Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils

Mokrane Bouzeghoub — Dimitre Kostadinov

Laboratoire PRiSM, Université de Versailles
45, avenue des Etats-Unis, 78035 Versailles
{Prénom.Nom@prism.uvsq.fr}

RÉSUMÉ. *Le but de la personnalisation est de faciliter l'expression du besoin de l'utilisateur et de lui permettre d'obtenir des informations pertinentes lors de ses accès à un système d'information. La pertinence de l'information se définit par un ensemble de critères et de préférences personnalisables spécifiques à chaque utilisateur ou communauté d'utilisateurs. Les données décrivant les utilisateurs sont souvent regroupées sous forme de profils. Le contenu du profil d'un utilisateur varie selon les approches et les applications. Les approches existantes répondent partiellement aux questions liées à la personnalisation, mais il manque un modèle donnant une vision globale sur les aspects de la prise en compte des préférences des utilisateurs. Dans cet article nous proposons un modèle de profil générique qui permet de classifier un grand nombre d'informations contenues dans les profils. Nous présentons également un ensemble d'opérateurs de manipulation de profils et une plateforme de gestion de profils.*

ABSTRACT. *The goal of personalization is to facilitate the expression of the need for a particular user and to enable him to obtain relevant information when he accesses an information system. The relevance of the information is defined by a set of criteria and preferences specific to each user or community of users. The data describing the users is often gathered in the form of profiles. The content of the profile vary according to the approaches and the applications. The existing approaches answer partially the questions related to the personalization, but it misses a model giving a global vision on the aspects of the taking into account of the preferences of the users. In this article we propose a generic model of profile, which enables the classification of the profile's contents. We also present some operators of profile manipulation and a platform for the profiles management.*

MOTS-CLÉS : *personnalisation, model générique de profil, opérateurs, manipulation de profils.*

KEYWORDS : *personalization, generic profile model, operators, profile manipulation.*

Cette recherche a été partiellement soutenue par le Ministère Délégué à la Recherche et aux Nouvelles Technologie, dans le programme ACI Masses de Données, projet #MD-33.

1. Introduction

L'accès à une information pertinente, adaptée aux besoins et au contexte de l'utilisateur est un challenge dans un environnement Internet ou GRID, caractérisé par une prolifération des ressources hétérogènes (données structurées, documents textuels, composants logiciels, images), conduisant à des volumes de données considérables. Au fur et à mesure que ce volume s'accroît et que les données se diversifient, les systèmes de recherche d'informations (moteurs web, SGBD, etc.) délivrent des résultats massifs en réponse aux requêtes des utilisateurs, générant ainsi une surcharge informationnelle dans laquelle il est souvent difficile de distinguer l'information pertinente d'une information secondaire ou même du bruit. La personnalisation de l'information constitue un enjeu majeur pour l'industrie informatique. Que ce soit dans le contexte des systèmes d'information d'entreprise, du commerce électronique, de l'accès au savoir et aux connaissances ou même des loisirs, la pertinence de l'information délivrée, son intelligibilité et son adaptation aux usages constituent des facteurs clés du succès ou du rejet de ces systèmes.

La personnalisation de l'information se définit, entre autres, par un ensemble de préférences individuelles représentées par des couples *(attribut, valeur)*, par des ordonnancements de critères ou par des règles sémantiques spécifiques à chaque utilisateur ou communauté d'utilisateurs. Ces préférences servent à décrire le centre d'intérêt de l'utilisateur, le niveau de qualité des données qu'il désire ou les modalités de présentation de ces données. L'ensemble de ces informations est représenté dans un modèle d'utilisateur appelé souvent *profil*. Un profil regroupe l'ensemble des connaissances nécessaire à une évaluation efficace des requêtes et à une production d'une information pertinente adaptée à chaque utilisateur. Il convient donc de distinguer la notion de *profil* de la notion de *requête*. Un profil peut être défini comme un modèle personnalisé d'accès à l'information alors qu'une requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil. Un profil a un caractère plus invariant que les requêtes même si le centre d'intérêt et les préférences de l'utilisateur peuvent évoluer.

La personnalisation de l'information a été particulièrement abordée dans la communauté Recherche d'Information (RI), la communauté Bases de Données (BD) et la communauté de l'interaction homme-machine (IHM). Dans le domaine de la RI, l'utilisateur fait partie du processus de personnalisation. L'évaluation d'une requête se fait généralement de façon interactive et incrémentale; à chaque itération, le système tient compte des informations collectées à partir des interactions précédentes avec l'utilisateur ou profite de l'expérience des autres utilisateurs (filtrage collaboratif). La personnalisation est ainsi définie comme un apprentissage réalisé à partir des préférences rendues par les utilisateurs à l'issue de la présentation des résultats successifs du système.

Dans le domaine des BD, il n'est pas courant d'intégrer l'utilisateur dans le processus de recherche d'informations. Une requête SQL contient en général

l'ensemble des critères jugés utiles à une sélection de données pertinentes. Les profils sont alors intégrés directement aux requêtes par les utilisateurs ou lors de la compilation de ces dernières; ils sont alors pris en compte en une seule fois durant le filtrage de l'information.

Dans le domaine des IHM, la notion de profil, souvent appelé modèle d'utilisateur, se focalise plus sur le niveau d'expertise et le métier de l'utilisateur afin de déterminer le type de dialogue que le système va avoir avec lui, les métaphores graphiques les plus appropriées ainsi que les modalités de livraison des résultats qu'il attend du système d'information. Dans une approche de type MVC, c'est le contrôleur qui prend généralement en charge la satisfaisabilité du profil.

L'objectif de cet article est de faire une classification des différentes connaissances constituant un profil et d'établir un cadre de référence pour définir et manipuler des profils. Ce cadre de référence se représente sous forme d'une typologie des informations nécessaires à la définition d'un profil dans un système d'accès personnalisé aux masses de données. La section 2 montre quelques exemples de profils, des exemples d'utilisation de profils dans les requêtes et des extensions des langages avec des préférences. Dans la section 3 nous présentons un modèle générique de profil et un ensemble d'opérations de gestion de profils. La plateforme de gestion de profils est présentée dans la section 4. La section 5 conclut sur les perspectives de recherche ouvertes par ce modèle générique.

