Knowledge Graph Refinement
A Survey of Approaches and Evaluation Methods
Semantic Web 8(3), 2017

Heiko Paulheim
Knowledge Graph: Definition

- This paper is about “Knowledge Graphs”
  - Review by Natasha Noy, June 2015:
    “Please define what a knowledge graph is – and what it is not.”
- Everybody talks about them, but what *is* a Knowledge Graph?
  - I don’t have a definition either...
Knowledge Graph: Definition

- Knowledge Graph definitions
- Many people talk about KGs, few give definitions

- Working definition: a Knowledge Graph
  - *mainly* describes instances and their relations in a graph
    - Unlike an ontology
    - Unlike, e.g., WordNet
  - Defines possible classes and relations in a *schema or ontology*
    - Unlike schema-free output of some IE tools
  - Allows for interlinking *arbitrary* entities with each other
    - Unlike a relational database
  - Covers *various* domains
    - Unlike, e.g., Geonames
Quantitative Overview on Knowledge Graphs

• Knowledge Graphs out there (not guaranteed to be complete)

<table>
<thead>
<tr>
<th>Name</th>
<th>Instances</th>
<th>Facts</th>
<th>Types</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBpedia (English)</td>
<td>4,806,150</td>
<td>176,043,129</td>
<td>735</td>
<td>2,813</td>
</tr>
<tr>
<td>YAGO</td>
<td>4,595,906</td>
<td>25,946,870</td>
<td>488,469</td>
<td>77</td>
</tr>
<tr>
<td>Freebase</td>
<td>49,947,845</td>
<td>3,041,722,635</td>
<td>26,507</td>
<td>37,781</td>
</tr>
<tr>
<td>Wikidata</td>
<td>15,602,060</td>
<td>65,993,797</td>
<td>23,157</td>
<td>1,673</td>
</tr>
<tr>
<td>NELL</td>
<td>2,006,896</td>
<td>432,845</td>
<td>285</td>
<td>425</td>
</tr>
<tr>
<td>OpenCyc</td>
<td>118,499</td>
<td>2,413,894</td>
<td>45,153</td>
<td>18,526</td>
</tr>
<tr>
<td>Google’s Knowledge Graph</td>
<td>570,000,000</td>
<td>18,000,000,000</td>
<td>1,500</td>
<td>35,000</td>
</tr>
<tr>
<td>Google’s Knowledge Vault</td>
<td>45,000,000</td>
<td>271,000,000</td>
<td>1,100</td>
<td>4,469</td>
</tr>
<tr>
<td>Yahoo! Knowledge Graph</td>
<td>3,443,743</td>
<td>1,391,054,990</td>
<td>250</td>
<td>800</td>
</tr>
</tbody>
</table>

• For a more detailed analysis of the public knowledge graphs see:
  – Ringler & Paulheim: “One Knowledge Graph to Rule Them All? Analyzing the Differences Between DBpedia, YAGO, Wikidata & co.”, presented at KI 2017
Knowledge Graph Refinement

- We see a lot of work by people who try to improve knowledge graphs
  - Rather than creating new ones
- Examples
  - Adding missing relations
  - Adding types to untyped entities
  - Identifying erroneous axioms
- We call this Knowledge Graph Refinement
Categorization of Approaches

• By goal
  – *Completeness*: Completion or augmentation
  – *Correctness*: Error detection or correction

• By target
  – Types
  – Relations
  – Literals
  – Interlinks
  – Schema

• Knowledge sources used
  – Internal: only the knowledge graph
  – External: using further sources

Given our instance-focused KG definition, we do not consider pure ontology refinement

Including humans
Categorization of Evaluation Methods

• Partial Gold Standard
  – For completion: collect all axioms that should exist for a set of entities
  – For correction: label a subset of axioms as correct or incorrect
  – some use external sources instead of manual labeling

• Knowledge Graph as Silver Standard
  – Usually used for completion
  – How well can we replicate the axioms in the knowledge graph?

• Retrospective Evaluation
  – Inspect the output of the completion/correction approach

• Orthogonal to the above: computational performance evaluation
  – Hard: measuring runtime and/or memory consumption
  – Soft: proof of concept on large knowledge graphs
## Categorization of Evaluation Methods

- Pros and cons of the three evaluation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial Gold Standard</td>
<td>• Reusable</td>
<td>• Cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Sampling for error detection</td>
</tr>
<tr>
<td>Knowledge Graph as Silver Standard</td>
<td>• Objective</td>
<td>• Prone to overfitting</td>
</tr>
<tr>
<td></td>
<td>• Large-scale evaluation feasible</td>
<td></td>
</tr>
<tr>
<td>Retrospective</td>
<td>• Suitable for unbalanced problems</td>
<td>• Not reusable</td>
</tr>
<tr>
<td></td>
<td>• In-depth analysis of results</td>
<td>• Direct comparison of approaches difficult</td>
</tr>
</tbody>
</table>
Survey Results

- 43 approaches analyzed in total
  - By goal and knowledge source

Surprise #2: Error correction hardly covered

Surprise #1
Survey Results

• 43 approaches analyzed in total
  – By goal and target
  – Multiple goals possible, but rare (4 out of 43)
Survey Results

• Evaluation methods
  – Multiple methods possible, but rare (3 out of 43)
  – Mixed methods for KG as silver standard: split validation, cross validation (various foldings), non-split validation

---

Completion Error Detection Both

8/11 are (easily) available :-)
Survey Results

• Evaluation Metrics
  – A clear trend to precision/recall (precision/total for retrospective)
  – With a lot of diversity in the long tail:
    accuracy, AUC-PR, AUC-ROC, RMSE, ...
Survey Results

• Computational performance
  – Relaxed counting: do the authors make a statement about performance?
  – Application on a large graph is also accepted as evidence for scalability

Note: “reported” is not the same as “it is scalable”!
Survey Results

- Knowledge Graphs used for evaluation
  - Clear tendency towards DBpedia
  - Not necessarily comparable!

So what do we know?

Where is Wikidata?
Survey Results – Further Observations

• We are talking about Knowledge Graphs
  – yet, genuinely graph-based approaches are rare
• Most approaches stop at the level of individual axioms
  – rarely any higher-level artifacts (e.g., common patterns of errors)
  – makes preservation of results difficult
• Completion approaches focus on existing entities
  – adding new entities (was) not very prominent
Final Words

• Make evaluation great again
  – Evaluate on multiple knowledge graphs
  – Carry out (systematic) scalability experiments
  – Evaluate also on synthetic datasets (Melo & Paulheim, ESWC 2017)

• Think about preserving results
  – Donate bugfixes to KG developers
  – Identify root causes, fix them (e.g., Paulheim & Gangemi, ISWC 2015)
  – Provide augmentations as LOD datasets
Knowledge Graph Refinement
A Survey of Approaches and Evaluation Methods
Semantic Web 8(3), 2017
Heiko Paulheim