Un modèle de qualité de l'information

Rami Harrathi*, Sylvie Calabretto* *

* LIRIS CNRS UMR 5205 - INSA de Lyon, Bâtiment Blaise Pascal 7, avenue Jean Capelle, *F-69621 Villeurbanne Cedex*

Rharrathi @yahoo.fr

**LIRIS CNRS UMR 5205 - INSA de Lyon, Bâtiment Blaise Pascal 7, avenue Jean Capelle, *F-69621 Villeurbanne Cedex*

Sylvie.Calabretto @insa-lyon.fr

Résumé. Ce travail s'intègre dans la problématique générale de la recherche d'information ; et plus particulièrement dans la personnalisation et la qualité d'information. Dans cet article nous proposons un modèle multidimensionnel de la qualité de l'information décrivant les différents facteurs de qualité influant sur la personnalisation de l'information. Ce 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 selon ses besoins et ses exigences en termes de qualité.

1 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

[1] Notre 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. Site Web: http://apmd.prism.uvsq.fr/ Partenaires: CLIPS-IMAG Grenoble, IRISA Lannion, IRIT Toulouse, LINA Nantes, LIRIS Lyon, PRiSM Versailles

aider à cette formulation, des travaux Bouzeghoub (2004**),** Zhu (2000) et Burgess (2002) 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 cet article nous proposons un modèle flexible de qualité de l'information décrivant les différents facteurs de qualité influant sur la personnalisation. Ce modèle va permettre de structurer les différents facteurs de qualité dans une hiérarchie afin d'assister l'utilisateur dans la construction de son propre profil selon ses besoins et exigences en terme de qualité.

Dans la section suivante, nous présentons un état de l'art sur les approches existantes sur la modélisation de la qualité des données. La section 3 sera consacrée à la présentation de notre modèle de qualité d'information. Enfin nous terminerons cet article par des conclusions et des perspectives.

2 Personnalisation et qualité d'information

La personnalisation de l'information s'exprime par un ensemble de critères et de préférences spécifiques à chaque utilisateur ou une communauté d'utilisateurs. Les données décrivant les préférences des utilisateurs sont souvent sauvegardées sous forme de profils. Parmi les données du profil on trouve une dimension relative à la qualité Bouzeghoub (2004). Afin de définir les facteurs de qualité relatifs à l'information influant sur la personnalisation, il est nécessaire d'analyser les différents travaux menés sur le thème de la modélisation de la qualité des données.

Modélisation de la qualité des données. 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 Wang (1997), Jarke (1997) et Berti (1999). Dans le cadre de la modélisation de la qualité des données, de nombreuses propositions ont été faites. 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 (TAB. 1). Dans Naumann (2000) les auteurs identifient trois approches d'analyse des critères de la qualité des données :

- approche orientée sémantique : elle est basée uniquement 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.

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 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. 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.

Auteurs	Dichotomie et caractérisation de la qualité des données
Wang, Strong et Kan (1997)	»Approche orientée sémantique » 4 Catégories » 13 Dimensions qualité de données
Jarke et Vassiliou (1997)	»Approche orientée objectif » 5 Facteurs qualité des entrepôts de données
Calabretto, Pinon, Poullet et Richez (1998)	»Approche orientée sémantique » 3 Critères de qualité d'information » 8 Critères de qualité des documents
Berti (1999)	»Approche orientée sémantique » 4 Catégories » 32 Critères de qualité des données multi-sources
Naumann et Rolker (2000)	»Approche orientée traitement » 3 Classes d'évaluation des critères » 11 Critères qualité de données
Zhu et Gauch (2000)	»Approche orientée sémantique » 5 Critères de qualité des pages web
Marotta (2002)	»Approche orientée traitement » 2 points de vue : système et utilisateur » 6 Catégories » 31 Critères

TAB. 1 – *Quelques approches de modélisation de la qualité des données.*

3 Proposition d'un modèle de qualité de l'information

3.1 Objectifs du modèle

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

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.

3.2 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, notre hiérarchie de qualité se décompose en un ensemble de dimensions (FIG. 1). Dans la suite nous proposons les différentes dimensions de la hiérarchie de qualité de l'information.

3.2.1 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. 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 :

- support de l'information : les facteurs de qualité liés aux documents.
- source de l'information : les facteurs de qualité liés aux fournisseurs de l'information (Base de données, Site Web, Bibliothèque numérique...).
- usage de l'information: les facteurs de qualité liés à l'usage des informations comme par exemple les formes de popularité (citation).

3.2.2 Dimensions système

Les dimensions système décrivent l'ensemble des critères de qualité vis-à-vis du système. En se basant sur le modèle de Naumann et Rolker (1999) nous proposons l'ensemble de critères préliminaires de la qualité (FIG. 1).

