Chapitre 1

Introduction générale

" Savoir ce que tout le monde sait, c'est ne rien savoir. Le savoir commence là où commence ce que le monde ignore. "

Remy de Gourmont, "*Promenades philosophiques*"

1.1 Contexte et problématique

Dans le contexte économique concurrentiel de nos jours, l'information joue un rôle crucial dans le quotidien des entreprises. L'acquisition, l'analyse et l'exploitation des informations sont devenues des choix stratégiques incontournables. La maîtrise de l'information est une compétence capitale pour toute entreprise voulant s'imposer dans les premiers rangs de son domaine d'activité. À la lumière de ces impératifs, les grands volumes de données de production, relatifs à l'activité de l'entreprise, sont devenus de véritables mines de connaissances. À partir de ce moment, de gros efforts sont à déployer pour maîtriser les grandes masses de données d'une part et pour extraire des connaissances potentielles à partir de ces données d'autre part.

Les *entrepôts de données* (*data warehouses*) ont apporté une solution adéquate et efficace au problème du stockage et de la gestion des données. Un entrepôt est une base centralisée de grands volumes de données, historisées, organisées par sujet et consolidées à partir de diverses sources d'informations [Inm96, Kim96]. En plus de sa vocation de stockage, la modélisation d'un entrepôt est complètement dédiée à l'analyse de ses données. En effet, les données d'un entrepôt sont sélectionnées pour construire des magasins de données (*data marts*) dédiées à une activité particulière. Les données sont alors organisées de façon multidimensionnelle selon des modèles *en étoile* ou en *flocons de neige* [CD97]. Ces modèles sont largement employés pour préparer les données à l'analyse. Ils permettent également de produire des vues de données communément appelées *cubes de données*. Un cube de données est constitué d'un ensemble de *cellules* où chaque cellule représente un *fait*. Ce dernier est décrit

par des descripteurs catégoriels selon plusieurs axes d'analyse, appelés *dimensions*, et est observé par un ou plusieurs indicateurs, appelés *mesures*. Alors qu'une mesure est souvent une valeur additive, une dimension repose sur un ensemble fini de *modalités* qui représentent des descripteurs catégoriels. Considérons, par exemple, une application de gestion de clientèles d'une entreprise commerciale. Le cube de la figure 1.1 montre un contexte d'analyse prévu à cet effet. Dans ce dernier, les axes d'analyse sont associés aux dimensions *Profession*, *Produit* et *Statut*. Chaque fait est représenté par une cellule décrite par un ensemble de modalités provenant des différentes dimensions. La cellule contient le *montant moyen des salaires* et l'*effectif des personnes*. Ces derniers indicateurs représentent les mesures du cube.

