Modèle de préférences contextuelles pour les analyses OLAP

Houssem Jerbi, Franck Ravat, Olivier Teste, Gilles Zurfluh

Université de Toulouse – IRIT (UMR 5505)
118, Route de Narbonne - 31062 Toulouse cedex 9 (France)
{jerbi, ravat, teste, zurfluh}@irit.fr

Résumé. Cet article présente un environnement pour la personnalisation des analyses OLAP afin de réduire la charge de navigation de l'utilisateur. Nous proposons un modèle de préférences contextuelles qui permet de restituer les données en fonction des préférences de l'utilisateur et de son contexte d'analyse.

1 Introduction

Les systèmes OLAP (On-Line Analytical Processing) permettent l'analyse de grands volumes de données issues des systèmes transactionnels de l'entreprise. Ils reposent le plus souvent sur des bases de données multidimensionnelles (BDM) qui organisent les données en sujets d'analyse appelés faits, et axes d'analyse appelés dimensions. L'analyse en ligne OLAP consiste à explorer intuitivement les BDM par l'application d'un ensemble d'opérateurs multidimensionnels (Abelló *et al.*, 2003), (Ravat *et al.*, 2008).

Les systèmes OLAP actuels ont peu de connaissances sur l'utilisateur. Ils ne tiennent pas compte des caractéristiques spécifiques de chaque utilisateur pour la restitution des données, à savoir ses objectifs et ses centres d'intérêts. Ceci oblige l'analyste à naviguer au sein des données par un enchainement d'opérations et une succession de résultats intermédiaires pour obtenir les données pertinentes à sa prise de décision (adaptées à ses besoins spécifiques d'analyse). L'analyse OLAP peut s'avérer alors une tâche fastidieuse qui dégrade les performances du processus d'analyse décisionnelle. Cette dégradation est aggravée par un coût d'exécution important des requêtes dans un environnement OLAP avec un grand nombre de dimensions (Choong *et al*., 2003). Notre objectif est de personnaliser l'exploration des BDM en restituant les données en fonction des préférences utilisateur et de son contexte d'analyse. Ceci permettrait de réduire la charge de navigation de l'utilisateur.

2 État de l'art

À notre connaissance seules deux propositions ont été développées sur la personnalisation de BDM. La première (Bellatreche *et al*., 2005) est centrée sur la personnalisation de la visualisation du résultat d'une requête : elle consiste à déterminer la partie du résultat qui répond aux préférences de l'utilisateur et à une contrainte de visualisation. La seconde proposition (Ravat *et al*., 2007) est centrée sur la personnalisation de l'affichage des paramètres en associant aux éléments du schéma de la BDM des poids reflétant l'intérêt de l'utilisateur.

Nous proposons un modèle de préférences contextuelles plus global que (Ravat *et al*., 2007) où les préférences portent seulement sur les données indépendamment du contexte de leur utilisation. Les préférences que nous définissons sont exploitées pour déterminer les données à restituer, contrairement à (Bellatreche *et al*., 2005) où les préférences impactent simplement les choix de visualisation.

3 Prise en compte de l'utilisateur dans l'analyse des données

Nous avons défini dans (Jerbi *et al.*, 2008) un modèle en constellation regroupant un ensemble de faits associés à des dimensions multi-hiérarchisées. *FIG. 1* présente un exemple de BDM permettant d'analyser les ventes. L'analyse OLAP consiste à explorer ces structures multidimensionnelles. Cette analyse suit un processus navigationnel (Dittrich *et al.*, 2005) consistant en un enchaînement incrémental et interactif d'opérations de manipulation : la rotation pour changer un axe de l'analyse, le forage pour visualiser les données à différents niveaux de granularité, et la sélection pour effectuer des restrictions sur les valeurs affichées.

Notre objectif est de simplifier la tâche de l'utilisateur en réduisant sa charge cognitive et son effort de navigation. Nous introduisons un modèle de préférences liées au contexte d'analyse. Les préférences modélisent les besoins spécifiques de l'utilisateur et ses habitudes d'analyse. Nous intégrons les préférences de l'utilisateur dans l'analyse de données afin d'anticiper sa stratégie de navigation.

