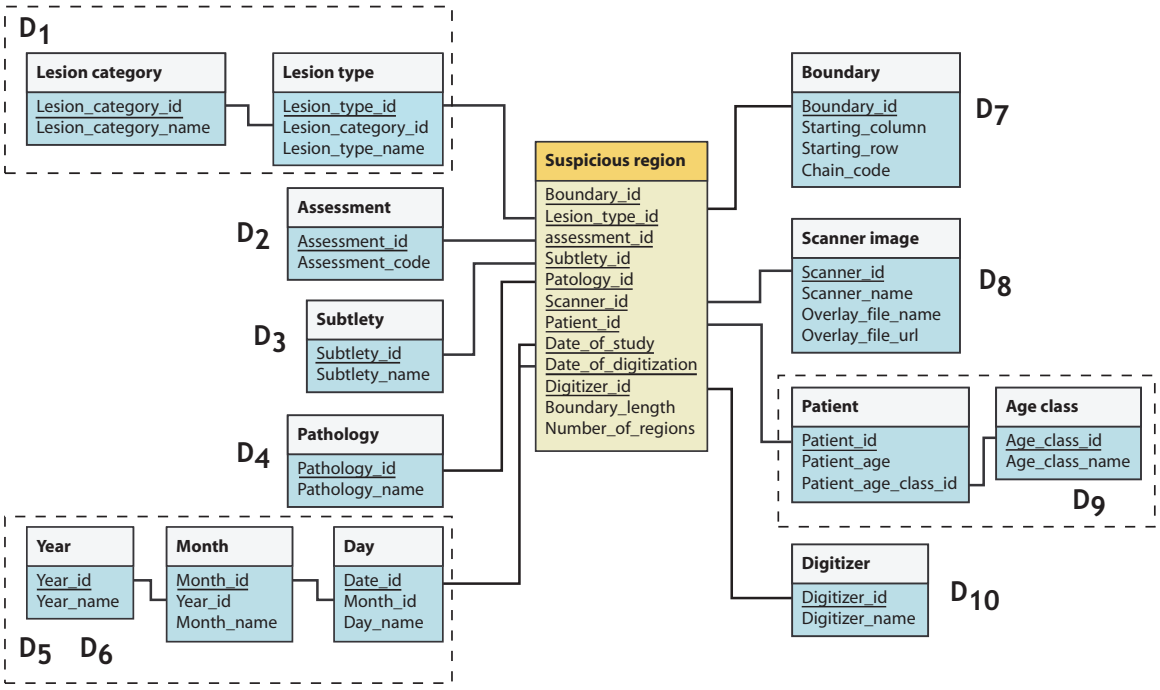


FIGURE 2.9 – Modélisation multidimensionnelle des régions suspectes

Contexte d'analyse. Parmi les dimensions et les niveaux hiérarchiques existants, nous retenons l'indice d'évaluation de la lésion et l'indice de subtilité de la lésion. Nous

1. http ://marathon.csee.usf.edu/Mammography/Database.html

retenons aussi le type de pathologie et le type de lésion présents chez la patiente. La classe d'âge de la patiente et le type de scanner utilisé pour l'examen sont les deux dernières dimensions du contexte d'analyse. La mesure correspond au nombre de zones suspectes détectées chez la patiente sur une mammographie. Au total nous avons 1 485 faits agrégés et 336 960 cellules dans le contexte d'analyse ainsi défini. Nous proposons une représentation sous forme de schéma en étoile du contexte d'analyse sur la figure 2.10. Ainsi dans ce contexte d'analyse, un utilisateur peut répondre à différentes questions, comme : "quel est le nombre de régions suspectes susceptibles d'être détectées selon la certitude du médecin lors de son évaluation et lorsque les autres dimensions sont fixées à une modalité particulière ?".

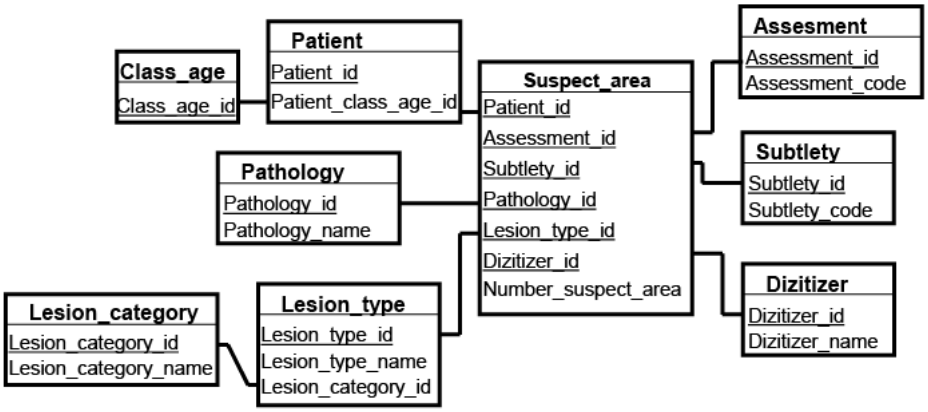


FIGURE 2.10 – Représentation multidimensionnelle du contexte d'analyse

Modèle de prédiction. Compte tenu du volume des données dans le contexte d'analyse défini ci-dessus, nous utilisons l'algorithme d'apprentissage AID pour construire l'arbre de régression. 70% des données du contexte d'analyse constituent l'échantillon d'apprentissage, soit 1 039 faits agrégés, les 30% restants, soit 446 faits agrégés, forment l'échantillon test. Nous avons paramétré la méthode avec un effectif minimal sur chaque sommet fixé à 15 faits agrégés et un nombre maximal de niveaux dans l'arbre à 10.

L'arbre de régression construit avec l'échantillon d'apprentissage comporte 8 sommets dont 6 feuilles (figure 2.11). Sur l'échantillon test, l'erreur moyenne de l'arbre est de 0,11 et la réduction de l'erreur est de 0,64 alors qu'elle était 0,71 dans l'échantillon d'apprentissage. Le modèle doit donc être exploité avec précaution.

Interprétation du modèle de prédiction. Les variables explicatives (dimensions) discriminantes sont l'indice d'évaluation du médecin et le type de scanner. Les autres

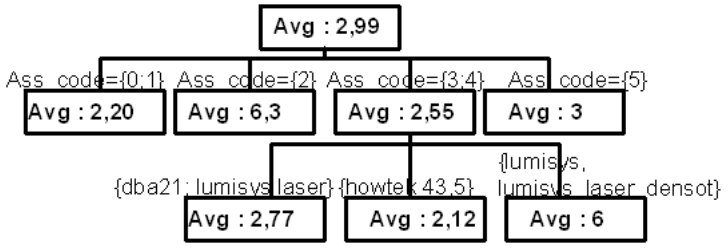


FIGURE 2.11 – Arbre de régression sur le cube de données.

dimensions ne sont donc pas des variables expliquant le nombre de régions suspectes détectées. Nous obtenons les 6 règles suivantes, correspondant à chacune des 6 feuilles terminales de l'arbre :

- $R_1((0 \vee 1) \Rightarrow 2,20 \ ; \ 7,12\% \ ; \ 0,87)$
- $R_2(2 \Rightarrow 6,03 \ ; \ 11,65\% \ ; \ 4,24)$
- $R_3(5 \Rightarrow 3 \ ; \ 13,76\% \ ; \ 1,47)$
- $R_4((dba21 \vee lumisys\_laser) \wedge (3 \vee 4) \Rightarrow 2,77 \ ; \ 40,62\% \ ; \ 1,48)$
- $R_5((howtek43.5) \wedge (3 \vee 4) \Rightarrow 2,12 \ ; \ 26,18\% \ ; \ 0,55)$
- $R_6((lumisys \vee lumisys\_laser\_densot) \wedge (3 \vee 4) \Rightarrow 6 \ ; \ 0,67\% \ ; \ 1,51)$

A titre d'exemple, la règle 4 est à interpréter de la sorte : si le type de scanner est "lumisys laser" ou "dba 21" et que l'indice d'évaluation du médecin est de 3 ou 4, alors le nombre de régions suspectes sera en moyenne 2,77. 40,62% des individus de l'ensemble d'apprentissage supportent cette règle et l'écart type est de 1,48.

Exploitation OLAP du modèle de prédiction. Nous avons choisi 5 cellules (5 faits inexistants) pour lesquelles nous souhaitons effectuer une prédiction de la mesure. Pour chacune d'elles nous avons comparé les modalités qui la décrivent aux antécédents des règles. Nous retrouvons en gras dans la dernière colonne du tableau 2.1, pour chacune des 5 cellules, le résultat obtenu avec le modèle. Par exemple, pour la 4ème cellule, la règle 4 est utilisée. Dans le cadre du *"What if analysis"*, la prédiction faite permet de répondre à la question : à combien de régions suspectes doit-on s'attendre si le médecin a un indice d'évaluation de 3, si la subtilité est de 2, si la tumeur est maligne, la lésion est de type calcification amorphe, si la patiente a entre 50 et 54 ans et si l'examen a été réalisé avec un scanner de type "lumisys laser" ?

De plus, on note pour la règle 2 un écart type élevé (4,24). Les prédictions pouvant être obtenues à partir de cette règle sont donc à prendre avec précaution. La validité des résultats peut par la suite être indiquée à l'utilisateur grâce à un code visuel de couleur, permettant de mettre en exergue les résultats à manipuler avec prudence.

Evaluation	Subtilité	Pathologie	Type de lésion	Classe d'âge	Scanner	Nombre de régions (AVG)
0	2	maligne	calcification amorphe	50 - 54 ans	lumisys laser	**2,2**
1	2	maligne	calcification amorphe	50 - 54 ans	lumisys laser	**2,2**
2	2	maligne	calcification amorphe	50 - 54 ans	lumisys laser	**6,03**
3	2	maligne	calcification amorphe	50 - 54 ans	lumisys laser	**2,77**
5	2	maligne	calcification amorphe	50 - 54 ans	lumisys laser	**3**

TABLE 2.1 – Valeurs prédites pour le nombre de régions suspectes.

Performances. Le temps de calcul de l'approche est en partie lié au type d'arbre de régression utilisé pour la construction du modèle de prédiction selon qu'il y a une phase d'élagage ou non. De plus, le contexte d'analyse défini par l'utilisateur peut aussi avoir un impact important sur les performances. Selon le niveau d'agrégation choisi, le volume des données peut varier considérablement et peut augmenter le temps de calcul lors de la phase d'apprentissage. Nous avons réalisé des tests où nous augmentions la taille du contexte d'analyse défini ($\times$ 2 *et* $\times$ 4). Avec des jeux de données de taille "raisonnable" en fouille de données (5 940 faits utilisés ici au maximum), nous n'avons pas remarqué de variation significative du temps (de l'ordre de la seconde). En revanche, nous avons noté des variations importantes de temps de calcul lorsque l'on utilise des variables explicatives avec beaucoup de modalités. Ceci a été le cas sur un test réalisé où nous avions plus de 1 000 modalités pour une dimension du cube. La variable était une expression textuelle de la taille et de la forme des zones suspectes analysées. Le temps de calcul est alors passé de quelques secondes à plusieurs dizaines de minutes. C'est donc un problème lié aux cas des données complexes, notamment textuelles, problème que nous aborderons dans le chapitre suivant.

2.4.4 Suite des travaux

Notre réflexion sur l'intégration de la prédiction dans l'OLAP étant relativement récente, nous encadrons actuellement des travaux pour :

- mettre en place des paramètres visuels indiquant les valeurs prédites, les cellules pouvant être prévues à un niveau d'agrégation supérieur et la qualité des prédictions ;
- traiter le cas où l'arbre de régression ne donne pas une prédiction plus précise que la moyenne globale de la mesure sur l'échantillon d'apprentissage ;
- explorer un niveau d'agrégation plus fin et ce en tenant compte des prédictions réalisées à un niveau supérieur ;
- poursuivre les expérimentations et tester l'ensemble des méthodes d'arbre de

régression.

Nous pensons qu'il faut aller au delà de ces perspectives immédiates, poursuivre notre réflexion et notre travail. Pour cela, nous avons commencé une collaboration avec l'Ecole Nationale des Sciences Appliquées d'Agadir au Maroc et une thèse en co-direction a démarré à l'automne 2009. Dans la thèse, nous cherchons à comment :

- partitionner le cube initial en sous-cubes denses selon une méthode de clustering ou de biclustrering ;
- construire et valider pour chaque sous-cube dense un modèle de prédiction ;
- choisir (par l'utilisateur) la cellule à prédire et non plus un contexte d'analyse ;
- déterminer le sous-cube dans lequel est la cellule désignée par l'utilisateur ;
- prédire la valeur de la cellule grâce au modèle du sous-cube.

Les deux premières étapes de ce processus se feraient préalablement à l'analyse en ligne. En revanche les trois étapes suivantes seraient "en ligne". L'idée de partitionner le cube en sous-cubes denses est une idée que l'on retrouve dans les travaux de compression du cube [BW00, MGCB07]. Ces sous-cubes denses pouraient être vus comme des régions intéressantes et servir d'ensemble d'apprentissage pour construire le modèle de prédiction.

Pour valider nos approches, nous avons entrepris et encadré le développement d'un prototype baptisé MiningCubes.

2.5 Plate-forme logicielle MiningCubes

Nos propositions d'extension des capacités de l'OLAP et nos propositions pour l'analyse en ligne des données complexes sont implémentées sous forme d'opérateurs dans la plate-forme MiningCubes. Une donnée complexe est souvent caractérisée par des sources hétérogènes, des supports différents et des représentations diverses. Nous avons fait le choix d'utiliser XML comme formalisme pour structurer, homogénéiser et entreposer les données complexes. MiningCubes se connecte à des cubes de données et permet de réaliser une analyse en ligne des données.

MiningCubes est une application Web développée avec le langage ASP. Avec une architecture de type Client/Serveur, l'application est composée d'un ensemble de modules : des modules de connexion aux données, des modules d'exploration des données et des modules d'analyse.

Les données sources peuvent provenir soit d'un cube de données sous *Analysis*

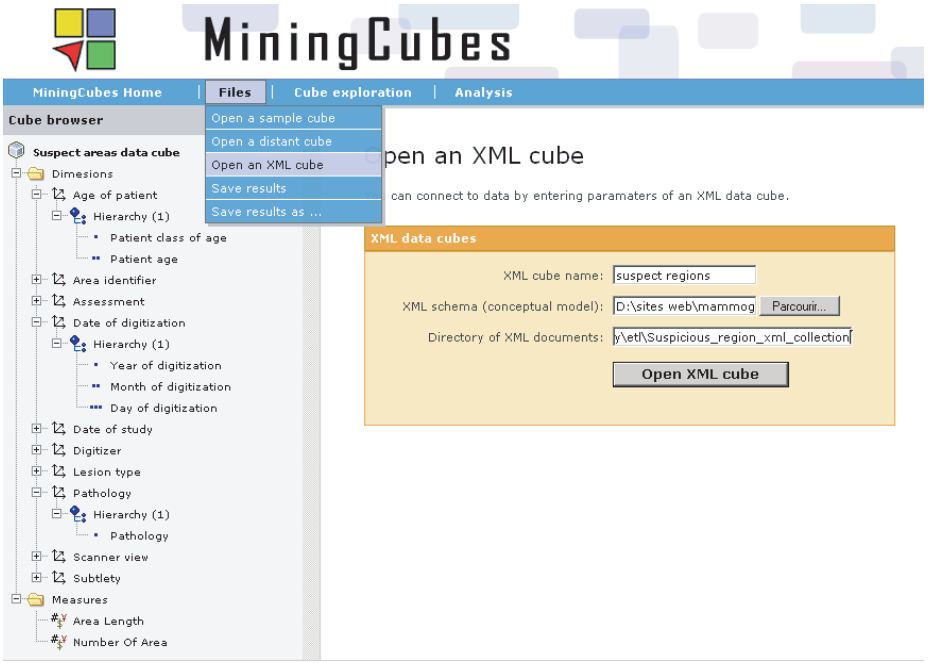


FIGURE 2.12 – Chargement d'un cube XML dans MiningCubes

Services de MicroSoft SQL Server soit d'un cube de données XML. Il y a deux modules de **connexion** aux données, un pour chaque type de connexion possible. Le deuxième module a pour objectif d'assurer la prise en compte des données complexes dans la plate-forme d'analyse ; il charge le schéma XML décrivant le cube ainsi que les documents XML (figure 2.12).

Les modules d'**exploration** des données multidimensionnelles permettent, avec des outils visuels, de naviguer dans un cube de données en se basant sur les opérations OLAP classiques telles que le *roll-up*, le *drill-down*, le *slice* et le *dice*.

Les modules d'**analyse** comprennent les implémentations de nos propositions de nouveaux opérateurs OLAP. Dans ces différents modules, il y a une interface de paramétrage qui assure la communication entre l'utilisateur et les modules de connexion aux données. Via cette interface, l'utilisateur connecte le module à un cube de données et choisit les dimensions, les mesures, les niveaux selon lesquels il souhaite travailler. Il construit ainsi le sous-cube à analyser et l'interface récupère les données correspondantes c'est-à-dire les faits OLAP.

Les modules d'analyse incluent des composantes spécifiques à chaque opérateur.

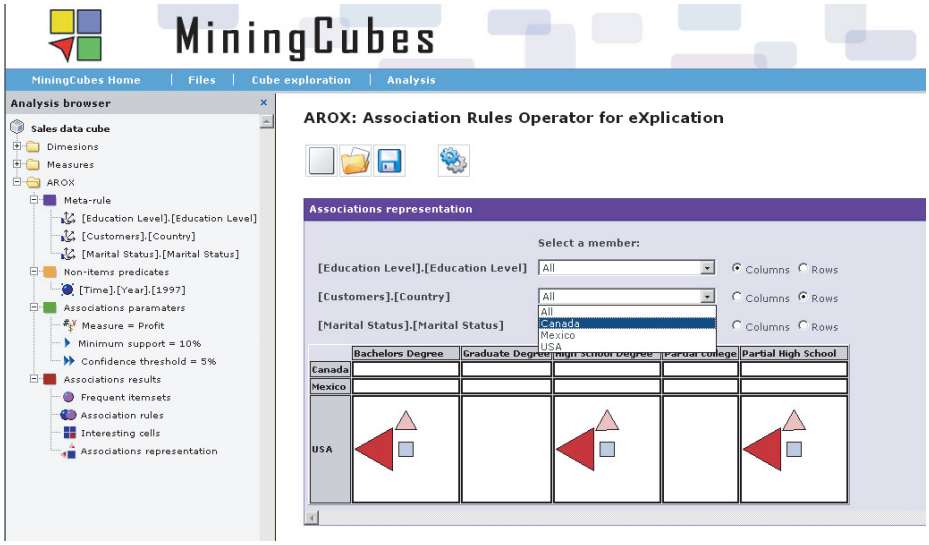


FIGURE 2.13 – Résultats d'une analyse dans MiningCubes

Généralement, il y a au moins une composante qui exécute la méthode de fouille de données et une qui permet de visualiser les résultats et les indicateurs d'aide à l'interprétation (figure 2.13).

A l'heure actuelle, la plate-forme MiningCubes est sur un serveur local dans un cadre expérimental.

Nous projetons la mise en ligne de la plate-forme pour faire de MiningCubes un fournisseur de services Web d'analyse en ligne de données complexes. MiningCubes est par conséquent en cours de réécriture. De plus, nous souhaitons faire de MiningCubes un projet ouvert dans lequel il sera possible d'accéder aux codes sources et de rajouter de nouveaux modules d'analyse en ligne et/ou de fouille de données.

2.6 Conclusion

Dans ce chapitre, nous avons proposé d'étendre l'OLAP à d'autres possibilités d'analyse que l'exploration interactive et la visualisation. Notre idée est de combiner les principes de l'OLAP à ceux de la fouille de données pour enrichir l'analyse en ligne.

Nous avons alors proposé deux nouvelles approches : AROX pour expliquer des phénomènes observés dans le cube à l'aide de règles d'association et OPReT pour

prévoir la valeur de la mesure de faits et réaliser une analyse de type *What If Analysis*. Ces deux extensions d'OLAP ont été développées sous forme de modules d'analyse dans la plate-forme Web MiningCubes.

Pour la recherche de règles d'association dans les données multidimensionnelles, aucun pré-traitement préalable n'est nécessaire sur les cubes de données. L'algorithme que nous proposons est une adaptation d'Apriori aux données multidimensionnelles. Il exploite la propriété d'anti-monotonie et repose sur une recherche ascendante des motifs fréquents qui est particulièrement adaptée aux données éparses. Nous utilisons une méta-règle inter-dimensionnelle afin de guider le processus de recherche de règles dans le cube. Nous modifions également les définitions du support et de la confiance d'une règle pour les rendre plus pertinentes dans le contexte de l'analyse en ligne. Afin de visualiser les règles extraites, nous en proposons un codage graphique.

En appliquant au coeur du processus OLAP une technique de prédiction basée sur les arbres de régression, nous proposons à l'utilisateur de se placer dans une démarche prédictive, et grâce à la discrimination des variables, dans une démarche explicative. Dans notre approche, l'utilisateur peut estimer la valeur de la mesure de faits inexistants et avoir des indicateurs de fiabilité des règles de décision. Une fois les valeurs estimées, l'utilisateur peut continuer sa navigation dans des niveaux prenant en compte les agrégats calculés.

Nos propositions sont une première démonstration de la faisabilité de combiner l'OLAP à la fouille de données pour faire significativement évoluer l'analyse en ligne. Elles sont une partie des réponses possibles à ce problème. Mais il existe beaucoup de méthodes de fouille et de nombreuses autres combinaisons possibles. L'idée n'est pas forcément de combiner toutes les techniques de fouille à l'OLAP mais de réfléchir aux extensions intéressantes dont devrait se doter l'analyse en ligne. Une autre voix de recherche serait de se tourner vers d'autres domaines pour en étudier les méthodes d'analyse.

Tous les problèmes abordés ainsi que les solutions proposées dans ce chapitre sont indépendants de la nature et du contexte des données. La problématique de l'enrichissement de l'analyse en ligne se pose y compris dans le cas des données classiquement stockées dans des entrepôts et traitées par l'OLAP. Cependant l'avènement des données complexes et leur nature spécifique posent de nouveaux problèmes de recherche et engendrent de nouveaux besoins en terme d'entreposage. De nouvelles approches pour l'intégration, la modélisation et l'analyse en ligne sont nécessaires quand on travaille avec des données complexes. Le chapitre suivant aborde le problème de l'analyse en ligne des données complexes.

Chapitre 3

Analyse en ligne des données complexes

Résumé

En utilisant l'exemple de l'analyse des publications d'un laboratoire de recherche, nous montrons dans ce chapitre comment on peut combiner l'OLAP, la fouille de données et la recherche d'information pour l'analyse en ligne des données complexes. Nous proposons de modéliser les données complexes de façon multidimensionnelle ainsi que trois opérateurs pour les visualiser (CoDaViC : Complex Data Visualization by Correspondences), les agréger (OPAC : Operator for Aggregation by Clustering) et les réarranger (ORCA : Operator for Reorganization by multiple Correspondence Analysis) dans le cube.

3.1 Introduction

L'avènement du Web, des données multimédias dans de nombreux domaines (médical, ...), des données dans différentes sources, a lancé un nouveau défi aux entrepôts de données et à l'analyse en ligne. Ces données sont qualifiées de complexes. Il n'existe pas de définition standard des données complexes, mais notre équipe propose de dire que les données sont complexes quand elles sont [BBDR03] :

- multi-format : données de types différents (numérique, symbolique, texte, image, son, vidéo, ...) et/ou
- multi-structure : données structurées, semi-structurées ou non structurées (vues relationnelles, données XML, ...) et/ou
- multi-source : données d'origine différente (bases de données réparties, données de production, scanners, satellites, enregistrements vidéos, comptes-rendus médicaux, résultats d'analyse, web, ...) et/ou

- multi-modale : un phénomène décrit par plusieurs canaux ou points de vue (radiographies et diagnostic audio d'un médecin pour évaluer l'état de santé d'un patient, données exprimées dans des échelles ou des langues différentes, ...) et/ou
- multi-version : données évolutives en terme de définition ou de valeur (bases de données temporelles, recensements périodiques avec critères évolutifs, ...).

