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

Yingjie Hu, Krzysztof Janowicz, Sathya Prasad and Song Gao

Abstract ArcGIS Online is a unified Web portal designed by Environment System Research Institute (ESRI). It contains a rich collection of Web maps, layers, and services contributed by GIS users throughout the world. The metadata about these GIS resources reside in data silos that can be accessed via a Web API. While this is sufficient for simple syntax-based searches, it does not support more advanced queries, e.g., finding maps based on the semantics of the search terms, or performing customized queries that are not pre-designed in the API. In metadata, titles and descriptions are commonly available attributes which provide important information about the content of the GIS resources. However, such data cannot be easily used since they are in the form of unstructured natural language. To address these difficulties, we combine data-driven techniques with theory-driven approaches to enable semantic search and knowledge discovery for ArcGIS Online. We develop an ontology for ArcGIS Online data, convert the metadata into Linked Data, and enrich the metadata by extracting thematic concepts and geographic entities from titles and descriptions. Based on a human participant experiment, we calibrate a linear regression model for semantic search, and demonstrate the flexible queries for knowledge discovery that are not possible in the existing Web API. While this research is based on the ArcGIS Online data, the presented methods can also be applied to other GIS cloud services and data infrastructures.

Keywords Metadata · Semantic search · Linked data · Geoportal · ArcGIS online

Y. Hu (✉) · K. Janowicz · S. Gao
STKO Lab, University of California Santa Barbara, Santa Barbara, CA, USA
e-mail: yingjiehu@umail.ucsb.edu

K. Janowicz
e-mail: jano@ucsb.edu

S. Gao
e-mail: sgao@uamil.ucsb.edu

S. Prasad
Applications Prototype Lab, ESRI Inc, Redlands, CA, USA
e-mail: sprasad@esri.com

F. Bação et al. (eds.), *AGILE 2015*, Lecture Notes in Geoinformation and Cartography, DOI 10.1007/978-3-319-16787-9_7

1 Introduction and Motivation

ArcGIS Online[1] is a geoportal developed by Environment System Research Institute (ESRI). It allows GIS users throughout the world to create, edit, and share geo-data, Web maps, services, and GIS tools (Dangermond 2009). To remain manageable, the plethora of ArcGIS Online resources (called *items*) are accompanied by a rich set of metadata, including titles, descriptions, and information about users, user groups, and so forth. Based on these metadata, one can browse through the collection of GIS resources, or sort them by features such as the popularity or date.

Currently, the data and metadata reside in data silos, and can be accessed via a RESTful Web API. However, only queries which satisfy pre-designed templates can be submitted to retrieve data. This hinders flexible knowledge discovery. For instance, if one wants to find out *"which users have produced highly rated maps about natural disasters in the USA"*, such a query has to be first hard-coded into the current API before it can be used. While it is possible to embed a small number of frequent queries, a GIS user can easily come up with a new customized search that has not been designed before. This limitation demands a solution that allows flexible and customized queries.

Meanwhile, as new GIS resources are being created every day, the existing keyword-based search is becoming increasingly limited for finding results that match a user's interests. For example, a search of *natural disasters in Oklahoma* would not be able to return maps about *tornados in Moore*, since the term *tornado* is not in the query and the system does not understand *Moore* is a city in *Oklahoma*. Thus, it is necessary to establish an intelligent search method that can retrieve maps based on the semantic and geographic meaning of the input query.

To enable semantic search as well as flexible knowledge discovery, ArcGIS Online resources should be annotated with machine readable terms which can characterize the map content. Titles and descriptions in the metadata can deliver important information to humans, but they cannot be directly used by machines. While ArcGIS Online also allows users to assign structured tags to maps, those tags are often incomplete or misleading due to the voluntary nature of the data.

To address these restrictions and thus improve the usability of Online GIS cloud services, three steps need to be taken: (I) the metadata provided by the users have to be enriched with machine readable terms; (II) all metadata have to be converted into a format which frees it from data silos and allows flexible queries; (III) a new user interface has to be developed to provide semantic search and enable interesting knowledge discovery. **The contributions of our work are as follows**:

- We present a workflow to enrich the original metadata with machine-readable concepts and named entities.
- We develop an ontology for ArcGIS Online and convert a sample of ArcGIS online metadata into Linked Data.

[1] http://www.arcgis.com.

- We design a semantic search function by expanding input queries and tuning a linear regression model.
- We discuss two flexible queries enabled by our solution, and show the knowledge that can be discovered from the Linked metadata.
- We implement a prototypical Linked-Data-driven Web portal for ArcGIS Online using the presented methods.

This work makes use of the Semantic Web technology stack, including the concepts (Berners-Lee et al. 2001), Linked Data principles (Bizer et al. 2009), Resource Description Framework (RDF) (Hitzler et al. 2011), and other techniques. Such technology stack has been used in existing works to facilitate knowledge discovery (Hu et al. 2013; Keßler et al. 2012). For a more detailed rationale on the use of Linked Data and semantics in GIScience, readers are recommended to (Janowicz et al. 2012). While we have used ArcGIS Online data in this research, the presented methods could also be generalized to other GIS cloud services.

