Utilisation de la théorie des sondages dans le cadre des OLAP

Sabine Goutier, Véronique Stéphan

Electricité de France Recherche et Développement
Département ICAME
1, av. du Général de Gaulle, 92 141 Clamart Cedex, France
{sabine.goutier, veronique.stephan}@edf.fr

Résumé. Dans le cadre de la théorie des sondages, le traitement de la non-réponse a donné lieu à différentes méthodologies reposant principalement sur la pondération ou sur l'imputation. L'objet de cet article est de montrer comment ce cadre formel statistique s'adapte naturellement au problème des valeurs manquantes dans le contexte des OLAP. Dans cette étude, nous nous limitons au cas des valeurs manquantes dans les dimensions, appelées alors dimensions creuses. La méthode d'ajustement est réalisée en intégrant un système de poids au sein du cube. La complexité algorithmique est fortement diminuée par la recherche d'un ensemble de systèmes de pondération minimum. Celle-ci, appelée méthode ROWN, est synthétisée et une validation expérimentale de l'évolution des estimations en fonction du support est présentée. Enfin, l'implémentation sous ORACLE EXPRESS est détaillée.
Mots-Clés : OLAP, cube de données, valeurs manquantes, redressement.

1. Introduction

Les technologies OLAP (OnLine Analytical Processing) permettent la consultation de grands volumes de données selon des directions multidimensionnelles (Codd 1993). Comme de nombreuses entreprises, Electricité De France (EDF) met en œuvre des OLAP pour la consultation de ses grandes bases de données. Au moyen des techniques OLAP, les marketeurs peuvent par exemple analyser les différents segments et tendances du marché, et répondre plus efficacement aux besoins opérationnels. Les données de cubes OLAP servent ainsi de support aux utilisateurs pour leur prise de décision. Il est donc indispensable de garantir la qualité des résultats fournis. En particulier, dans le cas de valeurs manquantes, un biais peut survenir par la seule prise en compte des valeurs renseignées pour le calcul des agrégats du cube.

S'agissant des données clients, l'existence dans les bases de valeurs manquantes est fréquente. Dans le contexte des bases EDF, les informations sont pour la plupart saisies lors de contacts avec le client. Hormis les informations tarifaires obligatoirement renseignées, les autres caractéristiques liées à un client (par exemple l'énergie du chauffage principal ou le type d'habitat) ne sont pas forcément (bien) remplies.

Le but de cet article est d'adapter les techniques de sondage pour améliorer la qualité des agrégats en présence de valeurs manquantes observées sur les dimensions du cube. Après avoir décrit le contexte de travail, nous présentons la repondération d'une dimension par