Qualité de données dans les entrepôts de données : élimination des similaires

Faouzi Boufarès*, Aïcha Ben Salem*,**, Sebastiao Correia**

*Laboratoire LIPN - UMR 7030 - CNRS, Université Paris 13
Av. J. B. Clément 93430 Villetaneuse France
{boufares, bensalem}@lipn.univ-paris13.fr,
**Société Talend, Rue Pagès 92150 Suresnes
{abensalem, scorreia}@talend.com

Résumé. Ce papier aborde la problématique de l'élimination des similaires (doublons non stricts) dans un entrepôt de données. En effet, la notion de la qualité de données présente un très grand enjeu pour une bonne gouvernance des données afin d'améliorer les interactions entre les différents collaborateurs d'une ou plusieurs organisations concernées. La présence de données en double ou similaires engendre des préoccupations importantes autour de la qualité des données. Un panorama des méthodes de calcul de distance de similarité entre les données ainsi que des algorithmes d'élimination des similaires sont exposés et comparés.

1 Introduction

Les travaux actuels sur l'extraction de connaissances à partir d'un environnement informationnel, qui se caractérise par de très grosses masses de données hétérogènes et distribuées dans les entrepôts de données (ED), se focalisent principalement sur la recherche d'information potentielle, utile et préalablement inconnue. La qualité de l'information recueillie dépend de celles des données. Prendre des décisions à partir de mauvaises informations peut nuire à l'organisation, d'où un coût de la non-qualité qui peut s'avérer très élevé. La construction d'un ED, issu de l'intégration de sources totalement hétérogènes de qualité variable, et d'outils d'aide à la décision issus de ces masses d'informations nécessite le développement de nouveaux outils d'extraction et de transformation de données (ETL - Extract Transform & Load). Ces derniers doivent, d'une part, prendre en compte l'hétérogénéité des données et leurs contraintes, et d'autre part, assurer la qualité du nouvel ensemble de données construit.
Nous avons présenté dans (Hamdoun et Boufarès, 2010) notre démarche pour intégrer des données hétérogènes. Nous avons développé un outil de construction et de maintenance d'un ED de sources hétérogènes HDIM (Heterogeneous Data Integration and Maintenance). Nous nous intéressons dans ce papier à la problématique de la qualité des données dans ces ED et plus précisément au problème d'élimination des doublons et des données similaires (Deduplication, Match and Merge problems (Hernandez et Stolfo, 1998)). Les données similaires sont des données qui ont des ressemblances au niveau de leurs valeurs. Le problème de l'élimination des données similaires (doublons non strictes) et de la fusion/intégration de données est très com-