2. Synthèse de l'état de l'art

Cette section a pour objectif d'illustrer, à travers différents travaux, la notion de profil ainsi que son utilisation dans l'accès à l'information. Nous montrerons d'abord comment certains profils sont définis dans leurs contextes applicatif et technologique. Nous montrerons ensuite comment ces profils sont utilisés pour enrichir les requêtes. Enfin, nous montrerons comment certains langages de requêtes prennent en compte les préférences définies dans les profils.

2.1 *Quelques exemples de profils*

Le contenu du profil d'un utilisateur varie selon les approches et les applications. Dans le domaine IHM, le profil contient, par exemple, des informations permettant au système d'adapter l'affichage des résultats selon les préférences de l'utilisateur. Un exemple simple d'un tel profil est celui utilisé par les fournisseurs de services Web: le profil d'un utilisateur est défini par un ensemble de données personnelles (nom, prénom, langue, genre, date de naissance, code postal, e-mail, profession, etc. dans le cas de Yahoo par exemple) et des catégories d'intérêts qui constituent sa page d'accueil (météo, horoscope, football, jeux, etc.). Le contenu de certaines catégories du centre d'intérêt peut être déduit à partir des données personnelles, comme le signe astral à partir de la date de naissance par exemple.

Dans le domaine de la RI, le profil de l'utilisateur décrit le plus souvent son centre d'intérêt et, de ce fait, est souvent confondu avec la requête de l'utilisateur. Ce profil est généralement défini à l'aide d'un vecteur de mots clés avec éventuellement un poids associé à chaque mot (Fereira et al., 2001), (Gauch et al., 1999), (Crabtree et al., 1998), (Croft et al., 2001). Par exemple, le profil d'un utilisateur intéressé par la personnalisation des données peut être présenté par le vecteur à trois termes suivant; le poids de chaque terme correspond généralement à sa fréquence d'apparition dans les documents:

Exemple 1: profil vectoriel
{(personnalisation, 0.7), (profil, 0.9), (modèle, 0.5)}

Dans le projet CASPER (Bradley et al., 2000) qui présente un moteur de recherche d'emploi, le profil d'un utilisateur est défini sous la forme de statistiques des actions qu'il a effectuées sur les offres d'emplois. L'intérêt de l'utilisateur pour une annonce est déterminé en fonction du temps qu'il a passé à la lire et du type d'action qu'il a effectuée dessus. Prenons par exemple un utilisateur qui a lu une annonce de travail, a postulé pour une autre et a envoyé une troisième à un ami :

Exemple 2 de profil :

Annonce	Action	Nombre de Clicks	Temps de lecture
job5	lire	1	234
job56	candidater	2	186
job45	envoyer à un ami	1	54

En utilisant un ensemble de règles de décision, le système va décider quelle annonce est pertinente pour l'utilisateur. Par exemple il peut considérer que les annonces 'job56' et 'job45' sont pertinentes pour l'utilisateur en raison des actions effectuées (candidater et envoyer à un ami). Par contre il n'a fait que lire l'annonce 'job5' qui sera considérée comme inintéressante.

Une autre approche de personnalisation est présentée par (Cherniack et al., 2003) qui donne le moyen d'exprimer des fonctions d'utilité (*utility*) sur un domaine d'intérêt (*domain*). La clause *DOMAIN* définit les sujets du centre d'intérêt de l'utilisateur et donne un nom abstrait à chaque objet. La clause *UTILITY* spécifie les valeurs relatives de l'intérêt de chacun de ces sujets en utilisant des équations d'utilité. Prenons par exemple le profil d'un voyageur dont le domaine d'intérêt est constitué de trois sujets: les companies de location de voitures (LV), les horaires des navettes (Na) et les cartes de direction routière (Di) (exemple 3 de profil). Dans ce profil il y a deux équations d'utilité pour exprimer l'importance relative des trois sujets. La fonction UPTO est un opérateur de définition de seuil. UPTO (x, m, n) signifie que les x premiers éléments du résultat ont une utilité égale à m et tous les autres une utilité égale à n. Sur l'exemple, *U(LV [# Di > 0]) = UPTO (2, 2, 0)* signifie que si le nombre de cartes de direction trouvées est supérieur à 0 ([# Di > 0]), les deux premières agences de location de voitures trouvées ont une utilité de 2

les autres n'en ont aucune. De même U(Na) = UPTO (1, 3, 0) signifie que l'utilité du premier horaire des navettes est de 3 et il est de 0 pour les autres.

Exemple 3 de profil :
PROFILE Voyageur
DOMAIN
LV = www.hertz.com (compagnies de location de voitures)
Na = « horaires des navettes qui vont de l'aéroport vers Boston»
Di = « cartes des directions de l'aéroport vers Boston »
UTILITY
U(LV [#Di > 0]) = UPTO (2, 2, 0)
U(Na) = UPTO (1, 3, 0)
END

Dans le domaine des BD, le profil de l'utilisateur contient des données qui expriment ses habitudes, des prédicats fréquemment utilisés dans ses requêtes ou des définitions d'ordre dans les prédicats (Koutrika et al., 2004). L'intérêt de l'utilisateur pour chacun de ces éléments est exprimé par un *degré* qui est un nombre réel compris entre 0 et 1. Prenons par exemple une BD dont le schéma est le suivant :

TRANSPORT (idT, moyen)
HOTEL (idH, nombre_étoiles, région)
VOYAGE (idV, prix, lieu_départ, lieu_arrivée, nombre_jours, idH, idT)
DEPART (idD, idV, date, heure)

L'exemple 4 décrit le profil d'un utilisateur qui habite à Paris, qui aime voyager pendant le week-end, descend d'habitude dans des hôtels au centre ville et préfère voyager en train plutôt qu'en car. Sur chaque expression du profil, considérée comme une sous requête, est ajouté un nombre compris entre 0 et 1 pour exprimer l'importante relative de cette expression par rapport aux autres. Ainsi la valeur 1 sur les trois premières expressions signifie que ces conditions doivent être toujours satisfaites. Les autres expressions expriment le fait que l'utilisateur a une plus forte préférence pour les hôtels situés au centre ville (d) que pour les voyages de deux jours (e) et qu'il préfère voyager en train (f) plutôt qu'en car (g).

