Accès personnalisé aux informations : approche dirigée par la qualité

Mokrane Bouzeghoub[†], Sylvie Calabretto[‡], Nathalie Denos[§], Rami Harrathi[‡], Dimitre Kostadinov[†], An-Te Nguyen[¥], Verónika Peralta[†]

† *Laboratoire PRiSM, Université de Versailles*
45 avenue des Etats-unis, 78035 Versailles Cedex
{Mokrane.Bouzeghoub, Dimitre.Kostadinov, Veronika.Peralta}@prism.uvsq.fr

‡ *Laboratoire LIRIS, INSA de Lyon*
7, avenue Jean Capelle, 69621 Villeurbanne Cedex
{Sylvie.Calabretto, Rami.Harrathi}@insa-lyon.fr

§ *Laboratoire d'Informatique de Grenoble (ex-CLIPS-IMAG)*
385 avenue de la BibliothèqueDomaine Universitaire 38400 Saint-Martin-d'Hères
Nathalie.Denos@imag.fr

¥ *Université des Sciences Naturelles de HoChiMinh-Ville, Vietnam.*
nate@hcmuns.edu.vn

RÉSUMÉ. Une solution à l'amélioration de la pertinence de l'information est la personnalisation des réponses fournies aux utilisateurs, selon leurs profils ou leurs préférences. Ainsi les requêtes de l'utilisateur, avant d'être exécutées, sont reformulées en tenant compte des éléments de profil. Les préférences utilisées pour cette reformulation peuvent être de différentes natures ; en particulier, elles peuvent concerner la qualité de l'information délivrée à l'utilisateur. Dans cet article, nous analysons l'impact des facteurs de qualité dans la personnalisation de l'information, puis nous détaillons leur incorporation dans un méta modèle de profil. Nous montrons comment la qualité peut impacter le cycle de vie d'une requête personnalisée, en considérant en particulier trois facteurs de qualité : la fraîcheur, l'exactitude et la popularité.

ABSTRACT. Data personalisation is one of the main solutions to improve relevance of data in information retrieval systems and database systems. Before being executed, user queries are reformulated on the basis of user profile preferences. These preferences may be specified on different knowledge, in particular on quality factors which characterize data delivered to the user. In this paper, we address the issue of quality factors and analyse their impact on the query execution life cycle. We illustrate this impact through three use-cases, respectively based on three quality factors: freshness, accuracy and popularity.

MOTS-CLÉS : personnalisation, qualité des données, facteurs de qualité, profil utilisateur, accès à l'information, cycle de vie d'une requête

KEYWORDS: personalization, data quality, quality factors, user profile, information access, query life cycle

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

1. Introduction

De nos jours, de nombreux systèmes fournissent un accès à des sources de données multiples, distribuées, autonomes et hétérogènes, disponibles sur Internet ou au sein d'une organisation. Les Moteurs de Recherche d'Informations, les Systèmes de Filtrage ou de Recommandations, les Systèmes de Médiation, les Entrepôts de Données, les Systèmes Pair à Pair et les Portails Web sont des exemples de ce type de systèmes. La notion de requête utilisateur prend ici un sens très large qui englobe à la fois la recherche ou le filtrage d'items au contenu faiblement structurés (données, documents, objets, etc.), qui relèvent du domaine de la Recherche d'Informations (RI), et l'interrogation de données structurées, qui relève du domaine des Bases de Données (BD). La quantité de données auxquelles ces systèmes ont accès ne cesse d'augmenter ce qui entraîne une croissance du volume de résultats que les utilisateurs se voient offrir en réponse à leurs requêtes. De même la qualité des données obtenues constitue un enjeu considérable, en particulier pour les décideurs qui, selon l'importance de leurs décisions, peuvent exiger des niveaux de qualité plus ou moins élevés.

La personnalisation est une réponse efficace au problème de la surcharge d'informations, d'une part, en injectant dans les requêtes des critères de filtrage plus restrictifs, et, d'autre part, en reformulant éventuellement les requêtes pour mieux tenir compte des centres d'intérêt et des préférences de l'utilisateur. Une sélection plus restrictive des données permet de limiter la surcharge d'informations et une reformulation de certaines parties de la requête permet de mieux cibler les résultats. Par exemple, la requête d'un utilisateur qui cherche des livres sur « Java » peut être personnalisée en sachant qu'il s'intéresse à la *programmation* et non à la *géographie* (centre d'intérêt), qu'il préfère lire *en anglais* (préférence de langue) et qu'il cherche les éditions *les plus récentes* avec une *bonne critique* (préférences de qualité). L'ensemble des données décrivant l'utilisateur et ses préférences sont souvent regroupées et stockées dans un profil utilisateur. Dans un tel contexte, la requête utilisateur exprime une demande d'information particulière tandis que le profil représente la partie relativement stable des besoins de l'utilisateur qui doit être prise en compte lors de l'évaluation de la requête.

Dans le présent article nous nous intéressons plus particulièrement aux « facteurs de qualité » qui influent sur la personnalisation de l'accès à l'information. Nous considérons la notion de qualité à la fois dans son sens objectif (précision ou complétude des données par exemple) mais aussi dans sa dimension subjective (réputation d'une source ou popularité des données par exemple). La personnalisation intervient aussi à deux niveaux : (i) un niveau de préférences explicites formulées directement dans le profil ou la requête utilisateur et (ii) un niveau de préférences implicites dérivées du comportement passé de l'utilisateur.

Dans cet article, notre premier objectif est de fixer et organiser les facteurs de qualité parmi les nombreux attributs de personnalisation. Après un résumé de l'état de l'art de l'utilisation des facteurs de qualité dans la personnalisation, nous montrons comment ces facteurs sont intégrés dans le méta modèle de profil proposé dans le projet Accès Personnalisé à des Masses de Données (APMD) (Bouzeghoub *et al.*, 2005). Notre second objectif est d'illustrer l'impact de la qualité sur le cycle de vie d'une

requête personnalisée. Enfin, nous montrons à travers quelques exemples de réalisation dans le projet, la mise en oeuvre d'un modèle de personnalisation basé sur la qualité.