Cette définition ne couvre pas la totalité des données complexes et pourrait être complétée par d'autres axes comme les traitements, le volume et la sémantique où la complexité porterait sur le contenu, l'interprétation et le contexte des données.

La spécificité des données complexes pose de nouveaux problèmes de recherche et engendre des besoins pour l'entreposage et l'analyse, en plus de ceux déjà existants pour des données non complexes (souvent numériques). Intégrer des sources différentes et des données complexes nécessite plus qu'une transformation des données en une représentation unique. Les sources de données ne sont pas forcément toutes structurées sous forme de bases de données mais peuvent être des corpus de documents, provenir du Web avec un rafraîchissement récurrent, être réparties, ... Le processus d'ETL ne peut plus se dérouler classiquement quand les données sont de type et de nature différents. Les données contenues dans les sources ne suffisent pas pour pouvoir modéliser de façon multidimensionnelle les données complexes. Les données complexes sont souvent décrites dans des documents XML, faut-il utiliser XML pour construire des entrepôts de documents ? Les opérateurs OLAP classiques ne sont pas adaptés quand on veut analyser des données contenant du texte, des images, du son ou décrites par des documents XML. De nouvelles approches pour l'intégration, la structuration ou modélisation et pour l'analyse des données complexes sont donc nécessaires.

Parmi tous ces défis, nous nous intéressons plus particulièrement à ceux posés par l'analyse en ligne des données complexes ce qui nous conduit à nous interroger également sur la modélisation multidimensionnelle de telles données.

D'après ses fondements, l'OLAP consiste à agréger des données détaillées selon des niveaux de granularité différents pour résumer et visualiser l'information contenue dans les données. D'autres opérations, plus structurelles, réorganisent les données, modifient les angles d'analyse pour faire émerger de l'information. En transposant ces objectifs aux données complexes, nous abordons les problèmes suivants :

- Comment agréger en ligne des données complexes ? Résumer l'information contenue dans des données complexes nécessite plus qu'une simple opération arithmétique.
- Comment visualiser en ligne l'information contenue dans les données complexes ?
- Comment organiser les cubes de données complexes pour améliorer la visualisation de l'information ? Les cubes de données complexes sont souvent épars et l'information est diffuse.

– Comment prendre en compte le contenu sémantique des données complexes ? Les données complexes véhiculent de la sémantique qu'il faut prendre en compte dans la modélisation et dans la construction du cube mais aussi dans l'analyse.

Pour comprendre les réflexions scientifiques menées et illustrer les propositions faites dans ce chapitre nous introduisons un nouvel exemple : l'analyse des publications d'un laboratoire de recherche.

Nous considérons qu'une publication scientifique est une donnée complexe. Une publication est associée à un document qui contient généralement un titre, le nom des auteurs, un résumé et le corps composé lui même de textes, de figures, de tableaux, ... En plus du document, d'autres informations sont associées à une publication comme la date et le support pour localiser la publication (conférence ou revue par exemple, éditeur, numéro et volume, nombre de pages, ...), la portée nationale ou internationale du support. Un auteur peut aussi être représenté par son nom, par sa photographie (image), ainsi que par d'autres attributs comme son statut (professeur, maître de conférences, doctorant, ...), les dates d'arrivée, de départ dans le laboratoire, ...

Les publications sont des entités sémantiques que nous souhaitons observer en tant que telles. Outre les éléments classiques comme l'année, le type de publication, les auteurs, le nombre de pages, l'utilisateur peut vouloir analyser aussi le contenu sémantique de la publication, c'est-à-dire le sujet sur lequel porte la publication.

Nous commençons par modéliser de façon multidimensionnelle les publications. Comme les publications contiennent entre autres choses un document texte et comme les données complexes peuvent être décrites par des documents XML, nous nous sommes orientés vers une combinaison de l'OLAP, de la recherche d'information et des documents XML. Nous nous inspirons des principes de la recherche d'information (*Information Retrieval*) pour la modélisation multidimensionnelle et pour la construction de cubes de données complexes, mais aussi pour l'analyse.

Dès la phase de modélisation, nous voulons capter une partie de la sémantique contenue dans les données. Nous associons à chaque publication plusieurs mots-clés. Plutôt que de demander aux auteurs d'attribuer eux-mêmes des mots-clés à leur publication, il nous paraît plus pertinent d'extraire automatiquement les mots-clés à partir du titre, du résumé ou du texte (corps) de la publication. Nous utilisons pour cela les principes de la recherche d'information (RI) où l'extraction automatique des mots-clés se fait lors de la phase d'indexation. Parmi les différentes techniques possibles d'indexation, nous avons choisi l'indexation conceptuelle et nous inspirons des travaux de Baziz [Baz05].

Une fois modélisées, les publications peuvent être analysées mais à l'aide d'opérateurs OLAP adaptés. Grâce à nos travaux combinant OLAP, fouille de données et recherche d'information, nous sommes en mesure de proposer trois adaptations de l'OLAP aux données complexes : CoDaViC (*Complex Data Visualization by Correspondences*) pour la visualisation des données complexes ; OpAC (*Operator for Aggregation by Clustering*) qui permet d'agréger de telles données et de construire ainsi une hiérarchie de dimension ; ORCA (*Operator for Reorganization by multiple Correspondence Analysis*) qui réorganise les modalités ou membres d'une dimension pour détecter des régions intéressantes. Ces propositions ont été publiées dans des conférences et des revues nationales ou internationales [BRBB04, BBR04, BBLR05, BBLR06d, BBLR06b, BBLR06a, BBLR07, MLB10, LB11].

Dans ce chapitre, nous commençons par présenter la modélisation multidimensionnelle des publications dans la section 3.2. Pour cela, nous nous intéressons aux travaux associant OLAP, recherche d'information et documents XML et nous exposons le modèle que nous avons adopté ainsi que le processus d'extraction de mots-clés. Nous présentons dans les sections 3.3, 3.4 et 3.5 les trois adaptations de l'OLAP aux données complexes pour visualiser, pour agréger des données complexes et pour ordonner les modalités d'une dimension. Nous terminons le chapitre en présentant dans la section 3.6 la plate-forme logicielle PUMA (*PUblication Modeling and Analysis*) dédiée à l'analyse en ligne des publications.

3.2 Modélisation des publications

3.2.1 OLAP, RI et documents

La RI ou recherche d'information consiste à chercher une information dans un ensemble ou un corpus de documents. L'utilisateur décrit à l'aide de mots-clés l'information qu'il cherche. Le système de RI évalue les documents par rapport à l'information cherchée, retourne les documents jugés comme les plus pertinents [SWY75]. La pertinence évalue dans quelle mesure le document est pertinent par rapport aux mots-clés exprimant l'information recherchée. Elle est souvent basée sur les fréquences locales (tf) et globales (idf) des mots-clés dans le document et dans le corpus de documents. L'évaluation du système de RI se fait par des indicateurs comme la précision et le rappel.

Depuis une dizaine d'années, quelques équipes de recherche s'intéressent à l'association de l'OLAP et de la RI. Les premiers travaux remontent au début des années 2000 [MLC+00, LGO03, MCDA03, PP03]. Plus récemment, l'équipe de Han introduit le terme de base de données textuelles multidimensionnelles

(*multidimensional text database*) en le distinguant des bases de données relationnelles et des bases de documents textes. Cette équipe propose des modèles, des hiérarchies de dimension et un prototype [LDH+08, ZZH09, SHZ+09, SYH09, YLS+09, DZL+10].

De plus, comme les données complexes et les documents sont souvent stockés sous forme de documents XML, cela nous conduit à nous intéresser aussi aux travaux de l'OLAP sur les documents XML. Selon [Tou07], il existe deux types de documents XML :

- Les documents XML orientés données représentant des données très structurées telles que le contenu d'une base de données relationnelle. La structure arborescente XML est utilisée pour décrire les données et les attributs, le contenu des données est inséré entre les balises.
- Les documents XML orientés documents principalement composés de texte tels que les versions électroniques des documents papiers. Ces documents ont une structure plus hétérogène et contiennent différents types de données tels que des images, du texte, des tableaux, ...

Compte tenu de nos problématiques, seuls les travaux portant sur les entrepôts XML orientés documents nous intéressent [Sul01, PHS05, PBAP05, PBAP07, KKL05, RTT07, RTTZ08b, RTTZ08a].

La combinaison de l'OLAP et de l'analyse des documents est une problématique abordée différemment selon les équipes de recherche. Il n'y a pas d'approche unifiée.

Nous constatons par exemple que le document peut jouer un rôle différent dans la modélisation multidimensionnelle et être :

- Un niveau ou une dimension. Le document peut être le 1er (et souvent unique) niveau d'une dimension textuelle, mais peu de travaux permettent une opération de navigation (avec des opérateurs spécifiques) sur cette dimension textuelle [SYH09, PHS05]. Ravat *et al.* introduisent la notion de dimension documentaire dont au moins l'une des hiérarchies modélise la structure logique du document (paragraphes, sections, chapitres, ...) [RTTZ08a].

- Associé à des concepts. Le document n'est pas toujours explicitement dans le modèle multidimensionnel, il est dans ce cas associé à un ou à des faits avec un lien [MCDA03, PP03]. De nombreux travaux utilisent le principe d'une dimension avec une hiérarchie de mots-clés ou de concepts décrivant le contenu du document ou le domaine s'y rapportant. Cette hiérarchie de concepts est soit issue d'une ontologie du domaine [LGO03, MCDA03] soit considérée comme textuelle et construite automatiquement à partir de mots-clés résumant le document [LDH+08, ZZH09, SHZ+09, SYH09, RTTZ08b]. L'utilisateur peut

alors faire des opérations de navigation sur cette dimension textuelle.

– Le fait. A notre connaissance, aucun des travaux ne considère le document comme le fait.

Selon les travaux, la mesure est différente et peut être :

– inexistante. Dans certains cas, il est difficile de définir une mesure et le fait est simplement une combinaison de toutes les dimensions sans mesure particulière. Généralement, dans un tel cas, la fonction d'agrégation COUNT est utilisée pour obtenir un dénombrement des faits.

– numérique (additive ou semi-additive) comme une fréquence de mots-clés [MLC+00, LGO03, LDH+08, BLP+10], un index inversé [LDH+08], un poids évaluant le lien ou la pertinence entre le fait et un document [MCDA03, PBAP05]. Ces mesures sont numériques et elles permettent de rester dans le contexte classique de l'OLAP et de pouvoir utiliser les opérateurs connus. On peut cependant regretter que ces mesures soient seulement structurelles sans prendre en compte la sémantique contenue dans le document.

– textuelle. Une mesure textuelle est à la fois non numérique et non additive [RTTZ08b]. Le contenu d'une mesure textuelle peut représenter un mot, un paquet de mots, un paragraphe voire un document complet. On peut distinguer plusieurs types de mesures textuelles : (1) une mesure textuelle brute dont le contenu correspond au contenu d'un document ; (2) une mesure textuelle élaborée dont le contenu est issu d'une mesure textuelle brute et ayant subi un certain nombre de prétraitements (retrait des mots outils, vides de sens, ...). Une mesure textuelle de type mot-clé est une mesure textuelle élaborée.

En s'inspirant des travaux de Park *et al.* [PHS05], Ravat *et al.* introduisent une mesure textuelle basée sur les mots-clés et deux fonctions d'agrégation : AVG KW pour effectuer une pseudo-moyenne des mots-clés en synthétisant un ensemble de mots-clés par un ensemble plus petit de mots-clés mais plus généraux [RTT07] ; TOP KEYWORDK qui renvoie à l'utilisateur les k principaux mots-clefs d'un ensemble de mots-clés à agréger [RTTZ08b].

Bringay *et al.* proposent une mesure adaptative basée sur les fréquences $tf \times idf$ pour extraire les mots-clés significatifs en tenant compte d'une hiérarchie d'une autre dimension [BLP+10]. Ils se démarquent des autres approches où les mots-clés extraits constituent le plus bas niveau d'une nouvelle dimension et où les mots-clés sont agrégés sans tenir compte des autres dimensions. La fonction d'agrégation de Bringay *et al.* ne retourne que les $top-k$ mots significatifs en fonction du niveau d'une autre dimension.

Les auteurs assortissent la mesure adaptative, d'une fonction d'agrégation dynamique des mots-clés permettant de faire émerger des nouveaux regroupements de mots-clés.

Une autre différence entre les travaux réside dans la modélisation multidimensionnelle selon qu'elle est classique ou spécifique. Ravat *et al.* proposent un modèle conceptuel multidimensionnel en galaxie adapté à l'analyse de documents XML contenant des données textuelles [RTTZ08a]. Ce modèle en galaxie repose sur un unique concept central de dimension. La notion de fait est supprimée. Une galaxie est un regroupement de dimensions liées entre elles par un ou plusieurs noeuds centraux ; chaque noeud modélise les dimensions compatibles pour une même analyse.

En résumé, les travaux diffèrent selon que :
- la modélisation multidimensionnelle est classique ou spécifique ;
- le fait est ou non un fait particulier autre que le fait lié au domaine étudié ;
- la mesure est classique (numérique) ou textuelle ;
- il existe ou non une hiérarchie de concepts pour prendre en compte le contenu sémantique du document ;
- il y a ou non une dimension textuelle ou documentaire ;
- des opérateurs OLAP adaptés soient proposés ;
- la place du document dans la modélisation ;
- la possibilité de faire une requête RI (requête avec des mots-clés ou des documents dans le cube OLAP) ;
- la prise en compte des documents XML.

Dans la modélisation des publications, certaines informations comme le temps, le support et les auteurs peuvent être structurées et hiérarchisées ; d'autres sont de nature textuelle. La modélisation multidimensionnelle des publications doit donc prendre en compte les informations structurées et textuelles, hiérarchisées ou non. Elle doit également supporter des opérations d'agrégation (comme le *roll-up* et le *drill-down*) non seulement sur les hiérarchies des dimensions classiques mais aussi sur les dimensions textuelles afin de pouvoir faire une navigation dans les données.

3.2.2 Modélisation multidimensionnelle

Dans notre modélisation multidimensionnelle le fait est la publication (figure 3.1). Nous choisissons de l'observer selon plusieurs dimensions : le temps, les auteurs, le support, les mots-clés et le document. Pour la dimension du temps, dans la date seule l'année nous intéresse et elle représente le plus bas niveau de granularité. Il est possible de regrouper les années en périodes. La dimension des auteurs contient deux niveaux hiérarchiques, permettant de regrouper les auteurs selon leur statut (professeur, maître de conférences, doctorant, ...). La dimension du support stipule

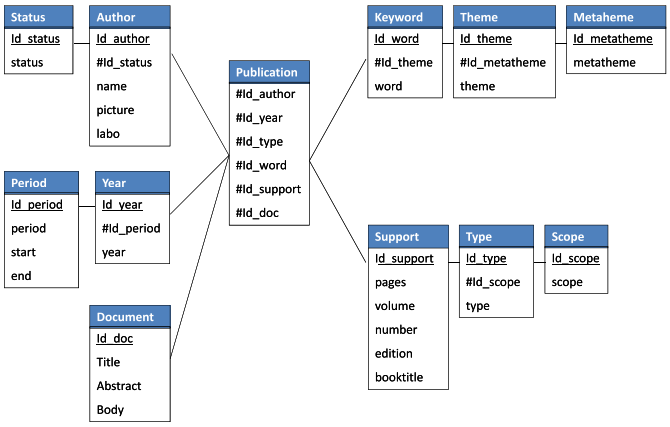


FIGURE 3.1 – Modélisation multidimensionnelle des publications

le nom de la conférence, de la revue, de l'ouvrage ainsi que le volume, le nombre de pages, ... Chaque support peut être agrégé selon son type (conférence, revue, chapitre d'ouvrage, mémoire de thèse, ...) et selon sa portée (nationale ou internationale). Avec les notations introduites dans le chapitre précédent, nous pouvons par exemple écrire que la dimension du support de publications D^5 contient trois niveaux : le niveau *Support* noté H_1^5, le niveau *Type* noté H_2^5 et le niveau *Scope* noté H_3^5. Le niveau *Scope* (H_3^5) contient deux modalités : *International*, notée a_1^{53} et *National*, notée a_2^{53}.

Les deux dernières dimensions sont des dimensions textuelles ou sémantiques avec les mots-clés associés à la publication (mots-clés que l'on peut regrouper en thème et en classes de thèmes ou «métathèmes») et le document en lui même.

Dans ce modèle, la complexité est liée, d'une part, à l'hétérogénéité des informations qui composent une publication. Pour la modéliser, nous introduisons des dimensions classiques avec des hiérarchies et des dimensions textuelles constituées du document ou d'une hiérarchie de concepts (mots-clés, thèmes et métathèmes). D'autre part, la complexité vient également de la structure particulière de certaines dimensions comme par exemple les hiérarchies des dimensions *Author* et *Keyword* qui ne sont pas strictes ou le statut d'un auteur qui peut évoluer dans le temps.

Un autre point de vue consisterait à dire que les faits pourraient aussi être les auteurs que l'on analyserait selon leurs publications, leur statut, ... De la même façon, les mots-clés pourraient être les faits que l'on analyserait selon les auteurs,

les années, ... On retrouve ici l'idée selon laquelle chaque dimension peut devenir fait (et vice versa) et la notion d'objets interchangeables dans la modélisation multidimensionnelle [BBBL10]. Cette idée est aussi modélisable avec le modèle en galaxie de Ravat *et al.* [RTTZ08a].

Comme dit précédemment, le fait à observer est la publication. C'est la combinaison de toutes les dimensions sans mesure particulière. Nous utilisons la fonction d'agrégation COUNT pour obtenir un dénombrement des faits. Certes, en l'absence de mesure, cette solution d'utiliser la fonction COUNT pour dénombrer les publications est toujours possible mais elle ne suffit pas car l'analyse qui en découle est pauvre. Nous cherchons alors d'autres moyens pour analyser les publications (proximité thématique, auteurs qui travaillent ensemble, ...). Nous considérons une publication comme une donnée complexe et nous cherchons un moyen d'en faire une analyse sémantique.

3.2.3 Extraction de mots-clés

Dans ce paragraphe, nous présentons rapidement le principe de l'indexation conceptuelle et justifions son choix. Nous exposons ensuite le processus d'extraction des mots-clés.

L'indexation d'un document consiste à transformer le document textuel en une représentation exploitable par le système de recherche d'information. Il faut extraire du texte un ensemble de descripteurs c'est-à-dire un ensemble de mots ou de termes (après suppression des mots grammaticaux, troncatisation, lemmatisation, ...). A l'aide de ces termes, il est possible de représenter chaque document Doc_i par un vecteur dans l'espace des descripteurs $Doc_i =< d_{i1}, d_{i2}, ..., d_{ik}, ..., d_{it} >$. d_{ik} est le poids du mot-clé k dans le document Doc_i et il dépend de la fréquence locale (tf : *term frequency*) et globale (idf : *inverse document frequency*) du mot-clé dans le document et dans le corpus de documents [SWY75]. Le poids d_{ik} est le produit des fréquences locales et globales :

$$d_{ik} = tf(k;i) \times idf(k)$$

où $tf(k;i)$ est la fréquence (nombre d'occurrences) du mot-clé k dans le document Doc_i ; $idf(k) = \log(\frac{N}{n_k})$ avec n_k le nombre de documents contenant le mot-clé k parmi les N documents du corpus.

En RI, l'indexation sémantique et l'indexation conceptuelle ont été introduites comme alternatives pour pallier les défauts de l'indexation classique. L'indexation conceptuelle se fonde d'une part sur des concepts tirés de ressources externes et d'autre part elle préfère indexer par des termes composés plutôt que par des mots simples :

- Un concept est une idée qui inclut tout ce qui est caractéristiquement associé à elle. En RI, les concepts sont généralement définis dans des structures conceptuelles comme les thésaurus, ontologies ou hiérarchies de concepts. Ces structures sont souvent construites manuellement par des spécialistes du domaine.
- Sémantiquement, on peut sentir l'intérêt d'extraire du document des termes composés de plusieurs mots et non pas comme dans l'indexation classique d'extraire seulement des mots simples. Selon Baziz, extraire des termes composés permet de réduire l'ambiguïté des mots, notamment lors de l'affectation à un concept [Baz05]. Les termes composés sont en général monosémiques (ils n'ont qu'un seul sens) même si les mots qui les composent peuvent être ambigus. Ainsi, plus il y a de concepts multi-mots, plus la désambiguïsation des termes est facile.

Dans ce contexte, Baziz propose une technique d'indexation conceptuelle guidée par une ressource externe [Baz05]. Il exploite les liens sémantiques pour identifier dans le texte du document, les termes qui correspondent aux concepts de la ressource externe. Cette technique a l'avantage de s'affranchir des aléas de la combinaison des groupes de mots et ne regroupe pas des mots seulement d'un point de vue syntaxique ou co-occurrent dans le texte. L'utilisation du vocabulaire contrôlé par la ressource externe élimine les «faux termes». Le problème n'est plus seulement de trouver les termes composés qui apparaissent fréquemment dans le document mais de trouver un moyen de les apparier aux concepts de la ressource externe. L'inconvénient de cette approche est certes qu'elle suppose que tous les concepts des documents soient couverts par la ressource et qu'elle soit suffisamment riche pour couvrir le domaine dont traitent les documents. Mais lorsqu'on est dans un domaine «fermé» avec un système spécifique, l'appariement des concepts peut être utilisé et donne de meilleurs résultats que l'indexation classique.

Notre ressource externe est un thésaurus avec tous les concepts couvrant le domaine de recherche du laboratoire. A chaque concept est associé un ensemble de mots synonymes qui peuvent désigner ce concept. L'objectif est d'extraire les termes fréquents du document et de regarder s'ils s'apparient avec les concepts du thésaurus. Le document est projeté sur le thésaurus : pour chaque terme candidat formé de mots adjacents dans le document, on interroge le thésaurus en regardant si le terme correspond à un concept ; s'il ne correspond pas à un concept, on interroge tous les mots associés au concept pour voir si le terme candidat peut s'apparier à un concept. Si le terme candidat correspond ou s'apparie à un concept, alors il devient un mot-clé du document. Après élimination des mots vides de sens, un terme candidat est donc défini comme un terme composé (1, 2 ou 3 mots) de mots adjacents dans le document et dont la fréquence est supérieure à un seuil. Un mot-clé est un terme candidat qui correspond ou s'apparie à un concept du thésaurus.

Comme dit précédemment, il est plus intéressant, sémantiquement, que les termes candidats soient des termes composés de plusieurs mots et non pas seulement de mots simples. Concernant la combinaison de mots, le terme le plus long correspondant à un concept est retenu. Par exemple la séquence (1) de mots "*l'analyse complexe des données*" génère, après élimination des mots vides de sens, 6 termes candidats : "*analyse*", "*complexe*", "*données*", "*analyse complexe*", "*complexe données*", "*analyse complexe données*". Le terme "*analyse données*" n'est pas un terme candidat de la séquence (1) car les mots ne sont pas adjacents dans la séquence. A l'inverse la séquence (2) "*l'analyse des données complexes*" génère comme candidats "*analyse*", "*données*", "*complexes*", "*analyse données*", "*données complexes*", "*analyse données complexes*". Si "**données complexes**" est un concept du thésaurus, alors seule la 2ème séquence aura comme mot-clé "*données complexes*" et les termes "*données*" et "*complexes*" ne seront plus des termes candidats.