Exemple 4 de profil :

{	TRANSPORT.idT = VOYAGE.idT	1	(a)
	HOTEL.idH = VOYAGE.idH	1	(b)
	VOYAGE.lieu_départ = 'Paris'	1	(c)
	HOTEL.région = 'centre ville'	0.9	(d)
	VOYAGE.nombre_jours = 2	0.7	(e)
	TRANSPORT.moyen = 'train'	0.7	(f)
	TRANSPORT.moyen = 'car'	0.5 }	(g)

Comme nous venons de le voir, il y a autant de définitions de profils que d'applications. Bien que les exemples de profils présentés dans cette section ne donnent qu'un aperçu de la multitude des informations contenues dans le profil d'un utilisateur, ils permettent de dégager quelques catégories de classification de ces informations comme par exemple le *domaine d'intérêt* (mots clés, prédicats de recherche, etc.), les *données personnelles* (nom, prénom, âge etc.) ou encore

l'historique du comportement de l'utilisateur (temps de lecture d'un document, nombre de clicks etc.). Nous reviendrons dans la section 3 sur ces catégories d'informations pour les compléter et les détailler.

2.2. *Utilisation des profils dans les requêtes*

Les données caractérisant un profil sont utilisées dans le processus de recherche d'informations afin de fournir à l'utilisateur des résultats pertinents pour sa requête. Nous allons examiner dans cette section comment les informations de profil sont exploitées dans l'accès aux données. Nous le montrerons à travers quelques exemples pris dans les domaines des IHM, de la RI et des BD.

Dans le domaine des IHM, la notion de requête n'existe pas sous forme langagière. Les systèmes utilisent des connaissances sur l'utilisateur (âge, niveau d'expertise, handicaps etc.) ou sur la technologie qu'il utilise (type du media, logiciels etc.) pour lui fournir une interface d'interaction adaptée. Un des objectifs de ces systèmes est de guider l'utilisateur dans ces recherches et de faciliter l'expression de ses besoins. Un exemple de tel système est 'Apt Decision' qui représente un agent de recherche d'appartements (Lieberman et al., 2001). Initialement l'utilisateur soumet un ensemble de critères de recherche (nombre de pièces, surface etc) et ensuite par le biais de l'interaction, le système guide l'utilisateur à travers les annonces disponibles. A chaque étape, le système analyse les actions que l'utilisateur effectue sur les annonces affichées pour lui proposer, dans la prochaine itération, des appartements conformes à ses préférences.

Dans le domaine de la RI, le profil est souvent utilisé pour remplacer la requête de l'utilisateur (Ferreira et al., 2001) ou pour rajouter des mots clés supplémentaires. Prenons par exemple le profil de l'utilisateur intéressé par la personnalisation des données (exemple 1 de la sous-section précédente). S'il soumet une requête contenant les mots clés : {techniques, personnalisation}, les termes de son profil qui n'apparaissent pas dans la requête (profil et modèle) seront ajoutés à celle-ci. La requête enrichie devient : {techniques, personnalisation, profil, modèle}. Ici le profil est présenté comme un vecteur à N dimensions où les dimensions sont définies par les termes les plus significatifs pour les documents recherchés, le système calcule le matching entre le profil et les mots clés significatifs extraits des documents en utilisant une technique basée sur la distance entre vecteurs à N dimensions. Seuls les documents dont le matching dépasse un certain seuil (spécifié par l'utilisateur) sont inclus dans le résultat.

Une autre approche consiste à exécuter la requête de l'utilisateur sans prendre en compte ses préférences et ensuite filtrer les résultats avant de les retourner à celui-ci. Par exemple, dans (Bradley et al., 2000), on utilise une technique de raisonnement par cas pour déterminer si une annonce est pertinente pour l'utilisateur. Prenons l'exemple 2 de profil et supposons que sur les trois annonces du profil de l'utilisateur il trouve les annonces job56 et job45 pertinents et l'annonce job5 ne l'intéresse pas. S'il soumet une requête de recherche d'annonces pour laquelle il y a

un ensemble d'annonces dans le résultat retourné, le système va décider pour chaque annonce du résultat si elle est pertinente ou pas pour l'utilisateur en la comparant aux annonces du profil. De cette manière le résultat est filtré et seules les annonces pertinentes (similaires à 'job56' et 'job45') sont retournées à l'utilisateur.

D'autres systèmes de personnalisation utilisent le filtrage collaboratif. Le principe de ces approches est de : (i) calculer la similarité entre le profil de l'utilisateur courant et les profils des autres utilisateurs; (ii) sélectionner les n profils les plus similaires au profil du client; (iii) faire des prédictions sur les éléments qui pourraient intéresser l'utilisateur en utilisant le contenu des n profils choisis. Un exemple de tel système est défini dans (Dai et al. 2000). Le profil d'un utilisateur est comparé à ceux des autres clients pour déterminer un ensemble d'éléments qui lui seront recommandés lors d'une session S.

Dans le domaine des BD, la requête de l'utilisateur contient l'ensemble des prédicats utiles à la sélection des informations pertinentes. Dans ce contexte, le travail mené dans (Koutrika et al., 2004) propose une approche d'enrichissement de la requête de l'utilisateur par de nouveaux critères de sélection contenus dans son profil. Ceci est fait en deux étapes : (i) sélection des k-meilleurs critères du profil de l'utilisateur et (ii) intégration des critères sélectionnés dans la requête en spécifiant le nombre minimal (L) de critères dont l'intérêt est inférieur à 1 qui doivent être satisfaits. Prenons par exemple le profil 4 présenté dans la section précédente, deux paramètres k=6 et L=2 et une requête initiale qui veut retrouver les séjours de trois jours à Madrid avec un départ le 10/11/2004:

```
SELECT V.idV
FROM VOYAGE V, DEPART D
WHERE        V.idV = D.idV AND D.date = '10/11/2004' AND
      V.lieu_arrivée = 'Madrid' AND V.nombre_jours = 3
```

