UNIVERSITE DE JENDOUBA
FACULTE DES SCIENCES JURIDIQUES, ECONOMIQUES
ET DE GESTION DE JENDOUBA

MEMOIRE

Présenté devant

L'Institut National des Sciences Appliquées de Lyon
(Ecole Doctorale Informatique et Information pour la Société)

Pour obtenir

Le Grade de Master
Spécialité : Informatique

Présenté Par

Rami HARRATHI

Facteurs de qualité et personnalisation de l'information

Encadré par

Sylvie CALABRETTO

Soutenu le 27 juin 2005

Remerciements

Ce travail de Master a constitué ma première expérience dans l'activité de recherche. Il n'aurait pas été aussi fructueux sans l'aide de plusieurs personnes. Je remercie tous ceux qui ont contribué, de prés ou de loin, à réaliser ce travail.

Je tiens à adresser mes remerciements aux membres du Laboratoire d'Informatique en Images et Systèmes d'informations ***LIRIS*** que j'ai pu côtoyer durant la période de mon stage et qui ont su rendre mon travail agréable.

Plus particulièrement, je tiens à remercier ***Mme Sylvie CALABRETTO*** pour son encadrement continu, pour les remarques constructives qu'elle m'a fournies ainsi que pour ses précieux conseils durant toute la période du stage.

Je remercie également la coopération entre l'INSA de LYON et Faculté des sciences juridiques, économiques et de gestion de JENDOUBA pour m'avoir offert la chance et la possibilité de faire ce stage de Master.

Ma gratitude, mon profond respect et mes remerciements à tous les membres du jury et à tous les rapporteurs pour leur attention consacrée à l'égard de mon travail.

À Dieu,

À l'âme de mon père, À Ma mère, Mes frères et sœurs …

Pour finir, un grand Merci à mes chers amis : Tarek et Naim pour leurs encouragements.

HARRATHI Rami

Table des matières

Introduction

Avec l'expansion d'Internet et du Web, on assiste à une prolifération des ressources hétérogènes (données structurées, documents textuels, composants logiciels, images), conduisant à des volumes considérables. Dans ce contexte les outils d'accès à l'information (moteurs Web, SGBD, etc.) délivrent, dans des temps de plus en plus longs, 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.

Une solution à l'amélioration de cette pertinence est la personnalisation ou l'adaptation des réponses fournies aux utilisateurs selon leurs profils c'est-à-dire selon leurs besoins et leurs préférences[1]. Ainsi la formulation du besoin d'information est devenue un des éléments clés pour obtenir des résultats pertinents dans un processus d'accès à l'information.
Pour aider à cette formulation, des travaux [BOU04] proposent d'introduire la notion de qualité. Il est par exemple possible de poser une requête en spécifiant des préférences extrinsèques en termes de qualité comme une réponse rapide ou une information fraîche. Ainsi on peut définir un profil qualité comme un ensemble de préférences ou besoins en termes de qualité d'information caractérisant un utilisateur ou groupe d'utilisateurs.

Dans ce cadre notre contribution porte sur la proposition d'un modèle de qualité de l'information décrivant les différents facteurs de qualité influant sur la personnalisation. Notre modèle va permettre de structurer les différents facteurs de qualité d'information dans une hiérarchie afin d'assister l'utilisateur dans la construction de son propre profil de qualité. Nous proposons aussi les différentes méthodes d'évaluation de la qualité, puis l'exploitation du profil qualité dans le processus d'accès aux informations.

Notre rapport se compose de quatre chapitres principaux : le premier chapitre présente un état de l'art sur les approches existantes sur la modélisation de la qualité des données d'une part et sur l'intégration de la qualité dans la personnalisation de l'information d'autre part. Dans le deuxième chapitre nous présentons notre modèle de qualité d'information en focalisant sur l'exploitation du modèle proposé dans la construction du profil qualité. Dans le troisième chapitre nous présentons les différentes méthodes d'évaluation de la qualité dans notre modèle. Enfin dans le quatrième chapitre nous présentons la manière d'exploitation du profil qualité dans le processus d'accès aux informations. Nous terminons notre rapport par une conclusion générale ainsi que quelques perspectives qui peuvent être abordées dans des travaux futurs.

[1] Notre travail se situe dans le cadre du projet ACI APMD (Accès Personnalisé à des Masses de Données).
Site Web: http://apmd.prism.uvsq.fr/
Partenaires: CLIPS Grenoble, IRISA Lannion, IRIT Toulouse, LINA Nantes, LIRIS Lyon, PRiSM Versailles

Chapitre 1
État de l'art

Afin de définir les facteurs de qualité relatifs aux données influant sur la personnalisation, il est nécessaire d'analyser les différents travaux menés sur le thème de la qualité des données.

Dans ce chapitre nous présentons les différentes approches concernant :

- la modélisation de la qualité des données.
- l'intégration de la qualité dans la personnalisation de l'information.

1. Modélisation de la qualité des données

1.1 Revue de littérature

La qualité des données est un domaine de recherche qui a suscité depuis longtemps un vif intérêt, mais qui émerge tout juste comme champ de recherche à part entière, tel que peuvent l'indiquer [STR97] [JAR97] [NAU99b] [NAU00] [MAR02]. Avec le nombre croissant des sources d'information disponibles sur le réseau, le problème de la qualité de leur contenu est devenu crucial.

Dans le cadre de la modélisation de la qualité des données, de nombreuses propositions ont été faites, ces propositions dépendent du point de vue de chaque auteur. La première difficulté réside dans l'absence de consensus sur la notion même de qualité. Tout le monde s'accorde en effet sur le fait que la qualité des données peut se décomposer en un certain nombre de dimensions, catégories, critères, facteurs, paramètres ou attributs, mais aucune définition ne fait aujourd'hui l'unanimité (tableau 1).

Dans [NAU00] les auteurs identifient trois approches d'analyse des critères de la qualité des données :

- approche orientée sémantique : elle est basée seulement sur la signification des critères. Cette approche est la plus intuitive (il s'agit d'une approche où les critères sont examinés de façon générale, c'est-à-dire séparés de tout cadre d'information).
- approche orientée traitement : elle classe les critères de qualité de l'information selon leur déploiement dans les différentes phases du traitement de l'information.
- approche orientée objectif : elle est caractérisée par une définition des objectifs de la qualité à atteindre et un classement des critères selon les objectifs définis.

Dans le tableau 1 nous présentons les caractéristiques de quelques approches de la modélisation de la qualité des données.

Auteurs	**Dichotomie et caractérisation de la qualité des données**			
Wang, Strong et Kan [STR97]	»Approche orientée sémantique » 4 Catégories » 13 Dimensions	**Catégorie**	**Dimension**	
		Qualité intrinsèque	exactitude, objectivité, crédibilité, réputation	
		Qualité d'accessibilité	accès, sécurité	
		Qualité contextuelle	pertinence, complétude, quantité de données	
		Qualité de la représentation	concision, cohérence, intégrité, facilité de compréhension	
Jarke et Vassiliou [JAR97]	»Approche orientée objectif » 5 Facteurs	**Facteur**		
		Facilité d'interprétation		
		Accessibilité		
		Utilité		
		Fiabilité		
		Validation		
Calabretto, Pinon, Poullet et Richez [CAL98]	»Approche orientée sémantique » 3 Critères de qualité d'information » 8 Critères de qualité des documents	**Critère**		
		Disponibilité		
		Fiabilité		
		Adaptabilité		
		Critère		
		Identifiabilité		
		Facilité d'exploitation		
		Crédibilité		
		Traçabilité		
		Compréhensibilité		
		Réutlisabilité		
		Portabilité		
		Flexibilité		
Berti [BER99]	»Approche orientée sémantique » 4 Catégories » 32 Critères	**Catégorie**	**Sous catégorie**	**Critère**
		Qualité de la gestion de la donnée par le système		accessibilité, confidentialité, facilité d'échange facilité de maintenance, facilité de manipulation facilité de recherche, sécurité...
		Qualité de la représentation de la donnée par le système		compréhension, concision, interprétation ...
		Qualité intrinsèque de la donnée		absence d'erreur, cohérence, exactitude, qualité de la source...
		Qualité relative de la donnée	Qualité contextuelle	*référentiel temporel discret : actualité*, fraîcheur ... *Continu :* traçabilité, variabilité, volatilité... *référentiel applicatif :* criticité, pertinence**,** utilité.
			Qualité relative à l'utilisateur	*référentiel cognitif :* fiabilité, originalité, rareté, vraisemblance *référentiel affectif* **:** intérêt, préférences...
			Qualité relative aux données homologues	comparaison, contradictions redondance, similarité...

Naumann, Leser et Freytag [NAU99a]	»Approche orientée traitement » 3 Classes » 11 Critères	**Classe**	**Critère**
		Spécifique source	facilité de compréhension, réputation, fiabilité, âge
		Spécifique requête	disponibilité, prix, consistance de représentation, temps de réponse, exactitude, relevance
		Spécifique attribut	complétude, quantité de données
Naumann et Rolker [NAU00]	»Approche orientée traitement » 3 Classes d'évaluation » 11 Critères	**Classe d'évaluation**	**Critère**
		Critères subjectifs	crédibilité, facilité, concision de représentation, d'interprétation, relevance, réputation, facilité de compréhension, utilité
		Critères objectifs	complétude, support client, documentation, objectivité, prix, fiabilité, sécurité, âge, vérifiabilité
		Critères du processus	exactitude, disponibilité, consistance de représentation, latence, temps de réponse
Zhu et Gauch [ZHU00]	»Approche orientée sémantique » 5 Critères de qualité des pages web	**Critère**	
		Actualité	
		Disponibilité	
		Autorité	
		Popularité	
		Exactitude	
Denos [DEN02]	»Approche orientée sémantique » 8 descripteurs de qualité des documents	**Descripteur**	
		Qualité scientifique	exactitude, précision, originalité, complétude, actualité, qualité de démonstration, qualité de la liste des références, qualité de la méthodologie
		Lisibilité	qualité du style d'écriture, qualité de la structure logique, adéquation des illustrations, absence des répétitions, clarté de l'expression des idées
		Public visé	niveau technique
		Fraîcheur	date de publication
		Autorité	réputation de l'auteur, réputation du journal ou de la conférence
		Disponibilité	longévité, imprimabilité
		Popularité	nombre des lecteurs, des citations
		Qualité d'identification	citabilité
Marotta [MAR02]	»Approche orientée traitement » 2 points de vue : - Système - Utilisateur » 6 Catégories » 31 Critères	**Catégorie**	**Critère**
		Qualité intrinsèque	crédibilité, précision, réputation, consistance, granularité, complétude, In ambiguïté, quantité de données
		Qualité contextuelle	pertinence, Fitness horizontal, Fitness vertical, fréquence de MAJ
		Qualité de la représentation	facilité de compréhension, concision de la représentation, syntaxique, sémantique
		Qualité d'accessibilité	privilèges, disponibilité, assistance, facile à localiser, temps de réponse, connectivité
		Contenu	exactitude, volume, utilité, complétude, fraîcheur
		Opérationnel	performance, accessibilité, facilité de compréhension.

Tableau 1 - Quelques approches de modélisation de la qualité des données.

1.2 Limites des modèles existants

L'inconvénient des approches proposées pour caractériser la qualité des données semble être une certaine rigidité qui parait ne laisser que relativement peu de choix à l'utilisateur, sans pour autant l'aider à construire un ensemble cohérent et minimal de critères de qualité ou bien l'assister dans leur spécification. En effet elles représentent la qualité comme une collection de critères.

La plupart des approches proposées de la modélisation de la qualité des données sont limitées dans leur applicabilité. Elles sont utiles seulement dans le domaine pour lequel elles ont été conçues ainsi la réutilisation de la définition de la qualité est limitée. En effet la majorité des modèles incorporent des critères de qualité les plus appropriés à leur domaine cible.