Le reste du document est organisé de la façon suivante : La section 2 présente un état de l'art sur l'usage des facteurs de qualité influant sur la personnalisation. Ensuite, la section 3 présente notre référentiel de profil en focalisant sur la partie décrivant la qualité, et la section 4 présente l'impact de la qualité sur le cycle de vie d'une requête. Finalement, la section 5 présente les conclusions de notre réalisation.

2. Etat de l'art sur l'usage des facteurs de qualité pour la personnalisation

La qualité des données est un domaine de recherche qui a suscité depuis longtemps un vif intérêt (Strong *et al*., 1997 ; Jarke *et al*., 1997 ; Gertz *et al*., 2004) mais qui prend une dimension cruciale ces dernières années en raison de la multiplicité des sources de données, de leur hétérogénéité et de leur évolutivité de plus en plus accélérée. Dans cette section, nous présentons un aperçu des travaux de modélisation et évaluation de la qualité des données ainsi que nos résultats de recherche passés obtenus dans ce domaine et nous focalisons sur la prise en compte de la qualité pour la personnalisation.

2.1. Bref rappel sur la qualité de l'information

La qualité de l'information a fait l'objet de nombreuses études mettant en évidence différents facteurs de qualité dont les définitions sont souvent nuancées par l'usage de termes aussi variés que catégorie, dimension, attribut, critère, mesure, paramètre, etc. S'il est souvent possible de rapprocher certaines notions, la standardisation des concepts et du vocabulaire reste à faire. Outre la définition des concepts, l'évaluation de la qualité constitue un des principaux problèmes de recherche; elle concerne la spécification et l'implémentation des procédures de mesure des différents facteurs caractérisant les sources de données ainsi que la combinaison de ces mesures pour obtenir une valeur agrégée interprétable par l'utilisateur.

Certaines approches se focalisent sur la définition de facteurs, par exemple (Redman, 1996 ; Wang *et al*., 1996). Par contre, les facteurs définis sont pertinents uniquement dans le domaine pour lequel ils ont été conçues, laissant relativement peu de possibilités de réutilisation ou de définition d'un sous-ensemble cohérent adapté à une application donnée. Certains travaux analysent et classifient les techniques de mesure, les métriques et les fonctions d'agrégation (Pipino *et al*., 2002 ; Naumann *et al*., 2000 ; Ballou *et al*., 1998). D'autres travaux montrent leur usage dans des contextes applicatifs précis comme, par exemple, la *fraîcheur* des données dans les systèmes de caching (Cho *et al*., 2000).

Quand nous interrogeons des sources de données multiples et hétérogènes, il est nécessaire de combiner plusieurs valeurs d'un facteur de qualité (mesurées pour les différentes sources) pour obtenir une valeur caractéristique du résultat de la requête. Par exemple, dans une requête qui intègre des données de différentes sources, la *disponibilité* des résultats peut être calculée comme une fonction de la disponibilité des

différentes sources. Malgré l'existence de techniques pour mesurer certains facteurs de qualité, la combinaison de valeurs a été peu traitée dans la littérature. Dans (Naumann *et al*., 1999), les auteurs proposent de combiner les valeurs de qualité des sources via des opérateurs arithmétiques (minimum, maximum, moyenne, somme, produit). La nécessité d'une algèbre pour la combinaison de valeurs de qualité a été soulignée également dans (Gertz *et al*., 2004). D'autres travaux analysent la combinaison de valeurs de qualité pour des facteurs spécifiques, par exemple l'*exactitude* et la *complétude* des données (Motro *et al*., 1998).

Aucun de ces travaux ne considère ni l'influence du contexte d'application ni les préférences des utilisateurs dans la combinaison. En effet, d'autres paramètres du contexte d'application, comme la nature des données, le domaine d'application, l'architecture du système, etc., peuvent influencer la qualité des résultats (Peralta, 2006). Par exemple, les coûts de traitement des données influencent le temps de réponse et la fraîcheur des données. De plus, différents utilisateurs peuvent avoir besoin de différentes stratégies de combinaison de valeurs de qualité. Par exemple, un utilisateur peut exiger que toutes les données soient exactes alors qu'un autre peut tolérer un certain taux d'erreur sur certains attributs (calcul d'une moyenne de l'exactitude).

2.2. Notre expérience de recherche sur la qualité

Cette section résume les travaux de recherche que nous avons menés dans le passé. Ces travaux ne s'inscrivaient pas nécessairement dans le cadre de la personnalisation de l'information. Certains de ces travaux sont exploratoires; d'autres ont conduit à la réalisation d'algorithmes et d'outils d'évaluation et d'amélioration de la qualité.

Les premiers travaux réalisés concernent l'évaluation de la qualité des publications académiques, en particulier la *qualité scientifique*, la *lisibilité*, la *fraîcheur* des données contenues, le *degré d'autorité*, la *disponibilité*, la *popularité* et la *qualité de l'identification*. Ils ont donné lieu à l'élaboration d'un modèle exhaustif (Denos, 2000a) qui a été mis en regard d'un contexte applicatif précis (archives de prépublication arXiv) pour établir la faisabilité pratique de l'exploitation de ces facteurs de qualité (Denos, 2000b). Suite à cette mise en œuvre, le retour des utilisateurs a été collecté, analysé et pris en compte pour affiner le modèle opérationnel de qualité pour cette application (Denos, 2002).

Dans (Bouzeghoub *et al*., 2004) et (Peralta, 2006), nous avons développé un canevas qui fournit un support formel pour l'évaluation de la qualité des données, en particulier la *fraîcheur* et l'*exactitude* des données. Ce canevas permet d'analyser les différentes définitions et métriques de ces facteurs, d'analyser les paramètres du système d'information qui les influencent, et de développer des algorithmes d'évaluation qui tiennent compte de ces paramètres. Nous avons également proposé des techniques d'amélioration de la fraîcheur et de l'exactitude lorsque les exigences utilisateur ne sont pas satisfaites.

