Chapitre 2

État de l'art

" S'il m'a été donné de voir un peu plus loin que les autres, c'est parce que j'étais monté sur les épaules de géants. "

Isaac Newton

2.1 Introduction

Le problème de la représentation des données est un enjeu important dans le problème du couplage entre l'analyse en ligne et la fouille de données. En effet, d'un côté, les algorithmes de fouille ne peuvent opérer que sur des données présentées sous la forme classique d'un tableau *attributs-valeurs* (connu aussi sous le nom de tableau *individus-variables*). De l'autre côté, dans le contexte d'un entrepôt de données, les données sont organisées selon une structure multidimensionnelle adaptée à l'analyse en ligne. Ainsi, la divergence des espaces de représentation des données propres aux deux domaines fait de la combinaison de l'analyse en ligne et de la fouille de données une tâche particulièrement délicate qui demande des adaptations préalables d'un côté comme de l'autre.

Imieliński et Mannila étaient les premiers qui se sont intéressés au problème général de l'intégration de l'ECD dans les systèmes de gestion de bases de données (SGBDs). Dans [IM96], les auteurs pensent déjà que la fouille dans les bases de données va aboutir à la création de nouveaux concepts, de nouvelles stratégies d'interrogation et de nouveaux langages de requêtes. Les auteurs prévoient même la naissance d'une *seconde génération* de systèmes de gestion de bases de données. Ils imaginent deux scénarii pour la suite des recherches dans ce domaine.

Le premier scénario résume une vision à court terme qui consiste à exploiter les outils actuellement disponibles dans les SGBDs et de greffer dessus des techniques d'apprentissage automatique telles que les règles d'association. Cependant, la plupart de ces techniques peuvent demander plusieurs passages sur les données ce qui alourdit

considérablement les temps de réponses. L'optimisation de ces temps de réponse est possible si on exploite les outils offerts par le SGBD tels que l'indexation, l'exécution parallèle des requêtes, les opérations d'agrégation et les capacités de stockage. Quelques projets de recherche qui ont abordé le couplage ECD et bases de données selon ce scénario ont déjà abouti à la commercialisation de produits tel que *IBM's Intelligent Miner*.

Quant au deuxième scénario, il illustre une évolution à long terme qui doit miser sur une intégration plus profonde de l'ECD dans l'architecture et les outils des SGBDs. Les systèmes KDDMS (*Knowledge and Data Discovery Management Systems*) constitueront la deuxième génération des SGBDs. Cette dernière va accumuler les expériences de la technologie des bases de données. Les requêtes, les compilateurs et les optimisateurs des KDDMS seront de plus en plus adaptés et serviront pour la fouille de données. Les auteurs rajoutent que, dans un KDDMS, les requêtes doivent jouer un double rôle : la génération des objets ECD à partir des données initiales, la réintégration et le stockage de ces objets dans la base de données. De plus, les requêtes ECD doivent être capables d'appréhender aussi bien les objets ECD que les objets classiques dans une base de données.

Depuis la fin des années 90, plusieurs travaux ont été proposés dans le cadre du couplage entre l'analyse en ligne et la fouille de données. Ces travaux ont abordé le problème selon des motivations et des approches différentes. D'une manière générale, nous distinguons trois grands groupes d'approches dans ce domaine :

1. le **premier groupe** d'approches consiste à *transformer* les données multidimensionnelles en données bi-dimensionnelles afin de les rendre exploitables par les algorithmes classiques de fouille ;
2. le **deuxième groupe** concerne des approches de type instrumental qui tirent partie des spécificités et des outils offerts dans les systèmes de gestion de bases de données multidimensionnelles (SGBDMs). Ces approches consistent à étendre les opérations OLAP ou les langages de requêtes SQL et à utiliser ces derniers comme instruments pour extraire et transmettre les données nécessaires pour la construction d'un modèle d'apprentissage ;
3. le **troisième groupe** comprend les approches qui ont pour but de *faire évoluer* les algorithmes de fouille de données et de les adapter aux espaces de représentations multidimensionnelles des données. Ainsi, selon ces approches, on peut appliquer des algorithmes *évolués* directement dans les cubes de données.

Nous détaillerons dans la suite les différents groupes d'approches. Certes, s'agissant d'un domaine de recherche en plein essor, nous essayons de présenter une liste la plus exhaustive possible des références traitant du couplage de la fouille de données et de l'analyse en ligne. Néanmoins, nous présentons les travaux les plus intéressants et qui répondent au mieux à la problématique étudiée.

2.2 Adaptation des données multidimensionnelles

Ce premier groupe d'approches consiste à faire un rapprochement entre les algorithmes classiques de fouille et les données multidimensionnelles moyennant l'adaptation de ces dernières.

2.2.1 Pré-traitement des données multidimensionnelles avec l'OLAP

Dans [CZC01], Chen *et al.* introduisent la plateforme IIMiner (*Integrated Interactive Data Miner*) pour la fouille des données hétérogènes qui proviennent de sources différentes. D'une manière générale, avec le développement de la technologie des entrepôts de données, les auteurs pensent qu'il est naturel de voir une émergence de projets visant l'intégration de la fouille de données avec les outils OLAP dans les systèmes décisionnels. Dans la plateforme proposée, les auteurs définissent un processus ECD selon lequel les entrepôts de données sont le support des données et la technologie OLAP permet d'effectuer des pré-traitements sur ces données. Ainsi, un processus ECD est une succession d'étapes prises en charge par l'entreposage de données, l'analyse en ligne OLAP et la fouille de données.

Dans la plateforme IIMiner, Chen *et al.* cherchent des corrélations entre les données de l'entrepôt. Pour cela, ils utilisent des opérations OLAP pour mettre en forme les données, concernées par l'apprentissage, selon un tableau individus–variables. Les auteurs emploient ensuite la méthode des réseaux bayésiens afin de découvrir et de représenter graphiquement les causalités des données.

Maedche *et al.* proposent également d'utiliser l'OLAP comme outil de pré-traitement pour des données de télécommunication [MHW00]. Leur approche combine les bases de données multidimensionnelles avec les systèmes classiques de fouille de données en utilisant les outils OLAP comme interface (voir figure 2.1). D'une manière générale, les auteurs affirment que plus le volume des données est grand, plus leur compréhension et leur pré-traitement deviennent difficiles. La vocation de l'analyse en ligne est de gérer et d'explorer des grands volumes de données. En plus, l'OLAP permet une bonne interaction entre l'utilisateur et la base de données.

Dans le cadre de leur application, Maedche *et al.* proposent donc de créer, à l'aide d'outils classiques de l'analyse en ligne, un processus flexible pour comprendre et nettoyer les grands volumes de données relatifs au domaine des télécommunications. Ces données nettoyées sont mises en forme tabulaire et sont chargées ensuite dans une composante de fouille de données. Dans [MHW00], les auteurs proposent d'utiliser la méthode des *k-means* pour classifier les abonnés du service téléphonique selon leurs profils de consommation.