La majorité des définitions proposées de la qualité des données ne distinguent pas le point de vue utilisateur et le point de vue système. Par exemple pour la fraîcheur des données on distingue la fraîcheur comme un point de vue utilisateur et la fréquence de mise à jour des données comme un point de vue système. Cette confusion rend difficile l'intégration de la qualité dans le processus d'exécution des requêtes.

On peut résumer les limites des travaux courants dans le domaine de la recherche sur la qualité des données dans les points suivants :

- un manque d'attention vis-à-vis du consommateur de la qualité (l'utilisateur).
- une confusion entre le point de vue utilisateur et le point de vue système.
- un choix divers de définitions de qualité des données.
- une réutilisation limitée des définitions courantes de la qualité des données.

2. Intégration de la qualité dans la personnalisation de l'information

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**.**

Parmi les données qui constituent un profil utilisateur on trouve une dimension relative à la qualité [KOS03]. Elle permet d'exprimer des préférences extrinsèques comme 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. Il est par exemple possible de poser une requête en spécifiant des préférences en termes de qualité comme une réponse rapide ou une information fraîche.

Dans le contexte de la personnalisation de l'information la qualité 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 utilisateur par rapport aux facteurs de qualité sont représentées sous forme des paramètres (poids).

Dans ce contexte peu de travaux intègrent la qualité de façon implicite. Dans la suite nous présentons quelques travaux qui intègrent la qualité dans les services de la personnalisation :

- le re-ordonnancement des résultats de la requête.
- le filtrage des résultats.

2.1 Re-ordonnancement des résultats de la requête

Le principe du re-ordonnancement est de modifier l'ordre d'affichage des résultats au client. 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 objets du résultat 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.

Un exemple des approches qui intègre des facteurs qualité est l'approche de **Zhu et Gauch** [ZHU00]. Les auteurs proposent 6 facteurs de qualité des pages web : *currency*, *availability*, *information-to-noise ratio*, *authority*, *popularity* et *cohesiveness*.
Dans cette approche le score d'une page web est calculé par la formule suivante :
$Sd = Sr * (ap *td + bp*ad + cp*id + dp*rd + ep*pd+ fp*cd)$ où

- Sd est le score final d'un document *d.*
- Sr est le degré de similarité normalisé retourné par l'algorithme de recherche.
- td, ad, id, rd, pd et cd sont les mesures de qualité du document normalisées qui correspondent respectivement à *currency*, *availability*, *information-to-noise ratio*, *authority*, *popularity* et *cohesiveness*.
- ap, bp, cp, dp, ep et fp sont les poids représentant l'importance de ces métriques dans la recherche d'information.

Les auteurs fusionnent le score d'un document avec la qualité du site source. Le score d'un document d retourné par un site i est donné par la formule suivante :
$S'_d = S_{d*} Gi$ où

- Sd est le score d'un document *d*.
- Gi est la qualité du site i donnée par cette formule :
  $Gi = Wi * (as' *Ti' + bs'*Ai' + cs'*Ii' + ds'*Ri' + es'*Pi' + fs'*Ci')$
- Wi est la quantité d'information
- Ti', Ai', Ii', Ri', Pi' et Ci' sont les mesures de qualité du site i normalisées qui correspondent respectivement à *currency*, *availability*, *information-to-noise ratio, authority*, *popularity*, *et cohesiveness*
- as', bs', cs', ds', es' et fs' sont les poids représentant l'importance de ces métriques dans la recherche d'information.

L'inconvénient de cette approche est que les auteurs ont défini des critères de qualité les plus appropriés à leur domaine et les préférences de l'utilisateur sont exprimées de façon implicite (sous forme de poids).

Une autre approche est celle de **Brin et Page** [BRI98]. Les auteurs définissent le rang d'une page web comme la probabilité de visite. Le rang d'une page est calculé par la formule suivante : $PR(A) = (1-d) + (PR(T1)/C(T1) + \ldots + (PR(Tn)/C(Tn))$ où

- T1...Tn sont des pages qui apprécient la page A (exemple citations).
- d est un facteur d'amortissement qui peut être placé entre 0 et 1.
- C(a) est défini par le nombre de liens dans la page A.

Ce travail ne définit pas explicitement des facteurs de qualité. Cependant, il ignore le fait que tous les utilisateurs sont différents, avec des besoins en termes de qualité différents. En effet les préférences en termes de qualité ne sont pas définies.

2.2 Filtrage des résultats de la requête

Le principe de base de cette approche est d'exécuter la requête sans prendre en compte la personnalisation et ensuite appliquer un post traitement sur le résultat afin d'éliminer les résultats non pertinents pour l'utilisateur. Le filtrage peut être fait soit en appliquant des requêtes supplémentaires sur le résultat, soit en traitant chaque élément séparément afin d'étudier sa pertinence.

Dans [BUR99] **Burgess, Gray** et **Fiddian** proposent une hiérarchie pour caractériser la qualité des données. Les principales dimensions de qualité sont coût, utilité et temps. Chaque dimension se décompose en sous-dimensions et critères de qualité.
Dans cette approche les préférences de l'utilisateur en terme de qualité sont exprimées par :

- Le choix des dimensions ou critères de qualité (représenté par des poids).
- 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 (Figure 1).

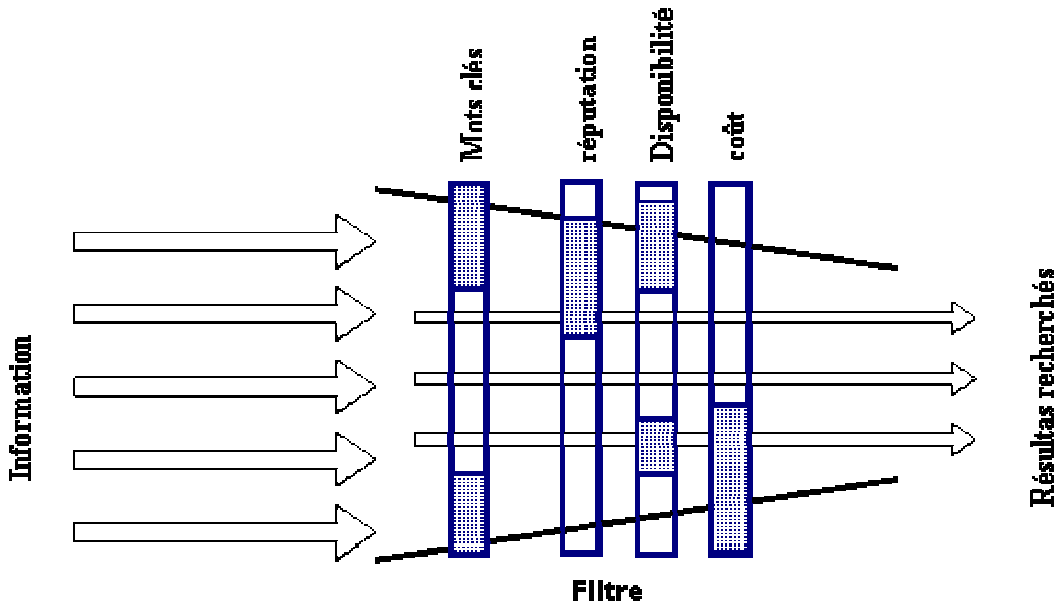


Figure 1- Filtrage d'information en utilisant des facteurs qualité [BUR99].

3. Conclusion

La plupart des approches proposées de la modélisation de la qualité des données sont caractérisées par un manque d'attention vis-à-vis de l'utilisateur. Peu de travaux sur la personnalisation utilisent de façon systématique la qualité.

Afin de déterminer les facteurs de qualité influant sur la personnalisation de l'information, il est nécessaire de fournir un modèle décrivant les différents facteurs de qualité et aidant à la construction d'une définition personnalisée de la qualité selon les exigences et les besoins d'un utilisateur ou d'un groupe d'utilisateur.

Chapitre 2
Proposition d'un modèle de qualité de l'information

Le but de notre stage est de définir les facteurs de qualité relatifs à l'information pouvant intervenir dans sa personnalisation.

Dans ce chapitre nous proposons un modèle de facteurs de qualité aidant à la construction d'une définition personnalisée de la qualité selon les exigences et les besoins d'un utilisateur ou d'un groupe d'utilisateurs.

1. Intégration du profil qualité dans la personnalisation de l'information

L'objectif 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 spécifiques à chaque utilisateur ou communauté d'utilisateurs.
Les données décrivant les utilisateurs sont souvent regroupées sous forme de profils. Parmi les données du profil on trouve une dimension relative à la qualité [KOS03]. Il est par exemple possible de poser une requête en spécifiant des préférences en termes de qualité comme une réponse rapide ou une information fraîche (Figure 1).

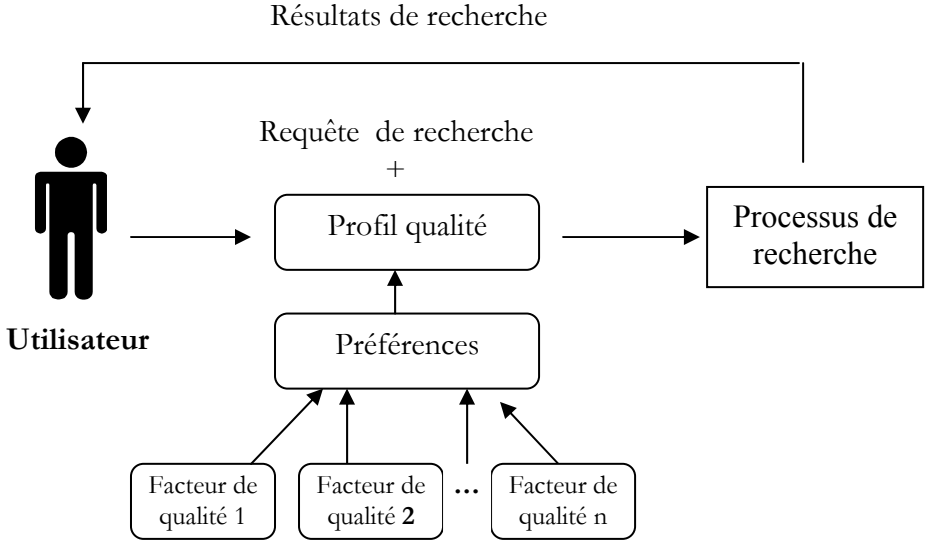


Figure 1- L'intégration du profil qualité dans la recherche information.

2. Objectifs du modèle :

Dans notre modèle la définition des facteurs de qualité influant sur la personnalisation de l'information repose principalement sur cette hypothèse :

Hypothèse : la définition de la qualité de l'information est relative à l'utilisateur.
La définition de la qualité est propre à l'utilisateur c'est- à- dire elle est relative à la satisfaction de ses besoins en termes de choix et d'appréciation des facteurs de la qualité (Figure 2).

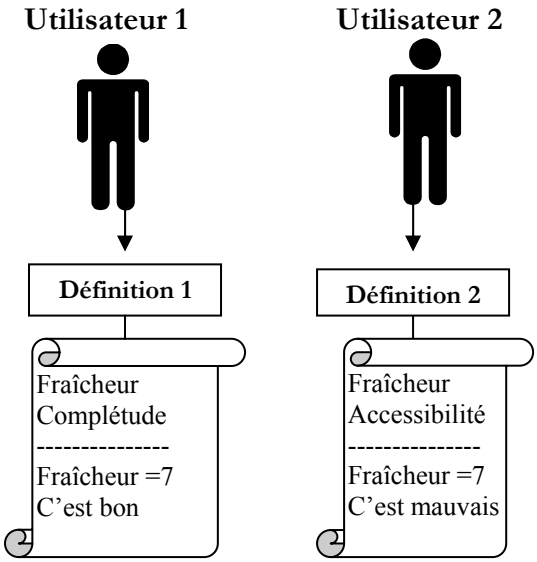


Figure 2- Hypothèse.