The remainder of this paper is organized as follows. Section 2 provides a brief description on ArcGIS Online. Section 3 discusses the workflow to extract metadata from the API, convert them into RDF, and enrich them with machine-readable terms. Section 4 presents a semantic search method which retrieves GIS resources based on semantic and geographic relevance. Section 5 employs two customized queries to demonstrate the flexible search enabled by the Linked-Data-driven solution. Section 6 describes the prototype implemented as a proof-of-concept. Section 7 summarizes this work and discusses future directions.

2 ArcGIS Online—A GIS Cloud Service

As a collaborative platform, ArcGIS Online enables GIS users throughout the world to create, edit, and share maps, services, and other GIS resources. ArcGIS Online contains a large variety of resources, including Web maps (consisting of a basemap and several layers), services (e.g., map service, feature service, geoprocessing service), as well as document-based data (e.g., shapefiles, CSV files). ArcGIS Online also contains a large number of registered users and user groups, e.g., a *transportation group*. Finally ArcGIS Online also provides a data sharing and reuse mechanism: users can integrate existing services into the maps instead of having to upload all data.

While there are datasets contributed by U.S. Geological Survey (USGS), Federal Emergency Management Agency (FEMA), and other authoritative institutes, a large proportion of the Web maps are volunteered geographic information (VGI). Similar to other VGI, (meta)data quality is one important issue that needs to be addressed (Goodchild and Glennon 2010). In this work, we mainly focus on enriching the metadata of Web maps and services, since it is directly related to the semantic search function which will be discussed later.

The metadata of Web maps and services are recorded in a semi-automatic manner. Information items, such as the map ID, creation date, and the creator's name, are generated by the system automatically, while the creator needs to manually type in a title, a short description (called *snippet*), and several tags. ArcGIS Online uses these tags as annotations to find maps according to a particular topic. The examples below show the titles, descriptions, and tags of three ArcGIS Online maps. While some of the tags are descriptive (e.g., *Thompsons Lake* and *tornadoes*), others are more difficult to interpret or even misleading. For example, the second map is tagged with *book*, while the map is actually about floods. While the tags of the first and the third map are more comprehensive, *landscape* could be one additional tag for the first map to characterize the type of *change*. Similarly, *natural disaster* could form an additional tag for the third map. Due to the voluntary nature, we cannot require users to provide a list that contain every possibly related tag, nor can we mandate the usage of certain pre-defined tags. However, map titles and descriptions often provide useful information about the content of a map, and therefore can be used to extract meaningful tags.

1. **Map title**: Landscape Change: Thompsons Lake, NY
   Snippet (**Description**): Minor landscape changes near Thompsons Lake in the Helderberg's of upstate New York
   Tags: Thompsons Lake, NY, Change, GIS
   URL: www.arcgis.com/home/item.html?id=849ae63adc2c446f9ba54c10a50fbd7b
2. **Map title**: Tragedy and Kindness: Brisbane Floods, January 2011
   Snippet (**Description**): This map shows pictures in Brisbane, Australia in the aftermath of the floods that occurred in January 2011
   Tags: book
   URL: www.arcgis.com/home/item.html?id=07845c87cd7e4f2eb2292b978267b6af
3. **Map title**: Moore, Oklahoma—Tornadoes from the 1950's to the 2000's-Copy
   Snippet (**Description**): Map showing tornadoes in Moore, Oklahoma from the 1950's to the 2000's decade by decade and classified by strength
   Tags: tornadoes, Fujita Scale, Tornado Alley
   URL: www.arcgis.com/home/item.html?id=45f31bea7f624766bf23827ec488d9a3

3 Data Conversion, Ontology Design, and Enrichment

In this section, we describe the process of converting a sample of the ArcGIS Online metadata into RDF and enriching the data with thematic terms and geographic entities extracted from map titles and descriptions.

3.1 ArcGIS Online Data Sample

ArcGIS Online data can be accessed and retrieved using the ArcGIS Online REST API.[2] In this work, we use a sample retrieved between 7 January 2013 and 9 January 2013. This sample contains information about 35,624 Web maps, 13,649 feature services, 5565 map services, 8582 Web mapping applications, 20,725 users, and 2052 user groups. These data can be divided into three categories: *GIS resources* (including maps, services, tools, and so forth), *ArcGIS Online users*, and their *user groups*. For our Linked Data conversion we are especially interested in the relations between those categories, such as: users create GIS resources; multiple users can belong to the same group; a single user can belong to multiple groups; if a user belongs to a group, then her public GIS resources also belongs to this group, and so forth. Following the ArcGIS Online terminology, we will refer to GIS resources as *items*.

3.2 Ontology for ArcGIS Online

An ontology formally restricts the interpretation of domain vocabulary towards their intended meaning, and can be considered as the backbone for data organization. A growing number of well-defined ontologies exist and have been used in many projects, e.g., Dublin Core (dc)[3] and Friend Of A Friend (foaf).[4] Reusing existing ontologies is generally a good practice to facilitate data exchange and integration (Heath and Bizer 2011). However, ArcGIS Online already has an established schema embedded in many of its existing functionalities. While it is possible to semantically align parts of the ArcGIS Online schema to existing ontologies (e.g., from *arcgis*:*owner* to *dc*:*creator*), such a translation may nevertheless bring compatibility issues that may require code revisions in other ArcGIS Online modules. Even more, as (to our best knowledge) there are no existing ontologies for Online GIS cloud services, we would have to import a wide variety of existing ontologies which often leads to unintended logical consequences (Janowicz et al. 2012). Therefore, we design a specific ontology for ArcGIS Online, which can be generalized to other GIS cloud services and aligned to existing ontologies (instead of importing them).