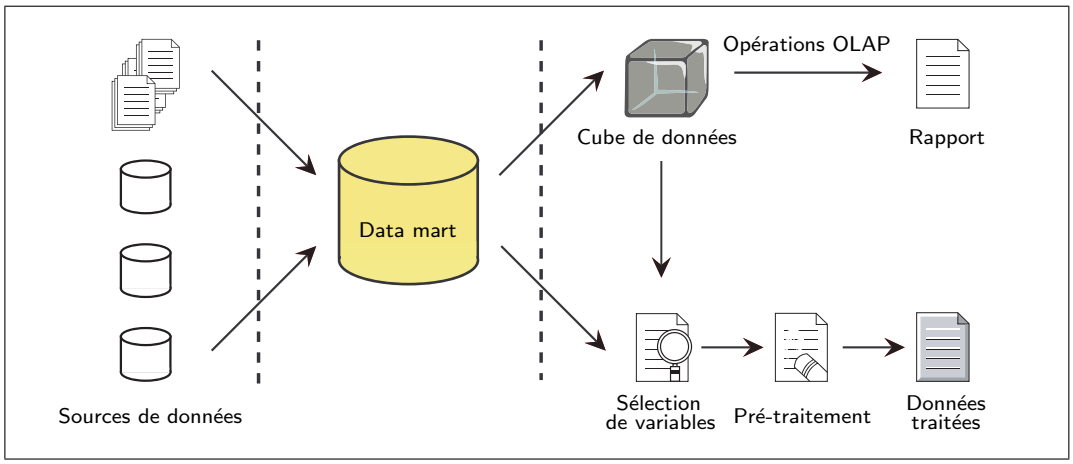

FIG. 2.1 – Pré-traitement des données avec les outils OLAP [MHW00]

2.2.2 Mise en forme des données multidimensionnelles avec l'OLAP

Goil et Choudhary affirment que les techniques de fouille de données peuvent être appliquées en conjonction avec les outils de l'analyse en ligne [GC99]. Ils soulignent également qu'une structure multidimensionnelle des données peut représenter une base d'apprentissage plus riche qu'une structure classique. Dans le cadre d'une plateforme parallèle PARSIMONY dédiée à l'analyse OLAP et la fouille de données, les auteurs proposent un classement dans les données multidimensionnelles par arbres de décision [GC99, GC01]. Cette approche consiste à utiliser les outils OLAP pour extraire, à partir d'un cube de données, une matrices de contingence pour chaque dimension et à chaque étape de la construction de l'arbre de décision. Ces matrices sont exploitées pour le calcul des *indices de Gini* afin de déterminer la variable d'éclatement de la prochaine itération.

Par ailleurs, Zaïane *et al.* proposent le système WebLogMiner qui emploie la fouille de données, l'entreposage et l'OLAP pour traiter et analyser les données des fichiers logs du Web (*web log records*) [ZXH98]. Dans cette proposition, les auteurs définissent un processus de traitement en quatre phases. La première consiste à filtrer à partir des fichiers logs les données les plus pertinentes au sens de l'analyse future. Ces données sont stockées dans une base de données relationnelles selon un schéma multidimensionnel. Dans la deuxième phase, un cube de données est construit à partir de la base relationnelle. Les opérations OLAP, tels que le forage vers le haut (*roll-up*) et vers le bas (*drill-down*), la sélection (*slice*) et la projection (*dice*), représentent les outils utilisés dans la troisième phase pour extraire et mettre en forme les informations à fouiller dans la phase suivante. Dans la phase de fouille, Zaïane *et*

al. mettent en œuvre une analyse de *séries temporelles*. Cette analyse concerne l'étude du trafic sur le réseau, les transactions, les séquences et les habitudes des internautes. Dans [ZXH98], les auteurs affirment que, avec la technologie OLAP, l'analyse des séries temporelles qu'ils proposent peut s'opérer sur plusieurs dimensions et plusieurs niveaux de granularité contrairement aux séries temporelles classiques.

2.2.3 Aplatissement et préparation des données d'un entrepôt

Dans un contexte d'extraction des règles d'association à partir des entrepôt de données, Tjioe et Taniar proposent des formalismes de pré-traitement des données multidimensionnelles avant la phase de recherche des *motifs fréquents* [TT05]. Ces formalismes préparent les données à fouiller d'une manière ciblée en vue de faciliter la recherche des motifs les plus intéressants au sens de l'analyse souhaitée par l'utilisateur. Les auteurs proposent quatre algorithmes de pré-traitement des données dans un cube : VAvg, HAvg, WMAvg, et ModusFilter. L'idée générale de ces algorithmes consiste à transformer les données d'un cube sous forme tabulaire dans un premier temps et d'élaguer dans un second temps les données inintéressantes ayant des valeurs inférieures à la moyenne par ligne ou par colonne. Les tableaux de données obtenus (*initialized tables*) sont ensuite utilisés comme entrée d'un algorithme classique de recherche de *motifs fréquents* et d'extraction de *règles d'association*.

Dans [Fu05], Fu pense que, dans un système d'aide à la décision, l'emploi d'un entrepôt de données et de l'analyse en ligne est une solution simpliste qui ne répond pas aux besoins de l'extraction des connaissances. Par conséquent, l'auteur propose une architecture d'un système intégré qui combine un SGBD pour les données multidimensionnelles, une composante OLAP et une composante OLAM (*Online Analytical Mining*). Comme le montre la figure 2.2, selon cette architecture, les utilisateurs peuvent soumettre des requêtes SQL, CQL ou DMQL (*Data Mining Query Language*) via une interface commune. La requête de l'utilisateur est ainsi analysée par un *parseur* qui va l'acheminer vers les différentes composantes du système. En cas d'une incohérence syntaxique de la requête, le parseur renvoie un message d'erreur.

Dans le cadre de ce système, l'auteur introduit aussi un classifieur, appelé CubeDT, qui construit des *arbres statistiques*. Un arbre statistique est une structure multidimensionnelle particulière inspirée des *arbres de décision* [FH00]. Cependant, l'algorithme CubeDT travaille sur des données extraites et aplaties par une composante de chargement (*Loader*) à partir d'un entrepôt de données via le serveur OLAP du système.