L'objectif de notre modèle est de fournir une définition des facteurs de qualité de l'information, afin de permettre à l'utilisateur de construire son propre profil de qualité et d'avoir ainsi une personnalisation au niveau de la définition et de l'évaluation de la qualité.

3. Approche multidimensionnelle pour la qualité

La définition des facteurs de qualité influant sur la personnalisation de l'information ne réside pas dans la définition des facteurs de qualité elle-même mais dans la structuration et la représentation de la qualité.
En se basant sur notre hypothèse, on propose un modèle de qualité hiérarchique orienté utilisateur. Notre hiérarchie de qualité se décompose en un ensemble de dimensions.
On identifie trois types de dimensions (Figure 3) :

- dimensions source : ce type de dimension décrit la source ou la provenance de la qualité comme source d'information ou support d'information. Elle se décompose en une ou plusieurs dimensions utilisateur ou système.
- dimensions utilisateur : dimensions d'agrégation personnalisables par l'utilisateur. Elle se décompose en une ou plusieurs dimensions utilisateurs ou système.
- dimensions système : c'est l'ensemble des critères de la qualité vis-à-vis du système.

Cette dichotomie nous permet de montrer une vue multidimensionnelle et hiérarchique de la qualité. Dans la suite nous proposons une hiérarchie de qualité afin d'assister l'utilisateur dans la construction de son propre profil qualité.

3.1 Dimensions source

On part du constat que s'il est difficile de garantir la qualité intrinsèque de l'information on peut déterminer a priori les sources de qualité (Figure 3) :

- support de l'information : les facteurs de qualité liés aux documents comme la fiabilité ou la fraîcheur.
- source de l'information : les facteurs de qualité du fournisseur de l'information comme par exemple la fraîcheur ou la précision.
- usage de l'information: les facteurs de qualité liés à l'usage des informations comme par exemple les formes de popularité (citation).

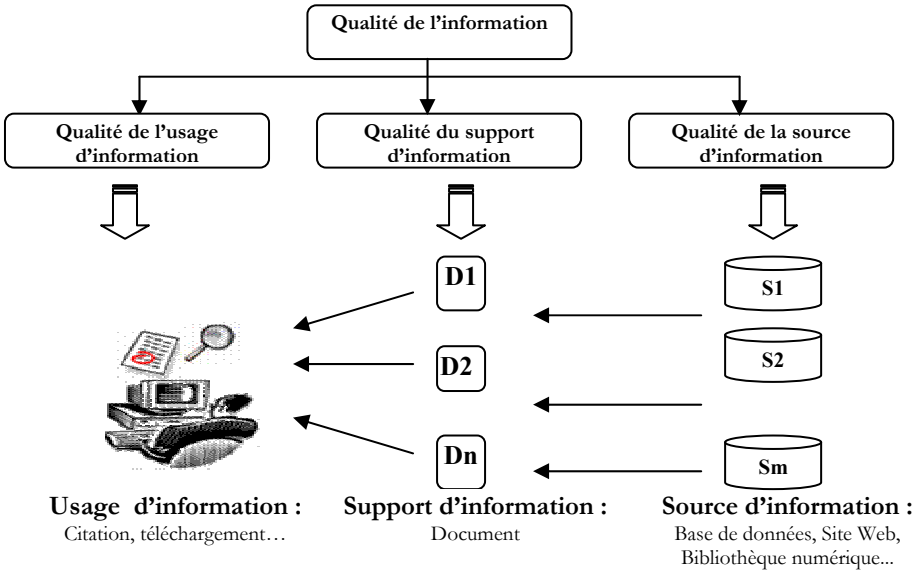


Figure 3- Principales dimensions source de la qualité de l'information.

3.2 Dimensions système

En se basant sur le modèle de **Naumann** et **Rolker** [NAU99b] on propose cet ensemble de critères préliminaires de la qualité :

Dimension source	Dimension système
Source d'information	Temps de réponse, Disponibilité, Prix, Débit réseau, Volume de données, Sécurité d'accès, Support client, Documentation, Concision de représentation, Facilité de compréhension , Relevance,Consistance de représentation Fréquence de mise à jour, Dernière mise à jour Complétude, Exactitude, Crédibilité, Utilité, Objectivité, Réputation, Vérifiabilité.
Support d'information	Temps de réponse, Disponibilité, Prix, Débit réseau, Volume de données, Sécurité d'accès, Support client, Documentation, Concision de représentation, Facilité de compréhension , Relevance, Consistance de représentation , Fréquence de mise à jour, Dernière mise à jour Complétude, Exactitude, Crédibilité, Utilité, Objectivité, Réputation, Vérifiabilité.
Usage d'information	Citation, recommandations, liens, téléchargements.

3.3 Dimensions utilisateur

☐ ***Les principales dimensions***

Dans [MAR02], **Marotta** distingue deux catégories de la qualité du point de vue utilisateur :

- Opérationnel : l'ensemble des facteurs de qualité liés aux mécanismes d'accès à l'information comme temps de réponse et coût.
- Contenu : l'ensemble des facteurs de qualité liés a l'information elle-même comme par exemple la fraîcheur et la complétude.

En s'appuyant sur la catégorisation de la qualité de **Marotta** nous proposons ces principales dimensions (Figure 5) :

- la qualité opérationnelle de la source d'information ou support d'information : l'ensemble des facteurs de qualité liés à l'accès à la source d'information ou support d'information.
- la qualité du contenu de la source d'information ou support d'information : l'ensemble des facteurs de qualité liés à la source d'information ou support d'information elle -même.
- la qualité opérationnelle de l'usage: les diverses formes de popularité liés à l'accès à l'information comme téléchargement ou liens.
- la qualité du contenu de l'usage: les diverses formes de popularité liés à l'appréciation du contenu de l'information comme citation.

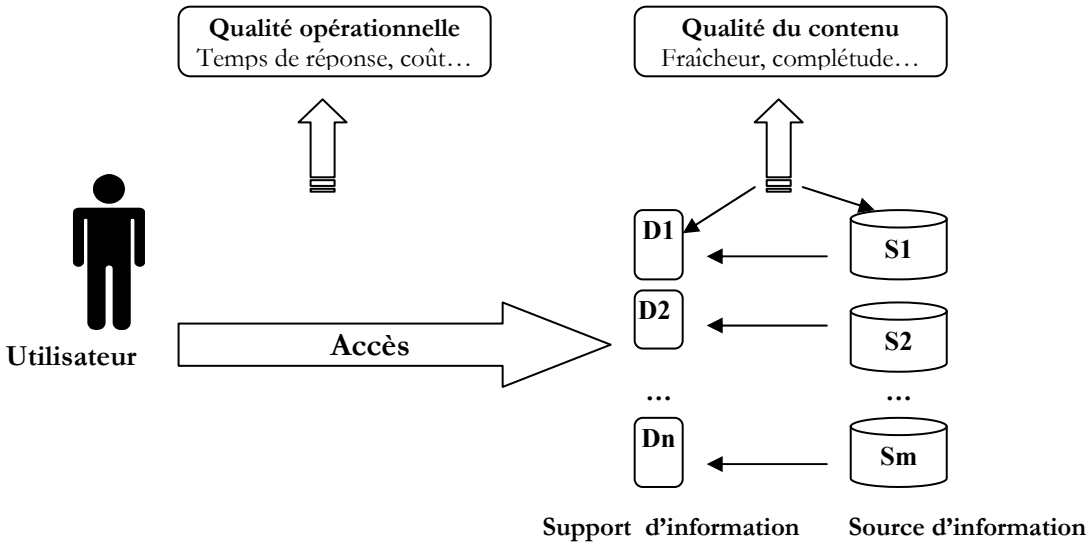


Figure 4- Principales dimensions utilisateur.

☐ ***Les sous-dimensions***

En raison du nombre de dimensions système disponibles dans notre modèle on a besoin d'une simple hiérarchie permettant à l'utilisateur de trouver facilement les dimensions système désirées d'où la proposition des dimensions suivantes :

- **Performance d'accès** : elle se décompose en ***Temps, Coût, Volume*** et ***Sécurité.***
- **Accessibilité** : elle se décompose en ***Assistance*** et ***Manipulation.***
- **Fraîcheur du contenu :** elle se décompose en ***Actualité*** et ***Âge*** [PER04].
- **Fiabilité du contenu** : elle se décompose en ***Complétude*** et ***Exactitude.***

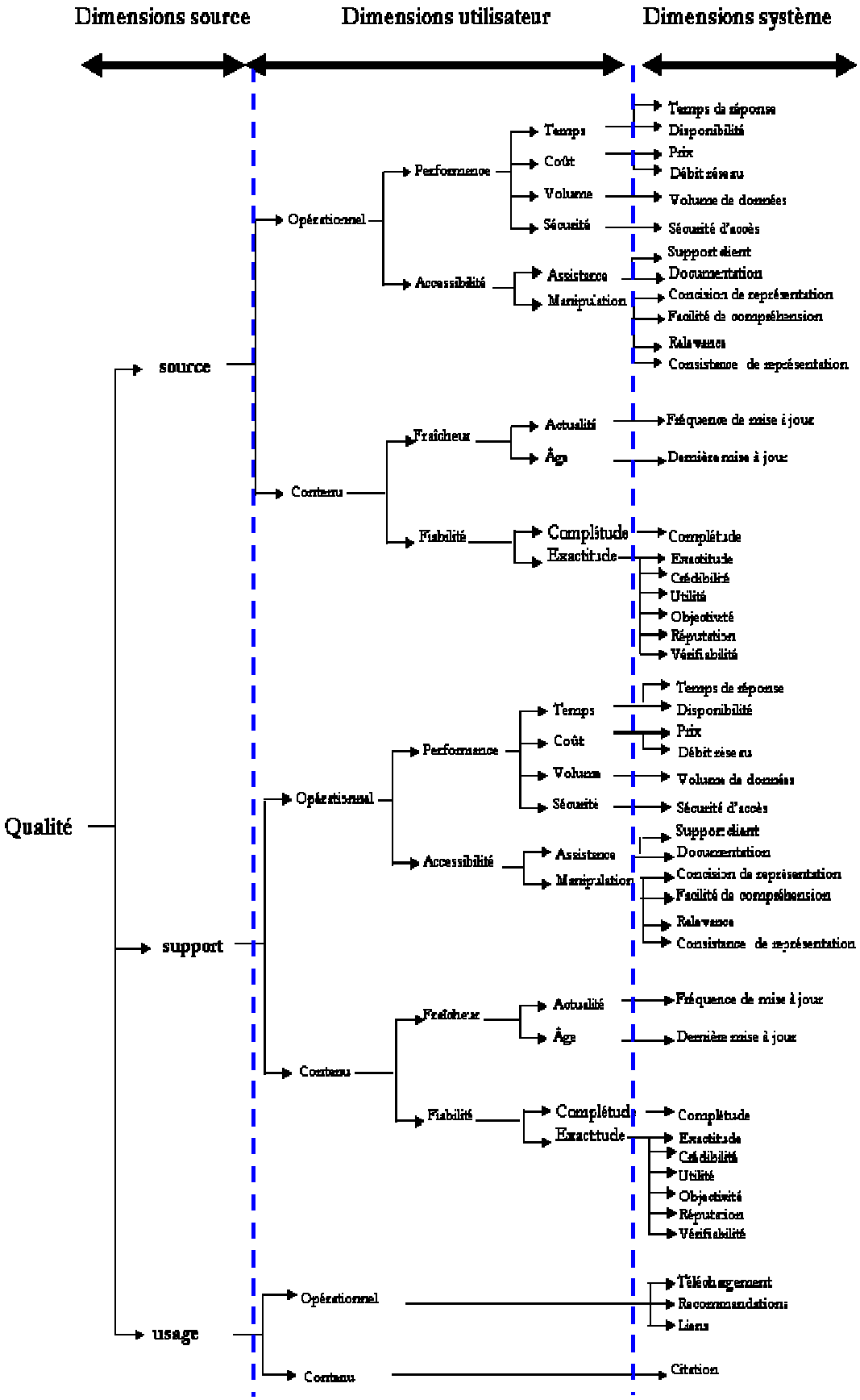


Figure 5- Une hiérarchie d'assistance.

