Automatically Constructing Geospatial Feature Taxonomies from *OpenStreetMap* Data

**Basel Shbita, Craig A. Knoblock**

*USC Information Sciences Institute*

18th IEEE International Conference on Semantic Computing (ICSC 2024)

02/06/2024
Agenda

• Intro
• Motivation
• Problem
• Approach
• Demo
• Evaluation
• Results & Discussion
• Related Work
• Future Work
• Conclusions
Intro

- accurate & comprehensive characterization of geospatial data in GIS
  - urban planning, route optimization, navigation systems, remote sensing ...
- structured taxonomy for geospatial features
Motivation

- **OpenStreetMap (OSM)** = rich source of geographic information
  - VGI (Volunteered Geographic Information)
    - relies on user contributions
  - geometries & attributes of both natural & urban features
  - limited...
    - no standardized taxonomy
    - heterogenous annotations
    - varying-granularity ("how specific")
    - inconsistent across regions
    - scale
  - can we still make use of this *noisy* data?
Formalizing the Problem

• How can we establish a comprehensive taxonomy of geospatial features from an unstructured crowdsourced groups of tags, automatically?

• Data-driven
  – “application” aware
  – “context” (region) aware
  – automatic
Approach

1. Construct base terminology:
   - Frequent non-informative terms:
     - apartments
     - building
     - driveway
     - house
     - highway
     - residential
     - service
   - Infrequent informative term:
     - residential_building

2. Build taxonomy:
   - Conflict resolution

3. Count parent-child relations:
   - Path frequency

Example:

```
{'apartments', 'building', 'driveway', 'highway', 'house', 'residential', 'service'}
```

```
<table>
<thead>
<tr>
<th>parent</th>
<th>child</th>
<th>counter</th>
</tr>
</thead>
<tbody>
<tr>
<td>building</td>
<td>house</td>
<td>15</td>
</tr>
<tr>
<td>highway</td>
<td>service</td>
<td>14</td>
</tr>
<tr>
<td>building</td>
<td>residential</td>
<td>33</td>
</tr>
<tr>
<td>highway</td>
<td>residential</td>
<td>22</td>
</tr>
<tr>
<td>building</td>
<td>apartments</td>
<td>2</td>
</tr>
<tr>
<td>service</td>
<td>driveway</td>
<td>5</td>
</tr>
</tbody>
</table>
```
Algorithm 1: Constructing a lightweight taxonomy.

Data: osmDataset
Result: taxonomyTree

for entity in osmDataset do
  tagPathsCounter[tagPath]++;  

for (tagPath,count) in tagPathsCounter.sort(order=descending) do
  insert_parent_child_pair(taxonomyTree, parent, child);

return taxonomyTree;
Demo

Usage


Automatically construct a lightweight taxonomy for geographic features using OpenStreetMap (OSM) data.

optional arguments:
-h, --help            show this help message and exit
--input INPUT         OSM dump (xml) input filename.
--output OUTPUT       Taxonomy tree (json) filename.
--threshold THRESHOLD Minimum frequency threshold per tag.
--blacklist BLACKLIST (txt) file with tags to ignore (one per line, as seen on OSM).
Evaluation

California USA (March 2023)
~150M instances
~10M tagged
1-16 tags (avg 2.3)

Greece (March 2023)
~40M instances
~2M tagged
1-13 tags (avg 2.1)
Results & Discussion

• California
Results & Discussion

• California v. Greece
Related Work

• Ontologies in Geospatial Data
  – Sun et al. [1]: Three-Level Ontology
    • manual
  – OSMonto [2]: Tag Hierarchies
    • explores tag relationships
  – WorldKG [3]: Geographic Knowledge
    • semantic representation

• Mapping OSM tags to Wikidata classes
  – Dsouza et al. [4]: neural architecture for tag-to-class mapping

Future Work

• Scalability
• Technology
  – ML & NLP for ambiguity & reconciliation
• User-centric
  – Incorporate user feedback
  – Tailor to specific applications
• Applications
  – Wider GIS integration
Conclusion

• Unsupervised & automatic approach for constructing lightweight geo-feature taxonomies from OpenStreetMap data
  – enhance OSM data usability
  – support data-driven analysis
  – improve geo-feature representation & categorization

• Source code available at:
  – https://github.com/basels/osm-taxonomy

Thank you for listening!

Questions?