Anonymiser des données multidimensionnelles à l'aide du coclustering

Françoise Fessant*, Tarek Benkhelif*,**, Fabrice Clérot *

*Orange Labs, Lannion
francoise.fessant, tarek.benkhelif, fabrice.clerot@orange.com,
**DUKe, LINA, Nantes
http://duke.univ-nantes.fr/

Résumé. Dans cet article, nous proposons une méthodologie pour anonymiser une table de données multidimensionnelles contenant des données individuelles (soit *n* individus décrits par *m* variables). L'objectif est de publier une table anonyme construite à partir d'une table initiale qui protège contre le risque de réidentification. En d'autres termes, on ne doit pas pouvoir retrouver dans les données publiées un individu présent dans la table originale. La solution proposée consite à agréger les données à l'aide d'une technique de coclustering, puis à utiliser le modèle produit pour générer une table de données synthétiques du même format que les données initiales. Les données synthétiques, qui contiennent des individus fictifs, peuvent maintenant être publiées. Les données produites sont évaluées en termes d'utilité pour différentes tâches de fouille (analyse exploratoire, classification) et de niveau de protection.

1 Introduction

Il y a une très forte demande économique et citoyenne pour l'ouverture des données que ce soit pour la recherche ou le marketing. Le secteur public en particulier, à travers ses instituts de statistique nationaux, de santé ou de transport est soumis à des pressions pour mettre à disposition du public autant d'informations que possible au nom de la transparence [1]. Les entreprises privées sont également concernées par la valorisation de leur données à travers l'échange ou la publication. Orange a ainsi récemment mis à disposition de la communauté scientifique différents jeux de communications mobiles collectées sur ses réseaux de Côte d'Ivoire ou du Sénégal, dans le cadre des challenges D4D (Data for Developement) dans un objectif de service à des projets de développement ou d'amélioration de politiques publiques (Blondel et al., 2012).

Quand les données publiées concernent des individus et comportent des données à caractère personnel, elles doivent être anonymisées. On peut définir l'anonymisation comme le processus par lequel des données sont rendues anonymes et à l'issue duquel elles ne peuvent plus être affectées ou rattachées à une personne en particulier. [2]

1. https ://www.republique-numerique.fr
2. Définition de l'Association Française des Correspondants à la protection des Données à caractère Personnel. Glossaire anonymisation de données de l'AFCDP du 23 mai 2007.