4. Représentation formelle d'un profil qualité

Un profil qualité est l'ensemble des préférences ou besoins en termes de qualité d'information d'un utilisateur ou groupe d'utilisateurs. En se basant sur notre hypothèse les préférences d'un utilisateur résident en :

- une définition personnalisée de la qualité : c'est l'utilisateur qui définit sa propre hiérarchie selon ses besoins et ses exigences.
- une évaluation personnalisée de la qualité : c'est l'utilisateur qui évalue et apprécie la qualité.

4.1 Définition personnalisée de la qualité

On définit :

- Q: dimension qualité.
  Elle représente la qualité générale de l'information.
- D_S: l'ensemble des dimensions source de la qualité.
- D_U: l'ensemble des dimensions utilisateur de la qualité.
- D_{Sy}: l'ensemble des dimensions système de la qualité.
- D_h : l'ensemble des dimensions qui constitue la hiérarchie de la qualité.
  $D_h = Q \cup D_S \cup D_U \cup D_{Sy}$
- A une fonction qui décrit les sous-dimensions de chaque dimension.
  $A: D_h \rightarrow D_h \quad d \rightarrow A(d) = \{d_1, d_2 \ldots d_n\}$

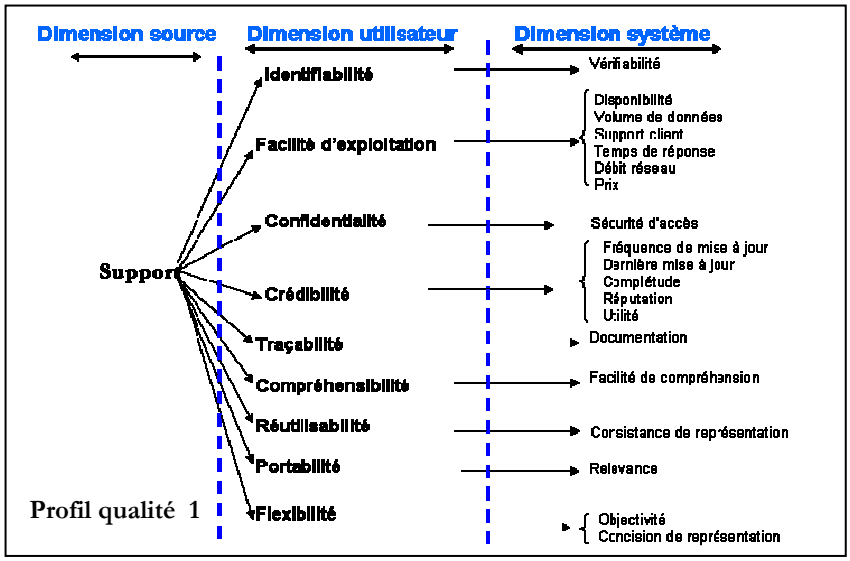
On définit une hiérarchie de la qualité **HQ** comme un ensemble de couples (d, A(d)) pour toute dimension d appartenant à D_h: $HQ = \{(d, A(d)) / d \in D_h\}$.
Cette hiérarchie est formée d'arbres qui vérifient ces propriétés :

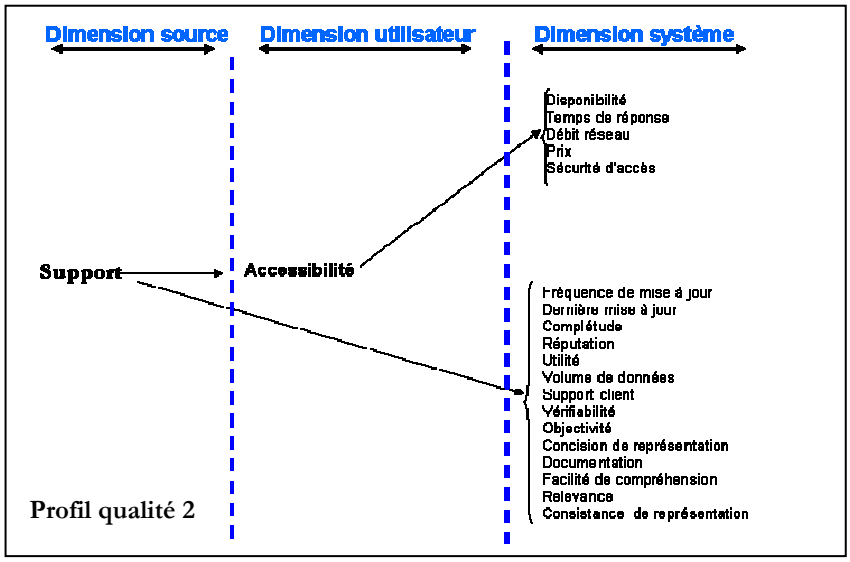
- *Propriété 1*: $\forall d_1 \in Dh, d_2 \in D_h / d_1 \neq d_2 \quad A(d_1) \cap A(d_2) = \emptyset$.
  Deux dimensions sont disjointes : elles n'ont pas de sous-dimensions communes.
- *Propriété 2*: $A(Q) = D_S$.
  La dimension qualité se décompose en dimensions source.
- *Propriété 3*: $\forall d \in D_S \quad A(d) \subset D_U \cup D_{Sy}$
  Une dimension source se décompose en sous-dimensions utilisateur ou système.
- *Propriété 4*: $\forall d \in Du \quad A(d) \subset D_U \cup D_{Sy}$.
  Une dimension utilisateur se décompose en sous-dimensions utilisateur ou système.
- *Propriété 5*: $\forall d \in D_{Sy} \quad A(d) = \emptyset$.
  Une dimension système n'a pas de sous-dimensions.
- *Propriété 6*: $\cup A(d) = D_h$.
  Couverture de toutes les dimensions de la hiérarchie.

Une définition personnalisée de la qualité est une hiérarchie personnalisée **HQ** où D_u est l'ensemble de dimensions définis par l'utilisateur, D_s et D_{Sy} sont fixées par notre modèle.

Par exemple dans la figure 6 on a trois définitions de la qualité :

- le profil qualité 1 : pour la dimension support, l'utilisateur définit huit dimensions et il détermine les sous-dimensions système qui la composent.
- le profil qualité 2 : pour la dimension support, l'utilisateur définit une seule dimension utilisateur et il détermine les sous-dimensions système qui la composent.
- le profil qualité 3 : pour la dimension support, l'utilisateur utilise notre hiérarchie d'assistance.





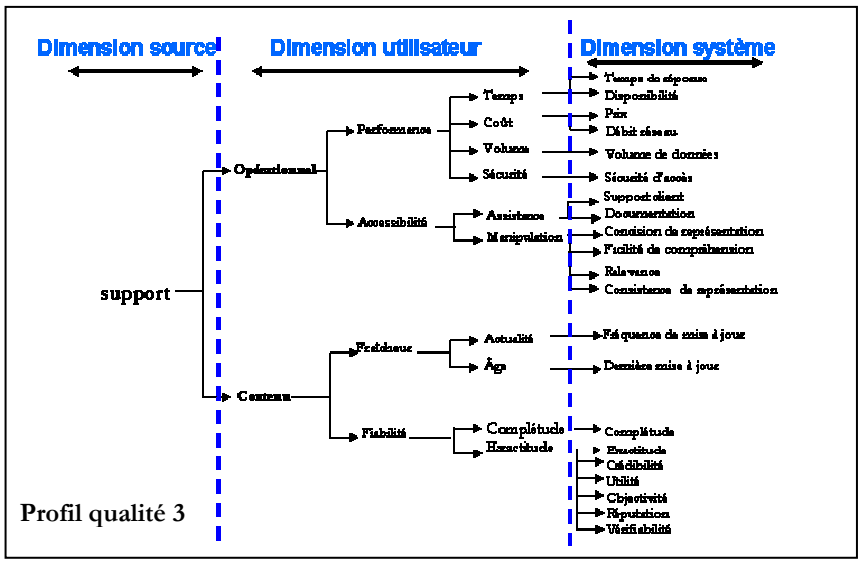


Figure 6- Profils qualité : définition.

4.2 Évaluation personnalisée de la qualité

Une évaluation personnelle de la qualité se définit comme une sous-hiérarchie de la définition de la qualité (Figure 7).

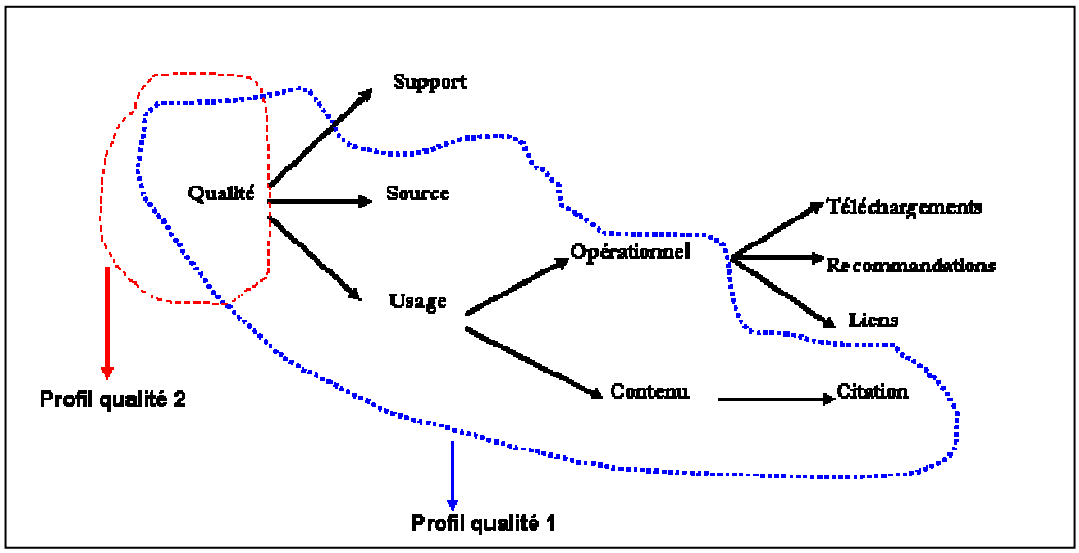


Figure 7- Profil qualité : Evaluation

On définit **Pref** (D, O, P, I) comme une préférence utilisateur par rapport à une dimension où

- D : dimension de qualité
- O : ordre de préférence (1 ou 0 : avec détail ou sans détail)
- P : poids attribué à D qui représente le degré d'importance de la dimension $0 \leq pi \leq 1$.
- I : intervalle de confiance / c'est l'intervalle de mesures de la qualité préférée.
- $\sum Pi=1$ la somme des poids attribués aux sous-dimensions de D vaut 1.

-Préférence d'ordre 1 : l'utilisateur s'intéresse à la dimension avec détail (les sous-dimensions). Les paramètres D, O, P et I des différentes sous-dimensions sont fixées explicitement par l'utilisateur (Figure 8).

