OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données

Riadh Ben Messaoud, Sabine Loudcher Rabaseda, Omar Boussaïd, Fadila Bentayeb

OpAC : Opérateur d'analyse en ligne basé sur une technique de fouille de données

Riadh Ben Messaoud*, Sabine Rabaseda**
Omar Boussaid**, Fadila Bentayeb*

ERIC – Université Lumière Lyon 2
5, avenue Pierre Mendès-France – 69676 Bron Cedex – France
http ://eric.univ-lyon2.fr

*{rbenmessoaud | bentayeb}@eric.univ-lyon2.fr
**{sabine.rabaseda | boussaid}@univ-lyon2.fr

Résumé. L'analyse en ligne OLAP (*On-line Analysis Processing*) et la fouille de données (*Data Mining*) sont deux champs de recherche qui ont connu, depuis quelques années, des évolutions parallèles et indépendantes. De récentes études ont montré l'importance et l'intérêt de l'association entre ces deux domaines scientifiques. A l'heure actuelle, on assiste à l'accroissement du besoin d'une analyse en ligne plus élaborée. Nous pensons que le couplage entre OLAP et la fouille de données pourra apporter des réponses à ce besoin. Dans cet article, nous proposons d'adopter ce couplage en vue de créer un nouvel opérateur, baptisé *OpAC* (*Opérateur d'Agrégation par Classification*), d'analyse en ligne des données multidimensionnelles. *OpAC* consiste particulièrement en l'agrégation sémantique des modalités d'une dimension d'un cube de données en se basant sur la technique de la classification ascendante hiérarchique.

Mots-Clés : Couplage, Analyse en ligne, Cubes de données, Fouille de données, Agrégation sémantique, Opérateur d'analyse, Classification ascendante hiérarchique.

1 Introduction

La gestion des grandes masses de données est devenue une tâche difficile et assez coûteuse à maintenir. Les entrepôts de données (*Data Warehouses*) ont apporté des solutions efficaces à ce problème [Inmon, 1996] [Kimball, 1996]. En effet, un entrepôt de données représente une structure informatique centralisée dans laquelle est emmagasiné un volume important de données historisées, organisées par sujets et consolidées à partir de diverses sources d'informations. En plus de sa vocation de stockage, un entrepôt vise aussi l'exploitation des données dans un processus d'analyse en ligne et d'aide à la décision. Des modèles particuliers, tels que le schéma en étoile ou le schéma en flocon de neige, ont été conçus afin de rendre les données d'un entrepôt prêtes à l'analyse. Ces modèles permettent de créer des vues multidimensionnelles des données appelées aussi cubes de données dont la vocation est l'analyse en ligne OLAP (*On-Line Analysis Processing*) [Chaudhuri et Dayal, 1997]. Ainsi, la grande capacité de stockage, la

structuration multidimensionnelle et les opérateurs OLAP font de l'entrepôt une plate-forme décisionnelle servant à l'analyse, la visualisation et la navigation dans les grandes masses de données.

D'autre part, la fouille de données (*Data Mining*) fait appel à des méthodes d'apprentissage pour induire des modèles de connaissances exprimés dans des formalismes valides et compréhensibles. Cependant, la fouille est un maillon qui ne peut être dissocié de la chaîne de l'extraction de connaissances à partir des données : *ECD* (*KDD : Knowledge Discovery in Databases*). La construction de tout modèle d'apprentissage demande un prétraitement et de mise en forme des données. En effet, les méthodes d'apprentissage automatique sont sensibles au bruit et ne peuvent opérer que sur des données obéissant à la forme classique des tableaux *"Individus-Variables"*. La prise en compte de ces propriétés donne à l'extraction des connaissances l'aspect d'une démarche décisionnelle robuste mais lourde dans son déploiement [Fayyad *et al.*, 1996].

L'analyse en ligne et la fouille de données sont considérées comme deux champs de recherche séparés qui ont connu parallèlement des évolutions indépendantes. A l'heure actuelle, nous pensons qu'un couplage entre les deux approches pourrait permettre une analyse en ligne plus élaborée dépassant la simple exploration des cubes de données. La robustesse de la fouille de données et la maniabilité de la structuration multidimensionnelle peuvent apporter des améliorations aux capacités de l'OLAP afin d'extraire des connaissances à partir des cubes de données. Notre but est d'enrichir l'analyse multidimensionnelle par une nouvelle forme d'agrégation élaborée basée sur une technique de fouille de données. Prenant en compte la structure multidimensionnelle des données et le besoin de les intégrer dans un processus d'analyse plus poussé, notre démarche consiste à développer un nouvel opérateur d'agrégation OLAP, baptisé *OpAC* (*Opérateur d'Agrégation par Classification*), basé sur la méthode *CAH* (Classification Ascendante Hiérarchique) [Lance et Williams, 1967].