Dans ce cas, sur les sept critères contenus dans le profil de l'exemple 4, seul celui qui exprime une préférence pour les voyages d'une durée de deux jours (VOYAGE.nombre_jours = 2 0.7) est conflictuel avec la requête et ne sera pas choisi. Les six autres sont intégrés dans la requête initiale qui devient :

```
SELECT V.idV
FROM VOYAGE V, DEPART D, HOTEL H, TRANSPORT T
WHERE        V.idV = D.idV AND T.idT = V.idT AND H.idH = V.idH AND
      D.date = '10/11/2004' AND V.lieu_arrivée = 'Madrid' AND
      V.nombre_jours = 3 AND V.lieu_départ = 'Paris' AND
      ( (H.région = 'centre ville' AND T.moyen = 'train') OR
        (H.région = 'centre ville' AND T.moyen = 'car') )
```

Les critères avec un poids égal à 1 (a, b, c) sont ajoutés à la requête comme des clauses conjonctives (T.idT = V.idT, H.idH = V.idH et V.lieu_départ = 'Paris'); les autres sont représentés par une disjonction de conjonctions ((H.région = 'centre ville' AND T.moyen = 'train') OR (H.région = 'centre ville' AND T.moyen = 'car')). On remarque qu'il reste une variante combinatoire des critères restants qui est T.moyen = 'train' et T.moyen = 'car'. Comme les deux prédicats sont contradictoires, ils ne sont pas inclus dans la requête.

Le contenu du profil d'un utilisateur peut être utilisé à différents moments du cycle de recherche d'informations. Il peut jouer le rôle de substituant de la requête initiale et donc devenir lui-même la requête à exécuter, il peut servir pour enrichir le contenu de la requête ou encore être utilisé pour adapter les résultats selon les modalités de présentation. En règle générale, il peut aussi influer sur l'exécution de la requête sans nécessairement intervenir dans l'expression de celle-ci.

2.3. *Langages de requêtes ouverts aux préférences définies dans les profils*

Si l'on veut faire une nette distinction entre la notion de profil et la notion de requête, il faudrait que les langages de requêtes soient à même d'intégrer certaines informations et préférences de profils. Depuis quelques années, les systèmes de bases de données proposent des langages de requêtes avancées dont certains concepts sont très appropriés à la prise en compte des informations de profil. Il en est ainsi des opérateurs approximatifs de type 'best of' ou 'arround', de poids attachés aux prédicats ou d'ordre imposé aux prédicats. Ces travaux puisent leurs sources dans les recherches sur les bases de données floues (Bosc et al., 1995), (Liu et al., 2001) et l'exécution flexible de requêtes (Parachar et al., 2004). Ainsi, une extension intéressante de SQL (PREFERENCE-SQL) permet d'exprimer bon nombre de préférences et est de ce fait très adaptée à la personnalisation de l'information. Pour illustrer les concepts mis en avant par ces langages, nous prenons trois exemples:

– *Introduction d'opérateurs dépendant du contexte*: Dans certaines requêtes, il est souvent intéressant d'exprimer des critères approximatifs du type 'environ 80m2, pas trop cher, bonne exposition, proche de Versailles'. Ce type de requête est particulièrement adapté à l'usage de profils dans la mesure où la sémantique des termes utilisés dépend de chaque utilisateur. Disposant de profils décrivant la sémantique de ces termes, en fixant par exemple des distances ou des intervalles, les systèmes offrant ces langages peuvent permettre une grande flexibilité dans l'accès aux données. La même requête émise par deux utilisateurs aux profils différents aura des réponses différentes. Certains systèmes comme PREFERENCE-SQL (Kießling, 2002) ne vont pas jusque là, ils donnent une interprétation spécifique à chaque clause et l'exécution d'une requête se fera de la même façon quel que soit l'utilisateur qui l'a émise. Mais la multiplication de tels systèmes conduira à une paramétrisation de ces sémantiques, sans doute dans les profils utilisateurs, pour accepter leur interopérabilité.

– *Introduction de préférences dans les critères de sélection*: Tous les résultats d'une requête n'ont pas la même importance pour l'utilisateur. La recherche des meilleurs résultats selon les valeurs des attributs sélectionnés, peut être réalisée en utilisant des formules de préférence (opérateur Winnow (Best) (Chomicki, 2002), opérateur Skyline (Borzsonyi et al., 2001) ou par un ordonnancement des valeurs préférées (préférences : POS, NEG, POS/NEG, POS/POS, EXPLICIT (Kießling, 2002)). Prenons par exemple une vue VOYAGE(Destination,Transport). La préférence des voyages en train par rapport à ceux en car peut être définie par l'expression : (d, train) > (d, car) pour une destination d. La même préférence peut

être exprimée en utilisant les ensembles de valeurs préférées: (Transport, POS1-set{train}, POS2-set{car}). Ici les meilleures valeurs pour l'attribut Transport sont celles contenues dans le premier ensemble positif (POS1-set). L'ensemble POS2-set contient les valeurs du même attribut qui sont moins préférées que celles du premier ensemble positif, mais plus pertinentes que toutes les autres valeurs qui ne figurent dans aucun des deux ensembles.

– *Introduction d'un ordre partiel entre les critères*: Certains langages permettent de spécifier une hiérarchie de préférences établissant un ordre partiel entre les critères de sélection. Dans PEFERENCE-SQL, cet ordre partiel est défini à l'aide de deux clauses complémentaires: la clause PREFERING qui fixe les choix initiaux et la clause CASCADING qui fixe les choix secondaires. Par exemple, si l'utilisateur de l'exemple précédent préfère descendre dans les hôtels 3 étoiles lors de ses voyages, en privilégiant ceux du centre ville, sa requête s'exprime de la façon suivante :

```
SELECT * FROM DESTINATION D, HOTEL H
WHERE V.idH = H.idH AND prix ARROUND 600
      PREFERING H.nombre_étoiles = 3   CASCADING H.région = 'centre ville'
```

La clause WHERE est exécutée sans l'influence des préférences. Ensuite la clause PREFERING est appliquée, suivie des clauses CASCADING qui sont exécutées les unes après les autres. Si, en appliquant une clause, le résultat devient nul, son effet est annulée et elle n'est pas prise en compte. Dans notre exemple, si plusieurs tuples satisfont les critères de la requête définis dans la clause WHERE, ceux qui contiennent des hôtels de trois étoiles sont retenus et finalement parmi eux, on choisit les tuples contenant des hôtels au centre ville.