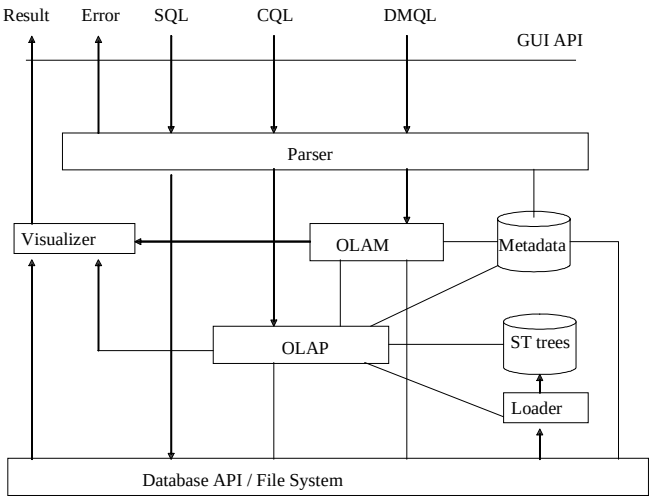

FIG. 2.2 – Architecture d'un système intégrant SGBD, OLAP et MOLAP [Fu05]

2.3 Extension de l'analyse OLAP et des langages de requêtes

Les origines de ce deuxième groupe d'approches de couplage entre l'OLAP et la fouille de données remontent aux propositions d'implantation de la fouille dans les bases de données relationnelles. En effet, nous estimons que l'utilisation des outils des SGBDMs pour la fouille des données multidimensionnelles s'inscrit dans une logique de continuité avec les efforts d'intégration de la fouille dans les SGBDs relationnels. À titre d'exemple, selon une approche relationnelle, Meo *et al.* [MPC96] ont proposé un opérateur SQL pour la recherche de règles d'association dans les bases de données relationnelles. Cet opérateur consiste en une extension de la syntaxe de SQL en y intégrant une nouvelle close MINE RULE. Dans [STA98], Sarawagi *et al.* ont largement étudié, moyennant une extension SQL, l'intégration de la découverte des règles d'association dans les SGBDs. Afin d'éviter des temps de traitements important engendrés par les entrées-sorties dans une base relationnelle, d'autres travaux ont tenté d'exploiter les outils propres aux SGBDs pour y intégrer la fouille. Par exemple, Bentayeb *et al.* [BDU04, UBDB04] ont proposé d'intégrer la fouille par arbre de décision ID3 [Qui86] à l'aide de procédures PL/SQL stockés dans Oracle.

2.3.1 Fouille de données en ligne

Les premières tentatives d'extension des outils OLAP pour la fouille de données remontent à 1997 avec les travaux de Han [Han97]. Ces travaux ont abouti à la

création du système DBMiner. Ce dernier est doté d'outils d'exploration graphique et de visualisation spatiale des cubes de données.

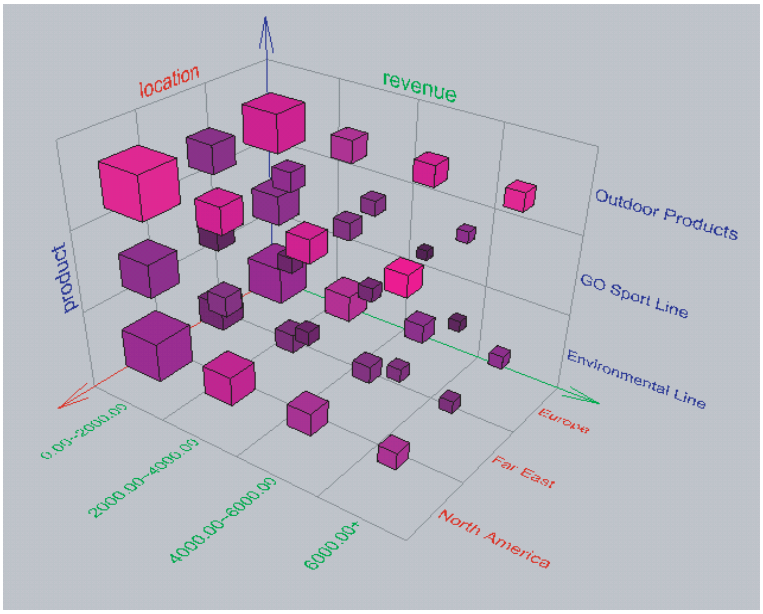

FIG. 2.3 – Exemple d'une exploration d'un cube à trois dimensions dans DBMiner [Han97]

Han définit la notion de l'*OLAP Mining* comme un mécanisme qui intègre des tâches de fouille de données dans des requêtes décisionnelles. Ce mécanisme peut s'appliquer à différents niveaux de granularité des données et à différentes parties d'un entrepôt de données [Han97, Han98]. Dans [HCC98], Han *et al.* introduisent déjà la terminologie *On-Line Analytical Mining* (OLAM) pour un processus d'analyse où les techniques de fouille sont utilisées, au même titre que les opérations OLAP, pour extraire des connaissances. Avec le processus OLAM, les auteurs prévoient même que les entrepôts de données seront, dans l'avenir, une large plateforme pour l'apprentissage automatique.

DBMiner repose essentiellement sur une instrumentation à l'aide des opérateurs OLAP en leur ajoutant des extensions aptes à simuler diverses techniques de fouille de données telles que la détection de *règles d'association*, la *caractérisation d'attributs*, le *classement*, la *prédiction*, etc. Cependant, à nos yeux, les références relatives à DBMiner [Han97, Han98, HCC98] décrivent plutôt le côté fonctionnel de ce dernier et manquent de précisions sur les formalismes et les fondements théoriques employés pour associer l'analyse en ligne et la fouille de données.

2.3.2 Fouille des bases de données multidimensionnelles

Chaudhuri, de *Microsoft Research*, propose une approche générale consistant à intégrer la fouille de données dans les SGBDMs à travers une extension du langage de requêtes SQL [Cha98]. L'auteur met en avant la possibilité de profiter de SQL et de l'aspect relationnel pour implanter des techniques de fouille dans les entrepôts de données. Selon [Cha98], avec une extension de SQL, les données d'un entrepôt peuvent être orientées vers la fouille et l'analyse en ligne. L'auteur rajoute que les SGBDMs, supportant le langage de requêtes SQL, fournissent un ensemble de primitives de données facilement exploitable par la plupart des algorithmes de fouille de données.

Chaudhuri définit les termes d'une fouille adéquate à la structure relationnelle des entrepôts de données (*Ad hoc Mining*). Ce processus consiste à appliquer la fouille sur des données extraites à la volée par des requêtes SQL appropriées. Des opérations OLAP appropriées peuvent aussi intervenir dans ce processus pour spécifier et extraire l'ensemble des données à fouiller. L'auteur suggère aussi d'exploiter les outils de sélection de variables afin de réduire la dimensionnalité des données et d'en extraire celles qui sont les plus intéressantes au sens de l'analyse souhaitée.

L'auteur pense que, avec une extension de SQL, il est possible d'intégrer des primitives de fouille de données dans le langage de requêtes, ce qui peut apporter des améliorations significatives des performances de la fouille. En guise de concrétisation de l'approche de Chaudhuri, *Microsoft Research* a intégré des outils logiciels (*middleware*) pour la fouille de données dans Microsoft SQL Server 7.0 (voir figure 2.4). Ces outils comprennent essentiellement des techniques de classement telles que les *arbres de décision* et les *réseaux bayésiens*.

