An Evolutionary Algorithm to Learn SPARQL Queries for Source-Target-Pairs
Finding Patterns for Human Associations in DBpedia

Jörn Hees

2016-11-20
EKAW 2016
Outline

• Background
• My Research (Demo)
• Graph Pattern Learning
• Evaluation
What are Human Associations?

- Mental connections between concepts
- What's the first thing that comes to your mind when thinking about ...?
- Example:
  - Dog
What are Human Associations?

- Mental connections between concepts
- What's the first thing that comes to your mind when thinking about ...?
- Example:
  - Dog: Cat, collar, leash, walk, fur, bark
What are Human Associations?

• Mental connections between concepts

• What's the first thing that comes to your mind when thinking about ... ?

• Example:
  • Dog: Cat, collar, leash, walk, fur, bark
  • House
What are Human Associations?

• Mental connections between concepts
• What's the first thing that comes to your mind when thinking about ... ?

• Example:
  • Dog: Cat, collar, leash, walk, fur, bark
  • House: Roof, door, window, flat, live
Example: DBpedia:Dog
Example: DBpedia:Dog

- Basic information (description, types)
Example: DBpedia:Dog

- **Basic information** (description, types)
- **Categorisation** (categories & types)
Example: DBpedia:Dog

- **Basic information** (description, types)
- **Categorisation** (categories & types)
- **Links to other datasets** (also Freebase)
Example: DBpedia:Dog

• **Basic information** *(description, types)*

• **Categorisation** *(categories & types)*

• **Links to other datasets** *(also Freebase)*

• **Inverse links** *(redirects, link from other datasets)*

https://w3id.org/associations
Example: DBpedia:Dog

- Basic information (description, types)
- Categorisation (categories & types)
- Links to other datasets (also Freebase)
- Inverse links (redirects, link from other datasets)
- Wiki-page-links (377 for Dog)
Example: DBpedia:Dog

- **Basic information** (description, types)
- **Categorisation** (categories & types)
- **Links to other datasets** (also Freebase)
- **Inverse links** (redirects, link from other datasets)
- **Wiki-page-links** (377 for Dog)
- **Associations?**
  - Dog:
    - Cat (✓), collar (✗), leash (✗), walk (✗), fur (✓), bark (✗)
Outline

• Background
• My Research (Demo)
• Graph Pattern Learning
• Evaluation
Outline

• Background

• My Research (Demo)

• Graph Pattern Learning

• Evaluation
My Research

• Research Question:

• Is it possible to learn patterns for Human Associations from Linked Data?
My Research

• Research Question:
  • Is it possible to learn patterns for Human Associations from Linked Data?
My Research

• Research Question:
  • Is it possible to learn patterns for Human Associations from Linked Data?
My Research

• Research Question:
  • Is it possible to learn patterns for Human Associations from Linked Data?
My Research

• Research Question:
  • Is it possible to learn patterns for Human Associations from Linked Data?

• Goal:
  • Given an input node predict the output node(s) we would associate
My Research

• Research Question:
  • Is it possible to learn patterns for Human Associations from Linked Data?

• Goal:
  • Given an input node predict the output node(s) we would associate

DEMO

https://w3id.org/associations
Demo
dbr:Linked_data

Top 10 predictions (method: target_occs)
- <http://dbpedia.org/resource/Semantic_Web> (7.000)
- <http://dbpedia.org/resource/Wikipedia> (3.000)
- <http://dbpedia.org/resource/Open_Semantic_Framework> (3.000)
- <http://dbpedia.org/resource/Tim_Berners-Lee> (2.000)
- <http://dbpedia.org/resource/Dereferenceable_Uniform_Resource_Identifier> (2.000)
- <http://dbpedia.org/resource/World_Wide_Web> (2.000)
- <http://dbpedia.org/resource/Hypertext_Transfer_Protocol> (2.000)
- <http://dbpedia.org/resource/Language> (2.000)
- <http://dbpedia.org/resource/Uniform_resource_identifier> (2.000)
- <http://dbpedia.org/resource/DBpedia> (2.000)

Top 10 predictions (method: precisions)
- <http://dbpedia.org/resource/Semantic_Web> (5.750)
- <http://dbpedia.org/resource/Tim_Berners-Lee> (1.450)
- <http://dbpedia.org/resource/Dereferenceable_Uniform_Resource_Identifier> (1.250)
- <http://dbpedia.org/resource/Language> (1.200)
- <http://dbpedia.org/resource/Data_set> (1.000)
- <http://dbpedia.org/resource/Serialization> (1.000)
- <http://dbpedia.org/resource/DBpedia> (0.833)
- <http://dbpedia.org/resource/World_Wide_Web> (0.750)
dbpedia:Semantic Web