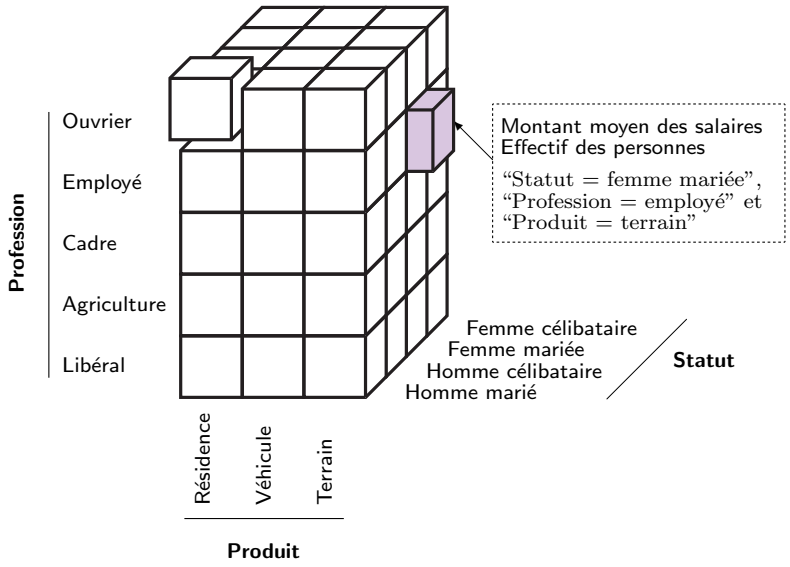

FIG. 1.1 – Exemple introductif d'un cube de données

D'une manière générale, une dimension comporte également plusieurs hiérarchies impliquant différents niveaux de précision dans la description des faits. Par exemple, il est possible d'organiser une dimension *temporelle* selon une hiérarchie à trois niveaux : *Jour* → *Mois* → *Année*. Il est aussi possible d'organiser cette dimension selon une autre hiérarchie à quatre niveaux : *Mois* → *Trimestre* → *Semestre* → *Année*. De telles hiérarchies permettent d'observer des indicateurs selon plusieurs niveaux de granularité et de construire des *agrégats* à partir des faits du niveau le plus fin. Ainsi, un cube de données représente une structure multidimensionnelle comprenant une organisation hiérarchique des données. Cette structure est simple à manœuvrer et est capable de supporter des opérations d'analyse en ligne OLAP (*On-Line Analytical Processing*). La technologie OLAP repose sur des outils pour la *visualisation*, la *structuration* et l'*exploration* des cubes de données. Classiquement, en vue de répondre

à des fins décisionnelles, l'OLAP fournit des moyens aux utilisateurs pour naviguer dans les données multidimensionnelles afin d'y découvrir des informations pertinentes.

Cependant, cette démarche exploratoire suppose une expertise suffisante de l'utilisateur lui permettant d'extraire les connaissances les plus intéressantes au regard de son domaine d'analyse. D'une manière générale, dans ce processus d'aide à la décision, c'est à l'utilisateur de trouver manuellement, en utilisant les outils de l'analyse en ligne, les connaissances potentielles contenues dans les données des cubes. En effet, la technologie OLAP offre des possibilités pour *visualiser* des faits, mais ne permet pas de *décrire* l'ordre d'importance ou les relations possibles entre ces faits. Elle permet aussi de *structurer* des faits selon des axes d'analyse, mais ne permet pas de les *classifier* ou de les regrouper selon leur ordre de proximité. Elle permet également d'*explorer* des faits, mais ne dispose pas de moyens pour *expliquer* les associations ou les implications entre ces faits.

D'un autre côté, la *fouille de données* (*data mining*) est une discipline qui a largement fait ses preuves depuis le début des années 90. Aujourd'hui, on peut considérer la fouille comme une nécessité imposée par le besoin des entreprises de valoriser les données qu'elles collectent dans leurs bases de données. La fouille de données emploie des méthodes d'apprentissage afin d'induire des modèles de connaissances exprimés dans des formalismes valides et compréhensibles. Les techniques de fouille de données ont été employées avec beaucoup de succès dans divers secteurs : la gestion de la relation client (*customer relationship management*), la gestion des connaissances (*knowledge management*) ou l'indexation des documents. Aucun domaine d'application n'est *a priori* exclu.

D'une manière générale, la fouille de données fait appel à deux champs disciplinaires. Alors que le premier est issu de la statistique et de l'analyse des données, le second champ trouve ses origines dans la reconnaissance de formes et dans l'apprentissage automatique. Dans la littérature, on distingue souvent trois grandes familles de techniques de fouille de données : (i) les techniques de *visualisation* et de *description* ; (ii) les techniques de *structuration* et de *classification* et (iii) les techniques d'*explication* et de *prédiction*.