L'article est organisé de la façon suivante. Dans la section 2, nous exposons un état de l'art des approches de couplage entre la fouille de données et l'analyse en ligne. La section 3 présente les objectifs de l'opérateur *OpAC*. Nous justifions, dans la section 4, le choix de la *CAH* comme technique d'analyse en ligne. Nous développons, dans la section 5, une formalisation décrivant les concepts théoriques de l'opérateur, notamment sa prise en charge des données multidimensionnelles. La section 6 est consacrée à l'implémentation d'un prototype qui valide notre approche. Finalement, la section 7 conclut l'article et évoque les évolutions possibles de ce travail.

2 Les approches de couplage entre la fouille de données et l'analyse en ligne

Peu de travaux de recherche traitent le couplage entre la fouille de données et l'analyse en ligne. Ceci s'explique par la priorité accordée à l'amélioration des techniques des deux domaines séparément. Cependant, trois grandes approches existent :

2.1 Extension des opérateurs OLAP

Il s'agit d'une approche instrumentale principalement due aux travaux de Han, auteur du système *DBMiner* [Han, 1997]. Ces travaux consistent à étendre le langage des requêtes OLAP pour simuler des techniques de fouille telles que la détection des règles d'association, la caractérisation d'attributs, la classification, la prédiction, etc. Cependant les références relatives à *DBMiner* [Han, 1997] [Han, 1998] décrivent plutôt le côté fonctionnel de ce dernier et ne donnent pas assez d'éclairages sur les procédés employés. Dans [Han *et al.*, 1998] on évoque la terminologie de la *"fouille en ligne"* (*OLAM : On-Line Analytical Mining*) désignant le mécanisme où les techniques d'apprentissage sont utilisées dans l'analyse en ligne. Chen, Dayal et al. proposent un prototype de suivi des habitudes de consommation sur le web. Ce prototype utilise les opérateurs OLAP pour générer des règles d'association à partir de plusieurs serveurs OLAP distribués [Chen *et al.*, 1999] [Chen *et al.*, 2000]. Les résultats de ces règles sont stockés dans des cubes spécifiques (*Association rule cubes*). Goil et Choudhary suggèrent l'extraction des connaissances à partir d'un cube de données en exploitant les fonctionnalités des opérateurs OLAP [Goil et Choudhary, 2001]. Leur approche s'est focalisée sur la détection des règles d'association à différents niveaux d'agrégation des dimensions du cube [Goil et Choudhary, 1998].

2.2 Adaptation des structures multidimensionnelles

Cette approche vise l'adaptation des données multidimensionnelles afin de les rendre intelligibles par les méthodes de fouille. Deux stratégies sont proposées :

La première consiste à utiliser les avantages des SGBDM (systèmes de gestion des bases de données multidimensionnelles) pour aider l'algorithme d'apprentissage pendant la construction de son modèle de connaissances. Laurent et al. proposent une coopération entre un SGBDM (*Oracle Express*), doté d'une capacité de calcul des agrégats complexes, avec un logiciel d'apprentissage (*Salammbô*) qui construit des arbres de décision flous [Laurent *et al.*, 2000] [Laurent, 2001]. Cette stratégie permet de transférer la gestion de la base d'apprentissage, les contraintes de stockage et de manipulation des données dans le SGBDM.

La deuxième stratégie consiste plutôt à transformer les données multidimensionnelles, en dehors du SGBDM, pour les adapter aux techniques de fouille. Pinto, Han et al. [Pinto *et al.*, 2001] proposent d'intégrer des informations multidimensionnelles dans les séquences de données et d'identifier, dans les nouvelles séquences, les motifs fréquents. En vue d'appliquer des arbres de décision sur un cube, Goil et Choudhary "aplatissent" les données du cube pour extraire des matrices de contingence pour chaque dimension et à chaque étape de construction de l'arbre [Goil et Choudhary, 2001]. Chen et al. proposent d'intégrer OLAP en tant qu'étape de prétraitement du processus d'ECD [Chen *et al.*, 2001]. Les données nettoyées et formatées peuvent, par la suite, être exploitées par les techniques de fouille de données.

2.3 Adaptation des algorithmes de fouille de données

La troisième approche consiste à modifier les algorithmes de fouille de données et à les utiliser directement dans l'environnement multidimensionnel. Palpanas pense

que cette approche serait tributaire d'un processus d'analyse élaboré apte à produire de nouvelles formes de connaissance plus intéressantes [Palpanas, 2000]. Sarawagi et al. proposent un outil d'identification des régions remarquables dans un cube de données. Leur idée consiste à intégrer un module statistique de régression multidimensionnelle (*Discovery-driven*) dans un serveur OLAP en vue de guider l'utilisateur pour détecter les valeurs remarquables à différents niveaux hiérarchiques d'un cube [Sarawgi *et al.*, 2001]. Dans [Sarawgi, 1999] [Sarawgi, 2001], l'auteur propose un nouvel outil *iDiff*, basé sur la programmation dynamique, capable de détecter aussi bien les valeurs remarquables que les raisons de leur présence dans un cube. Des travaux similaires, ont été réalisés par [Favero et Robin, 2001] pour la génération du langage naturel à partir des données multidimensionnelles.

