Nettoyage des données XML : combien ça coûte ?

Laure Berti-Équille

IRISA, Campus Universitaire de Beaulieu, 35042 Rennes cedex
berti@irisa.fr
http://www.irisa.fr

Résumé. L'objectif de cet article est de présenter un travail en cours qui consiste à proposer, implanter et valider expérimentalement un modèle pour estimer le coût d'un processus de nettoyage de documents XML. Notre approche de calcul de coût est basée sur une méthode par calibration selon une analyse probabiliste. Pour cela, nous proposons de calculer des probabilités de pollution et au préalable de détection des différents types de pollutions. Pour valider notre modèle, nous avons choisi de polluer artificiellement une collection de données XML avec l'ensemble des types d'erreurs possibles (erreurs typographiques, ajout de doublons, de valeurs manquantes, tronquées, censurées, etc.) et d'estimer, grâce au modèle proposé, le nombre et le coût des opérations nécessaires au nettoyage des données afin de proposer des stratégies de réparation ciblées et économes. Les expérimentations en cours ne sont pas rapportées dans cet article.

1 Introduction

Le nettoyage automatique des données se décompose classiquement en trois étapes : 1) examiner les données afin de détecter les incohérences, les données manquantes, les erreurs, les doublons, etc. 2) choisir les transformations pour résoudre les problèmes, 3) et enfin, appliquer les transformations choisies au jeu de données. La plupart des outils utilisés pour le nettoyage des données par *Extraction-Transformation-Loading* (*ETL*) permettent l'extraction d'expressions régulières et structures (*patterns*) à partir des données, ainsi que leur transformation et formatage par l'application de différentes fonctions (sélection, fusion, *clustering*, etc.) (Vassiliadis 2003) dont généralement, on ignore *a priori* le coût. Bien qu'il existe de nombreux travaux (Dasu 2003), (Winkler 2003), (Rahm 2000) outils et prototypes (Telcordia (Caruso 2000), AJAX (Galhardas 2001), Potter's Wheel (Raman 2001), Arktos (Vassiliadis 2000), IntelliClean (Low 2000), Tailor (Elfeky 2002)) développés pour « nettoyer » les données relationnelles, très peu de travaux à l'exception des récents travaux de Weis et Naumann (Weis 2004), ont jusqu'ici été menés pour le nettoyage de données XML et, à notre connaissance, aucun n'a abordé l'estimation du coût d'un nettoyage de données *a fortiori* pour des données XML. C'est dans ce cadre qu'a débuté notre travail dont l'objectif est de proposer, d'implanter et valider expérimentalement un modèle de coût global permettant d'estimer combien peut coûter un processus de nettoyage sur un document XML artificiellement pollué pour les besoins de nos expériences.

La suite de l'article s'organise de la façon suivante : la section 2 propose notre démarche illustrée par un exemple simple qui énumère les différents types de pollution possibles dans un document XML. La section 3 présente plus formellement notre modèle de coût avec ses