Toutefois, la fouille de données est une des étapes de la chaîne de traitement du processus d'*extraction des connaissances à partir des données* (ECD) (*Knowledge Discovery in Databases, KDD*). Habituellement, dans ce processus, la construction d'un modèle d'apprentissage demande une étape de pré-traitement des données étudiées. Le pré-traitement concerne la préparation des données, leur nettoyage, le traitement des données manquantes et la sélection d'attributs ou d'instances. Cette étape préparatoire résume une démarche de *structuration* robuste mais lourde dans son déploiement. Cependant, elle est toujours capitale dans un processus ECD vu que les techniques de fouille sont souvent sensibles aux bruits. L'information nécessaire à la construction d'un bon modèle de connaissances peut être disponible dans les données mais un bruit préalable, un choix inapproprié de variables ou un mauvais échantillon

d'apprentissage peut faire échouer l'opération. De plus, la fouille de données, dans sa définition restreinte, ne peut opérer que sur des représentations de données bi-dimensionnelles sous forme de tableaux "*individus-variables*".

D'une manière parallèle à l'évolution des entrepôts de données et des techniques de fouille, nous assistons ces dernières années à une prolifération de nouveaux formats de données. Cette situation vient en conséquence de l'avènement d'Internet, l'évolution des technologies de communication, l'étendue des sources d'informations et le développement des technologies multimédia. Aujourd'hui, les entreprises manipulent au quotidien des données non structurées, supportées par des formats variés et provenant de sources hétérogènes. Cette nouvelle génération de données, dites complexes, a suscité de nouvelles directions de recherche afin de répondre aux nouvelles exigences liées à leur stockage et à leur gestion [DBB+03]. Entre autres, la modélisation multidimensionnelle et l'entreposage des données complexes selon le format XML (*eXtensible Markup Language*) sont devenus aujourd'hui des sujets en plein essor dans la communauté des bases de données.

Cependant, du fait de leur nature complexe, l'intégration de ces données dans un processus décisionnel demeure toujours un problème difficile. En effet, contrairement aux données numériques, les données complexes ne se prêtent pas aisément à l'analyse en ligne. À titre d'exemple, avec les outils OLAP classiques, il est impossible de résumer des données textuelles ou des données images selon une *fonction d'agrégation* telles que la somme ou la moyenne. Toutefois, les données complexes constituent une source potentielle riche en connaissances. Par analogie à la fouille des données numériques, la *fouille des données complexes* est aussi devenue un champ de recherche à la fois actif et productif. Aujourd'hui, dans la littérature de la fouille de données, on parle de fouille des données textuelles (*text mining*) [FS06], de fouille d'images (*image mining*), de fouille de données multimédias (*multimedia mining*) [Dje02] et de fouille du Web (*Web mining*) [Cha02].

Dans cette thèse, nous proposons un processus d'aide à la décision qui associe la technologie OLAP avec les techniques d'extraction des connaissances. La combinaison de l'analyse en ligne et de la fouille de données s'avère une solution possible pour rehausser et enrichir le processus d'aide à la décision. L'analyse en ligne et la fouille de données sont deux domaines qui peuvent se compléter dans le cadre d'un processus d'analyse unifié. Ce couplage permet de tirer profit des points forts de chaque domaine et de combler les points faibles de chacun. En effet, alors que la fouille de données doit déployer une lourde étape préparatoire des données, l'analyse en ligne, essentiellement supportée par la technologie des bases de données, bénéficie d'emblée d'une structuration appropriée des données. Par ailleurs, ces données sont nettoyées et facilement manipulables par les opérations OLAP. En revanche, les outils de l'analyse en ligne se limitent à des tâches de visualisation et d'exploration des données. Alors que ces derniers demandent une expertise et un effort manuel de la part de l'utilisateur, afin de pouvoir extraire des informations intéressantes, la fouille

offre des moyens automatiques pour la découverte et l'évaluation de connaissances à partir des données. De plus, le couplage de l'OLAP et de la fouille de données est capable d'apporter des réponses satisfaisantes au problème de l'analyse en ligne des données complexes. Il permet d'étendre les capacités des opérateurs OLAP classiques aux données complexes en profitant de la validité de la fouille sur ces dernières.