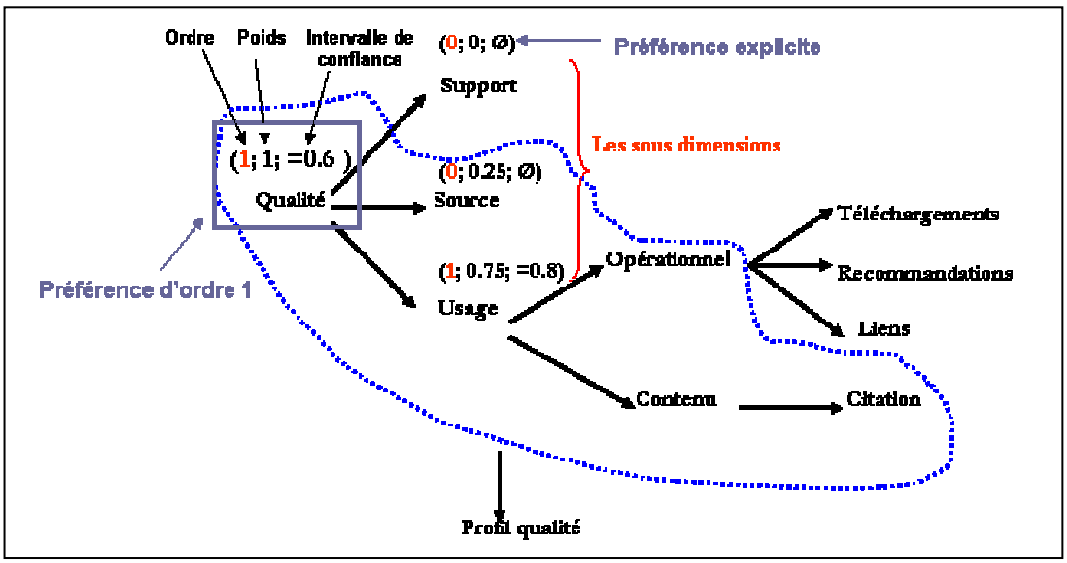


Figure 8- Préférence d'ordre 1

-Préférence d'ordre 0 : l'utilisateur s'intéresse à la dimension sans détail (les sous-dimensions) (Figure 9).

Hypothèse1 : les degrés d'importance des différentes sous-dimensions sont égaux.

En se basant sur cette hypothèse on déduit :

- Pour chaque d_i de A(D) = {d_1,…, d_n} alors le poids p_i =1/n.
- Ordre (d_i) =0 : l'utilisateur ne s'intéresse pas aux des détails.
- I de chaque d_i = Ø pas d'intervalle de confiance.

Exemple :

Pref (opérationnel ; 0; 0.4 ; Ø) signifie que l'utilisateur s'intéresse à la qualité opérationnelle sans détail avec un degré d'importance 0.4 par rapport à la qualité du contenu. Il s'agit d'une préférence explicite. Celle-ci va générer des préférences implicites : Pref (Téléchargements; 0; 0.33 ; Ø), Pref (Recommandations; 0; 0.33 ; Ø), Pref (Liens; 0; 0.33 ; Ø) (Figure 9)

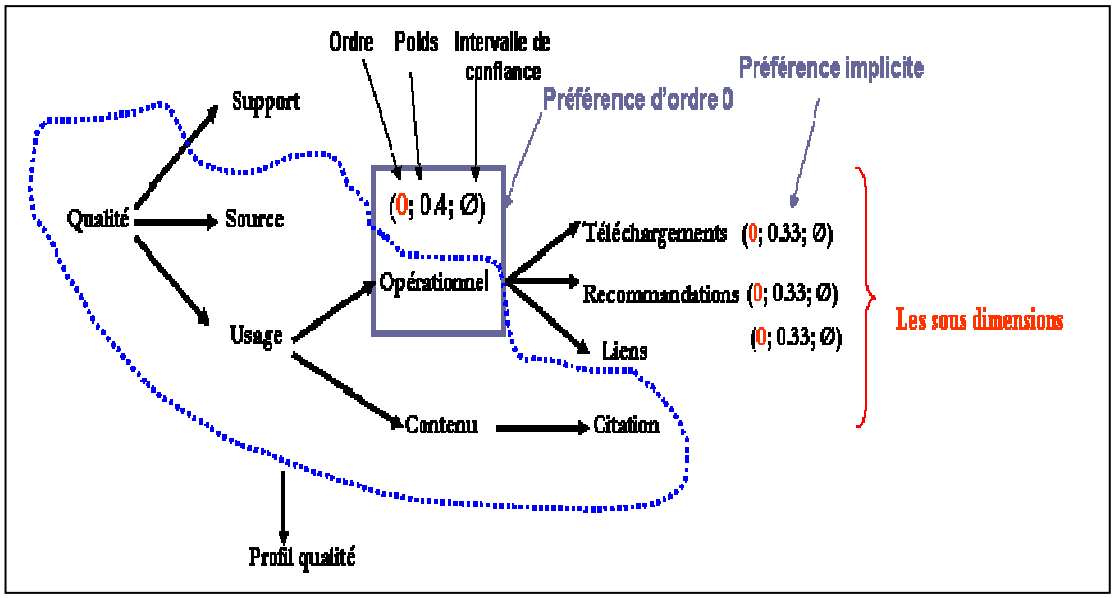


Figure 9- Préférence d'ordre 0

4.3 Définition formelle d'un profil qualité

Un profil qualité se définit comme un ensemble de préférences sur toutes les dimensions de la hiérarchie personnalisée **HQ** de la qualité.

Profil = {Pref (D, O, P, I) / D ∈ Dh}

Où

Dh : l'ensemble des dimensions de la hiérarchie de la qualité.

D: dimension de qualité.

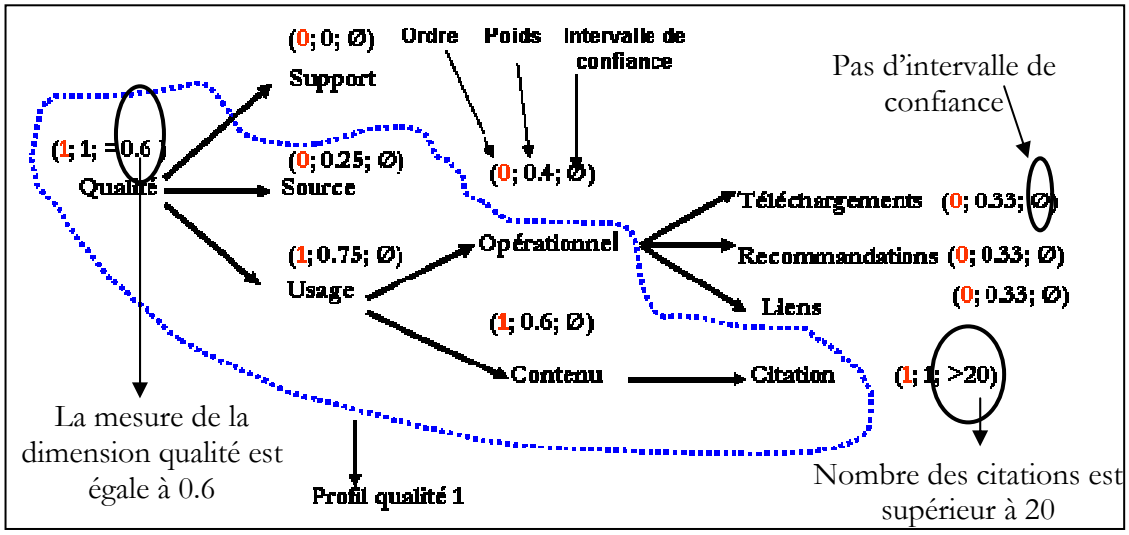
O: ordre de préférence.

P : poids attribué à D.

I : intervalle de confiance de la qualité.

Exemple :

Dans la figure 10 :
Profil 1= {**Pref**(Qualité, 1, 1,=0.6), **Pref** (Usage, 1, 0.75, Ø), **Pref** (Opérationnel, 0, 0.4, Ø), **Pref** (Contenu, 1, 0.6, Ø), **Pref** (Citation, 1, 1, >20), **Pref** (Téléchargements, 0, 0.33, Ø), **Pref** (Liens, 0, 0.33, Ø), **Pref** (Recommandations, 0, 0.33, Ø), **Pref**(Source,1, 1,=0.6), …}.



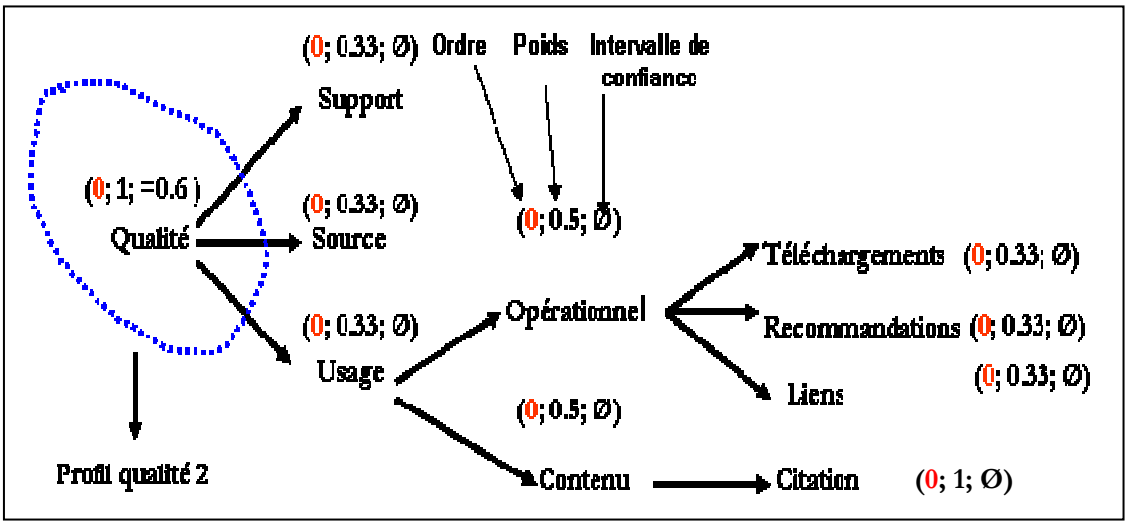


Figure 10- Profils qualité

5. Conclusion

Dans ce chapitre nous avons défini un modèle multidimensionnel de la qualité de l'information permettant à l'utilisateur de définir son profil de qualité. Dans le chapitre suivant nous étudions les différentes méthodes d'évaluation de la qualité dans notre modèle.

Chapitre 3
Méthodes d'évaluation de la qualité

Dans le chapitre précédent nous avons proposé un modèle de qualité de l'information permettant à l'utilisateur de construire son propre profil de qualité.

Dans ce chapitre nous abordons les différentes méthodes d'évaluation de la qualité dans la hiérarchie de la qualité proposée.

1. Méthodes de calcul de score de qualité

1.1 Méthodes d'analyse multicritère

L'analyse multicritère ou les méthodes d'aide à la décision multicritères désignent un ensemble de méthodes permettant d'agréger plusieurs critères avec l'objectif de sélectionner une ou plusieurs décisions. La différence entre ces méthodes se trouve soit dans la façon d'agréger les jugements pour choisir la décision la plus satisfaisante, soit dans la façon d'évaluer chacune des décisions en fonction des critères retenus.
Parmi les méthodes d'analyse multicritères les plus connues on peut citer :

- *la méthode SAW*
  La méthode SAW (Simple Additive Weighting) est une méthode multicritère caractérisée par une agrégation additive des critères par sommation pondérée.
- *la méthode TOPSIS*
  La méthode TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) est une méthode multicritère développée par Hwang et Yoon (1981) se basant sur une relation de dominance qui résulte de la distance par rapport à la solution idéale. Elle se caractérise par une agrégation compensatoire entre les critères
- *la méthode AHP*
  L'analyse procédurale hiérarchique AHP (Analytic Hierarchy Process), développée par Thomas Saaty (1971), est une méthode de prise de décision qui décompose le problème en structure hiérarchique de choix ou objectifs, par la suite en réduisant les décisions complexes en séries de simples paires de comparaison et en synthétisant les résultats.
- *la méthode DEA*
  L'analyse de l'enveloppement des données (Data Envelopment Analysis) est une méthode introduite par Thomas Charnes, Cooper et Rhodes (1978). Cette méthode détermine le degré d'efficacité d'une solution en résolvant un programme linéaire, pour trouver la valeur optimale. Dans la méthode DEA les variables du programme sont les poids des critères qui ne sont pas spécifiées par l'utilisateur.