A la lumière de cet état de l'art, nous constatons qu'aucune des approches précédentes n'a employé le couplage entre la fouille de données et l'analyse en ligne en vue d'étendre les fonctionnalités d'OLAP actuellement disponibles. Nous pensons que la structure multidimensionnelle des données peut apporter un contexte d'analyse ciblé pour la fouille de données. Ceci nous incite à définir une nouvelle génération d'opérateurs d'analyses en ligne basés sur des techniques de fouille. Notre approche associe l'aspect exploratoire d'OLAP à la démarche descriptive et prédictive de la fouille. Nous présentons, dans le cadre de cet article, un nouvel opérateur basé sur cette optique et permettant d'effectuer des analyses élaborées.

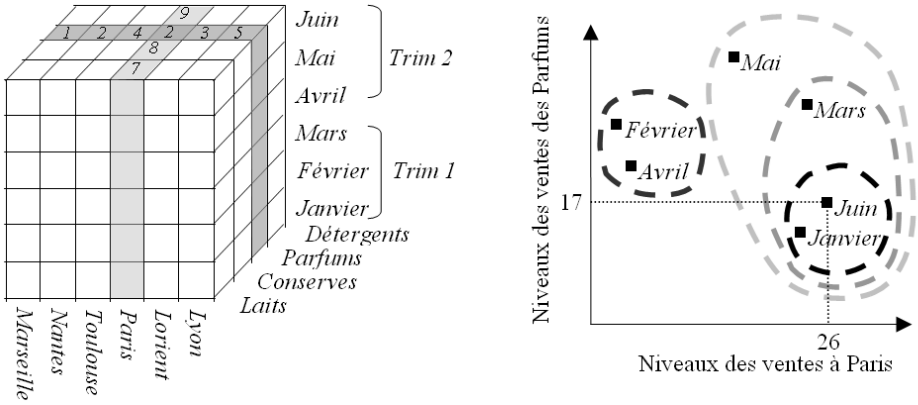
3 Objectifs de l'opérateur *OpAC*

La construction d'un cube de données cible un problème d'analyse précis. Le choix des dimensions et des mesures dépend des besoins de l'analyse. D'une manière générale, une dimension est organisée sur plusieurs hiérarchies traduisant différents niveaux de granularité. Chaque hiérarchie comporte un ensemble de modalités, et chaque modalité d'une hiérarchie agrège des modalités de l'hiérarchie immédiatement inférieure selon un ordre d'appartenance logique. Par exemple, une dimension temporelle peut être structurée en quatre niveaux hiérarchiques : *jours*, *mois*, *trimestres* et *années*.

Toutefois, la granularité d'une dimension est fortement dépendante du niveau de précision exigé par l'analyse. Par exemple, si l'analyse exclut les mesures quotidiennes, on peut limiter la dimension temporelle aux niveaux : *mois*, *trimestres* et *années*. En revanche, l'organisation des modalités d'une dimension est toujours régie par un ordre d'appartenance logique dicté par l'usage naturel des objets ou des concepts du monde réel. Par exemple, il est naturel de dire que la modalité *"1er Trimestre"* de la dimension temporelle contient les mois *"Janvier"*, *"Février"* et *"Mars"*.

Le cube de la Figure 1(a) est constitué de trois dimensions : *Localité géographique*, *Temps* et *Produit*. La dimension temporelle est organisée selon deux niveaux hiérarchiques : celui des *mois* et celui des *trimestres*.

L'idée de base de l'opérateur *OpAC* consiste à exploiter les mesures contenues dans un cube de données afin d'agréger les modalités d'une de ses dimensions. Si on veut agir sur la dimension *Temps*, les mois sont vus comme des individus qu'on peut décrire par



(a) *Contexte OLAP* (b) *Contexte classification*

FIG. 1 – *Principe de l'opérateur d'agrégation OpAC*

des mesures significatives provenant du cube. Comme le montre la Figure 1(b), on peut considérer *"Les ventes des Parfums"* et *"Les ventes à Paris"* comme des descripteurs des individus. Par exemple, d'après le cube de la Figure 1(a), le mois de *"Juin"* est caractérisé par 17 unités de ventes de *Parfums* et 26 unités de ventes à *Paris*. En adoptons une technique de classification, on agrège les mois les plus proches au sens des deux descripteurs. cités ci-dessus.