Figure 1 illustrates the major classes and relations of the developed ontology. *arcgis*:*Item* is a general class for all GIS resources, such as Web maps and map services. The particular type of the GIS resource is defined by the class *arcgis*:*Item-Type* whose instances include *arcgis:Web-Map*, *arcgis*:*Map-Service*, and *arcgis*:*Feature-Service*. If an *item* is a Web map, it also has links to its basemap and other

[2] http://resources.arcgis.com/en/help/arcgis-rest-api/index.html.

[3] http://dublincore.org/documents/dcmi-terms/.

[4] http://xmlns.com/foaf/spec/.

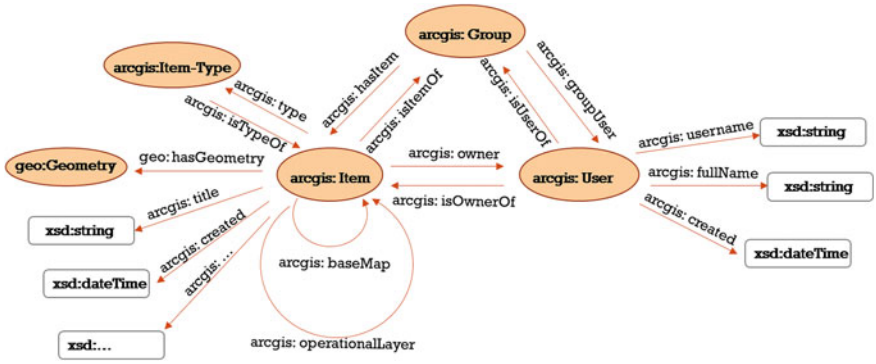

Fig. 1 Ontology for ArcGIS online (ellipses are classes, rectangles are literals)

layers through the relations of *arcgis*:*baseMap* and *arcgis*:*operationalLayer*. The geographic extent of an *arcgis*:*item* is represented using the class *geo*:*Geometry* from OGC's GeoSPARQL vocabulary,[5] which is defined to enable geographic queries based on SPARQL.[6] Figure 1 also shows the interactions among the classes *arcgis*:*Item*, *arcgis*:*User*, and *arcgis*:*Group*. For lack of space, we cannot discuss any axioms in detail here but refer the interested reader to our full implementation using the Web Ontology Language (OWL) at http://sejp.geog.ucsb.edu/esri/ontology.

3.3 Entity Naming

To publish and share high quality data on the Semantic Web, the naming of the entities (e.g., maps, users, and groups) should follow the established Linked Data principles (Heath and Bizer 2011; Keßler et al. 2012). While ArcGIS Online uses a hash string to identify its Web maps, it also provides globally unique HTTP URLs for users to access these GIS resources. Following Linked Data principles 1 and 2, we reuse these HTTP URLs to name the entities in the ArcGIS Online data. Below are some examples for the entity naming.

- **Web Map**: www.arcgis.com/sharing/rest/content/items/be9b7b9fb3514757ba5e6000aa4bd5ba
- **Feature Service**: services1.arcgis.com/10Nf6qqrwDJLkcL2/arcgis/rest/services/Rivers/FeatureServer

[5] http://schemas.opengis.net/geosparql/1.0/geosparql_vocab_all.rdf.

[6] SPARQL (http://www.w3.org/TR/sparql11-overview/) is the query language for graphed data, e.g., Linked Data, standardized by the World Wide Web Consortium (W3C).

- **Map Service**: tiles.arcgis.com/tiles/XWaQZrOGjgrsZ6Cu/arcgis/rest/services/CambridgeBasemap/MapServer
- **ArcGIS Online user**: www.arcgis.com/sharing/rest/community/users/ezgis76
- **ArcGIS Online group**: www.arcgis.com/sharing/rest/community/groups/a707bf7643cf47b89548d0a0184b6950

All of these entity names can be *dereferenced* (by appending "?f = json" to specify the output format), which leads to information about these GIS resources, users and groups. This practice follows the 3rd rule of the Linked Data principles: *published data resources should be self-descriptive*. Currently, we are also working on establishing external links from ArcGIS Online maps to *GeoNames* and *DBpedia* which will satisfy the 4th rule.

3.4 Enriching Data with Geographic Entities and Thematic Terms

Among the rich ArcGIS Online metadata, titles and descriptions often convey useful information about the map content. For example, given a map titled "Los Angeles population density", one can grasp the general idea of the map without having to look into the map. Titles and descriptions are represented in the form of natural language, which is easy for humans to read, but difficult for machines to process.

