Master Informatique, Mathématiques, Multimédia & Télécommunications
Spécialité « informatique »
Spécialité « Réseaux de Télécommunications, Multimédia et Automatique »

Proposition de Sujet de Stage de Master 2
2017-2018

Titre : Sur une nouvelle génération d'indicateurs de qualité de données

Title: On a new generation of data quality indicators

Entreprise : Laboratoire LIAS

Encadrant(s) : A. HADJALI (allel.hadjali@ensma.fr), B. Chardin (à confirmer)

Mots clés : Big data, qualité de données, mesures de qualité, réparation de données.

Contexte et problématique

L'existence d'anomalies et d'impuretés dans les données du monde réel est bien connue de nos jours. Dans [Kie 2016], leurs taux typiques sont estimés entre 10 à 30%. L'étude de la qualité des données reste donc un problème majeur car les données "impures ou impropres" (dirty data en anglais) peuvent conduire à des décisions incorrectes et à des analyses non fiables. La qualité des données [Fan 2012] désigne l'aptitude de l'ensemble des caractéristiques intrinsèques des données à satisfaire en vue de prise de décision ou de pilotage. La qualité est donc un concept complexe et multidimensionnel combinant plusieurs caractéristiques ou dimensions [Bert 2007].

L'avènement du Big data a exacerbé le problème lié à la qualité des données et a également ajouté de nouvelles dimensions. Une nouvelle vision du calcul d'indicateurs de qualité s'impose pour relever les défis posés par le Big data. La plupart des indicateurs de qualité proposés dans la littérature paraissent discutables d'un point de vue sémantique et sont de nature purement statistique. De plus, la façon dont ces mesures peuvent être utilisées en pratique et exploitées en lien avec des méthodes de nettoyage n'est pas claire.

Objectifs

L'objectif du stage est, dans un premier temps, de recenser les indicateurs de qualité qui existent dans la littérature. Puis, dans un second temps, revisiter certains de ces indicateurs en choisissant une à deux dimensions de la qualité. Ensuite, étudier les propriétés de ces indicateurs par rapport au processus de réparation. Des développements seront à réaliser pour

implémenter les algorithmes de calcul de ces indicateurs et des expérimentations sur les performances et le coût seront menées également.

Bibliographie

- [Kie 2016] Cornelia Kiefer, Assessing the Quality of Unstructured Data: An Initial Overview. LWDA 2016: 62-73.
- [Fan 2012] W. Fan and F. Geerts. Foundations of Data Quality Management. Morgan & Claypool Publishers, 2012.
- [Bert 2007] Laure Berti, Mémoire d'Habilitation, Université Rennes 1, 2007.

Lieu du stage : LIAS / ENSMA

UEs optionnelles conseillées : Cours IDD, AAW, AFGL

Technologies impliquées

- ***Base de données relationnelles*** : instances PostgreSQL (Sharing Data) pour le
- ***Langages de programmation*** **:** scripts Shell Bash, Java, JavaScript (Node.js)

SP2MI, Bvd Marie et Pierre Curie, Téléport 2, B.P. 30179, 86960 FUTUROSCOPE CEDEX
Téléphone : 05 49 49 74 71, Télécopie : 05 49 49 65 70
E-mail : Sylvie.Duclaud@sic.sp2mi.univ-poitiers.fr