Approximation des cubes OLAP et génération de règles dans les entrepôts de données

OLAP Cube Approximation and Rule Generation in Data Warehouses

Sami Naouali

LARIM, Université du Québec en Outaouais, Canada,

sami.naouali@uqo.ca.

Rokia Missaoui

LARIM, Université du Québec en Outaouais, Canada,

rokia.missaoui@uqo.ca.

Résumé

Cet article décrit une nouvelle approche d'approximation des résultats d'une requête OLAP soumise à un entrepôt de données. Cette approche est basée sur une adaptation de la théorie des ensembles approximatifs (rough set theory) aux données multidimensionnelles et offre de nouvelles possibilités d'exploration et d'extraction de connaissances à partir des cubes OLAP. L'objectif de cet article est d'intégrer des outils d'approximation dans les entrepôts de données dans le but de produire de nouvelles vues que l'on peut par la suite analyser et explorer en faisant appel à des opérateurs OLAP et/ou des algorithmes d'extraction de connaissances. Cette intégration permet alors à l'utilisateur de travailler soit en mode "strict" (ou restreint) en utilisant une approximation basse du cube OLAP, ou en mode "relâché" en utilisant une approximation haute de ce dernier. Le premier mode est utile dans le cas où la réponse à la requête est volumineuse, permettant ainsi à l'utilisateur de focaliser son attention sur un ensemble réduit de cellules fortement similaires. Le deuxième mode est utile dans le cas d'une requête retournant un ensemble réduit de cellules, permettant ainsi de relâcher les conditions de la requête afin d'élargir le volume du résultat.

Abstract

This paper presents a new approach toward approximate query answering in data warehouses. The approach is based on an adaptation of rough set theory to multidimensional data, and offers cube exploration and mining facilities. The objective of this work is to integrate approximation mechanisms and associated operators into data cubes in order to produce views that can then be explored using OLAP or data mining techniques. The integration of data approximation capabilities with OLAP techniques offers additional facilities for cube exploration and analysis. The proposed approach allows the user to work either in a restricted mode using a cube lower approximation or in a relaxed mode using cube upper approximation. The former mode is useful when the query output is large, and hence allows the user to focus on a reduced set of fully matching tuples. The latter is useful when a query returns an empty or small answer set, and hence helps relax the query conditions so that a superset of the answer is returned.

1. Introduction

Puisque les données d'un entrepôt proviennent de multiples sources de données hétérogènes et parfois peu fiables, les utilisateurs sont plus tolérants dans un environnement d'entrepôt de données et acceptent une certaine perte d'information et un léger écart par rapport aux données réelles. Dans le domaine des bases de données relationnelles, la problématique de traitement approximatif et flexible des requêtes a fait l'objet de plusieurs travaux (Andreasen, T., Motro, A., Christiansen, H., Larsen, H, 2002) ; (Chu et Chen, 1994) ; (Huh, Moon, Ahn et al. 2002) ; (Muslea, 2004). Ces travaux sont généralement axés sur la relaxation des conditions dans des requêtes retournant un résultat vide.

Dans un contexte d'entreposage de données multidimensionnelles, l'approximation des requêtes a été utilisée dans le but d'optimiser le calcul d'agrégation et par la suite de réduire le temps de réponse aux requêtes utilisateur au prix d'une certaine perte d'information. La plupart des travaux ont été réalisés en se basant sur des techniques d'échantillonnage (Babcock, Chauhuri et Das, 2003) , (Ganti, Lee et Ramakrishnan, 2000). Chakrabarti *et al.* (Chakrabarti, Garofalakis, et al., 2001) proposent une approche fondée sur le principe d'approximation par ondelettes (*wavelets*) et montrent qu'elle est plus efficace que l'échantillonnage. Dans ce même esprit, Fayyad *et al.* (Shanmugasundaram, Fayyad et Bradley, 1999) utilisent une distribution de la densité de probabilité des données pour proposer une représentation compacte des cubes de données réduisant ainsi leur espace de stockage et permettant une réponse approximative aux requêtes d'agrégation.

Les techniques basées sur le principe d'approximation par ondelettes ont été par la suite utilisées aussi bien pour des évaluations progressives de quelques requêtes OLAP assez spécifiques (Ambite, Shahabi, Schmidt et al., 2001) que pour faire de l'échantillonnage. Dans (Vitter et Wang, 1999). Ces techniques ont été utilisées dans le même objectif que (Shanmugasundaram, Fayyad et Bradley, 1999) pour générer une représentation compacte des cubes de données éparses et mieux gérer les requêtes d'agrégation de ces données.

Notre approche permet une approximation des requêtes OLAP de façon à retourner soit un sous-ensemble ou un sur-ensemble de l'ensemble exact des données répondant à la requête. Elle peut également être utilisée pour introduire une certaine tolérance (ex. vérification d'un sous-ensemble des conditions de la requête) lors de la manipulation des données bruitées pouvant être contenues dans les entrepôts.

Notre contribution consiste à offrir des mécanismes permettant une certaine flexibilité lors du traitement d'une requête ainsi qu'une exploration plus riche des cubes. Dans cet objectif, nous avons, primo, intégré les principes des ensembles approximatifs au contexte multidimensionnel des données dans le but de fournir des réponses approximatives aux requêtes et définir des concepts (partitions du cube en sous-ensembles de faits) en réponse aux requêtes des utilisateurs, secundo, proposé un enrichissement des techniques OLAP avec de nouveaux opérateurs apportant plus de flexibilité lors des interactions de l'utilisateur avec l'entrepôt de données, et tertio défini des vues matérialisées de données pour encapsuler et exploiter les réponses aux opérateurs d'approximation à des fins d'extraction de connaissances (segmentation de cubes et calcul des règles d'association). En conséquence,

l'approximation concerne non seulement la réponse aux requêtes mais également le résultat d'extraction de connaissances.