Malgré cet aspect complémentaire des deux domaines, historiquement, les travaux de recherche concernant l'analyse en ligne et ceux de la fouille de données ont été développés d'une manière indépendante. Une séparation a longtemps marqué les deux communautés de recherche. Ce n'est qu'à partir de la fin des années 90 que l'on a connu les premières publications montrant la nécessité d'unir les efforts des deux communautés [IM96, Par97, Pal00]. Par exemple, Parsaye souligne l'existence d'une complémentarité entre l'analyse en ligne et la fouille de données et démontre que les deux domaines peuvent coopérer dans un même processus décisionnel [Par97]. Depuis, plusieurs travaux de recherche ont abordé le problème du couplage des deux champs de recherche. D'une manière générale, nous distinguons trois types d'approches abordant ce problème : (i) la première consiste à adapter les données multidimensionnelles afin de les rendre exploitables par les algorithmes classiques de fouille ; (ii) la deuxième consiste à étendre les outils OLAP et les langages de requêtes des bases de données multidimensionnelles aux techniques de fouille et (iii) la dernière consiste à adapter les techniques classiques de fouille aux structures multidimensionnelles des données.

1.2 Objectifs et contributions

Dans le cadre de cette thèse, nous proposons de combiner l'analyse en ligne et la fouille de données afin de les intégrer dans un même processus d'aide à la décision. Le but de ce couplage est d'enrichir les capacités de l'analyse OLAP et de proposer aussi une solution au problème de l'analyse des données complexes. Nous mettons en place trois propositions couplant l'analyse en ligne et la fouille de données. Chacune de nos propositions correspond à une famille de techniques de fouille de données et à une manière d'opérer le couplage entre les deux domaines.

1. Notre première proposition s'inscrit dans le premier groupe des approches de couplage qui consiste à adapter les données multidimensionnelles pour les techniques de fouille. Nous utilisons une technique de *visualisation* et de *description*, à savoir l'*analyse des correspondances multiples* (ACM), en vue d'améliorer automatiquement l'organisation et la qualité de la représentation des données d'un cube. L'idée de base de notre contribution consiste à réorganiser les modalités dans les dimensions d'un cube de données selon des ordonnancements fournis par l'ACM. Grâce à cette réorganisation, il devient possible de fournir un point de vue intéressant homogénéisant au mieux le nuage des faits du cube. Nous apportons ainsi une solution au problème de la visualisation des données

engendré par la volumétrie et l'éparsité des données. Afin d'évaluer les résultats de notre contribution, nous proposons un indice pour la mesure de la qualité de la représentation des données multidimensionnelles. Ce dernier repose sur la notion de l'homogénéité de la répartition géométrique des faits dans l'espace de représentation d'un cube de données.

2. Notre deuxième contribution se base sur la *structuration* et la *classification* dans les données multidimensionnelles. Elle a pour objectif de fournir une nouvelle agrégation des faits d'un cube de données en utilisant la *classification ascendante hiérarchique* (CAH). Nous opérons l'association entre l'analyse en ligne et la CAH selon le deuxième type de couplage instrumental qui consiste à utiliser les opérateurs OLAP comme outils pour l'extraction des données nécessaires à la construction de l'algorithme de fouille. Notre proposition est capable de fournir des agrégats de données sémantiquement plus riches que les agrégats classiques de l'OLAP. Nous agrégeons particulièrement les modalités d'une dimension selon l'ordre de leur proximité et non pas selon l'ordre de leur appartenance hiérarchique établi lors de la phase de conception des axes d'analyses. Cependant, d'une manière générale, vu la nature *non supervisée* de la CAH, on ne dispose pas de connaissances *a priori* ni sur le nombre, ni sur la qualité des classes qu'on va obtenir. Afin d'aider l'utilisateur dans le choix du meilleur nombre d'agrégats fournis par la CAH, nous proposons un critère d'évaluation de la qualité des partitions en se basant sur la notion de la *séparabilité* des classes.
3. Notre troisième proposition rentre dans la catégorie des techniques d'*explication* et de *prédiction*. Elle fait l'objet d'un processus d'extraction de *règles d'association* à partir des cubes de données. Ce nouveau couplage entre les règles d'association et l'analyse en ligne repose sur une adaptation d'un algorithme, de type Apriori, pour extraire des connaissances directement à partir des données multidimensionnelles. Ainsi, nous mettons en place un cadre général pour une recherche de règles *inter-dimensionnelles* à partir des cubes de données. Dans ce cadre, nous utilisons les *méta-règles inter-dimensionnelles* en vue de guider le processus de fouille vers des contextes d'analyse ciblés. Afin de les adapter au contexte de l'analyse en ligne, nous proposons aussi une redéfinition du support et de la confiance des règles d'association en y intégrant les mesures du cube. Selon cette nouvelle définition, une règle d'association n'est pas évaluée selon le nombre des faits qui la supportent mais plutôt selon la somme des mesures de ces faits. En plus du support et de la confiance, nous employons deux autres critères descriptifs (le *Lift* et l'indice de *Loevinger*) pour évaluer l'intérêt des règles d'association découvertes. Pour valoriser les connaissances extraites, nous établissons un formalisme dédié à la visualisation des règles d'association inter-dimensionnelles en se basant sur les principes de la *sémiologie graphique*.

