LIPADE
Laboratoire d'informatique PAris DEscartes

Sujet de Master 2

Gestion de l'incertitude et de l'incohérence de données dans le fusion de Big Data

La gestion de l'incohérence et de l'incertitude de données est cruciale dans le contexte du « big data ». En effet, le « big data » met à disposition de quantités massives de données provenant de multiples sources indépendantes et par conséquent potentiellement conflictuelles comme DBpedia/Wikipedia, Freebase ou celles qu'on peut trouver sur le portail du gouvernement de France (*data.gouv.fr* - recensement de la population, transports, criminalité, etc.).

La recherche d'information dans le big data consiste à collecter, fusionner et réconcilier les données provenant de différentes sources. La fusion de données tente de rapprocher les données de différentes sources [Halevy 2001] au moyen des techniques de mapping/matching [Hassanzadeh 2013]. La reconciliation de données a été traitée dans plusieurs domaines, notamment en bases de données [MyDBMS 2009], en logique [Bruynooghe 2010] et en intégration de données [Sarma 2011].

La fiabilité de données peut être traitée suivant deux philosophies. La première consiste à nettoyer les données des sources en remplaçant les données peu fiables par les données plus fiables, éliminant ainsi les incohérences [Fan 2009]. Cette approche présente le risque de perte de données potentiellement correctes et utiles et peut ne pas être juridiquement applicables.

La seconde philosophie et qui est celle préconisée dans ces travaux consiste à maintenir les données telles quelles, avec leurs incohérences, et à calculer la fiabilité de ces données pendant l'évaluation des requêtes afin de restituer un résultat cohérent et le plus fiable possible [Bertossi 2011].

Dans ce master il s'agit de proposer une approche pour un raisonnement incertain lors de la fusion de données dans le big data et open data pour une application économico politique. Le candidat devra valider l'approche par une implantation technique.

Le stage se déroulera entre le Laboratoire LIPADE (équipe data intensive and kNowledge oriented systems group-diNo), Université Paris Descartes et le Département d'économie Science Po.

Prérequis :Base de données, probabilités, raisonnement logique.

Durée du stage : 6mois

Contacts

Salima Benbernou, Université Paris Descartes, salima.benbernou@parisdescartes.fr
Mourad ouziri, , Université Paris Descartes, mourad.ouziri@parisdescartes.fr

Références

[MyDBMS 2009] MayBMS: A System for Managing Large Uncertain and Probabilistic Databases. C. Koch. Chapter 6 of Charu Aggarwal, ed., *Managing and Mining Uncertain Data*, Springer-Verlag, 2009.

[Bertossi 2011] Leopoldo E. Bertossi, Solmaz Kolahi, Laks V. S. Lakshmanan: Data cleaning and query answering with matching dependencies and matching functions. ICDT 2011: 268-279

[Bruynooghe 2010] Maurice Bruynooghe, Theofrastos Mantadelis, Angelika Kimmig, Bernd Gutmann, Joost Vennekens, Gerda Janssens, Luc De Raedt: ProbLog Technology for Inference in a Probabilistic First Order Logic. ECAI 2010: 719-724

[Sarma 2011] Das Sarma.A, Dong, Y. Halevy: Uncertainty in Data Integration and Dataspace Support Platforms. Schema Matching and Mapping 2011: 75-108

[Fan 2009] W. Fan, F. Geerts, and X. Jia. Conditional dependencies: A principled approach to improving data quality. In A. P. Sexton, editor, BNCOD, volume 5588 of Lecture Notes in Computer Science, pages 8–20. Springer, 2009.

[Hassanzadeh 2013] Oktie Hassanzadeh, Ken Q. Pu, Soheil Hassas Yeganeh, Renée J. Miller, Lucian Popa, Mauricio A. Hernández, Howard Ho: Discovering Linkage Points over Web Data. PVLDB 6(6): 444-456 (2013)