Cet article est organisé comme suit. Nous présentons dans la section 2, les notions de base nécessaires pour la présentation de l'approche d'approximation des cubes. Un exemple illustratif est présenté en section 3. L'approche proposée est par la suite présentée en détail dans la section 4.

2. Contexte

2.1 Travaux connexes et motivations

Dans la section précédente, nous avons présenté une brève description des travaux traitant de l'approximation des requêtes dans les bases et les entrepôts de données. Dans ce qui suit, nous présentons un bref aperçu des travaux proposant des extensions au processus OLAP pour une meilleure exploitation des données de l'entrepôt en découvrant des phénomènes et structures cachés dans ces bases multidimensionnelles. Nous supposons que le lecteur dispose des connaissances de base sur les entrepôts de données. Des informations utiles sur le sujet se trouvent dans les travaux de Lakshmanan *et al.* (Gyssens et Lakshmanan, 1997), Lehner (Lehner, 1998), Pedersen et al. (Pedersen, Jensen, 1999), etc. De même, cette section n'est pas dédiée aux travaux proposant l'intégration dans un même système décisionnel des capacités OLAP avec des mécanismes de fouille de données tel que le système DBMiner (Han, Chiang, Chee et al., 1997).

Étant donnés l'importance de l'aspect temporel des données d'un entrepôt et l'intérêt que présente l'exploration de l'historisation de ces données multidimensionnelles, Ravat *et al.* (Ravat et Teste, 2001) proposent une modélisation à objets munie d'une algèbre regroupant des opérateurs issus de l'algèbre à objet et en particulier des opérateurs spécifiques à la modélisation temporelle proposée pour l'entrepôt. Ces opérateurs permettent de transformer les données de l'entrepôt en séries temporelles pour offrir différents points de vues sur les données de manière à enrichir leur analyse en tirant profit de leur aspect temporel.

S'appuyant sur le fait que les données d'un entrepôt sont souvent entachées d'imperfection et que les requêtes des utilisateurs des cubes OLAP sont dans la plupart du temps vaguement formulées, Laurent (Laurent, 2002) propose un cadre formel pour la mise en oeuvre d'un système de fouille de données floues conjointement avec des outils OLAP. Cette proposition concerne alors l'intégration du flou dans le processus de fouille de données, tant au niveau des bases multidimensionnelles qu'au niveau de leur intégration dans le processus global de fouille de données. Cette proposition permet, en particulier, le couplage d'algorithmes de construction d'arbres de décision flous et de génération de résumés flous avec une base de données multidimensionnelles floues.

C'est dans ce même esprit d'enrichissement des opérateurs OLAP que s'intègre le présent travail. Nous faisons appel aux capacités de la théorie des ensembles approximatifs à traiter des données imparfaites en vue de proposer des approximations aux résultats d'exécution des requêtes OLAP. Une extension des systèmes OLAP vers de telles capacités est sans aucun doute d'une extrême importance vue l'hétérogénéité des données source d'un entrepôt à laquelle s'ajoute une longue et lourde étape de prétraitement visant l'intégration des données selon un schéma d'entrepôt de données. Afin d'illustrer nos propos, prenons l'exemple de

deux cellules d'un cube OLAP qui se partagent les mêmes valeurs pour toutes les dimensions sauf une. Si on avait pu, lors de la création de la table des faits, introduire de la flexibilité dans le traitement des deux enregistrements (faits) très similaires, et plus précisément lors de leur comparaison, il nous aurait été possible d'agréger ces deux enregistrements en un seul permettant ainsi de compacter le cube.

2.2 Structure multidimensionnelle des données

La construction d'un cube de données à partir d'un entrepôt fait appel à l'ensemble des tables suivantes :

- Une table des faits (FT) qui sert à connecter n tables de dimension et à contenir m mesures. Elle comprend des coordonnées correspondant aux dimensions du cube et un contenu correspondant à ses mesures. Elle peut être une table de base (faisant partie du schéma multidimensionnel) ou calculée suite à l'application d'une ou plusieurs opérations d'analyse OLAP ou d'extraction de connaissances.
- Des tables de dimension qui sont associées à chacune des dimensions référencées dans la table des faits FT.
- Des tables qui sont reliées aux tables de dimension et qui sont utilisées pour représenter la hiérarchie des dimensions permettant ainsi d'exprimer les opérateurs de granularité (par exemple : le RollUp).

2.3 La théorie des ensembles approximatifs

La théorie des ensembles approximatifs, connue sous l'expression "*Rough Set Theory*" (RST), a été introduite au début des années 80 par Pawlak (Pawlak, 1982). Elle constitue un cadre mathématique approprié pour le traitement des concepts vagues aux frontières mal définies (Szladow et Ziarko, 1993). Son apparition fût une avancée très importante dans le domaine de l'intelligence artificielle et notamment en apprentissage inductif de concepts à partir de données incohérentes. Elle a été utilisée dans divers domaines tels que la médecine, l'industrie, la finance et le commerce.

Cette théorie est fondée sur les notions d'indiscernabilité et d'approximation. La première notion exprime le degré de similitude entre des objets tandis que la deuxième permet la description d'un concept (ensemble d'objets de l'univers) en tenant compte d'éventuelles imperfections des données manipulées.

Dans un contexte d'approximation, un système d'information A est représenté par un couple (U, A) où U est un ensemble fini et non vide d'objets appelé *univers* et A un ensemble fini et non vide d'attributs.

