

Détection des Données à Caractère Personnel dans les Bases Multidimensionnelles

Amine Mrabet, Ali Hassan, Patrice Darmon

Research & Innovation - Umanis
7, Rue Paul Vaillant Couturier, 92300 Levallois-Perret, France
{ amrabet, ahassan, pdarmon } @umanis.com

Résumé. Cartographier les données à caractère personnel dans la démarche de la de-identification est toujours un vrai pré-requis. Aujourd'hui avec les bases de données de grandes masses, nous sommes dans l'obligation d'automatiser l'étape de la détection dans cette démarche. Ce qui permet d'éviter les étapes chronophages, d'augmenter la précision de la détection et essentiellement permet de garantir la confidentialité. Pour toutes ces raisons nous avons proposé une nouvelle approche pour détecter les données personnelles. L'approche détaillée dans ce travail est adaptée aux bases de données multidimensionnelles. Nos méthodes utilisées dans cette approche sont sur deux niveaux. Nous proposons deux solutions de détection au niveau des données, et une solution au niveau des méta-données. Après détection des données à caractère personnel dans une base en utilisant les scores d'identifications, nous utilisons les scores de sensibilité afin d'évaluer la sensibilité totale de la base multidimensionnelle avant et après anonymisation.

1 Introduction

La protection de la vie privée est un droit humain fondamental. Ceci est reconnu par l'article 8 de la convention de sauvegarde des droits de l'homme et des libertés fondamentales (Conseil de l'Europe, 1948) qui assure le droit de chacun au respect de *"sa vie privée et familiale, de son domicile et de sa correspondance"*. De même, la Charte des droits fondamentaux de l'union européenne (European Commission for Justice, 2009) définit le *"respect de la vie privée et familiale"* (article 7) et ajoute un article spécifique sur la *"protection des données personnelles"* (article 8).

L'expression "données personnelles" a été définie dans la loi n 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. La définition de ces données était : *"toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres."* La Commission nationale de l'informatique et des libertés (CNIL) reprend la même définition et donne les exemples suivants d'une donnée à caractère personnel : *"un nom, une photo, une empreinte, une adresse postale, une adresse mail, un numéro de téléphone, un numéro de sécurité sociale, un matricule interne, une adresse IP, un identifiant de connexion informatique, un enregistrement vocal"*.