Exploiting the Semantic Web for the Automatic Extraction of Los Angeles City Data

Advisor: Letizia Tanca
Co-Advisor: Craig Knoblock

Author: Marianna Bucchi
Student ID: 898422
Introduction: Problem Identification

Large number of datasets available about the city of Los Angeles

Data access and integration are not easy in the Big Data environment

(Dong et al. 2015)

Departments within the city of Los Angeles are unaware of relevant datasets
Provide a **better access** to the data the city already has

**How?**

Automatically extracting the **content** from each dataset in a form of a class
Preliminary Notions

The Web was designed as an information space with the goal to be useful for human-to-human communication, while the **Semantic Web** approach develops methods and languages for expressing information in a machine-processable form.

*(Berners-Lee et al. 2001)*

**Ontologies** are stores of information accessible through queries on the Web, which describe the contextual relations between concepts and specify logical rules for reasoning about them.

*(Ismayilov et al. 2018)*
The Model Development

INPUT

First Phase
- First Model

Second Phase
- CTA Model
- CEA Model

Last Phase
- Frequency Model
- CTA Model
- CEA Model
- OpenStreetMap Model

OUTPUT

Wikidata Classes

DBpedia Classes

DBpedia Classes
From the observation of the datasets, we collected **some exclusion criteria** to isolate a single column from a dataset.
The Model Development

INPUT

First Phase
First Model

OUTPUT

Wikidata Classes

Second Phase
CTA Model
CEA Model

DBpedia Classes

Last Phase
Frequency Model
CTA Model
CEA Model
OpenStreetMap Model

DBpedia Classes
### RESULTS
- Meaningless classes in Wikidata
- Correctness in the generation of the candidates

#### First Phase: Wikidata Model

<table>
<thead>
<tr>
<th>Items</th>
<th>Q13220204</th>
<th>Q13360155</th>
<th>Q13410400</th>
<th>Q1496967</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autauga</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Baldwin</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Barbour</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Bibb</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Blount</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Bullock</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Butler</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Calhoun</td>
<td>2.0</td>
<td>2.0</td>
<td>1.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>
The Model Development

INPUT

First Phase
First Model

Second Phase
CTA Model
CEA Model

Last Phase
Frequency Model
CTA Model
CEA Model

OUTPUT

Wikidata Classes

DBpedia Classes

(Thawani et al. 2019)
Second Phase: the Column Type Annotation Model

Marianna Bucchi

Four Configurations
- Levenshtein Multiplied by TFIDF
- Weighted Average between Levenshtein and TFIDF
- Jaro-Winkler Multiplied by TFIDF
- Weighted Average between Jaro-Winkler and TFIDF

(Vembunarayanan, 2013)
Results Evaluation

\[ \text{Accuracy}_i = \frac{n_i}{D} \]

- \( n \) is the number of correct classes detected by the methodology \( i \) applied
- \( D \) is the total number of datasets evaluated (41)

<table>
<thead>
<tr>
<th>DATASET</th>
<th>OUTPUT</th>
<th>CLASS EXPECTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cultural Centers</td>
<td>Venue</td>
<td>Venue</td>
</tr>
<tr>
<td>Cultural Events</td>
<td>Museum</td>
<td>Event</td>
</tr>
<tr>
<td>Education Facilities</td>
<td>Organisation</td>
<td>Educational Institution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>Levenshtein multiplied by the TFIDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permissive Accuracy</td>
<td>0.36585</td>
</tr>
<tr>
<td>Restrictive Accuracy</td>
<td>0.31707</td>
</tr>
</tbody>
</table>
The Model Development

INPUT

First Phase
First Model

OUTPUT

Wikidata Classes

Second Phase
CTA Model
CEA Model

DBpedia Classes

Last Phase
Frequency Model
CTA Model
CEA Model
OpenStreetMap Model

DBpedia Classes
Last Phase: Final Model Development

Marianna Bucchi

MEASURE | CTA | Frequency
---|---|---
Permissive Accuracy | 0.36585 | 0.53658
Restrictive Accuracy | 0.31707 | 0.39024
Final Model Design

Marianna Bucchi
Performance Evaluations

<table>
<thead>
<tr>
<th>MEASURE</th>
<th>CTA + Frequency</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permissive Accuracy</td>
<td>0.56098</td>
<td>0.80487</td>
</tr>
<tr>
<td>Restrictive Accuracy</td>
<td>0.41463</td>
<td>0.56098</td>
</tr>
</tbody>
</table>
Conclusions

- Exploitation of Web ontologies
- A good reasoning capacity by integrating different approaches
- Seeking for efficiency
- Improve candidates’ generation phase (Manning et. al, 2008)
- Apply more advanced methodologies for column selection (Pham et. al, 2016)
- Introduce more sophisticated approaches for Frequency model
Thank you for the attention!

Any Questions?