GeoBench: a Geospatial Integration Tool for Building a Spatial Entity Matching Benchmark

A. Morana, T. Morel, B. Berjawi, F. Duchateau

Laboratoire d’InfoRmatique en Image et Systèmes d’information
UMR5205 CNRS / Université de Lyon / Université Claude Bernard Lyon 1 / INSA de Lyon
http://geobench.liris.cnrs.fr

Objectives
- For data integration practitioners: facilitate the building of spatial entity matching datasets
- For end-users: build a map with complete information about their favourite places

Motivations
- Evaluate and compare spatial entity matching approaches
- Build a characterized spatial dataset for machine learning purposes

Issues and Contributions
- Location-based services providers offer incomplete and/or contradictory data about tourist places
- Recent works are proposed to discover spatial entities that refer to the same place
- These works have been evaluated using different test protocols
- Datasets used for evaluation are not made fully available

Overview of GeoBench
GeoBench is a tool which serves to build a benchmark for spatial entity matching by facilitating the discovery and the integration of corresponding spatial entities.

GeoBench phases
- **Blocking Algorithm** aims at quickly identifying a subset of entities among all those available which likely represent the source entity $\varphi$.
- **Detecting differences** aims at classifying the terminological and spatial differences between the attributes of two entities.
- **Matching Algorithm** aims at computing a confidence score between the entities of the blocking phase and the initial source entity $\varphi$.
- **Integrating Corresponding Entities** offers the possibility to merge corresponding entities into a new integrated entity.

Specifications
- **Blocking**
  - Based on the coordinates of the source entity $\varphi$ and include all entities within a radius
  - Entities of the blocking area whose name shares a token with $\varphi$’s name and having the same type of $\varphi$
- **Matching**
  - Terminological: based on Levenshtein string similarity measure
  - Spatial: based on Euclidean distance

Example of heterogeneous between two LBS providers

Example of the matching process

Example of the integration process