Metadata Topic Harmonization and Semantic Search for Linked-Data-Driven Geoportals
-- A Case Study Using ArcGIS Online

Yingjie Hu\(^1\), Krzysztof Janowicz\(^1\), Sathya Prasad\(^2\), and Song Gao\(^1\)

\(^1\) STKO Lab, Department of Geography, U.C. Santa Barbara
\(^2\) Applications Prototype Lab, Esri Inc.
Outline

- Introduction and motivation
- Metadata Topic Harmonization
- Semantic Search on Linked Data
- Experiment and Evaluation
- Conclusions and Future Work
Introduction

- Geoportals are Web gateways that provide integrated access to geospatial resources

- Geoportals are key components of Spatial Data Infrastructure (SDI)

- Existing geoportals:
  - Data.gov
  - INSPIRE
  - California geoportal
  - ...
Introduction

- A typical *publish-find-bind* pattern has been used by many geoportals.

- Two factors that influence the search capability of a geoportal:
  - Quality of metadata
  - Search functionality
Introduction

• Quality of metadata
  - Multiple standards have been established to ensure the metadata quality, e.g., FGDC’s CSDGM and ISO 19115
  - However, data contributed to the same geoportal may be in different standards
Introduction

- Quality of metadata

- How to harmonize metadata in different standards?
  - Some elements can be automatically mapped using, e.g., NOAA’s metadata transformation tool
  - Some others have to be transformed manually, e.g., the topics
Introduction

• Search functionality

  - Traditional keyword-based search
    - Based on keyword matching
    - E.g., A search of “natural disaster” can only return maps which contain “natural” or “disaster”

  - Semantic search
    - Find maps based on the meaning of input query
    - E.g., return different disasters, such as wildfire, hurricane, earthquake…
Introduction

- Search functionality

  - The emerging of Linked-Data-driven Geoportals
    - Accommodate heterogeneous data using RDF data model
    - Graph-based data storage and browsing
    - Help discover the links hidden in data

  - Semantic search functionality for RDF data is not available for these novel geoportals
Metadata Topic Harmonization

- A machine learning based approach
  - A multi-label classification problem
  - One metadata can be associated with multiple topics
  - Based on titles and descriptions of each metadata entry
Metadata Topic Harmonization

- **LLDA (Labeled Latent Dirichlet Allocation)**
  - An extension of LDA by adding a component of supervised learning

- **Advantages of LLDA for topic harmonization compared with typical naïve Bayesian approach**
  - Considers each document as a mix of multiple topics
  - Robust estimation for prior probabilities of topics
  - Avoid overfitting for long descriptions
Semantic Search for Linked Data

• Query expansion
  - Extracting concepts and entities from the input query
  - Expanding them using related concepts and entities
    - Thematic concepts: Latent Semantic Analysis (LSA) and Wordnet
    - Geographic entities: Gazetteer service (Geonames)
Semantic Search for Linked Data

- Constructing Matching Features
  - Is this matching happens in title or in description?
  - Is this matching a thematic matching or geographic matching?
  - Is this an exact matching or a similar matching?
  - Resulted in 8 matching features (2 x 2 x 2)

<table>
<thead>
<tr>
<th>Title Thematic Exact match (TTE)</th>
<th>Title Thematic Similar match (TTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Geographic Exact match (TGE)</td>
<td>Title Geographic Similar match (TGS)</td>
</tr>
<tr>
<td>Snippet Thematic Exact match (STE)</td>
<td>Snippet Thematic Similar match (STS)</td>
</tr>
<tr>
<td>Snippet Geographic Exact match (SGE)</td>
<td>Snippet Geographic Similar match (SGS)</td>
</tr>
</tbody>
</table>
Semantic Search for Linked Data

- Constructing Matching Features
  - An additional feature: Thematic-Geo Interaction (TGI)

\[ TGI = (TTE + TTS + STE + STS) \times (TGE + TGS + SGE + SGS) \]

- Rationale for introducing this interaction feature:
  - Thematic or geo matching alone cannot determine the relevance
  - E.g., Searching “Crime in California”
  - “Crime in Florida” or “Waterbody in California” may not be what users want
  - “Robberies in Los Angeles” may be relevant

\[ R(q,m) = \lambda_1 TTE + \lambda_2 TTS + \lambda_3 TGE + \lambda_4 TGS + \lambda_5 STE + \lambda_6 STS + \lambda_7 SGE + \lambda_8 SGS + \lambda_9 TGI \]
Experiments and Evaluation

• Experimental data:
  - 26,917 metadata records from Data.gov in ISO 19115
  - 10,201 metadata records from ArcGIS Online

• Experiment procedure:
  • Use metadata from Data.gov to evaluate the performance of the LLDA-based workflow by comparing it with a naive Bayesian baseline
  • Apply the trained LLDA to the unstandardized ArcGIS Online data to automatically assign ISO 19115 topics
  • Test the semantic search function using human participant experiment experiment, and evaluate the quality of the search results.
Experiments and Evaluation

- Comparing the LLDA based approach with naïve Bayesian based approach
- Precision and recall curves
Experiments and Evaluation

• Human participant experiment
  - 7 human participants
  - Each person evaluate 10 queries and each query has 10 candidate maps
  - For each query and candidate, provide a score [0, 5]
Experiments and Evaluation

- Ten-fold cross validation using Pearson’s r
Semantic Search for Linked Data

- Embedding the semantic search to a geoportal
  - A SPARQL query to implement the regression model

```
SELECT ?item (COUNT(?titleThematicExact) AS ?TTE)
  (COUNT(?titleThematicSimilar) AS ?TTS)
  (COUNT(?titleGeoExact) as ?TGE)
  (COUNT(?titleGeoSimilar) as ?TGS)
  (COUNT(?snipThematicExact) as ?STE)
  (COUNT(?snipThematicSimilar) as ?STS)
  (COUNT(?snipGeoExact) as ?SGE)
  (COUNT(?snipGeoSimilar) as ?SGS)
  (( \lambda_1 \cdot ?TTE + \lambda_2 \cdot ?TTS + \lambda_3 \cdot ?TGE + \lambda_4 \cdot ?TGS + \lambda_5 \cdot ?STE + \lambda_6 \cdot ?STS + \\
\lambda_7 \cdot ?SGE + \lambda_8 \cdot ?SGS + \lambda_9 \cdot ?TGI) as ?ranking)
WHERE {
  OPTIONAL {
    ?item :hasTitleThematicTerm ?titleThematicExact .
    FILTER { ?titleThematicKey = :exactThematicTerm } }
  OPTIONAL {
    ?item :hasTitleThematicTerm ?titleThematicSimilar .
    FILTER { ?titleThematicSimilar = :expandedThematicTerm } }
  OPTIONAL {
    ?item :hasTitleGeoTerm ?titleGeoExact .
    FILTER { ?titleGeoExact = :exactGeoTerm } }
  OPTIONAL {
    ?item :hasTitleGeoTerm ?titleGeoSimilar .
    FILTER { ?titleGeoSimilar = :expandedGeoTerm } }
}
```
Interactive prototype

- [http://stko-exp.geog.ucsb.edu/linkedportal](http://stko-exp.geog.ucsb.edu/linkedportal)
Conclusions and Future Work

- Geoportals provide integrated access to geospatial resources
- The quality of metadata and the capability of the search function are two major factors affecting resource discovery
- We present a LLDA-based approach for harmonizing metadata topics, as well as enabled semantic search for RDF data
- Limitations and future work
  - Small scale human participants test need to be expanded
  - Increase the response efficiency of semantic search
Thank you!

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