

# Semantically-Enriched Search Engine for Geoportals: A Case Study with ArcGIS Online

Gengchen Mai, Krzysztof Janowicz, Sathya Prasad, Meilin Shi, Ling Cai, Rui Zhu, Blake Regalia, and Ni Lao

**Abstract** Many geoportals such as ArcGIS Online are established with the goal of improving geospatial data reusability and achieving intelligent knowledge discovery. However, according to previous research, most of the existing geoportals adopt Lucene-based techniques to achieve their core search functionality, which has a limited ability to capture the user's search intentions. To better understand a user's search intention, query expansion can be used to enrich the user's query by adding semantically similar terms. In the context of geoportals and geographic information retrieval, we advocate the idea of semantically enriching a user's query from both geospatial and thematic perspectives. In the geospatial aspect, we propose to enrich a query by using both place paratomy and distance decay. In terms of the thematic aspect, concept expansion and embedding-based document similarity are used to infer the implicit information hidden in a user's query. This semantic query expansion

---

Gengchen Mai  
STKO Lab, UC Santa Barbara  
e-mail: gengchen\_mai@geog.ucsb.edu

Krzysztof Janowicz  
STKO Lab, UC Santa Barbara  
e-mail: jano@geog.ucsb.edu

Sathya Prasad  
ESRI Inc, Redlands, CA, USA  
e-mail: sprasad@esri.com

Meilin Shi  
STKO Lab, UC Santa Barbara  
e-mail: meilinshi@ucsb.edu

Ling Cai  
STKO Lab, UC Santa Barbara  
e-mail: lingcai@ucsb.edu

Rui Zhu  
STKO Lab, UC Santa Barbara  
e-mail: ruizhu@geog.ucsb.edu

Blake Regalia  
STKO Lab, UC Santa Barbara  
e-mail: blake.regalia@gmail.com

Ni Lao  
SayMosaic Inc.  
e-mail: noon99@gmail.com

framework is implemented as a semantically-enriched search engine using ArcGIS Online as a case study. A benchmark dataset is constructed to evaluate the proposed framework. Our evaluation results show that the proposed semantic query expansion framework is very effective in capturing a user's search intention and significantly outperforms a well-established baseline – Lucene's practical scoring function – with more than 3.0 increments in DCG@K (K=3,5,10).

**Keywords:** Query Expansion · ArcGIS Online · Semantically Enriched Search Engine · Geoportal · Geographic Information Retrieval

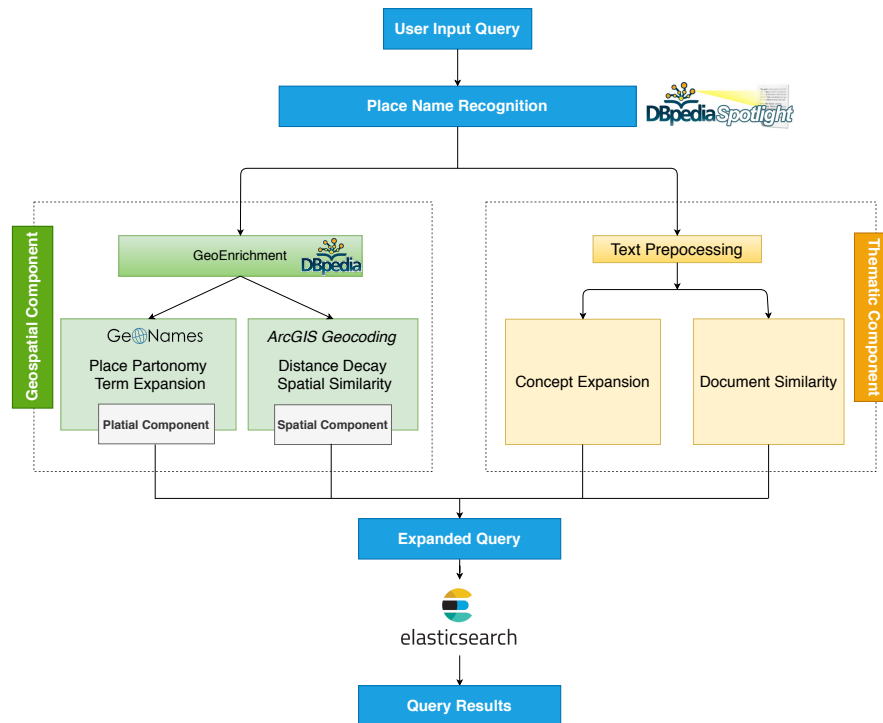


Fig. 1: The semantic query expansion framework

## 1 Introduction

The increasing growth of geospatial data poses a great challenge to data discovery, access, and maintenance (Jiang et al. 2018). In order to increase data reusability and facilitate geospatial knowledge discovery, many geoportals have been established

to provide integrated access to geospatial resources (Hu et al. 2015a). Examples of geoportals include the DataOne Data Catalog<sup>1</sup>, U.S. Geological Survey Science Data Catalog<sup>2</sup>, NASA Earth data Search<sup>3</sup>, ArcGIS Online, and so on.

The most important component of a geoportal is its search functionality, which is usually supported by geographic information retrieval (GIR) techniques. Generally speaking, information retrieval (IR) aims at finding relevant entries based on a user’s query. The entries can be documents, websites, services, maps, and so on, depending on the application scenarios. As a subfield of IR, geographic information retrieval (Jones & Purves 2008) adds space (and time) as additional dimensions to the traditional information retrieval problems (Janowicz et al. 2011). In addition to traditional thematic similarity, spatial (and temporal) similarity is considered when the relevance score between a user’s query  $q$  and an entry  $d$  is calculated.