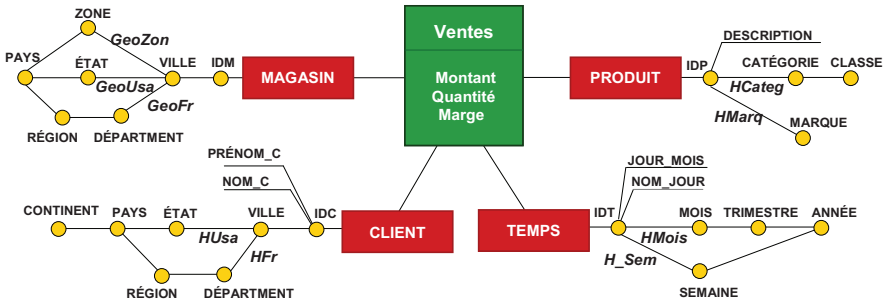


FIG. 1 - *Schéma d'une base de données multidimensionnelles.*

3.1 Modélisation du contexte d'analyse OLAP

Un contexte d'analyse représente un résultat intermédiaire d'une analyse OLAP. Chaque contexte d'analyse regroupe des composants du schéma en constellation (fait, mesures, dimensions et hiérarchies) ainsi que des instances de ce schéma (valeurs des paramètres et des mesures affichés). Nous distinguons deux catégories de composants :

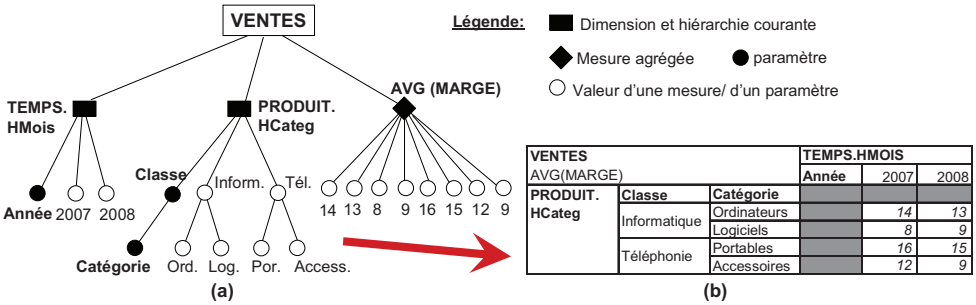
- des composants relatifs au sujet d'analyse : fait (F), mesure (m), valeur d'une mesure affichée (val_m), fonction d'agrégation des mesures (f^{Agreg}). Ces composants forment le contexte fait, noté C^F.
- des composants relatifs à chaque axe d'analyse du contexte : dimension (D) avec sa hiérarchie courante (H), paramètre ou attribut faible (p), valeur d'un attribut affiché (val_p). Ces composants forment le contexte dimension C^{Di}.

Définition. Un contexte d'analyse OLAP est défini par $C = \{C^F, C^{D_1}, \ldots, C^{D_n}\}$ où

- $C^F = F / (f^{Agreg}. m_j = val_m)^+$ est un contexte fait où f^{Agreg} est une fonction d'agrégation (SUM,…), $m_j \in M^F$ (mesures relatives au fait F), et $val_m \in Dom(m_j)$ [1],
- $\forall i \in [1..n]$, $C^{D_i} = D_i.H^i_j (/ p_k = val_p)^+$ est un contexte dimension où $p_k \in H^{Di}$ (attributs de D_i appartenant à la même hiérarchie H^{Di}), $val_p \in Dom(p_k)$.

Arbre de contexte. Un contexte d'analyse est représenté par une structure arborescente $A(N^C, A^C)$ (N^C est un ensemble de nœuds et A^C est un ensemble d'arcs) qui traduit la nature de la relation hiérarchique entre les différents éléments d'une analyse OLAP (cf. *FIG. 2* (a)). Il existe deux types de nœuds dans N^C: (1) des nœuds 'structures' (fait, mesure, dimension, attributs de dimension), et (2) des nœuds 'valeurs' (pour chaque attribut ou mesure).

Un intérêt de cette structure arborescente est son indépendance vis-à-vis du choix de visualisation. Ainsi, cette représentation interne est affichée à l'utilisateur sous diverses formes (tableau, diagramme, etc) facilitant l'interprétation des résultats. Une structure de visualisation très courante d'un contexte d'analyse est la Table Multidimensionnelle (TM) (Gyssens *et al*., 1997), qui centre l'analyse sur un fait suivant deux dimensions (cf. *FIG. 2* (b)).