Contrairement à l'agrégation au sens OLAP classique, basée sur le sens de l'appartenance logique des modalités, notre approche constitue une nouvelle forme d'agrégation *sémantique* qui tient compte des faits réels contenus dans un cube de données. Le but de l'opérateur *OpAC* est de pouvoir agréger les modalités selon leurs liens sémantiques et pas selon leurs liens logiques. Par exemple, dans la Figure 1(a), les mois de *"Janvier"*, *"Février"* et *"Mars"* forment un agrégat puisqu'ils appartiennent tous au premier trimestre de l'année. Alors que, dans la Figure 1(b), l'agrégation sémantique nous permet de constater que *"Janvier"* et *"Juin"* forment un agrégat plus significatif du point de vue de l'utilisateur puisqu'ils représentent des périodes particulières (niveaux de ventes semblables) concernant les ventes de *Parfums* à *Paris*.

4 Le choix de la classification ascendante hiérarchique

Contrairement aux modalités d'une dimension, qui sont organisées selon un ordre prédéfini, *OpAC* fournit des agrégats mettant en évidence les liens sémantiques entre les faits contenus dans les données. Cette forme d'agrégation permet de véhiculer des informations plus riches que celles fournies par l'agrégation classique d'OLAP. Prenant en compte ces objectifs, notre choix s'est porté sur la classification ascendante hiérarchique (*CAH*). Nous justifions ce choix par les points suivants :

1. Nous constatons l'existence d'une forte analogie entre les résultats de la *CAH* et la structuration d'une dimension d'un cube de données. De plus, les objectifs énoncés

pour notre opérateur correspondent bien à la stratégie de la *CAH*. Cette analogie est due, en grande partie, à l'aspect hiérarchique qui constitue un lien pertinent entre la *CAH*, la hiérarchie des dimensions d'un cube ainsi que la représentation des résultats de l'opérateur.

2. Contrairement à la *CDH* (Classification Descendante Hiérarchique), la *CAH* adopte une stratégie agglomérative partant de la partition la plus fine où chaque individu est vu comme une classe. Cette propriété permettra à *OpAC* d'inclure, dans ses résultats, les modalités les plus fines de la classification. De plus, la stratégie ascendante est plus rapide que la stratégie descendante. La complexité de la *CAH* est généralement polynomiale, tandis que celle de la *CDH* est exponentielle [Chavent *et al.*, 1999]. 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, tant dis qu'un algorithme descendant basé sur l'énumération complète évalue toutes les divisions des n individus en deux sous-ensembles non vides, soit $2^{n-1}-1$ possibilités.
3. Les résultats de la *CAH* sont compatibles avec l'esprit de l'analyse en ligne et peuvent être réutilisés par les opérateurs de navigation classiques d'OLAP. La *CAH* fournit plusieurs partitions d'individus où chaque partition correspond à un niveau hiérarchique. En passant d'un niveau de partitions au niveau qui lui est immédiatement supérieur, deux classes sont agrégées pour former un nouvel agrégat. Inversement, en passant d'un niveau de partition à un autre qui lui est immédiatement inférieur, un agrégat est divisé en deux classes. Ceci assure à *OpAC* un aspect exploratoire comparable à celui des opérateurs classiques *Roll-up* et *Drill-down*.

5 Formalisation de l'opérateur *OpAC*

Nous proposons un cadre formel pour la définition des individus et des variables de la classification à partir d'un cube de données. Des contraintes sont imposées afin d'assurer la validité statistique et logique des données extraites.

Notons Ω l'ensemble des individus et Σ l'ensemble des variables de la classification. Soit un cube de données $\mathcal{C}$ ayant d dimensions et m mesures. On note $D_1, \ldots, D_i, \ldots, D_d$ les dimensions de $\mathcal{C}$ et $M_1, \ldots, M_q, \ldots, M_m$ ses mesures.

Supposons que :
- Pour tout $i \in [1, d]$ la dimension D_i comprend n_i niveaux hiérarchiques. Notons h_{ij} le niveau hiérarchique j de D_i, avec $j \in [1, n_i]$;
- Pour tout $j \in [1, n_i]$ le niveau hiérarchique h_{ij} comprend l_{ij} modalités. Notons g_{ijt} la modalité t de h_{ij}, avec $t \in [1, l_{ij}]$;
- $\mathcal{G}(h_{ij})$ l'ensemble des modalités de h_{ij}.

Supposons que nous cherchons à agir sur la hiérarchie h_{ij}[1]. Statistiquement parlant,

[1] Le choix de h_{ij} dépend des besoins de l'analyse et des objectifs de l'utilisation de l'opérateur d'agrégation.