Une fois le document décomposé en termes (de longueur 1 à 3), est calculée pour chaque terme une fréquence qui détermine l'importance du terme dans le document. Les termes étant composés de plusieurs mots, Baziz propose d'étendre la fréquence tf pour prendre en compte la longueur (nombre de mots) du terme. Baziz fait deux hypothèses : (1) les termes composés de plusieurs mots sont plus riches sémantiquement que les mots pris séparément ("*donnée complexe*" a plus de sens que "*donnée*" et "*complexe*" pris séparément) ; (2) les mots composant les termes peuvent renvoyer à ces derniers même lorsqu'ils sont utilisés isolément après un certain nombre d'occurrences (par exemple dans un article sur les entrepôts de données, les auteurs peuvent utiliser en cours d'article le mot "*entrepôt*" pour désigner un "*entrepôt de données*"). Avec ces deux hypothèses, Baziz propose une nouvelle définition de la fréquence tf qu'il note cf et qu'il définit comme :

$$cf(k;i) = count(k) + \sum_{sk \subset subterms(k)} \left(\frac{length(sk)}{length(k)} \times count(sk)\right)$$

où $cf(k;i)$ est la fréquence locale du terme candidat k dans le document Doc_i ; $count(k)$ est le nombre de fois où le terme k apparaît dans le document ; $length(k)$ est la longueur du terme k c'est-à-dire le nombre de mots qui le composent ; sk est un sous-terme de k et $subterms(k)$ l'ensemble de tous les sous-termes issus du terme k.

Par exemple, $cf(entrepôt\ données\ complexes;i) = count(entrepôt\ données\ complexes) + 1/3\ count(entrepôt) + 1/3\ count(données) + 1/\ 3count(complexes) + 2/3\ count(entrepôt\ données) + 2/3\ count(données\ complexes)$.

Le poids du terme k est donné par $d_{ik} = cf(i;k) \times idf(k)$ avec $idf(k) = \log(\frac{N}{n_k})$ et n_k le nombre de documents contenant le terme k parmi les N documents du corpus.

La pondération $cf \times idf$ généralise la pondération classique $tf \times idf$ car dans le cas où le terme k est un mot simple, l'ensemble des sous-termes $subterms(k)$ est vide et $cf = tf$. La fréquence idf ne sert pas dans le processus d'extraction des mots-clés car on extrait les mots-clés document par document sans évaluer l'importance d'un mot-clé dans le corpus de documents.

Dans le thésaurus, tous les concepts sont en anglais. A chaque concept est associé un ensemble de mots, mots en anglais et en français, qui correspondent aux différentes formes possibles du concept. Cet ensemble de mots permet de gérer les problèmes de traduction (anglais/français) et en partie les problèmes de lemmatisation des mots. La lemmatisation consiste à remplacer les termes par leur forme canonique, comme par exemple l'élimination du "s" du pluriel, la suppression de la terminaison "ing" en anglais, ... A titre d'illustration, au concept **data warehouse** est associé l'ensemble {*data warehouses*, *data warehousing*, *warehouse*, *entrepôts données*, *entrepôt données*, *entrepôt*, *entreposer*, *entreposage*}.

En résumé, l'extraction des mots-clés d'un document se fait en plusieurs étapes :

1. Choix du texte i (titre, résumé et/ou corps de la publication).
2. Choix de la langue (anglais ou français).
3. Elimination des mots vides de sens et de la ponctuation.
4. Extraction des termes candidats :
   - décomposition du texte en termes k de longueur 1 à 3 mots ;
   - calcul de la fréquence $cf(k; i)$ pour chaque terme k dans le texte i ;
   - sélection des termes fréquents dont la fréquence $cf(k; i)$ est supérieure au seuil ($cf(k; i) \geq seuil$, seuil=2) ;
   - un terme fréquent devient un terme candidat.
5. Extraction des mots-clés :
   - projection des termes candidats sur le thésaurus ;
   - si un terme candidat s'apparie avec un mot ou un groupe de mots du thésaurus alors le concept (en anglais) associé à ce groupe de mots devient un mot-clé du texte. Tous les mots-clés sont en anglais, quelle que soit la langue du document.

Les mots-clés peuvent avoir été extraits du texte lors de l'alimentation des publications dans l'entrepôt de données ou bien ils peuvent être extraits à la volée lors de l'analyse par une fonction d'agrégation opérant sur des fractions de texte.

3.2.4 Suite des travaux

A l'heure actuelle, la hiérarchie de la dimension textuelle des mots-clés repose sur l'arbre hiérarchique des concepts du thésaurus. Elle est construite manuellement

par un expert du domaine. Lors de la recherche des mots-clés, les termes candidats doivent s'apparier à un concept qui est au niveau le plus détaillé de l'arbre, c'est-à-dire à un concept qui est une feuille de l'arbre. Une idée serait de construire automatiquement la hiérarchie des mots-clés. A l'instar des méthodes RankClus [SHZ+09] et NetClus [SYH09], on pourrait utiliser notre opérateur OPAC pour construire automatiquement la hiérarchie de la dimension textuelle. Cette construction pourrait se faire préalablement à l'analyse avec tous les faits contenus dans l'entrepôt ou pourrait se faire dynamiquement lors de la construction du cube et de l'analyse. Il conviendrait de comparer les résultats des méthodes RankClus, NetClus et OPAC.

Il faut également poursuivre la réflexion sur l'extraction des mots-clés, d'une part, pour gérer complètement les problèmes d'homonymie, de polysémie, de synonymie qui sont des cas particuliers d'ambiguïté ; d'autre part jusqu'à la création d'un noyau sémantique pour chaque document. L'objectif du modèle DocCore de [Baz05] est de représenter le contenu sémantique du document par un noyau sémantique, représentation analogue aux réseaux sémantiques.

Une fois les mots-clés extraits et les publications modélisées de façon multidimensionnelle, l'utilisateur a besoin d'une analyse OLAP adaptée. Les sections suivantes présentent nos travaux pour visualiser et agréger les données complexes ainsi que pour restructurer le cube et mettre en évidence les informations intéressantes.

3.3 Visualisation en ligne de données complexes

3.3.1 Positionnement et principe

Pour améliorer la visualisation dans les cubes, plusieurs travaux portent sur l'utilisation des techniques de *Visual Data Mining* et des principes du domaine de l'IHM (Interface Homme-Machine) dans l'OLAP. Maniatis *et al.* en dressent un panorama et proposent le modèle CPM (CPM : *Cube Presentation Model*) pour afficher des écrans dans une interface OLAP [MVSV03]. Généralement, les interfaces OLAP représentent un cube sous forme de table ou de tableau croisé. Pour tenter d'enrichir cette représentation standard, des outils plus sophistiqués proposent des alternatives visuelles avec les visualisations hiérarchiques (arbre de décomposition, arborescence, ...), les vues multi-échelles, les nuages de points interactifs [VM06]. Ces travaux constituent des avancées mais ils ne s'intéressent pas spécialement à la visualisation des données complexes dans l'OLAP. Dans le cas des cubes de données complexes, les faits représentent des données complexes, les dimensions peuvent comporter des descripteurs d'images, de textes, ... Les mesures ne sont pas forcément numériques ou peuvent ne pas exister. Compte tenu de ces particularités, les outils de

visualisation OLAP standard ne sont pas forcément adaptés et doivent être repensés. De plus, ils doivent permettre de visualiser aussi la sémantique contenue dans les données complexes.

Pour tenter d'apporter une solution à ce problème, nous nous sommes intéressés aux travaux sur la visualisation dans l'OLAP ou dans d'autres domaines qui, même s'ils ne traitent pas explicitement des données complexes, pourraient s'adapter à de telles données, s'adapter à l'OLAP ou prendre en le contenu sémantique des données.

A l'aide d'un test statistique, Ordonez et Chen cherchent dans un cube (de faible dimension) les cellule voisines qui ont des mesures significativement différentes [OC09]. Le test de comparaison de moyennes permet de détecter les cellules qui ont des valeurs très différentes alors qu'elles sont dans la même région du cube. L'idée proposée dans l'article est intéressante et montre une nouvelle fois l'intérêt d'utiliser des techniques de fouille de données pour enrichir les possibilités de l'OLAP. Cependant, l'utilisation d'un test statistique paramétrique oblige les auteurs à faire des hypothèses sur les données comme par exemple la distribution de probabilités de la mesure qui doit suivre une loi normale. Même si les auteurs qualifient cette hypothèse de faible et précisent qu'elle se vérifie généralement sur les données qu'ils utilisent, on peut penser que ce n'est pas le cas d'une façon générale. Malgré l'intérêt de l'approche et malgré le titre de l'article dans lequel elle est présentée, on peut penser qu'elle est plus une approche d'explication que de visualisation proprement dite.

Dans le contexte des applications OLAP sur le Web, Aouiche *et al.* s'intéressent particulièrement aux nuages de mots (*Tag clouds* en anglais) [ALG08]. Un nuage de mots est une représentation visuelle des mots-clés les plus utilisés dans un site Web. Généralement, les mots s'affichent dans une police de caractères d'autant plus grande qu'ils sont fréquement utilisés dans le site. Le nuage de mots-clés peut être considéré comme un résumé sémantique du site. Appliqué à l'OLAP, ce principe permet de représenter un cube sous forme d'un nuage où chaque mot-clé est une cellule et où la taille des mots-clés dépend de la valeur de la mesure. Cette nouvelle représentation du cube permet d'utiliser des opérations spécifiques aux nuages de mots comme trier les mots selon leur taille, enlever les mots de plus petites tailles, ... Pour aider l'utilisateur dans son exploration du cube, Aouiche *et al.* construisent des classes de mots-clés similaires et de réarrangent les modalités des dimensions [ALG08].

Dans une démarche à la croisée de la fouille de données et de la recherche d'information, Morin cherche à naviguer graphiquement dans des documents textes [Mor07]. Après avoir préparé les données (élimination des mots-outils, extraction des lemmes ou des mots-clés, ...) le tableau lexical (croisant les documents en ligne et les mots-clés en colonne) est analysé et visualisé à l'aide d'une analyse factorielle des correspondances.

Par rapport aux approches présentées, nous voulons créer un opérateur de visualisation (1) dans le contexte de l'analyse en ligne, (2) qui ne nécessite aucune

hypothèse sur les données, (3) qui soit adapté aux données complexes et (4) et qui prenne en compte le contenu sémantique des données complexes. Nous proposons l'opérateur CoDaViC : *Complex Data Visualization by Correspondences*. Il utilise les principes de l'analyse factorielle, ensemble de méthodes bien connues en fouille de données [Ben82, LMP04], pour visualiser les données complexes tout en mettant en évidence les points de vue intéressants pour l'analyse.

Le cube de données complexes avec plusieurs dimensions et avec la fonction d'agrégation COUNT, peut être vu comme un tableau de contingence ou d'occurrences. Il est alors possible d'utiliser une analyse des correspondances pour visualiser les faits.

L'analyse des correspondances produit des axes factoriels qui peuvent être utilisés comme nouvelles dimensions, à savoir des "dimensions factorielles". Ces axes ou dimensions constituent un nouvel espace de représentation dans lequel il est possible de projeter les faits ou les données complexes. Nous retenons seulement le premier plan factoriel (c'est-à-dire l'espace de représentation créé à partir des deux premiers axes produits par l'analyse des correspondances) qui est aussi appelé carte factorielle.

L'utilisation d'une méthode factorielle pour la visualisation se justifie pleinement car l'analyse des correspondances reprend l'objectif de navigation et d'exploration de l'OLAP. Notre principe peut être comparé à celui de l'analyse sémantique latente (LSA, de l'anglais *latent semantic analysis*) utilisée en recherche d'information [DDF$^+$90]. L'analyse sémantique latente et l'analyse factorielle des correspondances sont deux méthodes qui cherchent à réduire la dimension d'un espace pour mettre en évidence des liens qui peuvent exister dans les données. L'analyse factorielle des correspondances est préférée dans notre contexte car elle fournit en plus une représentation graphique synthétique ainsi que des indicateurs de qualité de la représentation graphique.

3.3.2 Démarche

Dans l'analyse des publications, supposons que l'utilisateur veuille synthétiser les travaux selon les mots-clés afin d'identifier les grands champs de recherche sur lesquels les chercheurs travaillent. De plus, l'utilisateur voudrait identifier les chercheurs qui travaillent sur les mêmes mots-clés. Pour répondre à ce besoin d'analyse, il faut fournir à l'utilisateur une visualisation des publications à travers les auteurs et les mots-clés, tout en permettant une navigation dans les données.

Pour cela nous proposons à l'utilisateur OLAP une démarche complète en plusieurs étapes :

- Choix du cube de données complexes,
- Construction du tableau de contingence,

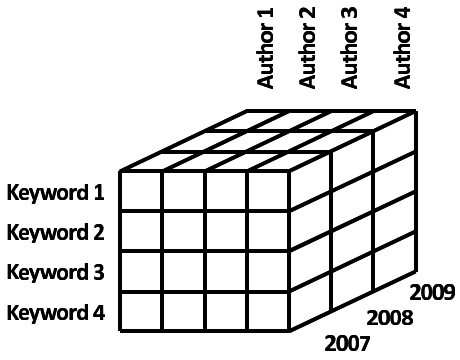


FIGURE 3.2 – Cube de données

- Réalisation de l'analyse factorielle des correspondances,
- Visualisation des données complexes sur la carte factorielle.

Cube ou sous-cube de données complexes. En fonction de ce qu'il veut analyser, l'utilisateur définit le sous-cube $\mathcal{C}'$ sur lequel il veut travailler.

Par exemple, l'utilisateur peut choisir de travailler dans le contexte des publications ayant été écrites pendant les années 2007 à 2009, par des auteurs qui ont le statut de professeur. Et dans ce contexte, l'utilisateur peut construire, comme dans la figure 3.2, un cube des publications selon les mots-clés, l'année de parution et le nom du premier auteur. Dans notre exemple, le sous-cube est alors défini par $(\Theta^1, \Theta^2, \Theta^3, \Theta^4)$= ({*full professor*},{*2007, 2008, 2009*},{*Keyword 1, Keyword 2, ..., Keyword 4*},{*Author 1, Author 2, ..., Author 4*}). M correspond au nombre de publications (*Count*).

Tableau de contingence. Classiquement, l'analyse factorielle des correspondances admet en entrée un tableau de contingence ou tableau d'occurrences. Dans notre démarche d'analyse en ligne de données complexes, l'idée est d'utiliser les opérateurs traditionnels OLAP pour construire ce tableau d'occurrences.

Dans le sous-cube $\mathcal{C}'$, l'utilisateur choisit deux niveaux hiérarchiques (c'est-à-dire il choisit un niveau pour deux dimensions différentes i et i'), niveaux sur lesquels il veut visualiser les données complexes. Soit Θ^i (respectivement $\Theta^{i'}$) l'ensemble des l (respectivement l') modalités choisies pour le niveau retenu de la dimension i (respectivement i'). Le tableau de contingence $\mathcal{T}$ a donc l lignes et l' colonnes dont

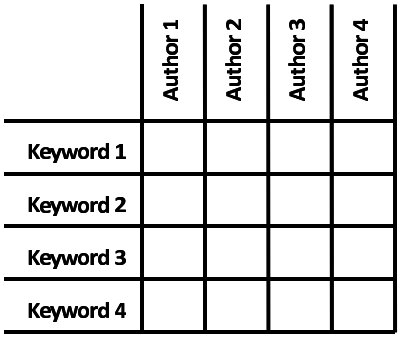


FIGURE 3.3 – Tableau d'occurrences

les intitulés sont donnés par $\{a_1^{ij}, ..., a_t^{ij}, ..., a_l^{ij}\}$ et $\{a_1^{i'j'}, ..., a_{t'}^{i'j'}, ..., a_{l'}^{i'j'}\}$. A chaque intersection de ligne t et de colonne t', sont dénombrés les faits ayant les modalités a_t^{ij} et $a_{t'}^{i'j'}$.

Dans notre exemple, le tableau de contingence correspond au croisement de l'ensemble des mots-clés avec l'ensemble des auteurs retenus dans le sous-cube. Étant donnée la structure de l'entrepôt de données, la démarche consiste à dénombrer les faits sur les trois années en effectuant un *roll-up* sur la dimension "année". Nous obtenons ainsi un tableau croisé admettant les mots-clés en ligne et les auteurs en colonne (figure 3.3). A l'intersection d'une ligne et d'une colonne, nous avons le nombre de publications écrites par un auteur sur un mot-clé. Ce tableau est prêt à être traité par une analyse factorielle.

Si la mesure utilisée est autre que le simple dénombrement et si la mesure est une mesure numérique, additive et n'ayant que des valeurs positives, alors il est possible de l'utiliser pour pondérer les faits dans le tableau de contingence. Le choix d'utiliser une telle mesure comme pondération est laissé à l'utilisateur.

Analyse factorielle. Réaliser une analyse factorielle des correspondances (AFC) consiste à projeter les données sur des axes synthétiques de manière à ce qu'un maximum d'informations s'exprime sur un nombre minimum d'axes orthogonaux. Il s'agit, comme toute méthode factorielle, de réduire la dimension de l'espace de représentation des lignes et celle de l'espace de représentation des colonnes. Les lignes et les colonnes jouent des rôles similaires. L'AFC permet la visualisation simultanée des projections des lignes et des colonnes dans un même plan factoriel. On peut alors interpréter les proximités entre les lignes et les colonnes.

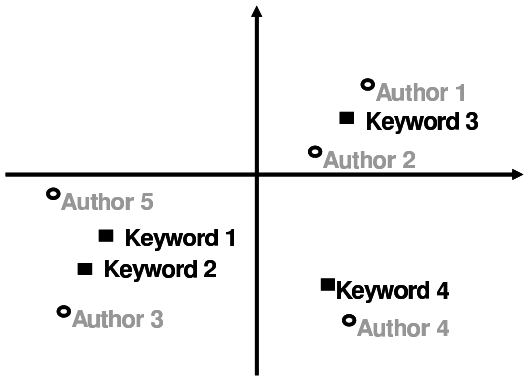


FIGURE 3.4 – Représentation sur le premier plan factoriel

Concrètement, la méthode démarre par le calcul des valeurs propres λ_α à partir desquelles sont déduits les vecteurs propres u_α qui définissent les axes factoriels F_α. Il y a $(min(l, l') - 1)$ valeurs propres, donc axes factoriels. Les deux premiers axes contiennent le plus d'informations, ils définissent donc le premier plan factoriel. Une fois les points lignes et colonnes projetés sur les axes, des indicateurs d'aide à l'interprétation permettent d'analyser les proximités entre les lignes et les colonnes. Les indicateurs les plus importants sont le poids, la contribution d'un point à l'inertie de l'axe et la qualité de représentation d'un point par l'axe (donnée par le $cosinus^2$). Pour donner une interprétation à un axe et pour analyser les proximités entre les points sur un axe, on s'intéresse aux points qui contribuent fortement à l'inertie de l'axe (dont la contribution est généralement trois fois supérieure à la contribution moyenne) et aux points qui sont bien représentés par l'axe (dont le $cosinus^2$ est élevé ou supérieur à 0,5).

Dans notre exemple, les publications sont représentées par les mots-clés et par les auteurs ; grâce à l'AFC, on va pouvoir dégager des associations entre les auteurs et les mots-clés. La figure 3.4 est une illustration très succinte de ce qu'on pourrait observer sur des données réelles : les mots-clés 1 et 2 sont très proches ce qui laisse supposer que les publications concernées ont été écrites par des auteurs communs. En revanche, les mots-clés 3 et 4 sont assez distants et par conséquent, paraissent relever de deux champs de recherche différents. Le placement des auteurs permet quant à lui d'avoir un aperçu de(s) champs de recherche sur les quel(s) ils travaillent. Par exemple, les auteurs 3 et 5 paraissent travailler sur des thèmes de recherche communs, tandis qu'ils s'opposent aux auteurs 1 et 2, qui eux-mêmes s'opposent à l'auteur 4. Nous obtenons ainsi une synthèse graphique des thèmes sur lesquels les auteurs travaillent et avec qui.

Visualisation. Les deux premiers axes factoriels sont retenus comme nouvelles dimensions factorielles, dans le sens où les coordonnées des faits projetés agissent comme des descripteurs de leur position sur ces axes. Une représentation graphique est construite (figure 3.5).

Comme expliqué précédemment, il est possible de donner une interprétation et du sens à la représentation graphique. Un outil interactif permet pour chaque point, c'est-à-dire mot-clé ou auteur, de visualiser ses indicateurs (contribution et $cosinus^2$). Les mots-clés et les auteurs ayant des indicateurs élevés sont représentés dans une autre couleur. Ainsi, l'utilisateur repère de suite les points les plus pertinents pour l'analyse. L'analyse factorielle facilite ainsi la lecture de l'information et apporte une aide automatique pour la compréhension et pour l'analyse de celle-ci. Par exemple, l'utilisateur peut identifier facilement les mots-clés les plus caractéristiques, les auteurs qui travaillent ensemble ou au contraire pas du tout ensemble et enfin des groupes d'auteurs qui travaillent sur certains mots-clés. A la demande de l'utilisateur, la photographie des auteurs peut se substituer à leur nom. Cela facilite encore la lecture du graphique et améliore la visibilité de l'information.

Pour respecter l'esprit de l'analyse OLAP, il est possible sur chaque auteur ou mot-clé d'effectuer un *drill-down* afin de voir les publications (à l'aide de leur titre) qui s'y rattachent. Dans un but explicatif, l'utilisateur a aussi la possibilité de projeter un niveau hiérarchique d'une autre dimension du cube. Les modalités de ce niveau seront projetés comme points dans l'espace factoriel mais elles n'auront pas participé à la construction des axes. Pour assurer une certaine cohérence, seuls les niveaux hiérarchiques dont les dimensions ne sont pas dans le sous-cube peuvent être utilisés comme éléments supplémentaires. Dans notre exemple, l'utilisateur pourrait utiliser comme élément supplémentaire le type de publications (journal, conférence, rapport technique, ...) et voir quels auteurs ou quels mots-clés publient ou sont publiés selon le type de support.

Notre approche permet de visualiser des données complexes tout en tenant compte de leurs proximités ou de leurs différences, elle intègre donc une partie de la sémantique contenue dans les données. Elle est adaptée aux données complexes et elle reprend l'esprit de l'OLAP avec des opérations d'agrégation possibles sur les éléments et une visualisation rapide et simple de l'information. Elle montre la faisabilité d'utiliser une méthode factorielle et de combiner une nouvelle fois les principes de l'OLAP et de la fouille de données. Cette combinaison est à nouveau reprise dans la section suivante, pour agréger les données complexes selon une hiérarchie.

