

Enabling Semantic Search and Knowledge Discovery for ArcGIS Online: -- A Linked-Data-Driven Approach

Yingjie Hu¹, Krzysztof Janowicz¹, Sathya Prasad², and Song Gao¹

¹ STKO Lab, Department of Geography, U.C. Santa Barbara

² Applications Prototype Lab, Esri Inc.

Outline

- **Introduction**
- **Data Conversion and Enrichment**
- **Semantic Search on Linked Data**
- **Knowledge Discovery using Flexible Queries**
- **Conclusions and Future Work**

Introduction

- **What is ArcGIS Online?**
 - A geoportal providing geospatial data and services
 - Contents are from both authorities and the general public
- **ArcGIS Online Metadata**
 - A rich amount of information, including titles, descriptions, tags, created dates, revised dates, ..., about the resources
 - Metadata data are available through Esri's public REST API

Introduction

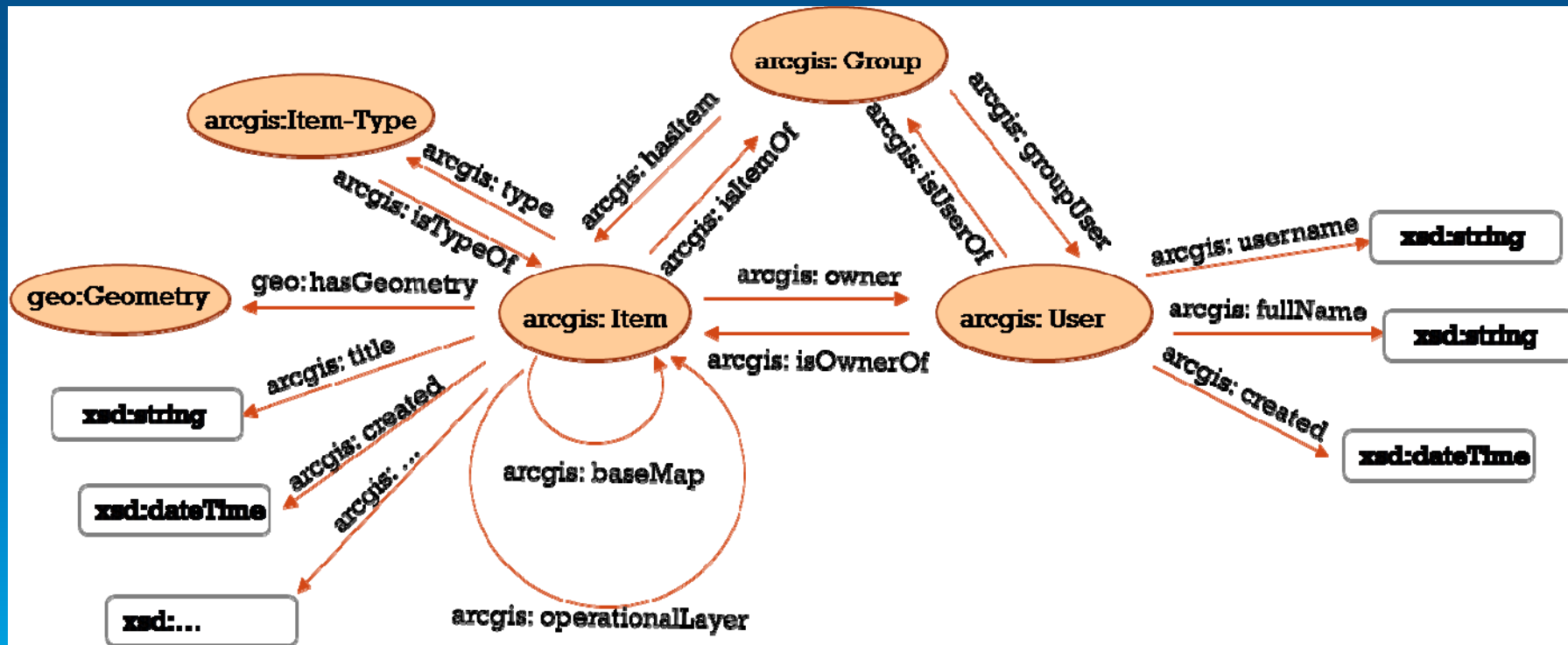
- **Limitations of ArcGIS Online map search**
 - **Content matching is based on keywords**
 - E.g., A search of “natural disaster” can only return maps which contain “natural” or “disaster”
 - But the user may want to find maps on different disasters, such as wildfire, hurricane, earthquake...
 - **Search is constrained by pre-designed REST API**
 - E.g., API allows searching maps based on created date
 - But what if the user wants to perform some customized queries, e.g., finding the basemaps which have been used most frequently