$\mathcal{G}(h_{ij})$ représente la population des individus du problème de la classification. Soit :

$$\Omega = \mathcal{G}(h_{ij}) = \{g_{ij1}, \ldots, g_{ijt}, \ldots, g_{ijl_{ij}}\}$$

Adoptons, à présent, les notations suivantes :
- $*$ un méta-symbole désignant l'agrégat total d'une dimension ;
- $\mathcal{G}$ l'ensemble des n-uplets des modalités des hiérarchies du cube $\mathcal{C}$ y compris les agrégats totaux des dimensions.

$$\mathcal{G} = \prod_{i=1}^{d} (\underbrace{\mathcal{G}(h_{ij})}_{j\in[1,n_i]} \cup \{*\}) = (\underbrace{\mathcal{G}(h_{1j})}_{j\in[1,n_1]} \cup \{*\}) \times \ldots \times (\underbrace{\mathcal{G}(h_{ij})}_{j\in[1,n_i]} \cup \{*\}) \times \ldots \times (\underbrace{\mathcal{G}(h_{dj})}_{j\in[1,n_d]} \cup \{*\})$$

On définit, maintenant, $\forall q \in [1, m]$ la mesure M_q en tant qu'une fonction de l'ensemble $\mathcal{G}$ dans l'ensemble des réels $\Re$.

$$M_q : \mathcal{G} \longrightarrow \Re$$

Reprenons l'exemple du cube de la Figure 1(a) composé de trois dimensions D_1 (la dimension temporelle), D_2 (la dimension géographique), D_3 (la dimension des produits) et d'une mesure (les niveaux de ventes d'une chaîne de magasins).

Dans ce cas :
- M_1(*Février 1999, Lyon,* $*$) désigne la mesure du niveau des ventes de tous les produits au mois de *Février* de l'année 1999 dans la ville de *Lyon*;
- M_1(*Février 1999,* $*$ *, Produits laitiers*) désigne la mesure du niveau des ventes des *Produits laitiers* dans toutes les localités géographiques au mois de *Février* de l'année 1999.

Rappelons que l'objectif de l'opérateur *OpAC* est d'établir une agrégation sémantique qui tient compte de la signification de l'information contenue dans les données d'un cube. Pour cela, nous considérons les mesures du cube comme des variables quantitatives décrivant la population $\mathcal{G}(h_{ij})$. Il faut, tout de même, respecter certaines contraintes logiques et statistiques fondamentales dans le choix de ces variables :

Première contrainte : Aucun niveau hiérarchique de la dimension retenue pour les individus ne doit être générateur des variables de la classification. En effet, décrire un individu par une variable exprimant une propriété qui le contient, ou qui l'agrège, n'aura aucun sens logique. Il serait insensé de vouloir décrire, par exemple, l'année *1999* par le niveau des ventes du mois de *janvier 1999* ou le niveau des ventes en *France* par celui de la ville de *Lyon*. Inversement, une variable qui spécifie des propriétés d'appartenance à un individu ne peut servir que pour la description de cet individu particulier. Par exemple, le niveau des ventes du mois de *janvier 1999* ne peut servir de descripteur que pour l'année *1999* et sera inutilisé pour la description des niveaux de ventes des autres années.

Seconde contrainte : Par dimension, on ne peut choisir qu'un seul niveau hiérarchique pour générer les variables. Cette contrainte est essentielle pour assurer l'indépendance des variables de la classification. En effet, la valeur d'une modalité peut s'obtenir par combinaison linéaire des valeurs des modalités qui lui appartiennent dans

la hiérarchie inférieure. Par exemple, la somme des valeurs des ventes pour chaque mois d'une année correspond bien à la valeur totale des ventes de l'année en question.

En conclusion, et en supposons que $\Omega = \mathcal{G}(h_{ij})$, les variables de la classification de l'opérateur appartiennent à l'ensemble suivant :

$$\Sigma \subset \left\{ \begin{array}{l} X/\forall t \in [1, l_{ij}],\ \underbrace{X(g_{ijt})}_{j\in[1,n_j]} = M_q(*,...,*,\ \underbrace{g_{ijt}}_{j\in[1,n_j]},*,...,*,\ \underbrace{g_{srv}}_{r\in[1,n_s]},*,...,*) \\ \text{avec } s \neq i,\ r \in [1, n_s] \text{ est unique pour chaque } s,\ v \in [1, l_{sr}] \text{ et } q \in [1, m] \end{array} \right\}$$

Le choix des variables dans cet ensemble dépend de la nature et des objectifs de l'analyse qu'on veut mener.

Pour mieux comprendre cette formalisation, revenons à l'exemple du cube de la Figure 1(a). Supposons que, pour des choix d'analyse, on souhaite classer les mois de l'année selon les niveaux des ventes par régions et/ou par produits. Dans ce cas, on retient les modalités du niveau des mois de la dimension D_1 comme individus statistiques. On aura donc :

$$\Omega = \{\ \textit{Janvier, Février, Mars, Avril, Mai, Juin}\ \}$$

En respectant la première contrainte suscitée, on ne peut plus réutiliser la dimension D_1 pour la génération des variables. De plus, et 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. Si, par exemple, on choisit le niveau des villes de la dimension D_2 pour générer les variables, on fait des agrégations totales (*Roll-up*) sur toutes les autres dimensions du cube outre la dimension D_1, retenue pour les individus, et D_2 retenue pour les variables. Dans cet exemple, on fait une agrégation totale sur D_3 (Figure 2). On obtient, un tableau de contingence exprimant les valeurs des ventes pour les modalités de D_1 croisées avec celles de D_2, c'est-à-dire les valeurs des ventes par ville pour chaque mois. De la même manière, on peut générer des variables à partir de D_3 en faisant une agrégation totale sur D_2.

