Exploitation de la spatialité sur des données textuelles hétérogènes provenant de Madagascar

Jacques Fize[1], Mathieu Roche[1], Maguelonne Teisseire[1]
[1]TETIS, Univ. Montpellier, APT, Cirad, CNRS, Irstea, Montpellier, France

Avec l'essor du Big Data, le traitement du Volume, de la Vélocité (croissance et évolution) et de la Variété de la donnée concentre les efforts des différentes communautés pour exploiter ces nouvelles ressources. Ces nouvelles ressources sont devenues si importantes, que l'on parle aujourd'hui du nouvel « or noir » [1]. Au cours des dernières années, le volume et la vélocité sont des aspects de la donnée qui sont maitrisés contrairement à la variété qui elle reste un défi majeur[2].

Nos travaux consistent à concevoir des modèles de représentation de données textuelles hétérogènes et les mesures de similarité permettant de les comparer. Les données textuelles sont dites hétérogènes par leur forme (format/structure de fichier) et par leur fond (variété de thématique, écriture, …). Nous proposons des méthodes pour comparer ces données selon trois aspects ou dimensions : la thématique (Quoi ?), la spatialité (Où ?) et la temporalité (Quand ?). À ce jour, nous avons décidé de nous focaliser sur le développement de modèles de représentation de la spatialité dans les textes. Dans ce contexte, nous avons conçu une structure de type graphe, appelée **STR[3]** pour **S**patial **T**extual **R**epresentation, composé des informations spatiales extraites : les entités spatiales et leurs relations spatiales. Une entité spatiale est une entité localisée dans l'espace et renseignée dans un référentiel (Geonames, OpenStreetMap, …). Les relations spatiales utilisées sont qualitatives : l'**adjacence**, *eg. France est adjacente à la Belgique*, et l'**inclusion**, *eg. Paris est inclus dans la région Ile-de-France*. La construction

Pour montrer le potentiel de ces modèles, nous travaillons sur différents jeux de données à notre de disposition, dont un jeu de données réelles liées aux activités de recherches du CIRAD à Madagascar et reposant sur la thématique de l'agroécologie.

[1] Arvind Singh, « Is Big Data the New Black Gold? », *Wired, http://www. wired. com/2013/02/is-big-data-the-new-black-gold*, 2013.
[2] Seref Sagiroglu et Duygu Sinanc, « Big data: A review », *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, 42–47, https://doi.org/10.1109/CTS.2013.6567202.
[3] Jacques Fize, Mathieu Roche, et Maguelonne Teisseire, « Spatial Textual Representation (STR) ou comment représenter la spatialité des donnnées textuelles », 2017, https://hal.archives-ouvertes.fr/hal-01643368/document.