Soit f une fonction permettant de retourner la valeur de l'attribut d'un objet quelconque de U :

$$\begin{array}{rcl} f : U \times A & \rightarrow & V \\ \forall o \in U \text{ et } a \in A,\ (o,a) & \longmapsto & f(o,a) \in V_a \end{array}$$

avec V_a le domaine de $a \in A$ et V l'union des domaines des attributs de A.

La relation d'indiscernabilité est une relation d'équivalence définie sur un sous-ensemble P d'attributs appelés « attributs de description » servant à décrire les objets de l'univers. Deux objets de l'univers sont indiscernables par rapport à P s'ils se partagent les mêmes valeurs [1]pour chaque attribut de P. Plus généralement, A est divisé en deux sous-ensembles d'attributs : un sous-ensemble C d'attributs de description (condition) utilisés pour décrire les objets de l'univers, et un sous-ensemble D d'attributs de décision (classification) utilisés pour partitionner l'univers en plusieurs concepts dans lesquels les objets de l'univers seront approximativement classifiés.

Etant donnés 2 objets o_i et $o_j \in U$, la relation d'indiscernabilité par rapport à $P \subseteq C$, $I_P \subset U \times U$ est définie comme suit

$$o_i \, I_P \, o_j \iff \forall q \in P \; f(o_i, q) = f(o_j, q)$$

Où, $f(o_i,q)$ est la valeur associée à l'attribut q pour l'objet o.

Puisque la relation d'indiscernabilité est une relation d'équivalence, l'univers U peut être partitionné en une ou plusieurs classes d'équivalence regroupant des objets indiscernables entre eux par rapport aux attributs pris dans P : $U=\{G_1,...,G_n\}$ avec $G_{i(1,...,n)}$ la i^e classe d'équivalence calculée à partir de U.

La RST permet, pour tout concept, le calcul de deux espaces d'approximation. Le premier, « approximation basse », contient des objets dont l'appartenance au concept est certaine. Le second espace, « *approximation haute* », peut contenir des objets n'appartenant pas au concept mais étant indiscernables avec un ou plusieurs objets de ce concept (figure 2). De ces deux espaces d'approximation dérivent d'autres espaces secondaires telles que la région douteuse d'un concept quelconque ainsi que les régions positive et négative des attributs de décision. Il nous est également possible de générer des règles de classification et de caractérisation. Les détails sur ces espaces d'approximation et règles seront fournis en section 4.

Une généralisation de la RST est appelée « α-Rough Set Theory » (α-RST/alpha-RST[1]) (Quafafou, 1997) ; (Naouali et Quafafou, 2004), où alpha représente le degré de similitude tolérée entre les objets selon un ensemble d'attributs de description. De plus amples détails sur cette généralisation de la RST seront donnés dans la section 4.

3. Illustration par un exemple

L'exemple suivant illustre l'approche proposée. Il consiste en une structure simplifiée d'un entrepôt de données appelé *Patents-OLAP* décrivant une application de demande et d'affectation de brevets. Le schéma du cube de données relie directement ou indirectement la table des faits aux dimensions suivantes (figure 1) :

- APDM (APplication Date in Month) : mois durant lequel le brevet a été demandé,

[1] Dans la suite de ce document, on le notera alpha-RST

- APDY (APplication Date in Year) : année correspondant à la demande du brevet,
- ISDM (ISsue Date in Month) : mois durant lequel le brevet a été accordé,
- ISDY (ISsue Date in Year) : année durant laquelle le brevet a été accordé,
- ET (Elapsed Time) : temps écoulé, en tranches de mois, entre la date de la demande et celle de l'obtention du brevet,
- INV (INVentor) : inventeur,
- AT (ATtorney or agent) : représentant légal de l'inventeur,
- ICL (International CLassification) : nomenclature internationale correspondant au produit faisant l'objet du brevet.

L'unique mesure (*nb_patents*) de la table des faits représente le nombre de brevets selon les dimensions retenues.

Comme la classification des brevets est établie manuellement par des agents du bureau des brevets, un brevet quelconque peut facilement appartenir à une classification différente de celle établie par une personne non experte du domaine des brevets. Pour illustrer ce fait, considérons un inventeur qui a demandé deux brevets : un pour un médicament et un autre pour l'emballage qu'il a mis au point spécialement pour ce médicament. L'inventeur a par la suite fait sa demande auprès d'un même représentant légal. Par conséquence, les deux demandes de brevets se partagent plusieurs propriétés communes (l'inventeur, le représentant légal, la date de demande et fort probablement la date d'obtention et le temps écoulé). Selon notre approche, et suivant ces mêmes variables caractéristiques des demandes de brevets, les deux demandes seraient indiscernables et appartiendraient par la suite à une même classification, ce qui n'est effectivement pas le cas car le médicament a été classifié par les experts dans le secteur pharmacie et son emballage dans le secteur papier et imprimerie. Ainsi donc, notre approche peut agir comme un mécanisme d'alerte ou de suggestion de révision de la classification proposée par les agents en mettant l'accent sur la similarité pouvant exister entre les faits de l'entrepôt selon un sous-ensemble des dimensions correspondant à une perspective spécifique de l'utilisateur.