Comme le montre la Figure 2, {*"Le niveau des ventes à Marseille"*, *"Le niveau des ventes à Nantes"*, *"Le niveau des ventes à Toulouse"*, *"Le niveau des ventes à Paris"*, *"Le niveau des ventes à Lorient"* et *"Le niveau des ventes à Lyon"*}, est un ensemble de variables possibles pour le problème de classification.

6 Implémentation

Pour valider notre approche, nous proposons un prototype[2] pour l'opérateur *OpAC*. L'implémentation a été réalisée en *Visual Basic* sous l'environnement *Windows XP Professional*. L'installation du serveur OLAP (*OLAP Services*) de *MS SQL Server* et du pilote *MSOLAP* est nécessaire pour l'exécution du prototype.

Le prototype de l'opérateur *OpAC* est conçu selon une architecture trois tiers : une interface de paramétrage, un module de chargement de données et un constructeur du modèle de classification.

[2] http ://bdd.univ-lyon2.fr/download/opac.zip

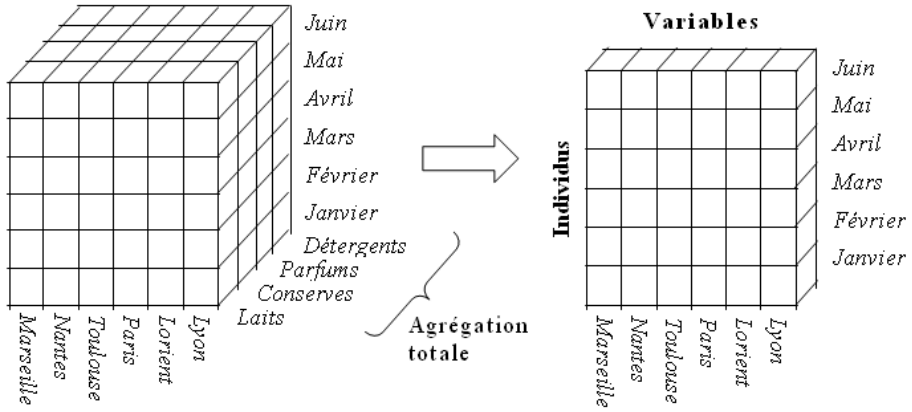


FIG. 2 – *Domaines des individus et des variables pour l'opérateur OpAC.*

L'interface de paramétrage permet le choix, de façon assisté, des dimensions et des mesures qui correspondent aux individus et aux variables de la classification. Cette assistance prend en compte les contraintes précisées dans la formalisation théorique. L'interface permet, également, le choix des paramètres de la *CAH* (la mesure de dissimilarité et le critère d'agrégation des classes).

Le chargeur des données assure trois tâches : il établit une connexion à un cube de données via le serveur OLAP externe ; il importe les informations qui décrivent la structure du cube (les noms des dimensions, des hiérarchies et des faits) ; il charge et met en forme les données à analyser dans un tableau dynamique *"Individus-Variables"*.

Le constructeur du modèle de classification assure non seulement la construction du modèle mais présente aussi les résultats de la classification à l'utilisateur sous forme de dendrogramme.

Le dendrogramme est accompagné par un résumé des données de l'analyse (les dimensions et les mesures, le nombre d'individus, le nombre de variables, etc.), ainsi qu'un rappel des paramètres du modèle de classification (mesure de dissimilarité et critère d'agrégation). Cependant, l'affichage et l'interprétation d'un dendrogramme deviennent de plus en plus difficiles avec l'augmentation du nombre des individus. Afin de contourner ce problème et d'assurer une présentation intelligible et interactive de l'information à visualiser, comme le montre la Figure 3, nous avons construit un outil visuel permettant à l'utilisateur de couper le dendrogramme à différents niveaux hiérarchiques. Cet outil permet, également, de réduire et d'agrandir la taille du dendrogramme par une fonction de mise en échelle. De plus, ce prototype apporte une assistance à l'utilisateur pour décider du nombre de classes, qui correspond au mieux à ses besoins d'analyse, par l'intermédiaire d'un outil d'évaluation de la qualité des partitions.