Dans (Calabretto *et al*., 1998), nous avons défini les critères de qualité d'un document en nous appuyant sur les critères de qualité des données. Ces critères sont *identifiabilité*, *facilité d'exploitation*, *crédibilité*, *traçabilité*, *compréhensibilité*,

réutilisabilité, *portabilité* et *flexibilité*. Dans (Harrathi *et al*., 2006), nous avons proposé un modèle de qualité hiérarchique orienté utilisateur. Cette hiérarchie compte trois niveaux : (1) *niveau système* qui décrit les métriques de qualité calculables au niveau du système ; (2) *niveau utilisateur* qui décrit les facteurs de qualité désirées par l'utilisateur qui sont calculés ou agrégés à partir des métriques système ; (3) *niveau source de la qualité* qui classifie ces facteurs dans 3 catégories (sources, support et usage). Par exemple, la *fréquence de mise à jour* (métrique système) s'utilise pour calculer la *fraîcheur* (facteur utilisateur) qui est un facteur de la catégorie *support*.

2.3. Exploitation de la qualité dans la personnalisation

Dans les travaux sur l'accès personnalisé aux données, la qualité de l'information n'a pas été prise en compte de façon explicite. Les facteurs de qualité sont utilisés comme des attributs dans la description des données et les préférences des utilisateurs sur ces attributs de qualité s'expriment sous forme de poids ou de seuils. Dans la suite nous présentons quelques travaux qui intègrent la qualité dans certaines activités de personnalisation : sélection de sources de données, sélection de plans d'exécution, ré-ordonnancement des résultats, filtrage des résultats.

2.3.1 Sélection de sources de données

Quand plusieurs sources de données fournissent les mêmes types de données, la sélection de sources consiste à choisir les sources qui fournissent les « meilleures » données afin de répondre à une requête utilisateur. La qualité des données a été utilisée comme critère de sélection dans plusieurs approches qui proposent la construction d'un score multicritère pour ordonner les sources de données (Naumann, 1998). Plusieurs méthodes de sélection multicritères sont comparées : (i) la méthode SAW (Simple Additive Weighting) propose la réalisation d'une somme pondéré en assignant différents poids aux différents critères ; (ii) la méthode AHP (Analytic Hierarchy Process) propose également un score mais assigne les poids en construisant une hiérarchie de critères et en définissent leurs importances relatives ; (iii) la méthode TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) propose de calculer la meilleure et la pire solution et de calculer leurs distances respectives par rapport aux sources candidates ; (iv) et la méthode DEA (Data Envelopment Analysis) construit une sorte de frontière (enveloppe d'efficience) en comparant les valeurs de qualité des sources : les sources sélectionnées sont celles appartenant à l'enveloppe. Dans tous ces travaux, les préférences de qualité se réduisent à la définition des poids pour les critères.

Dans (Mihaila *et al*., 2000), à chaque source est associé un ensemble de descripteurs de qualité qui décrivent le contenu, la complétude, la fraîcheur, la fréquence des mises à jour et la granularité d'une portion de la source. La sélection des sources de données est faite au travers de requêtes sur les métadonnées (ex. obtenir les 10 meilleures sources qui contiennent des données sur la météo à Paris avec une fraîcheur inférieure à 3 jours). Notons que les préférences de qualité font partie de la requête et sont données de façon explicite par l'utilisateur ; ceci s'explique par l'absence de la notion de profil utilisateur.

2.3.2 Sélection des plans d'exécution

Dans un contexte de BD relationnelles, les facteurs de qualité peuvent être exploités durant la compilation d'une requête pour choisir le meilleur plan d'exécution. Plus précisément, différents plans accédant à différentes sources sont comparés et ordonnés par rapport à la qualité des données. Dans (Naumann *et al*., 1999), des valeurs de différents facteurs de qualité sont évaluées et un score est agrégé pour chaque plan par le biais d'une somme pondérée. Le but consiste à choisir le plan (ou un sous-ensemble de plans) qui possède la meilleure qualité afin de l'exécuter. La notion de personnalisation est sous-jacente à l'exploitation de la qualité; elle se limite à donner des poids aux facteurs. Dans (Braumandl, 2003), l'auteur propose également la comparaison de différents plans d'exécution d'une requête (selon les sources accédées et les serveurs sur lesquels sont exécutées les requêtes) afin de choisir le plus adapté par rapport aux critères : *fraîcheur*, *complétude*, *taille du résultat*, *temps de réponse* et *coût* (au sens financier). La notion de personnalisation consiste à choisir des seuils pour ces critères ; les plans qui ne vérifient pas ces critères ne sont pas considérés.

2.3.3 Ré-ordonnancement des résultats d'une requête

Le principe du ré-ordonnancement des résultats d'une requête consiste à transformer l'ordre issu du calcul pour l'adapter à l'ordre préféré par l'utilisateur. Il s'agit d'un post traitement qui exploite les préférences de l'utilisateur. Par exemple, dans (Zhu *et al*., 2000), le score d'une page Web est calculé par la formule suivante : Sd = Sr * Q où Sr est le degré de similarité normalisé retourné par l'algorithme de recherche et Q est une somme pondérée des valeurs de qualité de la page. Les auteurs proposent 6 facteurs de qualité des pages Web : *fraîcheur*, *disponibilité*, *rapport signal/bruit*, *autorité*, *popularité* et *cohésion*. Les poids représentent l'importance de ces métriques dans la RI. Le score d'un document est fusionné avec la qualité du site source. L'inconvénient de cette approche est son manque de généralisation et le mode implicite des préférences de l'utilisateur.

2.3.4 Filtrage des résultats de la requête

Dans les systèmes de RI, le principe de base du filtrage de résultats est d'exécuter la requête sans prendre en compte la personnalisation et d'appliquer ensuite un post traitement sur le résultat afin d'éliminer les résultats non pertinents pour l'utilisateur. Le filtrage peut être réalisé soit en appliquant des requêtes supplémentaires sur le résultat, soit en traitant chaque élément séparément afin d'étudier sa pertinence. Dans (Burgess *et al*., 2002), les auteurs proposent une hiérarchie pour caractériser la qualité des données. Les principales dimensions de la qualité sont le *coût*, l'*utilité* et le *temps*. Chaque dimension se décompose en sous-dimensions et facteurs de qualité. Les préférences de l'utilisateur en termes de qualité sont exprimées par : (i) le choix des dimensions ou facteurs de qualité (représenté par des poids) ; (ii) le choix du seuil de qualité. Les auteurs définissent une taxonomie de la qualité afin de filtrer les informations. Les informations qui ont une qualité inférieure au seuil sont éliminées.

