Geotagging Non-Spatial Concepts

Amgad Madkour, Walid G. Aref, *Mohamed Mokbel, ** Saleh Basalamah

Purdue University, USA
* University of Minnesota, Twin-Cities, USA
** Umm Al-Qura University, KSA
A user wants to identify points of interest (POI) on the map that match his query ...

Query 1: Find locations that are responsible for pollution
Query 2: Find locations that are related to crime
Query 3: Find locations that are associated with health

Question:
How to find locations on the map that are related to non-spatial concepts?
KEY OBSERVATIONS

• Using the *semantic information* associated with concepts for identifying relations between spatial and non-spatial concepts

• Probing the *textual co-occurrences* of spatial and non-spatial concepts for identifying relations between spatial and non-spatial concepts

• Generalizing the relatedness based on the *concepts type* instead of relatedness between two specific concepts
  • **Example:**
    • **Query**: Find locations related to *Research* in the *United States*
    • **Expected Output**: Display all locations of type ‘School’, ‘University’ within *United States*
CHALLENGES

• How to *represent co-ocurrences* of spatial and non-spatial concepts within the same textual resource

• How to *infer the types* of spatial concepts that are semantically related to the non-spatial query concept

• How to *evaluate* given that there is no known dataset for type-relatedness between spatial and non-spatial concepts
CONTRIBUTION

• We propose CGTag, a system for geotagging a non-spatial concept query with spatial concepts based on type relatedness

• We propose a semantic query-processing algorithm that utilizes several Linked-Data-based filtering strategies

• We propose an evaluation method for type relatedness in addition to a baseline to determine the correctness of the results
REPRESENTING CO-OCCURRENCES

• **Hypothesis**
  • “All concepts mentioned in the same textual resource are implicitly related to each other”
  • **Example**: A single text document can have (Pollution – Factory – Industry – Waste)

• A clique can be used to represent the concepts co-occurrences
  • Vertex \(\rightarrow\) Concept
  • Edge \(\rightarrow\) Weighted Relation

• **Using Cliques**
  • To indicate a *single* (initially) co-occurrence between the concepts and each other
COMPONENTS

• **Information Extraction**
  • Identification, disambiguation, entity linking
  • Example: `<dbpedia.org/resource/Barack_Obama>Obama</>`

• **Graph Construction**
  • Construct a **local graph** (document level)
  • A clique is used to represent a single (initially) co-occurrence of a concept with other concepts in the same document

• **Knowledge Store**
  • Online Mode: Answer user queries
  • Offline Mode: Store the result of the local graph construction to a **global graph**
COMPONENTS

• **Semantic Query Processor (SQP)**
  • Infer the types of spatial concepts in the global graph that are most related to the non-spatial concept query

• **Parameters**
  • **Input**: the non-spatial concept query
  • **Output**: a location of interest

• **Filtering steps**
  • **Co-occurrence threshold** – Co-occurrence frequency/weight
  • **Linked Data properties** – Ontology Type, Spatial Information
    • **Example**: *Type:Building* - Superclass of (Hotel, Restaurant, Shopping Mall, Castle, HistoricBuilding)
  • **Similarity Filtering** - Pairwise document similarity between the textual resources of the concepts (TF-IDF as a representation)
• **Type Filtering of Non-Spatial Concepts**
  • Determine the spatial concepts that have a type that matches the types deduced by the semantic query processor
  
  • If a location is specified in the query, then the location acts as a filtering criteria for the spatial concepts

  • **Example**: Semantic Query Processor proposes: “Art”
  • Spatial linking module attempts to match the type “Art” against the types of geo-tagged resources.
  • If location is specified such as “NYC” then the linking is restricted to “NYC” only
EXPERIMENTAL SETUP

• **CGTag** is evaluated based on two overlapping factors:

  • **Query processing filtering efficiency**
    • Each filtering criteria is evaluated separately and then in combination with each other
    • The number of remaining concepts are observed after each filter has been applied