VENTES AVG(MARGE)			TEMPS.HMOIS		
			Année	2007	2008
PRODUIT. HCateg	Classe	Catégorie			
	Informatique	Ordinateurs		14	13
		Logiciels		8	9
	Téléphonie	Portables		16	15
		Accessoires		12	9

FIG. 2 - *Exemple de contexte d'analyse des marges des ventes.*

Exemple. *FIG. 2* présente un exemple de contexte d'analyse des données issues de la BDM en *FIG. 1*. Le contexte présenté est $C = \{C^F, C^{D1}, C^{D2}\}$ où C^F = Ventes / AVG.Marge $\in \{14, 13, 8, 9, 16, 15, 12, 9\}$, C^{D1} = Produit.HCateg / Classe $\in$ {Informatique, Téléphonie} / Catégorie $\in$ {Ordinateurs, Logiciels, Portables, Accessoires}, C^{D2} = Temps.HMois / Année $\in$ {2007, 2008}.

Le passage d'un contexte d'analyse à un autre se fait à l'aide d'une opération OLAP. Dans notre cadre formel, le passage d'un arbre de contexte à un autre est effectué par une opération de manipulation d'arbre. Nous distinguons trois types d'opérations :

- *Création de l'arbre*. Elle permet de créer le contexte d'analyse initial correspondant au début d'une analyse.
- *Suppression d'un nœud.* Cette opération élimine un axe d'analyse, une mesure, un paramètre, une valeur de paramètre ou de mesure. La suppression d'un nœud engendre la suppression de tous ses descendants. Notons que la suppression de la valeur d'un paramètre entraine la suppression des valeurs de mesures dépendantes.

[1] *Dom*(*e*) est l'ensemble des valeurs de e

- *Ajout d'un sous arbre*. Il s'agit d'insérer un nouvel axe d'analyse (avec éventuellement ses niveaux de granularités et leurs valeurs), une nouvelle mesure, un nouveau paramètre ou une valeur de paramètre ou de mesure.

Les opérations de transformation de l'arbre de contexte nécessitent une mise à jour des nœuds valeurs des mesures. En effet, la suppression ou l'insertion d'un axe d'analyse ou du paramètre le plus détaillé d'une dimension (qui est lié aux nœuds valeurs) nécessite un nouveau calcul de l'agrégation des mesures.

Les opérations de manipulation d'arbre de contexte sont indépendantes des mécanismes des opérations OLAP pour lesquelles il n'existe pas un langage standard.

A chaque opération OLAP exprimée par l'utilisateur correspond une ou plusieurs opérations de transformation de l'arbre ; par exemple, une opération de rotation de dimension (Ravat *et al.*, 2008) correspond à une suppression du nœud dimension suivie de l'insertion d'un sous arbre.

3.2 Modélisation des préférences contextuelles

Une préférence modélise les besoins spécifiques d'un utilisateur en définissant ses priorités d'analyse dans un contexte particulier. Le *contexte d'une préférence* représente le cadre d'analyse dans lequel elle s'intégre. Conceptuellement, il s'agit d'un fragment de l'arbre de contexte d'analyse (Jerbi *et al.*, 2008).

Les *préférences* utilisateur que l'on considère concernent (1) les axes d'analyse (dimensions) et (2) les niveaux de granularité de l'analyse selon une dimension.

Définition. Soit une constellation CS. Une *préférence inter-dimensions*, notée $P^{C}_{k} = (CS, >_p, C)$, est un ordre partiel strict sur l'ensemble D^C des dimensions de la constellation CS qui sont connectées au même fait, où $>_p \subseteq D^C \times D^C$. C désigne le contexte de la préférence.