Despite the success of GIR in academia, in practice, the core search functionality of most existing geoportals is still based on Apache Lucene or Elasticsearch (Jiang et al. 2018). These Lucene-based engines use a term frequency-inverse document frequency (TF-IDF) approach to compute the similarities between a user’s query and document entries, which is insufficient to completely capture a user’s search intention. For example, when a user searches for *natural disaster in California* (Query  $q_1$ ), (s)he is probably more interested in a document which describes the Kincade Fire that burned in Sonoma County on Oct. 23rd, 2019 since wild fires are a type of natural disaster and Sonoma County is a subdivision of California. However, if this document contains neither the term “natural disaster” nor “California”, a Lucene-based model will give a zero relevance score between this document and the Query  $q_1$ , thus resulting in a low recall. This highlights the necessity of understanding the user’s search intentions both semantically and spatially in a (G)IR system.

According to Dominich (2008), IR can be formally defined as:

$$IR = m[\mathcal{R}(D, (q, \langle I, \mapsto \rangle))] \quad (1)$$

where  $m$  is the degree of relevance;  $\mathcal{R}$  is the relevance relationship;  $D$  is a set of (document) entries;  $q$  is the user’s query;  $I$  and  $\mapsto$  are implicit and inferred information. The most challenging part in this equation is the question of how to obtain the implicit and inferred information  $I, \mapsto$  based on user queries. *Query expansion* techniques, which add terms and conditions to a user query with the goal of improving the query-object relevance score (Vechtomova 2009), can be utilized to semantically take the user’s search intention into account.

The traditional query expansion focuses on semantically-enriching a user’s query from a thematic perspective. In the context of geoportals (e.g., ArcGIS Online) we argue that a user’s query should be expanded (or semantically-enriched) from two perspectives: thematic and geospatial. In the thematic aspect, a query can be enriched/expanded by adding thematically similar concepts/terms. For example, as for Query  $q_1$ , some highly related topics of “natural disaster” such as *earthquake, wild*

<sup>1</sup> <https://search.dataone.org/data>

<sup>2</sup> <https://data.usgs.govdatacatalog>

<sup>3</sup> <https://search.earthdata.nasa.gov/search>

*fire*, *flood*, and *hurricane* can be added to the original query. In a geoportal, extra attention should be paid to the geospatial aspect. Geospatially related terms can be added to the query. For example, as for Query  $q_1$ , we can consider adding the names of the subdivisions of California to the query. Since this process relies on the place hierarchy, we call it *patial* query expansion. Moreover, the spatial scopes of the query and entries can also be used to compute the spatial similarity between them. After being enriched/expanded from these two perspectives, the new query is applied to the geoportal in the hope of improving the recall of the GIR system.

Note that the core idea of query expansion is to minimize the mismatch between a user query and candidate entries so that the recall of the IR system is improved. A similar idea can be applied when we calculate spatial similarities between a user's query and entries. Most of the traditional spatial similarity measures are based on topological relations between the spatial scopes of the user's query and an entry. For example, Jiang et al. (2018) defined the spatial similarity between a query  $q$  and a document entry  $d$ , denoted as  $Sim(q, d)$ , based on their geographic scopes  $Area(q)$ ,  $Area(d)$  as well as their intersection  $Area(q \cap d)$  (See Equation 2).

$$Sim(q, d) = \left( \frac{Area(q \cap d)}{Area(q)} + \frac{Area(q \cap d)}{Area(d)} \right) \times 0.5 \quad (2)$$

According to Equation 2, if  $Area(q \cap d) = 0$ , then  $Sim(q, d) = 0$  which means if the intersection of the geographic footprints of  $q$  and  $d$  is zero, the spatial similarity score is zero. This may lead to a loss of valuable spatial proximity information in many scenarios. To give a concrete example, if a user searches for *Weather in Los Angeles* (Query  $q_2$ ), a map  $d_1$  about *Temperature in Oxnard* should be considered more relevant than, say,  $d_2$  which is about *Temperature in Southern Africa*. However, since the both geographic scopes of Oxnard and Southern Africa do not intersect with the footprint of Los Angeles ( $Area(q_2 \cap d_1) = 0$  and  $Area(q_2 \cap d_2) = 0$ ), we will have  $Sim(q_2, d_1) = 0$  and  $Sim(q_2, d_2) = 0$  according to Equation 2 which does not match our intuition.

In other words, it might be better to utilize a distance decay function here instead and minimize the mismatch between the current query  $q_2$  and  $d_1$ . Inspired by this observation, we utilize a Gaussain kernel distance decay function to compute the spatial similarity between the spatial scopes/geographic footprints between the query and documents. Using a distance decay function to optimize the query-document relevance is also related to work on query relaxation in the context of geographic question answering (Mai et al. 2019).

**The research contributions of this work are as follows:**

1. We propose a semantic query expansion framework for geoportals which enriches a user's query from both thematic and geospatial aspects.
2. We develop a semantically-enriched search engine prototype for ArcGIS Online by implementing the proposed query expansion framework.
3. We collect a benchmark dataset to evaluate the presented framework against a widely used baseline model - Lucene's practical scoring function. The evalua-

tion results show that our semantic query expansion framework outperforms the baseline by a significant margin.

The remainder of this work is structured as follows. In Sec. 2, several work about geographic information retrieval are discussed. Next, we present our query expansion framework and describe each component of this system in Sec. 3. Particularly in Sec. 3.1 we discuss about the reproducibility of our work and provide guidelines related to data sets and software that facilitate future research along this line. In Sec. 4, we introduce a benchmark dataset we collect to evaluate our GIR framework and then discuss the evaluation results. Finally in Sec. 5 we conclude our work and discuss the future research directions.