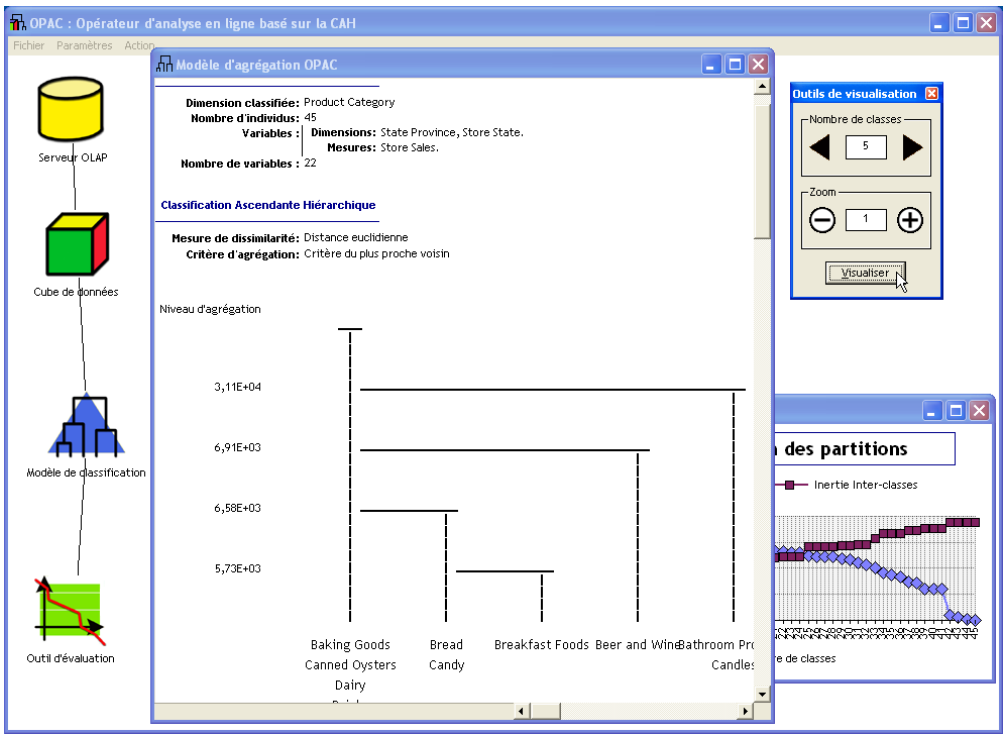


FIG. 3 – *Le prototype de l'opérateur OpAC.*

7 Conclusion

L'objectif de notre travail était de réaliser un couplage entre la fouille de données et la technologie OLAP pour répondre à des besoins d'analyse en ligne élaborée. Notre idée était d'intégrer une technique de fouille, la *CAH*, dans la structure multidimensionnelle des données et de créer un nouvel opérateur d'agrégation en ligne.

Notre démarche a commencé par la spécification des objectifs visés par l'opérateur *OpAC* et le choix d'une technique de classification adéquate. Partant de l'analogie existante entre les hiérarchies des dimensions d'un cube de données et celles définies pour la *CAH*, le choix de cette dernière nous parut évident. Une formalisation théorique de l'opérateur a été proposée pour définir les individus et les variables de la classification. *OpAC* se distingue par sa capacité d'agréger les modalités d'une dimension d'un cube selon leurs liens sémantiques. Les opérateurs OLAP classiques agrègent les modalités d'une dimension selon des liens logiques prédéfinis. En revanche, *OpAC* permet de créer de nouveaux agrégats qui reflètent plutôt des faits réels contenus dans le cube. Afin de valider notre démarche, nous avons implémenté un prototype doté d'un aspect visuel et interactif permettant une navigation et une synthèse des données conforme à l'analyse en ligne. Notre opérateur constitue une voie possible pour une analyse en ligne plus élaborée que celle des opérateurs existants. Le choix de la classification ascendante

hiérarchique n'exclut nullement l'utilisation d'autres méthodes de classification. Nous pensons que l'utilisation d'autres techniques de fouille de données permettra d'établir de nouveaux modèles d'apprentissage en ligne sur les données multidimensionnelles.

Finalement, l'opérateur, dans sa version actuelle, est sujet à plusieurs perfectionnements possibles [Messaoud, 2003]. Nous projetons l'amélioration de l'outil d'évaluation pour mieux apprécier la qualité des agrégats générés par l'opérateur. Nous envisageons également apporter plus de polyvalence à notre opérateur afin qu'il puisse traiter aussi bien les cubes de données numériques que les cubes de données complexes telles que l'image, le texte et le son.

Références

[Chaudhuri et Dayal, 1997] S. Chaudhuri et U. Dayal. An overview of data warehousing and olap technology. *SIGMOD Record*, 26 :65–74, 1997.

[Chavent *et al.*, 1999] M. Chavent, C. Guinot, Y. Lechevallier, et M. Tenenhaus. Méthodes divisives de clustering et segmentation non supervisée : recherche d'une typologie de la peau humaine saine. *Revue de Statistique Appliquée*, XLVII :87–99, 1999.