Dans [CFB97, CFB99], Chaudhuri *et al.* affirment que la construction des *arbres de décision* qu'ils proposent est capable de prendre en compte un grand volume de données relationnelles. Les accès aux données sont assurés par des requêtes SQL à chaque étape de construction d'un nœud de l'arbre. Une requête consiste à calculer le nombre d'enregistrements, dans la table de faits de l'entrepôt, dont les attributs satisfont les conditions du nœud en question.

D'une manière générale, Chaudhuri *et al.* proposent deux solutions possibles pour intégrer un arbre de décision dans un SGBDM. La première consiste à générer à la volée pour chaque nœud de l'arbre une requête SQL pour extraire les données nécessaires à ce nœud. La seconde solution consiste à générer des requêtes SQL pour créer des tables intermédiaires dans la base de données. C'est-à-dire, chaque nœud de l'arbre va correspondre à une table spécifique dans la base de données.

Dans [GC98a, GC98b], Goil et Choudhary présentent une plateforme parallèle de construction des cubes à partir des bases de données relationnelles. Ils proposent également d'intégrer des *règles d'association* dans les cubes construits. Les auteurs exploitent les opérations OLAP pour extraire les agrégats pré-calculés du cube afin de calculer le *support* et la *confiance* des *règles d'association*. Les auteurs proposent

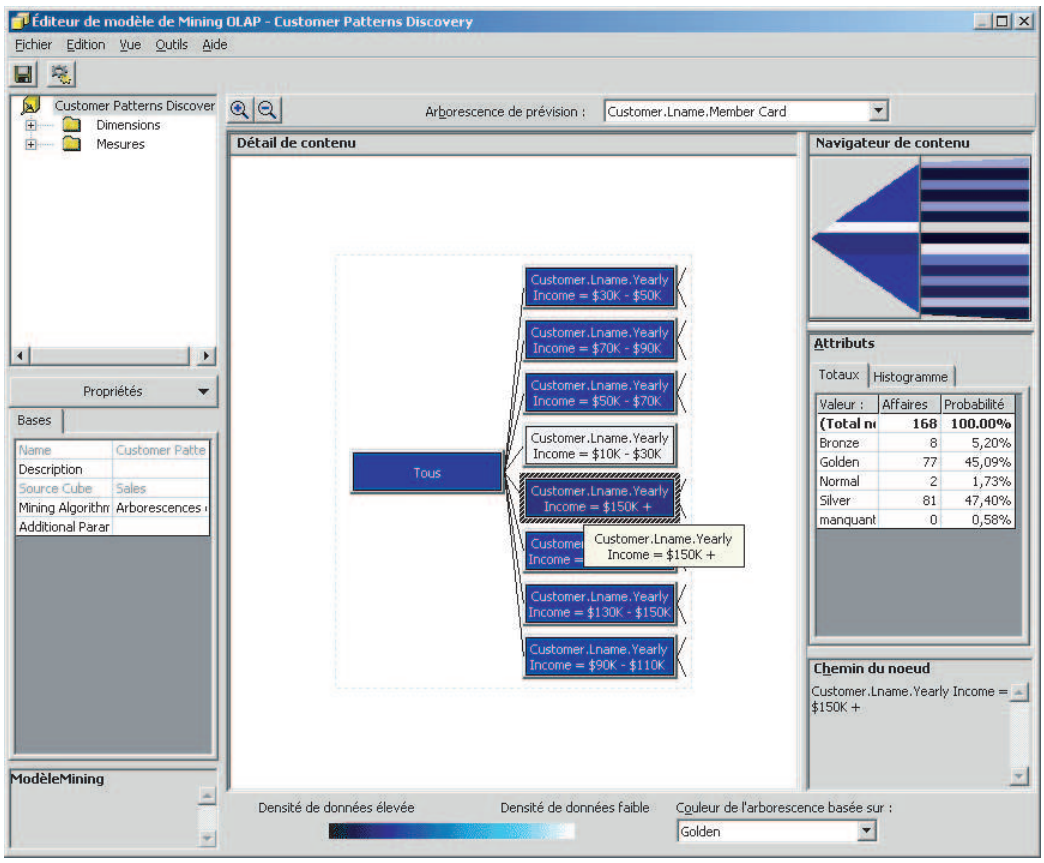

FIG. 2.4 – Arbre de décision dans Microsoft SQL Server 7.0

aussi d'utiliser ces agrégats pour le calcul de l'indice de l'écart à l'indépendance d'une règle. Ainsi, selon cette approche, des opérations OLAP sont utilisées comme outils pour extraire et calculer les différents critères d'une règle.

2.3.3 Utilisation des SGBDMs pour la découverte des connaissances

Laurent *et al.* proposent aussi d'exploiter le langage de requêtes des SGBDMs pour étendre ces derniers à des méthodes d'apprentissage automatique. Pour cela, les auteurs font coopérer un SGBDM et un système de construction d'*arbres de décision flous* [LGM00, LBMD+00]. L'objectif de cette coopération est de tirer profit des spécificités et des avantages du SGBDM pour aider l'algorithme d'apprentissage pendant la construction de son modèle de connaissances. Ainsi, la méthode de fouille n'a plus à gérer la base d'apprentissage et ne prend plus en charge les contraintes de stockage et de manipulation des données. Ces tâches sont assurées par le SGBDM.

Tout de même, les auteurs expliquent que dans ce type d'association entre SGBDMs et fouille de données, le problème principal réside dans l'équilibrage des rôles des deux composantes. En particulier le SGBDM peut soit envoyer à l'algorithme de fouille des agrégats simples ; soit calculer des agrégats plus complexes, incluant des opérations statistiques ou logiques par exemple, et d'en transmettre les résultats à l'algorithme de fouille. La première possibilité permet d'intégrer diverses techniques de fouille vu que la majorité de celle-ci s'appuient sur des agrégats de type comptage des données et gèrent elles-mêmes les traitements spécifiques de ces résultats. Cependant, cette solution demande un grand nombre de requêtes au niveau du SGBDM et par conséquent, un grand nombre d'échanges entre le SGBDM et l'algorithme de fouille. Les auteurs qualifient le niveau de ces échanges de *niveau élémentaire*, où les valeurs associées à chacune des modalités de chacun des attributs sont transmises par le SGBDM. La situation alternative suppose que le SGBDM est assez puissant pour effectuer des traitements capables de générer des calculs complexes d'agrégats partiels. Cette solution a l'avantage de diminuer les échanges entre SGBDM et algorithme de fouille, mais limite considérablement le nombre de techniques d'apprentissage qui peuvent être intégrées. Selon les auteurs, les échanges se font à un *niveau intégré* où une valeur globale pour chaque attribut est transmise par le SGBDM au système d'apprentissage.