Dans les systèmes de BD le filtrage est réalisé durant l'exécution de la requête, les critères de filtrage faisant partie intégrante de l'expression de la requête. Dans

(Koutrika *et al*., 2005), la requête initiale de l'utilisateur est enrichie avec des prédicats de son profil. La notion de qualité fait référence au processus de production des données et plus particulièrement au temps de réponse des requêtes enrichies. L'idée principale de cette approche est qu'en ajoutant un nouveau prédicat à la requête, la pertinence des résultats augmente, mais en contrepartie le temps de réponse croît et le nombre de résultats retournés diminue. Par conséquent, les paramètres qui influent sur le choix des prédicats du profil sont la pertinence des résultats obtenus (importance des prédicats ajoutés), le temps de réponse de la requête, et le nombre de résultats. L'approche présentée est la personnalisation sous contrainte qui consiste à optimiser la valeur d'un des paramètres tout en satisfaisant des contraintes sur les autres.

3. Méta modèle de profils

Nous décrivons ici de façon plus fine la problématique de la personnalisation de l'accès à l'information, sous l'éclairage de la notion de profil définie dans le projet APMD, et en focalisant sur les éléments du profil relevant des « facteurs de qualité ».

3.1. Rappel du méta modèle de profil

Un des objectifs du projet APMD est d'étudier les différents types d'informations constituant le profil utilisateur ainsi que les techniques de leur exploitation dans un environnement BD ou RI. Dans ce contexte il est indispensable de disposer d'une infrastructure générique de définition et de gestion de profils. Le profil utilisateur est représenté par six dimensions ouvertes capables d'accueillir la plupart des informations décrivant l'utilisateur et ses préférences (Bouzeghoub *et al*., 2005) :

– La dimension « Données personnelles » comprend toutes les informations caractéristiques de l'utilisateur qui sont relativement stables dans le temps, par exemple : nom, prénom, sexe, date de naissance, langues parlées, handicaps, etc.

– La dimension « Domaine d'intérêt » correspond aux critères dits « de contenu » : elle exprime les caractéristiques générales des informations que l'utilisateur souhaite obtenir du système.

– La dimension « Préférences de livraison » couvre les contraintes souples intervenant entre la production des informations et leur mise à disposition à l'utilisateur comme par exemple les dimensions de son écran.

– La dimension « Sécurité » comprend toutes les informations liées aux privilèges ou au secret concernant les données, l'utilisateur ou les processus de traitement.

– La dimension « Historique des interactions » contient des préférences manifestées implicitement ou explicitement lors de recherches d'information passées, au travers des interactions avec les résultats de la requête.

– La dimension « Qualité » regroupe les exigences de qualité de l'utilisateur relatives aux informations, aux sources fournissant les informations et aux traitements appliqués à ces informations. 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.

Dans la prochaine sous-section, nous nous focalisons sur la dimension Qualité.

3.2. Détail de la dimension qualité

L'usage du méta modèle de qualité se fait à deux niveaux: (i) un niveau de définition initiale d'un profil, par sélection des dimensions et des attributs génériques, (ii) un niveau d'instanciation et de raffinement du profil ainsi constitué. C'est seulement à l'issue de la seconde étape que le profil peut être exploité lors de la compilation ou de l'évaluation des requêtes.

Les préférences de qualité de l'utilisateur peuvent être exprimées de différentes manières selon les besoins des applications qui les utilisent. Elles peuvent concerner un facteur particulier ou un ensemble de facteurs. La représentation la plus courante de préférences de qualité est la spécification des valeurs de qualité que l'utilisateur souhaite obtenir. Ceci est fait généralement par des semi-intervalles, en spécifiant une borne supérieure ou inférieure (ex. précision > 60%), ou par des ensembles flous (en donnant la fonction d'appartenance). Une deuxième approche d'expression de préférences de qualité est la définition de relations d'ordre. Une relation d'ordre peut être donnée explicitement (ex. « j'ai plus confiance dans la source S_1 que dans S_2 ») ou définie implicitement par des fonctions permettant de dériver cet ordre (ex. « je préfère une source S_i à une autre S_j si S_i a une plus grande précision »). Dans ce dernier cas, l'ordre peut être établi selon un ou plusieurs facteurs de qualité.

Un autre type de préférences concerne la combinaison (agrégation) de facteurs de qualité. Dans la majorité des travaux, cette préférence est exprimée par l'attribution d'un poids à chaque facteur de qualité pour faire une somme pondérée. Ainsi l'utilisateur peut déterminer le degré d'importance de chaque facteur dans la méthode d'évaluation de la qualité. Une alternative à cette approche est d'appliquer les préférences selon un ordre prédéfini (ex. on préfère les éléments qui ont une meilleure fraîcheur et en cas d'égalité ceux qui ont une meilleure exactitude). Finalement, des expressions plus complexes peuvent être utilisées pour décrire les préférences de qualité (ex. requêtes SQL like (Mihaila *et al*, 2000)).

Afin de pouvoir exprimer des préférences sur des facteurs de qualité, l'utilisateur doit être capable de comprendre la signification de chaque facteur et des métriques utilisées pour l'évaluer. Etant donné la diversité des définitions proposées dans la littérature (parfois juxtaposées), il est nécessaire de les analyser et de faire ressortir des taxonomies qui aident la compréhension de l'utilisateur. Dans la suite, nous proposons une synthèse des familles de facteurs de qualité auxquelles nous nous sommes intéressées : la fraîcheur, l'exactitude et la popularité.

3.2.1 Fraîcheur

Intuitivement, le concept de fraîcheur introduit l'idée d'âge des données : Les données sont-elles suffisamment fraîches pour les utilisateurs ? Les données sont-elles

obsolètes ? Une certaine source de données a-t-elle les données les plus récentes ? Quand les données ont-elles été produites ? La fraîcheur représente une famille de facteurs de qualité, chacun représentant un certain aspect de fraîcheur et ayant ses propres métriques. Nous distinguons deux facteurs de fraîcheur :

– *Actualité* (Segev *et al*.,1990): L'actualité mesure la distance ou le décalage entre l'extraction et la livraison de données. Par exemple, en regardant un solde bancaire, nous voulons savoir le moment où il a été extrait de la banque, peu importe quand il a été mis à jour.