Introduction

- **What we did in this work**
 - Re-organizing and enriching ArcGIS Online Metadata using Linked Data principles
 - Addressing the two limitations of search
 - Semantic search based on Linked Data
 - Flexible queries for knowledge discovery
- **While using ArcGIS Online as the experimental platform, the proposed methods can be applied to other geoportals**

Data Conversion and Enrichment

- Ontology for ArcGIS Online



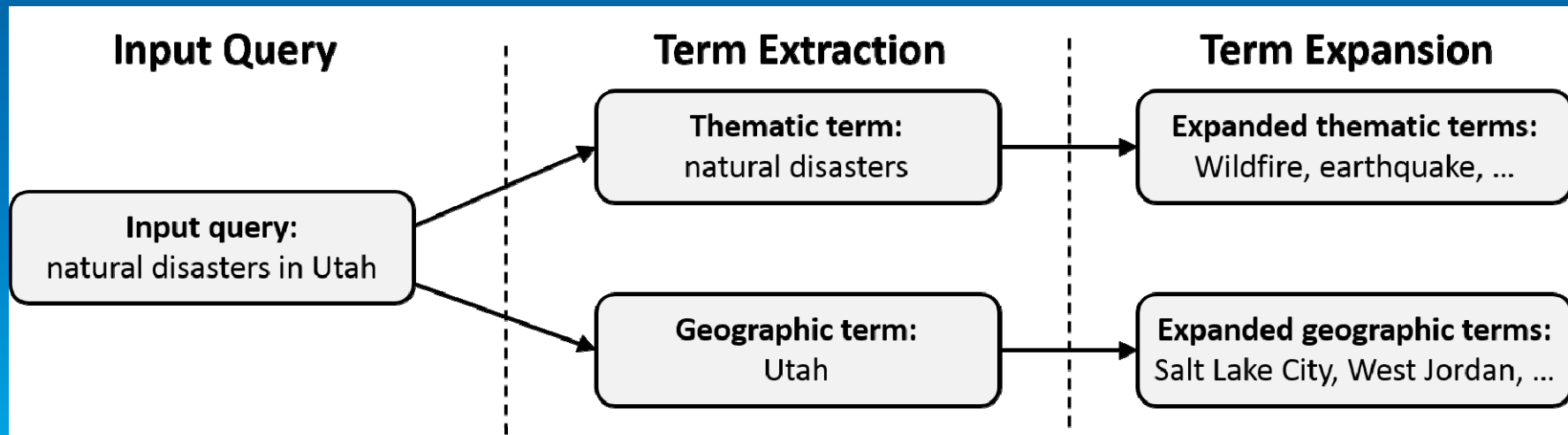
Data Conversion and Enrichment

- **Enriching metadata with entities from textual titles and descriptions**
 - **Problem: original tags from general users are often incomplete or contain errors**
 - **Tools: DBpedia Spotlight and OpenCalais**
- **Example:**
 - **Map title: Tragedy and Kindness: Brisbane Floods, January 2011**
 - **Snippet : This map shows pictures in Brisbane, Australia in the aftermath of the floods that occurred in January 2011**
 - **Original tags: book**
 - **Enriched tags:**
 - **Title thematic terms: flood, january, kind, natural disaster, tragedy**
 - **Title geo-terms: brisbane**
 - **Snippet thematic terms: aftermath, flood, january, natural disaster**
 - **Snippet geo-terms: australia, brisbane, brisbane,_australia**

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**

- Find a matching between the expanded input query and the enriched metadata
- Is this matching happens in title or in description?
- Is this matching a thematic matching or geographic matching?
- Is this a exact matching or a similar matching?
- Resulted in 8 matching features (2 x 2 x 2)

Title Thematic Exact match (TTE)

Title Geographic Exact match (TGE)

Snippet Thematic Exact match (STE)