Bien que la notion de profil n'existe pas dans les approches d'extension des langages de requêtes, ces langages ouvrent la voie à la personnalisation de l'accès aux données. Une partie du profil utilisateur (principalement le centre d'intérêt) peut être automatiquement intégrée aux requêtes par un processus de réécriture et ainsi simplifier l'expression des requêtes par l'utilisateur. L'enrichissement progressif des langages de requêtes permet à l'avenir d'en faire aussi des langages de définition de profils.

3. Vers une représentation multidimensionnelle d'un profil

Comme nous l'avons vu précédemment à travers quelques exemples, le profil d'un utilisateur peut contenir différents types d'informations: des données démographiques, la description d'un centre d'intérêt, des statistiques de comportement, des préférences définies par des fonctions d'utilité ou des ordres partiels, ou enfin des modalités de présentation des informations recherchées. Nous avons vu aussi comment ces informations de profil peuvent être utilisées pour enrichir les requêtes ultérieures de l'utilisateur, guider le système dans l'évaluation de ces requêtes, inférer des procédures d'adaptation du système à l'environnement de l'utilisateur ou à son niveau d'expertise.

Le projet APMD (Accès Personnalisé à des Masses de Données) a pour objectif d'étudier de façon systématique et aussi générique que possible les différents types d'informations constituant un profil ainsi que les techniques sous-jacentes à leur exploitation aussi bien dans un environnement de RI que de BD. Pour supporter cette étude, il est indispensable de disposer d'une infrastructure générique de définition et de gestion de profils. C'est l'objectif de cette section.

La définition d'un profil peut se faire de deux façons différentes, éventuellement complémentaires: par instanciation d'un modèle générique de profils ou par spécification au moyen d'un langage de description de profils. La première approche est plus simple, elle consiste en une identification et une modélisation conceptuelle des différents attributs de profils. Elle a l'avantage d'une implémentation rapide mais a l'inconvénient d'être bornée par les seuls types d'informations prévus. La seconde approche est plus ouverte mais il est plus difficile de trouver un langage uniforme en raison de la grande diversité des informations constituant un profil. En tout état de cause, même l'approche langage nécessite une identification des types d'information constituant un profil; c'est la raison pour laquelle nous abordons la première approche en priorité dans cette section. Elle sera complétée par diverses opérations de gestion de profils.

3.1. *Modèle générique de profils*

La classification, l'organisation et la structuration des données de profils est un élément clé si on veut avoir une vision globale de la personnalisation. Différents travaux ont abordé cet aspect sans le couvrir dans son ensemble. Par exemple, P3P (W3)[1], standard pour la sécurisation des profils, permet de définir des classes distinguant entre les *attributs démographiques* (identité, données personnelles), *les attributs professionnels* (employeur, adresse, type) et *les attributs de comportement* (trace de navigation. Dans (Amato et al., 1999), les auteurs proposent un modèle de profil pour les utilisateurs d'une bibliothèque digitale composé de cinq catégories d'informations : *Données Personnelles* (identité), *Données Collectées* (contenu, structure et provenance des documents), *Données de Livraison* (moment et moyen de livraison), *Données de Comportement* (interactions de l'utilisateur avec le système), *Données de Sécurité* (conditions d'accès aux informations du profil). Ces tentatives de structuration sont louables mais insuffisantes pour couvrir le champ de la personnalisation. Par ailleurs, elles se contentent de catégoriser les informations de profil, mais sont difficilement extensibles.

Poursuivant la classification de (Amato et al., 1999), notre objectif est de proposer un ensemble de dimensions ouvertes, capables d'accueillir la plupart des informations caractérisant un profil. La figure 1 donne un aperçu de la structure générique d'un profil. Chaque dimension est constituée d'un ensemble d'attributs dont les valeurs peuvent être simples (valeur numérique ou symbolique) ou complexes (expression logique, fonction d'utilité ou ordre de préférence par

[1] W3, Platform for Privacy Preferences (P3P) project, http://www.w3.org/P3P/

exemple). Certaines dimensions sont organisées en sous-dimensions selon la nature de leurs attributs. Un attribut du profil est défini par un nom, un type, une expression de préférence et une sémantique.

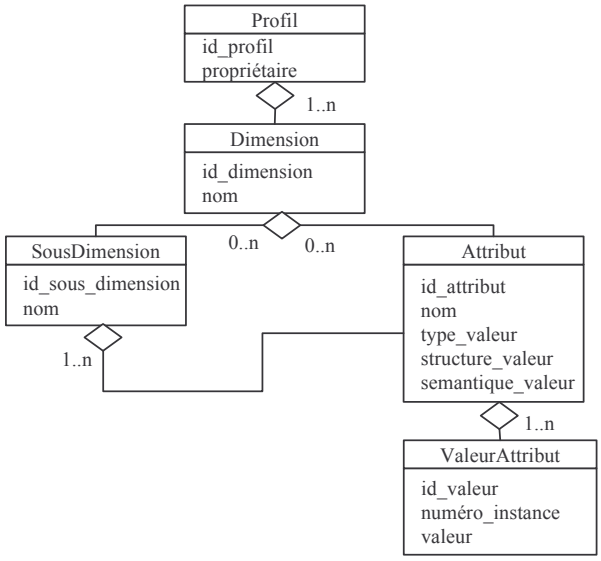

Figure 1. *Modèle conceptuel d'un profil*

– *Le type* peut être l'un des types couramment utilisés: entier, réel, chaîne de caractères, intervalle, ensemble, etc.,

– *L'expression de préférence* peut être de plusieurs natures: un vecteur de termes pondérés, une expression logique (requête), une fonction d'utilité, un ordre partiel entre un ensemble de termes ou une composition de tous ces éléments. Un langage de description des préférences est défini selon la nature des préférences.

– *La sémantique* permet de définir le concept représenté par l'attribut lorsque son sens est lié au contexte d'utilisation. Par exemple, la notion de meilleure fraîcheur des données est définie comme étant la valeur minimale alors que la notion de meilleure fiabilité des données est définie comme étant la valeur maximale. Par ailleurs, les expressions de préférences d'un attribut peuvent utiliser des opérateurs dont la définition est également dépendante de l'utilisateur ou du contexte. C'est le cas des opérateurs souvent utilisés comme 'environ', 'autour de', 'proche de', 'optimiste', 'pessimiste', etc. La sémantique d'un attribut peut être définie à l'aide d'une fonction, d'une ou plusieurs règles ou d'un graphe conceptuel (une ontologie).