Dans le cadre de leur approche, Laurent *et al.* développent une application qui fait coopèrer le SGBDM Oracle Express avec le logiciel d'apprentissage Salammbô qui construit des *arbres de décision flous*. Afin d'améliorer les performances de leur application, les auteurs adoptent le *niveau intégré* d'échange où les calculs complexes sont exécutés au niveau du SGBDM. Dans le processus de construction d'un arbre de décision flou, Salammbô prépare une requête spécifiant la position courante dans l'arbre et l'envoie au SGBDM. Ce dernier interroge le cube de données concerné et calcule les entropies pour chacune des dimensions du cube et les renvoie. Les auteurs précisent que le calcul de l'entropie peut s'effectuer grâce à des opérations OLAP, tels que le forage vers le haut (*roll-up*), la sélection (*slice*) et la projection (*dice*), ou grâce aussi à des programmes internes disponibles dans Oracle Express.

2.3.4 Extension de l'OLAP à la découverte des connaissance

Dans [NNQ04], Naouali *et al.* affirment que les algorithmes d'extraction de connaissances, classiquement prévus pour les données tabulaires, ne sont pas adaptés pour le contexte de données multidimensionnelles caractérisées principalement par leur aspect hiérarchique.

Naouali *et al.* proposent d'extraire des *motifs fréquents* à partir de la table de faits d'un cube de données, où chaque motif représente un fait OLAP. Les auteurs utilisent un algorithme de découverte des motifs fréquents adapté aux grandes bases de données caractérisées par des données denses [TNBP00]. Selon les auteurs, ces

motifs fréquents permettent de mettre en évidence des *liens sémantiques* traduisant des relations intéressantes entre les cellules du cube étudié. Une segmentation de ces motifs fréquents par algorithmes génétiques est aussi proposée. Ainsi l'approche établit des classes de liens sémantiques entre les cellules du cube.

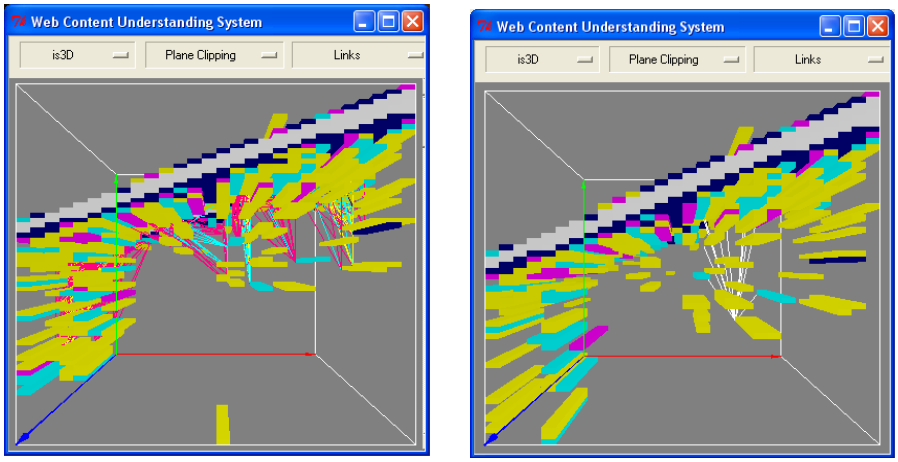

FIG. 2.5 – Exploration des liens sémantiques entre les cellules d'un cube [NNQ04]

Les auteurs proposent aussi d'étendre les fonctionnalités de la technologie OLAP afin de prendre en compte graphiquement ces nouvelles connaissances dans le cube de données étudié. Cette approche consiste à fournir à l'utilisateur une nouvelle représentation visuelle des cubes incluant les liens sémantiques existant dans ces derniers. Cette représentation est aussi munie d'outils de navigation, d'exploration et d'analyse en ligne (voir figure 2.5). Ainsi, en explorant les liens sémantiques des cellules d'un cube, OLAP est capable de fournir de nouvelles capacités de découverte de connaissances dans les structures multidimensionnelles des données.

Dans [MJBN06], Missaoui *et al.* établissent un cadre de couplage entre les entrepôts de données et l'extraction des *motifs fréquents* et des *règles d'association*. Cette proposition repose principalement sur la notion des *trellis de concepts*.

Dans cette approche, un *trellis de concepts* est exploité afin de représenter graphiquement les données d'un entrepôts. Comme le montre l'exemple de la figure 2.6, une telle représentation permet de mettre en valeur les relations existantes entre un ensemble d'*objets* (transactions) et un ensemble d'*attributs* (éléments d'une base de transactions). Ces relations sont, par la suite, exploitées en vue d'extraire des *motifs fréquents fermés* et des *règles d'association* à partir des données de l'entrepôt.

Les auteurs définissent aussi un nouveau type d'opérateurs dédiés à une *fouille de données à la demande* [MJBN06]. Ces opérateurs emploient des mécanismes, semblables à ceux des opérateurs OLAP classiques, permettant de naviguer dans

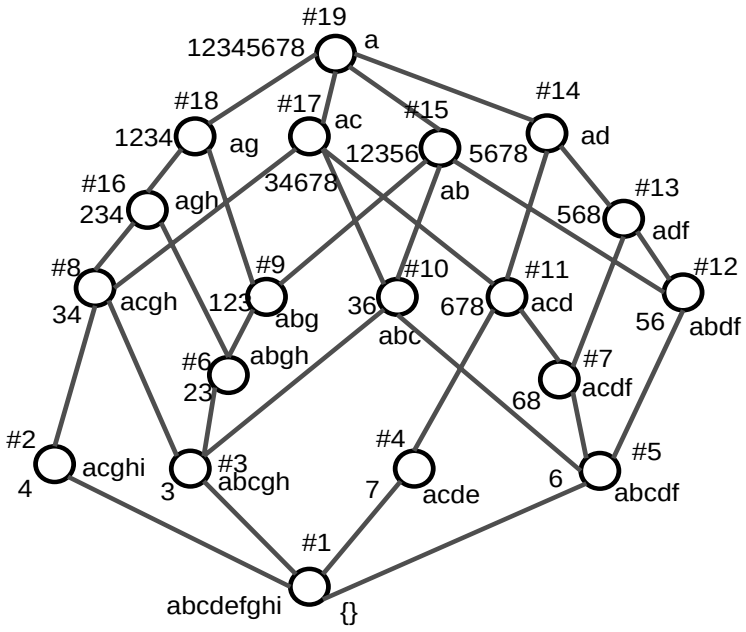

FIG. 2.6 – Exemple d'un trellis de concepts représentant des transactions [MJBN06]

un trellis de concepts par *sélection*, *projection* et *jointure de trellis*. Cette extension des opérateurs OLAP est aussi supportée par des outils visuels. Ainsi, un utilisateur est capable de visualiser et d'explorer le résultat de l'extraction des motifs fréquents fermés et des règles d'association dans le trellis de concepts.