Top 10 predictions (method: target_occurs)

1. [http://dbpedia.org/resource/World_Wide_Web] (15.000)
2. [http://dbpedia.org/resource/Web_search_engine] (10.000)
4. [http://dbpedia.org/resource/Blog] (7.000)
5. [http://dbpedia.org/resource/Semantic_Web] (7.000)
6. [http://dbpedia.org/resource/Web_2.0] (7.000)
7. [http://dbpedia.org/resource/Artificial_intelligence] (6.000)
8. [http://dbpedia.org/resource/Cascading_Style_Sheets] (6.000)
9. [http://dbpedia.org/resource/Tim_Berners-Lee] (5.000)
10. [http://dbpedia.org/resource/Language] (4.000)

Top 10 predictions (method: precisions)

1. [http://dbpedia.org/resource/World_Wide_Web] (7.316)
4. [http://dbpedia.org/resource/Web_search_engine] (3.051)
5. [http://dbpedia.org/resource/Artificial_intelligence] (2.977)
7. [http://dbpedia.org/resource/Web_2.0] (1.791)
8. [http://dbpedia.org/resource/Tim_Berners-Lee] (1.568)
10. [http://dbpedia.org/resource/Provenance] (1.500)
My Research

• Research Question:

  • Is it possible to learn patterns for Human Associations from Linked Data?

• Dataset of "Semantic Associations" needed
My Research

• Research Question:
  • Is it possible to learn patterns for Human Associations from Linked Data?

• Dataset of "Semantic Associations" needed

https://w3id.org/associations
Edinburgh Associative Thesaurus

• Association corpus (1973) G. Kiss, C. Armstrong, R. Milroy, J. Piper
  • For each stimulus asked 100 ppl for a response
  • Strong responses became stimuli of next round

https://w3id.org/associations
Edinburgh Associative Thesaurus

- Association corpus (1973) G. Kiss, C. Armstrong, R. Milroy, J. Piper
  - For each stimulus asked 100 ppl for a response
  - Strong responses became stimuli of next round
- ~ 790 K associations (free text)
- Graph: (|V| = 23 K, |E| = 325 K)
Edinburgh Associative Thesaurus

- Association corpus (1973) by G. Kiss, C. Armstrong, R. Milroy, J. Piper
  - For each stimulus asked 100 ppl for a response
  - Strong responses became stimuli of next round
- ~790 K associations (free text)
- Graph: ($|V| = 23$ K, $|E| = 325$ K)
  - ~5000 strong associations (>19x)
  - ~167.4 K raw associations

https://w3id.org/associations
Edinburgh Associative Thesaurus

- **Difficulty:** Free text
- **Mapped to Semantic (DBpedia) Entities**
  - "pupil" - "school"

- **Semi-Automatic Mapping Approach**
  - 727 verified distinct “Semantic Associations”
  - ~25.5 K raw associations

https://w3id.org/associations
Semantic Associations Dataset

- (Raw) EAT as RDF (1.7 M triples)
- 727 verified distinct Semantic Associations
### Semantic Associations Dataset

- **727 verified distinct Semantic Associations**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbr:Cow</td>
<td>dbr:Milk</td>
</tr>
<tr>
<td>dbr:Camping</td>
<td>dbr:Tent</td>
</tr>
<tr>
<td>dbr:Expense</td>
<td>dbr:Money</td>
</tr>
<tr>
<td>dbr:Bed</td>
<td>dbr:Sleep</td>
</tr>
<tr>
<td>dbr:Pupil</td>
<td>dbr:Eye</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- **Not readily modelled in DBpedia!**
- **Not one property!**
Outline

• Background

• My Research (Demo)

• Graph Pattern Learning

• Evaluation
Outline

• Background
• My Research (Demo)
  • Graph Pattern Learning
• Evaluation
Data Analysis

• Local Linked Data Endpoint:
  • Central Datasets
  • \(\sim8\) G triples
  • SPARQL Queries
• Scalability Issues
First Analysis

• Node Degrees:
  • Avg: 3643
• Excluding big nodes?
  • Would cause bias!
• Exclude in-links?
  • Directionality depends on modelling!