Suite à ces contributions théoriques, nous mettons en place une plateforme logicielle générale, appelée MiningCubes. Il s'agit d'une application Web dotée d'interfaces ergonomiques, faciles à utiliser et adaptées au contexte de l'analyse en ligne. Dans cette plateforme, nous développons des modules pour concrétiser et valider, sur un plan technique, nos approches de couplage entre l'analyse en ligne et la fouille de données. Elle intègre aussi des solutions techniques qui utilisent XML comme formalisme pour la représentation d'un contexte d'analyse relatif à des données complexes.

En complément de ces travaux, nous proposons une méthodologie générale, entièrement basée sur XML, appelée X-Warehousing, pour l'entreposage des données complexes. Avec cette méthodologie, il est possible de concevoir et de créer des *cubes XML* traduisant des objectifs d'analyse concernant des données de type complexe. En guise d'un cas d'application, nous utilisons des données médicales concernant le dépistage du cancer du sein. Ces données représentent des dossiers de patientes où chaque dossier comprend des sources hétérogènes et éparpillées sur plusieurs types de supports. Dans une phase préparatoire, nous décrivons ces données complexes dans des documents XML. Nous employons ensuite X-Warehousing pour mettre en place un *cube XML de mammographies*. Nous exploitons après ce cube et mettons en œuvre, dans notre plateforme MiningCubes, notre approche d'agrégation par classification.

Enfin, à la lumière de ces travaux, nous avons été amenés à réfléchir à un cadre théorique qui définit une analyse combinant la technologie OLAP et la fouille de données dans un processus décisionnel unifié. Ainsi, nous mettons en place les premières bases d'un cadre formel général pour le problème du couplage entre l'analyse en ligne et la fouille de données. Nous proposons un *modèle de données multidimensionnelles* supportant un noyau minimal fermé d'une *algèbre OLAP*. Ce dernier inclut des opérateurs de *structuration* et de *navigation*. Il est donc possible, avec ces opérateurs, de construire un cube de données, de manipuler sa structure et d'explorer son contenu en vue de répondre à toute sorte de besoins d'analyse OLAP possibles. Nous tentons également d'étendre cette algèbre à une nouvelle génération d'opérateurs de *fouille de données en ligne*. En se basant sur nos différentes expériences de couplage entre l'analyse en ligne et la fouille de données, nous proposons une première formalisation de deux opérateurs de fouille de données en ligne. Le premier, appelé OrCA, est un opérateur de *réorganisation d'un cube de données par une ACM* qui consiste en une formalisation algébrique de notre première contribution. Le second opérateur, appelé OpAC, est dédié à l'*agrégation par une CAH dans un cube de données* et fait l'objet d'une formalisation de notre deuxième contribution.