Nous distinguons principalement six dimensions dans la définition d'un profil : Données personnelles, Centre d'intérêt, Qualité attendue, Préférences de livraison, Sécurité et Historique des interactions de l'utilisateur.

Les données personnelles

Les données personnelles sont la partie statique du profil. Elles comprennent l'identité de l'utilisateur (nom, prénom, numéro de sécurité sociale, etc.), des données démographiques (âge, genre, adresse, situation familiale, nombre d'enfants, etc.), les contacts personnels et professionnels et d'autres informations comme le numéro de la carte bancaire ou de la carte Vitale. Les données personnelles sont relativement stables dans le temps et ne demandent pas de mise à jour automatique par le gestionnaire de profils. Dans plusieurs approches ces données ne jouent pas un rôle dans le processus de recherche d'information, mais servent souvent comme monnaie d'échange contre les services de personnalisation ou les services d'accès. C'est le cas des systèmes de vente sur Internet qui demandent aux utilisateurs de fournir des formulaires de données personnelles qui sont ensuite utilisées pour faire des statistiques pour mieux cibler les clients. Mais dans certains cas, il est possible de dériver des données personnelles des informations qui compléteront le profil. Par exemple, à partir de la date de naissance, on peut déduire le signe astral (et proposer ensuite un horoscope); à partir de l'adresse, on peut déduire les services culturels et sportifs de la région et les utiliser lors des requêtes sur ces thématiques, etc.

Le centre d'intérêt

Le centre d'intérêt exprime le domaine d'expertise de l'utilisateur ou son périmètre d'exploration. Il peut être défini par un ensemble de mots clés (concepts) ou un ensemble d'expressions logiques (requêtes). Dans de nombreuses approches, l'importance de chaque concept est définie par une pondération des mots clés du centre d'intérêt. L'ontologie du domaine complète la définition du centre d'intérêt en explicitant la sémantique de certains termes. Par exemple, on peut explicitement définir que 'BD' signifie 'bases de données' dans le profil et non 'bande dessinée', que dans le contexte du profil 'client' et 'consommateur' sont synonymes. Le centre d'intérêt peut être vu comme une présélection virtuelle qui réduit la masse d'informations à prendre en compte. On peut rapprocher le centre d'intérêt de la notion de vue en BD. Par conséquent toute requête émise par l'utilisateur sera enrichie avec les mots clés ou les prédicats des requêtes définissant le centre d'intérêt. Le centre d'intérêt peut être corrélé avec les données personnelles et s'enrichir par déduction de certaines informations comme nous l'avons vu précédemment.

La qualité attendue

La qualité est un des facteurs clés de la personnalisation ; elle permet d'exprimer des préférences extrinsèques sur l'origine de l'information, sa précision, sa fraîcheur, sa durée de validité, le temps nécessaire pour la produire ou la crédibilité de sa source. Les attributs de cette dimension expriment la qualité attendue ou

espérée; elle sera confrontée à la qualité effective produite par le système de recherche d'informations. Il faut noter que la qualité d'un produit informationnel ne se mesure pas toujours sur le produit lui-même, mais quelquefois sur sa source de production ou son processus de production. La qualité d'une information factuelle (précision de l'adresse d'une personne) ne se mesure pas comme la qualité d'un agrégat statistique (moyenne des revenus des français). Le mode de production et les opérandes de la production peuvent êtres déterminants dans ce dernier cas. La qualité attendue peut donc être définie par des expressions de préférences sur un facteur de qualité donné. Par exemple l'utilisateur peut exprimer sa préférence de la source S1 devant la source S2 en terme de fraîcheur. Ceci se traduit par l'ajout dans son profil d'une sous-dimension '*ExpressionQualité*' où le type d'expression est *préférence de source*, l'attribut de qualité est *fraîcheur* et l'expression de préférence est *S1>>S2*.

Les préférences de livraison

Les préférences de livraison concernent d'abord tout ce qui est lié aux modalités de présentation des résultats en fonction de la plateforme (média de livraison), de la nature et du volume des informations délivrées, des préférences esthétiques ou visuelles de l'utilisateur. A ces modalités de présentation, on peut ajouter les modalités d'exécution, décrivant le moment d'exécution d'une requête (mode pull ou push), la manière de notifier les résultats (différé, immédiat par exemple) et la quantité de résultats que l'on souhaite recevoir (les Top k, les premiers calculés, l'ensemble des résultats mais à la fin du calcul, etc.).

La sécurité

La sécurité est une dimension fondamentale du profil. Elle peut concerner les données que l'on interroge ou modifie, les informations que l'on calcule, les requêtes utilisateurs elles-mêmes ou les autres dimensions du profil. La sécurité des données peut être exprimée par des niveaux de sécurité prédéfinis qui dépendent de la hiérarchie des vues autorisées, par des clauses d'octroi ou de révocation de droits à la SQL, par le support d'identification ou de stockage (carte à puce, certificat web etc.) et par des moyens de transmission utilisés (protocoles, cryptage). Les niveaux de sécurité peuvent concerner différents types d' «objets» comme les catégories du profil, les résultats des requêtes, mais aussi un processus de traitement ou une fonction de calcul. La sécurité du processus exprime la volonté de l'utilisateur de cacher un traitement qu'il effectue. Ceci peut être fait en définissant le degré de visibilité de certaines opérations.

L'historique des interactions de l'utilisateur

Cette dimension contient entre autre ce qu'on appelle communément le 'feedback' de l'utilisateur. Elle regroupe l'ensemble des informations collectées sur le comportement de l'utilisateur, que ces informations soient directement fournies par lui (feedback explicite) ou qu'elles soient récupérées ou dérivées à son insu (feedback implicite). Il peut s'agir par exemple de l'exclusion des résultats qu'il

n'aime pas, du nombre de clicks qu'il a effectué sur le lien d'une page ou du nombre et de la nature des requêtes qu'il a émises. Les informations de l'historique des requêtes de l'utilisateur ne sont pas utilisées directement dans le processus de personnalisation, mais sont analysées afin d'extraire les raisons de ce comportement qui sont ensuite soit ajoutées aux profils (dans le cas où elle ne sont pas présentes), soit utilisées pour la mise à jour des valeurs de profil.

