Enriching Top-down Geo-ontologies Using Bottom-up Knowledge Mined from Linked Open Data

Yingjie Hu and Krzysztof Janowicz
STKO Lab, Department of Geography, U.C. Santa Barbara
Outline

• Introduction
• Workflow
• Experiment
• Conclusions
Introduction

• Geo-ontologies play important roles in GIScience
  • Enhance semantic interoperability
  • Improve geographic information retrieval
  • Support spatial decision making
  • ...

• A top-down approach for developing geo-ontologies
  • Pros: captures valuable expert knowledge; provides concise and meaningful terms
  • Cons: the derived ontology may be biased towards the opinions of the participating experts; or may be incomplete
Introduction

• Linked Open Data (LOD) cloud: a fast evolving data resource

• Merits of the LOD cloud:
  • A rich amount of data from both authorities and the general public
  • A lot of data are about geographic places: DBpedia, Geonames, LinkedGeoData, ...
  • Data are structured using Resource Description Framework (RDF)
Workflow

- A workflow for extracting bottom-up knowledge
- A concept learning approach
Workflow

• Three-stage process for extracting knowledge
  • 1. Extracting common properties
     • Properties only in positive instances
     • Properties shared by both instances
  • 2. Filtering and constructing properties
     • Filter out irrelevant properties, e.g., leaderTitle
     • Construct potentially relevant properties, e.g., population density
  • 3. Identifying distinguishable properties and classification thresholds
Workflow

• Three-stage process for extracting knowledge
  • 3. Identifying distinguishable properties and classification thresholds
    • Segment instances in a property into an increasing numbers of groups
    • Calculate entropy for each segmentation
    • Information gain before and after the property has been segmented

\[
\text{entropy}(X) = - \sum_{i=\{pos, neg\}} P(x_i) \log P(x_i)
\]

\[
\text{IG} = \text{entropy}_b(X) - \text{entropy}_a(X)
\]

(A) a property with clear cut

(B) a property with mixed instances
Experiment

• An example geographic concept (*port city*) and a possible top-down ontology

• A sample dataset from DBpedia
  • **Target category:** Port cities and towns of the United States Atlantic coast and Port cities and towns of the United States Pacific coast
  • **Positive instances:** 49 cities which have been classified into these two categories by Wikipedia users
  • **Negative instances:** 29 inland U.S. cities randomly selected
Experiment

• A Java program developed to identify common properties
  • Properties shared by at least 95% of positive instances and no more than 5% of negative instances: is dbpedia-owl:homeport of. dbpedia:Ship
  • Properties shared by at least 95% of both positive and negative instances

• Filtering irrelevant properties and constructing new properties
  • Filtering out irrelevant properties, e.g., names of the celebrities...
  • Constructing a new property, waterLandPercentage

<table>
<thead>
<tr>
<th>dbpedia-owl:areaTotal</th>
<th>dbpedia-owl:areaLand</th>
<th>dbpedia-owl:areaWater</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:populationTotal</td>
<td>dbpedia-owl:elevation</td>
<td>waterLandPercentage</td>
</tr>
</tbody>
</table>
Experiment

- Examining the information gain for each property under different numbers of segmentations
Experiment

• Aggregate the values of positive instances to derive thresholds

\[
elevation \leq 49.36 \\
waterLandPercentage \geq 11.79\%
\]

• Evaluation: does the extracted knowledge make sense?
  • An unseen dataset from DBpedia to test the extracted knowledge
  • 38 cities from Germany (21 positive and 17 negative)
Experiment

- Evaluating the extracted knowledge using unseen cities
Conclusions and Future work

• Top-down geo-ontologies capture valuable expert knowledge but may be biased or incomplete
• The rich amount of data from the LOD cloud provide a resource to mine geographic knowledge
• This study presents a preliminary framework to extract bottom-up knowledge from Linked Datasets

• Limitations and future work:
  • The selection of positive and negative instances
  • Regional variability of geographic concepts
Thank you!

Yingjie Hu
yingjiehu@umail.ucsb.edu
http://geog.ucsb.edu/~hu