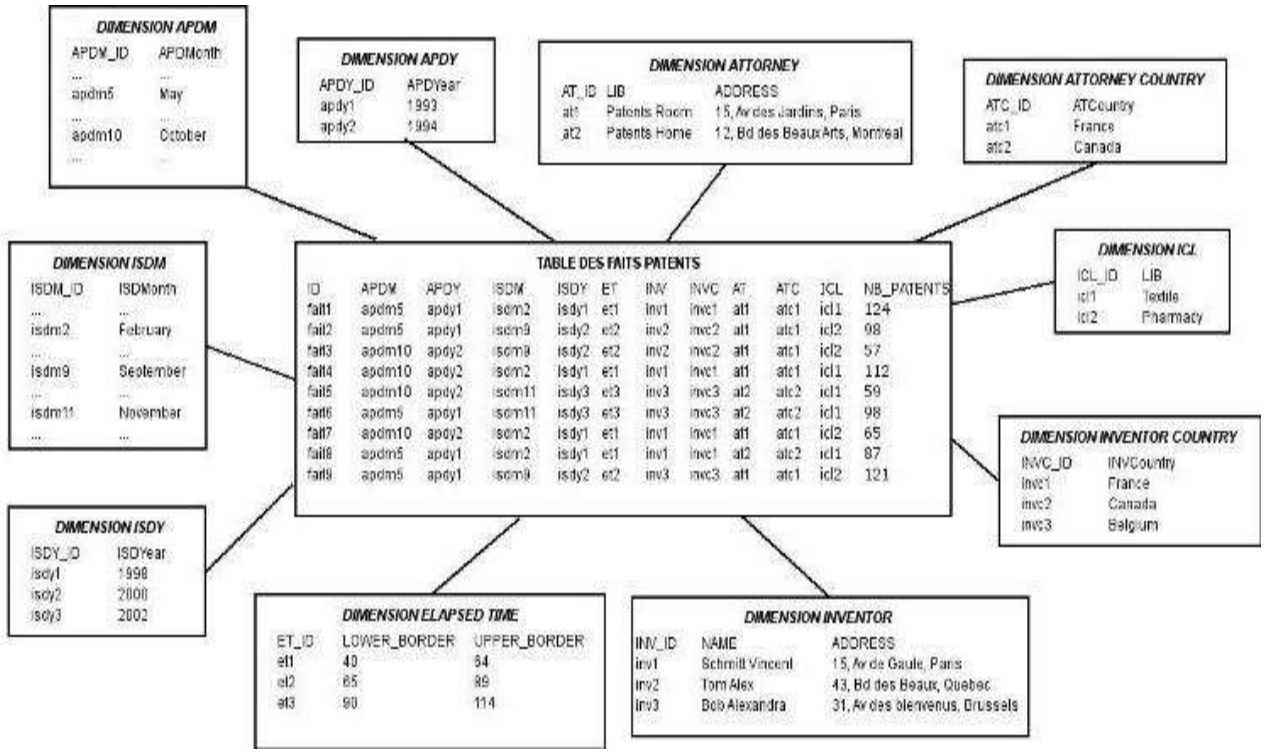

Figure 1. Schéma en flocon de neige et état du cube de données Patents-OLAP

4. Approximation des cubes

Notre contribution consiste, en premier lieu, à intégrer les notions de base de la alpha-RST au contexte multidimensionnel des données d'un entrepôt. Par la suite, des mécanismes de calcul de réponses approximatives aux requêtes utilisateur ainsi que des requêtes d'analyse OLAP et d'extraction de connaissances peuvent être exploités.

4.1 Adaptation de la alpha-RST aux données multidimensionnelles

La alpha-RST offre plus de flexibilité que la RST dans la mesure où l'indiscernabilité entre deux objets est évaluée selon un sous-ensemble des attributs de description pris dans P plutôt que l'ensemble P en entier. Dans cette section, nous présentons une adaptation de la théorie des ensembles approximatifs (Pawlak, 1982) telle que généralisée dans (Quafafou, 1997) aux données multidimensionnelles en proposant de nouveaux opérateurs d'approximation de concepts. Ces opérateurs sont basés sur les notions d'approximations basse et haute d'un concept, les régions positive et négative et les règles de classification et de caractérisation. Ceci peut être perçu comme un enrichissement des techniques OLAP avec des mécanismes de gestion de l'incertitude dans les données manipulées. Cet enrichissement permet d'introduire une certaine flexibilité lors de l'interrogation des données multidimensionnelles.

Dans un contexte multidimensionnel, l'univers U correspond à la table des faits FT et les *alpha-attributs* clé de FT sont des pointeurs logiques vers les tables de dimension. Les dimensions sont partitionnées selon deux ensembles distincts : les dimensions de description C utilisées pour décrire les faits, et les dimensions de décision D utilisées pour des fins de classification et de prédiction. À une même dimension de décision correspondent autant de

concepts cibles que de membres (valeurs) de cette dimension. Par la suite, le calcul des approximations pour une dimension de décision consiste à déterminer l'appartenance des faits de l'entrepôt à chacun des concepts cibles associés à cette dimension selon les dimensions contenues dans C.

Exemple 1 En utilisant l'exemple Patents-OLAP, on peut prendre la dimension ICL comme dimension de décision, et utiliser le reste des dimensions pour caractériser les faits de l'entrepôt. Notre objectif étant par la suite de déterminer l'appartenance d'un brevet quelconque à une certaine classification internationale non seulement selon la nature du produit faisant l'objet de la demande du brevet mais aussi selon un certain nombre de dimensions. Les dimensions retenues dans cet exemple sont la date de demande et la date d'obtention du brevet, le laps de temps écoulé entre ces deux dates, l'inventeur ainsi que son représentant légal. Il nous est par la suite possible de mener une certaine comparaison (superposition) entre la classification établie par les experts et les résultats d'approximation de cette classification tels que retournés par notre approche.