3.2.3 Dimensions utilisateur

Les dimensions utilisateurs sont des dimensions d'agrégation personnalisables par l'utilisateur. Elles se décomposent en une ou plusieurs dimensions utilisateurs ou système. En s'inspirant de la catégorisation de la qualité de Marotta (2002) nous proposons les *principales dimensions utilisateur* suivantes :

- la qualité opérationnelle : l'ensemble des facteurs de qualité liés à l'accès à la source d'information ou support d'information.
- la qualité du contenu : 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.

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 souhaitées d'où la proposition des *sous-dimensions utilisateur* 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 Peralta (2004).
- Fiabilité du contenu : elle se décompose en Complétude et Exactitude.

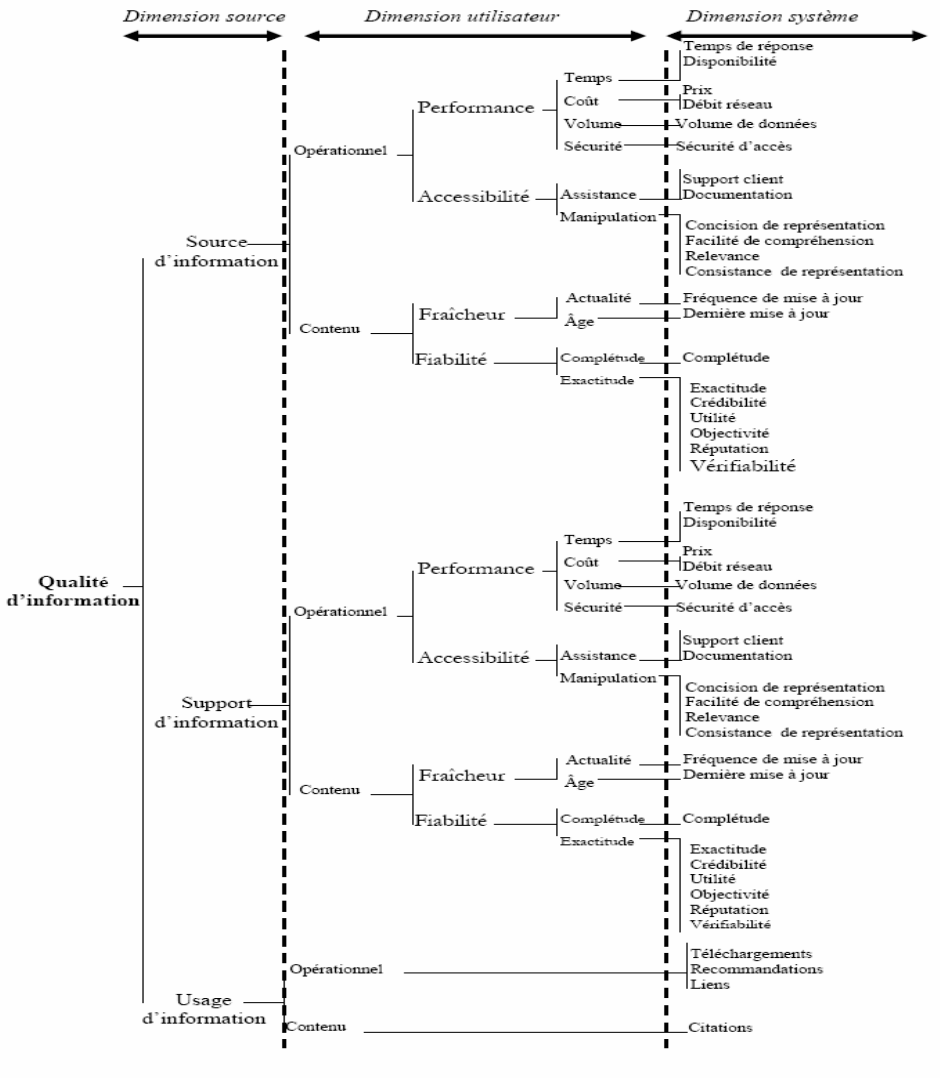

FIG. 1 – *Hiérarchie de dimensions de la qualité d'information.*

4 Conclusion et perspectives

Dans ce travail, nous avons présenté un modèle flexible de qualité 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. En termes de perspectives à notre travail nous comptons : établir les métriques et les méthodes d'évaluation des différentes dimensions de qualité ; proposer un modèle formel de représentation et construction d'un profil qualité.

Références

Berti L. (1999). *Qualité de données multi sources et recommandation multicritère*, INFORSID 99.

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

Burgess M., Alex Gray W. et Fiddian N. (2002). *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 ,103-113.

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

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

Marotta A(2002). *Quality Management in MSIS*, Technical Report INCO TR-03-03. ISSN 0797-6410.

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

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

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

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

Zhu X. et Gauch S. (2000). *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 .288–295.

Summary

This work is included in the general problems of information retrieval and more particularly in personalization and quality of information. In this paper we propose a multidimensional model of information quality describing various quality factors influencing the information personalization. This model makes it possible to structure the various information quality factors in a hierarchy in order to assist the user in the construction of his own profile according to its requirements in term of quality.