Dans Liu [LZBX06], Liu *et al.* utilisent également la technologie OLAP pour explorer des règles d'association. Les *règles d'association* sont extraites à partir de la base de données de *Motorola*. L'objectif de ces règles et de détecter les causes d'échec des appels téléphoniques.

Cependant, dans ce cas d'application, les règles extraites sont nombreuses, ce qui pose un problème dans la phase d'exploitation et l'évaluation de l'intérêt des connaissances extraites. De plus, une règle et plus intéressante lorsqu'elle est comparée à d'autre règles complémentaires. Ainsi, Liu *et al.* propose une plateforme d'exploration des règles d'association basée sur la technologie OLAP. Selon un formalisme introduit dans [LZBX06], les règles d'association sont structurées selon l'organisation multidimensionnelle des cubes OLAP. En se basant sur des outils visuels et des opérateurs classiques d'OLAP, la plateforme permet aux utilisateurs d'explorer et de naviguer dans les règles d'association extraites à partir des données téléphoniques.

2.4 Adaptation des techniques de fouille de données

La troisième approche fait appel à un emploi direct des algorithmes de fouille dans les données multidimensionnelles. Dans ce cas, un travail d'adaptation de ces algorithmes est nécessaire pour établir l'interaction nécessaire entre l'algorithme de fouille et les données multidimensionnelles.

2.4.1 Extraction des connaissances à partir des cubes de données

Peu nombreux sont les travaux qui ont abordé le couplage de l'analyse en ligne et de la fouille de données selon cette approche. Palpanas explique ceci par la nouveauté relative de la technologie OLAP et par la concentration des efforts de recherche sur le domaine de la fouille de données [Pal00].

Palpanas pense que, devant la richesse des données multidimensionnelles et leur modélisation adéquate pour le domaine décisionnel, une analyse approfondie de ces données, basée sur la fouille de données, donnerait lieu à des modèles de connaissances plus valorisants que le cas de la fouille classique. L'auteur affirme qu'une analyse complète doit intégrer aussi bien les opérations OLAP que les techniques de fouille dans un seul processus de découverte des connaissances. Dans ce processus, l'OLAP doit constituer un automate qui propose à l'analyste des pistes pour guider sa tâche d'exploration des données multidimensionnelles. Tout de même, Palpanas affirme que la structure multidimensionnelle peut servir de source pour l'extraction de modèles de connaissances plus riches et qui sont introuvables dans les données tabulaires.

Palpanas propose des directions de recherche prometteuses pour l'intégration de la fouille dans des structures multidimensionnelles des données. Il affirme également que les algorithmes d'apprentissage finiront par s'adapter aussi bien aux opérateurs OLAP qu'à la structure multidimensionnelle et hiérarchique des données. Ceci les rendra capables de produire des connaissances à différents niveaux de granularité de l'information [Pal00].

D'une manière semblable, Parsaye [Par97] propose un système théorique, appelé OLAP Data Mining System, évoluant dans un espace hybride formé par des données et des agrégats. Ce système comprend trois composantes : une base de données relationnelle pour l'entreposage des données, un système MOLAP ou ROLAP pour la structuration et l'accès aux données et une composante de découverte de connaissances dans les données multidimensionnelles (*multidimensional discovery engine*).

2.4.2 Généralisation des règles d'association aux cubes de données

Imieliński *et al.* proposent une adaptation des règles d'association au contexte des données multidimensionnelles. Dans [IKA02], les auteurs introduisent le concept des

cubegrades (*cubes de données différentielles*). Il s'agit d'une généralisation des cubes de données et des règles d'association. Un cubegrade est un formalisme qui calcule le différentiel des mesures agrégées d'un cube de données par passage d'un *cube source* à un *cube cible*. Un tel passage peut correspondre à une opération de spécification (*drill-down*), de généralisation (*roll-up*) ou de permutation d'une modalité dans une dimension (*switch*). Par exemple, un cubegrade permet de voir de combien est affectée la moyenne des âges des consommateurs de *pain* quand on spécialise la population à celle des consommateurs de *pain* et de *lait*. En d'autres termes, un cubegrade exprime de combien un agrégat d'un cube de données peut varier lors de modifications de structure sur ce cube.

Selon Imieliński *et al.*, les cubegrades sont des atomes de connaissances qui expliquent le comportement des agrégats dans différents segments d'une base de données. Ils considèrent aussi les cubegrades comme une nouvelle formulation des connaissances hybrides qui combinent à la fois règles d'association et analyse en ligne. Un langage de requête, appelé CGQL (*CubeGrades Query Language*), a été également introduit dans [IKA02] pour interroger les cubegrades dans une base de données multidimensionnelles.

Dans [DHL+01], Dong *et al.* ont repris les travaux de Imieliński *et al.* et ont proposé des améliorations dans le concept des cubegrades. À cet effet, les auteurs introduisent la notion des constrained gradients qui respecte une *contrainte de significativité*. Une telle contrainte permet de contourner le problème de volumétrie des cubes de données à fouiller. Ainsi, la recherche des cubegrades se limite à la partie *significative* du cube qui satisfait la contrainte.

Classiquement, la recherche des cubegrades consiste à comparer chaque cellule dans un *cube source* avec les autres cellules dans le *cube cible*. Dong *et al.* soulignent que, même avec la contrainte de significativité, les cubegrades générés demeurent toujours nombreux. Par conséquent, les auteurs proposent aussi de prendre en compte une deuxième *contrainte probabiliste* permettant de restreindre la recherche des constrained gradients.

Les auteurs ajoutent que, dans une analyse OLAP, on ne s'intéresse souvent qu'à certains niveaux de changements entre la cellule source et la cellule cible. Par exemple, un utilisateur ne s'intéresse qu'aux cellules dont la moyenne augmente de plus de 40%. Les auteurs, introduisent un seuil pour les mesures des cellules à choisir. Les paires de cellules dont les mesures varient avec des taux supérieurs au seuil sont appelées *cellules gradients* (*gradient cells*) et le seuil est appelé la contrainte du gradient (*gradient constraint*). L'algorithme LiveSet-Driven algorithm est également proposé dans [DHL+01] pour la recherche des constrained gradients selon les trois contraintes développées.

2.4.3 Modèles statistiques dans les cubes de données

Dans [SAM98], Sarawagi *et al.* proposent un outil d'identification des régions remarquables dans les cubes de données. Habituellement, pour détecter des exceptions ou des valeurs aberrantes dans un cube, un utilisateur est amené à explorer un espace de données multidimensionnelles. Cette tâche devient de plus en plus difficile avec les cubes de données volumineux. Pour remédier ce problème, Sarawagi *et al.* introduisent un modèle statistique intégré dans un serveur OLAP (*Discovery-driven*) pour assister l'utilisateur dans sa tâche d'exploration des cubes de données. Ce modèle a pour vocation de guider l'utilisateur à détecter les motifs des données remarquables suivant plusieurs dimensions et à différents niveaux de granularité.