Définition. Une *préférence intra-dimension* P^{H}_{k} est définie par $P^{H}_{k} = (H, >_p, C)$ où H est une hiérarchie d'une dimension D, $>_p$ est un ordre partiel strict sur A^H (ensemble des attributs de la hiérarchie H) où $>_p \subseteq A^H \times A^H$, et, C est un contexte de préférence

Exemple. Le directeur des ventes désire avoir une vision générale des marges des ventes au cours du temps. Pour cela, il préfère visualiser la dimension Client par pays puis par continent. Par contre, il est intéressé par une vision plus détaillée (région puis département) dans le contexte d'analyse des ventes de l'année en cours. Ses préférences sont traduites par :

- $P^{HFr}_{1} = (HFr, >_p, C_1)$, tel que : *pays* $>_p$ *continent*, $C_1 = \{Ventes/AVG(Marge), Temps\}$
- $P^{HFr}_{2} = (HFr, >_p, C_2)$, tel que : *région* $>_p$ *département*, $C_2 = \{Ventes/AVG(Marge), Temps.HMois / Année = 2008\}$

Le directeur marketing est lui intéressé aux marges des ventes en fonction du temps par ville puis par département puis par région. Ceci est traduit par une préférence intra-dimension $P^{HFr}_{3} = (HFr, >_p, C_1)$, tel que : *ville* $>_p$ *département* $>_p$ *région*.

4 Personnalisation des analyses OLAP

Notre approche consiste à intégrer les préférences contextuelles de l'utilisateur afin de réduire sa charge de navigation. Le résultat d'une opération OLAP, qui est classiquement restitué à l'utilisateur, est gardé en interne. Le système l'enrichi, voire le transforme en fonction des préférences utilisateur. Le mécanisme de personnalisation d'une opération OLAP est décrit comme suit (cf. *FIG.3*):

- La TM initiale détermine le contexte d'analyse initial CA_{init}.
- Le système transforme CA_{init} en fonction des opérations de manipulation d'arbre. Le résultat est un contexte intermédiaire (CA_{interm}) qui n'est pas affiché à l'utilisateur.
- Le système sélectionne les préférences impliquées par CA_{interm} par appariement avec les contextes des préférences.
- L'arbre de CA_{interm} est transformé en fonction des préférences sélectionnées. Le contexte d'analyse résultat (CA_{Res}) est affiché à l'utilisateur.

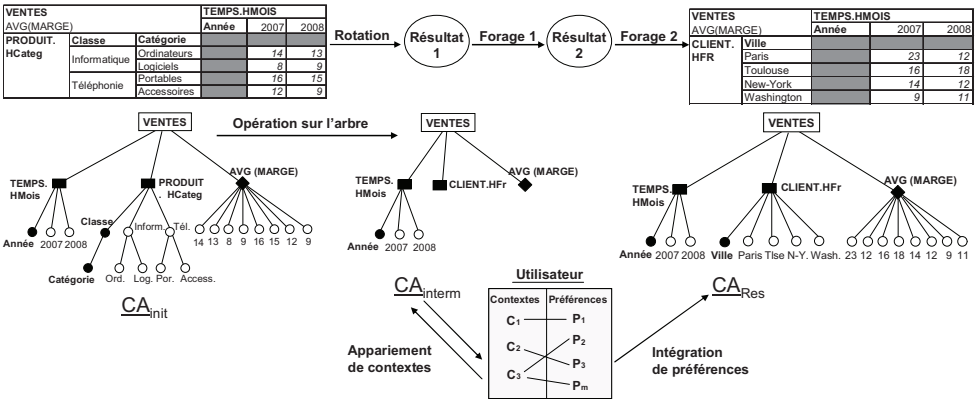


FIG.3 - *Mécanisme de personnalisation d'une analyse OLAP.*

Exemple. Considérons le contexte d'analyse représenté en *FIG. 2*. Le directeur marketing désire changer l'axe d'analyse Produit par l'axe Client. Cette opération se traduit par la suppression du nœud « Produit » avec ses descendants et l'insertion d'un nouveau nœud « Client ». Le système intègre le paramètre « Ville » conformément à la préférence P^{H}_{3}. Dans un fonctionnement classique d'une rotation (Ravat *et al.*, 2008), le système affiche le paramètre de plus haute granularité de la dimension (Année). L'utilisateur doit procéder ensuite à des opérations de forage sur la dimension Client pour avoir les données par ville (cf. *FIG.3*).