– *Age* (Wang *et al*., 1996) **:** L'age mesure la distance ou le décalage entre la création (ou mise à jour) et la livraison des données. Il est indépendant du moment de l'extraction des données. Par exemple, si nous obtenons une recommandation des 10 meilleurs CDs, nous nous intéressons à la date de création de cette liste, peu importe quand elle a été extraite.

Le Tableau 1 montre les métriques de fraîcheur proposées dans la littérature ; une étude détaillée est présentée dans (Bouzeghoub *et al*., 2004).

Facteur	Métrique	Définition
Actualité	Actualité	Le temps passé depuis l'extraction des données (borné par la fréquence d'extraction)
	Obsolescence	Le nombre de transactions de mise à jour d'une source depuis l'extraction des données
	Ratio de fraîcheur	Le taux de données qui sont à jour
Age	Age	Le temps passé depuis la création ou la mise à jour des données (borné par la fréquence de mise à jour)

Tableau 1. *Métriques de fraîcheur*

3.2.2 Exactitude

L'exactitude est liée à la correction et la précision avec laquelle les données du monde réel sont représentées dans un système d'information (Gertz *et al*., 2004) : Les données, correspondent-elles au monde réel ? Comportent-elles des erreurs ? Le niveau de détail des données est-il acceptable ? L'exactitude représente également une famille de facteurs de qualité, nous présentons ci-dessous ceux habituellement utilisés en BD :

– *Correction sémantique* (Shanks *et al*., 1999) : La correction sémantique exprime le niveau auquel les données représentent des états du monde réel. Par exemple, si une base de données nous indique l'adresse d'un client, nous voulons savoir si le client habite vraiment à cette adresse. La correction sémantique implique la comparaison des données par rapport au monde réel ou par rapport à un référentiel considéré comme valide.

– *Correction syntaxique* (Naumann *et al*., 1999) : La correction syntaxique exprime la présence d'erreurs syntaxiques dans les données, par exemple des fautes d'orthographe ou des discordances de format. La correction syntaxique implique l'exécution de règles de vérification. Exemples de règles : « les téléphones internes de la compagnie ont 5 chiffres », « les noms des rues se trouvent dans le catalogue des rues ».

– *Précision* (Redman, 1996) : La précision mesure le niveau de détail des données. Par exemple, le poids d'une personne peut être stocké en kilos (ex. 88 kg.) ou avec plus de précision (ex. 88,25 kg.). D'une façon analogue, une date peut être représentée par l'année, par le jour précis ou même en incluant l'heure. Nous avons besoin d'une hiérarchie de représentations pour décider quand une valeur est plus précise qu'une autre.

Le Tableau 2 liste les métriques d'exactitude proposées dans la littérature ; une étude détaillée est présentée dans (Peralta, 2006).

Facteur	Métrique	Description
Correction sémantique	Taux de correction sémantique	Le taux de données qui ne correspondent pas au monde réel
	Taux de valeurs inexistantes	Le taux de données qui n'existent pas dans le monde réel
	Taux de valeurs inexactes	Le taux de données qui contiennent des erreurs de représentation pour certains attributs
	Degré de correction sémantique	Le degré de confiance dans la correction sémantique des données
	Déviation de correction sémantique	La distance sémantique entre une donnée et son correspondant dans le monde réel
Correction syntaxique	Taux de correction syntaxique	Le taux de données qui satisfont les règles syntaxiques
	Déviation de correction syntaxique	La distance syntaxique entre une donnée et une valeur de référence considérée comme correcte
Précision	Échelle	La précision associée à l'échelle de mesure
	Erreur standard	La déviation standard d'un ensemble de mesures
	Granularité	Le nombre d'attributs utilisés pour représenter un concept simple

Tableau 2. *Métriques d'exactitude*

3.2.3 Popularité

La popularité fait référence au degré d'utilisation effective de l'information par les utilisateurs. Ce facteur présente un grand intérêt car il permet de capturer des items qui sortent du champ habituel de contenu auquel s'intéresse l'utilisateur a priori. Il peut donc contribuer à contrecarrer l'effet dit « entonnoir », en apportant des items inattendus sous l'angle de leur strict « contenu », et néanmoins appréciés pour leurs qualités reconnues par toute une communauté d'utilisateurs.

La popularité peut être évaluée de différentes façons. Dans les contextes répandus d'accès en ligne aux items, la consultation et les téléchargements sont de bons indicateurs de popularité. Lorsque ces systèmes permettent aux utilisateurs d'évaluer les items de façon explicite (ratings), ces évaluations constituent une base intéressante pour évaluer la popularité, de même que les recommandations que les utilisateurs se font entre eux, de façon moins formelle dans les forums. De façon plus spécifique, dans le contexte de la publication académique, la notion de popularité se mesure au travers des citations croisées entre les publications (Goodrum *et al.*, 2001) ; en effet, cette activité de citation est une bonne indication de l'usage réel des documents par les

membres de la communauté académique concernée. De façon analogue, certains moteurs de recherche s'appuient sur les liens entre sites pour intégrer un facteur de popularité dans leur évaluation de la pertinence (Page *et al.*, 1998). Enfin, dans le cadre du e-commerce, ce sont les achats qui traduisent la popularité des items (Schafer *et al.*, 2001). Le Tableau 3 liste les métriques de popularité mentionnées.

Facteur	Métrique	Description
Popularité	Nombre de consultations	Le nombre de consultations des données
	Rating	L'évaluation ou recommandation donnée par un groupe d'utilisateurs
	Nombre de citations	Le nombre de citations référençant les données
	Quantité d'achats	Le volume d'achats

Tableau 3. *Métriques de popularité*

4. Impact de la qualité sur le cycle de vie d'une requête

Comme nous l'avons dit plus haut, la notion de requête prend ici un sens très large : elle reflète la demande d'information d'un utilisateur, sachant que la réponse attendue comporte des données factuelles ou des objets véhiculant une information plus complexe. Cette section décrit l'impact de la qualité sur le cycle de vie d'une requête.

4.1. Rappel du cycle de vie d'une requête

Dans cette section nous rappelons les quatre étapes du cycle de vie d'une requête. La personnalisation intervient à chacune des étapes en tenant compte des préférences de l'utilisateur décrites dans son profil.