<table>
<thead>
<tr>
<th>Node</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbp:Animal</td>
<td>400624</td>
</tr>
<tr>
<td>dbp:Insect</td>
<td>195058</td>
</tr>
<tr>
<td>dbp:France</td>
<td>190047</td>
</tr>
<tr>
<td>dbp:India</td>
<td>181119</td>
</tr>
<tr>
<td>dbp:Italy</td>
<td>132719</td>
</tr>
<tr>
<td>dbp:Village</td>
<td>132400</td>
</tr>
<tr>
<td>dbp:Plant</td>
<td>126731</td>
</tr>
<tr>
<td>dbp:Scotland</td>
<td>71828</td>
</tr>
<tr>
<td>dbp:Paris</td>
<td>64232</td>
</tr>
<tr>
<td>dbp:Switzerland</td>
<td>55471</td>
</tr>
</tbody>
</table>

https://w3id.org/associations
Graph Pattern Learning

- First idea: Shortest paths
  - Many false positives (associations)
  - Problems with high degrees
    - Super-nodes (owl:Thing, Lists, countries, cities)
      - Everything is connected with paths of length 2 ;)

![Diagram](https://w3id.org/associations)
Graph Pattern Learning

• First idea: Shortest paths
  • Many false positives (associations)
  • Problems with high degrees
    • Super-nodes (owl:Thing, Lists, countries, cities)
      • Everything is connected with paths of length 2 ;)
• Problems due to modelling
  • Nearly linear parts / chains
    • (owl:sameAs, freebase, temporal properties)
Graph Pattern Learning

- Shortest paths
Good Graph Patterns?

• How often is a response reached? $\max$

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbr:Cow</td>
<td>dbr:Milk</td>
</tr>
<tr>
<td>dbr:Camping</td>
<td>dbr:Tent</td>
</tr>
<tr>
<td>dbr:Expense</td>
<td>dbr:Money</td>
</tr>
<tr>
<td>dbr:Bed</td>
<td>dbr:Sleep</td>
</tr>
<tr>
<td>dbr:Pupil</td>
<td>dbr:Eye</td>
</tr>
</tbody>
</table>
Good Graph Patterns?

- How often is a response reached? $\text{max}$
- How many other nodes are reached? $\text{min}$

Stimulus
- dbr:Cow
- dbr:Camping
- dbr:Expense
- dbr:Bed
- dbr:Pupil

Response
- dbr:Milk
- dbr:Tent
- dbr:Money
- dbr:Sleep
- dbr:Eye
Good Graph Patterns?

• How often is a response reached? max
• How many other nodes are reached? min
• How many nodes need to be expanded? min

Stimulus
- dbr:Cow
- dbr:Camping
- dbr:Expense
- dbr:Bed
- dbr:Pupil

Response
- dbr:Milk
- dbr:Tent
- dbr:Money
- dbr:Sleep
- dbr:Eye

https://w3id.org/associations
Good Graph Patterns?

- How often is a response reached? \( \max \)
- How many other nodes are reached? \( \min \)
- How many nodes need to be expanded? \( \min \)
- How long does a query take? \( \min \)

**Stimulus**
- dbr:Cow
- dbr:Camping
- dbr:Expense
- dbr:Bed
- dbr:Pupil

**Response**
- dbr:Milk
- dbr:Tent
- dbr:Money
- dbr:Sleep
- dbr:Eye
Evolutionary Algorithm

- Fitness function: Good Graph Pattern?
- Individuals: SPARQL BGP patterns

```sparql
{  
}
```

[Diagram showing RDF graph with nodes for `?source`, `?target`, and edge labeled `?p` with property `dbo:Animal` for each node.]
Evolutionary Algorithm

- Fitness function: Good Graph Pattern?
- Individuals: SPARQL BGP patterns
- Mutation:
Evolutionary Algorithm

• Fitness function: Good Graph Pattern?
• Individuals: SPARQL BGP patterns
• Mutation:
  • Add

```
?source
  rdf:type Animal
  ?p  
?target
  rdf:type

?n
  ?q
```

https://w3id.org/associations
Evolutionary Algorithm

• Fitness function: Good Graph Pattern?
• Individuals: SPARQL BGP patterns
• Mutation:
  • Add / delete triples
Evolutionary Algorithm

- Fitness function: Good Graph Pattern?
- Individuals: SPARQL BGP patterns
- Mutation:
  - Add / delete triples
  - Property or node ↔ variable
Evolutionary Algorithm

- Fitness function: Good Graph Pattern?
- Individuals: SPARQL BGP patterns
- Mutation:
  - Add / delete triples
  - Property or node ↔ variable
Evolutionary Algorithm

- Fitness function: Good Graph Pattern?
- Individuals: SPARQL BGP patterns
- Mutation:
  - Add / delete triples
  - Property or node ↔ variable
- Mating: Exchange of triples & unifying vars
- Selection: Tournament
Evolutionary Algorithm

- Starting Population:
  - Randomised length paths between \( \text{source} \) and \( \text{target} \)
Evolutionary Algorithm