Therefore, our goal is to extract meaningful terms from titles and descriptions to characterize the content of maps. In contrast to plain text documents, the content of a map can often be divided into two parts: the thematic part and the geographic part. Examples in our sample dataset include maps entitled *"New York Population Density"*, *"California Fires"*, and *"Hurricans in Florida"*, to name but a few. Consequently, our extraction and enrichment process differentiates thematic and geographic terms. This differentiation is important for the functionality of semantic search, as thematic similarity and geographic similarity need to be treated separately (Jones et al. 2001).

We use two Linked-Data-driven and semantically-enabled Web services to extract and differentiate thematic and geographic terms: *DBpedia Spotlight* (Mendes et al. 2011) and *OpenCalais* (Butuc 2009). *DBpedia Spotlight* is an automatic annotation system that can label out the terms that have corresponding entries in *DBpedia* (Auer et al. 2007; Bizer et al. 2009). An important feature of *DBpedia Spotlight* is its capability to disambiguate a term that has multiple matching entries based on the term's context. For example, the term *Santa Barbara* can refer to a

place[7] but also a TV series.[8] To find the most likely *DBpedia resource* for *Santa Barbara*, DBpedia Spotlight applies the TF-IDF (term frequency- inverse document frequency) similarity matching between the surrounding text (i.e., the map titles or descriptions) and the corresponding *DBpedia* content and then ranks the resources according to the matching score. Such disambiguation is possible as DBpedia uses rich ontology, and therefore places and TV series can be distinguished by their types.

Similar to DBpedia Spotlight, *OpenCalais* can extract and semantically categorize entities. While typically Spotlight is able to extract most of the important thematic concepts and geographic entities for our sample data, *OpenCalais* complements those results with *broader terms*. For instance, it will extract *natural disaster*, if *earthquake* is present in a map's title or description. Thus, we employ both services for the enrichment process. Additionally, we also differentiate between the entities extracted from the map titles and those extracted from the descriptions. The list below shows the three ArcGIS Online examples (discussed in Sect. 2) with previous and newly added tags.

1. **Map title**: Landscape Change: Thompsons Lake, NY
   Previous Tags: Thompsons Lake, NY, Change, GIS
   After Enrichment:

   Title thematic terms: change, lake, landscape
   Title geo-terms: Thompson
   Descriptions thematic terms: lake, landscape, minor, Thompsons lake
   Descriptions geo-terms: New York, Thompson, upstate new york

2. **Map title**: Tragedy and Kindness: Brisbane Floods, January 2011
   Previous Tags: book
   After Enrichment:

   Title thematic terms: flood, January, kind, natural disaster, tragedy
   Title geo-terms: Brisbane
   Descriptions thematic terms: aftermath, flood, January, natural disaster, picture
   Descriptions geo-terms: Australia, Brisbane, Brisbane,_Australia

3. **Map title**: Moore, Oklahoma—Tornadoes from the 1950's to the 2000's-Copy
   Previous Tags: Tornadoes, Fujita Scale, Tornado Alley
   After Enrichment:

   Title thematic terms: 1950, 2000, natural disaster, Tornado
   Title geo-terms: Moore, Moore,_Oklahoma, Oklahoma
   Des. thematic terms: 1950, 2000, classified, decade, natural disaster, strength, Tornado
   Descriptions geo-terms: Moore, Moore,_Oklahoma, Oklahoma

[7] http://live.dbpedia.org/page/Santa_Barbara,_California.
[8] http://live.dbpedia.org/page/Santa_Barbara_(TV_series).

Finally, based on the developed ontology, the naming schema, and the data enrichment process, we convert the ArcGIS Online sample to RDF triples using a customized script developed on top of the Jena API[9] and store the Linked Data in the GeoSPARQL-enabled Parliament triple store (Battle and Kolas 2012). The newly extracted terms are inserted into Parliament as triples, and are linked to the corresponding maps to complete the enrichment process. It is worth to note that the original tags voluntarily contributed by users are no longer used due to their varied completeness and potential errors.

4 Semantic Search for Maps

In this section, we discuss our approach to enabling semantic search based on the enriched metadata. Semantic search attempts to understand the meaning of the user's input query, thereby improving the search results (Guha et al. 2003; Zhou et al. 2007). This differs from traditional keyword search which is based on the occurrence of syntactic matches (Tran et al. 2007).

4.1 Query Expansion

The first step towards semantic search is to expand the natural language query from the user to cover related terms. Similar to the data enrichment process, we use *DBpedia Spotlight* and *OpenCalais* to dynamically extract thematic terms and geographic entities, which provide the basic terms for query expansion. The expansion of thematic terms and geographic entities should be treated differently. For the thematic terms, it is important to identify the terms which have similar meaning but different syntaxes, whereas for the geographic entities, place hierarchies and spatial proximity should be taken into account (Jones et al. 2001).

For the expansion of the thematic aspect, we use the UMBC Semantic Similarity Service (Han et al. 2013), which is based on a combination of Latent Semantic Analysis (LSA) (Landauer and Dumais 1997) and knowledge from WordNet (Miller 1995). Given a thematic term, the UMBC Service can find the top n semantically similar terms based on a similarity score ranking. This allows us to also find maps about *reservoirs* if a user searches for *lakes*. For the expansion of geographic entities, we use the GeoNames gazetteer service to find the top 10 nearby and related places. Thus, if the user's query contains *California*, popular places related to California, such as *San Fransisco* and *Los Angeles*, will also be returned as related entities. The list below shows an example for expanding a user's query.