En s'appuyant sur les mécanismes d'approximation de concepts offerts par la théorie des ensembles approximatifs, on peut générer deux classes dans lesquelles un brevet quelconque peut être classé. La première est une classe "stricte" contenant des brevets indiscernables entre eux et auxquels ne correspond aucun autre brevet qui leur soit indiscernable tout en étant en dehors de cette classe. La deuxième est une classe "relaxée" contenant ces mêmes brevets ainsi que ceux appartenant également au même concept traité mais étant indiscernables avec des brevets n'appartenant pas au concept. Ces derniers sont également pris dans ce second espace d'approximation. L'utilisateur peut par la suite décider, selon ses besoins et selon le degré de similitude toléré qu'il a dû fixer, de considérer telle ou telle classe pour décrire le concept faisant l'objet de l'approximation, plutôt que de considérer l'ensemble entier des brevets qu'un mécanisme "classique" de requête aurait retourné.

Puisque la dimension cible ICL possède deux valeurs possibles ; textile et pharmacie (figure 1), deux concepts cibles sont dégagés, i.e., ICL = textile et ICL = pharmacie.

Considérons à présent deux faits x, y ∈ FT, I_p^{alpha} une relation binaire entre deux faits quelconques, appelée également relation d'indiscernabilité établie pour FT par rapport à un ensemble de dimensions de description P ⊆ C et définie comme suit :

$$x I_P^{\alpha} y \Leftrightarrow \frac{|\{q \in P : f(x,q) = f(y,q)\}|}{|P|} \geq \alpha$$

Où alpha qui prend ses valeurs dans l'intervalle [0,1] définit la proportion minimale de dimensions de description que les faits x et y doivent partager pour être considérés comme indisernables (pour alpha = 1, $I_p^{alpha} = I_p$)

Exemple 2 Soit alpha = 0.75, et P = C = {APDM, ET, INVC, ATC}. Les deux faits $fait_2$ et fait9 sont indiscernables (conférer figure 1).

Il est important de noter ici que Ipalpha agit différemment de la relation Ip en partitionnant la table des faits traitée non pas en classes d'équivalence mais plutôt en classes de recouvrement. Ceci est dû au fait que dans une même classe, deux faits quelconques peuvent ne pas être

indiscernables mais chacun l'est de son côté avec un même troisième fait de la même classe. Notons par contre que pour alpha =1, les classes ainsi formées sont des classes d'équivalence.

L'approximation basse de $X \subseteq FT$ par rapport à P, qu'on note $\underline{P}(X)$, est l'union des classes de recouvrement G_k définies selon I_p^{alpha} et incluses dans X :

$$\underline{P}(X) = \bigcup_{G_k \subseteq X} G_k$$

En d'autres termes, l'approximation basse d'un sous-ensemble de faits ne peut contenir que les faits dont l'appartenance au concept cible est certaine. En effet, pour n'importe quel $x_i \in \underline{P}(X)$, on est sûr que $x_i \in X$.

L'approximation haute de X par rapport à P, qu'on note P(X), est l'union des classes de recouvrement définies selon I_p^{alpha} et contenant au moins un fait de X :

$$\overline{P}(X) = \bigcup_{G_k \cap X \neq \emptyset} G_k$$

L'approximation haute de X peut par la suite contenir des faits qui ne sont pas forcément contenus dans X mais qui sont indiscernables avec au moins un fait de X. Ainsi donc, les approximations haute et basse d'un concept X correspondent respectivement à l'intérieur et à la fermeture de cet ensemble dans la topologie générée par la relation d'indiscernabilité tel qu'illustré par la figure 2.

La région douteuse de $X \subseteq FT$, qu'on note P(X), est définie comme suit :

$$\tilde{P}(X) = \overline{P}(X) - \underline{P}(X)$$

Exemple 3 : En considérant toujours le même exemple illustratif, les classes de recouvrement extraites à partir de la table des faits par rapport au sous-ensemble de dimensions de description P={APDM, ET, ATC, INVC} et pour alpha = 0.75 sont {$fait_1$, $fait_4$, $fait_7$, $fait_8$} , {$fait_2$, $fait_3$, $fait_9$}, et {$fait_5$, $fait_6$}.

Pour le concept cible (ICL = textile), l'ensemble des faits lui correspondant est {$fait_1$, $fait_4$, $fait_5$, $fait_6$, $fait_8$}

Son approximation basse est l'union des classes de recouvrement incluses dans l'ensemble de ses exemples, i.e., $\underline{P}$ (ICL = textile) = {$fait_5$, $fait_6$}. Son approximation haute est l'union des classes de recouvrement présentant une intersection non vide avec ses exemples, i.e. P (ICL = textile) = {$fait_1$, $fait_4$, $fait_5$, $fait_6$, $fait_7$, $fait_8$}. La région douteuse de ce concept cible contient les faits qu'on ne peut pas classer sans ambiguïté mais qui sont indiscernables avec des faits de ce concept, i.e., P (ICL = textile) = {fait1, fait4, fait7, fait8}.

La figure 2 illustre la manière avec laquelle les brevets sont distribués selon les espaces d'approximation correspondant à ce concept. Le signe "+" illustre les brevets appartenant certainement au concept, le signe "-" illustre ceux n'appartenant pas au concept et la

différence entre ces deux régions délimitées par des pointillés, représente la régions douteuse, i.e., les brevets dont l'appartenance au concept cible est possible mais pas certaine.