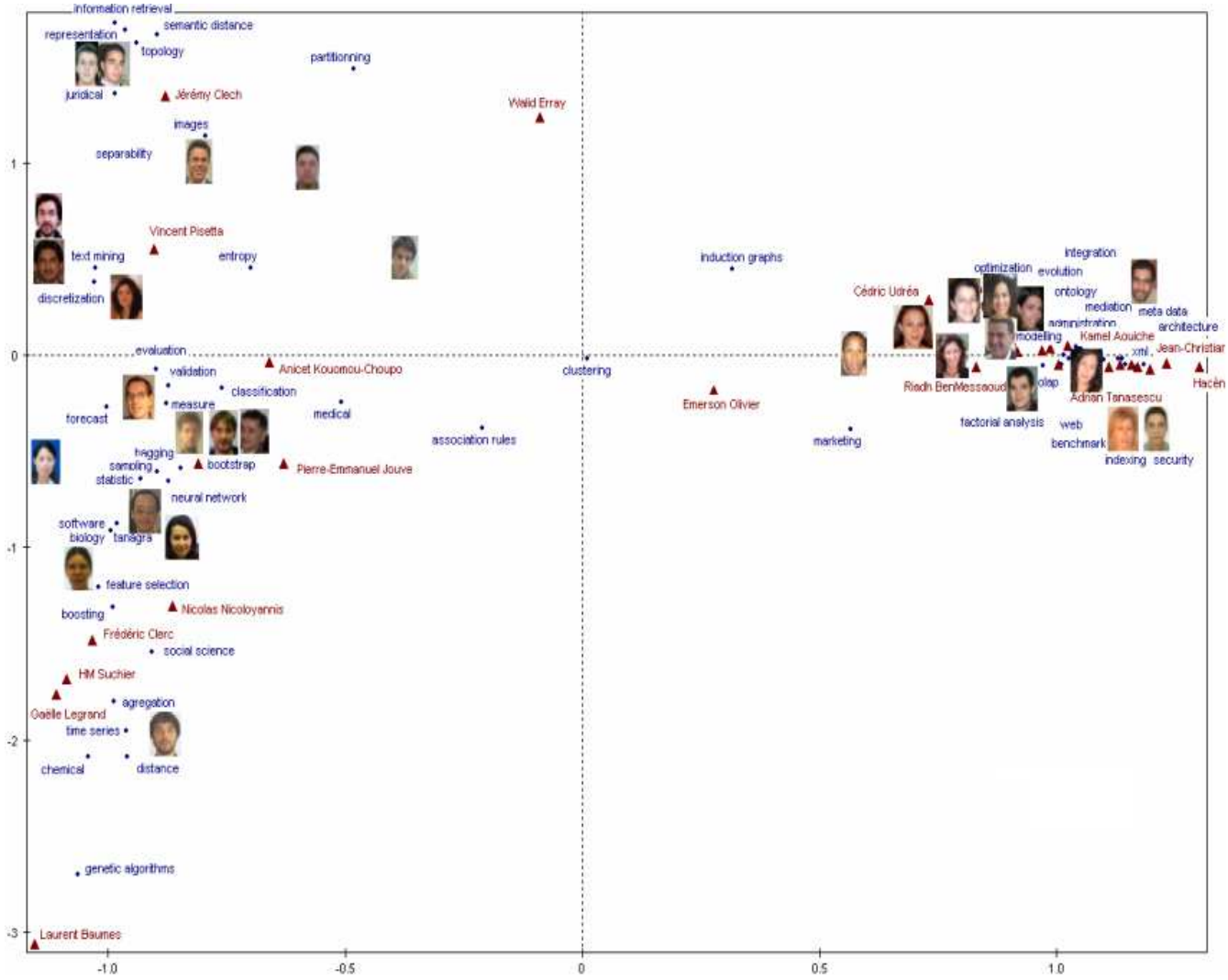


FIGURE 3.5 – Visualisation des publications

3.4 Agrégation OLAP par classification

3.4.1 Motivation et objectif

Classiquement, les dimensions dans un modèle multidimensionnel peuvent être structurées selon une ou plusieurs hiérarchies exprimant des niveaux d'abstraction différents. Par exemple, une dimension de localisation géographique peut être structurée en plusieurs niveaux : la ville, le pays, le continent, ... Le niveau *Pays* présente une granularité supérieure à celle du niveau *Ville* et chaque modalité du niveau *Ville* est rattachée à une modalité du niveau *Pays*. Il en est de même pour le niveau *Pays* et le niveau *Continent*.

Cette organisation des modalités selon l'ordre hiérarchique des dimensions est une organisation structurelle qui est mise en place lors de la modélisation conceptuelle de l'entrepôt de données. L'ordre hiérarchique est donné soit par l'ordre naturel des choses soit par les experts du domaine étudié. Cette organisation structurelle des modalités ne permet malheureusement pas de rendre compte de la sémantique ou des liens de proximité des faits. Par exemple, dans le contexte d'une entreprise de vente à distance, on considère classiquement que la *France*, l'*Italie* et l'*Espagne* appartiennent au même groupe de pays, à savoir l'*Europe*. Une hiérarchie classique ne permet pas de considérer, par exemple, que la *France* et le *Canada* doivent être dans un même groupe de pays où le niveau des ventes d'ouvrages littéraires francophones est semblable. L'agrégation classique de la *France*, l'*Italie* et l'*Espagne* en *Europe* est structurelle alors que l'agrégation de la *France* avec le *Canada* serait sémantique et beaucoup plus intéressante en terme d'analyse.

Ce problème s'amplifie lorsque l'on travaille avec des données complexes. Au sens classique du terme, il n'existe parfois pas de hiérarchie définie a priori pour agréger les données complexes et il n'existe pas de fonction d'agrégation adaptée. Par exemple, dans l'analyse des publications, plusieurs questions se posent : comment agréger les publications ? Comment définir une hiérarchie pour les étudier à des niveaux de granularité différents ? Intuitivement, il devrait être possible de regrouper en un groupe les publications qui se ressemblent selon les auteurs, les thèmes sur lesquels elles portent, les types de support, ... Ce problème d'agrégation de données correspond à un problème de classification automatique, bien connu dans le domaine de la fouille de données. En effet, il s'agit de faire émerger des groupes ou classes d'objets semblables au sens d'une métrique donnée.

Dans cette section, nous proposons une approche pour l'agrégation des données complexes et la création d'une hiérarchie de dimension. Nous agrégeons les faits d'un cube selon leur proximité et non plus selon un ordre hiérarchique prédéterminé. Pour cela, nous utilisons la Classification Ascendante Hiérarchique (CAH) pour construire

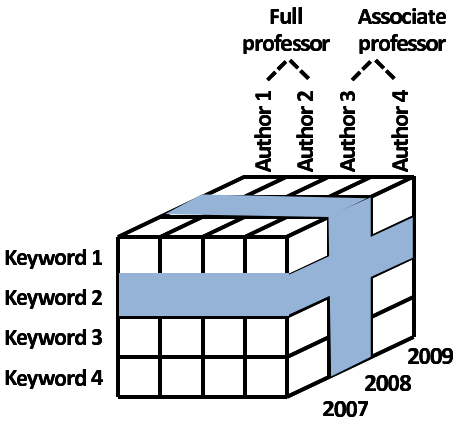


FIGURE 3.6 – Agrégation classique

des classes correspondant à des agrégats du cube. La succession de partitions fournie par la CAH nous donne une hiérarchie de dimension. Nous créons l'opérateur OpAC (*Operator for Aggregation by Clustering*). Il permet une agrégation sémantique des données complexes.

Par exemple, l'utilisateur peut vouloir travailler sur le cube des publications de la figure 3.6. Ce cube est constitué de trois dimensions : les auteurs, les mots-clés et l'année de la publication. Dans ce cube, la fonction d'agrégation COUNT est utilisée pour dénombrer le nombre de publications dans chaque cellule. La dimension des auteurs comporte une hiérarchie avec deux niveaux : le nom des auteurs et leur statut avec par exemple *Full professeur*, *Assistant professor*, *PhD student*, *Associate professor*, ...

Les auteurs peuvent être agrégés selon leur statut, par exemple les auteurs 1 et 2 sont agrégés ensemble pour former l'agrégat des professeurs. Ce regroupement est structurel. Nous pensons qu'il peut aussi être intéressant pour l'utilisateur de regrouper les auteurs avec un point de vue plus scientifique, comme par exemple d'agréger les auteurs selon les sujets scientifiques sur lesquels ils travaillent et selon les années. Supposons que l'utilisateur veuille considérer seulement le nombre de publications écrites en 2008 sur le mot-clé 2 (figure 3.6). En utilisant une technique de classification on obtient la figure 3.7 sur laquelle on peut constater que les auteurs 1 et 4 sont proches, au sens des deux descripteurs choisis par l'utilisateur (mot-clé 2 et année 2008), car ils ont écrit beaucoup de publications en 2008 sur le mot-clé 2. Les auteurs 3 et 5 sont également proches mais parce qu'ils ont écrit beaucoup de publications en 2008 ne portant pas ou peu sur le mot-clé 2. Scientifiquement, il serait plus intéressant de regrouper l'auteur 1 avec l'auteur 4 et non pas les auteurs 1 et 2

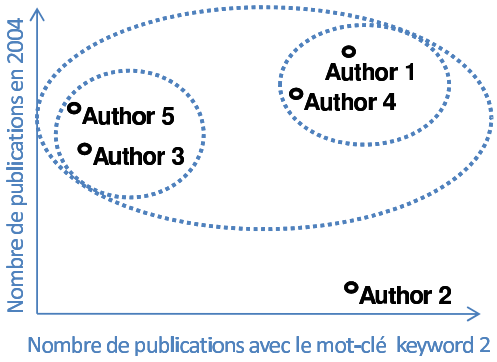


FIGURE 3.7 – Agrégation avec OPAC

ensemble sur le simple principe qu'ils sont des professeurs.

Contrairement à l'agrégation OLAP classique basée sur la structure des données, notre approche constitue une nouvelle forme d'agrégation sémantique qui prend en compte le contenu des données. Les agrégats ainsi créés mettent en évidence les liens réels entre les faits. Cette forme d'agrégation permet d'obtenir des informations sémantiquement plus riches que celles fournies par l'agrégation classique d'OLAP.

Comme méthode de classification, nous choisissons d'utiliser la Classification Ascendante Hiérarchique (CAH) pour les raisons suivantes :

- Parmi les méthodes de classification existant en fouille de données, nous avons besoin d'une méthode hiérarchique donnant une hiérarchie de partitions. Cette hiérarchie de partitions reprend l'esprit de l'analyse en ligne et est analogue à la structure d'une dimension d'un cube. Elle peut être parcourue par des opérateurs d'agrégation classiques (*roll-up* et *drill-down*). Chaque partition correspond à un niveau hiérarchique. En passant d'un niveau de partition à celui qui lui est immédiatement supérieur, deux classes sont agrégées ensemble pour former un nouvel agrégat. Inversement, en passant d'un niveau de partition à celui qui lui est immédiatement inférieur, un agrégat est divisé en deux classes.
- Parmi les méthodes de classification hiérarchique, la stratégie ascendante est plus rapide que la stratégie descendante (classification descendante hiérarchique-CDH). La complexité de la CAH est généralement polynomiale, tandis que celle de la CDH est exponentielle [CGLT99]. En effet, lors de la première étape d'une méthode ascendante, il faut évaluer toutes les agrégations possibles de deux individus parmi n, soit $n(n-1)/2$ possibilités, tandis qu'un algorithme descendant évalue toutes les divisions des n individus en deux sous-ensembles

non vides, soit $2^{n-1} - 1$ possibilités.

3.4.2 Démarche

Individus et variables de la classification. En fonction de ce qu'il veut analyser, l'utilisateur définit comme précédemment le sous-cube $\mathcal{C}'$ sur lequel il veut travailler.

Nous notons Ω l'ensemble des individus et Σ l'ensemble des variables de la classification à définir.

Supposons que nous cherchions à agir sur les modalités du niveau hiérarchique H_j^i de la dimension D^i. Ce choix est fait par l'utilisateur en fonction de ses besoins d'analyse et objectifs. C'est donc à l'utilisateur de fixer la dimension D^i, le niveau hiérarchique H_j^i et les modalités qu'il souhaite classifier dans l'ensemble $\mathcal{A}^{ij}$.

Ainsi, statistiquement parlant, l'ensemble des individus Ω à agréger par la CAH correspond à l'ensemble des modalités choisies par l'utilisateur dans $\mathcal{A}^{ij}$:

$$\Omega \subset \mathcal{A}^{ij} = \{a_1^{ij}, \ldots, a_t^{ij}, \ldots, a_l^{ij}\}$$

Par exemple, dans le cube de la figure 3.6, D^1 est la dimension des auteurs, D^2 celle des mots-clés et D^3 celle du temps. M_q correspond au nombre de publications. L'utilisateur choisit d'agréger les auteurs, c'est-à-dire les modalités du niveau (H_1^1) de la dimension (D^1). Dans ce cas, l'ensemble Ω des individus à classifier appartiennent à l'ensemble $\mathcal{A}^{11} = \{$*Author 1, Author 2, Author 4, Author 4,*$...\}$.

Soit $\mathcal{A}$ l'ensemble des modalités des hiérarchies du cube $\mathcal{C}$ y compris les agrégats totaux des dimensions :

$$\mathcal{A} = \prod_{i=1}^{d} \mathcal{A}^{ij} = \mathcal{A}^{1j} \times \ldots \times \mathcal{A}^{ij} \times \ldots \times \mathcal{A}^{dj}$$

On considère qu'une mesure numérique M du cube peut s'écrire selon une fonction de l'ensemble $\mathcal{A}$ dans l'ensemble des réels $\mathbb{R}$.

$$M : \mathcal{A} \longrightarrow \mathbb{R}$$

Par exemple, $M($*Auhtor 1, Keyword 4, All*$)$ donne le nombre de publications écrites par l'auteur 1 et comportant le mot-clé 4 durant toutes les années.

Notre objectif est de proposer une agrégation sémantique, c'est-à-dire une agrégation qui tienne compte de l'information contenue dans les faits. Pour cela, nous considérons les mesures numériques du cube comme des variables quantitatives décrivant la population des individus Ω. Cependant, certaines contraintes logiques ou statistiques sont à respecter dans le choix de ces variables :

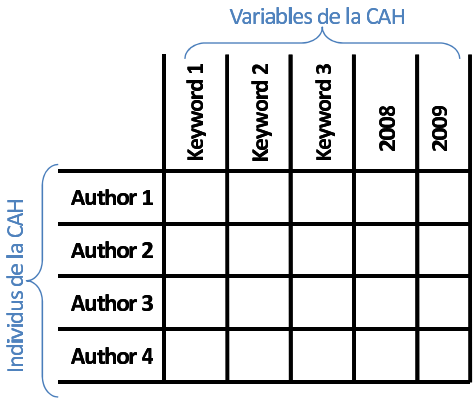


FIGURE 3.8 – Individus et variables de la classification

- Aucun niveau hiérarchique de la dimension retenue pour les individus ne doit générer des variables de la classification.
- Par dimension, on ne peut choisir qu'un seul niveau hiérarchique pour générer des variables. Cette contrainte est essentielle pour assurer l'indépendance des variables de la classification.

Dans notre exemple, d'après la première contrainte, on ne peut plus utiliser la dimension D^1 pour la génération des variables. De plus, en respectant la seconde contrainte, on ne peut choisir qu'un seul niveau hiérarchique de D^2 et/ou de D^3 comme générateur de variables. Dans la figure 3.8, pour générer les variables, on a choisi pour le niveau des *mots-clés* de la dimension D^2 les trois premiers et pour le niveau des *années* de la dimension D^3 les années 2008 et 2009.

Classification ascendante hiérarchique. Les données extraites sont mises sous la forme d'un tableau individus-variables (figure 3.8) désigné par X, une matrice d'ordre (n, p). Les n lignes de X représentent les individus de Ω et les p colonnes de X représentent les variables de Σ. Soit S la matrice des distances. S est une matrice carrée symétrique dont le terme général s_{ij} correspond à la distance entre l'individu i et l'individu j.

La technique de la CAH produit une succession de partitions avec des classes des n individus selon les p variables de Σ. Dans chaque classe, les individus sont les plus semblables possible. Les classes sont les plus différentes possibles entre elles.

Le principe de l'algorithme consiste à créer, à chaque étape, une partition obtenue en agrégeant deux à deux les éléments les plus proches. On désigne par élément à la fois les individus et les groupes d'individus générés par l'algorithme. L'algorithme

ne fournit pas une partition en k classes d'un ensemble de n individus mais une hiérarchie de partitions, se présentant sous la forme d'un arbre appelé dendrogramme et contenant $n-1$ partitions. L'intérêt de cet arbre est qu'il peut donner une idée du nombre de classes existant effectivement.

Il y a différentes manières de calculer la distance entre les individus et différentes règles pour évaluer la distance entre les groupes (règles appelées critères d'agrégation), d'où un nombre important de variantes de cette technique. Ces choix reviennent à l'utilisateur. Les distances les plus connues sont la distance *euclidienne*, la distance de *Chebychev*, la distance de *Manhattan* et la distance de *Minkowski*. Les critères d'agrégation usuels sont celui du *saut minimal*, du *saut maximal*, *saut moyen* et le critère de *Ward*.

L'algorithme de la classification ascendante hiérarchique peut être résumé par les étapes suivantes :

- **Etape 0** : les n individus de la matrice X sont affectés chacun à des classes distinctes. La partition P_0 est constituée de n classes. La matrice S de toutes les distances est calculée selon la distance choisie.
- **Etape 1** : on construit la matrice S de distances entre les n éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient la partition P_1 à $(n-1)$ classes.
- **Etape 2** : on construit une nouvelle matrice S des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement $(n-1)$ éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une partition P_2 avec $(n-2)$ classes qui englobe la première.
- **Etape n-1** : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'une seule classe regroupant tous les individus et qui constitue la dernière partition à une classe.

La hiérarchie des classes peut être représentée par un arbre, appelé dendrogramme, où en ordonnée figure la valeur des distances correspondant aux différentes agrégations. En coupant l'arbre par une droite horizontale, on obtient une partition, d'autant plus fine que la section est proche des éléments terminaux. Le choix de l'endroit où couper l'arbre, c'est-à-dire le choix du nombre k de classes ($1 \leq k \leq n$), revient à l'utilisateur. Malheureusement, la CAH ne donne aucune indication sur la qualité et la pertinence des classes fournies. Par conséquent, il est souvent difficile pour l'utilisateur de choisir la meilleure partition au sens de son analyse. Le choix de cette partition est encore plus difficile quand l'utilisateur est face à un grand nombre d'individus à classifier. Des critères d'évaluation des classes ou des agrégats sont donc nécessaires.

Evaluation des agrégats de modalités. Dans la littérature, plusieurs travaux ont proposé des critères d"évaluation de la qualité des résultats de la classification [War63, LMP04, LFAH04]. Mais il n'existe pas de critère universel pour l'évaluation des classes. Chaque mesure de qualité dépend fortement de sa propre définition et des orientations d'analyse que sous-entend l'utilisateur [LFAH04]. Par conséquent, dans notre approche, nous proposons d'utiliser plusieurs critères d'évaluation des classes fournies par la CAH ; critères qui expriment des points de vue différents de la qualité des classes.

Pour la formulation de ces critères, nous adoptons les notations suivantes :

- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ l'ensemble des individus à classer ;
- chaque individu ω est caractérisé par un poids $P(\omega)$ et est décrit par p variables quantitatives $V_1, V_2, \ldots, V_p$;
- soit $k \in \{0, \ldots, n-1\}$ l'indice des itérations, ou des partitions, de la CAH. A une itération k correspond une partition P_k à $(n-k)$ classes notées $A_1, A_2, \ldots, A_{n-k}$. P_k est la partition courante de Ω ;
- n_i est le nombre d'individus de la classe A_i ;
- la classe A_i est caractérisée par son poids $P(A_i) = \sum_{\omega \in A_i} P(\omega)$;
- $G(A_i) = \frac{1}{P(A_i)} \sum_{\omega \in A_i} P(\omega)V(\omega)$ est le centre de gravité de la classe A_i ;
- $G = \sum_{\omega \in \Omega} P(\omega)V(\omega)$ est le centre de gravité de tous les individus de l'ensemble Ω ;
- d représente la distance euclidienne et d^2 représente la distance euclidienne carrée.

Comme critères, nous utilisons celui de l'inertie intra et inter-classes [LMP04] et celui de la méthode de Ward [War63]. Nous proposons également un nouveau critère basé sur la notion de la séparabilité des classes.

Le critère de l'inertie intra et inter-classes se fonde sur l'idée de minimiser la distance entre les individus d'une même classe et de maximiser celle des groupes entre eux c'est-à-dire de maximiser les distances entre les centres de gravité des différentes classes.

Pour une classe A_i, l'inertie intra-classe est définie :

$$I(A_i) = \sum_{\omega \in A_i} P(\omega) d(V(\omega), G(A_i))$$

Cette inertie intra-classe est la somme pondérée des écarts des individus par rapport au centre de gravité de leur classe. Elle mesure le degré d'homogénéité de la classe : plus elle est petite, plus la classe est homogène.

L'inertie intra-classe totale d'une partition P_k, notée $I_{intra}(k)$, est égale à la somme des inerties des $(n-k)$ classes :

$$I_{intra}(k) = \sum_{i=1}^{n-k} I(A_i)$$

L'inertie inter-classes d'une partition P_k, notée $I_{inter}(k)$ est la somme pondérée des distances séparant les centres de gravité $G(A_i)$ des classes A_i du centre de gravité G de Ω :

$$I_{inter}(k) = \sum_{i=1}^{n-k} P(A_i) d(G(A_i), G)$$

Cette inertie permet de mesurer la dissemblance des classes, plus elle est grande, plus les classes sont éloignées.

D'après le théorème de *Huygens*, la somme des deux inerties est une constante égale à l'inertie totale du nuage des individus de Ω :

$$I_{intra}(k) + I_{inter}(k) = I(\Omega)$$

On peut aussi montrer que l'inertie intra-classe (respectivement, inter-classes) est globalement croissante (respectivement, décroissante) en fonction de l'indice des partitions k. Le critère de l'inertie intra et inter-classes consiste à calculer, pour chaque partition de la CAH, les inerties intra et inter-classes. La détection d'un changement important de la valeur d'une des inerties d'une itération à l'autre est un indicateur pertinent pour arrêter le processus d'agrégation. Le nombre de classes de l'itération précédant le changement important de valeur d'une des deux inerties est un bon compromis entre des classes homogènes et différentes entre elles. Le critère des inerties est cependant un critère globalement monotone qui, dans certaines situations, n'offre pas de comparaisons pertinentes sur la qualité des différentes partitions de la CAH.

Le critère de la méthode de Ward évalue différemment la qualité des classes en mesurant le coût d'agrégation en passant d'une partition à une autre dans le processus de construction de la CAH. Le critère mesure la variation de l'inertie interne quand deux classes A_i et A_j sont agrégées. A chaque itération, le coût d'agrégation est calculé selon la distance euclidienne carrée entre les centres de gravité des classes à agréger pondérés par leur poids respectifs :

$$W(A_i, A_j) = \frac{n_i n_j}{n_i + n_j} d^2(G(A_i), G(A_j))$$

A une itération k, une forte valeur du critère de *Ward* indique une importante variation de l'inertie interne quand on passe de la partition $k-1$ à la partition k. Cette variation suggère à l'utilisateur retenir la partition de l'itération $k-1$ avec $(n-k+1)$ classes.

Les deux critères présentés sont liés à l'inertie et se reposent sur l'idée de maximiser l'homogénéité des individus dans les classes. Afin de fournir à l'utilisateur un point de vue complémentaire, nous proposons un nouveau critère pour mesurer la qualité

des agrégats de la CAH en se basant sur le principe de la séparabilité des classes.

Ce nouveau critère se fonde sur le principe de la séparabilité des classes introduit par Zighed *et al.* [ZLM02] et utilise les graphes de voisinage. Un graphe de voisinage permet de visualiser l'existence ou l'absence de liaisons entre les individus statistiques de Ω. Un graphe est formé d'un ensemble de sommets, qui représentent les individus, reliés entre eux par des arêtes (arcs non orientés) [BG88]. Un sommet est relié à un autre par une arête s'il est voisin de ce dernier selon une structure de voisinage telles que la structure des *k-plus proches voisins*, la structure de *Gabriel*, la structure des *voisins relatifs* ou la structure des polyhèdres de *Delaunay*.

Pour notre critère de séparabilité des classes, nous utilisons le *graphe de Gabriel* [GS69]. Dans un graphe de *Gabriel*, deux individus représentés par les sommets A et B sont reliés par une arête si l'hypersphère de diamètre AB ne contient aucun sommet de Ω. La figure 3.9 (a) montre un exemple de graphe de *Gabriel* construit pour des individus décrits par deux variables X_1 et X_2.

Soit g_Ω le graphe de *Gabriel* construit sur l'ensemble Ω des individus à classifier selon les variables de la classification Σ.