[9] https://jena.apache.org/.

- **User query**: Vacations in Hawaii
  Extracted Terms:

  Thematic term: Vacation
  Geo term: Hawaii
  Expanded Terms:

  Thematic terms: holiday, honeymoon, leisure, picnic, getaway, sabbatical, spring break, camping, leave, resort
  Geo terms: Honolulu, Hawaii County

4.2 Constructing Matching Features

Based on the expanded queries and the enriched map data, we construct matching features to quantify the relevance between a query and the candidate maps. 8 matching features have been constructed using the thematic concepts and geographic entities extracted from titles and descriptions. To avoid mismatches due to minor syntax variations (e.g., *library* and *libraries*, or *California* and *California*), we make use of the stemming technique, and convert terms to lowercase. Detailed explanation for each feature is listed as below:

- **Title Thematic Exact matching** (**TTE**): the number of matches between the original user input thematic terms and the thematic terms in the titles of candidate maps (e.g., *vacation* from the query and *vacation* from the map title).
- **Title Thematic Similar matching** (**TTS**): the number of matches between the expanded thematic terms from the user's query and the thematic terms from the titles of candidate maps (e.g., the term *holiday* expanded from the input term *vacation* and *holiday* in the map title).
- **Title Geographic Exact matching** (**TGE**): the number of matches between the geographic entities from the original user input query and the geographic entities from the titles of candidate maps (e.g., *California* from the input query and *California* from the map title).
- **Title Geographic Similar matching** (**TGS**): the number of matches between the expanded geographic entities and the geographic entities from the titles of candidate maps (e.g., *Los Angeles* expanded from the input term *California* and *Los Angeles* in the title).
- **Description Thematic Exact matching** (**DTE**): the number of matches between the original input thematic terms and the thematic terms in the descriptions of candidate maps (e.g., *water body* from the query and *water body* in the map description).
- **Description Thematic Similar matching** (**DTS**): the number of matches between the expanded thematic terms and the thematic terms in the descriptions of candidate maps (e.g., *lake* expanded from the input term *water body* and *lake* in the map description).

- **Description Geographic Exact matching** (**DGE**): the number of matches between the geographic entities from the original input and the geographic entities in the descriptions of candidate maps (e.g., *California* from input query and *California* in the map description).
- **Description Geographic Similar matching** (**DGS**): the number of matches between the expanded geographic entities and those in the descriptions of candidate maps (e.g., *Los Angeles* expanded from *California* in the input query and *Los Angeles* in the description).

In addition, an interaction variable, namely **Thematic-Geo Interaction** (TGI), has been introduced, which is defined as the sum of thematic matching scores multiplying the sum of geographic matching scores; see Eq. 1.

$$TGI = (TTE + TTS + DTE + DTS) \times (TGE + TGS + DGE + DGS) \quad (1)$$

As denoted by the name, TGI captures the interactions between thematic matches and geographic matches. TGI will have a positive value only when both thematic and geographic matches exist; otherwise it will be zero.

The rationale for introducing this 9th feature is that a good result for map search often needs to have both thematic and geographic matches. Consider a query for *Drugs and Crime in California*. A map that has a high number of thematic matches (and thus a high thematic matching score) but is about *Drugs and Crime in Spain* may not be of interest to the user. On the contrary, a map that has only one thematic match, e.g., *drugs*, but also the geographic match with *California* is more likely to be considered as a good match for a user's query. In fact, we will test this assumption in the evaluation section.

4.3 Ranking Results Using a Linear Regression Model

While we have constructed 9 matching features, a method is necessary to combine these matching features and quantify the relevances between an input query and the candidate maps. Specifically, such a method should satisfy two criteria: (1) correctly rank the relevance between a query and a candidate map; (2) can be easily embedded into a SPARQL query (since RDF has been employed to interlink the data).

We propose to use a regression model to combine the 9 matching features. Such a model can satisfy the above criteria: it can provide fair ranking and can be easily integrated into a SPARQL query. Equation 2 shows the regression model.

$$\begin{aligned} R(q,m) = {} & \lambda_1 TTE + \lambda_2 TTS + \lambda_3 DTE + \lambda_4 DTS + \lambda_5 TGE \\ & + \lambda_6 TGS + \lambda_7 DGE + \lambda_8 DGS + \lambda_9 TGI \end{aligned} \quad (2)$$

where $R(q,m)$ represents the ranking score between query q and map m. $TTE, TTS, \ldots, TGI$ are the 9 matching features respectively, and $\lambda_1, \lambda_2, \ldots, \lambda_9$

are the coefficients for each matching feature. It is worth to note that we deliberately design this regression model without a constant term. This is because when no match exists, $R(q, m)$ should be 0. Therefore, the constant term can be considered as 0, and we force the model to pass through the origin of the axis.