De la même façon, le concept cible (ICL = pharmacie) dont l'ensemble des exemples est {fait2, fait3, fait7, fait9} admet comme approximation basse P (ICL = pharmacy) = {fait2, fait3, fait9} et comme approximation haute P(ICL=pharmacy) = {fait1, fait2, fait3, fait4, fait7, fait8, -fait9}. Sa région douteuse est P(ICL=pharmacie) = {fait1, fait4, fait7, fait8}.

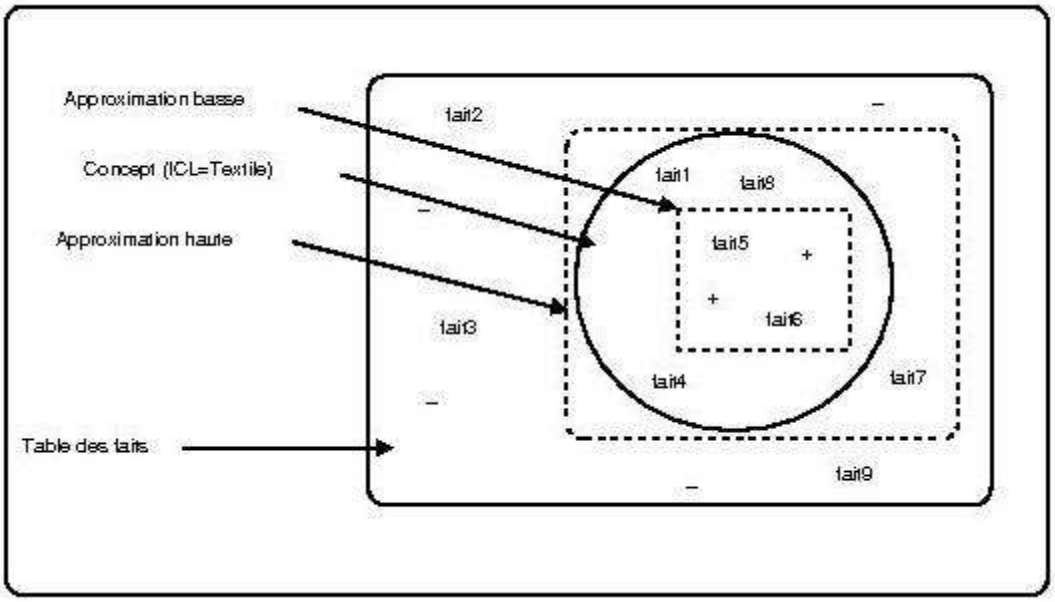

Figure 2. Espaces d'approximation correspondant au concept ICL=Textile

Considérons $d_1,\ldots, d_n$ les valeurs (membres) de la dimension cible d. La région positive de d, qu'on note $POS_p(d)$, par rapport au sous-ensemble de dimensions de description P contient les faits qu'on peut classer avec certitude et sans ambiguïté dans l'un des concepts cibles déterminés pour cette dimension. Elle est définie par :

$$POS_P(d) = \bigcup_{i(\forall i:1\ldots n)} \underline{P}(d_i)$$

Exemple 4 Pour la dimension cible ICL, la région positive est POS_p(ICL) = {fait2, fait3, fait5, fait6, fait9}.

La région négative de d par rapport à P contient les faits dont on est sûr de leur non appartenance à aucun des concepts cibles correspondant à d. Elle est définie comme suit :

$$NEG_P(d) = FT - \bigcup_{i(\forall i:1\ldots n)} \overline{P}(d_i)$$

Exemple 5 Dans notre exemple, NEG_p(ICL) = ϕ.

Une fois que les approximations basses et hautes de chacun des concepts cibles sont calculées, des règles d'association peuvent être générées. L'approximation basse permet de générer des

règles dites de classification ou de décision, définies à partir de la description des faits dont l'appartenance au concept cible ne pose aucune ambiguïté. Les règles extraites à partir de l'approximation haute d'un concept cible quelconque sont dites règles de description (caractérisation) du moment qu'elles sont basées sur la description de faits dont l'appartenance au concept cible est possible mais sans nécessairement être certaine.

Une règle de classification est une expression de la forme $\varphi(d=v_d)$ extraite à partir de l'approximation basse, tandis qu'une règle de description est exprimée par $(d=v_d)\varphi$ extraite à partir de l'approximation haute où d est une dimension de décision, v_d une valeur quelconque (membre) de cette dimension et φ est une disjonction de prédicats définissant des descriptions de faits. Chaque prédicat peut être simple comme il peut être une conjonction de prédicats $c = v_c$ avec $c \in C$ (dimension de description) et v_c une valeur attribuée à c par le fait courant.

Exemple 6 Les règles de classification correspondant à notre exemple illustratif sont :

$$(APDM = Octobre \vee Mai) \wedge (ET = 90 \text{ à } 114 \text{ mois}) \wedge (INVC = Belgique) \wedge (ATC = Canada) \longrightarrow (ICL = Textile)$$

$$((APDM = Mai \vee Octobre) \wedge (ET = 65 \text{ à } 89 \text{ mois}) \wedge (INVC = Canada) \wedge (ATC = France)) \vee ((APDM = Mai) \wedge (ET = 65 \text{ à } 89 \text{ mois}) \wedge (INVC = Belgique) \wedge (ATC = France)) \longrightarrow (ICL = Pharmacie)$$

La première règle peut être exprimée comme suit : « **si** le brevet a été soumis en octobre ou en mai et son acceptation est survenue dans un délai compris entre 90 et 114 mois et le pays de l'inventeur est la Belgique et le pays du représentant légal est le Canada, alors sa classification internationale est Textile.
Cette règle est générée à partir de l'approximation basse P(ICL = textile) = {fait5, fait6}.

