Nettoyage de données guidé par la sémantique inter-colonnes

Houda Zaidi*,*** Faouzi Boufarès** Yann Pollet*

*CEDRIC, CNAM, 2 Rue Conté, 75003 Paris,France
houda.zaidi@cnam.fr, yann.pollet@cnam.fr
**LIPN, Université Sorbonne Paris Cité, 93430, Villetaneuse, France
faouzi.boufares@lipn.univ-paris13.fr
***RIADI, Université de la Manouba, Manouba 2010, Tunisie

De nos jours, il est intéressant de développer de nouveaux outils d'intégration et de manipulation de données (ETL) afin d'aider à mieux comprendre la sémantique et la structure des données manipulées Boufarès et al. (2013), Ben Salem (2015). Nous réalisons ce travail en collaboration avec la société Talend (éditeur d'un ETL). La première partie du projet a traité des anomalies inter-lignes une fois la sémantique de la colonne est connue et ses anomalies corrigées. La deuxième phase du projet consiste à découvrir d'éventuels liens sémantiques inter-colonnes afin de corriger d'autres types d'anomalies . La vérification des contraintes de dépendances permettra de corriger les anomalies telles que les valeurs nulles et certaines dépendances fonctionnelles. La reconnaissance sémantique des données est présentée dans le premier paragraphe. La section deux aborde l'étape de nettoyage de données intra et inter-colonnes.

(1) La Catégorisation sémantique des données consiste à déterminer le sens de chaque colonne d'une source de données S. En effet, pour pouvoir qualifier une donnée syntaxiquement incorrecte, il faudrait l'évaluer dans son contexte. Plusieurs exemples peuvent illustrer nos propos : (i) La chaîne de caractères "Pari" ne peut être considérée incorrecte syntaxiquement que s'il s'agit du nom en français de la ville "Paris" ; (ii) Les mots "Pékin" et "Beijing" désignent la même chose dans deux langues différentes, s'il l'on sait qu'il s'agit de noms de villes. "Beijing" pourrait être considérée sémantiquement incorrecte si la langue dominante est le français ; (iii) Les deux chaînes de caractères "16-10-1996" et "10-16-1996" représentent la même information de type date définie par une expression régulière. Le format n'est pas le même. Pour ce faire nous utilisons des connaissances stockées dans un réferentiel appelé dictionnaire de données (DD), Zaidi et al. (2015), identifiées (i) par extension, c'est une liste donnée à priori tels que des noms de villes ou des mots clés ; (ii) par intention qui sont des connaissances qui vérifient des propriétés telles que des expressions régulières (un Email ou une date). Chaque catégorie correspond à un seul type de données (String, Nombre ou Date). **La reconnaissance de la structure sémantique de données** (le processus de catégorisation) renvoie un nom sémantique (une catégorie) à chaque colonne, une sous-catégorie (la langue), un type de données (domaine syntaxique), des contraintes (intra et inter-colonnes) et des commentaires. La reconnaissance sémantique consiste à trouver des similarités entre les données de S et celles de DD afin d'inférer la catégorie de chaque colonne en utilisant des mesures de distance de similarité avec les méthodes "s'écrit comme" et "se prononce comme" telles que Jaro-Winkler et Soundex. **La reconnaissance de dépendances sémantiques inter-colonnes**