## 2 Related Work

The idea of query expansion is to reformulate a user's query by adding semantically related concepts (Azad & Deepak 2019) to minimize the query-object mismatch and increase the recall of an IR system. This typically comes at the expense of reducing the precision. Generally speaking, query expansion techniques can be classified into two categories: global analysis and local analysis (Azad & Deepak 2019). As for global analysis, the expansion terms are selected based on manually built knowledge bases, knowledge graphs, or large corpora. Finding semantically related terms based on word embedding (Mikolov et al. 2013, Mai et al. 2018) or topic modeling (Hu et al. 2015b) is an example. Local analysis refers to query expansion methods that select expansion terms based on the retrieved documents of the initial user's query. Example models include relevance feedback (Rocchio 1971) and pseudo-relevance feedback (Buckley et al. 1995). In this work, we adopt the global analysis method and use word embedding to select semantically related terms of query terms.

Many query expansion techniques are not directly applicable for geospatial terms. For example, it is more reasonable to select geospatially related terms based on place hierarchies (e.g., from a digital gazetteer) rather than using word embedding models. This suggests a need for separately handling geospatial aspect in a query expansion task. For instance, Huang et al. (2008) classified queries into two types - location sensitive and location non-sensitive - and then handled them by using different query expansion techniques.

In the field of geographic information retrieval, there are a few works aiming at ranking documents based on both textual and spatial relevance such as the multi-dimensional scattered ranking method proposed by Van Kreveld et al. (2005). Our work follows a similar research direction but also add *patial* similarity to the ranking algorithm.

In addition to query expansion, another line of work for building a semantically-enriched search engine for geoportals is to enrich the metadata. For example, Hu et al. (2015a) converted the metadata of ArcGIS Online items into Linked Data and then enriched the metadata to enable semantic search. Similar to our idea, Hu et al. (2015a) also considered the semantic enrichment in two aspects: thematic and geospatial. However, converting data into another format for semantic enrich-

ment requires additional processing steps, storage, and maintenance to keep both data sources in sync. In this work, we focus on enabling semantic search by using query expansion techniques in which the underlying data storage (e.g., Elasticsearch, Apache Lucene) remains unchanged.

### 3 Method

In this section, we will first describe the dataset and project setup in Section 3.1. Next, we describe our semantic query expansion framework in detail. The proposed framework is composed of two major components - geospatial component and thematic component - which focus on different aspects. Figure 1 shows the overall architecture of the proposed framework. We will present each component below with the example query *Chicago traffic* (Query  $q_3$ ).

#### 3.1 Data and Software Availability

Developed by Environment System Research Institute (ESRI), ArcGIS Online is one of the best-known web geoportals. It contains a collection of web maps, data layers, tools, services, and applications contributed from different GIS users all over the world (Hu et al. 2015a). Elasticsearch<sup>4</sup>, a widely used search and analytic engine, is utilized to store the metadata of these ArcGIS Online items and support the portals searching functionality. The metadata of each ArcGIS Online item has different fields such as “id”, “title”, “snippet”, “description”, “type”, “location” (point), “coordinates” (the bounding box) and so on. The core search functionality of ArcGIS Online is based on Lucene’s query-document similarity function which is computed based on term frequency and inverse document frequency (TF-IDF) scoring such as Lucene’s practical scoring function<sup>5</sup>, Okapi BM25, and so on. Therefore, Lucene’s practical scoring function is a natural baseline for our semantic query expansion framework.

In order to establish an evaluation dataset for our search engine prototype, we collect 53,404 items using the ArcGIS Online RESTful API which contains 1) all items published by Esri or its related organizations before September 2017; 2) all items published on ArcGIS Online between June and September in 2014 and 2017.

We use Elasticsearch to host all the retrieved ArcGIS Online items. The proposed semantic query expansion framework will serve as a middle layer as shown in Figure 1 to semantically-enrich the current user query. The expanded query will be sent to the established Elasticsearch index to get relevant ArcGIS Online items. The motivation here is to enable semantic search functionality on top of a portal such as ArcGIS Online without changing the underlying layers, e.g., data storage. In order to evaluate the proposed semantic query expansion framework and compare it with the baseline, namely Lucene’s practical scoring function, we also conduct a human

---

<sup>4</sup> <https://www.elastic.co/>

<sup>5</sup> <https://www.elastic.co/guide/en/elasticsearch/guide/2.x/practical-scoring-function.html>

participant test to get query-document relevance scores through Amazon Mechanical Turk sandbox<sup>6</sup>. Detail description about this benchmark dataset can be found in Section 4.2. The data and source code are available at <sup>7</sup> including 1) the evaluation benchmark dataset; 2) the source code of our query expansion framework. The established database is hosted by Elasticsearch 5.4.0<sup>8</sup> with a vector scoring plugin<sup>9</sup> to enable word embedding computation.

### 3.2 Query Preprocessing: Place Name Recognition

```

select ?place ?lat ?long ?label ?area ?geoid {
OPTIONAL {
  ?place geo:lat ?lat.
  ?place geo:long ?long.
}
OPTIONAL {
  ?place rdfs:label ?label.
  FILTER(lang(?label) = "en")
}
OPTIONAL {
  ?place <http://dbpedia.org/ontology/PopulatedPlace/
*2      ↪ areaTotal> ?area.
}
OPTIONAL {
  ?place owl:sameAs ?geoid .
  FILTER(CONTAINS(str(?geoid), 'geonames'))
}
VALUES ?place {
  <http://dbpedia.org/resource/Chicago>
}
}

```

Listing 1: An example of DBpedia SPARQL query generated in the GeoEnrichment step to get more information about the identified places.