Snippet Geographic Exact match (SGE)

Title Thematic Similar match (TTS)

Title Geographic Similar match (TGS)

Snippet Thematic Similar match (STS)

Snippet Geographic Similar match (SGS)

Semantic Search for Linked Data

- **Constructing Matching Features**

- **An additional feature: Thematic-Geo Interaction (TGI)**

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

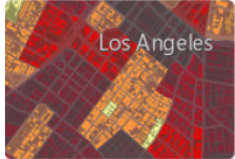

- **Rationale for introducing this interaction feature:**
 - Thematic or geo matching alone cannot determine the relevance
 - E.g., Searching “Crime in California”
 - “Crime in Florida” or “Waterbody in California” may not be what users want
 - “Robberies in Los Angeles” may be relevant

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

Semantic Search for Linked Data

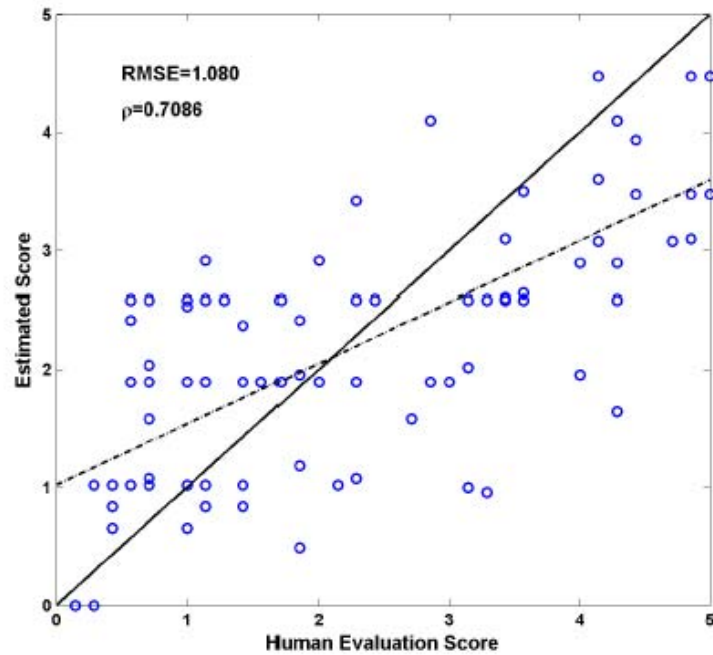
- Estimating parameters through human participant test
 - 7 human participants
 - Each person evaluate 10 queries and each query has 10 candidate maps
 - For each query and candidate, provide a score [0, 5]

Query 3: "california population density"

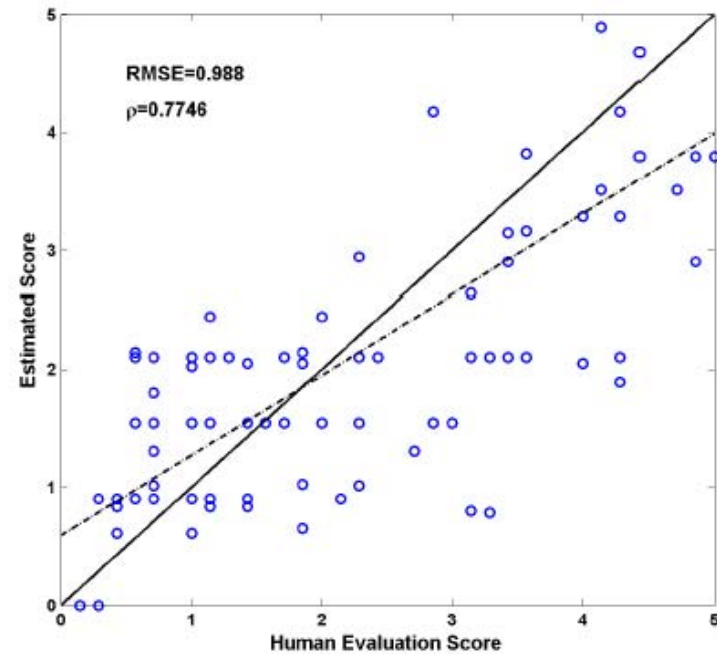
Map	Link of the Map
<p>3.1</p>  <p>Los Angeles Population Density This map emphasizes areas with the highest population density (more than 50,000 persons per square kilometer).</p>	<p>http://www.arcgis.com/home/webmap/viewer.html?webmap=4971065a7a734e31a7079ace59a19f27</p>
<p>3.2</p>  <p>California earthquake faults showing USGS fault data with focus on California</p>	<p>http://www.arcgis.com/home/webmap/viewer.html?webmap=3489cb99f9174edb86dda62b6550772b</p>