Le calcul de notre critère se fait pendant la construction de la hiérarchie de partitions. A chaque itération k de la CAH (figure 3.9 (b)), le graphe de voisinage g_{A_i} des individus est construit dans chaque classeA_i (figure 3.9 (c)). Il est à remarquer que pour une partition donnée, l'union des sous-graphes engendrés par les classes A_i ($i \in \{1, \ldots, n-k\}$) de cette partition ne correspond pas forcément au graphe complet de l'ensemble des individus de Ω : $\cup_{i=1}^{n-k}\{g_{A_i}\} \neq g_\Omega$.

Soit e_{ij} l'arête reliant le sommet i (représentant l'individus ω_i) au sommet j (représentant l'individu ω_j) dans un graphe de voisinage. Chaque arête e_{ij} est associée à un poids $P(e_{ij})$ égale à l'inverse de la distance euclidienne qui sépare les deux sommets ω_i et ω_j, $P(e_{ij}) = \frac{1}{d(\omega_i,\omega_j)}$.

Le poids d'une arête mesure le *degré de séparabilité* des deux sommets. En effet, deux sommets séparés par une grande distance correspondent à deux individus peu semblables et qui sont donc facilement séparables dans un processus de classification ; le poids de l'arête est donc faible. Inversement deux sommets séparés par une petite distance caractérisent deux individus semblables et donc moins facilement séparables ; le poids de l'arête est donc important.

A chaque itération de la CAH (figure 3.9 (b)), notre critère calcule la somme des poids des nouvelles arêtes construites dans les graphes de voisinage des classes de la partition en cours (figure 3.9 (c)). Soit $\mathcal{E}^k$ l'ensemble des nouvelles arêtes construites à l'itération k de la CAH. On note $J(k)$ la somme des poids des nouvelles arrêtes construites par les graphes de voisinage à l'itération k :

$$J(k) = \sum_{e \in \mathcal{E}^k} P(e)$$

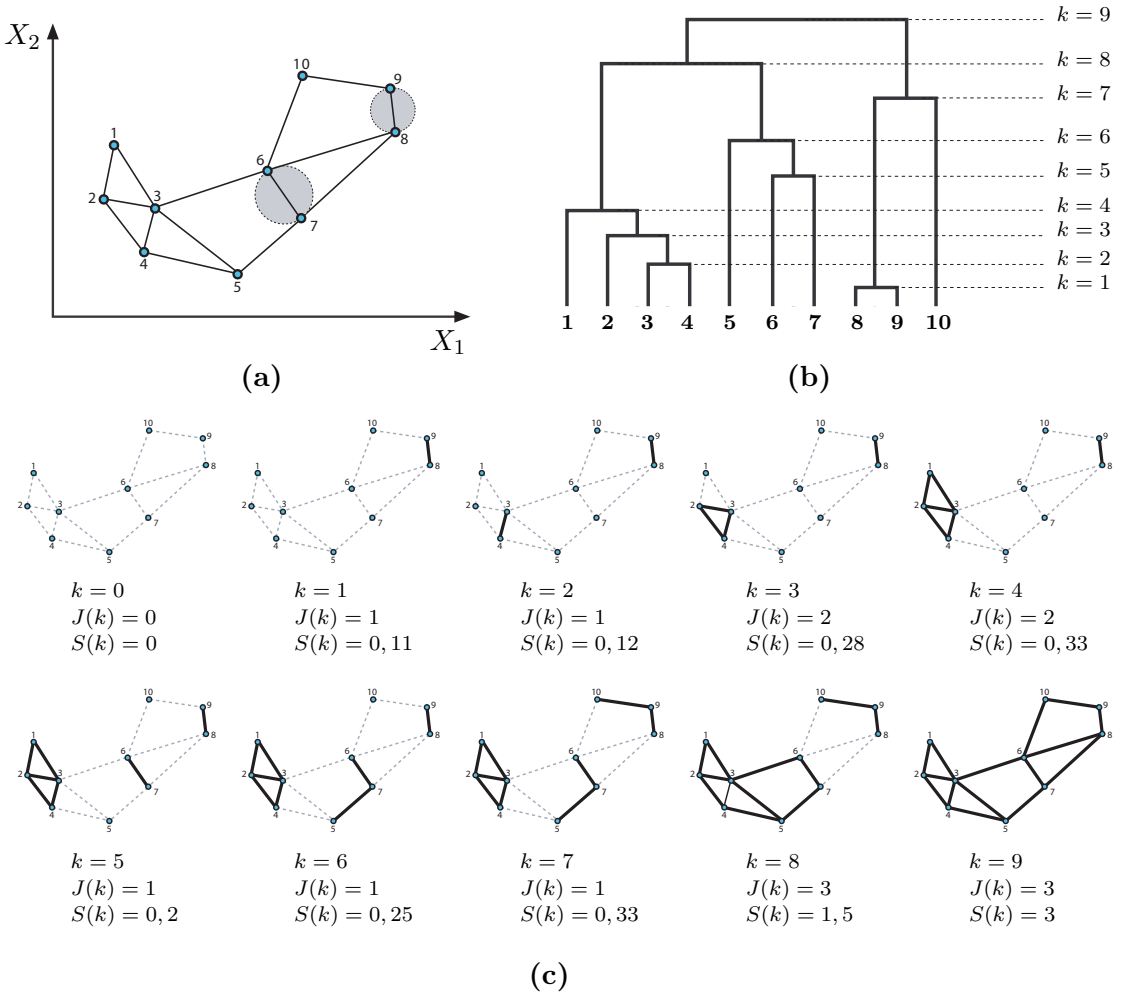


FIGURE 3.9 – Principe du critère de séparabilité des classes

Le but de notre critère est d'évaluer la séparabilité des classes pour chaque partition fournie par la CAH. Deux classes ont plus de chance d'être facilement séparables quand elles sont reliées par des arêtes dont la somme des poids est relativement faible. Cependant, les nouvelles arêtes construites, même si leurs poids rendent compte du degré de séparabilité des classes de la partition en cours, ne tiennent pas compte du nombre de classes de cette partition alors que nous pensons qu'il faut prendre en compte ce nombre.

Pour une partition k, notre critère de séparabilité des classe $S(k)$ est :

$$S(k) = \frac{J(k)}{n-k} = \frac{\sum_{e \in \mathcal{E}^k} P(e)}{n-k}$$

$S(k)$ calcule la moyenne des poids des arêtes nouvellement construites au moment où l'algorithme de la CAH passe de l'itération $(k-1)$ à l'itération k. Ce critère est relativement simple à interpréter. Une faible valeur de $S(k)$ indique qu'en passant de la partition $(k-1)$ à la partition k, des arêtes de faibles poids ont été construites. Par conséquent, les classes regroupées à cette étape ont plus de chance d'être séparables. Ainsi, dans ce cas, on peut préférer la partition $(k-1)$ – dont les classes présentent une meilleure propriété de séparabilité – à la partition k. Dans l'exemple, $S(k)$ atteint une valeur relativement faible à la partition $k = 5$. Ce constat peut amener l'utilisateur à préférer la partition précédente $k = 4$ qui se compose de six classes.

Pour les données complexes modélisées sous forme de cube, nous proposons d'agréger les données en prenant en compte leur sémantique. La Classification Ascendante Hiérarchique nous permet de construire automatiquement des agrégats significatifs de faits ainsi qu'une hiérarchie de dimension. L'agrégation est sémantique car elle tient compte des proximités entre les données. Pour aider l'utilisateur dans l'évaluation des agrégats nous proposons différents critères dont un basé sur le principe de séparabilité des classes.

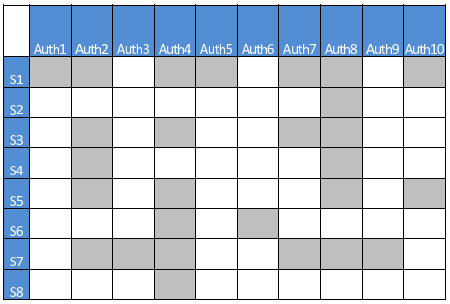
3.5 Réorganisation d'un cube par analyse factorielle

3.5.1 Objectif et positionnement

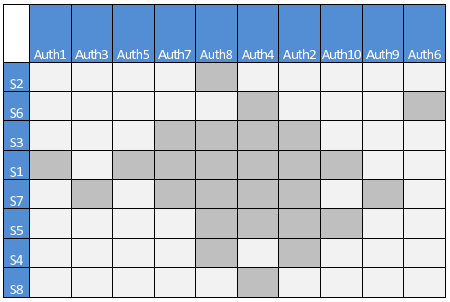
L'OLAP permet de naviguer dans un cube dans le but de trouver des informations intéressantes. C'est à l'utilisateur de trouver par lui même ces informations ainsi que les régions intéressantes du cube à explorer plus précisément. Cette exploration intuitive n'est pas une tâche aisée, peut ne pas donner de résultats intéressants car elle ne repose que sur les connaissances de l'utilisateur et prend d'autant plus de temps que le cube est volumineux et épars. Cette difficulté est encore plus importante quand on analyse des données complexes. Les cubes de données complexes sont souvent épars : dans une représentation multidimensionnelle du cube, à l'intersection de différentes modalités de dimensions, il n'existe pas forcément de faits correspondants. Cette éparsité peut être accentuée par la présence d'un grand nombre de dimensions et/ou d'un grand nombre de modalités dans chacune des dimensions. De plus, quand les modalités des dimensions sont ordonnées selon un ordre pré-établi (ordre naturel, ordre chronologique, ordre alphabétique, ...), les cellules pleines peuvent être éparpillées dans l'espace des dimensions du cube.

Prenons l'exemple de la figure 3.10 qui présente un cube de données à deux dimensions : les supports des publications $(S1, \ldots, S8)$ et les auteurs $(Auth1, \ldots, Auth10)$. Les cellules grisées sur la figure sont pleines et contiennent comme valeur le nombre de faits alors que les cellules blanches sont vides. La répartition des cellules pleines

dans la figure 3.10(a) ne se prête pas facilement à une interprétation car l'information est éparpillée dans l'espace de représentation. En revanche, dans la figure 3.10(b), les cellules pleines sont concentrées dans une zone centrale du cube. Cette représentation offre de meilleures possibilités de comparaison et d'analyse des cellules pleines.



(a)



(b)

FIGURE 3.10 – Exemple de deux représentations d'un espace de données

Notons que les deux représentations de la figure 3.10 correspondent au même cube de données. La représentation (b) est obtenue par simples permutations des lignes et des colonnes de la représentation (a).

L'utilisateur a donc besoin d'outils pour le guider vers les régions intéressantes du cube et pour réorganiser intelligemment les dimensions ou les modalités du cube. Notre objectif est de proposer une approche qui permet de détecter les faits intéressants ainsi que leurs liens et qui permet de les représenter de manière appropriée pour améliorer la visualisation de l'information dans un cube. Notre objectif n'est pas

de diminuer l'éparsité d'un cube multidimensionnel, mais d'atténuer l'effet négatif de cette éparsité sur la visualisation. L'idée est de réorganiser la disposition des cellules pleines et de les séparer le mieux possible des cellules vides dans l'espace de représentation du cube de données.

La plupart des travaux qui s'intéressent à l'espace de représentation du cube, portent très souvent sur des aspects d'optimisation (stockage, compression, temps de réponse, ...). Malgré l'importance des questions d'optimisation, notre approche ne traite pas ces aspects ; elle a un objectif différent. Nous ne présentons donc pas ces travaux dans le mémoire. En revanche, à notre connaissance, très peu de travaux s'intéressent au problème de visualisation de l'information. Néanmoins, citons les travaux de Choong *et al.* [CLLM04, CLM03] où les auteurs utilisent les règles floues (combinaison d'un algorithme de règles d'association et de la théorie des sous-ensembles flous) afin de faciliter la visualisation et la navigation dans l'espace de représentation du cube. Leur approche consiste à identifier et à construire des blocs de données similaires au sens de la mesure du cube. Cependant, cette approche ne prend pas en compte le problème d'éparsité du cube.

Notre idée de réorganiser les modalités d'un niveau est reprise par Sureau *et al.* Pour une meilleure visualisation de l'information, ils proposent des réorganiser les modalités d'un niveau soit selon des heuristiques basées sur la dissimilarité entre les modalités soit selon un algorithme génétique [SBV09]. Notre approche est considérée comme une approche heuristique. L'algorithme génétique fait évoluer le cube afin de maximiser un critère d'évaluation de la visualisation. Les algorithmes heuristiques sont cependant plus compétitifs que l'algorithme génétique dans le cadre d'une analyse OLAP car ils donnent un meilleur ratio gain de visualisation / temps de calcul. A terme les auteurs veulent intégrer leurs algorithmes de réorganisation dans leur plateforme de réalité virtuelle VRMiner afin de disposer d'un environnement complet pour l'OLAP.

Nous proposons une méthode de réorganisation des données multidimensionnelles en exploitant les principes de l'Analyse des Correspondances Multiples (ACM) [Ben82]. Nous préconisons d'utiliser l'ACM car cette méthode descriptive de fouille de données fournit une représentation graphique synthétique d'une grande quantité de données décrites par des variables qualitatives. Elle synthétise l'information et elle met en évidence les données intéressantes ainsi que les liens qui les caractérisent. L'ACM est ici utilisée comme un outil d'aide à la construction du cube pour une meilleure visualisation. Quand l'ordre initial des dimensions ne donne pas une bonne visualisation, les résultats de l'ACM permettent de réorganiser les modalités de chaque dimension du cube et de mettre ainsi en évidence les régions intéressantes pour l'analyse. L'utilisateur pourra alors se focaliser sur l'analyse des faits de ces régions.

Le fonctionnement de notre approche, baptisée ORCA (*Operator for Reorganization by multiple Correspondence Analysis*), est schématisé dans la figure 3.11.

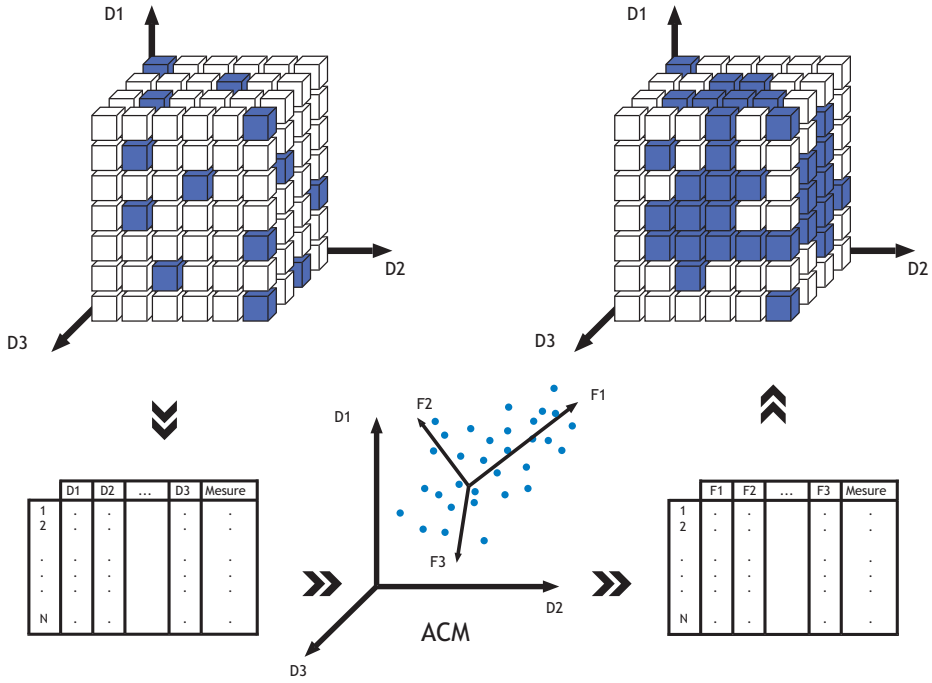


FIGURE 3.11 – Réorganisation d'un cube de données par approche factorielle

La première étape consiste à transformer les données initiales du cube en un tableau *individus-variables* avec un codage disjonctif complet. Ensuite, l'ACM est effectuée sur les données ainsi transformées. Nous obtenons alors des axes factoriels qui représentent au mieux les faits et qui traduisent des relations entre les modalités des dimensions du cube. Les associations entre modalités sont exploitées afin d'arranger les modalités dans chaque dimension du cube selon deux façons : selon les *projections* sur les axes factoriels selon les *valeurs-test* des modalités.

3.5.2 Démarche

L'utilisateur choisit les d dimensions D^i qui l'intéresse, il fixe un niveau hiérarchique H_j^i pour chacune de ces dimensions et il sélectionne une mesure quantitative M. Soit $\mathcal{C}'$ le sous-cube ainsi défini par l'utilisateur et n le nombre de faits contenus dans ce cube. Dans le but d'alléger les notations, nous assimilons volontairement, dans cette section, une dimension D^i à son niveau hiérarchique H_j^i. Ainsi, on notera que chaque dimension D^i contient l_i modalités catégorielles au lieux de l_{ij}. Soit donc

$\{a_1^i, \ldots, a_t^i, \ldots, a_{l_i}^i\}$ l'ensemble des modalités de la dimension D^i pour un niveau fixé. On note aussi $l = \sum_{i=1}^{d} l_i$ le nombre total de toutes les modalités dans $\mathcal{C}'$.

Tableau de contingence et tableau disjonctif complet. Pour pouvoir réaliser une ACM sur le cube $\mathcal{C}'$, il faut transformer ce dernier et l'exprimer sous forme d'un tableau disjonctif complet puis du tableau de contingence de Burt (ou hypercube de contingence).

Pour chaque dimension D^i ($i \in \{1, \ldots, d\}$), est générée une matrice Z^i à n lignes et l_i colonnes. Z^i est telle que sa $k^{ième}$ ligne contient $(l_i - 1)$ fois la valeur 0 et une fois la valeur 1 dans la colonne correspondant à la modalité prise par le fait k. Z^i est un sous-tableau disjonctif qui décrit la partition des n faits induite par les modalités de la dimension D^i.

En juxtaposant les d matrices Z^i, nous construisons la matrice Z à n lignes et l colonnes. $Z = [Z^1, Z^2, \ldots, Z^i, \ldots, Z^d]$ est un tableau disjonctif complet qui décrit les n faits du cube $\mathcal{C}'$ par un codage binaire. La figure 3.12 (a) représente un exemple simple de cube à 3 dimensions $D^1 : \{Aut1, Aut2\}$, $D^2 : \{T1, T2\}$ et $D^3 : \{S1, S2, S3\}$. Ce cube est transformé en tableau disjonctif complet Z dans la figure 3.12 (b).

Id	D^1	D^2	D^3	M_1
1	$Aut1$	$T2$	$S1$	9
2	$Aut2$	$T2$	$S3$	5
3	$Aut2$	$T1$	$S2$	6
4	$Aut1$	$T1$	$S3$	7

(a)

	Z						
	Z^1		Z^2		Z^3		
Id	$Aut1$	$Aut2$	$T1$	$T2$	$S1$	$S2$	$S3$
1	1	0	0	1	1	0	0
2	0	1	0	1	0	0	1
3	0	1	1	0	0	1	0
4	1	0	1	0	0	0	1

(b)

FIGURE 3.12 – Transformation d'un cube en tableau disjonctif complet

A partir du tableau disjonctif complet Z, nous construisons le tableau $B = Z'Z$, où Z' désigne la transposée de Z. B est une matrice carrée symétrique d'ordre l qui rassemble les croisements deux à deux de toutes les dimensions du cube $\mathcal{C}'$. B est donc une juxtaposition des tableaux de contingence des dimensions croisées deux par deux ou d'une dimension croisée avec elle-même.

B est appelé tableau de *Burt* ou hypercube de contingence.

Par exemple, la figure 3.13 (b) représente le tableau de contingence de *Burt* obtenu à partir du tableau disjonctif complet Z de la figure 3.13 (a).

$$Z = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad B = Z'Z = \begin{pmatrix} 2 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 2 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 2 \end{pmatrix}$$

(a) (b)

FIGURE 3.13 – Transformation d'un tableau disjonctif complet en tableau de *Burt*

Analyse des Correspondances Multiples. Soit X une matrice carrée, symétrique, diagonale et d'ordre l, ayant les mêmes éléments diagonaux que le tableau de contingence de *Burt* B et des zéros ailleurs. Afin de trouver les axes factoriels, l'ACM consiste à diagonaliser la matrice $S = \frac{1}{d}Z'ZX^{-1}$ dont le terme général est $s_{tt'} = \frac{1}{dz_{.t'}} \sum_{k=1}^{n} z_{kt} z_{kt'}$

Après diagonalisation de la matrice S, l'ACM fournit $(l-d)$ valeurs propres notées λ_α. A chaque valeur propre λ_α correspond un vecteur propre u_α, vecteur directeur de l'axe factoriel α.

Soit $\varphi^i_{\alpha t}$ la coordonnée ou la projection de la modalité a^i_t sur l'axe α.

La contribution de la modalité a^i_t dans la construction de l'axe α est évaluée par :

$$Cr_\alpha(a^i_t) = \frac{z^i_{.t} {\varphi^i_{\alpha t}}^2}{nd\lambda_\alpha}$$

où $z^i_{.t} = \sum_{k=1}^{n} z^i_{kt}$ correspond au nombre de faits dans le cube $\mathcal{C}'$ ayant la modalité a^i_t ; $z^i_{.t}$ est donc le poids de la modalité a^i_t dans le cube $\mathcal{C}'$. La contribution $Cr_\alpha(a^i_t)$ représente la part d'inertie due à la modalité a^i_t dans la construction de l'axe factoriel F_α.

La contribution d'une dimension D^i dans la construction du facteur α est la somme des contributions de ses modalités :

$$Cr_\alpha(D^i) = \sum_{t=1}^{l_i} Cr_\alpha(a^i_t) = \frac{1}{nd\lambda_\alpha} \sum_{t=1}^{l_i} z^i_{.t} {\varphi^i_{\alpha t}}^2$$

On repère ainsi les dimensions du cube qui ont le plus contribué à la définition de chaque axe factoriel de l'ACM. La contribution d'une dimension à un axe factoriel est un indicateur de liaison entre la dimension et le facteur.

Compte tenu de notre objectif de réarrangement des modalités des dimensions, nous introduisons la notion de valeur-test d'une modalité a^i_t [LMP04]. Cette notion

va permettre de tester si la position d'une modalité sur l'axe α est due au hasard ou non, c'est-à-dire si la modalité est proche ou non du centre de gravité (origine de l'axe α).

Soit $\psi_{\alpha k}$ la coordonnée ou la projection du fait k sur l'axe α.

Si les $z^i_{.t}$ faits ayant la modalité a^i_t dans le cube $\mathcal{C}'$ ont été choisis au hasard, parmi l'ensemble des n faits du cube, alors la moyenne des coordonnées des $z^i_{.t}$ faits sur l'axe α, notée $Y^i_{\alpha t}$ avec

$$Y^i_{\alpha t} \;=\; \tfrac{1}{z^i_{.t}} \textstyle\sum_k \psi_{\alpha k}$$

est une variable centrée (de moyenne nulle) et de variance

$$\mathrm{VAR}(Y^i_{\alpha t}) \;=\; \tfrac{n-z^i_{.t}}{n-1}\;\tfrac{\lambda_\alpha}{z^i_{.t}}$$

Sous cette hypothèse, appelée H_0 ou hypothèse nulle du test, la variable $Y^i_{\alpha t}$ suit une loi normale.

La valeur-test de la modalité a^i_t sur l'axe α s'écrit :

$$V^i_{\alpha t} = \sqrt{z^i_{.t}\;\tfrac{n-1}{n-z^i_{.t}}}\;\varphi^i_{\alpha t}$$

Une valeur-test $V^i_{\alpha t}$ mesure en nombre d'écarts-types la distance entre la modalité a^i_t et l'origine de l'axe α. Ainsi, la position d'une modalité est significativement intéressante sur un axe α si sa valeur-test est supérieure à 2 en valeur absolue. On considère alors que la position de la modalité n'est pas due au hasard et qu'elle est éloignée du centre de gravité (origine de l'axe) avec un risque d'erreur égal à 5%.

