Matching Geospatial Instances

Heshan Du[1], Natasha Alechina[1], Michael Jackson[1], Glen Hart[2]

[1] University of Nottingham
[2] Ordnance Survey of Great Britain

The work presented in this paper extends our work on matching formal and informal geospatial ontologies [1], aimed to realize the synergistic use of authoritative and crowd-sourced geospatial information. A geospatial instance is an object which has a certain and verifiable location (geometry, topographic footprint), as well as a meaningful label (for example, Victoria Shopping Centre in Nottingham, UK). The source of examples in this paper are: The OpenStreetMap (OSM) [2] and the Ordnance Survey of Great Britain (OSGB) [3].

Different geospatial instances may have the same purely lexical information and the same classification in terms of an ontology, but different locations. For example, there may be several restaurants called 'Prezzo Ristorante' in the same city. Therefore, when matching geospatial instances, it is essential to use location information. However, few of existing ontology matching or data interlinking methods can match spatial instances effectively.

There is also a choice in representing objects such as shopping centres as a collection of parts or as a single instance. For example, Victoria Centre is represented as a collection of shops and other businesses in OSGB and as a single instance in OSM. In order to produce a meaningful correspondence between instances in OSGB and in OSM, we propose to use 'partOf' relation (mereological partOf in geometry and having similar labels). If for two instances a and b we get 'partOf' relations in both directions, we generate a hypothesis that a and b belong to 'sameAs'.

Here we propose a new method for establishing 'sameAs' and 'partOf' relations between geospatial instances from different ontologies. In crowd-sourced data, there is an increased possibility of error in measurement, and tendency to simplify shapes of buildings. For this reason, in our method we use buffers to compare geometries of instances. The size of the buffer (intuitively, 'the margin of error' σ) is a parameter which can be arrived at experimentally or within reason set arbitrarily. Intuitively the optimal value of σ corresponds to the maximal deviation between two representations of the same object in two different data sets. In the case of OSGB and OSM data for Nottingham, this is experimentally determined as 20m. The new method has four steps.

Step 1: Extracting geometry sets. An ABox contains facts (geometry, lexical and semantic classification information) about geospatial instances. We extract a set of geometries, G_i, from all the spatial instances in each ABox A_i, $i = 1, 2$.

Step 2: Matching geometry sets. For two sets of geometries, G_1, G_2, a level of tolerance α and tolerance for the second best β, we generate the best two candidate matches for each geometry in G_1 if they exist in G_2, and the best two candidate matches for each geometry in G_2 if they exist in G_1. The candidates are selected by comparing minimal buffers. The buffer of a geometry g, $\mathit{buffer}(g, \sigma) = \{p : \exists p' \in g.\ \mathit{distance}(p, p') \leq \sigma\}, (\sigma > 0)$. For two geometries g and h, the minimal buffer of h containing g is $\mathit{buffer}(h, \sigma)$ such that $g \subseteq \mathit{buffer}(h, \sigma)$ and for all $\sigma' < \sigma$, $g \not\subseteq \mathit{buffer}(h, \sigma')$. For any geometry g, the minimal buffer α ($\alpha \leq \sigma$) of its best candidate o_1 ($g \subseteq \mathit{buffer}(o_1, \alpha)$) is the smallest among those of all the candidates. We generate a 'buffered part of' (BPT) relation between each geometry and its candidates, i.e. $(g, o_1) \in BPT(\sigma)$.

Step 3: Comparing labels. We use string comparison, including equality, inclusion, abbreviation and edit distance to check whether the labels, such as names or addresses, of two instances are similar. If there is no pair of labels of spatial instances s_1, s_2 that are similar, then their lexical information is incompatible, $(s_1, s_2) \in LF$. Otherwise, their lexical information is compatible, $(s_1, s_2) \in LT$.

For every pair of spatial instances s_1, s_2, if $(g_1, g_2) \in BPT(\alpha)$ (where g_i is the geometry of s_i, $i = 1, 2$) and $(s_1, s_2) \in LT$, then $(s_1, s_2) \in partOf$ possibly holds, and we will add it to the initial instance mapping M.

Step 4: Verifying initial instance mapping M using semantic classification information. It is part of ontology (ABox and TBox) matching process, presented in [1].

We implement the method described above as part of GeoMap [1]. From the studied area (2km sq) of Nottingham city centre, 713 geospatial individuals of 47 types are added to OSGB Buildings and Places ontology from the OSGB Address Layer 2 and the OSGB Topology Layer [3], 253 geospatial individuals of 39 types are added into OSM ontology automatically from the building layer of OSM data. The ground truth instance mapping is obtained from manually matching all the instances in the two ontologies. It contains 286 'partOf' relations, and 73 'sameAs' relations can be inferred. The data used is available on `http://www.cs.nott.ac.uk/~hxd/GeoMap.html`. We compare the performance of our method with LogMap [4] and KnoFuss [5]. The precisions of mappings produced by GeoMap, LogMap and KnoFuss are 1, 0.24 and 0.18 respectively, and the recalls are 0.95, 0.38, 0.25 respectively. The precision and recall of GeoMap are much higher, mainly because LogMap and KnoFuss cannot make effective use of location information.

References

1. Du, H., Alechina, N., Jackson, M., Hart, G.: Matching Formal and Informal Geospatial Ontologies. In: Geographic Information Science at the Heart of Europe. Lecture Notes in Geoinformation and Cartography. Springer (2013) 155–171
2. OpenStreetMap: http://www.openstreetmap.org (2012)
3. Ordnance Survey: http://www.ordnancesurvey.co.uk/oswebsite (2012)
4. Jiménez-Ruiz, E., Grau, B.C.: LogMap: Logic-Based and Scalable Ontology Matching. In: International Semantic Web Conference (1). (2011) 273–288
5. Nikolov, Andriy and Uren, Victoria and Motta, Enrico: KnoFuss: a Comprehensive Architecture for Knowledge Fusion. In: Proceedings of the 4th International Conference on Knowledge Capture. (2007) 185–186