Semantic Search for Linked Data

- Evaluating the fitness of the model



(a) Without the interaction variable



(b) With the interaction variable

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)
(((?TTE+?TTS+?STE+?STS)*(?TGE+?TGS+?SGE+?SGS)) as ?TGI)
((  $\lambda_1$ *?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) 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 ) }
```

Knowledge Discovery using Flexible Queries

- **Which Basemaps Are Most Popular?**

- **Popularity based on the times of view**

```
SELECT DISTINCT ?baseMap ?numViews
WHERE { ?baseMap arcgis:isBaseMapOf ?item .
        ?baseMap arcgis:numViews ?numViews }
ORDER BY DESC(?numViews) LIMIT 10
```

- *World Boundaries and Places* map has been viewed most frequently

- **Popularity based on the times of actual usage**

```
SELECT ?baseMap (count(distinct ?item) as ?usedTimes)
WHERE { ?baseMap arcgis:isBaseMapOf ?item }
GROUP BY ?baseMap
ORDER BY DESC(?usedTimes) LIMIT 10
```

- *World Topographic Map* is the one that have been used
 - Used 13,507 times; significantly more than the usage of the *World Boundaries and Places* map 2,855 times.

Knowledge Discovery using Flexible Queries

- **Which Group Has Created Most Maps about California?**
 - Using a query based on GeoSPARQL

```
SELECT DISTINCT ?group (count(?item) as ?itemCount)
WHERE { ?group arcgis:type arcgis:ArcGIS-Group .
        ?group arcgis:hasItem ?item .
        ?item geo:hasGeometry ?itemGeo .
        ?itemGeo geo:asWKT ?wkt
        Filter (geof:sfWithin(?wkt, Polygon((-125 42, -120 42,
-120 39, -114 34, -114 32,
-120 32, -125 42))^sf:wktLiteral)) }
```

The group is a Web GIS class from the University of California, Riverside

Interactive prototype

- A sample of ArcGIS Online dataset from July 2013 to Sep. 2013
- <http://stko-exp.geog.ucsb.edu/linkedarcgis/>

The screenshot shows the 'Linked Data Portal for ArcGIS Online' interface. At the top, there are navigation tabs: 'Semantic Search', 'Knowledge Discovery', 'GeoSPARQL', 'Statistics', and 'About'. The main heading is 'Semantic Search', with a subtext: 'This module provides semantic search based on the exported linked map data (currently 35,624 maps in total)'. Below this is a search bar containing the text 'natural disasters in utah' and a green 'Search' button. The results are organized into three columns:

- Geo & thematic matching: 5 results**
 - The Great Utah Shake Out**: Utah ShakeOut 2013 - Salt Lake City Segment 7.0 (Exercise). FEMA Hazus Earthquake Analysis of Potential Earthquake Risk Resulted from a Salt Lake City Segment M 7.0.
 - Utah Water Supply**: Utah Drought Conditions/Phases.
- Thematic matching: 124 results**
 - Moore, Oklahoma-Tornadoes 2013**: US Historical Tornadoes.
 - Occurrence of flood events in Europe 1998 - 2008**: Occurrence of flood events in Europe.
- Geospatial matching: 21 results**
 - Utah basemaps**: 7 Utah AGRC basemaps.
 - Utah Urban Tree Canopy**: A study of urban tree canopy in several counties in Utah.

Conclusions and Future Work

- **ArcGIS Online is an online geoportal which contain data from both authoritative agencies and the general public**
- **This work converted a sample of ArcGIS Online metadata into Linked Data**
- **Enabled semantic search and flexible queries on the RDF-based ArcGIS metadata**

- **This work can be improved in several aspects:**
 - **The small-scale human participant test can be expanded**
 - **The efficiency of the search need to be increased**
 - **Themes can be added to the maps to enable facet search**

Thank you!

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