Arrangement des modalités. Nous proposons deux façons de réarranger les modalités, soit selon leurs projections, soit selon leurs valeurs-test.

L'arrangement de modalités selon leur projection consiste à chercher pour chaque dimension initiale D^i le meilleur axe factoriel α c'est-à-dire celui qui est le mieux expliqué par les modalités de cette dimension. Nous cherchons à maximiser la valeur de $\lambda_\alpha Cr_\alpha(D^i)$. Soit α^* l'indice maximisant cette quantité. Nous récupérons ensuite les coordonnées $\varphi^i_{\alpha^* t}$ des modalités a^i_t de la dimension D^i sur l'axe factoriel α^*. Selon un tri croissant de ces coordonnées, nous obtenons un nouvel ordre des a^i_t dans la dimension D^i. Cet arrangement a pour effet de concentrer les cellules pleines au centre du cube et d'éloigner les cellules vides vers les extrémités.

L'arrangement de modalités selon leur valeurs-test consiste, pour chaque dimension, à trier ses modalités selon l'ordre croissant des valeurs-test sur le premier axe

factoriel α_1, puis sur le deuxième axe factoriel α_2,..., et ainsi de suite sur les axes retenus par l'utilisateur.

Qualité de représentation. Afin d'évaluer le gain induit par l'arrangement des modalités des dimensions, nous proposons un indice permettant de mesurer l'homogénéité de la disposition des cellules dans le cube. Nous considérons que plus les cellules pleines (ou bien vides) sont concentrées, plus le cube est homogène.

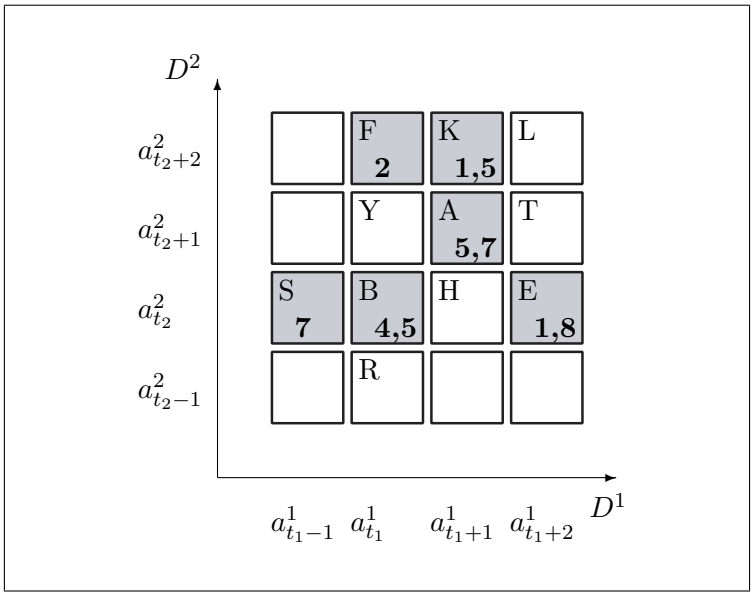


FIGURE 3.14 – Exemple en 2 dimensions de la notion de voisinage des cellules

Les modalités des dimensions constituent les coordonnées des cellules dans le cube. Soit $A = (a^1_{t_1}, \ldots, a^i_{t_i}, \ldots, a^d_{t_d})$ une cellule dans le cube $\mathcal{C}$, avec $i \in \{1, \ldots, d\}$ et $t_i \in \{1, \ldots, l_i\}$. t_i est l'indice de la modalité que prend la cellule A pour la dimension D^i. Toutes les modalités des dimensions D^i sont géométriquement ordonnées dans le cube selon l'ordre des indices t_i. La modalité $a^i_{t_i-1}$ précède $a^i_{t_i}$, qui, à son tour, précède $a^i_{t_i+1}$ (cf. figure 3.14). L'ordre des indices t_i correspond à l'ordre dans lequel sont arrangées les modalités de la dimension D^i.

Soit $A = (a^1_{t_1}, \ldots, a^i_{t_i}, \ldots, a^d_{t_d})$ et $B = (b^1_{t_1}, \ldots, b^i_{t_i}, \ldots, b^d_{t_d})$ deux cellules du cube. Les cellules A et B sont voisines si les coordonnées de B vérifient : $b^i_{t_i} = a^i_{t_i-1}$ ou $b^i_{t_i} = a^i_{t_i}$ ou $b^i_{t_i} = a^i_{t_i+1}$. Dans le cas où $\forall i \in \{1, \ldots, d\}$ $b^t_{t_i} = a^i_{t_i}$, B n'est pas considérée comme une cellule voisine de A car $B = A$.

Dans l'exemple de la figure 3.14, la cellule B est voisine de A. Y est aussi voisine de A. En revanche, les cellules S et R ne sont pas voisines de A.

Nous appelons voisinage de A, noté $\mathcal{V}(A)$, l'ensemble de toutes les cellules qui sont voisines de A. Dans notre exemple, le voisinage de la cellule est $\mathcal{V}(A) = \{F, K, L, Y, T, B, H, E\}$.

La similarité des cellules A et B, notée $\delta(A, B)$, est définie comme :

$$\delta(A,B) \quad = \quad \begin{cases} 1 - (\frac{||A|-|B||}{\max(\mathcal{C})-\min(\mathcal{C})}) & \text{si } A \text{ et } B \text{ sont pleines} \\ 0 & \text{sinon} \end{cases}$$

où $||A|-|B||$ est la valeur absolue de la différence des mesures contenues dans A et B. $\max(\mathcal{C})$ (respectivement, $\min(\mathcal{C})$) est la valeur maximale (respectivement, la valeur minimale) de la mesure dans $\mathcal{C}$, avec $\min(\mathcal{C}) \neq \max(\mathcal{C})$.

Dans le cube de la figure 3.14, où les cellules grises sont pleines et les cellules blanches sont vides, $(\max(\mathcal{C}) = 7)$ et $(\min(\mathcal{C}) = 1,5)$. Par conséquent, la similarité des cellules A et B est $\delta(A, B) = 1 - (\frac{|5,7-4,5|}{7-1,5}) \simeq 0,78$. En revanche, la similarité des cellules A et Y est nulle car la cellule Y est vide.

Notre définition de la similarité de deux cellules n'est pas applicable quand toutes les cellules du cube ont la même valeur de mesure. Dans ce cas là, on pose $\delta(A, B) = 1$.

La similarité de A à son voisinage, notée $\Delta(A)$, est la somme des similarités de la cellule A avec toutes ses cellules voisines :

$$\Delta(A) = \textstyle\sum_{B \in \mathcal{V}(A)} \delta(A, B)$$

Par exemple, la similarité au voisinage de la cellule A vaut :

$\Delta(A) = \delta(A,F)+\delta(A,K)+\delta(A,L)+\delta(A,T)+\delta(A,E)+\delta(A,H)+\delta(A,B)+\delta(A,Y)$
$\Delta(A) = 1 - (\frac{|5,7-2|}{7-1,5}) + 1 - (\frac{|5,7-1,5|}{7-1,5}) + 0 + 0 + 1 - (\frac{|5,7-1,8|}{7-1,5}) + 0 + 1 - (\frac{|5,7-4,5|}{7-1,5}) + 0$
$\Delta(A) \simeq 1,64$

Nous appelons, indice d'homogénéité brut du cube $IHB(\mathcal{C})$, la somme des similarités de tous les couples de ses cellules (à la fois pleines et voisines) :

$$IHB(\mathcal{C}) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq \mathsf{NULL}}} \sum_{B \in \mathcal{V}(A)} \delta(A, B) = \sum_{\substack{A \in \mathcal{C} \\ |A| \neq \mathsf{NULL}}} \Delta(A)$$

Dans l'exemple de la figure 3.14, l'indice d'homogénéité brut du cube vaut $IHB(\mathcal{C}) = \Delta(F) + \Delta(K) + \Delta(A) + \Delta(S) + \Delta(B) + \Delta(E) \simeq 6,67$.

Par construction, la valeur de cet indice croit en fonction de la qualité de représentation du cube. Plus les cellules d'un cube sont homogènes en terme de voisinage et de similarité, plus la valeur de l'indice d'homogénéité brut est grande. La représentation la plus homogène possible correspond au cas où le cube ne contient pas de cellules vides et que toutes les cellules ont des mesures égales. Dans ce cas, les similarités aux voisinages sont toutes égales à 1. Par conséquent, l'indice d'homogénéité brut atteint sa valeur maximale $IHB_{max}(\mathcal{C})$:

$$IHB_{max}(\mathcal{C}) = \sum_{A \in \mathcal{C}} \sum_{B \in \mathcal{V}(A)} 1$$

Nous définissons l'indice d'homogénéité du cube $\mathcal{C}$, noté $IH(\mathcal{C})$, comme le rapport de l'indice d'homogénéité brut par l'indice d'homogénéité maximale :

$$IH(\mathcal{C}) = \frac{IHB(\mathcal{C})}{IHB_{max}(\mathcal{C})} = \frac{\sum\limits_{\substack{A \in \mathcal{C} \\ |A| \neq \mathsf{NULL}}} \Delta(A)}{\sum\limits_{A \in \mathcal{C}} \sum\limits_{B \in \mathcal{V}(A)} 1}$$

Il mesure la qualité de représentation d'un cube de données. Cette qualité est d'autant meilleure que les cellules pleines et ayant des mesures proches sont géométriquement voisines et rassemblées dans certaines régions du cube. Par exemple, sachant que l'indice d'homogénéité brut maximum de cube $\mathcal{C}$ de la figure 3.14 est $IHB_{max}(\mathcal{C}) = 84$, l'indice d'homogénéité est dans ce cas égal à $IH(\mathcal{C}) = \frac{6,67}{84} \simeq 0,08$.

Seule, la valeur de l'indice d'homogénéité ne permet pas d'évaluer la qualité "universelle" de représentation du cube. En revanche, avec cet indice, on peut mesurer l'apport d'une réorganisation des dimensions en évaluant le gain de la qualité induit par cette réorganisation.

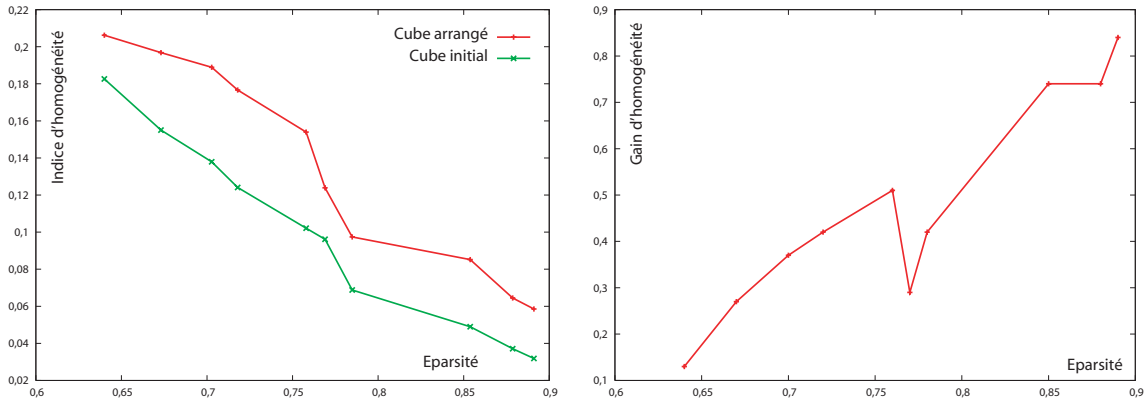
Pour mesurer l'apport de l'arrangement des modalités sur la représentation du cube, nous calculons le gain d'homogénéité g :

$$g = \frac{IH(\mathcal{C}_{arr}) - IH(\mathcal{C}_{ini})}{IH(\mathcal{C}_{ini})}$$

où $IH(\mathcal{C}_{ini})$ est l'indice d'homogénéité de la représentation du cube initial et $IH(\mathcal{C}_{arr})$ est celui de la représentation réorganisée selon notre méthode.

Evaluation Afin de mesurer l'impact de l'éparsité des données, nous avons différents essais avec plusieurs échantillons aléatoires provenant d'un cube initial comportant

199 523 faits. En faisant varier le taux d'échantillonnage, nous faisons varier l'éparsité du cube. Les expériences ont été menées sur les deux types de réorganisation que nous proposons (figures 3.15 et 3.16).



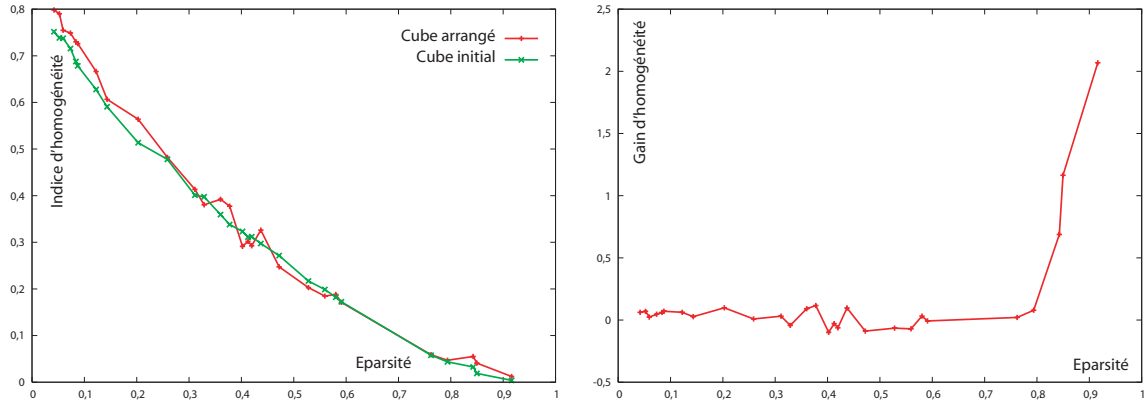
(a) **(b)**

FIGURE 3.15 – Indice d'homogénéité (a) et gain d'homogénéité (b) en fonction de l'éparsité : arrangement selon les projections des modalités

Quel que soit le mode d'arrangement, les valeurs de l'indice d'homogénéité du cube initial et du cube arrangé sont décroissantes en fonction de l'éparsité du cube (figures 3.15 (a) et 3.16 (a)). En effet, la valeur indice avec le nombre de cellules pleines dans le cube. De plus, pour des taux d'éparsité élevés (supérieurs à 60%) l'indice d'homogénéité est plus grand dans le cube arrangé. Nous réalisons toujours un gain d'homogénéité lors de l'arrangement du cube.

D'après les figures 3.15 (b) et 3.16 (b), plus l'éparsité est importante (supérieure à 60%), plus le gain en homogénéité augmente en général. En effet, quand le cube est épars, nous avons une meilleure marge de manoeuvre pour réorganiser les données. Parfois, pour de fortes éparsités, le gain peut fléchir localement. Ceci est lié à la structure des données : si les données du cube initial sont déjà dans une représentation homogène, l'application de notre méthode n'apportera pas de gain considérable. Et enfin nous remarquons dans la figure 3.16 (b) que, pour de faibles valeurs de l'éparsité, le gain d'homogénéité oscille autour de zéro. En effet, quand les données sont denses, le ré-arrangement des modalités a peu d'intérêt et apporte peu ou pas.

En résumé, notre approche réorganise les faits dans l'espace de représentation que constitue le cube et est d'autant plus efficace lorsque ce dernier est volumineux et épars. Les axes factoriels obtenus par l'ACM donne une autre disposition du cube dans laquelle les faits sont beaucoup moins éparpillés et regroupés selon leur contenu.



(a) **(b)**

FIGURE 3.16 – Indice d'homogénéité (a) et gain d'homogénéité (b) en fonction de l'éparsité : arrangement selon les valeurs-test des modalités

Ils permettent de mettre en évidence des régions que l'utilisateur peut explorer plus particulièrement. Cela aide l'utilisateur dans sa démarche exploratoire et facilite la navigation dans le cube avec une navigation plus pertinente.

Notre approche est concrétisée sous forme d'un module d'analyse dans la plate-forme MiningCubes. Ainsi, en plus des opérateurs classiques de structuration du cube, l'utilisateur dispose d'un nouvel opérateur pour réorganiser intelligemment et automatiquement le cube.

En plus de la plate-forme MiningCubes dans laquelle sont implémentés les opérateurs d'analyse en ligne des données complexes, une deuxième application, dédiée uniquement à l'analyse des publications, est développée.

3.6 Plate-forme logicielle PUMA

Afin d'expérimenter concrètement nos approches et nos premières solutions pour l'analyse en ligne des données complexes sur l'exemple des publications, nous avons entrepris le développement d'une deuxième application logicielle. Intitulé PUMA (*PUblication Modeling and Analysis*), le prototype permet de :

- Alimenter et rafraîchir l'entrepôt des publications,
- Extraire les mots-clés pour chaque nouvelle publication,
- Construire un cube,
- Faire une analyse en ligne des publications.

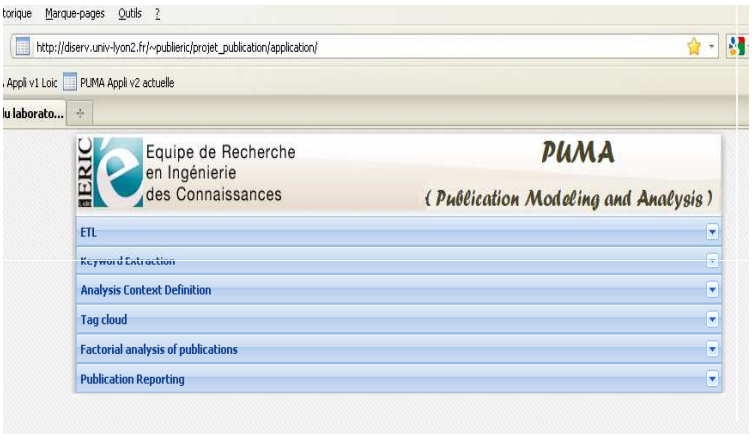


FIGURE 3.17 – Application PUMA (PUblication Modeling and Analysis)

PUMA est une application Web *open source* et est développée en PHP ; l'entrepôt de données est construit avec MySQL. L'application utilise également le logiciel R et son package FactoMiner. Elle est conçue selon une architecture trois tiers avec une interface de paramétrage, un serveur d'application et un serveur de base de données. Depuis l'interface de paramétrage, l'utilisateur alimente l'entrepôt, définit son contexte d'analyse en créant un cube et paramètre les analyses qu'il souhaite réaliser. D'un point de vue technique, l'interface graphique est gérée par le framework ExtJS et avec le support Ajax. D'un point de vue utilisateur, l'application se présente sous forme de modules dans un accordéon à déployer pas à pas (figure 3.17).

Un module d'**ETL** permet de gérer l'alimentation et le rafraîchissement de l'entrepôt. Pour chaque publication, les informations comme le titre, les auteurs, le nom du support, l'année, ... sont dans une base de données du laboratoire. Le module d'ETL extrait ces informations de la base et les charge dans l'entrepôt. De plus, à chaque publication correspond un document qui est sous la forme d'un fichier PDF. Une autre fonctionnalité du module d'ETL est de convertir le fichier PDF en fichier texte, d'épurer le texte (élimination des figures, des références bibliographiques, ...) et d'extraire le résumé et le corps du document. Ces éléments viennent alimenter l'entrepôt.

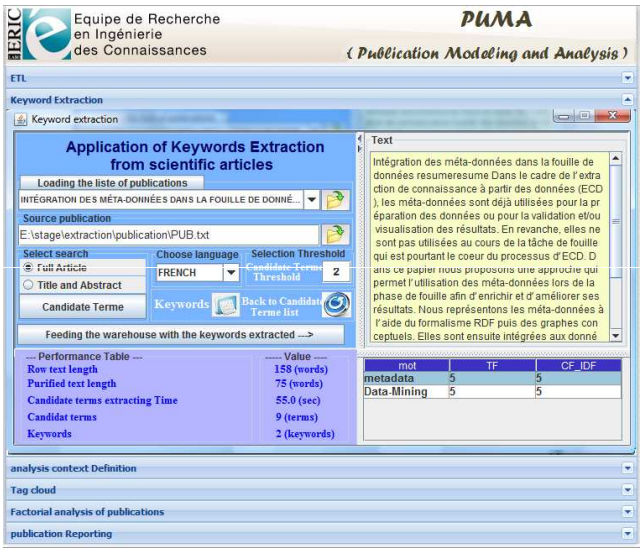


FIGURE 3.18 – Extraction des mot-clés dans PUMA

Le module de **génération de mots-clés** extrait automatiquement les mots-clés pour chaque publication et les insère dans l'entrepôt (figure 3.18). L'utilisateur choisit d'extraire les mots-clés depuis le titre, le résumé ou le corps de la publication. Il choisit également la langue du texte (français ou anglais). Puis le module élimine les mots vides de sens, la ponctuation, ... et décompose le texte en termes. Le module extrait alors les termes candidats et les projette sur le thésaurus pour obtenir les mots-clés. Les mots-clés ainsi extraits sont insérés dans l'entrepôt pour chaque publication.

Le module de définition du **contexte d'analyse** permet de construire le cube ou le sous-cube sur lequel l'utilisateur veut faire des analyses. L'utilisateur choisit les dimensions et niveaux respectifs sur lesquels il veut travailler. Il peut ensuite générer le nuage des mots-clés associés aux publications présentes dans le cube. Dans ce nuage, les mots s'affichent dans une police de caractères d'autant plus grande qu'ils ont été sujets de publications.

Une fois le contexte d'analyse défini (c'est-à-dire le cube ou sous-cube construit), il est possible de lancer une analyse en ligne des publications avec les opérateurs de visualisation, d'agrégation ou de ré-arrangement (figure 3.19).

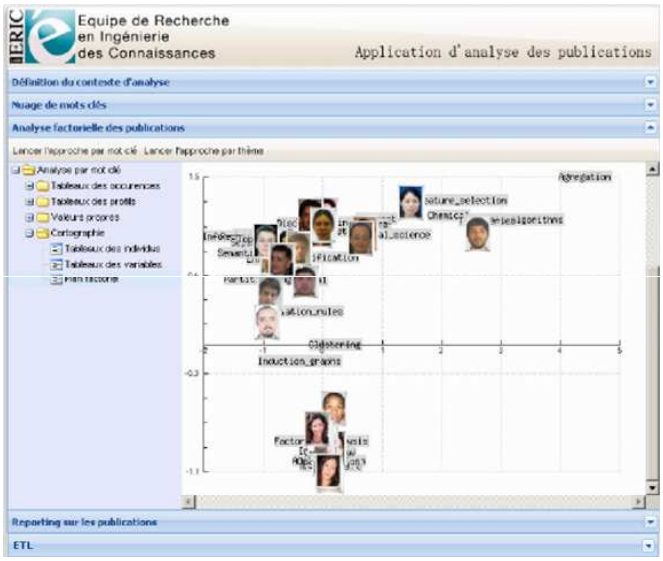


FIGURE 3.19 – Visualisation des publications dans PUMA

Dans la plate-forme PUMA, la possibilité de faire de la recherche d'information dans l'entrepôt pourrait être ajoutée. On pourrait donner à l'utilisateur la possibilité de chercher, grâce à une requête, les publications portant sur certains mots-clés. Le système retournait les publications les plus pertinentes.

