Analyse et évaluation de la qualité des données migratoires

Encadrants

CRIL, Lens : Said Jabbour, jabbour@cril.fr Lakhdar Sais, sais@cril.fr

LIPADE, Paris : Salima Benbernou, salima.benbernou@parisdescartes.fr , Mourad Ouziri , mourad.ouziri@parisdescartes.fr

Migrinter, Poitiers : Nelly Robin, nelly.robinsn@orange.fr, Cyril Roussel, cyril cyrille.roussel@univ-poitiers.fr

Contexte du projet QDoSSI

Les migrations internationales ont pris dans le monde contemporain une ampleur inédite. A l'heure où *les sciences et technologies du numérique* jouent un rôle fondamental dans la construction et l'analyse des parcours migratoires, de nouveaux défis sont ainsi posés à la communauté scientifique en termes d'évolution des concepts, des méthodes et des outils utiles à l'analyse et à la compréhension des phénomènes migratoires. A la croisée de ces enjeux sociétaux et scientifiques, le projet QDoSSI s'appuie sur un réseau de recherche, qui associe des pratiques scientifiques propres aux sciences humaines et sociales et à l'informatique, allant des bases de données à la fouille de données en passant par l'intelligence artificielle. Il s'agit d'interroger autrement et d'exploiter des bases de données imparfaites, fortement hétérogènes et évolutives portant sur *des espaces migratoires et des temporalités variables*.

Notre champ d'analyse porte sur des données, a priori hétéroclites, collectées par les laboratoires MIGRINTER et CEPED ; il s'agit notamment : (1) de registres administratifs, médicaux et judiciaires, (2) de corpus juridiques européens, d'Afrique de l'Ouest et des Balkans, carrefours importants des circulations migratoires vers l'Europe, (3) des récits de vies, notamment des mineurs en mobilité sur les routes transsahariennes, des demandeurs d'asile le long du corridor des Balkans et des migrants de Calais, (4) des recensements et enquêtes effectués récemment auprès des mineur(e)s à Mayotte, de migrants qualifiés au Mexique et des personnes déplacées et réfugiées (Syrie et Kurdistan irakien). A titre d'exemple, dans le domaine des migrations internationales, les données administratives permettent l'observation des relations entre plusieurs acteurs : États, migrants et groupes criminels. Ces données peuvent être mises en perspective avec des récits de vies collectés sur les routes migratoires, riches en informations contextualisées, et des corpus juridiques, nationaux et internationaux, définissant les droits des migrants, des demandeurs d'asile et des réfugiés.

Objectif du stage

Les données mobilisées dans le cadre de QDoSSI sont hétérogènes et n'ont jamais été traitées comme éléments de la même matrice : bases de données structurées, documents textuels (récits de vies, biographies migratoires), documents (corpus juridiques). Avec des temporalités et des représentativités variables, elles renseignent sur des phénomènes migratoires observés sur des espaces géographiques différents. Enfin, elles présentent diverses imperfections, comme la présence de données manquantes, de données mal orthographiées ou encore de données entachées d'incertitudes et d'imprécisions. *Ceci pose le problème de la mesure, de l'intégration et de l'évaluation de l'impact de la qualité des données sur les résultats d'analyses dans un contexte spatio-temporel évolutif.*

L'objectif de ce stage est de quantifier plus précisément la qualité des différentes bases de données cibles. Ceci passe par la définition de mesures de la qualité des données avant et après prétraitement. Il s'agit ensuite d'analyser les résultats obtenus par les différentes techniques de fouille de données en donnant des indicateurs de qualité/erreurs. Le challenge est d'y associer un modèle interactif et cyclique (prétraitement, analyse, et validation). Ces indicateurs conjugués aux attributs impliqués dans chaque traitement envisagé seront exploités pour mieux cibler les réparations. *Nous envisageons de développer un modèle à base de contraintes où les incqualitohérences représentent les violations du modèle.*

Travail à réaliser

Le travail de stage qui sera effectué dans le cadre du projet *QDoSSI* s'articulera autour des tâches suivantes :

- Faire une étude bibliographique des différentes approches traitant de la qualité des données
- Dans un deuxième temps, établir des mesures pertinentes de la qualité dans le contexte des données migratoires.
- Implémenter et quantifier ces mesures sur les différentes bases de données
- Proposer et implanter des approches capables de réparer certaines types d'imperfections.

Références

[1] THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT : Defining Data Quality Dimensions. DAMA UK Working Group on "Data Quality Dimensions

[2] Dhana Rao, Venkat N. Gudivada, Vijay V. Raghavan: Data quality issues in big data. 2654-2660

[3] Khalifeh AlJadda, Mohammed Korayem, Trey Grainger: Improving the quality of semantic relationships extracted from massive user behavioral data. 2951-2953

[4] Data Quality Assessment," Communications of the ACM, April 2002. pp. 211-218, Leo Pipino, Yang Lee, and Richard Wang.

[5] "Sample-Based Quality Estimation of Query Results in Relational Database Environments," IEEE Transactions on Knowledge and Data Engineering, May 2006, Vol.18, No.5, pp 639-650, D.P. Ballou, I.N. Chengalur-Smith, and R.Y. Wang.

[6] Impact of information quality and decision-maker quality on decision quality: A theoretical model," Decision Support Systems, Oct 99, Vol. 26, Issue 4, pp. 275-287. Srinivasan Raghunathan.

[7] Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. 2009. Overview and Framework for Data and Information Quality Research. *J. Data and Information Quality* 1, 1, Article 2 (June 2009), 22 pages. DOI=http://dx.doi.org/10.1145/1515693.1516680

[8] Craig W. Fisher, Eitel J. M. Lauria, and Carolyn C. Matheus. 2009. An Accuracy Metric: Percentages, Randomness, and Probabilities. *J. Data and Information Quality* 1, 3, Article 16 (December 2009), 21 pages. DOI=http://dx.doi.org/10.1145/1659225.1659229