– *Reformulation* : Chaque requête utilisateur Q est reformulée en exploitant le profil de l'utilisateur P et les descriptions des sources de données. Il y a deux techniques principales de reformulation de requêtes : l'enrichissement et la réécriture. L'enrichissement a pour objectif d'intégrer des éléments du centre d'intérêt du profil utilisateur à la requête initiale afin de mieux cibler les informations recherchées. La réécriture a pour objectif de traduire une requête utilisateur (exprimée sur un schéma global) en un ensemble de requêtes exprimées sur les sources de données. Notons que plusieurs sources de données peuvent contenir des informations de même nature et par conséquent il est possible d'obtenir un ensemble de reformulations de la requête initiale.

– *Optimisation* : L'optimisation consiste principalement à compiler la requête reformulée Q' (ou l'ensemble de reformulations) et à élaborer un plan d'exécution. Dans un contexte multi-sources, pendant cette phase une requête est décomposée en un ensemble de sous-requêtes, chacune exprimée sur une source unique.

– *Exécution* : Une fois le plan d'exécution choisi, les sous-requêtes sont évaluées et leurs résultats sont intégrés. La fonction de correspondance (appariement) peut faire l'objet d'adaptations en vue de la personnaliser selon le profil.

– *Présentation* : Avant d'être restitués à l'utilisateur, les résultats obtenus peuvent subir une étape de préparation afin d'être présentés selon les préférences de l'utilisateur. L'adaptation de la présentation des résultats peut se traduire par leur ré-ordonnancement, le choix des modalités de livraison (nombre de résultats par page, disposition à l'écran, nombre maximum de résultats etc.) et du média de livraison (e-mail, fax, etc.) ou encore par la notification de l'arrivée des résultats.

Une fois les résultats restitués à l'utilisateur, l'historique de ses activités peut être analysé pour mettre à jour son profil.

La dimension « Qualité » du profil peut être exploitée à différentes étapes du cycle de vie de la requête :

– pendant la reformulation, pour sélectionner les sources de données conformes aux préférences de l'utilisateur et ainsi restreindre l'espace de recherche,

– pendant l'optimisation, pour choisir la stratégie d'exécution la plus adaptée,

– pendant l'exécution, comme un ensemble de facteurs prenant part au calcul du score de pertinence affecté aux objets sélectionnés,

– pendant la présentation des résultats, pour organiser les résultats en tenant compte de leur capacité à répondre à une demande de l'utilisateur.

4.2. Quelques travaux d'APMD sur la personnalisation guidée par la qualité

Dans le projet APMD (ACI Masses de Données), l'objectif est de définir un cadre général pour la personnalisation de l'accès aux informations, et de l'instancier au travers de techniques de personnalisation couvrant largement le champ ainsi défini. Dans cette section nous décrivons brièvement quelques résultats de l'utilisation de la qualité pour la personnalisation, obtenus dans le projet APMD.

4.2.1 Sélection des sources de données selon les préférences de qualité

L'étape de reformulation de requête produit un ensemble de reformulations (requêtes) qui interrogent différentes sources. Dans ce contexte il est intéressant de prendre en compte la qualité des données des sources et de la confronter aux préférences de l'utilisateur afin d'augmenter la pertinence des résultats. Les mesures de qualité des données des différentes sources doivent être agrégées afin d'obtenir une qualification du résultat d'une reformulation. Ce calcul dépend de plusieurs paramètres : la qualité des données sources, la nature des données, les opérateurs de la requête, le contexte d'application, etc.

Notre approche pour l'évaluation de la qualité (Peralta, 2006) consiste à propager les valeurs de qualité le long du graphe algébrique correspondant à la requête, en les combinant avec les autres paramètres de l'application (ex. coûts, décisions de conception). L'évaluation peut être faite en partant des sources de données vers la racine du graphe qui représente le résultat de la requête (propagation de la qualité des sources) ou inversement, en partant de la racine vers les sources (propagation des préférences des utilisateurs). Les deux types de propagation ont des objectifs

différents. D'un coté, la propagation des préférences de l'utilisateur est faite dans une perspective de sélection des sources qui permettent de satisfaire ces préférences. On obtient des contraintes de qualité pour les sources, c'est-à-dire, les conditions que chaque source de données doit remplir afin de satisfaire les demandes des utilisateurs. De l'autre coté, la propagation de la qualité des sources permet d'estimer les valeurs de qualité des résultats de la requête. Les valeurs obtenues peuvent être comparées aux exigences de l'utilisateur afin de choisir les reformulations qui satisfont au mieux les attentes de celui-ci. Dans certains cas il est possible que les préférences de l'utilisateur ne puissent pas être satisfaites. La propagation de la qualité des sources jusqu'aux utilisateurs permet également de découvrir les bornes optimales des valeurs de qualité que le système d'information est capable de fournir.

4.2.2 Prise en compte de la subjectivité des facteurs : la popularité personnalisée

On peut envisager la popularité sous un angle plus subjectif que celui décrit plus haut, dans la mesure où elle implique le jugement émis par des personnes. Ce facteur peut ainsi être personnalisé, en ne prenant en compte que les indications de popularité émanant de personnes proches de l'utilisateur considéré. Il faut pour cela définir la notion de « communauté », qui regroupe les utilisateurs qui ont quelque chose qui les rapproche. Dans cette perspective, la valeur de l'attribut « popularité » pour un item ou un document n'est pas la même pour tous les utilisateurs, car le calcul se basera sur des données de référence différentes. Par exemple, un film peut être très populaire auprès des jeunes de 18 à 25 ans, et pas du tout auprès des 25-40 ans.

Pour étendre ce principe, le modèle des « espaces de communautés » (Nguyen *et al.*, 2006a) permet de définir, à partir d'un ensemble d'utilisateurs, les communautés formées selon divers critères de rapprochement (âge, profession, localisation géographique, centre d'intérêt affichés, opinions formulées, etc.), qui trouvent leur origine dans certaines dimensions du profil utilisateur : Données personnelles (l'âge des personnes, leur profession, leur localisation géographique), Domaine d'intérêt, Historique des interactions (les opinions formulées sur les items rencontrés par le passé). Dans ce cadre, un utilisateur U est associé à une communauté pour chacun des critères de rapprochement : cela donne lieu au « vecteur de positionnement », qui contient l'identificateur de la communauté à laquelle U appartient dans chacun des espaces de communautés. Chaque élément du vecteur traduit des préférences de façon implicite. Par exemple, pour l'espace « Age » les utilisateurs sont regroupés en communautés (« Enfants », « Ados », « Adultes », « Seniors », etc.). Voici un exemple de vecteur de positionnement : [Age = Ados ; Profession = Tertiaire ; Localisation géographique = France ; Centres d'intérêt = Films d'aventure ; Goûts = Communauté #23].