3.7 Conclusion

Les données complexes ne peuvent pas être abordées seulement par les techniques classiques d'entreposage et d'analyse en ligne. Des approches spécifiques ou adaptées sont nécessaires. Ne pouvant traiter tous les aspects, nous nous sommes focalisés sur l'analyse en ligne afin d'étudier pourquoi et comment il fallait adapter l'OLAP aux données complexes. En reprenant les fondements de l'OLAP et en y intégrant les spécificités des données complexes, les travaux présentés dans ce chapitre abordent des problèmes d'agrégation et de visualisation des données complexes ; de réorganisation du cube pour une meilleure visualisation des régions intéressantes ; de prise en compte de la sémantique des données complexes dans la modélisation multidimensionnelle et dans l'analyse en ligne.

Pour apporter des premières solutions, nous associons les principes de l'OLAP, de la fouille de données et de la recherche d'information.

Le recours aux méthodes factorielles (AFC et ACM) nous permet de visualiser les faits dans un cube et de trouver des régions intéressantes en réorganisant les dimensions du cube. Les méthodes factorielles détectent dans un cube volumineux et épars, les faits ou les modalités de dimensions intéressants ; mettent en évidence les liens entre les faits ou dimensions ; représentent de façon synthétique ces informations. Ceci permet de prendre en compte la sémantique des données pour leur visualisation et leur réorganisation.

Les principes de la classification hiérarchique ascendante nous aident à faire une agrégation sémantique des faits selon une hiérarchie, agrégation particulièrement adaptée quand il s'agit de données complexes. Les faits similaires ou se ressemblant sont regroupés ensemble pour former des agrégats. Une succession de regroupements permet de construire une hiérarchie de dimension.

Pour l'analyse en ligne des documents textes, le modèle multidimensionnel présenté dans ce mémoire n'est qu'une proposition faite pour l'étude des publications. Il faut poursuivre pour traiter plus complètement l'ensemble des problèmes liés à l'analyse OLAP des documents textes : (1) les faits doivent avoir non seulement des dimensions classiques mais aussi une dimension textuelle. L'utilisateur doit aussi pouvoir naviguer sur la hiérarchie textuelle avec une hiérarchie adaptée. (2) Une ou des mesures textuelles doivent permettre de résumer le contenu d'un document en fonction du niveau de concepts. La mesure textuelle doit être assortie d'une fonction d'agrégation adaptée. (3) Une matérialisation efficace des cubes contenant des documents et des mesures textuelles doit être étudiée.

Dans l'exemple présenté dans ce chapitre, les données complexes comportent des documents textes et nous dissertons beaucoup autour des problèmes posés par les données comportant du texte. Cependant, les documents textes ne sont qu'une instance de données complexes. Nous devons étendre nos réflexions à l'ensemble des données complexes. Les trois approches présentées pour l'analyse en ligne des données complexes sont aussi adaptées et pertinentes quand les données complexes comportent des images, des vidéos, des aspects géographiques, ... Elles ne dépendent pas de la nature des données complexes mais sont applicables dès que l'on dispose de descripteurs des données complexes . Le lecteur pourra se reporter à la thèse de Riadh Ben Messaoud où un exemple de données complexes comportant des images est traité.

Chapitre 4

Conclusion générale

4.1 Bilan des travaux

Lorsque les données sont complexes, le processus d'entreposage et d'analyse en ligne doit être repensé. Les données complexes ne peuvent pas être traitées seulement par les techniques classiques, elles nécessitent des approches adaptées et/ou spécifiques pour l'intégration, la modélisation et l'analyse en ligne. Ces problématiques constituent des problèmes de recherche non encore résolus ou que partiellement et ont donnée naissance à un nouveau champ de recherche. Dans ce contexte, nous nous intéressons plus personnellement aux questions liées à l'analyse en ligne des données complexes.

L'OLAP est une méthode de structuration et d'analyse de l'information dans un processus décisionnel. Son intérêt est de permettre une navigation interactive dans les données, une visualisation rapide de l'information et une exploration de la structure multidimensionnelle des données. Mais les constats faits sur les limites d'une telle analyse navigationnelle et exploratoire et l'avènement des données complexes militent pour une évolution ou adaptation de l'OLAP. Il devient nécessaire de : (1) enrichir les possibilités de l'analyse OLAP en la dotant de nouvelles possibilités ; (2) créer une analyse en ligne adaptée aux données complexes ; (3) faire évoluer l'OLAP vers une analyse sémantique des données.

Pour aborder cette problématique, nous avons identifié les problèmes et choisi de traiter les questions d'agrégation et visualisation des données complexes, de réorganisation du cube pour identifier des régions d'analyse intéressantes, et d'étendre l'OLAP à des possibilités d'explication et de prédiction. Dans toutes ces questions, nous avons également essayé de tenir compte de la sémantique véhiculée par les données. Des premières solutions sont trouvées grâce à une combinaison des principes de l'OLAP, de la fouille de données et de la recherche d'information.

Pour introduire une analyse explicative dans l'OLAP, nous avons choisi de faire une recherche guidée de règles d'association dans le cube. Cela nous a conduit à modifier la définition du support et de la confiance d'une règle. Les arbres de régression nous permettent de proposer à l'utilisateur de faire de la prédiction dans le cube et d'avoir ainsi une démarche de type *What If Analysis*.

Pour l'analyse des données complexes, deux méthodes factorielles (AFC et ACM) rendent possible la visualisation des faits dans un cube et la détection de régions intéressantes en réorganisant les dimensions du cube. Nous proposons également une agrégation sémantique des faits et une nouvelle hiérarchie de dimension construite automatiquement grâce aux principes d'une méthode de classification (CAH). Ces propositions sont illustrées par un exemple réel d'analyse de publications scientifiques.

Pour tester nos propositions, nous avons conçu deux plates-formes logicielles : MiningCubes et PUMA (*PUblication Modeling and Analysis*) qui est dédiée à l'analyse en ligne des publications.

A court terme, nos travaux vont trouver une application immédiate dans le nouveau projet de collaboration entre le laboratoire ERIC et le laboratoire ICAR (Interactions, Corpus, Apprentissages, Représentations, Lyon 2-ENS-CNRS). Ce projet interdisciplinaire doit permettre d'identifier automatiquement des phénomènes complexes (remerciements, ouverture, développement, sortie de conflit, plaisanteries familières, plainte, ...) dans des interactions orales. Le projet utilisera la base de données en ligne CLAPI (Corpus de LAngue Parlée en Interaction) qui rassemble des corpus oraux, chacun contenant un ensemble de transcriptions représentées par des documents XML. Ces corpus sont des données complexes. Il s'agira de définir un modèle multidimensionnel, des mesures et des dimensions appropriées aux transcriptions et de construire un entrepôt de documents. De plus, pour l'identification des phénomènes complexes dans les corpus, des méthodes d'analyse appropriées seront conçues et mises en oeuvre. Dans tous ces aspects, il faudra tenir compte de la sémantique contenue dans les corpus linguistiques.

Nos travaux scientifiques des ces sept dernières années sont une première démonstration de la faisabilité de combiner l'OLAP à d'autres techniques comme la fouille de données et la recherche d'information pour faire significativement évoluer l'analyse en ligne et s'adapter aux données complexes. Elles sont une partie des réponses possibles à cette problématique. Mais il existe beaucoup de méthodes de fouille et d'autres méthodes d'analyse. L'idée n'est pas forcément de combiner toutes ces possibilités à l'OLAP mais de réfléchir à l'évolution que devrait avoir l'analyse en ligne.

Après 20 ans d'existence et avec une activité de recherche soutenue, l'intérêt et la vocation de l'OLAP ne sont plus à démontrer. Mais nos réflexions scientifiques de ces dernières années, nous font penser et dire que, pour l'analyse en ligne des données complexes, l'OLAP doit dépasser ses propres limites et doit évoluer. Avec des données complexes, l'OLAP doit certes s'adapter à leur structure mais doit aussi prendre en compte la sémantique contenue dans ces données. Ce point constitue un véritable verrou scientifique mais qui est que partiellement abordé par la communauté scientifique. Il faudrait également identifier tous les problèmes posés par les données complexes et ce quels que soient leur nature, contexte ou spécificités.

Nous voulons poursuivre nos travaux dans cette voie et faire évoluer l'OLAP vers une nouvelle génération d'analyse en ligne : l'OLAP sémantique. Cette direction constitue un nouveau domaine de recherche à lui seul.

4.2 Projet de recherche : l'OLAP sémantique

L'OLAP est un processus d'analyse composé de plusieurs phases, allant de la construction de l'espace d'analyse, à la navigation et visualisation. On pourrait y ajouter l'extraction de connaissances. Avec les données complexes, ce processus devrait être un processus sémantique mais les problèmes majeurs sont de :

– **Couvrir tous les types et toutes les caractéristiques des données complexes**.
L'OLAP tente de s'adapter aux données complexes depuis plusieurs années. Plusieurs équipes travaillent sur l'association des Systèmes d'Information Géographique, des entrepôts de données et de l'analyse OLAP pour créer le SOLAP (*Spatial* OLAP) [BRP07, MZ08, BTML07, BH09]. Les données spatiales sont une forme de données complexes. Dans un cube de données spatiales, les dimensions et les mesures peuvent contenir des composantes spatiales ou géométrique.
Un autre exemple de données complexes peut être celui des flots de données (*data stream*). Dans de telles données, les analystes souhaitent détecter des changements dynamiques et essayent de faire une analyse en ligne des flots. On parle de fouille de flots de données multidimensionnelles, d'*OLAPing Stream Data* ou de *Stream cube* [CDH$^+$02, HCD$^+$05, PLP09].
Le XOLAP (ou XML OLAP) cherche à faire des analyses OLAP classiques sur des documents XML tout en tenant compte des spécificités des documents XML (hiérarchies multiples, imbriquées, incomplètes, ...)[WLHG05, HMD08].
Ces déclinaisons de l'OLAP sont des exemples d'adaptation de l'OLAP aux différents types de données mais ces adaptations portent souvent que sur la structure des données et non pas sur leur contenu. De plus, il faut aller au delà

pour être capable de couvrir tous les types de données complexes, par exemple peu ou quasiment aucuns travaux ne traitent de l'OLAP sur des images sur le son ou, alors que les images sont une composante importante des données complexes.

- **Modéliser les données complexes, leur sémantique et leurs liens**.
  Comme le dit Pei [Pei03], il est parfois difficile de définir la mesure et la fonction d'agrégation dans le cas de données complexes, alors que dans les données classiques, une mesure est toujours définie et est assortie d'une fonction d'agrégation. Il n'est pas toujours possible ou approprié d'en faire autant lorsque l'on travaille sur des données complexes. Il n'existe pas forcément de mesure, de mesure agrégeable ou de hiérarchie explicite prédéfinie dans les données complexes.
  Le modèle multidimensionnel en étoile ou en flocon de neige ne permet pas de représenter les liens sémantiques existants entre les données (même classiques). Ce modèle est un dérivé du modèle entités/associations dans sa déclinaison la plus réduite. Dans le modèle multidimensionnel, les liens entre les tables ne sont que des liens de calcul alors que dans un modèle entité/associations, il y a des liens autres, des associations qui expriment un lien sémantique entre plusieurs entités.
  Dans le modèle multidimensionnel, les mesure sont généralement numériques ou syntaxique/structurelle (par exemple la fréquences de mots-clés). Ces mesures ne sont pas suffisantes pour des domaines où l'analyse du contenu est nécessaire. Les réponses passeront certainement par la définition d'une nouvelle génération de modèle multidimensionnel (sans forcément tous les principes du schéma classique) et par l'introduction dans ce modèle de mesures sémantiques. Il sera certainement difficile de trouver un modèle générique car le modèle dépendra du contenu des données complexes.

- **Analyser en ligne les données complexes**. La principale contribution de nos travaux porte sur ce verrou mais ce ne sont que des premières propositions. En plus de tous les questions évoquées dans ce mémoire, il y a encore de nombreux points à envisager et résoudre.
  Comme le montre Lakshmanan *et al.* dans un article précurseur, les techniques et les opérateurs classiques de l'OLAP ne permettent pas d'appréhender la sémantique contenue dans un cube [LPZ03]. Ils n'arrivent pas à saisir les relations sémantiques qui existent entre les cellules. Par exemple, si deux cellules se ressemblent au sens de la mesure, aucun opérateur OLAP classique ne permet de le mettre en évidence ou de l'expliquer. Il faudrait aussi pouvoir naviguer "sémantiquement" dans le cube. Dans des cubes de taille importante, l'utilisateur OLAP ne sait pas sur quelle dimension il peut être intéressant de faire une opération de *roll-up* ou de *drill-down*. En fonction de l'information

contenue dans les cellules, il pourrait être pertinent de compresser le cube selon sa sémantique. La plupart des méthodes de compression de cube proposent une compression "syntaxique" du cube sans prendre en compte la sémantique du cube.
Compte tenu de la nature des données, et chacun dans leur domaine, le *text mining*, le *data mining*, l'*image mining*, le *Web mining*, ... permettent d'extraire des connaissances à partir des données. L'idée serait d'intégrer toutes ces techniques dans un seul cadre d'analyse et de proposer une démarche et des opérateurs adaptés aux données complexes.

– **Intégrer les connaissances de l'utilisateur dans le processus d'analyse**.
Une fois le cube construit, si l'utilisateur n'a pas une grande connaissance du domaine (à la fois du métier et de l'OLAP) alors l'analyse du cube ne donnera rien. Dans l'OLAP, la navigation devrait dépendre aussi de l'utilisateur, de son profil, de son métier. Des collègues de notre équipe travaillent sur une approche pour guider l'utilisateur dans son exploration du cube afin de trouver des informations pertinentes [BFB08]. Ils utilisent le principe de la recommandation et de la personnalisation de requêtes. De travaux relativement récents étudient la recommandation et la personnalisation dans les entrepôts de données [BGM$^+$, FBB07] mais cette problématique est peu abordée dans l'analyse en ligne et traite partiellement des données complexes. Par exemple, quand un utilisateur navigue dans le cube, tous les sous-cubes ne sont pas forcément intéressants pour lui, il faudrait pouvoir détecter les régions intéressantes selon son profil.
Dans l'OLAP sémantique, l'objectif serait de mettre l'utilisateur au coeur du processus pour la définition des besoins d'analyse et pour l'exploitation des résultats. Il serait assister dans l'analyse ; son profil et ses connaissances seraient pris en compte pour filtrer les résultats et ne voir que les plus pertinents pour lui.

L'OLAP sémantique doit traiter ces problèmes. Dans ce nouveau cadre d'analyse, on pourrait également essayer d'unifier et de formaliser la définition des données complexes. Certains auteurs parlent d'objets complexes composés de données hétérogènes que l'on doit analyser comme un tout. La notion d'objet complexe pourrait encapsuler celle de données complexes, mais il faudra définir précisément cette notion et montrer en quoi elle constitue un apport pour modéliser la sémantique des données complexes.

Ces verrous scientifiques nous conduisent à créer, au sein de notre équipe, un nouveau thème de recherche et ils seront abordés à travers plusieurs thèses. Les prochaines années devraient permettre de définir et de formaliser l'OLAP sémantique des objets complexes, de recenser les verrous à la fois théoriques, méthodologiques et technologiques, de proposer des approches et de développer des outils adaptés.

Bibliographie

[ALG08] Aouiche K., Lemire D., Godin R., « Web 2.0 OLAP : From Data Cubes to Tag Clouds », in *Proceedings of the* 4^{th} *International Conference onWeb Information Systems and Technologies (WEBIST 08)*, pp. 5–12, Madeira, Portugal. 2008.

[Baz05] Baziz M., *Indexation conceptuelle guidée par ontologie pour la recherche d'information*, Thèse de doctorat, Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, Toulouse, France. 2005.

[BBBL10] Boukraa D., Boussaid O., Bentayeb F., Loudcher S., « OLAP Operators For A Complex Object-Based Multidimensional Model », *International Journal of Data Mining and Business Intelligence (DMBI)*, pp. 34–46. 2010.

[BBDR03] Boussaid O., Bentayeb F., Darmont J., Rabaseda S., « Vers l'entreposage des données complexes : structuration, intégration et analyse », *Ingénierie des Systèmes d'Information (RSTI série ISI)*, 8(5-6) :79–107. 2003.

[BBDR05] Bentayeb F., Boussaid O., Darmont J., Rabaseda S.L., « 1ère journée francophone sur les entrepôts de données et l'analyse en ligne (EDA'05) », Lyon, France. 2005.

[BBLR05] BenMessaoud R., Boussaid O., Loudcher-Rabaseda S., « Evaluation of a MCA-Based Approach to Organize Data Cubes », in *Proceedings of the* 14^{th} *ACM International Conference on Information and Knowledge Management (CIKM'2005)*, pp. 341–342, Bremen, Germany : ACM Press. October – November 2005.

[BBLR06a] BenMessaoud R., Boussaid O., Loudcher-Rabaseda S., « A Data Mining-Based OLAP Aggregation of Complex Data : Application on XML Documents », *International Journal of Data Warehousing and Mining*, 2(4) :1–26. 2006.

[BBLR06b] BenMessaoud R., Boussaid O., Loudcher-Rabaseda S., « Efficient Multi-dimensional Data Representation Based on Multiple Correspondence Analysis », in *Proceedings of the* 12^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2006)*, pp. 662–667, Philadelphia, USA : ACM Press. August 2006.

[BBLR06c] BenMessaoud R., Boussaid O., Loudcher-Rabaseda S., « Mining Association Rules in OLAP Cubes », in *Proceedings of the* 1^{st} *International Conference on Innovations in Information Technology (IIT'2006)*, Dubai, UAE : IEEE Communications Society. November 2006.

[BBLR06d] BenMessaoud R., Boussaid O., Loudcher-Rabaseda S., « Using a Factorial Approach for Efficient Representation of Relevant OLAP Facts », in *Proceedings of the* 7^{th} *International Baltic Conference on Databases and Information Systems*

(DB&IS'2006), pp. 98–105, Vilnius, Lithuania : IEEE Communications Society. July 2006.

[BBLR07] BenMessaoud R., Boussaid O., Loudcher-Rabaseda S., *A multiple correspondence analysis to organize data cubes*, volume 155(1) de *Databases and InformationSystems IV - Frontiers in Artificial Intelligence and Applications*, pp. 133–146, IOS Press. 2007.

[BBR04] BenMessaoud R., Boussaid O., Rabaseda S., « A New OLAP Aggregation Based on the AHC Technique », in *Proceedings of the 7^{th} ACM International Workshop on Data Warehousing and OLAP (DOLAP'2004)*, pp. 65–72, Washington D.C., VA, USA : ACM Press. November 2004.

[BDBLR08] Boussaid O., Darmont J., Bentayeb F., Loudcher-Rabaseda S., « Warehousing complex data from the Web », *International Journal of Web Engineering and Technology*, 4(4) :408–433. 2008.

[Ben82] Benzécri J., *L'analyse des données (tome 2 : l'analyse des correspondances)*, Dunod, Paris, 4ème édition. 1982.

[Ben06] BenMessaoud R., *Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes*, Thèse de doctorat, Université Lumière Lyon 2, Lyon, France. Novembre 2006.

[Ber99] Bertin J., « Graphics and graphic information processing », pp. 62–65. 1999.

[BFB08] Bentayeb F., Favre C., Boussaid O., « A User-driven Data Warehouse Evolution Approach for Concurrent Personalized Analysis Needs », *Integrated Computer-Aided Engineering (ICAE), Special Issue*. 2008.

[BFOS84] Breiman L., Friedman J.H., Olshen R.A., Stone C.J., *Classification and Regression Trees*. 1984.

[BG88] Barthelémy J., Guenoche A., *Les arbres et les représentations des proximités*, Paris : Masson. 1988.

[BGM+] Bellatreche L., Giacometti A., Marcel P., Mouloudi H., Laurent D., « A Personalization Framework for OLAP Queries », in *Proceedings of the 8^{th} ACM International Workshop on Data Warehousing and OLAP (DOLAP'2005)*.

[BH09] Bédard Y., Han J., *Geographic Data Mining and Knowledge Discovery*, chapitre Fundamentals of Spatial Data Warehousing for Geographic Knowledge Discovery, Taylor & Francis. 2009.

[Bla05] Blanchard J., *Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association*, Thèse de doctorat, Ecole Polytechnique de l'Université de Nantes, Nantes, France. Novembre 2005.

[BLP+10] Bringay S., Laurent A., Poncelet P., Roche M., Teisseire M., « Bien cube, les données textuelles peuvent s'agréger », in $10^{ème}$ *Conférence d'Extraction et Gestion des Connaissances (EGC 2010)*, pp. 585–596, Hammamet, Tunisie. janvier 2010.

[BLRBM06] BenMessaoud R., Loudcher-Rabaseda S., Boussaid O., Missaoui R., « Enhanced Mining of Association Rules from Data Cubes », in *Proceedings of the 9^{th} ACM International Workshop on Data Warehousing and OLAP (DOLAP'2006)*, pp. 11–18, Arlington, USA : ACM Press. November 2006.

[BLRMB07] BenMessaoud R., Loudcher-Rabaseda S., Missaoui R., Boussaid O., *OLEMAR : an On-Line Environment for Mining Association Rules in Multidimensional*

Data, volume 2 de *Advances in Data Warehousing and Mining*, Idea Group Publishing. 2007.

[BMM+11] Bentayeb F., Maiz N., Mahboubi H., Favre C., Loudcher S., Harbi N., Boussaid O., Darmont J., *Business Intelligence Applications and the Web : Models, Systems and Technologies, IGI Book. To appear*, chapitre Innovative Approaches for efficiently Warehousing Complex Data from the Web. 2011.

[BMS97] Brin S., Motwani R., Silverstein C., « Beyond Market Baskets : Generalizing Association Rules to Correlations », in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'1997)*, pp. 265–276, Tucson, Arizona, USA : ACM Press. May 1997.

[BNBLRB08] Bodin-Niemczuk A., BenMessaoud R., Loudcher-Rabaseda S., Boussaid O., « Vers l'intégration de la prédiction dans l'OLAP », in $8^{ème}$ *Conférence d'Extraction et Gestion des Connaissances (EGC 08)*, pp. 203–204, Sofia Antipolis, France. Janvier 2008.

[BRBB04] BenMessaoud R., Rabaseda S., Boussaid O., Bentayeb F., « OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données », in $4^{ème}$ *Conférence d'Extraction et de Gestion des Connaissances (EGC'2004)*, pp. 35–46, Clermont-Ferrand, France. Janvier 2004.

[BRP07] Bédard Y., Rivest S., Proulx M.J., *Data Warehouses and OLAP : Concepts, Architectures and Solutions*, chapitre Spatial Online Analytical Processing : Concepts, Architectures and Solutions from a Geomatics Engineering Perspective, pp. 298–319, Idea Group Inc. February 2007.

[BT06] Boussaid O., Trousse B., « 3ème atelier Fouille de données complexes (FDC'06) », Lille, France. 2006.

[BTML07] Bimonte S., Tchounikine A., Miquel M., Laurini R., « Vers l'intégration de l'analyse spatiale et multidimensionnelle », in *Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2007)*, Clermont Ferrand, France. 2007.

[BW00] Barbara D., Wu X., « Using loglinear models to compress datacubes », in *Proceedings of the First International Conference on Web-Age Information Management (WAIM 00)*, pp. 311–322, London, UK : Springer-Verlag. 2000.

[CCLR05] Chen B.C., Chen L., Lin Y., Ramakrishnan R., « Prediction Cubes », in *Proceedings of the* 31^{st} *International Conference on Very Large Data Bases (VLDB 2005)*, pp. 982–993, Trondheim, Norway : ACM Press. August - September 2005.

[CCS93] Codd E., Codd S., Salley C., « Providing OLAP (On-line Analytical Processing) to User-Analysts : An IT Mandate », Technical report, Hyperion Solutions Corporation, San Jose. 1993.

[CDH+02] Chen Y., Dong G., Han J., Pei J., Wah B., Wang J., « Online Analytical Processing Stream Data : Is It Feasible ? », in ACM, editor, *Proceedings of the* 7^th *Workshop on the Research Issues in Data Mining and Knowledge Discovery (DMKD 2002)*, Madison, USA. 2002.