On trouvera dans (Kostadinov 2003) une description plus détaillée des attributs de profil que nous avons recensés à travers notre état de l'art.

3.2. *Opérations sur les profils*

La gestion de profils consiste en plusieurs opérations:

– la création d'un profil qui peut se faire de façon interactive, par apprentissage à partir des actions passées ou par importation,

– l'évolution d'un profil qui maintiendra le profil toujours conforme à l'environnement de son exploitation,

– l'importation d'un profil impliquant un test de validité par rapport au modèle générique (validité sémantique),

– l'intégration de plusieurs profils pour former un nouveau profil d'utilisateur ou de communauté d'utilisateurs,

– la validation d'un profil par rapport à l'environnement technique dans lequel il serait utilisé (validité opérationnelle),

– l'appariement de profils qui est une opération fondamentale utilisée par la plupart des opérations précédentes.

La description détaillée de ces opérations est au delà des limites de cet article; nous allons nous focaliser uniquement sur l'opération d'appariement en raison de son caractère fondamental pour les autres opérations. Cette opération compare deux profils pour s'assurer de leur équivalence stricte, isoler leurs parties communes ou énumérer leurs différences. Cette opération est donc constituée de trois primitives distinctes qui s'appliquent aussi bien aux types des profils qu'aux valeurs des profils:

– le test d'équivalence de deux profils (*Equivalence* (P_1, P_2)),

– le calcul de l'intersection de deux profils (*Match* (P_1, P_2)),

– le calcul des différences entre deux profils (*MisMatch* (P_1, P_2))

Le tableau de la figure 2 résume de façon informelle les définitions de ces primitives. P_i représente un profil i, P_i*T_i représente le type du profil i (structure arborescente des dimensions, sous-dimensions et attributs d'un profil), P_i*V_i représente la valeur du profil i (Pi*T_i avec des valeurs associées aux attributs).

	Types	Valeurs
Équivalence	Équivalence(P_1*T_1,P_2*T_2)→ bool Deux profils ont même type s'ils ont les mêmes dimensions, sous-dimensions et attributs	Équivalence(P_1*V_1,P_2*V_2)→bool Deux profils sont équivalents par leurs valeurs s'ils ont même type et si tous leurs attributs possèdent deux à deux les mêmes valeurs
Match	Match (P_1*T_1, P_2*T_2) → P_3 rend un profil P_3 dont les dimensions, sous-dimensions et attributs appartiennent à la fois à P_1*T_1 et à P_2*T_2.	Match (P_1*V_1, P_2*V_2) → P_4 restriction du profil P_3 aux seuls attributs ayant deux à deux les mêmes valeurs dans P_1*V_1 et P_2*V_2.
MisMatch	MisMatch (P_1*T_1, P_2*T_2) → P_5 rend un profil P_5 = (P_1 U P_2) – Match (P_1*T_1, P_2*T_2)	MisMatch (P_1*V_1, P_2*V_2) → P_6 rend un profil P6 contenant les attributs de P1 qui appartiennent aussi à P3 et qui ont des valeurs disjointes dans P_1*V_1 et P_2*V_2

Figure 2. *Tableau récapitulatif des opérateurs de manipulation de profils*

4. Architecture et fonctionnalités de la plateforme de gestion de profils

L'objet de la plateforme de gestion de profils est de supporter le modèle générique présenté dans la section précédente ainsi que l'ensemble des opérations définies sur ce modèle. Cette plateforme servira pour créer un catalogue de profils qui seront utilisés dans différents scénarios, soit pour mesurer la pertinence d'un système par rapport à un échantillon de profils, soit pour comparer certains systèmes sur la base de profils types, soit pour créer des scénarios de personnalisation en composant plusieurs profils complémentaires. Dans le projet APMD[2], cette plateforme peut servir d'environnement intégrateur des outils développés dans le projet.

La plateforme de gestion de profils implémente le modèle de profil générique et les opérateurs d'appariement de profils définis dans la section précédente. Son architecture générale est donnée dans le contexte architectural du projet APMD (Figure 3).

Le modèle générique de profils ainsi que ses instances sont implémentés dans une base de données Oracle. La connexion entre la base de données et son

[2] Site web du projet APMD: Accès Personnalisé à des Masses de Données, ACI Masses de Données, http://apmd.prism.uvsq.fr/

environnement applicatif est faite, au choix, par une API SQL pour une vision relationnelle de la base ou une API XQuery pour une vision XML de la base. Le gestionnaire de profils est constitué de l'ensemble des opérations définies dans la section précédente. Une interface graphique permet de créer et de manipuler manuellement les profils.