1.2 Sélection d'une méthode de calcul de score

Dans notre modèle les préférences utilisateur sont exprimées essentiellement par une affectation de poids aux dimensions de la hiérarchie de la qualité et la stratégie d'évaluation de la qualité repose essentiellement sur le calcul de score de la qualité des dimensions à partir des scores des sous-dimensions. En se basant sur la comparaison de **Naumann** [NAU98] et l'étude de complexité de **Burgess** [BUR03] notre choix de la méthode de calcul se fait selon les propriétés suivantes :

- interaction : l'utilisateur peut exprimer des préférences sur les critères (affectation d'un poids numérique aux critères de qualité).
- dominance : une source d'information domine l'autre si elle est égale ou meilleure pour tous les critères et meilleure au moins pour un critère.
- type de résultat : on distingue deux types de résultats : ordonnancement (score de qualité) ou classement.
- critères positifs et négatifs : la méthode distingue entre les critères négatifs (prix, début réseau, temps de réponse…) et les critères positifs (disponibilité, complétude …).
- complexité : la complexité de l'algorithme de calcul de scores de qualité.

Propriété	SAW	TOPSIS	AHP	DEA
Interaction	poids	poids	décisions	-
Dominance	oui	oui	oui	oui
Type de résultat	ordonnancement	ordonnancement	ordonnancement	classement
Critères positifs et négatifs	oui	non	non	non
Complexité	O (n.m)	O (n.m)	exponentielle	polynomiale

Tableau 1- Comparaison des méthodes d'analyses multicritères.

D'après l'étude comparative des méthodes d'analyses multicritères (tableau 1) on conclut que la méthode **SAW** nous permet d'avoir une meilleure sélection des propriétés utilisées ultérieurement dans notre stratégie d'évaluation de la qualité, à savoir l'affectation du poids numérique, l'ordonnancement des résultats, la distinction entre critère positif et négatif ainsi que la complexité. C'est pour cela, que nous la proposons comme une méthode de calcul de score de qualité dans notre modèle.

1.3 Méthode SAW (Simple Additive Weighting)

L'algorithme de calcul de score dans la méthode SAW se décompose en trois étapes :

☐ **Étape 1** : *création de la matrice de décision*

La première étape est la création de la matrice de décision qui représente les données disponibles et les critères de qualité (tableau2).

	Critères			
Alternatives	C_1	C_2	…	C_n
Poids relatives	W_1	W_2		W_n
A_1	$\mathbf{d_{11}}$	$\mathbf{d_{12}}$	**…**	$\mathbf{d_{1n}}$
A_2	$\mathbf{d_{21}}$	$\mathbf{d_{22}}$	**…**	$\mathbf{d_{2n}}$
…	**…**	**…**	**…**	**…**
A_m	$\mathbf{d_{m1}}$	$\mathbf{d_{m2}}$	**…**	$\mathbf{d_{mn}}$

Tableau 2- Matrice de décision.

A_i est la $i^{éme}$ alternative, Cj est le $j^{éme}$ critère et dij est la mesure de la performance de la $i^{éme}$ alternative pour le $j^{éme}$ critère. Dans notre modèle, les critères sont les dimensions, et les alternatives sont les sources ou support d'information.

☐ **Étape 2** : *création de la matrice de décision normalisée*

Cette étape consiste à normaliser les poids et les mesures de performance afin d'obtenir des valeurs entre 0 et 1. La normalisation se fait selon les formules suivantes :

Pour les poids : $w'_j = \dfrac{w_j}{\sum_{j=1}^{m} w_j}$

Pour les mesures :

$$v_{ij} = \frac{d_{ij} - d_j^{\min}}{d_j^{\max} - d_j^{\min}} \quad \text{Pour les critères positifs}$$

$$v_{ij} = \frac{d_j^{\max} - d_{ij}}{d_j^{\max} - d_j^{\min}} \quad \text{Pour les critères négatifs}$$

$$\text{Où} \quad d_j^{\max} \equiv \max_i [d_{ij}] \text{ et } \quad d_j^{\min} = \min_i [d_{ij}]$$

☐ **Étape 3** : *calcul du score de chaque alternative.*

Le score de qualité de chaque alternative est donné par cette formule :

$\text{Score}(A_i) = \sum w_j v_{ij} \qquad 0 \le \text{Score}(A_i) \le 1$

2. Stratégies d'évaluation de la qualité

Notre stratégie d'évaluation consiste à calculer les scores des différentes dimensions de la hiérarchie de la qualité définie dans un profil qualité. Dans notre définition d'un profil l'utilisateur peut attribuer un poids et un intervalle de confiance à chacune des dimensions de qualité selon l'importance qu'il lui accorde. Ainsi notre stratégie doit prendre en compte les préférences définies dans un profil.

On propose deux stratégies de calcul du score des différentes dimensions :

2.1 Stratégie 1 : évaluation des dimensions à partir des sous-dimensions qui la composent.

C'est la démarche la plus intuitive. Elle repose sur les étapes suivantes (Figure 1) :

①Évaluation des dimensions système.
②Évaluation des dimensions utilisateur.
③Évaluation des dimensions source.
④Évaluation de la qualité de l'information.

Les poids attribués aux différentes dimensions sont définis par le profil qualité.

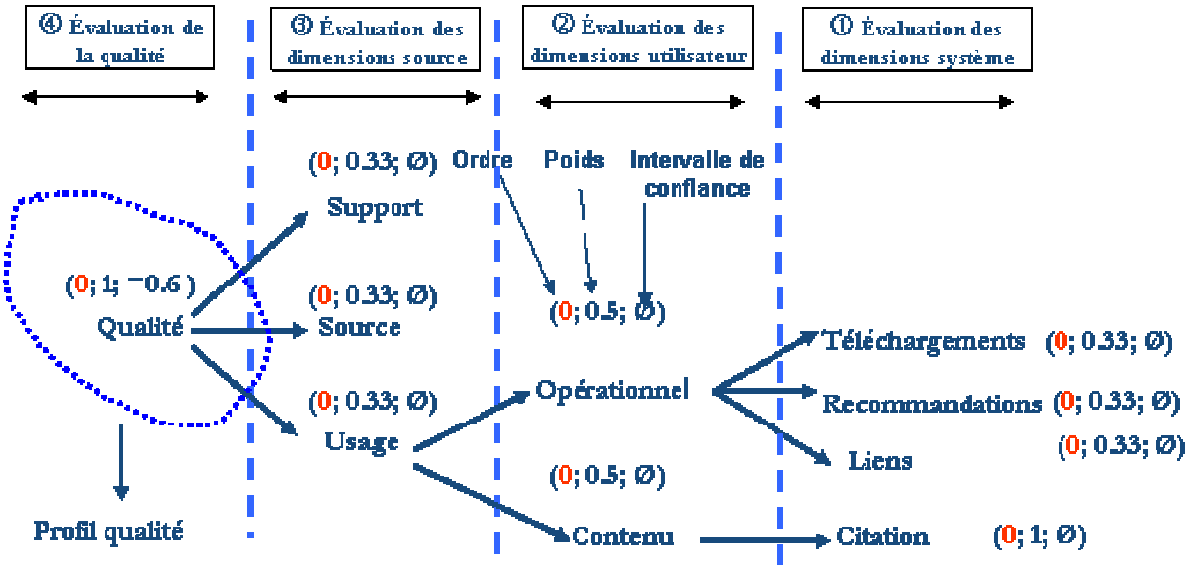


Figure 1- Évaluation des dimensions à partir des sous-dimensions.

2.2 Stratégie 2 : évaluation des dimensions à partir des dimensions système qui la composent.

La démarche d'évaluation repose sur les étapes suivantes (Figure2):

① Évaluation des dimensions système.

② Évaluation des dimensions source, utilisateur et qualité de l'information à partir des dimensions système.

Les poids attribués aux différentes dimensions système sont dérivés à partir de la propagation de poids (Figure 2). La somme des poids vaut 1.

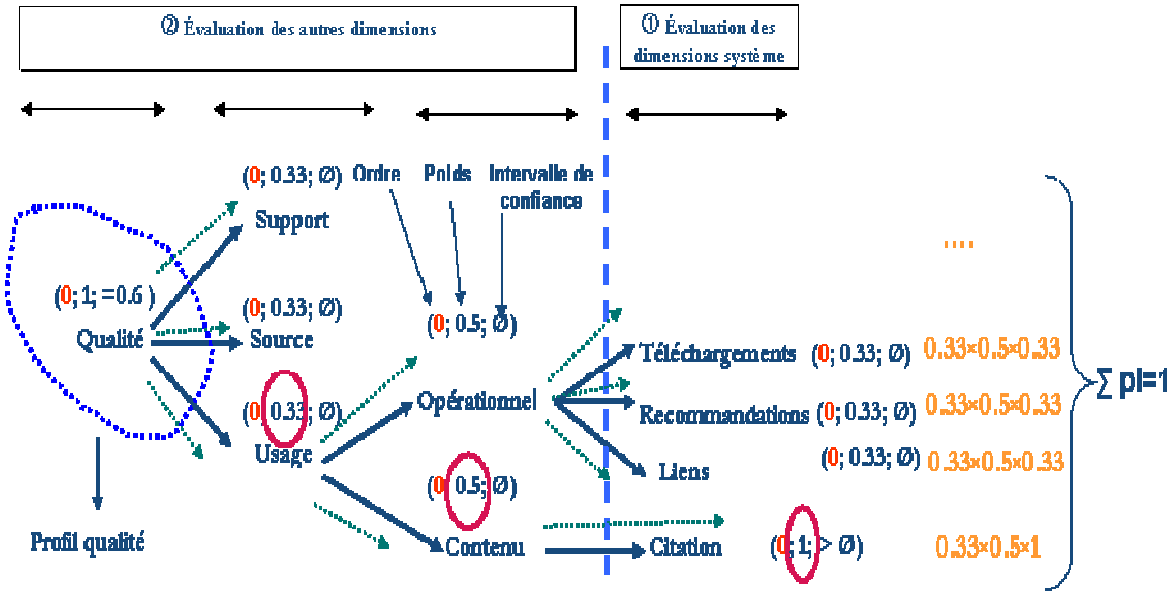


Figure 2- Évaluation des dimensions à partir des dimensions système.

□ **Exemple**

Pour la dimension *Usage* les nouveaux poids des différentes sous-dimensions système :
- $Poids'_{Citation} = Poids_{usage} * Poids_{Contenu} * Poids_{Citation}$ (0.5*1=0.5).
- $Poids'_{Liens} = Poids_{usage} * Poids_{Contenu} * Poids_{Liens}$ (0.33*0.5=0.17).
- $Poids'_{Téléchargements} = Poids_{usage} * Poids_{Contenu} * Poids_{Téléchargements}$ (0.33*0.5=0.17).
- $Poids'_{Recommandations} = Poids_{usage} * Poids_{Contenu} * Poids_{Recommandations}$ (0.33*0.5=0.17).

$$Poids'_{Citation} + Poids'_{Liens} + Poids'_{Téléchargements} + Poids'_{Recommandations} = 1$$

Pour la *qualité générale* les nouveaux poids des différentes sous-dimensions système :
- $Poids'_{Citation} = Poids_{usage} * Poids_{Contenu} * Poids_{Citation}$ (0.33*0.5*1=0.17).
- $Poids'_{Liens} = Poids_{usage} * Poids_{Contenu} * Poids_{Liens}$ (0.33*0.5*0.33=0.05).
- $Poids'_{Téléchargements} = Poids_{usage} * Poids_{Contenu} * Poids_{Téléchargements}$ (0.33*0.5*0.33=0.05).
- $Poids'_{Recommandations} = Poids_{usage} * Poids_{Contenu} * Poids_{Recommandations}$ (0.33*0.5*0.33=0.05).
- …

2.3 Sélection d'une stratégie

Notre choix d'une stratégie d'évaluation de la qualité se fait selon ces propriétés **:**

- complexité : complexité de calcul de score de qualité pour toutes les dimensions de la hiérarchie.
- propagation d'erreur : on calcule l'erreur de calcul sur les différentes dimensions en fonction des erreurs sur les dimensions système.
-

Par exemple pour la hiérarchie de qualité d'usage, où ε_j est l'erreur absolue sur les dimensions système : téléchargement, recommandations, liens et citation, on a établi ce tableau :

Stratégie	Complexité	Propagation d'erreur
Stratégie 1	O (6.n)	$\sum_{j=1}^{j=4} w_j \varepsilon_j$
Stratégie 2	O (8.n)	$\sum_{j=1}^{j=4} w_j \varepsilon_j$

On remarque que la propagation d'erreur est la même dans les deux stratégies, même les résultats de calcul de score sont les mêmes. Cela est dû à la linéarité de la fonction de calcul de score dans la méthode SAW. Au niveau de la complexité, dans la stratégie 2 elle est supérieure à celle de la stratégie 1. En effet dans la stratégie 1 le nombre de dimensions diminue à chaque étape d'évaluation de la qualité. C'est pourquoi on la propose comme une stratégie d'évaluation pour notre modèle.