• Starting Population:
  • Randomised length paths between `?source` and `?target`
• Keeping the population healthy
  • Re-introduce basic / good fit patterns
Pattern Coverage

• Multiple runs of Evolutionary Algorithm

• After good patterns are found, refocus on remaining source-target pairs

Stimulus
- dbr:Cow
- dbr:Camping
- dbr:Expense
- dbr:Bed
- dbr:Pupil

Response
- dbr:Milk
- dbr:Tent
- dbr:Money
- dbr:Sleep
- dbr:Eye
Pattern Coverage

- Multiple runs of Evolutionary Algorithm
- After good patterns are found, refocus on remaining source-target pairs

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbr:Cow</td>
<td>dbr:Milk</td>
</tr>
<tr>
<td>dbr:Camping</td>
<td>dbr:Tent</td>
</tr>
<tr>
<td>dbr:Expense</td>
<td>dbr:Money</td>
</tr>
<tr>
<td>dbr:Bed</td>
<td>dbr:Sleep</td>
</tr>
<tr>
<td>dbr:Pupil</td>
<td>dbr:Eye</td>
</tr>
</tbody>
</table>
Pattern Coverage

• Multiple runs of Evolutionary Algorithm

• After good patterns are found, refocus on remaining source-target pairs

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbr:Cow</td>
<td>dbr:Milk</td>
</tr>
<tr>
<td>dbr:Camping</td>
<td>dbr:Tent</td>
</tr>
<tr>
<td>dbr:Expense</td>
<td>dbr:Money</td>
</tr>
<tr>
<td>dbr:Bed</td>
<td>dbr:Sleep</td>
</tr>
<tr>
<td>dbr:Pupil</td>
<td>dbr:Eye</td>
</tr>
</tbody>
</table>
Pattern Coverage

- Multiple runs of Evolutionary Algorithm
- After good patterns are found, refocus on remaining source-target pairs

**Stimulus**
- dbr:Cow
- dbr:Camping
- dbr:Expense
- dbr:Bed
- dbr:Pupil

**Response**
- dbr:Milk
- dbr:Tent
- dbr:Money
- dbr:Sleep
- dbr:Eye
Pattern Coverage

• Multiple runs of Evolutionary Algorithm
  • After good patterns are found, refocus on remaining source-target pairs

Stimulus
- dbr:Cow
- dbr:Camping
- dbr:Expense
- dbr:Bed
- dbr:Pupil

Response
- dbr:Milk
- dbr:Tent
- dbr:Money
- dbr:Sleep
- dbr:Eye

https://w3id.org/associations
Pattern Coverage

- Multiple runs of Evolutionary Algorithm
- After good patterns are found, refocus on remaining source-target pairs

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbr:Cow</td>
<td>dbr:Milk</td>
</tr>
<tr>
<td>dbr:Camping</td>
<td>dbr:Tent</td>
</tr>
<tr>
<td>dbr:Expense</td>
<td>dbr:Money</td>
</tr>
<tr>
<td>dbr:Bed</td>
<td>dbr:Sleep</td>
</tr>
<tr>
<td>dbr:Pupil</td>
<td>dbr:Eye</td>
</tr>
</tbody>
</table>
Learned Graph Patterns

https://w3id.org/associations
Info from BabelNet
Info from WikiData
Info from WikiData
Info from BabelNet
Outline

- Background
- My Research (Demo)
- Graph Pattern Learning
- Evaluation
Outline

- Background
- My Research (Demo)
- Graph Pattern Learning
- Evaluation
Evaluation

• How good are the learned patterns?
  • Difficult to evaluate directly

• Indirect objective approach:
  • Are they good for prediction?
  • Training/Test set split
  • Clustered similar GPs
  • Given a stimulus from the test set, what’s the rank of the true response in the prediction results?
Evaluation Results

The graph shows the recall at k (recall@k) for various measures across different values of k. Each line represents a different measure:
- f_measures
- gp_precisions
- precisions
- scores
- target_occs
- hits_bidi
- hits_in
- hits_out
- indeg_bidi
- indeg_in
- indeg_out
- outdeg_bidi
- outdeg_in
- outdeg_out
- pagerank_bidi
- pagerank_in
- pagerank_out