[CDH+06] Chen Y., Dong G., Han J., Pei J., Wah B., Wang J., « Regression Cubes with Lossless Compression and Aggregation », *IEEE Transactions on Knowledge and Data Engineering*, 18(12) :1585–1599. 2006.

[CGLT99] Chavent M., Guinot C., Lechevallier Y., Tenenhaus M., « Méthodes divisives de classification et segmentation non supervisée », *Revue de Statistique Appliquée*, XLVII(4) :87–99. 1999.

[Cha98] Chaudhuri S., « Data Mining and Database Systems : Where is the Intersection ? », *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 21(1) :4–8. 1998.

[Che98] Cheng S., *Statistical Approaches to Predictive Modeling in Large Databases*, Master's thesis, Simon Fraser University, British Columbia, Canada. February 1998.

[CLLM04] Choong Y., Laurent A., Laurent D., Maussion P., « Résumé de cube de données multidimensionnelles à l'aide de règles floues », in des Nouvelles Technologies de l'Information R., editor, *4èmes Journées Francophones d'Extraction et de Gestion des Connaissances (EGC'2004)*, volume 1, pp. 95–106, Clermont-Ferrand, France. January 2004.

[CLM03] Choong Y., Laurent D., Marcel P., « Computing Appropriate Representations for Multidimensional Data », *Data & knowledge Engineering Journal*, 45(2) :181–203. 2003.

[CRST06] Chen B.C., Ramakrishnan R., Shavlik J.W., Tamma P., « Bellwether Analysis : Predicting Global Aggregates from Local Regions », in *Proceedings of the* 32^{nd} *International Conference on Very Large Data Bases (VLDB 2006)*, pp. 655–666, Seoul, Korea : ACM Press. September 2006.

[DBB+03] Darmont J., Boussaid O., Bentayeb F., Rabaseda S., Zellouf Y., *Web Multiform Data Structuring for Warehousing*, volume 22 de *Multimedia Systems and Applications*, pp. 179–194, Kluwer Academic Publishers. 2003.

[DDF+90] Deerwester S., Dumais S., Furnas G., Landauer T., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the Society for Information Science*, 41(6) :391–407. 1990.

[DZL+10] Ding B., Zhao B., Lin C.X., Han J., Zhai C., « TopCells : Keyword-based search of top-k aggregated documents in text cube », in *Proceedings of the* 26^th *International Conference on Data Engineering (ICDE 2010)*, pp. 381–384, Long Beach, USA : IEEE. 2010.

[FBB07] Favre C., Bentayeb F., Boussaid O., *A Survey of Data Warehouse Model Evolution*, Encyclopedia of Database Technologies and Applications, Second Edition (to appear), Idea Group Publishing. 2007.

[FPSS96] Fayyad U., Piatetsky-Shapiro G., Smyth P., « Knowledge Discovery and Data Mining : Towards a Unifying Framework », in *Proceedings of the* 2^{nd} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'1996)*. 1996.

[GC98] Goil S., Choudhary A., « High Performance Multidimensional Analysis and Data Mining », in *Proceedings of the High Performance Networking and Computing Conference (SC'1998)*, Orlando, Florida, USA. November 1998.

[Gin21] Gini C., « Measurement of Inequality of Incomes », *The Economic Journal*, 31(121) :124–126. March 1921.

[GMH09] Guran M., Mehanna A., Hussein B., « On-Line Analytical Mining for Advanced Business Intelligence Solution », Scientific Bulletin, Series C, University Politehnica of Bucharest. 2009.

[GR08] Golfarelli M., Rizzi S., « UML-Based Modeling for What-If Analysis », in *Proceedings of the* 10^{th} *International Conference on Data Warehousing and Knowledge Discovery (DAWAK 2008)*, pp. 1–12, Turin, Italy. 2008.

[GRP06] Golfarelli M., Rizzi S., Proli A., « Designing What-If Analysis : towards a methodology », in *Proceedings of the 9th International Workshop on Data Warehousing and OLAP (DOLAP 2006)*, pp. 51–58, Arlington, USA. 2006.

[GS69] Gabriel K., Sokal R., « A New Statistical Approach to Geographic Variation Analysis », *Systematic Zoology*, 18 :259–278. 1969.

[Han97] Han J., « OLAP Mining : An Integration of OLAP with Data Mining », in *Proceedings of the 7th IFIP Conference on Data Semantics*, Leysin, Switzerland. October 1997.

[HCC98] Han J., Chee S., Chiang J., « Issues for On-Line Analytical Mining of Data Warehouses », in *Proceedings of the 1998 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'1998)*, pp. 2 :1–2 :5, Seattle, Washington, USA. June 1998.

[HCD+05] Han J., Chen Y., Dong G., Pei J., Wah B., Wang J., Cai Y., « Stream Cube :An Architecture for Multidimensional analysis of Data Streams », *Distributed and Parallel Databases*, 18 :173–187. 2005.

[HLL+04] Huang H., Lechevallier Y., Lauro N.C., Saporta G., Verde R., Wang H., « Complex Data Analysis », Beijing, China. 2004.

[HMD08] Hachicha M., Mahboubi H., Darmont J., « Expressing OLAP operators with the TAX XML algebra », in *3rd International Workshop on Database Technologies for Handling XML Information on the Web (DataX-EDBT 08)*. March 2008.

[HPDW01] Han J., Pei J., Dong G., Wang K., « Efficient computation of Iceberg cubes with complex measures », in *SIGMOD '01 : Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pp. 1–12, New York, NY, USA : ACM Press. 2001.

[HWD+02] Han J., Wang J., Dong G., Pei J., Wang K., « CubeExplorer : online exploration of data cubes », in *SIGMOD '02 : Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pp. 626–626, New York, NY, USA : ACM Press. 2002.

[IKA02] Imieliński T., Khachiyan L., Abdulghani A., « Cubegrades : Generalizing Association Rules », *Data Mining and Knowledge Discovery*, 6(3) :219–258, ISSN 1384-5810. 2002.

[Inm96] Inmon W.H., *Building the Data Warehouse*, John Wiley & Sons. 1996.

[KHC97] Kamber M., Han J., Chiang J., « Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes », in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD'1997)*, pp. 207–210, Newport Beach, CA, USA : The AAAI Press. August 1997.

[Kim96] Kimball R., *The Data Warehouse Toolkit*, John Wiley & Sons. 1996.

[KKL05] Keith S., Kaser O., Lemire D., « Analyzing Large Collections of Electronic Text Using OLAP », in *Proceedings of the Atlantic Provinces Council on the Sciences (APICS 2005)*, Wolfville, Canada. October 2005.

[LB11] Loudcher S., Boussaid O., « OLAP Visualization Operator for Complex Data », in 23rd *International Conference on Advanced Information Systems Engineering - Forum (CAISE'11)*, pp. 139–146, London, UK. june 2011.

[LDH+08] Lin C.X., Ding B., Han J., Zhu F., Zhao B., « Text Cube : Computing IR Measures for Multidimensional Text Database Analysis », in *Proceedings of the 8th*

International Conference on Data Mining (ICDM 2008), pp. 905–910, Pisa, Italy : IEEE. 2008.

[LFAH04] Lamirel J., Francois C., AlShehabi S., Hoffmann M., « New Classification Quality Estimators for Analysis of Documentary Information : Application to Patent Analysis and Web Mapping », *Scientometrics*, 60(3) :445–562. 2004.

[LGO02] Lee J., Grossman D., Orlandic R., « MIRE : A Multidiemsional Information Retrieval Engine for Structured Data and Text », in *Proceedings of the International Conference on Information Technology : Coding and Computing (ITCC'02)*, IEEE. 2002.

[LGO03] Lee J., Grossman D., Orlandic R., « An Evaluation of the Incorporation of a Semantic Network into a Multidimensional Retrieval Engine », in *Proceedings of the* 12^{th} *International Conference on Information and Knowledge Management*, pp. 575–575, New York, USA : IEEE. 2003.

[LMP04] Lebart L., Morineau A., Piron M., *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 3ème édition. 2004.

[Loe47] Loevinger J., « A Systemic Approach to the Construction and Evaluation of Tests of Ability », *Psychological Monographs*, 61(4). 1947.

[LPZ03] Lakshmanan L.V., Pei J., Zhao Y., « Efficacious Data Cube Exploration by Semantic Summarization and Compression », in Press A., editor, *Proceedings of the* 29^{th} *International Conference of Very Large Data Bases (VLDB'2003)*, pp. 1125–1128, Berlin, Germany. 2003.

[LVL05] Lallich S., Vaillant B., Lenca P., « Parametrised Measures for the Evaluation of Association Rules Interestingness », in *Proceedings of the* 6^{th} *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'2005)*, pp. 220–229, Brest, France. May 2005.

[LVL06] Lenca P., Vaillant B., Lallich S., « On the Robustness of Association Rules », in *Proceedings of the 2006 IEEE International Conference on Cybernetics and Intelligent Systems (CIS'2006)*, pp. 596–601, Bangkok, Thailand : IEEE Computer Society. June 2006.

[MCDA03] Mothe J., Chrisment C., Dousset B., Alaux J., « DocCube : multi-dimensional visualisation and exploration of large document sets », *Journal of the American Society for Information Science and Technology*, 54(7) :650–659. 2003.

[MGCB07] Missaoui R., Goutte C., Choupo A.K., Boujenoui A., « A Probabilistic Model for Data Cube Compression and Query Approximation », in *Proceedings of the* 10^{th} *ACM International Workshop on Data Warehousing and OLAP (DOLAP'2007)*, pp. 33–40, Lisbon, Portugal : ACM Press. November 2007.

[MLB10] Mabit L., Loudcher S., Boussaid O., « Analyse en ligne d'objets complexes avec l'analyse factorielle », in $10^{ème}$ *Conférence d'Extraction et Gestion des Connaissances (EGC 2010)*, pp. 381–386, Hammamet, Tunisie. janvier 2010.

[MLC+00] McCabe C., Lee J., Chowdhory A., Grossman D., Frieder O., « On the Design and Evaluation of a Multi-dimensional Approach to Information Retrieval », in *Proceedings of the* 23^{rd} *International ACM Conference on Research and Development in Information Retrieval (SIGIR00)*, pp. 363–365, ACM Press. 2000.

[Mor07] Morin A., « Recherche d'information et analyse factorielle des correspondances : étude des publications récentes sur la malaria », in *Colloque International de*

Statistique Appliquée pour le Développement en Afrique (SADA 07)), Cotonou, Bénin. 2007.

[MRL+09] Mahboubi H., Ralaivao J.C., Loudcher S., Boussaid O., Bentayeb F., Darmont J., *Data Warehousing Design and Advanced Engineering Applications : Methods for Complex Construction*, chapitre X-WACoDa : An XML-based approach for Warehousing and Analyzing Complex Data, pp. 38–54, Advances in Data Warehousing and Mining, Hershey, PA, USA : IGI Publishing. August 2009.

[MS63] Morgan J., Sonquist J., « Problems in the analysis of survey data, and a proposal », *Journal of the American Statistical Association*, 58(302) :415–434. 1963.

[MVSV03] Maniatis A.S., Vassiliadis P., Skiadopoulos S., Vassiliou Y., « Advanced visualization for OLAP », in *Proceedings of the* 6^{th} *ACM International Workshop on Data Warehousing and OLAP (DOLAP'2003)*, pp. 9–16, ACM. 2003.

[MZ08] Malinowski E., Zimanyi E., *Advanced Data Warehouse Design : From Conventional to Spatial and Temporal Applications*, Springer. 2008.

[OC09] Ordonez C., Chen Z., « Exploration and Visualization of OLAP Cubes with Statistical Tests », in *Proceedings of the* 15^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Workshop on Visual Analytics and Knowledge Discovery*, pp. 46–55, Paris, France. 2009.

[PBAP05] Pérez J.M., Berlanga R., Aramburu M.J., Pedersen T.B., « A Relevance-Extended Multi-Dimensional Model for a Data Warehouse Contextualized with Documents », in *Proceedings of the* 8^{th} *ACM International Workshop on Data warehousing and OLAP (DOLAP'2005)*, pp. 19–28, ACM. 2005.

[PBAP07] Pérez J.M., Berlanga R., Aramburu M.J., Pedersen T.B., « R-Cubes : OLAP Cubes Contextualized with Documents », in *Proceedings of the* 23^{rd} *International Conference on Data Engineering (ICDE 2007)*, pp. 1477–1478, IEEE. 2007.

[PCL+05] Plantevit M., Choong Y., Laurent A., Laurent D., Teisseire M., « M^2SP : Mining Sequential Patterns Among Several Dimensions », in *Proceedings of the* 9^{th} *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'2005)*, Lecture Notes in Computer Science, pp. 205–216, Porto, Portugal : Springer-Verlag. October 2005.

[Pei03] Pei J., « A General Model for Online Analytical Processing of Complex Data », in *Proceedings of the* 22^{nd} *International Conference on Conceptual Modeling (ER'03)*, Chicago, USA. October 2003.

[Phi88] Philippakis A., « Structured what if analysis in DSS models », in *Proceedings of the* 21^{st} *Hawaii International Conference on System Sciences (HICSS 1988)*, volume 3, pp. 366–370. 1988.

[PHS05] Park B.K., Han H., Song I.Y., « XML-OLAP : A multidimensional Analysis Framework for XML Warehouses », in *Proceedings of the* 7^{th} *International Conference on Data Warehousing and Knowledge Discovery (DaWaK'2005)*. August 2005.

[PK01] Palpanas T., Koudas N., « Entropy Based Approximate Querying and Exploration of Datacubes », in *Proceedings of the* 13^{th} *International Conference on Scientific and Statistical Database Management (SSDBM'01)*, pp. 81–90, Fairfax, Virginia, USA : IEEE Computer Society. July 2001.

[PKM05] Palpanas T., Koudas N., Mendelzon A., « Using Datacube Aggregates for Approximate Querying and Deviation Detection », *IEEE Transactions on Knowledge and Data Engineering*, 17(11) :1465–1477. November 2005.

[PLP09] Pitarch Y., Laurent A., Poncelet P., « Multidimensional Data Streams Summarization Using Extended Tilted-Time Windows », in *Proceedings of the* 5^{th} *International Symposium on Frontiers of Information Systems and Network Applications (FINA 2009)*, Bradford, UK. 2009.

[PP03] Priebe T., Pernul G., « Towards Integrative Enterprise Knowledge Portals », in *Proceedings of the* 12^{th} *International Conference on Information and Knowledge Management (CIKM 3003)*, pp. 216–223, ACM. 2003.

[RALT06] Rizzi S., Abelló A., Lechtenbörger J., Trujillo J., « Research in data warehouse modeling and design : dead or alive ? », in *Proceedings of the* 9^{th} *ACM International Workshop on Data Warehousing and OLAP (DOLAP'2006)*, pp. 3–10. 2006.

[RC07] Ramakrishnan R., Chen B., « Exploratory Mining in Cube Space », *Data Mining and Knowledge Discovery,*, 15(1) :29–54. 2007.

[RTT07] Ravat F., Teste O., Tournier R., « OLAP Aggregation Function for Textual Data Warehouse », in *International Conference on Enterprise Information Systems (ICEIS 2007)*, pp. 151–156, DISI INSTICC Press. 2007.

[RTTZ08a] Ravat F., Teste O., Tournier R., Zurfluh G., « Designing and Implementing OLAP Systems from XML Documents », *Annals of Information Systems, New Trends in Data Warehousing and Data Analysis*, 3 :295–315. novembre 2008.

[RTTZ08b] Ravat F., Teste O., Tournier R., Zurfluh G., « A Top keyword extraction method for OLAP document », in *International Conference on Data Warehousing and Knowledge Discovery (DAWAK 2008)*, volume 5182, pp. 55–64, Springer Verlag, LNCS. 2008.

[SAM98] Sarawagi S., Agrawal R., Megiddo N., « Discovery-driven Exploration of OLAP Data Cubes », in *Proceedings of the* 6^{th} *International Conference on Extending Database Technology (EDBT'1998)*, pp. 168–182, Valencia, Spain : Springer. Mars 1998.

[SBV09] Sureau F., Bouali F., Venturini G., « Optimisation heuristique et génétique de visualisations 2D et 3D dans OLAP : premiers résultats », in *5èmes Journées francophones sur les Entrepôts de Données et l'Analyse en ligne (EDA 09), Montpellier*, volume B-5 de *Revue des Nouvelles Technologies de l'Information*, pp. 62–75, Toulouse : Cépaduès Editions. Juin 2009.

[SEE+11] Sair A., Erraha B., Elkyal M., Loudcher S., Boussaid O., « Prédiction dans les cubes de données OLAP », in *Conférence Méditerranéenne sur l'Ingénierie Sure des Systèmes Complexes (MISC'11*, Agadir, Maroc. mai 2011.

[SHZ+09] Sun Y., Han J., Zhao P., Yin Z., Cheng H., Wu T., « RankClus : integrating clustering with ranking for heterogeneous information network analysis », in *Proceedings of the* 12^{th} *International Conference on Extending Database Technology (EDBT 2009)*, pp. 565–576, ACM. 2009.

[Sul01] Sullivan D., *Document Warehousing and Text Mining*, John Wiley & Sons. 2001.

[SWY75] Salton G., Wong A., Yang C.S., « A vector space model for automatic indexing », *Communications of the ACM*, 18(11) :613–620. 1975.

[SYH09] Sun Y., Yu Y., Han J., « Ranking-based clustering of heterogeneous information networks with star network schema », in *Proceedings of the* 15^{th} *ACM International Conference on Knowledge Discovery and Data mining (KDD 2009)*, pp. 797–806, ACM. 2009.

[Tou07] Tournier R., *Analyse en ligne (OLAP) de documents*, Thèse de doctorat, Université Paul Sabatier, Toulouse III, Toulouse, France. Décembre 2007.

[TT05] Tjioe H., Taniar D., « Mining Association Rules in Data Warehouses », *International Journal of Data Warehousing and Mining*, 1(3) :28–62. 2005.

[VM06] Vinnik S., Mansmann F., « From analysis to interactive exploration : Building visual hierarchies from OLAP cubes », in *Proceedings of the* 14^{th} *International Conference on Extending Database Technology (EDBT'2006)*, pp. 496–514, Munich, Germany : Springer. 2006.

[War63] Ward J., « Hierarchical Grouping to Optimize an Objective Function », *Journal of the American Statistical Association*, 58 :236–244. 1963.

[WLHG05] Wang H., Li J., He Z., Gao H., « OLAP for XML data », in *Proceedings of the* 1^{st} *International Conference on Computer and Information Technology (CIT2005)*, pp. 233–237, Shangai, China : IEEE Computer Society. 2005.

[YCLM10] Yaghlane B.B., Cleuziou G., Lebbah M., Martin A., « 7ème atelier Fouille de données complexes (FDC'10) », Hammamet, Tunisie. 2010.

[YLS+09] Yu Y., Lin C.X., Sun Y., Chen C., Han J., Liao B., Wu T., Zhai C., Zhang D., Zhao B., « iNextCube : information network-enhanced text cube », *Proceedings of the VLDB Endowment*, 2(2) :1622–1625. 2009.

[Zhu98] Zhu H., *On-Line Analytical Mining of Association Rules*, Master's thesis, Simon Fraser University, Burnaby, British Columbia, Canada. December 1998.

[ZLM02] Zighed D., Lallich S., Muhlenbach F., « A statistical Approach for Separability of Classes », in *Statistical Learning, Theory and Applications*, Paris, France. 2002.

[ZR02] Zighed D.A., Rakotomalala R., *Data Mining*, volume H3 744 de *Techniques de l'ingénieur*, pp. 1–26, Editions Techniques de l'Ingénieur. 2002.

[ZRT08] Zighed D.A., Ras Z.W., Tsumoto C.S., « The 4th International Workshop on Mining Complex Data (MCD'08) », Pisa, Italy. 2008.

[ZSD02] Zaiane O., Simoff S., Djeraba C., « International Workshop on Mining Multimedia and Complex Data », . 2002.

[ZZH09] Zhang D., Zhai C., Han J., « Topic Cube : Topic Modeling for OLAP on Multidimensional Text Databases », in *Proceedings of the SIAM International Conference on Data Mining (SDM 2009)*, pp. 1123–1134, Sparks, USA : SIAM. 2009.

Liste des figures

Table des matières

Résumé

L'analyse en ligne OLAP permet une navigation interactive dans les données, une visualisation rapide de l'information et une exploration de la structure multidimensionnelle des données. Une des limites est de se restreindre à des aspects exploratoires et navigationnels. De plus, avec l'avènement des données complexes (données multi-format et/ou multi-structure et/ou multi-source et/ou multi-modale et/ou multi-version), l'analyse en ligne doit s'adapter à la nature spécifique de ces données tout en gardant l'esprit de l'OLAP. Les opérateurs OLAP sont définis pour des données classiques et sont souvent inadaptés quand il s'agit de données complexes par exemple composées de textes, images, son ou vidéos. Les limites de l'OLAP ainsi que la spécificité des données complexes nécessitent une évolution ou adaptation de l'OLAP. Il devient nécessaire de : (1) enrichir les possibilités de l'analyse OLAP en la dotant de nouvelles possibilités ; (2) créer une analyse en ligne adaptée aux données complexes ; (3) faire évoluer l'OLAP vers une analyse sémantique des données.

Dans cette vaste problématique, nous choisissons de traiter les questions d'agrégation et visualisation des données complexes, de réorganisation du cube pour identifier des régions d'analyse intéressantes, et d'étendre l'OLAP à des possibilités d'explication et de prédiction. Pour toutes ces questions, nous essayons également de tenir compte de la sémantique véhiculée par les données. Pour apporter des premières solutions, nous orientons vers une combinaison des principes de l'OLAP, de la fouille de données et de la recherche d'information.

Afin d'introduire une analyse explicative dans l'OLAP, nous faisons une recherche guidée de règles d'association dans le cube. Cela nous conduit à modifier la définition du support et de la confiance d'une règle. Les arbres de régression nous permettent de proposer à l'utilisateur de faire de la prédiction dans le cube et d'avoir ainsi une démarche de type *What If Analysis*.

Pour l'analyse des données complexes, deux méthodes factorielles (AFC et ACM) rendent possible la visualisation des faits dans un cube et la détection de régions intéressantes en réorganisant les dimensions du cube. Nous proposons également une agrégation sémantique des faits et une nouvelle hiérarchie de dimension construite automatiquement grâce aux principes d'une méthode de classification (CAH).

Nos propositions sont une première démonstration de la faisabilité de combiner l'OLAP à d'autres techniques comme la fouille de données et la recherche d'information pour faire significativement évoluer l'analyse en ligne et s'adapter aux données complexes. L'OLAP a commencé à s'adapter à leur structure et à leur spécificité (XOLAP - XML OLAP, SOLAP - *spatial* OLAP). Mais il faut aller au delà et nous pensons qu'un des défis est d'extraire et d'analyser (en ligne) la sémantique contenue dans les données complexes. Ce point constitue un véritable verrou scientifique mais qui est que partiellement abordé par la communauté scientifique. Il faudrait également identifier tous les problèmes posés par les données complexes et ce quels que soient leur nature, contexte ou spécificités.

Nous voulons poursuivre nos travaux dans cette voie et faire évoluer l'OLAP vers une nouvelle génération d'analyse en ligne : l'OLAP sémantique. Les problèmes majeurs à traiter seront comment : (1) modéliser toutes les formes de données complexes, leur sémantique et leurs liens ; (2) analyser en ligne les données complexes ; (3) Intégrer les connaissances de l'utilisateur dans le processus de l'analyse ?

Mots-clés : entrepôt de données, cubes de données, analyse en ligne, données complexes, fouille de données, recherche d'information, OLAP sémantique.