3. Conclusion

Dans ce chapitre nous avons défini les différentes méthodes d'évaluation de la qualité d'information dans la hiérarchie. Dans le chapitre suivant nous allons étudier l'impact de la qualité sur l'accès aux informations.

Chapitre 4

Intégration de la qualité dans le processus d'accès à l'information

Dans ce chapitre nous déterminons l'impact de la qualité sur la personnalisation de l'information c'est-à-dire selon le profil de l'utilisateur ou selon ses exigences explicites (en termes de qualité de informations), le processus d'accès aux informations ne présentera que les informations de qualité les plus adaptées.

1. Techniques d'accès à l'information

Les techniques d'accès à l'information permettent à un individu d'obtenir des informations répondant à ses besoins. Nous pouvons les regrouper en deux grands groupes [TCH04] :

- celles qui reposent sur une approche « service au comptoir » ou « pull » : qui consistent à renvoyer des informations répondant à une demande explicite d'un individu. C'est le cas de la Recherche d'Information.
- celles qui reposent sur une approche « service à domicile » ou « push » : qui consistent à renvoyer automatiquement à un individu des informations qui pourraient l'intéresser, sans qu'il n'en ait fait explicitement la demande. C'est le cas du Filtrage (ou Recommandation) d'Information.

Le processus de Recherche d'Information repose sur l'expression du besoin d'un individu au travers d'une requête formulée dans un langage libre plus ou moins structuré. En réponse à cette requête, un appariement est réalisé entre les termes (ou mots-clés) d'indexation de la requête et ceux des informations pré-indexées par le système. La recherche d'information est principalement basée sur le principe d'un appariement optimal, de type vectoriel ou probabiliste. Enfin, le système propose traditionnellement à l'individu les informations pertinentes sous forme d'une liste ordonnée selon leur degré de pertinence décroissant.

Alors que la Recherche d'Information (RI) est une tâche très interactive, celle du Filtrage d'Information (FI) est relativement passive car l'utilisateur ne formule pas explicitement ses besoins au travers d'une requête (ou expression d'un besoin ponctuel) comme c'est le cas en RI. En Filtrage d'Information, on utilise plutôt une représentation de l'utilisateur appelé profil utilisateur pour lui envoyer des informations. Elles sont ensuite comparées aux différents profils disponibles pour déterminer ceux auxquels elles correspondent.

2. Filtrage multidimensionnel d'informations selon leurs qualités

Dans cette partie nous proposons l'exploitation des profils utilisateurs en termes de qualité dans le processus de filtrage d'informations. Les informations sont filtrées selon les différentes dimensions de la hiérarchie de la qualité définie dans le profil qualité de l'utilisateur.

2.1 Principe de l'approche de filtrage

Notre approche de filtrage repose sur les étapes suivantes (Figure 1):

- *Filtrage des sources d'information* : Dans notre hiérarchie de la qualité d'information on a défini une dimension *source d'information.* Cette dimension représente la qualité du fournisseur d'information (site web, bibliothèque numérique, base de données…). Notre approche est de filtrer les sources d'informations qui contiennent potentiellement les réponses des requêtes selon les préférences en termes de qualité. Les sources d'information dont la qualité des différentes dimensions de la source n'appartient pas à l'intervalle de confiance défini dans le profil qualité sont éliminées.

- *Filtrage des résultats de la requête* : Dans notre hiérarchie de la qualité d'information on a défini les dimensions : *support d'information, usage d'information et la qualité générale.* Ces dimensions représentent respectivement la qualité du document, le support physique de l'information, la qualité d'usage du document et sa qualité générale. Notre approche vise à filtrer les documents, qui représentent les résultas de la requête de recherche, selon les préférences en termes de qualité. Les documents dont la qualité des différentes dimensions du support ou usage n'appartient pas à l'intervalle de confiance défini dans le profil qualité sont éliminés.

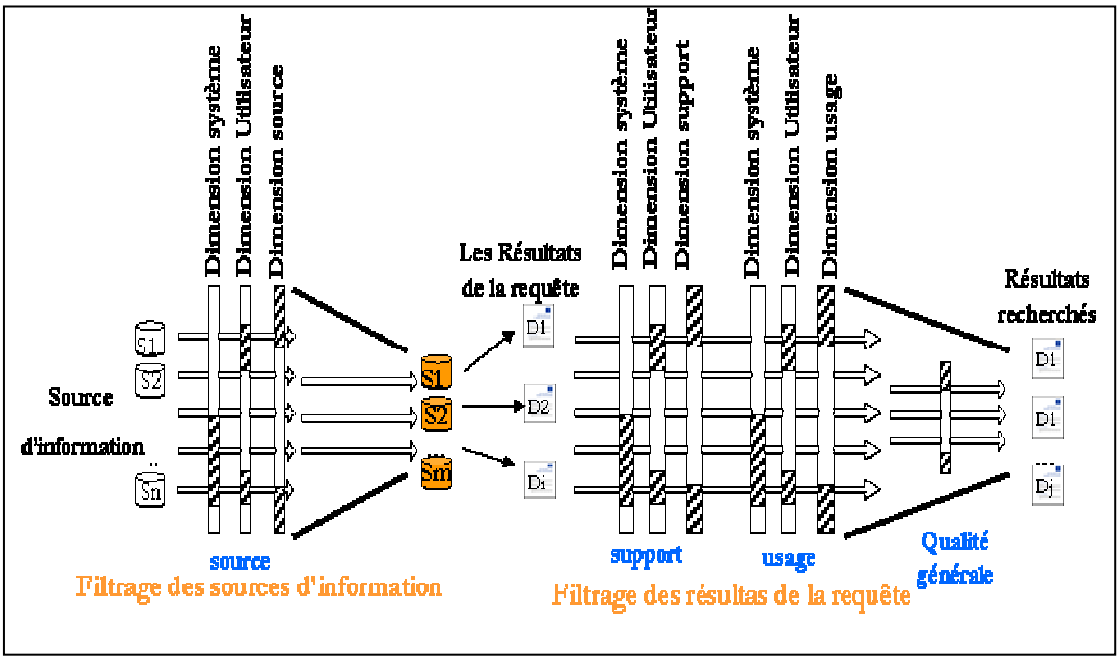


Figure 1- Filtrage d'informations.

Afin d'optimiser le filtrage on évalue tout d'abord les sous-dimensions puis on applique le filtrage selon un ordre croissant des intervalles de confiance des sous-dimensions.

2.2 Exemple

En se basant sur la stratégie d'évaluation de la qualité et la méthode de calcul de score SAW défini dans le chapitre 3 on présente dans cet exemple comment intégrer le profil qualité dans le filtrage d'information. Le tableau 1 présente un exemple des préférences exprimées dans un profil qualité.

Dimension source	Dimension utilisateur	Poids	Intervalle de confiance	Dimension système	Poids	Intervalle
Source d'informations Poids= 4 Intervalle ≥0,5	Accessibilité	9	[0,3 ; 1]	Temps de réponse	6	[2 4]
				Disponibilité	5	Ø
				Prix	4	<5
				Débit réseau	7	Ø
	Fraîcheur	3	>0,1	Fréquence de MAJ	2	Ø
				Dernière MAJ	8	Ø
Support d'informations Poids= 8 Intervalle ≥0,5	Accessibilité	3	>0,4	Temps de réponse	2	Ø
				Disponibilité	2	Ø
				Prix	10	Ø
				Débit réseau	2	Ø
Usage d'informations Poids= 6 Intervalle = Ø	Popularité	9	Ø	Téléchargements	5	Ø
				Recommandations	5	Ø
				Liens	5	Ø
				Citations	10	>20

Tableau 1- Exemple d'un profil de qualité

□ **Filtrage des sources d'information :**

⇒ *Filtrage des sources d'information selon la hiérarchie de qualité source.*

- Exemple d'évaluation des dimensions système des sources d'information.

Source	Temps de réponse	Disponibilité	Prix	Débit réseau	Fréquence de MAJ	Dernière MAJ
S_1	4	4	1	20	5	14
S_2	7	3	4	30	3	10
S_3	8	5	5	20	4	12
S_4	3	7	6	15	5	18
S_5	2	9	7	8	10	20
S_6	5	10	2	25	2	30
S_7	6	5	2	30	3	15
S_8	1	6	2	40	5	10
S_9	1	8	3	10	6	20

- Le filtrage des sources d'information selon la dimension système *Temps de réponse* (intervalle de confiance le plus petit) donne *S_1, S_4, S_5, S_8 et S_9.*
- Le filtrage des sources d'information *S_1, S_4, S_5, S_8 et S_9* selon la dimension système *Prix* donne *S_1, S_8 et S_9.*

- Évaluation du score de qualité de la dimension utilisateur *accessibilité* à partir des sous-dimensions système : *Temps de réponse, Disponibilité, Prix* et *Débit réseau.*

Matrice de décision					**Matrice de décision normalisée**				**Calcul du score**
Source	Temps de réponse	Disponibilité	Prix	Début réseau	Temps de réponse	Disponibilité	Prix	Débit réseau	Score de qualité
Poids	6	5	4	7	0,273	0,227	0,182	0,318	
S_1	4	4	1	20	0,000	0,000	1,000	0,667	**0,394**
S_8	1	6	2	40	1,000	0,500	0,500	0,000	**0,477**
S_9	1	8	3	10	1,000	1,000	0,000	1,000	**0,818**

- Le filtrage des sources d'information selon l'intervalle de confiance de la dimension *accessibilité* (**I** = [0,3 1]) donne S_1, S_8 *et* S_9.
- Évaluation du score de la qualité de la dimension utilisateur *Fraîcheur* à partir des sous - dimensions systèmes *Fréquence de MAJ* et *Dernière MAJ.*

Matrice de décision			**Matrice de décision normalisée**		**Calcul du score**
Source	Fréquence de MAJ	Dernière MAJ	Fréquence de MAJ	Dernière MAJ	Score de qualité
Poids	2	8	0,200	0,800	
S_1	5	14	0,000	0,600	**0,480**
S_8	5	10	0,000	1,000	**0,800**
S_9	6	20	1,000	0,000	**0,200**

- Le filtrage des sources d'information selon l'intervalle de confiance de la dimension *fraîcheur* (**I** > 0,1) donne S_1, S_8 *et* S_9.
- Évaluation du score de la qualité de la dimension *Source* à partir des sous- dimensions utilisateur *accessibilité* et *fraîcheur.*

Matrice de décision			**Matrice de décision normalisée**		**Calcul du score**
Source	Accessibilité	Fraîcheur	Accessibilité	Fraîcheur	Score de qualité
Poids	9	3	0,750	0,250	
S_1	0,394	0,480	0,394	0,480	**0,415**
S_8	0,477	0,800	0,477	0,800	**0,558**
S_9	0,818	0,200	0,818	0,200	**0,664**

- Le filtrage des sources d'information selon l'intervalle de confiance de la dimension *Source* (**I** ≥0,5) donne S_8 *et* S_9.