Given a query such as *Chicago traffic*, we need to first split it into a geospatial aspect and a thematic aspect. A place name recognition service (e.g., DBpedia Spotlight<sup>10</sup>) is utilized to recognize the toponyms appearing in the query (in this case the

<sup>6</sup> <https://www.mturk.com/>

<sup>7</sup> <https://github.com/gengchenmai/arcgis-online-search-engine>

<sup>8</sup> <https://www.elastic.co/blog/elasticsearch-5-4-0-released>

<sup>9</sup> <https://github.com/MLnick/elasticsearch-vector-scoring>

<sup>10</sup> <https://www.dbpedia-spotlight.org/>

city of Chicago) and then link it to the corresponding entities (`dbo:Chicago`) in a knowledge graph such as Wikidata or DBpedia. The identified places are then handled by the geospatial query expansion component and the rest of the query is sent to the thematic query expansion component.

### 3.3 Geospatial Query Expansion Component

The geospatial query expansion component focuses on improving the *platial* and *spatial* similarity between a user’s query and a candidate ArcGIS Online item.

In order to facilitate the following query expansion process, we first enrich the identified geographic entities with additional information such as geographic coordinates, place names, total area, and their GeoNames identifier (See Listing 1). We call this GeoEnrichment step (See Figure 1).

#### 3.3.1 Platial Component

The platial component focuses on finding similar geographic terms based on the place hierarchy. We use the GeoNames<sup>11</sup> service to get the top  $K$  subdivisions of the identified places. For example, we can add *Belmont Cragin* and *Englewood* as expanded geographic terms to the expanded query of Query  $q_3$ . Here, the platial similarity between a query  $q$  and an ArcGIS item  $d_o$ , denoted as  $Sim_{platial}(q, d_o)$ , is defined as

$$Sim_{platial}(q, d_o) = \sum_{p_i \text{ in } q} W_{geo}(p_i) \sum_{p_{ij} \in Q_{platial}(p_i) \cup \{p_i\}} W_{platial}(p_i, p_{ij}) \sum_{f_k \text{ in } d_o} W_f(f_k) M(p_{ij}, f_k) \quad (3)$$

Here  $p_i$  refers to the  $i$ th identified place from  $q$ ;  $W_{geo}(p_i)$  is the relative importance of place  $p_i$  among all the identified places and  $\sum_{p_i \text{ in } q} W_{geo}(p_i) = 1$ ;  $Q_{platial}(p_i)$  refers to the set of expanded geographic terms;  $W_{platial}(p_i, p_{ij})$  indicates the importance of  $p_{ij} \in Q_{platial}(p_i) \cup \{p_i\}$  with respect to the corresponding place  $p_i$ ;  $W_f(f_k)$  indicates the weight of matching one specific metadata field  $f_k$  since matching some fields such as “title” is much more important than matching other fields such as “description” and  $\sum_{f_k \text{ in } d_o} W_f(f_k) = 1$ ;  $M(p_{ij}, f_k)$  indicates the number of matches of the expanded geographic term  $p_{ij}$  in the current field  $f_k$ .

#### 3.3.2 Spatial Component

The spatial component measures the spatial similarity between a query  $q$  and item  $d_o$ . Frontiera et al. (2008) discussed different geometric approaches to accessing spatial similarity and most of them are computed based on the topological relationships between the geographic scopes of query  $q$  and item  $d_o$ . An example of similarity

<sup>11</sup> <https://www.geonames.org/>



measures is Jaccard similarity index (Jaccard 1912). Some non-topological relation based spatial similarity indices also exist such as Hausdorff Distance.

In this work, we use a distance decay approach with Gaussian kernels. Each identified place has a Gaussian kernel which is placed at the center of its bounding box. The bandwidth of a kernel is determined based on the bounding box of the corresponding place. The intuition comes from Tobler’s First Law of Geography: the relatedness between query  $q$  and item  $d_o$  decreases with respect to their distance. Here ArcGIS Geocoding API is utilized to obtain the bounding boxes of the identified places. The spatial similarity  $Sim_{spatial}(q, d_o)$  is defined in Equation 4 where  $Gauss(p_i, d_o)$  is the Gaussian score between identified place  $p_i$  and item  $d_o$ . The impact of different spatial similarity measures on the performance of this semantic query expansion framework will be left for future work.

$$Sim_{spatial}(q, d_o) = \sum_{p_i \text{ in } q} W_{geo}(p_i) Gauss(p_i, d_o) \quad (4)$$

### 3.4 Thematic Query Expansion Component

As the name indicates, thematic query expansion focuses on minimizing the query-item mismatch from a thematic, i.e., topic-based, point of view. To achieve this, we adopt two approaches: concept expansion and embedding-based document similarity. We will discuss each of them below.

Before performing thematic query expansion, some text preprocessing steps such as tokenization, word lemmatization, and stop word removal have been taken to extract thematic concepts/terms from the user’s query such as *natural*, *disaster* in Query  $q_1$  and *traffic* in Query  $q_3$ .

#### 3.4.1 Concept Expansion Component

The idea of concept expansion is to find thematically similar terms to the query terms and add them to the expanded query clause. This is a common way to do query expansion (Jiang et al. 2018, Hu et al. 2015b). Unlike the previous work in GIR which use semantic knowledge base (Jiang et al. 2018) or topic modeling (Hu et al. 2015b) to find thematically similar terms, we use word embedding technique (Mikolov et al. 2013) to achieve this. A similar approach has been used in developing academic search engine (Mai et al. 2018). Given the term *traffic*, word embedding model finds thematically similar terms such as *congestion*, *rail*, *train*, *roads*, and so on.