To estimate the coefficients, we designed an experiment with test queries. In this experiment, 7 participants were invited to evaluate the search results for 10 different queries. For each query, we provide a search phrase (e.g., "California population density") and 10 maps. These 10 maps were manually selected to match the following cases: a combination with both thematic and geographic matches, only geographic matches, only thematic matches, and no match at all. Each participant was asked to compare the maps with their corresponding queries, and rank the degree of matching from 0 (not matching at all) to 5 (perfect matching). In total, we have collected 700 data records, and combined the results with the 900 matching feature scores (9 scores for each of the 100 maps). It is worth mentioning that a detailed study on user preferences is out of scope here and the topic has been extensively studied in the search and information retrieval literature. Here, we are only interested in estimating the relative importance of each matching feature to ensure cognitively meaningful results.

Based on the data from the human-participant experiment, we derive values for the coefficients in the regression model. To evaluate the necessity of including the thematic-geo interaction variable, we construct two regression models and will evaluate them respectively.

4.4 Evaluation

We evaluate the regression-based ranking model using the two criteria, namely the correctness of the ranking result and the convenience of being embedded into a SPARQL query.

To evaluate the correctness of ranking, we compare the ranking scores computed by the regression model with the average scores from human judgments. Two statistics, root-mean-square error (RMSE) and Pearson's correlation coefficient, have been used to quantify the closeness between the two. Figure 2 shows the comparison results.

The solid lines in the above two figures represent the reference line, and indicate the perfect consistence between the estimated ranking score and human judgments. The dotted lines represent the actual relation between the estimated and ground truth ranking. As can be seen, including the interaction variable brings higher correlation coefficient as well as lower RMSE. The dotted line in Fig. 2b is also closer to the reference line than that in Fig. 2a. The correlation coefficient in Fig. 2b is 0.7746 ($p < 0.01$) which indicates a high consistence between the estimated ranking and the average human judgments.

This regression-based ranking model can also be integrated into a Linked-Data-driven geoportal without much difficulty. To demonstrate this, we implement this

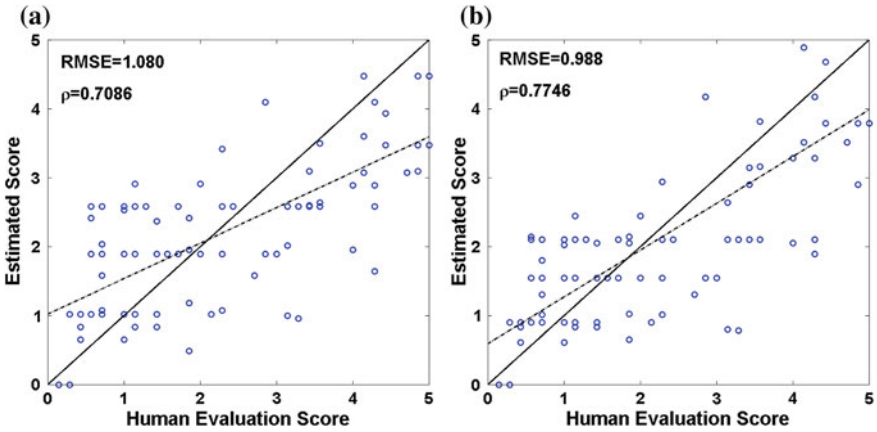

Fig. 2 Comparing estimated ranking scores with human judgments **a** without the interaction variable, **b** with the interaction variable

ranking model as a single SPARQL query (shown in Listing 1). Such a SPARQL query can be directly embedded into a system's existing search module without having to change other parts of the system.

```
SELECT ?item (COUNT(?titleThematicExact) AS ?TTE)
(COUNT(?titleThematicSimilar) AS ?TTS)
(COUNT(?descThematicExact) as ?DTE)
(COUNT(?descThematicSimilar) as ?DTS)
(COUNT(?titleGeoExact) as ?TGE)
(COUNT(?titleGeoSimilar) as ?TGS)
(COUNT(?descGeoExact) as ?DGE)
(COUNT(?descGeoSimilar) as ?DGS)
(((?TTE+?TTS+?DTE+?DTS)*(?TGE+?TGS+?DGE+?DGS)) as ?TGI)
(( λ1*?TTE + λ2*?TTS + λ3*?TGE + λ4*?TGS + λ5*?STE + λ6*?STS +
λ7*?SGE + λ8*?SGS + + λ9*?TGI) as ?ranking)
WHERE { OPTIONAL {
    ?item :hasTitleThematicTerm ?titleThematicExact .
    FILTER ( ?titleThematicKey = :exactThematicTerm ) }
  OPTIONAL {
    ?item :hasTitleThematicTerm ?titleThematicSimilar .
    FILTER ( ?titleThematicSimilar = :expandedThematicTerm ) }
  OPTIONAL {
    ?item :hasDescThematicTerm ?descThematicExact .
    FILTER ( ?descThematicExact = :exactThematicTerm ) }
  OPTIONAL {
    ?item :hasDescThematicTerm ?descThematicSimilar .
    FILTER ( ?descThematicSimilar = :expandedThematicTerm ) }
  OPTIONAL {
    ?item :hasTitleGeoTerm ?titleGeoExact .
    FILTER ( ?titleGeoExact = :exactGeoTerm ) }
```

```
        OPTIONAL {
         ?item :hasTitleGeoTerm ?titleGeoSimilar .
         FILTER ( ?titleGeoSimilar = :expandedGeoTerm ) }
        OPTIONAL {
         ?item :hasDescGeoTerm ?descGeoExact .
         FILTER ( ?descGeoExact = :exactGeoTerm ) }
        OPTIONAL {
         ?item :hasDescGeoTerm ?descGeoSimilar .
         FILTER ( ?descGeoSimilar = :expandedGeoTerm ) }
      } GROUP BY ?item ORDER BY Desc(?ranking) LIMIT 200
```

Listing 1: SPARQL query for estimating the relavance of resources and result ranking.