□ **Filtrage des résultats de la requête**

⇒ *Filtrage des résultas selon la hiérarchie de la qualité d'usage.*

- Exemple d'évaluation des dimensions système de l'usage d'information.

Source	Document	Téléchargements	Recommandations	Citations	Liens
S_8	D_1	4	1	30	5
	D_2	3	4	30	3
	D_3	5	5	20	4
	D_4	7	6	15	5
	D_5	9	7	8	10
S_9	D_6	10	2	25	2
	D_7	5	2	30	3
	D_8	6	2	40	5
	D_5	8	3	10	6

- Le filtrage des documents selon la dimension système *Citations* donne D_1, D_2, D_6, D_7 *et* $D8$.

- Évaluation du score de la qualité de la dimension utilisateur *Popularité à* partir des sous-dimensions système : *Téléchargements, Recommandations, Citations et Liens.*

Matrice de décision					**Matrice de décision normalisée**				**Calcul du score**
Document	Télécharg	Recomm	Cita	Liens	Télécharg	Recomm	Cita	Liens	Score de qualité
Poids	5	5	10	5	0,200	0,200	0,400	0,200	
D_1	4	1	30	5	0,143	0,000	0,333	1,000	**0,362**
D_2	3	4	30	3	0,000	1,000	0,333	0,333	**0,644**
D_6	10	2	25	2	1,000	0,333	0,000	0,000	**0,267**
D_7	5	2	30	3	0,286	0,333	0,333	0,333	**0,124**
D_8	6	2	40	5	0,429	0,333	1,000	1,000	**0,819**

- Le filtrage des documents selon la dimension *utilisateur* donne *D_1, D_2, D_6, D_7 et D8* (pas d'intervalle de confiance).
- Le score de la dimension *usage* est la même que la dimension *Popularité* (une seule dimension la compose).
- Filtrage des documents selon la dimension usage donne *D_1, D_2, D_6, D_7 et D8* (pas d'intervalle de confiance).

⇒ *Filtrage des résultas selon la hiérarchie de la qualité du support*

- Le filtrage des documents selon les dimensions *système* donne *D_1, D_2, D_6, D_7 et D8* (pas d'intervalle de confiance).
- Évaluation du score de la qualité de la dimension utilisateur *accessibilité* à partir des sous-dimensions systèmes *Temps de réponse, Disponibilité, Prix* et *Début réseau.*

Matrice de décision					**Matrice de décision normalisée**				**Calcul du score**
Document	Temps de réponse	Disponibilité	Prix	Début réseau	Temps de réponse	Disponibilité	Prix	Débit réseau	Score de qualité
Poids	2	2	10	2	0,200	0,200	0,400	0,200	
D_1	4	1	30	5	0,750	0,000	1,000	0,556	**0,788**
D_2	3	4	30	3	1,000	0,000	0,833	0,556	**0,715**
D_6	10	2	25	2	0,500	0,429	0,667	0,833	**0,637**
D_7	5	2	30	3	0,000	0,857	0,933	0,000	**0,690**
D_8	6	2	40	5	0,250	1,000	0,000	1,000	**0,281**

- Le filtrage des documents selon l'intervalle de confiance de la dimension *accessibilité* (**I** ≥0.4) donne *D_1, D_2, D_6 et D_7.*

⇒ *Filtrage des résultas selon la qualité générale du document*

- Évaluation du score de la dimension *qualité* à partir des sous-dimensions source : *source d'information, usage d'information, support d'information.*

Matrice de décision				**Matrice de décision normalisée**			**Calcul du score**
Document	Source	Support	Usage	Source	Support	Usage	Score de qualité
Poids	4	8	6	0,222	0,444	0,333	
D_1	0,558	0,788	0,362	0,558	0,788	0,362	**0,595**
D_2	0,558	0,715	0,644	0,558	0,715	0,644	**0,656**
D_6	0,664	0,637	0,267	0,664	0,637	0,267	**0,520**
D_7	0,664	0,690	0,124	0,664	0,690	0,124	**0,496**

- Le filtrage des documents selon la dimension *Qualité* donne *D_1, D_2, D_6, D_7 et D8* (pas d'intervalle de confiance).

3. Re-ordonnancement des résultats de la requête

Le principe du re-ordonnancement est de modifier l'ordre d'affichage des résultats au client. 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 objets du résultat (documents ou informations) 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.

On propose d'intégrer la qualité dans le re-ordonnancement des résultats de la requête (Tableau 2). Notre fonction de rang se base sur la fusion du score du document, retourné par l'algorithme de recherche, et sa qualité. 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.
- Q_d est la qualité du document d.

Document	Q_d	S_d	Ancien Rang	S'_d	Nouveau Rang
D_1	0,595	0,788	**1**	**0,469**	**2**
D_2	0,656	0,755	**2**	**0,495**	**1**
D_6	0,520	0,690	**3**	**0,358**	**3**
D_7	0,496	0,655	**4**	**0,325**	**4**

Tableau 2 - Exemple de re-ordonnancement des résultats de la requête

Conclusion et Perspectives

Dans ce travail, nous avons présenté un modèle flexible de qualité de l'information décrivant les différents facteurs influant sur la personnalisation de l'information.

La multi–dimensionnalité de la hiérarchie de la qualité proposée permet à l'utilisateur d'obtenir différents points de vue selon différentes dimensions et selon différents niveaux de « curiosité » personnalisables vis-à-vis de la qualité d'information.

La flexibilité de notre modèle permet de capturer les paramètres du profil de façon automatique à travers la génération des préférences implicites. Ainsi l'utilisateur peut construire un ensemble cohérent et minimal de critères de qualité satisfaisant ses besoins.

En créant une taxonomie des dimensions de la qualité dans le filtrage des informations on a montré également l'efficacité et la personnalisation de l'approche proposée de l'intégration de la qualité dans le processus d'accès aux informations.

En termes de perspectives à notre travail nous comptons : établir les métriques des différents critères de qualité (dimensions système dans notre modèle) en se focalisant sur les différentes méthodes d'évaluation et d'extraction automatique des métriques de qualité (citations, fréquence de mise à jour, temps de réponse …) ; valider l'approche proposée de l'intégration de la qualité dans le processus d'accès aux informations par des expérimentations et des tests sur des applications de recherche d'information à grande échelle (web).

Une autre perspective est l'amélioration de l'approche de la construction du profil qualité en mettant l'accent sur la construction et la mise à jour les paramètres du profil d'un utilisateur (intervalles de confiance, construction des dimensions utilisateur).

L'utilisation d'une taxonomie de dimensions de qualité permet de fournir des informations adaptées aux besoins en termes de qualité des usagers c'est-à-dire des informations pertinentes et personnalisées.

L'objectif a été de fournir en premier lieu un modèle de qualité d'information assistant l'utilisateur dans la construction de son propre profil de qualité d'une part et une approche d'intégration de la qualité dans le processus d'accès aux informations d'autre part. Il reste néanmoins à vérifier, par expérimentations, l'impact réel de celle-ci sur les informations restituées.

Bibliographie

[BER99] Berti L., *Qualité de données multi sources et recommandation multicritère*, INFORSID 99.

[BOU04] Bouzeghoub M., et Kostadinov D., *Une approche multidimensionnelle pour la personnalisation de l'information*, RapportPRiSM, Versailles, France, 2004.

[BRI98] Brin S. et Page L., *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Proceedings of the 7th International World Wide Web Conference (WWW7), Brisbane, Australia, p.107-117, 1998.

[BUR02] Burgess M., Alex Gray W. et Fiddian N., *Establishing Taxonomy of Quality for Use in Information Filtering*, Proceedings of the 19th British National Conference on Databases (BNCOD 2002), Lecture Notes in Computer Science: Advances in Databases (LNCS 2405), Sheffield, UK, July, p.103-113, 2002.

[BUR03] Burgess M., *Using Multiple Quality Critiera to Focus Information Search Results*, PhD Thesis Cardiff University, UK, September, 2003.

[CAL98] 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, p.37-52, 1998.

[DEN02] Denos N., *QCT and SF services in Torii: Human Evaluations of Documents Benefit to the Community*, in Proceedings of the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (AH2002) Workshop on "Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives", Stefano Mizzaro and Carlo Tasso (Eds.), Malaga, Spain, p.105-114, 29/5-31/5, 2002.

[JAR97] Jarke M. et Vassiliou Y., *Data warehouse quality design: A review of the DWQ project*, In Proceedings of the International Conference on Information Quality (IQ), Cambridge, MA, 1997.

[KOS03] Kostadinov D., *Personnalisation de l'information et gestion des profils utilisateurs*, Rapport de DEA, Université de Versailles, France, 2003.

[MAR02] Marotta A., *Quality Management in MSIS*, Technical Report INCO TR-03-03. ISSN 0797-6410., Septembre, 2002.

[NAU98] Naumann F., *Data Fusion and Data Quality*, Seminar on New Techniques et Technologies for Statistics, 1998.

[NAU99a] Naumann F., Leser U., et Freytag J.C., *Quality-driven integration of heterogenous information systems*, In Proceedings of the International Conference on Very Large Databases (VLDB), Edinburgh, 1999.

[NAU99b] Naumann F. et Rolker C., *Do metadata models meet IQ requirements?*, In Proceedings of the International Conference on Information Quality (IQ), p. 99-114, Cambridge, MA, 1999.

[NAU00] Naumann F. et Roker C., *Assessment Methods for Information Quality Criteria*, Proceedings of the International Conference on Information Quality (IQ2000) Cambridge, MA 2000.

[PER04] Peralta V. et Bouzeghoub M., *On the evaluation of data freshness in data integration systems*, 20èmes Journées de Bases de Données Avancées (BDA'2004). Montpellier, FRANCE, Octobre 2004.

[STR97] Strong D., Lee Y. et Wang R., *Data quality in context*, Communications of the ACM, vol. 40, no. 5, p. 103-110, 1997.

[TCH04] Tchienehom P., *Architecture de Recherche et de Recommandation d'Information à base de Profils*, INFORSID 2004.

[ZHU00] Zhu X. et Gauch S., *Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web*, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece p.288–295, 2000.

Facteurs de qualité et personnalisation de l'information

Résumé :

Ce travail se situe dans le cadre du projet ACI APMD (Accès Personnalisé à des Masses de Données) dont l'objectif est de mener une réflexion globale sur la personnalisation et la qualité de l'information dans un environnement à grande échelle. Notre contribution porte sur la proposition d'un modèle multidimensionnel de la qualité de l'information décrivant les différents facteurs de qualité influant sur la personnalisation de l'information. Notre modèle permet de structurer les différents facteurs de qualité de l'information dans une hiérarchie afin d'assister l'utilisateur dans la construction de son propre profil de qualité. Nous présentons également les différentes méthodes d'évaluation de la qualité de l'information dans notre modèle ainsi qu'une approche d'exploitation du profil qualité dans le processus d'accès aux informations.

Mots-clés : Personnalisation, Qualité d'information, Profil qualité, Évaluation de la qualité, Accès à l'information.

Abstract :

This work is included in the ACI project APMD (Personalized Access to Massive Data) which aims are to carry out a cross global reflection in customization and information quality in a large scale environment. Our contribution consists in a proposition of multidimensional model of the information quality describing the different quality factors that impact in the information personalization. Our model allows structuring the different information quality factors in hierarchy in order to assist the user in the construction of their own quality profile. We present also different evaluation methods of the information quality in our model and our approach of exploitation of quality profile in the information access process.

Key-words: Personalization, Information quality, Quality profile, Quality evaluation, Information access.

Rami HARRATHI