Nous détaillons dans la suite l'étape de sélection des préférences.
Cette étape vise à déterminer les préférences pour l'enrichissement du résultat d'une requête. Les préférences candidates dépendent du contexte d'analyse CA_{interm} généré par la requête. S'il existe une préférence dont le contexte est égal à CA_{interm}, elle est prise en compte, sinon, le système cherche celle dont le contexte est inclus dans CA_{interm}. Le problème se ramène alors à un problème de résolution de contexte : un contexte C (représenté par l'arbre de contexte A) est « inclus » dans C' (représenté par A') si son arbre est un sous-arbre de A'. Nous avons défini un algorithme qui détermine la liste de ces préférences (Jerbi *et al.*, 2008).

S'il existe plusieurs préférences candidates, le système sélectionne la préférence dont le contexte couvre le plus le contexte intermédiaire CA_{interm}.

Définition. Soit C un contexte et $E = \{C_1, C_2, \ldots\}$ un ensemble de contextes inclus dans C. On dit que C_i couvre le plus C si C_i admet plus d'éléments en commun avec C, c'est-à-dire si l'arbre de C_i admet le plus de sommets (ou d'arcs) qui appartiennent à celui de C.

Exemple. Nous considérons le contexte d'analyse représenté en *FIG. 2*. Supposons que le directeur des ventes désire effectuer la même manipulation que le directeur marketing

(voir exemple précédent). Les préférences P^{H}_{1} et P^{H}_{2} sont deux préférences candidates. Le contexte de P^{H}_{2} couvre plus CA_{interm}. L'enrichissement se fait conformément à P^{H}_{2}. Le système affiche les données par régions des clients. Ainsi le directeur marketing et le directeur des ventes obtiennent des résultats différents tenant compte des objectifs d'analyse de chacun.

5 Conclusion

Nous avons proposé un cadre pour la personnalisation des analyses OLAP basé sur un modèle de préférences contextuelles. Ce modèle repose sur les concepts de contexte d'analyse et de préférence. Le contexte d'analyse est représenté par un arbre indépendant de la visualisation des données. Nous avons défini des opérations de transformation du contexte qui permettent de passer d'un contexte à un autre indépendamment des choix de présentation à l'utilisateur. Le résultat d'une requête produit un contexte intermédiaire qui est enrichi par les préférences de l'utilisateur. La sélection de ces préférences est assurée par un mécanisme d'appariement entre contextes et leur intégration permet de générer un contexte résultat qui est plus adapté aux besoins de l'utilisateur. Ceci réduit la charge de navigation qui est souvent nécessaire dans les systèmes OLAP classiques.

Nous poursuivons nos travaux par l'étude de techniques d'apprentissage et de mise à jour automatique des préférences utilisateurs.

Références

Abelló, A., Samos, J., Saltor, F. (2003). Implementing operations to navigate semantic star schemas. Intl. Workshop DOLAP'03, pp. 56-62.

Bellatreche, L., Giacometti, A., Marcel, P., Mouloudi, H., Laurent, D. (2005). A personalization framework for OLAP queries. Intl. Workshop DOLAP'05, pp. 9-18.

Choong Y. W., Laurent D., Marcel P. (2003). Computing Appropriate Representations for Multidimensional Data. Int. Journal DKE, Volume 45, pp.181-203.

Dittrich J. P., Kossmann D., Kreutz A. (2005). Bridging the gap between OLAP and SQL. 31st Intl. Conf. on Very Large Data Bases, pp.1031-1042.

Gyssen, M., Lakshmanan, L.A. (1997). Foundation for multi-dimensional databases. 23rd Intl. Conf. on Very Large Data Bases, pp. 106-115.

Jerbi, H., Ravat, F., Teste, O., Zurfluh, G. (2008). Management of context-aware preferences in Multidimensional Databases. 3rd Intl. Conf. on Digital Information Management, Londres, Angleterre, pp. 669-675.

Ravat, F., Teste, O., Zurfluh, G. (2007). Personnalisation de bases de données multidimensionnelles. In Proc. XXIVème Congrès INFORSID, pp. 121-136.

Ravat, F., Teste, O., Tournier, R., Zurfluh, G. (2008). Algebraic and graphic languages for OLAP manipulations. Intl. Journal of Data Warehousing and Mining, Volume 4, p.17-46.

Summary

In this paper we define a framework to personalize users' analysis. We provide a context-aware preference model by which users are provided with relevant data according to their preferences as well as their analysis context.