Pour instancier un élément du vecteur de positionnement, il faut définir des classes d'utilisateurs reflétant leur proximité relativement au critère en question. Selon la nature des attributs de profil impliqués, il faut appliquer une technique adéquate de classification des utilisateurs. Pour les critères simples, les communautés sont formées en regroupant les utilisateurs selon les valeurs des attributs de profil impliqués. Par exemple, la communauté « Ados », regroupe les utilisateurs ayant entre 12 et 18 ans sur la base de l'âge indiqué dans la dimension « Données personnelles » du profil. Pour

les critères complexes comme « Centres d'intérêt » ou « Goûts », qui s'appuient respectivement sur la dimension « Centre d'intérêt » et « Historique des interactions », la définition d'une mesure de proximité entre les utilisateurs est nécessaire aux méthodes de classification non supervisée utilisées pour former les communautés. Dans (Nguyen *et al.*, 2005), nous avons proposé de combiner l'algorithme des fourmis artificielles avec l'algorithme des K-moyennes pour faciliter la visualisation en 2D des classes obtenues.

Il est fréquent que le vecteur de positionnement soit incomplet ou incorrect pour certains utilisateurs, par exemple pour les nouveaux utilisateurs dont l'historique d'interaction est vide. Pour améliorer ces vecteurs, nous avons proposé une méthode d'induction fondée sur la théorie des ensembles d'approximation permettant de compléter les vecteurs de positionnement incomplets selon la stratégie conduisant à la meilleure qualité d'induction (Nguyen *et al.*, 2006b).

Dans le cycle de vie de la requête, ce vecteur peut intervenir à l'étape de *présentation des résultats*, en intégrant au score des items un facteur lié à la popularité, ce facteur traduisant la propension des utilisateurs proches de U à apprécier les items candidats. Ce vecteur peut aussi intervenir lors de *l'exécution de la requête*, non pas pour restreindre le champ de l'espace de recherche, mais pour l'élargir. En effet, il fait émerger des items au nom de leur popularité auprès d'utilisateurs proches de U alors même qu'ils seraient ignorés si l'on s'en tenait aux critères de sélection inhérents à la requête.

4.2.3 Prise en compte de la qualité dans la présentation des résultats

Notre approche vise à filtrer les documents résultant de l'exécution d'une requête selon les préférences de l'utilisateur en termes de qualité. Les documents dont la valeur de qualité des facteurs n'appartient pas à un intervalle de confiance défini dans le profil utilisateur sont éliminés. Afin d'optimiser le filtrage, nous évaluons tout d'abord les facteurs de qualité puis nous appliquons le filtrage selon un ordre croissant des intervalles de confiance associés à ces facteurs. En se basant sur la stratégie d'évaluation de la qualité et la méthode de calcul de score SAW (Simple Additive Weighting), nous présentons sur un exemple comment intégrer la partie qualité du profil dans le filtrage d'information et le ré-ordonnancement des résultats. Le Tableau 4 donne un exemple de filtrage des résultats selon la qualité générale du document. L'évaluation du score d'un document se fait par agrégation des scores élémentaires associés à la source d'information, l'usage et le support de cette information.

Matrice de décision				**Calcul du score**
Document	Source	Support	Usage	Score de qualité
Poids	0,222	0,444	0,333	
D_1	0,558	0,788	0,362	**0,595**
D_2	0,558	0,715	0,644	**0,656**
D_6	0,664	0,637	0,267	**0,520**
D_7	0,664	0,690	0,124	**0,496**

Tableau 4. *Exemple de calcul de score*

Le principe du ré-ordonnancement consiste à modifier l'ordre d'affichage des résultats à l'utilisateur. Il s'agit d'un post traitement qui, étant donné les éléments retournés par une requête, essaie de trouver une manière d'échanger leurs emplacements en fonction des préférences de l'utilisateur sans pour autant négliger l'ordre qui a été attribué aux documents par le moteur de recherche. L'échange de l'ordre d'apparition des éléments des résultats est effectué généralement en appliquant une fonction qui permet de calculer le nouveau rang de l'objet.

Nous proposons d'intégrer la qualité dans le ré-ordonnancement des résultats en introduisant une fonction de rang basée sur la fusion du score du document, retourné par l'algorithme de recherche, et sa qualité (Harrathi, 2005). Le nouveau score d'un document est obtenu par la formule suivante : $S'_d = S_d * Q_d$ où Sd est le degré de similarité normalisé retourné par l'algorithme de recherche et Q_d est la qualité du document d.

5. Conclusion

La prise en compte des profils utilisateurs dans les systèmes d'accès à l'information apporte une solution au problème de la surcharge informationnelle subie par l'utilisateur lors d'une recherche sur Internet ou dans des fédérations de bases de données dans les entreprises (redondance et multiplicité des données, bruit). Elle améliore aussi la pertinence de l'information restituée en intégrant dans les critères de sélection diverses préférences de l'utilisateur. Elle permet enfin une plus grande adéquation au contexte de requêtage de l'utilisateur, notamment le lieu géographique, le moment d'émission de la requête, le type de terminal utilisé, etc.

La qualité de l'information ne fait habituellement pas partie des systèmes d'accès à l'information. L'évaluation des requêtes se fait sans tenir compte de la fraîcheur des données ou de leur précision, de la confiance accordée aux sources de données ou de la crédibilité des informations délivrées. L'un des objectifs du projet APMD est d'intégrer cette dimension qualité dans le processus d'accès personnalisé. La dimension qualité impacte différentes étapes du cycle de vie d'une requête. Dans cet article, nous avons décrit les principaux éléments de cette dimension qualité et avons montré à travers quelques exemples de réalisations dans le projet APMD l'impact de la qualité sur la personnalisation.

6. Bibliographie

Ballou D., Wang R., Pazer H., Tayi G., "Modelling Information Manufacturing Systems to Determine Information Product Quality", *Management Science*, vol. 44, nº 4, 1998.