  • **The accuracy of the type relatedness**
    • Presented 9 evaluators with 30 arbitrarily selected non-spatial concept queries.
    • Given a non-spatial concept, the objective is to understand what would be the expected types of spatial concepts in the result.
      • **Example**: Fishing is more related to ‘Island’ and ‘City’ types than ‘School’ and ‘University’
EXPERIMENTAL SETUP

• Collections and Datasets
  • Wikipedia: 178K articles
  • DBPedia: Rich medium for interlinking the concepts mentioned in Wikipedia with other collection
  • LinkedGeodata: Spatial Information + Interlinking dataset
    • An interlinking dataset indicates what a resource in one dataset corresponds to in another dataset
      • Example: Obama (DBPedia) → Obama (Wikipedia)

• Baseline
  • The co-occurrence threshold is used as the baseline

• Concept Extraction
  • DBPedia Spotlight – Provides identification, disambiguation and entity linking
EXPERIMENTAL SETUP

• Queries

<table>
<thead>
<tr>
<th>Query</th>
<th>Airport</th>
<th>City</th>
<th>Island</th>
<th>Mountain</th>
<th>School</th>
<th>Stadium</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Fishing</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Broadcasting</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Use Case:

• **Online Phase:** Find locations related to **Science** in the **United States**
• **Semantic Query Processor Output:** School, University
• **Type Filter Output:** Show all locations of type ‘**School**’, ‘**University**’ within **United States**
EXPERIMENTAL SETUP

• Interlinking Dataset
  • We use the criteria as the target ‘Type’ for the queries

<table>
<thead>
<tr>
<th>Criteria</th>
<th>USA</th>
<th>GERMANY</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airport</td>
<td>3128</td>
<td>27</td>
<td>109</td>
</tr>
<tr>
<td>City</td>
<td>8469</td>
<td>7409</td>
<td>4521</td>
</tr>
<tr>
<td>Island</td>
<td>92</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>Mountain</td>
<td>887</td>
<td>76</td>
<td>587</td>
</tr>
<tr>
<td>School</td>
<td>2026</td>
<td>7</td>
<td>154</td>
</tr>
<tr>
<td>Stadium</td>
<td>55</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>University</td>
<td>70</td>
<td>4</td>
<td>25</td>
</tr>
</tbody>
</table>
RESULTS

• **Type Relatedness Evaluation**
  • All linked data filters in addition to the co-occurrence similarity provide the highest accuracy across 3 datasets

<table>
<thead>
<tr>
<th>Technique</th>
<th>USA</th>
<th>UK</th>
<th>Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked Data without Similarity</td>
<td>0.43</td>
<td>0.43</td>
<td>0.42</td>
</tr>
<tr>
<td>Linked Data with Similarity</td>
<td>0.69</td>
<td>0.7</td>
<td>0.78</td>
</tr>
<tr>
<td>Co-occurrence Threshold (3)</td>
<td>0.78</td>
<td>0.68</td>
<td>0.06</td>
</tr>
<tr>
<td>Linked Data without Similarity + Threshold (3)</td>
<td>0.53</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>Linked Data with Similarity + Threshold (3)</td>
<td><strong>0.72</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.76</strong></td>
</tr>
</tbody>
</table>
RESULTS

• **Query processing filtering efficiency**
  
  • **Evaluated:**
    
    • LET: Linked Data Type Expansion
    • LTP: Linked Data Type Pruning
    • LSIM: Similarity Pruning
    • LSP: Spatial Pruning
    • THR(2): Co-occurrence filtering with weight 2
    • Linked Data (all) filters

  • **Result:**
    
    • Linked Data filtering (all) + co-occurrence achieves the highest filtering efficiency while still maintaining a good accuracy
CONCLUSION

• Presented **CGTag**, a system for discovering type relatedness between spatial and non-spatial concepts

• Demonstrates how co-occurrences can be used as a means for discovering implicit relationships between non-spatial and spatial concepts

• Presented a query-processing algorithm that identifies the spatial types related to a query-specified non-spatial concept

• The type-relatedness accuracy averages at 73%
Thank you