4.2 Algorithme

L'algorithme que nous proposons pour le calcul d'approximation de concepts à partir de données multidimensionnelles permet de calculer en une seule fois tous les espaces approximatifs (approximations haute et basse et région douteuse) correspondant à chacun des concepts cibles déterminés pour la dimension cible. Il permet également le calcul des régions positive et négative ainsi que les règles de classification et celles de description correspondant à la dimension cible.

L'algorithme utilise pour ceci les tables de dimension ainsi que la table des faits, prend en considération le sous-ensemble P de dimensions de description, la dimension de décision d ainsi qu'une valeur pour alpha dans l'intervalle [0, 1].

4.3 Explication de l'algorithme

L'algorithme se compose principalement de deux étapes. Lors de la première étape (lignes 8 à 20), l'algorithme détermine les concepts cibles associés à la dimension de décision. Un concept cible est noté Di ce qui est équivalent à d = di avec d la dimension cible et di la ième valeur de d. Par la suite, et pour chacun de ces concepts cibles, l'algorithme détermine la liste

de ses exemples qui sont les faits vérifiant la valeur en question de d (ligne 10). Ensuite, et pour chaque exemple du concept cible, on détermine sa classe de recouvrement à la ligne 12. Puis, si cette classe de recouvrement est incluse dans l'ensemble des exemples du concept courant, alors cette classe est ajoutée à la vue décrivant l'approximation basse du concept cible traité. Sinon elle est ajoutée à la vue décrivant son approximation haute. A la fin de cette première étape et avant de passer au concept suivant, l'algorithme calcule (ligne 19) la vue représentant la région douteuse du concept cible traité.

Après avoir traité tous les concepts cibles, on génère pour la dimension cible les régions positive et négative ainsi que les règles de classification et de description à partir des vues créées lors de la première étape.

Comme indiqué plus haut, l'algorithme proposé calcule les espaces d'approximation qu'il stocke dans des vues partageant la même structure que la table des faits initiale. Ces vues deviennent par la suite des composants importants de l'entrepôt de données, et systématiquement reliées aux tables de dimension. Encore mieux, ces vues permettent à l'utilisateur de générer de nouveaux cubes OLAP encapsulant les approximations basse et haute ainsi que la région douteuse de chaque concept cible ainsi que les régions positive et négative de la dimension cible. Quant aux règles d'association, elles sont stockées dans des tables relationnelles indépendamment du schéma de l'entrepôt.

Algorithm 1 α-approximations

input
TF : table des faits avec m dimensions et k mesures ;
$DIM = \{dim_1, \ldots, dim_m\}$: ensemble de m tables de dimension ($DIM = C \cup \{d\}$) ;
P : sous-ensemble de dimensions de description : $P \subseteq C \subset DIM$;
d : dimension de décision avec n membres ;
α : coefficient déterminant le degré de similitude entre les faits ;
BEGIN{α-approximations}
for i **from** 1 **to** m **do** $D_i = calculer\_concepts\_cible(TF, d)$;/*D_i concept cible déterminé pour d */
for all D_i *in* $D_1, \ldots, D_n$ **do**
  $D_i^{\oplus} := calculer\_exemples(D_i, TF)$; /* $D_i^{\oplus}$ l'ensemble des faits, dits exemples, vérifiant le concept cible D_i */
  for all $f \in D_i^{\oplus}$ **do**
    $[f]_{I_P^{\alpha}} := calculer\_classe\_recouvrement(f, TF, P, \alpha)$; /*$[f]_{I_P^{\alpha}}$ est la classe de recouvrement du fait f déterminée selon I_P^{α} */
    if $[f]_{I_P^{\alpha}} \subseteq D_i^{\oplus}$ **then**
      $insérer([f]_{I_P^{\alpha}}, Vue\_Approximation\_Basse\_D_i)$;
    else
      $insérer([f]_{I_P^{\alpha}}, Vue\_Approximation\_Haute\_D_i)$;
    end if
  end for
  $Vue\_RégionDouteuse\_D_i := Vue\_Approximation\_Haute\_D_i - Vue\_Approximation\_Basse\_D_i$;
end for
$Vue\_POS_P(d) := \bigcup_{i=1}^{n} Vue\_Approximation\_Basse\_D_i$;
$Vue\_NEG_P(d) := U - \bigcup_{i=1}^{n} Vue\_Approximation\_Haute\_D_i$;
$table\_Règles\_Décision\_d := générer\_règles\_décision(Vue\_POS_P\_d)$;
$table\_Règles\_Description\_d := générer\_règles\_description(\bigcup_{i=1}^{n} Vue\_Approximation\_Haute\_D_i)$;
END{α-approximations}

Les avantages de l'algorithme proposé sont :

- le calcul des espaces d'approximation correspondant à une dimension cible quelconque et par rapport à un sous-ensemble de dimensions de description,
- la similarité relative entre les faits par le biais du paramètre alpha,
- la transparence dans l'extraction de connaissances et requêtes OLAP approximatives à partir des vues créées.

4.4 Nouveaux opérateurs pour l'approximation de concepts

Le mot clé Select est utilisé pour permettre à l'utilisateur d'avoir les espaces d'approximation et les règles d'associations ci-dessus discutés une fois l'algorithme exécuté et les tables et vues générées. Les nouveaux opérateurs que nous proposons sont :

- select lower dimension_cible = valeur_cible
- select upper dimension_cible = valeur_cible
- select boundary_region dimension_cible = valeur_cible

- select characteristic_rules dimension_cible = valeur_cible
- select classification_rules dimension_cible = valeur_cible
- select positive_region dimension_cible
- select negative_region dimension_cible

Ces opérateurs agissent comme des filtres sur les cubes OLAP existants en retournant seulement les cellules correspondant à la requête d'approximation posée par l'utilisateur. Ce dernier peut par la suite poursuivre son exploration des données en se limitant au nouveau cube OLAP ainsi généré et en lui appliquant des opérateurs d'analyses OLAP et/ou d'extraction de connaissances. De tels mécanismes enrichissent le processus OLAP en lui rajoutant de nouvelles capacités et en intégrant des aspects OLAP et de nouveaux mécanismes d'extraction de connaissances.