The x-axis represents the value of k, ranging from 0 to 100, and the y-axis represents the recall@k, ranging from 0.0 to 1.0.
## Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>Recall@1</th>
<th>Recall@2</th>
<th>Recall@3</th>
<th>Recall@4</th>
<th>Recall@5</th>
<th>Recall@10</th>
<th>MAP</th>
<th>NDCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>outdeg in</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.042</td>
<td>0.097</td>
<td>0.029</td>
<td>0.105</td>
</tr>
<tr>
<td>outdeg out</td>
<td>0.069</td>
<td>0.125</td>
<td>0.153</td>
<td>0.153</td>
<td>0.167</td>
<td>0.181</td>
<td>0.126</td>
<td>0.209</td>
</tr>
<tr>
<td>outdeg bidi</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.014</td>
<td>0.056</td>
<td>0.125</td>
<td>0.045</td>
<td>0.131</td>
</tr>
<tr>
<td>indeg in</td>
<td>0.056</td>
<td>0.111</td>
<td>0.153</td>
<td>0.167</td>
<td>0.181</td>
<td>0.306</td>
<td>0.129</td>
<td>0.207</td>
</tr>
<tr>
<td>indeg out</td>
<td>0.056</td>
<td>0.125</td>
<td>0.153</td>
<td>0.153</td>
<td>0.194</td>
<td>0.194</td>
<td>0.121</td>
<td>0.200</td>
</tr>
<tr>
<td>indeg bidi</td>
<td>0.042</td>
<td>0.069</td>
<td>0.111</td>
<td>0.139</td>
<td>0.139</td>
<td>0.194</td>
<td>0.104</td>
<td>0.205</td>
</tr>
<tr>
<td>pagerank in</td>
<td>0.069</td>
<td>0.125</td>
<td>0.153</td>
<td>0.194</td>
<td>0.194</td>
<td>0.292</td>
<td>0.140</td>
<td>0.219</td>
</tr>
<tr>
<td>pagerank out</td>
<td>0.056</td>
<td>0.097</td>
<td>0.153</td>
<td>0.153</td>
<td>0.167</td>
<td>0.208</td>
<td>0.117</td>
<td>0.195</td>
</tr>
<tr>
<td>pagerank bidi</td>
<td>0.056</td>
<td>0.069</td>
<td>0.111</td>
<td>0.139</td>
<td>0.153</td>
<td>0.236</td>
<td>0.113</td>
<td>0.219</td>
</tr>
<tr>
<td>hits in</td>
<td>0.014</td>
<td>0.028</td>
<td>0.042</td>
<td>0.069</td>
<td>0.083</td>
<td>0.111</td>
<td>0.046</td>
<td>0.095</td>
</tr>
<tr>
<td>hits out</td>
<td>0.056</td>
<td>0.056</td>
<td>0.111</td>
<td>0.125</td>
<td>0.153</td>
<td>0.181</td>
<td>0.102</td>
<td>0.181</td>
</tr>
<tr>
<td>hits bidi</td>
<td>0.014</td>
<td>0.042</td>
<td>0.042</td>
<td>0.056</td>
<td>0.069</td>
<td>0.125</td>
<td>0.050</td>
<td>0.110</td>
</tr>
<tr>
<td>scores</td>
<td>0.236</td>
<td>0.278</td>
<td>0.375</td>
<td>0.389</td>
<td>0.389</td>
<td>0.556</td>
<td>0.323</td>
<td>0.413</td>
</tr>
<tr>
<td>gp precisions</td>
<td>0.250</td>
<td>0.319</td>
<td>0.417</td>
<td>0.500</td>
<td>0.528</td>
<td>0.639</td>
<td>0.365</td>
<td>0.457</td>
</tr>
<tr>
<td>precisions</td>
<td>0.250</td>
<td>0.361</td>
<td>0.444</td>
<td>0.486</td>
<td>0.528</td>
<td>0.625</td>
<td>0.371</td>
<td>0.460</td>
</tr>
<tr>
<td>target occs</td>
<td>0.278</td>
<td>0.319</td>
<td>0.458</td>
<td>0.528</td>
<td>0.528</td>
<td>0.611</td>
<td>0.381</td>
<td>0.466</td>
</tr>
<tr>
<td>f measures</td>
<td><strong>0.306</strong></td>
<td>0.347</td>
<td><strong>0.472</strong></td>
<td>0.500</td>
<td><strong>0.542</strong></td>
<td>0.611</td>
<td><strong>0.399</strong></td>
<td><strong>0.479</strong></td>
</tr>
</tbody>
</table>
Summary

- **Goal**
  Learning Graph Patterns for Associations

- **Datasets**

- **Evolutionary Algorithm**
  Learns SPARQL Patterns for Source-Target-Pairs (> 60% Top-10 Accuracy)
Future Work

• Apply Evolutionary Algorithm
  • to other datasets
  • to other types of relations

• Extensions:
  • Work on Literals
  • Include FILTER
Discussion

Thanks for your attention

Questions?

https://w3id.org/associations