Le fondement du modèle se base essentiellement sur la comparaison des valeurs prédites des cellules avec leur contenu réel. Une combinaison avec les différentes dimensions de ces cellules est envisagée pour la vérification de l'aberrance du contenu. Statistiquement, la prédiction de la valeur d'une cellule est assurée par une *modélisation log-linéaire* qui construit des équations expliquant la valeur prédite en fonction des agrégats de ses dimensions. Cependant, l'implémentation de cette approche n'est pas évidente du moment où elle doit tenir compte des différentes dimensions d'un cube, ainsi que les différents agrégats de chaque dimension et de l'ensemble des combinaisons possibles de ces dimensions. À ce propos, dans [SAM98], les auteurs utilisent des méthodes d'optimisation qui réduisent les coûts de traitements.

Une amélioration de ces travaux a été proposée par Sarawagi dans [Sar99, Sar01]. Cette amélioration concerne une meilleure automatisation de l'analyse par l'emploi de la programmation dynamique. Cette automatisation est garantie par un nouvel opérateur, appelé iDiff, qui détecte les régions remarquables et explore les raisons de présence de ces régions dans un cube de données. Ces raisons sont exprimées, sous forme de tableaux sommaires, en fonction des valeurs d'autres cellules du cube appartenant à des niveaux d'agrégation plus fins et en corrélation avec les cellules de départ. Un prototype est implémenté pour cet opérateur sur le serveur DB2/OLAP d'IBM. Des expérimentations ont démontré un niveau de performance acceptable en fonction du nombre de tuples et des granularités choisies.

Dans [RF01], Favero et Robin ont adopté une approche semblable à celle de Sarawagi. Ils proposent le système HYSSOP (*HYpertext Summary System of On-line analytical Processing*) pour générer automatiquement des statistiques quantitatives extraites à partir des cubes de données. Ces statistiques sont exprimées en langage naturel intégrant des liens hypertextes. Les auteurs considèrent que l'association entre la fouille de données et l'analyse en ligne est capable d'accomplir des analyses quantitatives du contenu d'un cube [FR00, RF01]. Ils proposent une composante de fouille de données (*Content Dertermination*), intégrée dans HYSSOP, qui concrétise cette approche en utilisant les hiérarchies du cube pour classifier les données. Les

résultats de ce module sont pris en charge par un générateur de langage naturel (*Natural Language Generation*) afin fournir des résumés textuels compréhensibles par l'humain.

2.5 Conclusion et positionnement

À la lumière de cet état de l'art, nous faisons la distinction entre trois grandes approches traitant du problème du couplage de l'analyse en ligne et de la fouille de données. Dans la suite, en vue de positionner nos contributions, nous exposons une synthèse de l'ensemble des travaux existants. Cette synthèse repose sur une organisation thématique qui croise les trois approches, que nous avons détectées, avec les trois familles de techniques de fouille de données, à savoir : (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*.

Adaptation des données multidimensionnelles			
	Type de techniques de fouille		
Proposition	Visualisation et description	Structuration et classification	Explication et prédiction
Chen *et al.*			Réseaux Bayésiens
Maedche *et al.*		k-means	
Goil et Choudhary			Arbres de décision
Zaïane *et al.*			Séries temporelles
Tjioe et Taniar			Règles d'association
Fu			Arbres de décision
Notre proposition	**ACM**		

TAB. 2.1 – Comparaison des propositions de couplage de l'OLAP et de la fouille de données selon la première approche

La première approche de couplage de l'analyse en ligne et de la fouille de données regroupe les travaux qui préconisent la transformation des données multidimensionnelles en données tabulaires. Cette approche, bien que simple et intuitive, permet tout de même d'extraire des connaissances à partir de données provenant de structures multidimensionnelles. Cependant, d'une manière générale, la transformation des données multidimensionnelles en données tabulaires présente le risque de faire perdre à ces dernières leur aspect hiérarchique.

De plus, comme le montre le tableau comparatif 2.1, mise à part la proposition de Maedche *et al.* [MHW00] où les auteurs font de la *classification* des consommateurs selon leurs profils, toutes les autres propositions utilisent des méthodes d'*explication*

et de *prédiction* telles que les *réseaux bayésiens*, les *arbres de décision* et les *règles d'association*.

Selon cette première approche, nous couplons l'analyse en ligne avec une méthode factorielle dédiée à la *visualisation* et à la *description* [MRB05]. Concrètement, nous utilisons l'analyse des correspondances multiples (ACM) dans le but d'améliorer la représentation des faits dans un cube de données [MBR06d, MBR06b]. Dans une phase préparatoire, les données du cube sont transformées en *tableau disjonctif complet* selon un codage binaire approprié. L'application de l'ACM, sur ce dernier, fournit une réorganisation des modalités dans les dimensions du cube. Grâce à cette réorganisation, nous parvenons à fournir des points de vue intéressants qui homogénéisent au mieux le nuage des faits dans le cube. Ainsi, notre proposition permet de pallier le problème, souvent rencontré, de la visualisation des données multidimensionnelles engendré par la volumétrie et l'éparsité des ces dernières [MAF05]. En plus, afin de valider l'apport de la réorganisation du cube, nous proposons un indice d'homogénéité pour la mesurer la qualité de représentation des données multidimensionnelles [MBR05].

Extension de l'analyse OLAP et des langages de requêtes			
	Type de techniques de fouille		
Proposition	Visualisation et description	Structuration et classification	Explication et prédiction
Han *et al.*			Règles d'association Arbres de décision
Chaudhuri *et al.*			Arbres de décision Réseaux Bayésiens
Goil et Choudhary			Règles d'association
Laurent *et al.*			Arbres de décision flous
Naouali *et al.*			Motifs fréquents
Missaoui *et al.*			Motifs fréquents fermés Règles d'association
Liu *et al.*			Règles d'association
Notre proposition		**CAH**	

TAB. 2.2 – Comparaison des propositions de couplage de l'OLAP et de la fouille de données selon la deuxième approche

La deuxième approche est instrumentale et consiste à exploiter ou à étendre des outils existants à des tâches de fouille de données. Cette extension peut porter sur les SGBDMs, les langages de requêtes SQL ou les opérations OLAP.

Cette approche est intéressante car elle permet d'intégrer la fouille de données dans un SGBDM [Cha98] ou dans des modules d'analyse annexes [CFB97, CFB99]. Elle permet aussi d'établir une coopération entre un SGBDM et un logiciel externe

pour la fouille de données [LGM00, LBMD+00]. Le langage de requêtes SQL est donc utilisé afin d'assurer la communication entre la source de données et l'algorithme de fouille. Profitant de sa capacité d'interroger de grandes bases de données, SQL permet d'extraire rapidement les données nécessaires à chaque étape de construction des modèles d'apprentissage. Par exemple, dans [CFB97, CFB99], pour chaque nœud d'un arbre de décision, une requêtes SQL est formulée à la volée.