Equation 5 shows the thematic similarity between  $q$  and  $d_o$  based on concept expansion  $Sim_{concept}(q, d_o)$ . Here,  $t_i$  indicates a thematic term in the user’s query such as *traffic*.  $W_{thematic}(t_i)$  means the normalized weight of  $t_i$  among all thematic query terms and  $\sum_{t_i \text{ in } q} W_{thematic}(t_i) = 1$ .  $T_{w2v}(t_i)$  indicates the set of thematically similar terms of  $t_i$  based on a pretrained word embedding model such as GloVe

(Pennington et al. 2014) and  $W_{w2v}(t_i, t_{ij}) = \frac{\text{cosine}(t_i, t_{ij})}{\sum_{t_{ix} \in T_{w2v}(t_i) \cup \{t_i\}} \text{cosine}(t_i, t_{ix})}$  indicates normalized weight of term  $t_{ij}$  with respect to  $t_i$  based on their cosine similarity.  $M(t_{ij}, f_k)$  refers to the number of matches of the expanded thematic term  $t_{ij}$  in the current field  $f_k$ .

$$Sim_{concept}(q, d_o) = \sum_{t_i \text{ in } q} W_{thematic}(t_i) \sum_{t_{ij} \in T_{w2v}(t_i) \cup \{t_i\}} W_{w2v}(t_i, t_{ij}) \sum_{f_k \text{ in } d_o} W_f(f_k) M(t_{ij}, f_k) \quad (5)$$

### 3.4.2 Embedding-Based Document Similarity Component

Instead of explicitly matching the expanded thematic terms to ArcGIS Online items, the embedding-based document similarity compares query  $q$  and item  $d_o$  in the hidden word embedding space. Equation 6 shows how the similarity score is defined.  $E_{query}(q) = \sum_{t_i \text{ in } q} Word2Vec(t_i)$  is the embedding of query  $q$  which is computed by simply adding the word embeddings of each thematic terms in the query  $q$ .  $E_{doc}(d_o)$  is the document embedding of  $d_o$  which is computed based on TF-IDF weighted word embedding of each terms in its title, snippet, and description.

$$Sim_{doc}(q, d_o) = \text{cosine}(E_{query}(q), E_{doc}(d_o)) \quad (6)$$

### 3.5 Expanded Query Construction

The overall similarity between a query  $q$  and an ArcGIS Online item  $d_o$  is a weighted sum of all four components: platial (place-based) component, spatial component, concept expansion component, and embedding-based document similarity component.  $\lambda_{platial}$ ,  $\lambda_{spatial}$ ,  $\lambda_{concept}$ , and  $\lambda_{doc}$  are their corresponding weights.

$$Sim(q, d_o) = \lambda_{platial} * Sim_{platial}(q, d_o) + \lambda_{spatial} * Sim_{spatial}(q, d_o) + \lambda_{concept} * Sim_{concept}(q, d_o) + \lambda_{doc} * Sim_{doc}(q, d_o) \quad (7)$$

In practice, each component can be written as a collection of function score query clauses in Elasticsearch. Figure 2 shows an example of Elasticsearch query constructed after the proposed semantic query expansion framework for the given *Chicago traffic* query. Each component is highlighted. Executing this expanded query in the established Elasticsearch index will give us the final search result.

```

{ "min_score": 1.2288,
  "query":
  { "function_score":
    { "query": { "match_all": {} },
      "boost": 1,
      "functions": [
        // Spatial Component
        { "gauss":
          { "location":
            { "origin": [ -87.63244999999995, 41.884250000000065 ],
              "scale": "27.463337102094016km",
              "offset": "0km",
              "decay": 0.5 } },
          "weight": 0.2 },
        // Platial Component
        { "filter": { "match": { "title": "Chicago" } }, "weight": 0.025 },
        { "filter": { "match": { "title": "albany park" } },
          "weight": 0.0025000000000000005 },
        .....
        // Concept Expansion Component
        { "filter": { "match": { "title": "traffic" } },
          "weight": 0.031847459297063896 },
        { "filter": { "match": { "title": "congestion" } },
          "weight": 0.022992405698700064 },
        .....
        // Embedding-Based Document Similarity Component
        { "script_score":
          { "script":
            { "inline": "payload_vector_score",
              "lang": "native",
              "params":
                { "field": "doc2vec",
                  "vector":
                    <QUERY EMBEDDING>,
                  "cosine": true } } },
          "weight": 0.2 },
        ],
      "score_mode": "sum",
      "boost_mode": "sum"
    }
  }
}

```

Fig. 2: An example of Elasticsearch query derived from our semantic query expansion framework.

## 4 Experiment

### 4.1 Semantically-Enriched Search Engine

Based on the presented semantic query expansion framework in Section 3, we develop a semantically-enriched search engine prototype for ArcGIS Online on top of the established Elasticsearch index. Figure 3 is a screenshot of the developed system in which the radio buttons *Semantic Search* and *Lucene* correspond to our semantic query expansion based GIR model and the baseline - Lucene's practical scoring function based IR model which we will call it Lucene baseline in the fol-

lowing. This web interface is available through here <sup>12</sup> A mobile application is also developed based on AppStudio for ArcGIS (See Figure 4) .