5. Conclusion

Nous avons présenté une approche d'approximation de concepts dans les cubes OLAP par l'intégration de notions puisées de la théorie des ensembles approximatifs. Le but est de calculer des approximations de la réponse à une requête utilisateur en permettant à ce dernier de considérer soit un ensemble "restreint" ne contenant que les cellules qu'on peut classer sans ambiguïté dans la réponse à la requête, soit un ensemble "relaxé" pouvant éventuellement contenir des cellules qu'on ne peut pas classer avec certitude dans le concept. Le premier cas est utile surtout quand la réponse à une requête utilisateur est très volumineuse, d'où l'intérêt de la réduire, tandis que la deuxième alternative est utile surtout dans le cas où la réponse à la requête utilisateur est vide ou presque, d'où l'intérêt de relâcher les conditions de la requête pour que des cellules puissent être classées dans le concept recherché. Notre approche permet également la génération de règles de classification et de description pour des fins de prédiction ou d'association.

Nos travaux actuels consistent à explorer différentes alternatives de réduction de la dimensionnalité au sein des cubes de données ainsi que l'approximation de ces derniers par des techniques statistiques de prédiction.

Bibliographie

Ambite, J. L., Shahabi, C., Schmidt, R. R., Philpot, A. (2001). Fast approximate evaluation of olap queries for integrated statistical data. Proceedings of the First National Conference on Digital Government Research.

Andreasen, T., Motro, A., Christiansen, H., Larsen, H. (2002). On measuring similarity for conceptual querying. Proceedings of the 5th International Conference on Flexible Query Answering Systems, FQAS'2002, London, UK, 100–111.

Chu, W. W., Chen, Q. (1994). A structured approach for cooperative query answering. IEEE Transactions on Knowledge and Data Engineering, 6(5): 738–749.

Ganti, V., Lee, M. L., Ramakrishnan, R. (2000). Icicles: Self-tuning samples for approximate query answering. Proceedings of the 26th International Conference on Very Large Data Bases, VLDB'2000, Morgan Kaufmann Publishers Inc. 176–187

Gyssens, M., Lakshmanan, L. V. S. (1997). A foundation for multi-dimensional databases. In VLDB'1997: Proceedings of the 23rd International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc. 106–115.

Han, J., Chiang, J. Y., Chee, S., Chen, J., Chen, Q., Cheng, S. et al. (1997). Dbminer: A system for data mining in relational databases and data warehouses. Meeting of Minds, CASCON'1997, Toronto, Canada, November, 249–260.

Huh, S. Y., Moon, K. H., Ahn, J. K. (2002). Cooperative query processing via knowledge abstraction and query relaxation. Advanced topics in database research, vol. 1, Idea Group Publishing, 211–228.

Laurent, A. (2002). Bases de Données Multidimensionnelles Floues et leur Utilisation pour la fouille de données. PhD thesis, Université Paris 6, France.

Lehner, V. (1998) Modelling large scale olap scenarios. Proceedings of the 6th International Conference on Extending Database Technology, Springer-Verlag, 153–167.

Muslea, I. (2004). Machine learning for online query relaxation. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Datamining, KDD'2004, ACM Press, 246–255.

Naouali, S., Quafafou, M. (2004). Rough sql : Approximation base querying for pragmatic olap. Proceedings of the IEEE International Conference on Information and Communication Technologies: from Theory to Applications, ICTTA'2004.

Pawlak, Z. (1982). Rough sets. International Journal of Computer and Information Sciences, 341–356.

Pedersen, T. B., Jensen, C. S. (1999). Multidimensional data modeling for complex data. Proceedings of the 15th International Conference on Data Engineering, ICDE'1999, IEEE Computer Society, 336.

Quafafou, M. (1997). alpha-rst: A generalization of rough sets theory. Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing, RSSC'1997.

Ravat, F., Teste, O. (2001). Modélisation et manipulation de données historisées et archivées dans un entrepôt orienté objet. 17ème Journées Bases de Données Avancées, Cépadues Editions - ISBN 2-85428-570-0, 243-256.

Szladow, A., Ziarko, W. (1993). Rough sets - working with imperfect data. In AI Expert, vol. 8, no. 7, 36-41.

Babcock, B., Chaudhuri, S., Das, G. (2003). Dynamic sample selection for approximate query processing. Proceeding of the SIGMOD'03 Conference, 539–550.

Chakrabarti, K., Garofalakis, M., Rastogi, R., Shim, K. (2001). Approximate query processing using wavelets. The VLDB Journal, 10(2-3). 199–223.

Shanmugasundaram, J., Fayyad, U., Bradley, P. S. (1999). Compressed data cubes for olap aggregate query approximation on continuous dimensions. Proceeding of the KDD'1999 Conference, 223–232.

Vitter, J. S., Wang, M. (1996). Approximate computation of multidimensional aggregates of sparse data using wavelets. Proceeding of the SIGMOD'99 Conference, 193–204.

Naouali, S. (2004). Enrichissement d'Entrepôts de Données par la Connaissance : Application au Web. PhD thesis, Université de Nantes, France.