Selon cette approche, la technologie OLAP peut-être exploitée pour extraire des agrégats de données nécessaires à la recherche des règles d'association dans les cubes de données [GC98a, GC98b]. Les opérateurs OLAP peuvent aussi faire l'objet d'une extension à une fouille de données en ligne [Han97, Han98, HCC98]. De plus, avec ses capacités classiques d'exploration et de navigation, l'OLAP peut devenir un instrument utile pour la validation des connaissances extraites à partir des données multidimensionnelles [TNBP00, NNQ04, MJBN06, LZBX06].

Comme le résume le tableau comparatif 2.2, tous les travaux, qui abordent le problème du couplage selon cette approche, se limitent à des techniques d'*explication* et de *prédiction* telles que les *arbres de décision*, les *réseaux bayésiens*, les *motifs fréquents* ou les *règles d'association*.

Dans le cadre de cette approche instrumentale, nous associons l'analyse en ligne à une technique de *structuration* et de *classification*. Nous utilisons la classification ascendante hiérarchique (CAH) pour améliorer la qualité d'agrégation dans les cubes de données. Notre proposition exploite des opérateurs OLAP d'exploration, tels que le forage vers le haut (*roll-up*) et le forage vers le bas (*drill-down*), en vue d'extraite les individus et les variables nécessaires à la classification. Nous classifions particulièrement les modalités d'une dimension d'un cube selon leurs ressemblances. Nous agrégeons ensuite les faits du cube selon les classes de modalités obtenues. Ainsi nous sommes capables de fournir des agrégats de données sémantiquement plus riche que les agrégats classiques de l'OLAP [MRBB04, MBR04]. Nous proposons également une évaluation de la séparabilité des classes fournies par les partitions de la CAH afin d'assister l'utilisateur dans le choix du meilleur nombre d'agrégats [MBR06a].

Quant à la troisième approche, elle se base sur l'adaptation des algorithmes de fouille aux données multidimensionnelles. Bien que récente et ayant peu d'applications concrètes, cette approche est aussi intéressante car elle permet d'extraire des connaissances directement à partir des cubes de données, ce qui permet de prendre en compte l'aspect multidimensionnel et hiérarchique des données dans la construction d'un modèle d'apprentissage. Nous pensons que, dans l'avenir, cette approche est capable de créer une nouvelle génération de techniques de *fouille de données multidimensionnelles*.

Dans le cadre de cette approche, comme le résume le tableau comparatif 2.3, à l'image des propositions purement théoriques de Palpanas [Pal00] et de Parsaye [Par97], il n'y a pas beaucoup de travaux qui ont concrétisé cet aspect de couplage. Nous considérons que les cubegrades de Imieliński *et al.* [IKA02], les

Adaptation des techniques de fouille de données			
	Type de techniques de fouille		
Proposition	Visualisation et description	Structuration et classification	Explication et prédiction
Palpanas			
Parsaye			
Imieliński *et al.*			Cubegrades
Dong *et al.*			Constrained gradients
Sarawagi *et al.*			Modèle log-linéaire
Favero et Robin		Analyses quantitatives	
Notre proposition			**Règles d'association**

TAB. 2.3 – Comparaison des propositions de couplage de l'OLAP et de la fouille de données selon la troisième approche

constrained gradients de Dong *et al.* [DHL+01] et l'opérateur iDiff de Sarawagi [Sar99, Sar01] sont les seules propositions qui tentent véritablement d'adapter la fouille aux données multidimensionnelles.

Dans le cadre de cette troisième approche, nous utilisons une méthode d'*explication* dans les cubes de données. Notre proposition consiste à adapter la recherche des règles d'association aux données multidimensionnelles. Ainsi, nous mettons en place un nouvel algorithme, de type Apriori, capable d'extraire des règles d'association directement à partir d'une structure multidimensionnelle sans avoir recours à une transformation tabulaire des données initiales. Notre algorithme repose sur une fouille de données pilotée par les besoins de l'utilisateur via la définition d'une méta-règle [MRBM06]. Il se base également sur une nouvelle définition du support et de la confiance des règles d'association adaptée au contexte de l'analyse en ligne [MBR06c]. De plus, nous proposons une visualisation graphique, basée sur la *sémiologie graphique*, afin de valoriser les connaissances véhiculées par les règles extraites.

Réorganisation des cubes de données par une approche factorielle

Résumé

Dans ce chapitre, nous exposons notre première proposition traitant du couplage entre l'analyse en ligne et la fouille de données. Elle s'inscrit dans le premier groupe des approches de couplage qui consiste à adapter les données multidimensionnelles pour les techniques de fouille.

Notre proposition permet d'apporter une solution au problème de la visualisation des données engendré par l'éparsité des données. En se basant sur les résultats d'une analyse des correspondances multiples (ACM), nous tentons d'atténuer l'effet négatif de l'éparsité en réorganisant différemment les cellules d'un cube de données. À travers ce couplage entre l'OLAP et l'ACM, nous construisons un espace de représentation se prêtant mieux à l'analyse et dans lequel les faits du cube sont regroupés le mieux possible.

Sommaire

3.1 Introduction 33
3.2 Objectifs et motivations 36
3.3 Représentation des données multidimensionnelles 38
3.4 Notations générales 40
3.5 Définitions et formalisation 42
3.6 Arrangement des modalités 47
3.7 Évaluation de la qualité de représentation multidimensionnelle 49
3.8 Études de cas 53
3.9 Expérimentations et performances 60
3.10 Conclusion et perspectives 63

Publications

[MAF05] Messaoud R.B., Aouiche K., Favre C., « Une approche de construction d'espaces de représentation multidimensionnels dédiés à la visualisation », in $1^{ère}$ *journée francophone sur les Entrepôts de Données et l'Analyse en ligne (EDA'2005)*, *Revue des Nouvelles Technologies de l'Information*, pp. 34–50, Lyon, France : Cépaduès Editions. Juin 2005.

[MBR05] Messaoud R.B., Boussaid O., Rabaséda S.L., « 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.

[MBR06b] Messaoud R.B., Boussaid O., Rabaséda S.L., « Efficient Multidimensional 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, PA, USA : ACM Press. August 2006.

[MBR06d] Messaoud R.B., Boussaid O., Rabaséda S.L., « 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.

[MRB05] Messaoud R.B., Rabaséda S., Boussaid O., « L'analyse factorielle pour la construction de cubes de données complexes », in $2^{ème}$ *atelier Fouille de Données Complexes (FDC'2005)*, pp. 53–56, Paris, France. Janvier 2005.