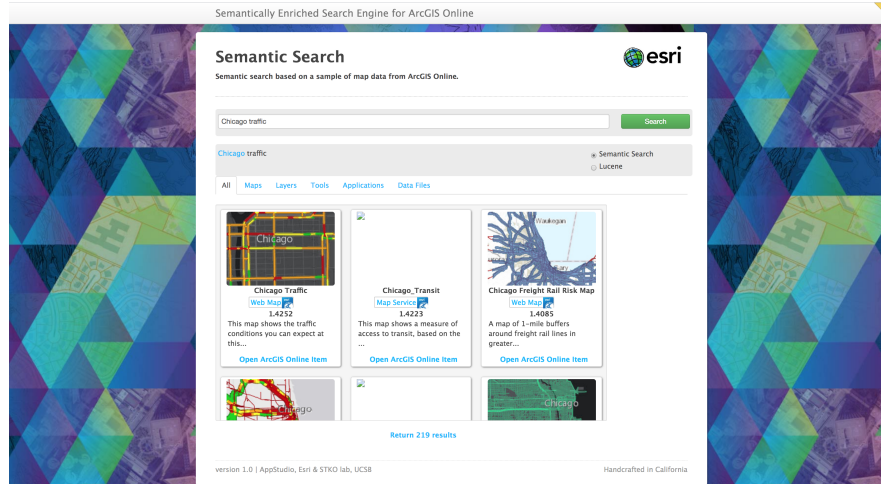


Fig. 3: A web interface for the semantically-enriched search engine prototype for ArcGIS Online.

## 4.2 Evaluation

A collection of user search logs is an ideal benchmark dataset to evaluate the presented framework as well as the Lucene baseline as Jiang et al. (2018) did. As the search logs are not available for the current project, we decide to build our own evaluation dataset. The benchmark dataset construction process can be summarized as follows:

1. We collect a query set which consists of 20 queries. All queries can be seen in Table 1. The first 10 queries are obtained from Hu et al. (2015b), while we manually generate another 10 queries based on the topics and geographic coverage of the collected ArcGIS Online items.
2. For each query, we get the top 10 search results from our semantic query expansion model as well as the Lucene baseline.
3. We create a survey form for each query and each model. Each survey form consists of one query and 10 random ordered ArcGIS Online items. Users are then asked to judge the relevance between the query and each item on an ordinal scale,

<sup>12</sup> <http://stko-testing.geog.ucsb.edu:3010/>

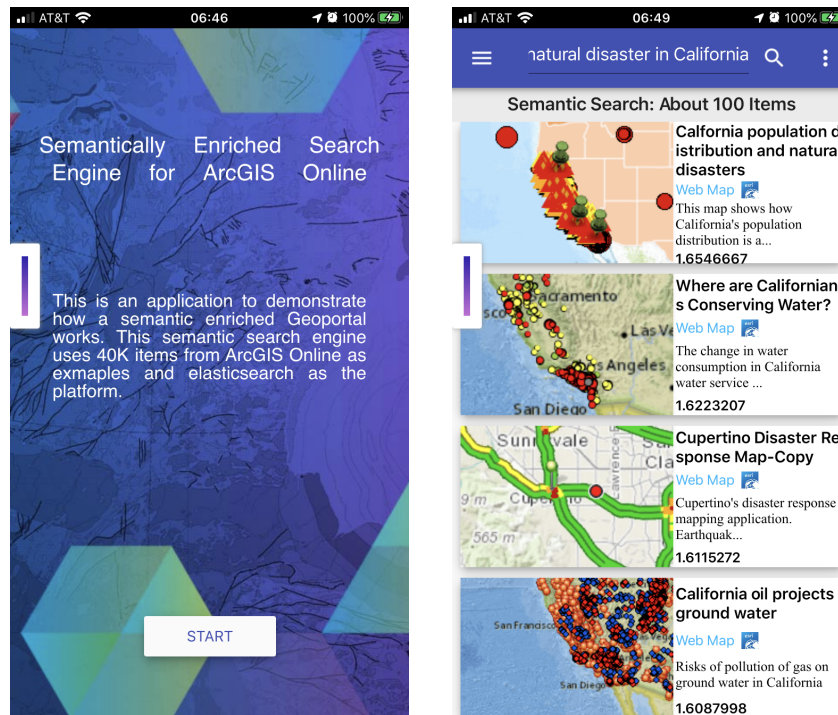


Fig. 4: A mobile application for the semantically-enriched search engine prototype for ArcGIS Online.

with labels such as “Perfect” (4), “Good” (3), “Some Relevance” (2), “Fair” (1), and “Bad” (0). The numbers in () are used as the corresponding relevance score. An example survey form can be seen in Figure 5.

- To host these surveys, a crowd-facing Web interface is developed and deployed on Amazon Mechanical Turk sandbox environment.
- Eight users completed these surveys who are from different departments of a US university.

In total, we have 40 survey forms, 20 for each GIR model, completed by 8 different accessors. The average relevance score among these 8 accessors’ results is treated as the relevance score  $rel$  between a query and an item in one form.

Discounted Cumulative Gain at top K rank (DCG@K) (Carterette & Jones 2008, Järvelin & Kekäläinen 2002) is a typical evaluation metric for information retrieval system. DCG is the weighted sum of “gains” of presenting a specific item. The weight is a *discounted* factor by ranking an item at a particular position. For IR systems, DCG at top K rank is defined as shown in Equation 8 in which  $rel_i$  indicates the relevance score between a query and an item, the said gain, and  $\frac{1}{\log_2 i}$  is the discounted factor based on the current rank  $i$ .

**Survey Instructions** (Click to expand)

For this task, the meaning of each rating level means:

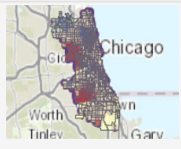
- **Perfect** (4): means the results is exactly what the user is looking for, and user will be satisfied if we only show this results to him/her.
- **Good** (3): means the results is very relevant to what the user is looking for, but it may (or may not) be exactly what the user is looking for (because either the user's intent is ambiguous, or multiple results may meet the user's need).
- **Some Relevance** (2): means the result is relevant to the query, but it is possible that the result is not what the user is looking for. However, showing it to the user is acceptable.
- **Fair** (1): means the result is not that relevant, but it is OK to show to the user if we don't have better results.
- **Bad** (0): it is really embarrassing to show this results to users. The user will question the quality of our system, or even be pissed off.