5 Flexible Queries for Knowledge Discovery

The existing Web API of ArcGIS Online only allows users to submit queries satisfying pre-designed templates. As a sample of ArcGIS Online metadata has been converted into Linked Data, they automatically support flexible queries without requiring additional hard coding in the Web API. In this section, we employ two scenarios to demonstrate these user-defined queries that can be performed, as well as some interesting knowledge that can be discovered. While only two queries are shown in this paper, we also provide more than 20 additional examples in the *knowledge discovery* menu of the implemented ArcGIS Online Linked Data Web interface (see Sect. 6).

5.1 Which Basemaps Are Popular?

ArcGIS Online allows users to browse the available basemaps, and records the number of times that each basemap has been **viewed** by users. Based on this number, one might assume that the most popular basemap is the one that has been viewed by most people. However, given the newly interlinked metadata, we can also count the times that each basemap has actually been **used**. In this scenario, we compare the results based on these two definitions of *popularity*. The below SPARQL query returns the result based on the times of views:

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

The result of the above query indicates that the *World Boundaries and Places* map has been **viewed** most frequently. However, making the number of usages as the criterion for popularity may lead to a different ranking. Below is the corresponding SPARQL query:

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

Interestingly, the result indicates that the *World Topographic Map* is the one that have been **used** most times in other maps. In fact, it has been used 13,507 times which is significantly more than the usage of the *World Boundaries and Places* map (2855 times).

5.2 Which Group Has Created Most Maps About California?

In this scenario, we demonstrate the additional capabilities of GeoSPARQL, an OGC standard language for querying geographic RDF data. It allows users to query and summarize data based on not only non-spatial attributes but also geographic extents. As an example, we search for the user group that has created the highest number of maps about California.

```
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))  }
GROUP BY ?group ORDER BY DESC(?itemCount) LIMIT 100
```

In the above query, we first define a polygon element to approximate the boundary of California. We then use this polygon as the extent limit for a geographic filter based on the topological relation *within*. ArcGIS Online items that fall within the boundary of California will be counted for each group, and the query returns the top 100 groups that have created maps about California. The result shows that the No.1 group is a Web GIS class from the University of California, Riverside.

6 Implementation

A prototypical Linked-Data-driven Web portal for ArcGIS Online has been implemented using the presented methods, and can be accessed via http://stko-exp.geog.ucsb.edu/linkedarcgis/. Based on the semantically annotated and enriched Linked Data, the portal provides the following capabilities:

- **Semantic Search**. This function enables the search of maps based on the enriched semantic interpretation of queries. Figure 3 shows an example of searching for *natural disasters in Utah*, in which the system returns maps about flood, tornado and other disasters. To increase the clarity of the search results, we use three columns to separately show maps that have both thematic and geographic matches, only thematic matches, and only geographic matches.
- **Knowledge Discovery**. This module shows additional examples for flexible queries automatically enabled by the Linked Data. We design a simplified user interface which deliberately hide the technique details, but interested users can click the *SPARQL* button to check the SPARQL statements used behind the scene.
- **GeoSPARQL**. This module demonstrates the capability of OGC's GeoSPARQL in supporting geographic queries on Linked Data. Users can search maps by inputting a thematic term (e.g., fire), and specifying a geographic area (e.g., California). While results will be shown as thumbnails, a map will also be shown at the bottom of the page to indicate the geographic extents of the search results.
- **Statistics**. This module gives a numeric summary of the exported and converted ArcGIS Online sample data.