1.3 Organisation de la thèse

Ce mémoire de thèse est organisé comme suit. Le chapitre 2 présente un *état de l'art* général du couplage de l'analyse en ligne et la fouille de données. Nous exposons les trois types d'approches que nous avons détectés. Nous relatons, pour chacune de ces approches, le contexte, les motivations et les travaux réalisés. Nous présentons également une synthèse permettant de positionner nos contributions au regard de l'existant.

Le chapitre 3 fait l'objet de notre approche de *réorganisation des cubes de données par une approche factorielle*. Nous évoquons les objectifs et les motivations qui nous ont poussés à choisir l'ACM. Nous exposons en détail les formalismes et les différentes étapes de notre approche. Nous insistons particulièrement sur l'adaptation des données multidimensionnelles pour l'ACM. Nous présentons également l'indice de la qualité des données multidimensionnelles que nous proposons. Ce chapitre inclut aussi des études de cas ainsi que des résultats expérimentaux évaluant la performance de notre réorganisation des cubes.

Nous évoquons, dans le chapitre 4, notre approche d'*agrégation par classification dans les cubes de données*. Nous montrons l'intérêt de notre approche et motivons le choix de la CAH comme une technique d'agrégation dans le contexte des données multidimensionnelles. Nous détaillons les principes de notre démarche, notamment le choix des individus et des variables de la classification. Nous présentons deux critères possibles pour l'évaluation de la qualité des classes fournies par la CAH. En complément à ces derniers, nous exposons notre critère de séparabilité des classes.

Dans le chapitre 5, nous abordons notre approche d'*explication dans les cubes de données par règles d'association*. Nous détaillons les principaux travaux ayant intégré les règles d'association dans les structures multidimensionnelles. Nous décrivons notre cadre général concernant la recherche guidée des règles d'association inter-dimensionnelles. Nous présentons aussi l'algorithme, basé sur Apriori, que nous avons amélioré et adapté pour les cubes de données. Des résultats expérimentaux sont aussi fournis afin d'apprécier les performances de notre algorithme. Ce chapitre inclut aussi un exposé sur notre proposition de visualisation des règles inter-dimensionnelles dans l'espace multidimensionnel d'un cube de données.

Le chapitre 6 présente l'*implémentation* de la plateforme MiningCubes ainsi que le *cas d'application aux données complexes*. Nous décrivons l'architecture générale de notre plateforme et détaillons ensuite ses modules prévus pour nos différentes approches de couplage entre l'analyse en ligne et la fouille de données. Nous introduisons le jeu des données médicales sur lequel nous déroulons notre méthodologie d'entreposage X-Warehousing afin de construire le cube XML de mammographies. Nous exposons les différentes étapes de l'agrégation par classification sur ce cube et discutons les résultats obtenus.

Le chapitre 7 trace les grandes lignes que nous suivons *vers un cadre formel général*. Nous survolons les principales algèbres OLAP existantes. Nous exposons notre modèle de données multidimensionnelles et détaillons les formalismes des opérateurs algébriques de notre noyau minimal fermé. Nous présentons aussi la formalisation de nos deux opérateurs de fouille de données en ligne et montrons la possibilité d'étendre notre algèbre à une base théorique générale dédiée au couplage de l'analyse en ligne et de la fouille de données.

Enfin, le chapitre 8 conclut ce mémoire en présentant un bilan général de l'ensemble de nos contributions et en évoquant de nouvelles perspectives de recherche.

État de l'art

Résumé

Dans ce chapitre, nous abordons une recherche bibliographique et nous exposons une synthèse des travaux qui traitent du problème du couplage entre l'analyse en ligne OLAP et la fouille de données. Notre synthèse est organisée selon une vision thématique qui distingue trois grandes approches pour ce couplage.

Dans une première section, nous distinguons et introduisons les différentes approches dans l'état de l'art actuel. Chacune des sections suivantes expose en détail le contexte, les motivations et les travaux réalisés dans chaque approche. Dans la dernière section de ce chapitre, nous concluons et précisons le positionnement des travaux de notre thèse par rapport à l'existant.

Sommaire