**Query: Chicago traffic**

**Item 1. [Open ArcGIS Online Item]**



**Vehicle Theft in Chicago - 2011**

**Type:** Web Map

**Tags:** Vehicle Thefts in Chicago

**Snippet:** Vehicle Thefts in Chicago

**Description:**

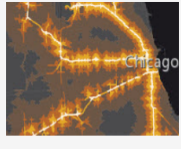
Click to see description

**Please select whether the above ArcGIS Online item is a good search result of the above query:**

Perfect  
 Good  
 Some Relevance  
 Fair  
 Bad

**Item 2. [Open ArcGIS Online Item]**



**Chicago Highway Access**

**Type:** Web Map

**Tags:** highway; access; freeway; interstate; exit; Urban.Chicago.Movement.Highway\_Access; urban; movement

**Snippet:** Areas that are within 10 minutes of an exit are emphasized on this map, to give an indication of how accessible neighborhoods are by highway.

**Description:**

Click to see description

**Please select whether the above ArcGIS Online item is a good search result of the above query:**

Perfect  
 Good  
 Some Relevance  
 Fair  
 Bad

Fig. 5: An example of the crowd-facing Web survey form we developed to collect query-item relevance scores.

$$DCG_K = rel_1 + \sum_{i=2}^K \frac{rel_i}{\log_2 i} \quad (8)$$

We choose DCG@3, DCG@5, and DCG@10 as the evaluation metrics and Table 1 shows the evaluation results of both our semantic search model and Lucene baseline on each query. Some interesting observations can be made based on Table 1:

1. By comparing the average DCG scores, our semantic search model outperforms Lucene baseline by a significant margin.
2. In 17 out of 20 queries, the semantic search model outperforms the Lucene baseline with  $\Delta DCG@K > 3$ .
3. As for the two queries (Query 2 and Query 8), the semantic search model provides relatively similar DCG scores ( $< 1$ ).
4. The only query in which our semantic search model performs clearly worse is Query 10 - *Crimes in Tennessee*. After examining the top 10 search results the two models, we find that:
  - a. All top 10 search results of Lucene baseline are crime maps about other places such as New York, Miami, or world wide crime reports. Basically Lucene baseline fetches these items based on the thematic similarity.
  - b. 9 out of 10 search results of semantic search model are about other topics in Tennessee such as public health, energy, banking while one item is about crimes in neighboring states. As for these 9 items, 7 of them do not contain any place names in their title, snippet, or description but with spatial footprints close to the center of Tennessee. This implies that semantic search model finds these items mostly based on spatial similarity.
  - c. There is actually no correct answer about the crime in Tennessee.
  - d. However, based solely on these observations we cannot conclude that people pay more attention to thematic similarity than spatial similarity. That is because this bias may be caused by the design of the survey form in which thematic similarity is relatively easy to judge, while spatial similarity is rather difficult as users need to click the link and go to the web map to see the geographic scopes of an item.
  - e. These observations raise an interesting question. How to design an appropriate survey form for evaluating GIR systems in contrast more general IR systems.

## 5 Conclusion

In this work, we present a semantic query expansion framework for geographic information retrieval systems. It enriches a user's query from both geospatial and thematic perspectives. Two components are developed for each perspective. By using ArcGIS Online as an example, we develop a semantically enriched search engine prototype by following the proposed query expansion framework. We constructed

Table 1: Evaluation results of the semantic query expansion model and baseline

Query	Lucene Baseline			Semantic Search		
	DCG@3	DCG@5	DCG@10	DCG@3	DCG@5	DCG@10
1 New York water	1.35	1.91	4.07	<b>5.90</b>	<b>8.20</b>	<b>11.78</b>
2 California fire	<b>9.12</b>	<b>11.27</b>	<b>14.81</b>	8.25	10.67	14.36
3 California population density	3.72	5.18	7.26	<b>6.97</b>	<b>9.38</b>	<b>12.88</b>
4 Vacation in Hawaii	3.85	5.05	7.93	<b>8.60</b>	<b>11.54</b>	<b>15.50</b>
5 Florida flood	8.53	10.71	13.12	<b>8.70</b>	<b>10.38</b>	<b>14.60</b>
6 Weather in Iowa	3.30	5.44	7.40	<b>5.51</b>	<b>7.97</b>	<b>11.67</b>
7 Chicago traffic	6.55	7.41	10.36	<b>8.81</b>	<b>11.60</b>	<b>15.55</b>
8 Libraries in Montana	<b>9.40</b>	<b>12.57</b>	<b>15.30</b>	9.29	12.56	15.26
9 Natural disasters in Utah	3.18	5.45	8.30	<b>7.22</b>	<b>8.82</b>	<b>10.85</b>
10 Crimes in Tennessee state	<b>5.03</b>	<b>7.54</b>	<b>11.90</b>	1.74	1.97	2.92
11 California transportation	5.28	5.95	7.36	<b>6.44</b>	<b>8.73</b>	<b>12.68</b>
12 Agriculture in Michigan	6.32	7.03	8.61	<b>8.69</b>	<b>9.98</b>	<b>12.36</b>
13 California weather	6.11	8.27	10.49	<b>6.93</b>	<b>9.18</b>	<b>12.42</b>
14 Tourist attraction in LA	1.57	2.03	3.43	<b>6.43</b>	<b>8.18</b>	<b>11.35</b>
15 Hurricane in Louisiana	4.18	5.64	9.33	<b>7.11</b>	<b>9.22</b>	<b>13.20</b>
16 Universities in Boston	2.14	2.68	4.30	<b>5.66</b>	<b>7.23</b>	<b>9.10</b>
17 Hospitals in New York	1.90	2.63	4.41	<b>5.82</b>	<b>8.70</b>	<b>12.17</b>
18 Grocery store in Seattle	6.12	8.17	11.28	<b>10.40</b>	<b>13.93</b>	<b>16.99</b>
19 Highways in Los Angeles	2.22	2.92	4.19	<b>7.64</b>	<b>9.09</b>	<b>10.26</b>
20 Air pollution of New York	6.88	8.37	9.76	<b>7.04</b>	<b>9.55</b>	<b>12.71</b>
<b>Average DCG</b>	4.84	6.31	8.68	<b>7.16</b>	<b>9.34</b>	<b>12.43</b>