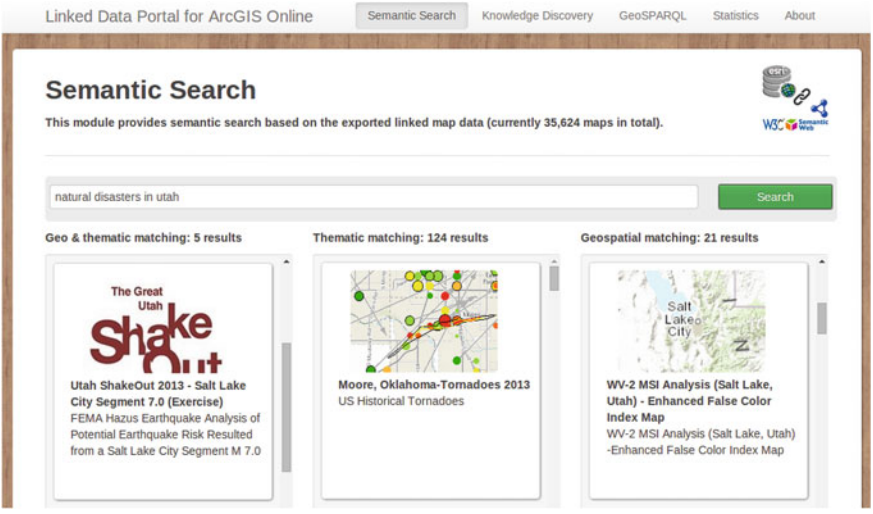

Fig. 3 A screenshot for the prototypical linked-data-driven web portal

7 Conclusions and Future Work

With the fast growth of online geoportals and the wide availability of GIS resources, there is an increasing demand for intelligent and flexible search to help users efficiently find data. Our work is an effort towards this direction. Based on ArcGIS Online, a large geoportal and online cloud service, we present a workflow for converting the metadata using the Linked Data principles and enriching them with machine-readable terms extracted from titles and descriptions. We design a semantic search function for Linked Data by expanding users' input queries and tuning a linear regression model with human participants. An evaluation experiment has been conducted to assess both the correctness and the usability of the presented semantic search function. As a sample of metadata from ArcGIS Online has been converted into Linked Data, they automatically support flexible queries without requiring pre-designing and hard-coding in a Web API. We use two scenarios to demonstrate the flexible queries that can be submitted to discover new knowledge from the data. An online prototype has been implemented using the presented methods as a proof-of-concept.

While we have taken a Linked-Data-driven approach in this work, it is worth to clarify that some techniques used in our workflow, such as query expansion and entity extraction, do not necessarily require a Linked Data approach. However, Linked Data automatically enables flexible and customized queries which significantly expand the searching capability that a GIS user can have. Thus, we consider Linked Data as an indispensable cornerstone in our solution. This research also has several limitations. For example, the coefficients of the regression model for semantic search are derived from a small number of participants. While the evaluation shows a fair performance in the search results, a larger scale human participant test is nevertheless necessary to further calibrate the model. In addition, since external services, such as *DBpedia Spotlight*, have been used to expand the queries in real time, the response time of the semantic search varies. While the online system is still a prototype, the search speed could be improved by integrating those external services as part of the entire system. Finally, so far we have extracted thematic concepts and geographic entities from the map titles and descriptions, and a next-step research could focus on inferring topic categories (e.g., whether this map is about *transportation* or *agriculture*) from the textual descriptions, thereby further enriching the voluntary metadata.

Acknowledgments This work is a collaborative effort from UCSB STKO Lab and ESRI Applications Prototype Lab. The authors would like to thank Jack Dangermond, Hugh Keegan, Dawn Wright, as well as the three anonymous reviewers for their constructive comments and feedbacks.

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In: *The semantic web* (pp. 722–735). Berlin, Springer.

Battle, R., & Kolas, D. (2012). Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web, 3*(4), 355–370.

Berners-Lee, T., Hendler, J., Lassila, O. (2001) *The semantic web* (pp. 29–37). Scientific American (2001).

Bizer, C., Heath, T., & Berners-Lee, T. (2009a). Linked data—the story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., & Cyganiak, R. (2009b). DBpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web, 7*(3), 154–165.

Butuc, M. G. (2009). Semantically enriching content using opencalais. *EDITIA, 9*, 77–88.

Dangermond, J. (2009). *GIS: Design and evolving technology*. ESRI, Fall: ArcNews.

Goodchild, M. F., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth, 3*(3), 231–241.

Guha, R., McCool, R., Miller, E. (2003) Semantic search. In: *Proceedings of the 12th International Conference on World Wide Web*, (pp. 700–709). ACM.

Han, L., Kashyap, A., Finin, T., Mayfield, J., Weese, J. (2013) *UMBC ebiquity-core: Semantic textual similarity systems* (p. 44). Atlanta, Georgia, USA.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology, 1*(1), 1–136.

Hitzler, P., Krotzsch, M., Rudolph, S. (2011). *Foundations of semantic web technologies*. Boca Raton, CRC Press.

Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., Hitzler, P. (2013). A linked-data-driven and semantically-enabled journal portal for scientometrics. In: *The semantic web–ISWC* 2013 (pp. 114–129). Berlin, Springer.

Janowicz, K., Scheider, S., Pehle, T., & Hart, G. (2012). Geospatial semantics and linked spatiotemporal data–past, present, and future. *Semantic Web, 3*(4), 321–332.

Jones, C.B., Alani, H., Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In: *Spatial information theory*, (pp. 322–335). Berlin, Springer.

Keßler, C., Janowicz, K., Kauppinen, T.: spatial@ linkedscience–Exploring the research field of GIScience with linked data. In: *Geographic Information Science* (pp. 102–115). Berlin, Springer (2012).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Mendes, P.N., Jakob, M., Garca-Silva, A., Bizer, C. (2011). DBpedia spotlight: Shedding light on the web of documents. In: *Proceedings of the 7th International Conference on Semantic Systems* (pp. 1–8). ACM.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Communications of the ACM, 38* (11), 39–41.

Tran, T., Cimiano, P., Rudolph, S., Studer, R. (2007). Ontology-based interpretation of keywords for semantic search. In: *The Semantic Web* (pp. 523–536). Berlin, Springer.

Zhou, Q., Wang, C., Xiong, M., Wang, H., Yu, Y. (2007) Spark: Adapting keyword query to semantic search. In: *The Semantic Web* (pp. 694–707). Berlin, Springer.