Bouzeghoub M., Peralta V., "A Framework for Analysis of Data Freshness", *Actes du 1st Int. Workshop on Information Quality in Information Systems (IQIS)*, Paris, France, 2004.

Bouzeghoub M., Kostadinov D., "Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils", *Actes de la 2ème Conférence en Recherche d'Informations et Applications (CORIA)*, Grenoble, France, 2005.

Braumandl R., "Quality of Service and Optimization in Data Integration Systems", *Actes du GI-Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW)*, Germany, 2003.

Burgess M., Alex Gray W., Fiddian N., "Establishing Taxonomy of Quality for Use in Information Filtering", *Actes de la* 19th British National Conference on Databases (BNCOD 2002), Sheffield, UK, 2002, p.103-113.

Calabretto S., Pinon J.M., Poullet L. et Richez M.A, "De la qualité de l'information à la qualité de la documentation", Document Numérique, vol.12, nº.1, 1998, p. 37-52.

Cho J., Garcia-Molina H., “Synchronizing a database to improve freshness”, *Actes de la ACM Int. Conf. on Management of Data (SIGMOD)*, Dallas, USA, 2000, p. 117-128.
Denos N., Quality Control Tools User Requirements - Part 1, Deliverable D2.1.1, Report UR-R1-QCT1 on WP2 Task 2.1, TIPS European project IST-1999-10419, July 2000.
Denos N., Quality control tools specifications, Deliverable D3.1.3, Report SSR-QCT on WP3 Task 3.3, TIPS European project IST-1999-10419, 2000.
Denos N., “QCT and SF services in Torii: Human Evaluations of Documents Benefit to the Community”, *Actes du Workshop on Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives*, Malaga, Spain, 2002, p.105-114.
Gertz M., Tamer Ozsu, M., Saake, G., Sattler, K., “Report on the Dagstuhl Seminar: Data Quality on the Web”, *SIGMOD Record*, vol. 33, nº 1, March 2004.
Goodrum A. A.; McCain K. W., Lawrence S., Giles C. L., “Scholarly Publishing in the Internet Age: A Citation Analysis of Computer Science Literature”, *Information Processing & Management*, vol. 37 nº 5, 2001, p. 661-75.
Harrathi R., Facteurs de qualité et personnalisation de l’information, Mémoire de master, Institut National des Sciences Appliquées de Lyon, Juin 2005.
Harrathi R., Calabretto S., “Un modèle de qualité de l'information”, *Actes des Journées Extraction et Gestion de Connaissances (EGC)*, Lille, France, 2006, p. 299-304.
Jarke M., Vassiliou Y., “Data warehouse quality design: A review of the DWQ project”, *Actes de la* Int. Conf. on Information Quality (IQ), Cambridge, USA, 1997.
Koutrika G., Ioannidis Y. E., “Constrained Optimalities in Query Personalization”, *Actes de* ACM SIGMOD, Baltimore, USA, 2005.
Mihaila G. A., Rashid L., Vidal M. E., “Using Quality of Data Metadata for Source Selection and Ranking”, *Actes du 3rd Int. Workshop on the Web and Databases (WebDB)*, USA, 2000.
Motro A., Rakov I., “Estimating the quality of databases”, *Actes de la 3rd Int. Conf. on Flexible Query Answering Systems (FQAS)*, Roskilde, Denmark, 1998.
Naumann F., “Data Fusion and Data Quality”, *Actes du Seminar on New Techniques et Technologies for Statistics*, Sorrento, Italy, 1998.
Naumann F., Leser U., et Freytag J.C., “Quality-driven integration of heterogenous information systems”, *Actes de la Int. Conf. on Very Large Databases (VLDB)*, Edinburgh, 1999.
Naumann F., Rolker C., “Assessment Methods for Information Quality Criteria”, *Actes de la Int. Conf. on Information Quality (IQ)*, Cambridge, USA, 2000.
Nguyen A.-T., Denos N., Berrut C., “Cartes de communautés pour l’adaptation interactive de profils dans un système de filtrage d’information”, *Actes du Congrès INFORSID*, Grenoble, France, 2005, p. 253-268.
Nguyen A.-T., Denos N., Berrut C., “Modèle d’espaces de communautés basé sur la théorie des ensembles d’approximation dans un système de filtrage hybride”, *Actes de la* Conf. en Recherche Information et Applications (CORIA), Lyon, France, 2006, p. 303-314.
Nguyen A.-T., Denos N. et Berrut C., “Exploitation des données "disponibles à froid" pour améliorer le démarrage à froid dans les systèmes de filtrage d'information”, *Actes du Congrès INFORSID*, Hammamet, Tunisie, 2006, p. 81-95.
Page L., Brin S., Motwani R., and Winograd T., The PageRank Citation Ranking: Bringing Order to the Web*,* Technical report, Stanford Digital Library Technologies Project, 1998.
Peralta, V., Data Quality Evaluation in Data Integration Systems, PhD Thesis, Université de Versailles, France and Universidad de la República, Uruguay, 2006.
Pipino L.L., Lee Y.W., Wang, R., “Data Quality Assessment”, *Communications of the ACM*, vol. 45, nº 4, April 2002.
Redman T., *Data Quality for the Information Age*, Artech House, 1996.
Schafer J. B., Konstan J. A., Riedl J. 2001. “E-Commerce Recommendation Applications”, *Data Min. Knowl. Discov*, vol. 5, nº 1-2, 2001, p. 115-153.
Segev A., Weiping F., “Currency-Based Updates to Distributed Materialized Views”, *Actes de la* 6th Int. Conf. on Data Engineering (ICDE), Los Angeles, USA, 1990.
Shanks G., Corbitt B., “Understanding Data Quality: Social and Cultural Aspects”, *Actes de la 10th Australasian Conf. on Information Systems*, Wellington, New Zealand, 1999.
Strong D., Lee Y. et Wang R., “Data quality in context”, *Communications of the ACM*, vol. 40, nº 5, 1997, p. 103-110.
Wang R., Strong D., “Beyond accuracy: What data quality means to data consumers”, *Journal on Management of Information Systems*, vol. 12, nº 4, 1996, p. 5-34.
Zhu X., Gauch S., “Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web, *Actes de la 23rd ACM SIGIR Conf.*, Greece, 2000, p. 288–295.