[Chen *et al.*, 1999] Q. Chen, U. Dayal, et M. Hsu. A distributed olap infrastructure for e-commerce. In *Fourth IFCIS Conference on Cooperative Information Systems (CoopIS'99)*, Edinburgh, Scotland, Semptember 1999.

[Chen *et al.*, 2000] Q. Chen, U. Dayal, et M. Hsu. An olap-based scalable web access analysis engine. In *2nd International Conference on Data Warehousing and Knowledge Discovery (DAWAK'2000)*, London, UK, September 2000.

[Chen *et al.*, 2001] M. Chen, Q. Zhu, et Z. Chen. An integrated interactive environment for knowledge discovery from heterogeneous data resources. *Information and Software Technology*, 43 :487–496, 2001.

[Favero et Robin, 2001] E.L Favero et J. Robin. Using olap and data mining for content planning in natural language generation. *Lecture Notes in Computer Science*, 1959 :164–175, 2001.

[Fayyad *et al.*, 1996] U.M. Fayyad, G.P. Shapiro, P. Smyth, et R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, 1996.

[Goil et Choudhary, 1998] S. Goil et A. Choudhary. High performance multidimensional analysis and data mining. In *High Performance Networking and Computing Conference (SC'98)*, Orlando, USA, November 1998.

[Goil et Choudhary, 2001] S. Goil et A. Choudhary. Parsimony : An infrastructure for parallel multidimensional analysis and data mining. *Journal of parallel and distributed computing*, 61 :285–321, 2001.

[Han *et al.*, 1998] J. Han, S. Chee, et J.Y. Chiang. Issues for on-line analytical mining of data warehouses. In *SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'98)*, Seattle, Washington, June 1998.

[Han, 1997] J. Han. Olap mining : An integration of olap with data mining. In *The IFIP Conference on Data Semantics*, Leysin, Switzerland, October 1997.

[Han, 1998] J. Han. Toward on-line analytical mining in large databases. *SIGMOD Record*, 27 :97–107, 1998.

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

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

[Lance et Williams, 1967] G.N. Lance et W.T. Williams. A general theory of clustering sorting strategies : Clustering systems. *The Computer Journal*, pages 271–277, 1967.

[Laurent *et al.*, 2000] A. Laurent, S. Gançarski, et C. Marsala. Coopération entre un système d'extraction de connaissances floues et un système de gestion de bases de données multidimensionnelles. In Cepaduès editions La Rochelle, editor, *Rencontres Francophones sur la Logique Floues et ses Applications*, pages 325–332, Octobre 2000.

[Laurent, 2001] A. Laurent. De l'olap mining au f-olap mining. *Revue Extraction des connaissances et apprentissage (ECA)*, 1 :189–200, 2001.

[Messaoud, 2003] R. Ben Messaoud. *Construction d'un opérateur d'analyse en ligne des données complexes basé sur une technique de fouille de données*. Mémoire de dea, Université Lumière Lyon 2, 2003.

[Palpanas, 2000] T. Palpanas. Knowledge discovery in data warehouses. *SIGMOD Record*, 29 :88–100, 2000.

[Pinto *et al.*, 2001] H. Pinto, J. Han, J. Pei, K. Wang, Q. Chen, et U. Dayal. Multi-dimensional sequential pattern mining. In *On Information and Knowledge Management (CIKM'01)*, Atlanta, USA, November 2001.

[Sarawgi *et al.*, 2001] S. Sarawgi, R. Agrawal, et N. Megiddo. Discovery-driven exploration of olap data cubes. In *The 6^{th} Int'l Conference on Extending Database Technology (EDBT)*, Valencia, Spain, Mars 2001.

[Sarawgi, 1999] S. Sarawgi. Explaining differences in multidimensional aggregates. In *The 25^{th} Int'l Conference on Very Large Databases (VLDB)*, Edinburgh, Scotland, September 1999.

[Sarawgi, 2001] S. Sarawgi. idiff : Informative summarization of differences in multidimensional aggregates. *Data Mining And Knowledge Discovery*, 5 :213–246, 2001.

Summary

For a few years, on-line analysis processing (OLAP) and data mining have known parallel and independent researches evolutions. Some recent studies have shown the interest of the association of these two fields. Currently, we attend the increase of a more elaborated analysis's need. We think that the idea of coupling OLAP and data mining will be able to fulfill this need. We propose to adopt this coupling in order to create a new operator, called *OpAC*, for multidimensional on-line analysis. The main idea of *OpAC* consists of using the agglomerative hierarchical clustering to achieve a semantic aggregation on the attributes of a data cube dimension.

Keywords : Coupling, On-line analysis processing, Data cubes, Data mining, Semantic aggregation, Analysis operator, Agglomerative hierarchical clustering.