a benchmark dataset to evaluate the proposed GIR model as well as a widely used baseline model - Lucene’s practical scoring function model. The results demonstrate that our semantic query expansion model significantly outperforms the Lucene baseline, thereby highlighting the effectiveness of our proposed approach.

As for future research, we want to improve the efficiency of the presented semantic query expansion framework. We also want to investigate other ways to measure spatial similarity such as Space2Vec (Mai et al. 2020). In addition, we are interested in evaluating the impact of different spatial similarity measures on the performance of GIR systems more generally. Moreover, we plan to investigate the question of whether the added geospatial aspect of GIR will affect the way how we evaluate the system.

## Acknowledgments

This presented work was partially done while the first author was interning at Esri Inc.. This work is partially funded by Esri Inc. and the NSF award 1936677 C-Accel Pilot - Track A1 (Open Knowledge Network): *Spatially-Explicit Models, Methods, And Services For Open Knowledge Networks*. We thank four Ph.D. students from



UC Santa Barbara for evaluation data annotations: Jingyi Xiao, Ning Zhang, Haoxin Zhou, and Yao Xuan.

## References

- Azad, H. K. & Deepak, A. (2019), 'Query expansion techniques for information retrieval: a survey', *Information Processing & Management* **56**(5), 1698–1735.
- Buckley, C., Salton, G., Allan, J. & Singhal, A. (1995), 'Automatic query expansion using smart: Trec 3', *NIST special publication sp* pp. 69–69.
- Carterette, B. & Jones, R. (2008), Evaluating search engines by modeling the relationship between relevance and clicks, in 'Advances in Neural Information Processing Systems', pp. 217–224.
- Dominich, S. (2008), *The modern algebra of information retrieval*, Springer.
- Frontiera, P., Larson, R. & Radke, J. (2008), 'A comparison of geometric approaches to assessing spatial similarity for gir', *International Journal of Geographical Information Science* **22**(3), 337–360.
- Hu, Y., Janowicz, K., Prasad, S. & Gao, S. (2015a), Enabling semantic search and knowledge discovery for arcgis online: A linked-data-driven approach, in 'AGILE 2015', Springer, pp. 107–124.
- Hu, Y., Janowicz, K., Prasad, S. & Gao, S. (2015b), 'Metadata topic harmonization and semantic search for linked-data-driven geoportals: A case study using arcgis online', *Transactions in GIS* **19**(3), 398–416.
- Huang, S., Zhao, Q., Mitra, P. & Giles, C. L. (2008), Hierarchical location and topic based query expansion., in 'AAAI', pp. 1150–1155.
- Jaccard, P. (1912), 'The distribution of the flora in the alpine zone. 1', *New phytologist* **11**(2), 37–50.
- Janowicz, K., Raubal, M. & Kuhn, W. (2011), 'The semantics of similarity in geographic information retrieval', *Journal of Spatial Information Science* **2011**(2), 29–57.
- Järvelin, K. & Kekäläinen, J. (2002), 'Cumulated gain-based evaluation of ir techniques', *ACM Transactions on Information Systems (TOIS)* **20**(4), 422–446.
- Jiang, Y., Li, Y., Yang, C., Hu, F., Armstrong, E. M., Huang, T., Moroni, D., McGibbney, L. J. & Finch, C. J. (2018), 'Towards intelligent geospatial data discovery: A machine learning framework for search ranking', *International journal of digital earth* **11**(9), 956–971.
- Jones, C. B. & Purves, R. S. (2008), 'Geographical information retrieval', *International Journal of Geographical Information Science* pp. 219–228.
- Mai, G., Janowicz, K. & Yan, B. (2018), Combining text embedding and knowledge graph embedding techniques for academic search engines., in 'Semdeep/NLIWoD@ ISWC', pp. 77–88.
- Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L. & Lao, N. (2020), Multi-scale representation learning for spatial feature distributions using grid cells, in 'The Eighth International Conference on Learning Representations', openreview.

- Mai, G., Yan, B., Janowicz, K. & Zhu, R. (2019), Relaxing unanswerable geographic questions using a spatially explicit knowledge graph embedding model, *in* 'AG-ILE: The 22nd Annual International Conference on Geographic Information Science', Springer, pp. 21–39.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Advances in neural information processing systems', pp. 3111–3119.
- Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global vectors for word representation, *in* 'Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)', pp. 1532–1543.
- Rocchio, J. (1971), 'Relevance feedback in information retrieval', *The Smart retrieval system-experiments in automatic document processing* pp. 313–323.
- Van Kreveld, M., Reinbacher, I., Arampatzis, A. & Van Zwol, R. (2005), 'Multi-dimensional scattered ranking methods for geographic information retrieval', *GeoInformatica* **9**(1), 61–84.
- Vechtomova, O. (2009), 'Query expansion for information retrieval', *Encyclopedia of database systems* pp. 2254–2257.