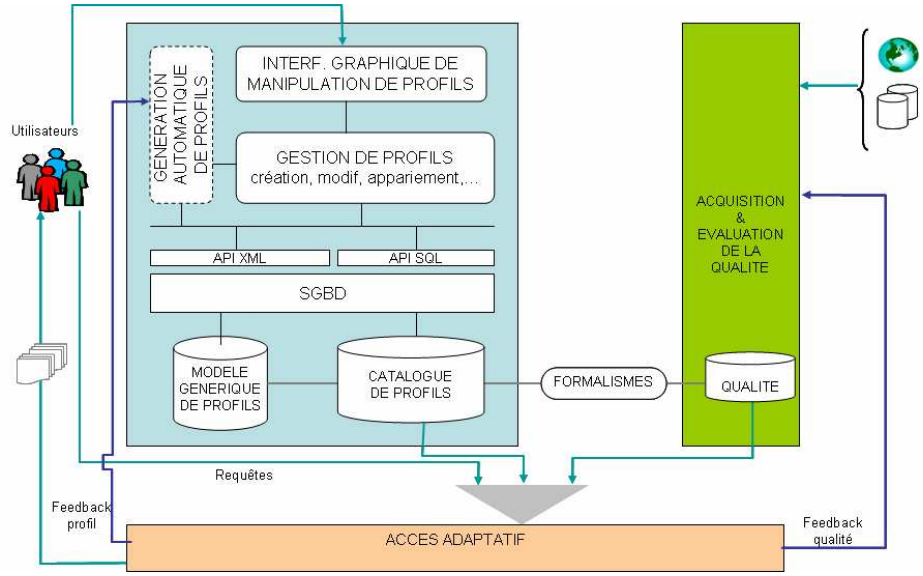

Figure 3. *Détail du gestionnaire de profils dans l'architecture APMD*

La définition d'un type de profil pour un utilisateur particulier se fait par recopie de tout ou d'une partie du modèle générique. Par conséquent, la création d'un profil utilisateur est faite en deux étapes : (i) choix de la structure du profil et (ii) attribution de valeurs aux attributs du profil. La sélection de la structure d'un profil peut être faite de trois manières différentes : recopie de l'ensemble du modèle générique, choix de la même structure qu'une instance de profil déjà existante ou sélection d'une structure personnalisée à partir du modèle générique. La deuxième étape de la création d'un profil revient à attribuer des valeurs aux attributs. Ceci peut être fait par le module de génération automatique, par apprentissage ou datamining par exemple, ou manuellement par l'utilisateur via l'interface graphique. La construction automatique d'un profil est un processus complexe qui dépasse le cadre de cet article et ne fera pas l'objet d'une discussion.

Le gestionnaire de profils offre des fonctionnalités de mise à jour classiques qui sont : l'insertion, la suppression et la modification de la structure et des valeurs des profils. La seule règle à respecter est que le modèle de profil générique doit inclure

l'union des structures de toutes les instances de profils. Pour maintenir le profil générique le plus complet possible et pour préserver la typologie des données (le nom d'un utilisateur est une chaîne de caractères dans l'ensemble des instances de profils utilisateurs), chaque insertion d'une nouvelle dimension, sous-dimension ou attribut dans une instance de profil se fait par le modèle générique. Par exemple si un utilisateur veut ajouter un attribut « *revenu* » à ces données personnelles, il est obligé de créer cet attribut dans le modèle générique de profil utilisateur pour ensuite le recopier dans son profil.

Le gestionnaire de profils présenté dans cette section permet la manipulation manuelle des profils. Cependant un des objectifs de la personnalisation est de faire intervenir l'utilisateur le moins souvent possible. C'est la raison pour laquelle le gestionnaire a été conçu comme une API qui permet son utilisation par un autre programme via les opérateurs et les fonctions définis dans cette API.

5. Conclusion

L'élaboration d'un modèle de profil générique est le premier pas vers la construction de systèmes de personnalisation capables de délivrer des informations selon les préférences d'un utilisateur. Nous avons défini un modèle générique de profil qui permet de structurer les informations nécessaires pour la description des préférences d'un utilisateur. Ce modèle est composé de six dimensions qui regroupent un grand nombre de paramètres nécessaires à divers scénarios de personnalisation. Il est complété par un ensemble d'opérations de gestion de profils qui serviront à une manipulation directe ou via des API des profils. Nous avons également présenté un prototype de gestion de profils qui implémente le modèle générique et les opérateurs associés. Cette plateforme va servir dans un premier temps à créer un catalogue de profils qui sera utilisé pour tester différents scénarios de personnalisation. Dans un second temps, elle peut constituer un cadre d'intégration des prototypes développés dans le cadre du projet APMD.

6. Bibliographie

Amato G., Straccia U., « User Profile Modeling and Applications to Digital Libraries », *In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*, Paris, France, 1999

Borzsonyi S., Kossmann D., Stocker K., « The Skyline Operator », *In Proceedings of the IEEE Conference on Data Engineering*, pages 421-430, Heidelberg, Germany, 2001

Bosc P., Pivert O., « SQLf : A Relational Database Language for Fuzzy Querying », *IEEE Transactions on Fuzzy Systems*, vol. 3, No 1, pp 1-17, 1995

Bradley K., Rafter R., Smyth B., « Case-Based User Profiling for Content Personalisation », *In: Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, Trento, Italy, August 2000

Cherniack M., Galvez Ed., Franklin M., Zdonik St., « Profile-Driven Cache Management », *In: Proceedings of the 19th International Conference on Data Engineering*, Bangalore, India, 2003

Chomicki J., « Querying with Intrinsic Preferences », *In: Proceeding of the 8th International Conference on Extending Database Technology*, Prague, Czech Republic, 2002

Croft W. B., Cronen-Townsend St., Lavrenko V., « Relevance Feedback and Personalization: A Language Modelling Perspective », *In: Proceedings of the Second DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries*, Dublin City University, Ireland, 18-20 June 2001

Crabtree B., Soltysiak S., « Automatic learning of user profiles-towards the personalisation of agent services », *BT Technol J.* Vol 16 No 3, July 1998

Dai H., Mobasher B., Luo T., Nakagawa M., « Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization », *Data Mining and Knowledge Discovery*, 61-82, 2002

Ferreira J., Silva A., « MySDI: A Generic Architecture to Develop SDI Personalised Services », *In: Proceedings of the 3rd International Conference on Enterprise Information Systems*, Setubal, Portugal, July 7-10, 2001

Gauch S., Pretschner A., « Ontology Based Personalized Search », *In: Proceeding of the 11th IEEE Intl. On Tools with Artificial Intelligence*, pp. 391-398, Chicago, November 1999

Kießling W., « Foundations of Preferences in Database Systems », *In: Proceedings of the 28th Conference on Very Large Data Bases*, Hong Kong, China, 2002

Koutrika G., Ioannidis Y., « Personalization of Queries in Database Systems », *In: Proceedings of the 20th International Conference on Data Engineering*, Boston, Massachusetts, USA, April, 2004

Kostadinov D., Personnalisation de l'information et gestion des profils utilisateurs, Rapport de DEA, Université de Versailles, France, 2003

Lieberman H., Shearin S., « Intelligent Profiling by Example », *In: Proceedings of the 2001 International Conference on Intelligent User Interfaces*, Santa Fe, USA, January 2001

Liu C., Yang Q., Zhang W., Wu J., Yu C., Nakajima H, Rishe N. D., « Efficient Processing of Nested Fuzzy SQL Queries in a Fuzzy Database », *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, No 6, November/December 2001

Parachar M., Schmidt C., « Enabling Flexible Queries with Guarantees in P2P Systems », *IEEE Internet Computing*, vol. 8, No 3, May/June 2004