Appariement d'entités spatiales avec l'outil GeoAlign

Nelly Barret

LIRIS — Université Claude Bernard Lyon 1

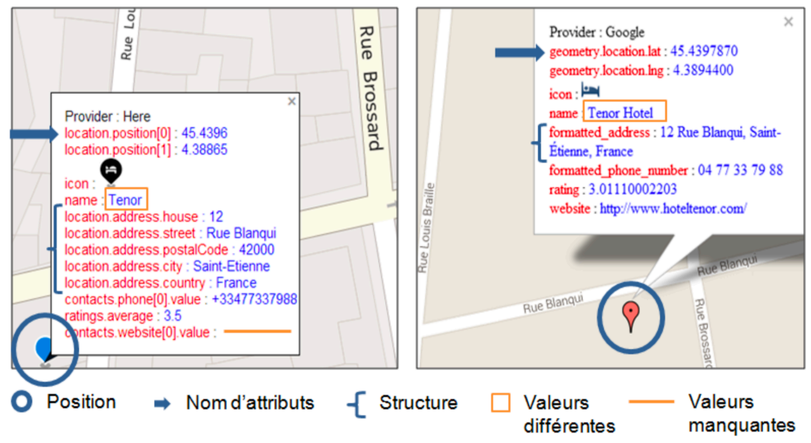
Contexte

- Différents fournisseurs cartographiques de POI (restaurants, hôtels…)
- Plusieurs entités spatiales pour un POI : différences, incohérences, incomplétude

Comment obtenir des informations complètes, minimales et à jour pour un POI ?

→ Appariement et fusion des entités correspondantes

→ Évaluation des algorithmes d'appariement (avec ou sans réalité terrain)



Travaux existants

GeoDDupe, Sehgal : appariement d'entités spatiales

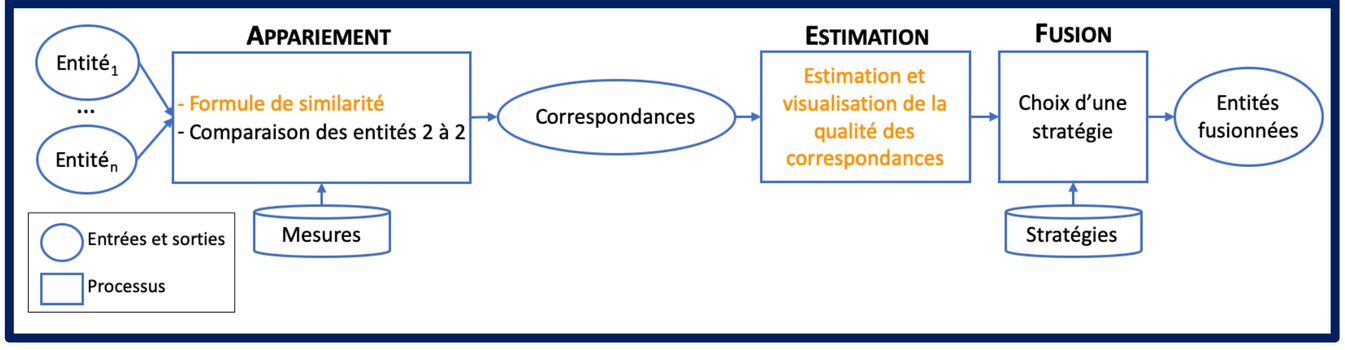
→ Formule de similarité fixée et peu personnalisable

GeoBench : construction d'un benchmark

→ Jeu de données limité et statique, opérations manuelles

Hyunmo Kang, et al. Geoddupe: a novel interface for interactive entity resolution in geospatial data. In IV 2007.
Vivek Sehgal, et al. Entity resolution in geospatial data integration. In GIS 2006.
Anthony Morana, et al. GeoBench: a geospatial integration tool for building a spatial entity matching benchmark. In SIGSPATIAL 2014.

Aperçu de l'approche GeoAlign



Formule de similarité personnalisable

- Combinaison d'une liste de *tokens* : somme pondérée de mesures de similarité sur les attributs
- Un seuil de décision

$$\sum_{i=1}^{n} poids_i * sim_i(attribut_i) > seuil$$

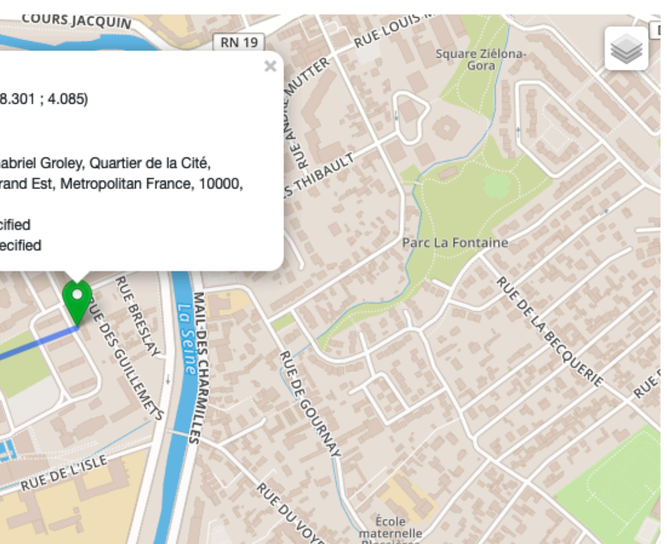
Estimation de la qualité des correspondances (sans réalité terrain)

Intuition : une correspondance a plus de chance d'être correcte quand plusieurs formules de similarité la détectent et que ces formules sont différentes

Calcul de la dissimilarité entre formules de similarité

- Deux tokens sont dans un *groupe de tokens* s'ils portent sur le même attribut et si leurs mesures de similarité appartiennent à la même catégorie, e.g. terminologique, spatiale…
- Calcul de la dissimilarité Δj à partir d'un groupe de tokens
  - Calcul de l'écart-type des poids
  - Pondération de l'écart-type par une formule basée sur une hyperbole
  - Calcul de la dissimilarité globale (moyenne des dissimilarités Δj)

$$\Delta j = \begin{cases} j = 1 \; si \; taille(groupe) = 1 \\ \left(\frac{\sigma_j}{\sigma_{max_j}} - 0.5\right) * c_j + 0.5 \; sinon, \text{où } c_j = 1 - \frac{0.25}{n_j - 1.5} \end{cases}$$



Prototype GeoAlign

Perspectives

- Expérimentations pour vérifier la pertinence de l'estimation de la qualité des correspondances détectées
- Généralisation avec une combinaison à base de règles (besoin d'assistance ou d'apprentissage pour construire la formule)